AN INTRODUCTION TO STATISTICS

J. Calvin Berry
Mathematics Department
University of Louisiana at Lafayette

August 2008 edition
(corrected July 2014)

7242014

Table of Contents

Chapter 1
Introduction

1.1 Basic ideas

Statistical methods deal with properties of groups or aggregates. In many applications the entity of primary interest is an actual, physical group (population) of objects. These objects may be animate (*e.g.*, people or animals) or inanimate (*e.g.*, farm field plots, trees, or days). We will refer to the individual objects that comprise the group of interest as **units**. In certain contexts we may refer to the unit as a population unit, a sampling unit, an experimental unit, or a treatment unit.

In order to obtain information about a group of units we first need to obtain information about each of the units in the group. A **variable** is a measurable characteristic of an individual unit. Since our goal is to learn something about the group, we are most interested in the **distribution of the variable**, *i.e.*, the way in which the possible values of the variable are distributed among the units in the group.

When the units are actual, physical objects we define the **population** as the collection of all of the units that we are interested in. In most applications it is unnecessary or undesirable to examine the entire population. Thus we define a **sample** as a subset or part of the population for which we have or will obtain data. The collection of observed values of one or more variables corresponding to the individual units in the sample constitute the **data**. Once the data are obtained we can use the distributions of the variables among the units in the sample to characterize the sample itself and to make inferences or generalizations about the entire population, *i.e.*, inferences about the distributions of these variables among the units in the population.

When discussing the distribution of a variable we need to consider the structure possessed by the possible values of the variable. This leads to the following classification of variables into four basic types.

A **qualitative** variable (categorical variable) classifies a unit into one of several possible categories. The possible values of a qualitative variable are names for these categories. We can distinguish between two types of qualitative variables. A qualitative variable is said to be **nominal** if there is no inherent ordering among its possible values. The sex of a person (female or male) and the color of a person's eyes (blue, brown, *etc.*) are examples of nominal qualitative variables. If there is an inherent ordering of the possible values of a qualitative variable, then it is said to be **ordinal**. The classification of a student (freshman, sophomore, junior, or senior), the ranking of a unit with respect to several size classes (small, medium, or large), and the degree to which a person agrees with a statement

(recorded as strongly disagree, disagree, neutral, agree, or strongly agree) are examples of ordinal qualitative variables.

A **quantitative** variable (numerical variable) assigns a meaningful numerical value to a unit. Because the possible values of a quantitative variable are meaningful numerical quantities, they can be viewed as points on a number line. Therefore, it makes sense to talk about where the values of a quantitative variable are located on the number line, whether one value is larger than another, and how far apart two values are. If the possible values of a quantitative variable correspond to isolated points on the number line, then there is a discrete jump between adjacent possible values and the variable is said to be a **discrete** quantitative variable. The most common example of a discrete quantitative variable is a count such as the number of babies in a litter of animals or the number of plants in a field plot. If there is a continuous transition from one value of the variable to the next, then the variable is said to be a **continuous** quantitative variable. For a continuous quantitative variable there is always another possible value between any two possible values, no matter how close together the values are. In practice all quantitative variables are discrete in the sense that the observed values are rounded to a reasonable number of decimal places. Thus the distinction between a continuous quantitative variable and a discrete quantitative variable is often more conceptual than real. If a value of the variable represents a measurement of the size of a unit, such as height, weight, or length, or the amount of some quantity, then it is reasonable to think of the possible values of the variable as forming a continuum of values on the number line and to view the variable as continuous.

The values of ordinal variables are often recorded using numerical codes (ranks) such as 1:strongly disagree, 2:disagree, 3:neutral, 4:agree, or 5:strongly agree. This sort of coding of an ordinal variable does not make it quantitative. For example, the fact that these rankings are equally spaced points on the number line does not necessarily mean that the difference between 1:strongly disagree and 2:disagree is the same as the difference between 4:agree and 5:strongly agree. Therefore, the common practice of treating such ranking variables as quantitative must be used with caution and the fact that the values of the variable are simply ranks must be taken into account when interpreting an analysis of such a ranking variable.

We can also classify variables with respect to the roles they play in a statistical analysis. That is, we can distinguish between response variables and explanatory variables. A **response variable** is a variable that measures the response of a unit to natural or experimental stimuli. A response variable provides us with a measurement or observation that characterizes a unit with respect to a characteristic of primary interest. An **explanatory variable** is a variable that can be used to explain, in whole or in part, how a unit responds

to natural or experimental stimuli. This terminology is clearest in the context of an experimental study. Consider an experiment where a unit is subjected to a treatment (some combination of conditions) and the response of the unit to the treatment is recorded. A variable that describes the treatment conditions is called an explanatory variable, since it may be used to explain the outcome of the experiment. A variable that measures the outcome of the experiment is called a response variable, since it measures the response of the unit to the treatment. An explanatory variable may also be used to subdivide a group so that the distributions of a response variable can be compared among subgroups.

In some applications, such as experimental studies, the population is best viewed as a hypothetical population of values of one or more variables. For example, suppose that we are interested in the effects of an alternative diet on weight gain in some population of experimental animals. We might conduct an experiment by randomly assigning animals to two groups; feeding one group a standard diet and the other group the alternative diet; and then recording the weight gained by each animal over some fixed period of time. In this example we can envision two hypothetical populations of weight gains: The population of weight gains we would have observed if all of the animals were given the standard diet; and, the population of weight gains we would have observed if all of the animals were given the alternative diet.

Statistics is often defined as a collection of methods for collecting, describing, and drawing conclusions from data. Methods for collecting data fall under the heading of sampling and experimentation; we will discuss these topics in Chapter 4. Descriptive statistical methods are used to describe the distributions of the values of variables among the units in a sample, *i.e.*, to gain insight about the sample. We will discuss univariate (one variable) descriptive statistical methods in Chapters 2 and 3 and bivariate (two variables) descriptive methods in Chapter 9. Inferential statistical methods are used to make inferences or generalizations, based on the data from the sample, about the distributions of the values of variables among the units in the population, *i.e.*, to gain insight about the population based on information obtained from the sample. Inferential methods are probabilistic in the sense that they are based on probability models for the distributions of variables. The majority of this book deals with inferential statistics; probability models are introduced in Chapter 4a.

We will use the following simple example to clarify the concepts and definitions from above. The data presented in Table 1 were collected on the first day of classes during the Spring 1999 semester. These data provide information about the 67 students who were present on the first day of classes for two sections of the statistics course Stat 214 at the University of Louisiana at Lafayette. Aside from being grouped by section, the data are

Table 1. Statistics 214 class data, spring 1999.

line	section	classification	sex	age	height	weight	siblings	BMI
1	1	senior	male	21	69	170	1	25.10
2	1	junior	male	25	71	165	3	23.01
3	1	junior	female	25	62	160	2	29.26
4	1	freshman	male	18	72	162	1	21.97
5	1	junior	female	22	63	170	1	30.11
6	1	freshman	female	18	64	110	2	18.88
7	1	freshman	female	18	60	103	1	20.11
8	1	freshman	female	18	68	135	3	20.52
9	1	sophomore	female	19	62	105	5	19.20
10	1	freshman	male	18	74	190	2	24.39
11	1	sophomore	female	20	70	150	1	21.52
12	1	senior	female	21	61	116	1	21.92
13	1	freshman	female	18	65	150	3	24.96
14	1	freshman	female	19	64	140	4	24.03
15	1	freshman	male	18	68	130	2	19.76
16	1	freshman	female	18	63	110	2	19.48
17	1	sophomore	female	21	62	125	1	22.86
18	1	freshman	female	18	63	115	2	20.37
19	1	freshman	female	19	64	135	3	23.17
20	1	freshman	female	18	69	155	1	22.89
21	1	sophomore	female	20	65	110	2	18.30
22	1	sophomore	female	19	68	140	1	21.28
23	1	freshman	female	47	66	110	1	17.75
24	1	sophomore	female	20	70	145	2	20.80
25	1	freshman	female	20	61	140	5	26.45
26	1	freshman	female	18	63	180	0	31.88
27	1	junior	male	22	70	175	2	25.11
28	1	freshman	female	18	63	120	1	21.25
29	1	senior	female	22	68	170	2	25.85
30	1	freshman	female	18	66	125	3	20.17
31	1	junior	male	22	75	205	2	25.62
32	1	freshman	female	18	67	110	1	17.23
33	1	senior	male	22	68	135	1	20.52
34	1	senior	female	22	64	185	2	31.75
35	1	freshman	female	41	61	96	1	18.14
36	1	junior	female	22	59	95	5	19.19

This table is continued on the next page.

Table 1. Statistics 214 class data (continuation).

line	section	classification	sex	age	height	weight	siblings	BMI
37	2	junior	female	20	66	110	1	17.75
38	2	junior	male	20	72	180	1	24.41
39	2	junior	female	21	66	120	1	19.37
40	2	sophomore	female	21	61	105	3	19.84
41	2	freshman	female	18	68	134	7	20.37
42	2	freshman	female	28	66	130	4	20.98
43	2	sophomore	female	26	64	135	4	23.17
44	2	sophomore	female	19	64	117	1	20.08
45	2	freshman	female	20	66	140	4	22.59
46	2	junior	female	20	64	130	1	22.31
47	2	senior	female	48	66	140	3	22.59
48	2	junior	female	22	67	115	2	18.01
49	2	sophomore	female	19	66	170	2	27.44
50	2	freshman	male	18	66	190	3	30.66
51	2	sophomore	female	21	67	135	4	21.14
52	2	freshman	female	20	68	140	2	21.28
53	2	sophomore	female	19	62	115	2	21.03
54	2	sophomore	female	20	60	110	2	21.48
55	2	freshman	male	18	72	185	3	25.09
56	2	senior	male	23	72	190	2	25.77
57	2	senior	male	24	69	170	4	25.10
58	2	junior	male	21	72	140	3	18.98
59	2	junior	female	20	65	112	2	18.64
60	2	junior	female	21	62	130	1	23.77
61	2	freshman	female	18	64	120	1	20.60
62	2	sophomore	female	25	66	145	2	23.40
63	2	junior	male	19	65	156	6	25.96
64	2	freshman	female	18	67	125	0	19.58
65	2	junior	female	44	66	165	4	26.63
66	2	sophomore	male	19	71	155	3	21.62
67	2	sophomore	female	19	62	133	2	24.32

presented in no particular order. These data correspond to a convenience sample of students which may or may not be representative of some larger population of students. Values are provided for eight variables: the section the student was registered in (1 or 2); the classification of the student (freshman, sophomore, junior, or senior); the sex of the student (female or male); the age of the student (in years); the height of the student (in inches); the weight of the student (in pounds); the number of siblings the student had $(0, 1, 2, \ldots)$; and the body mass index (BMI) of the student. The derived or constructed

variable BMI (in kg/m^2) is the weight of the student (in kilograms) divided by the square of the student's height (in meters).

The sex of a student (with possible values of female and male) and the section the student was registered in (with possible values 1 and 2) are nominal qualitative variables. The classification of a student (with possible values of freshman, sophomore, junior, and senior) is an ordinal qualitative variable. The other variables are quantitative. The number of siblings that the student had (with possible values of $0, 1, 2, \ldots$) is inherently discrete. The other quantitative variables, age (in years), height (in inches), weight (in pounds), and BMI (in kg/m^2) can be viewed as continuous variables.

The section that the student was registered in was included as a potentially interesting explanatory variable which could be used to divide these students into two subgroups so that the distributions of the other variables for these subgroups could be compared. For an initial analysis of these data we would probably view all of the other variables as response variables. That is, a first analysis might consist of examination of the distributions of these response variables for the entire group or comparisons of these distributions by section. After looking at the overall distributions of the variables we might also want to group the students by sex (treat the sex of a student as an explanatory variable) and compare the distributions of height, weight, and BMI for the two sexes.

1.2 Some examples

This section contains a collection of examples which will be used in exercises and as examples in the sequel.

Example. DiMaggio and Mantle. Joe DiMaggio and Mickey Mantle were two well known baseball players. DiMaggio played center field for the New York Yankees for 13 years and was succeeded by Mantle who played center field for 18 years. There has been some argument about which of these two players was better at hitting home runs. The data given in Table 2 are the numbers of home runs hit by the player during each of the seasons he played. For each player these numbers of home runs are listed in order by the seasons he played.

Table 2. Home run data.

Joe DiMaggio:	29 46 32 30 31 30 21 25 20 39 14 32 12
Mickey Mantle:	13 23 21 27 37 52 34 42 31 40 54 30 15 35 19 23 22 18

Example. Weed seeds. C. W. Leggatt counted the number of seeds of the weed *potentilla* found in 98 quarter–ounce batches of the grass *Phleum praetense*. This example is taken from Snedecor and Cochran, *Statistical Methods*, Iowa State, (1980), 198; the original source is C. W. Leggatt, *Comptes rendus de l'association international d'essais de*

semences, **5** (1935), 27. The 98 observed numbers of weed seeds, which varied from 0 to 7, are summarized in Table 3.

Table 3. Weed seed frequency distribution.

number of seeds	frequency
0	37
1	32
2	16
3	9
4	2
5	0
6	1
7	1
total	98

Example. Vole reproduction. An investigation was conducted to study reproduction in laboratory colonies of voles. This example is taken from Devore and Peck, *Statistics*, (1997), 33; the original reference is the article "Reproduction in laboratory colonies of voles", *Oikos*, (1983), 184. The data summarized in Table 4 are the numbers of babies in 170 litters born to voles in a particular laboratory.

Table 4. Vole baby frequency distribution.

number of babies	frequency
1	1
2	2
3	13
4	19
5	35
6	38
7	33
8	18
9	8
10	2
11	1
total	170

Example. Wooly–bear caterpillar cocoons. A study was conducted to investigate the relationship between air temperature and the temperature inside a wooly–bear

caterpillar cocoon. It seems quite reasonable to expect the temperature inside a cocoon to be higher than the air temperature (outside the cocoon). The data given in Table 5 are pairs of air and cocoon temperatures made on 12 days at a location in the high arctic region. Each cocoon temperature is actually the average of two cocoon temperatures. This example comes from Kevan, P.C., T.S. Jensen, and J.D. Shorthouse, "Body temperatures and behaviorial thermoregulation of high arctic wooly–bear caterpillars and pupae (*Gynaephora rossii,* Lymantridae: Lepidoptera) and the importance of sunshine", *Arctic and Alpine Research*, **14**, (1982).

Table 5. Wooly–bear temperature data.

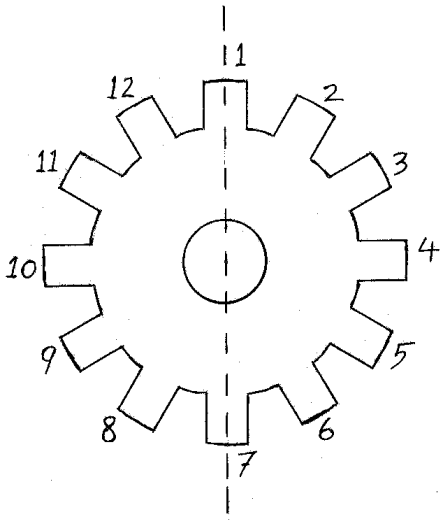
Day	Cocoon temp	Air temp	Day	Cocoon temp	Air temp
1	15.1	10.4	7	3.6	1.7
2	14.6	9.2	8	5.3	2.0
3	6.8	2.2	9	7.0	3.0
4	6.8	2.6	10	7.1	3.5
5	8.0	4.1	11	9.6	4.5
6	8.7	3.7	12	9.5	4.4

Example. Homophone confusion and Alzheimer's disease. A study was conducted to investigate the relationship between Alzheimer's disease and homophone spelling confusion. A homophone pair is a pair of words with the same pronunciation having different meanings and spellings. Twenty patients with Alzheimer's disease were asked to spell 24 homophone pairs (given in random order) and the number of homophone confusions, *e.g.* spelling doe given the context bake bread dough, was recorded for each patient. One year later, the same patients were again asked to spell the same 24 homophone pairs and the number of homophone confusions was again recorded. The data given in Table 6 are the numbers of homophone confusions at the two times of measurement for the 20 Alzheimer's patients. This example comes from Neils, J., D.P. Roeltgen, and F. Constantinidou, "Decline in homophone spelling associated with loss of semantic influence on spelling in Alzheimer's disease", *Brain and Language*, **49**, (1995).

Table 6. Alzheimer's homophone confusion data.

Patient	Time 1	Time 2	Patient	Time 1	Time 2
1	5	5	11	7	10
2	1	3	12	0	3
3	0	0	13	3	9
4	1	1	14	5	8
5	0	1	15	7	12
6	2	1	16	10	16
7	5	6	17	5	5
8	1	2	18	6	3
9	0	9	19	9	6
10	5	8	20	11	8

Example. Gear tooth strength. The data used in this example were published by B. Gunter, "Subversive data analysis, Part II: More graphics, including my favorite example", *Quality Progress*, Nov., 1988, 77–78. This description is adapted from Wild and Seber, *Chance Encounters*, Wiley, (2000), 118. These data concern gear blanks purchased by the Ford Motor Company. Ford engineers found that the teeth on these gears were breaking at too low a stress. The data given below are the impact strengths (in lb–ft) required to break a gear tooth. Each gear had 12 equally spaced teeth. The position



numbers for these teeth begin with 1 at 12 o'clock and proceed in a clockwise direction. The tooth positions are important since they are related to the position of the tooth in the mold used to make the gear. Teeth 1 and 7 are distinguishable; but, teeth located symmetrically about a line drawn through positions 1 and 7 are not, since these positions depend on which face of the gear is upward. Thus, observations for pairs of teeth in a symmetrical position about a line through position 1 and 7 are grouped in Table 7.

Table 7. Gear tooth strength data.

gear position						
1	**2 & 12**	**3 & 11**	**4 & 10**	**5 & 9**	**6 & 8**	**7**
1976	2425	2228	2186	2228	2431	2287
1916	2000	2347	2521	2180	2250	2275
2090	2251	2251	2156	2114	2311	1946
2000	2096	2222	2216	2365	2210	2150
2323	2132	1940	2593	2299	2329	2228
1904	1964	1904	2204	2072	2263	1695
2048	1750	1820	2228	2323	2353	2000
2222	2018	2012	2198	2449	2251	2006
2048	1766	2204	2150	2300	2275	1945
2174		2144	2311	2078	1958	2006
1976		2305	2102	2150	2185	2209
2138		2042	2138	2377		2216
2455		2120	1982	2108		1934
1886		2419	2042	2257		1904
2246		2162	2030	2383		1958
2287		2251	2216	2323		1964
2030		2222	2305	2246		2066
2210			2204	2251		2222
2084			2198	2156		2066
2383			2204	2419		1964
2132			2162	2329		2150
2210			2120	2198		2114
2222			2108	2269		2125
1766			2030	2287		2210
2078			2180	2330		1588
1994			2251	2329		2234
2198			2210	2228		2210
2162			2216			2156
1874			2168			2204
2132			2210			1641
2108			2341			2263
1892			2000			2120
1671			2132			2156

Example. Immigrants to the United States. The data concerning immigrants admitted to the United States summarized by decade as raw frequency distributions in Table 8 were taken from the *2002 Yearbook of Immigration Statistics*, USCIS,

(`www.uscis.gov`). Immigrants for whom the country of last residence was unknown are omitted.

Table 8. Region of last residence for immigrants to USA.

	period		
region	**1931–1940**	**1961–1970**	**1991–2000**
Europe	347,566	1,123,492	1,359,737
Asia	16,595	427,692	2,795,672
North America	130,871	886,891	2,441,448
Caribbean	15,502	470,213	978,787
Central America	5,861	101,330	526,915
South America	7,803	257,940	539,656
Africa	1,750	28,954	354,939
Oceania	2,483	25,122	55,845
total	528,431	3,321,634	9,052,999

Example. Cholesterol levels in Guatemalans. This example is taken from Devore and Peck, *Statistics*, 3 ed., (1997), Duxbury, p. 23. The original source is "The Blood Viscosity of Various Socioeconomic Groups in Guatemala" in *The American Journal of Clinical Nutrition*, Nov., 1964, 303–307. The Institute of Nutrition of Central America and Panama measured the serum total cholesterol levels for a group of 49 adult, low–income rural Guatemalans and for a group of 45 adult, high–income urban Guatemalans. The serum total cholesterol levels (in mg/dL) are provided in Table 9.

Table 9. Guatemalan cholesterol data.

Rural group cholesterol levels (in mg/dL).

95	108	108	114	115	124	129	129	131	131
135	136	136	139	140	142	142	143	143	144
144	145	146	148	152	152	155	157	158	158
162	165	166	171	172	173	174	175	180	181
189	192	194	197	204	220	223	226	231	

Urban group cholesterol levels (in mg/dL).

133	134	155	170	175	179	181	184	188	189
190	196	197	199	200	200	201	201	204	205
205	205	206	214	217	222	222	227	227	228
234	234	236	239	241	242	244	249	252	273
279	284	284	284	330					

1.3 Exercises

For each of the examples in Section 1.2 define or identify the following:

1. The unit.

2. The group(s) of interest.

3. The variable(s) and the possible values of the variable(s).

4. The type of variable(s) (nominal qualitative, ordinal qualitative, discrete quantitative, or continuous quantitative).

Chapter 2
Descriptive Statistics I: Tabular and Graphical Summary

2.1 Generalities

Consider the problem of using data to learn something about the characteristics of the group of units which comprise the sample. Recall that the distribution of a variable is the way in which the possible values of the variable are distributed among the units in the group of interest. A variable is chosen to measure some characteristic of the units in the group of interest; therefore, the distribution of a variable contains all of the available information about the characteristic (as measured by that variable) for the group of interest. Other variables, either alone or in conjunction with the primary variable, may also contain information about the characteristic of interest. A meaningful summary of the distribution of a variable provides an indication of the overall pattern of the distribution and serves to highlight possible unusual or particularly interesting aspects of the distribution. In this chapter we will discuss tabular and graphical methods for summarizing the distribution of a variable and in the following chapter we will discuss numerical summary methods.

Generally speaking, it is hard to tell much about the distribution of a variable by examining the data in raw form. For example, scanning the Stat 214 data in Table 1 of Chapter 1 it is fairly easy to see that the majority of these students are female; but, it is hard to get a good feel for the distributions of the variables which have more than two possible values. Therefore, the first step in summarizing the distribution of a variable is to tabulate the frequencies with which the possible values of the variable appear in the sample. A **frequency distribution** is a table listing the possible values of the variable and their frequencies (counts of the number of times each value occurs). A frequency distribution provides a decomposition of the total number of observations (the sample size) into frequencies for each possible value. In general, especially when comparing two distributions based on different sample sizes, it is preferable to provide a decomposition in terms of relative frequencies. A **relative frequency distribution** is a table listing the possible values of the variable along with their relative frequencies (proportions). A relative frequency distribution provides a decomposition of the total relative frequency of one (100%) into proportions or relative frequencies (percentages) for each possible value.

Many aspects of the distribution of a variable are most easily communicated by a graphical representation of the distribution. The basic idea of a graphical representation of a distribution is to use area to represent relative frequency. The total area of the graphical representation is taken to be one (100%) and sections with area equal to the relative frequency (percentage) of occurrence of a value are used to represent each possible value of the variable.

2.2 Describing qualitative data

In this section we consider tabular and graphical summary of the distribution of a qualitative variable. Assuming that there are not too many distinct possible values for the variable, we can summarize the distribution using a table of possible values along with the frequencies and relative frequencies with which these values occur in the sample. Recall that a frequency distribution provides a decomposition of the total number of observations (the sample size) into frequencies for each possible value; and, a relative frequency distribution provides a decomposition of the total relative frequency of one (100%) into proportions or relative frequencies (percentages) for each possible value. In most applications, and especially for comparisons of distributions, it is better to use relative frequencies rather than raw frequencies. When forming a relative frequency distribution for a nominal qualitative variable we can list the possible values of the variable in any convenient order. On the other hand, the possible values of an ordinal qualitative variable should always be listed in proper order to avoid possible confusion when reading the table.

Table 1 contains the frequency distributions and relative frequency distributions of the two qualitative variables sex and classification for the Stat 214 example. Notice that the possible values of the ordinal variable, classification of the student, are listed in proper order to avoid possible confusion when reading the table.

Table 1. Relative frequency distributions for the sex and classification distributions in the Stat 214 example.

Sex distribution.

sex	frequency	relative frequency
female	51	.761
male	16	.239
total	67	1.000

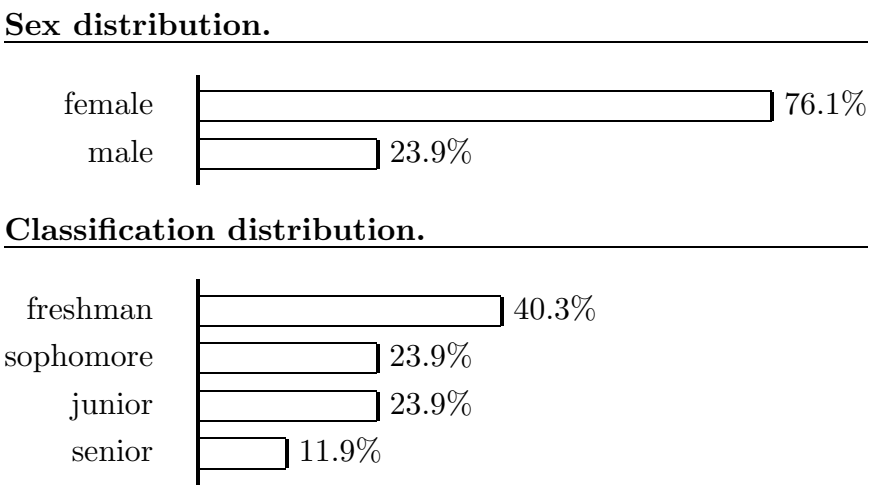
Classification distribution.

classification	frequency	relative frequency
freshman	27	.403
sophomore	16	.239
junior	16	.239
senior	8	.119
total	67	1.000

Bar graphs summarizing the sex and classification distributions for the Stat 214 example are given in Figure 1. Again, to avoid confusion, the possible classification values are presented in proper order. A **bar graph** consists of a collection of bars (rectangles) such that the combined area of all the bars is one (100%) and the area of a particular bar is the relative frequency of the corresponding value of the variable. Two other common forms for such a graphical representation are segmented bar graphs and pie graphs. A

segmented bar graph consists of a single bar of area one (100%) that is divided into segments with a segment of the appropriate area for each observed value of the variable. A segmented bar graph can be obtained by joining the separate bars of a bar graph. If the bar of the segmented bar graph is replaced by a circle, the result is a pie graph or pie chart. In a **pie graph** or pie chart the interior of a circle (the pie) is used to represent the total area of one (100%); and the pie is divided into slices of the appropriate area or relative frequency, with one slice for each observed value of the variable.

Figure 1. Bar graphs for the sex and classification distributions in the Stat 214 example.



For these two sections of Stat 214 it is clear that a large majority (76.1%) of the students are female. There are two simple explanations for the predominance of females in this sample: The proportion of females among all undergraduate students at this university is roughly 65%; and, the majors which require this particular course traditionally attract more females than than males. Turning to the classification distribution, it is clear that relatively few (11.9%) of the students in these sections are seniors. This aspect of the classification distribution is not surprising, since Stat 214 is a 200 level (nominally sophomore) course. It is somewhat surprising, for a 200 level course, to find that the most common classification is freshman (40.3%). We might wonder whether these characteristics of the sex and classification distributions are applicable to both sections of Stat 214. The section variable can be used as an indicator variable (a qualitative explanatory variable used for grouping observations) to divide the sample of 67 students into the group of 36 students in section 1 and the group of 31 students in section 2. The sex and classification distributions for these two sections are summarized in Tables 2 and 3 and Figures 2 and 3. Notice that, because of rounding of the relative frequencies, the sum of the relative frequencies is not exactly one in Table 3.

Table 2. Relative frequency distributions for sex, by section.

section 1

sex	frequency	relative frequency
female	28	.778
male	8	.222
total	36	1.000

section 2

sex	frequency	relative frequency
female	23	.742
male	8	.258
total	31	1.000

Figure 2. Bar graphs for sex, by section.

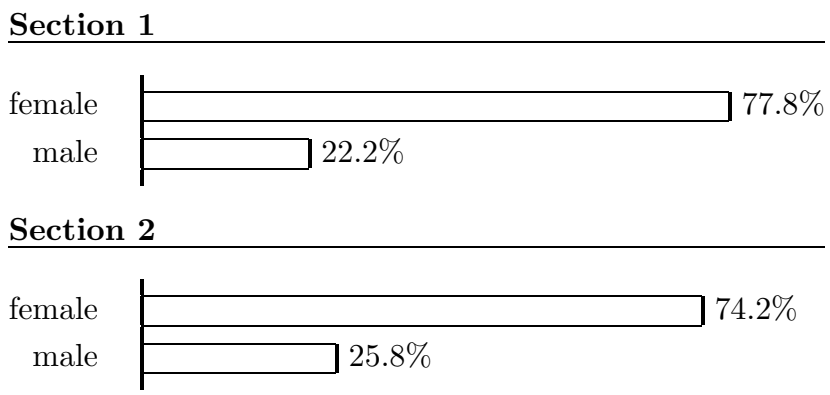


Table 3. Relative frequency distributions for classification, by section.

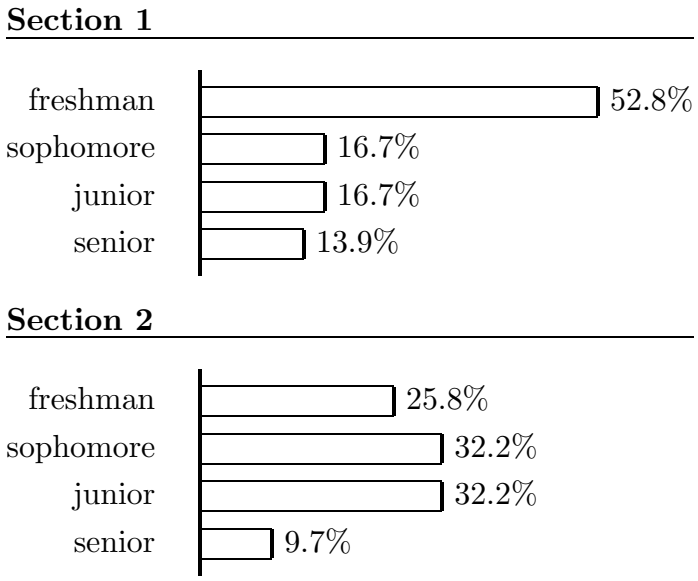
section 1

classification	frequency	relative frequency
freshman	19	.528
sophomore	6	.167
junior	6	.167
senior	5	.139
total	36	1.001

section 2

classification	frequency	relative frequency
freshman	8	.258
sophomore	10	.322
junior	10	.322
senior	3	.097
total	31	.999

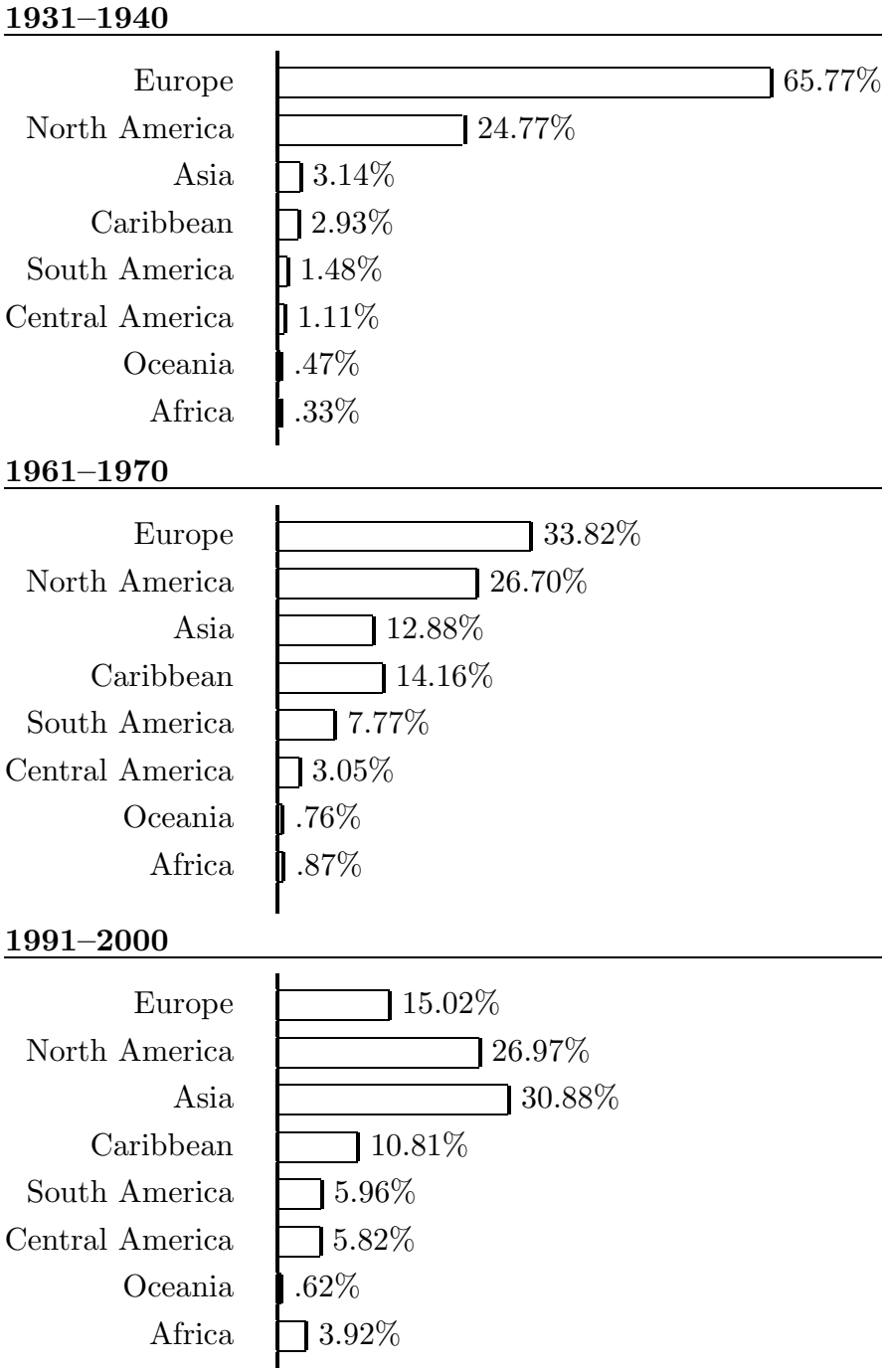
Figure 3. Bar graphs for classification, by section.



The sex distributions for the two sections are essentially the same; in both sections approximately 75% of the students are female. On the other hand, there is a clear difference between the two classification distributions. The section 1 classification distribution is similar to the combined classification distribution with a very large proportion of freshmen. In fact, more than half (52.8%) of the students in section 1 are freshmen. This predominance of freshmen does not happen in section 2 where there is not a single dominant classification value. The most common classifications for section 2 are sophomore and junior, with each of these classifications accounting for 32.2% of the students. In summary, we find that for section 1 the majority of students (52.8%) are freshmen but that for section 2 the majority (64.4%) of the students are sophomores (32.2%) or juniors (32.2%). It is interesting to notice that in both sections the proportions of sophomores and juniors are equal.

Example. Immigrants to the United States. The data concerning immigrants admitted to the United States summarized by decade as raw frequency distributions in Section 1.2 were taken from the *2002 Yearbook of Immigration Statistics*, USCIS, (`www.uscis.gov`). Immigrants for whom the country of last residence was unknown are omitted. For this example a unit is an individual immigrant and these data correspond to a census of the entire population of immigrants, for whom the country of last residence was known, for these decades. Because the region of last residence of an immigrant is a nominal variable and its values do not have an inherent ordering, the values in the bar graphs (and relative frequency distributions) in Figure 4 have been arranged so that the percentages for the 1931–1940 decade are in decreasing order.

Figure 4. Region of last residence for immigrants to USA, by decade.



Two aspects of the distributions of region of origin of immigrants which are apparent in these bar graphs are: The decrease in the proportion of immigrants from Europe; and, the increase in the proportion of immigrants from Asia. In 1931–1940 a large majority (65.77%) of the immigrants were from Europe but for the later decades this proportion steadily decreases. On the other hand, the proportion of Asians (only 3.14% in 1931–1940)

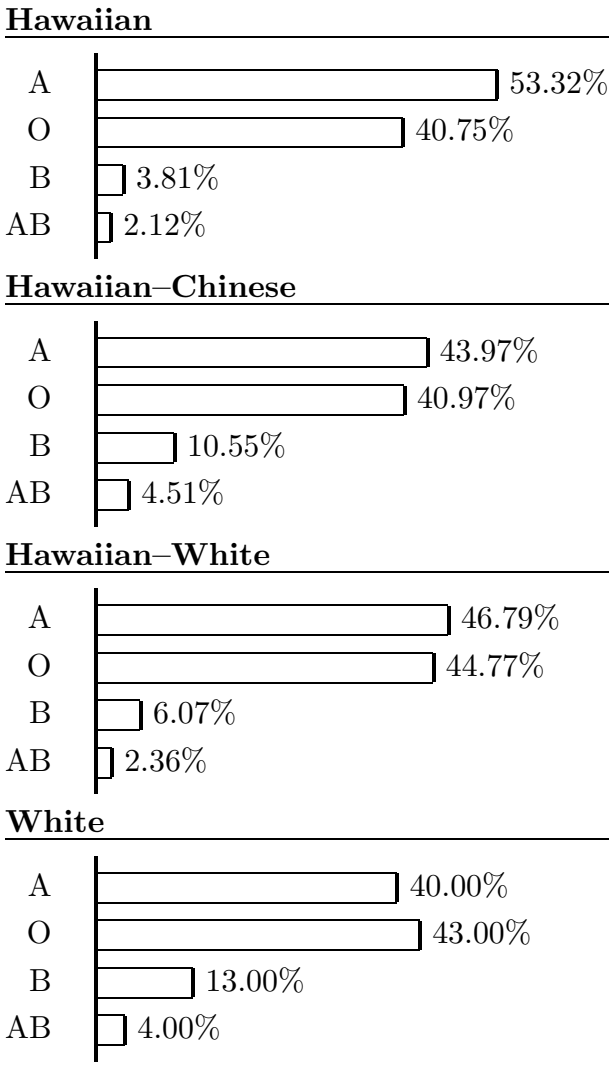
steadily increases to 30.88% in 1991–2000. Also note that the proportion of immigrants from North America is reasonably constant for these three decades. The patterns we observe in these distributions may be attributable to several causes. Political, social, and economic pressures in the region of origin of these people will clearly have an impact on their desire to immigrate to the US. Furthermore, political pressures within the US have effects on immigration quotas and the availability of visas.

Example. Hawaiian blood types. This example is based on the description in Moore and McCabe, *Introduction to the Practice of Statistics*, Freeman, (1993) of a study discussed in A.E. Mourant, *et al.*, *The Distribution of Blood Groups and Other Polymorphisms*, Oxford University Press, London, 1976. The Blood Bank of Hawaii cross–classified 145,057 individuals according to their blood type (A, AB, B, O) and their ethnic group (Hawaiian, Hawaiian–Chinese, Hawaiian–White, White). The frequencies for each of the 16 combinations of the 4 levels of these two qualitative variables are given in Table 4. This sample of individuals is most likely a convenience sample of blood donors. We will use the classification of an individual by ethnic group as an explanatory (indicator) variable and consider the (conditional) distributions of the nominal qualitative variable blood type for the four ethnic groups. The four columns corresponding to the ethnic groups in Table 4 provide the conditional (frequency) distributions for these groups. Because there is no inherent ordering among the blood types the arrangement of the blood types in Figure 5 is such that the percentages for the Hawaiian group are in decreasing order. The conditional (relative frequency) distributions of blood type for the ethnic groups summarized in Figure 5 clarify the differences in the distributions of blood type for these four ethnic groups.

Table 4. Blood type and ethnic group observed frequencies.

	ethnic group				
blood type	Hawaiian	Hawaiian–Chinese	Hawaiian–White	White	total
A	2490	2368	4671	50008	59537
O	1903	2206	4469	53759	62337
B	178	568	606	16252	17604
AB	99	243	236	5001	5579
total	4670	5385	9982	125020	145057

Figure 5. Conditional distributions of blood type by ethnic group.



2.3 Describing discrete quantitative data

The tabular representations used to summarize the distribution of a discrete quantitative variable, *i.e.,* the frequency and relative frequency distributions, are defined the same as they were for qualitative data. Since the values of a quantitative variable can be viewed as points on the number line, we need to indicate this structure in a tabular representation. In the frequency or relative frequency distribution the values of the variable are listed in order and all possible values within the range of the data are listed even if they do not appear in the data.

First consider the distribution of the number of siblings for the Stat 214 example. The relative frequency distribution for the number of siblings is given in Table 5.

Table 5. Relative frequency distribution for number of siblings.

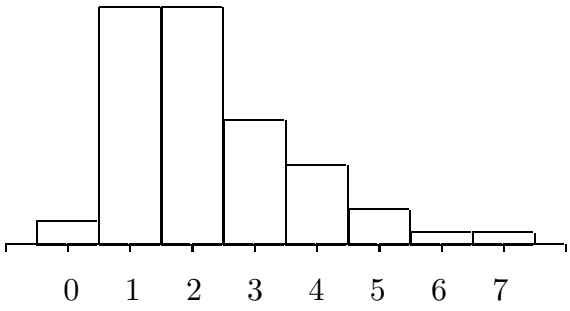
number of siblings	frequency	relative frequency
0	2	.030
1	21	.313
2	21	.313
3	11	.164
4	7	.104
5	3	.045
6	1	.015
7	1	.015
total	67	.999

We will use a graphical representation called a histogram to summarize the distribution of a discrete quantitative variable. Like the bar graph we used to represent the distribution of a qualitative variable, the histogram provides a representation of the distribution of a quantitative variable using area to represent relative frequency. A **histogram** is basically a bar graph modified to indicate the location of the observed values of the variable on the number line. For ease of discussion we will describe histograms for situations where the possible values of the discrete quantitative variable are equally spaced (the distance between any two adjacent possible values is always the same).

Consider the histogram for the number of siblings for the Stat 214 example given in Figure 6. This histogram is made up of rectangles of equal width, centered at the observed values of the variable. The heights of these rectangles are chosen so that the area of a rectangle is the relative frequency of the corresponding value of the variable. There is not a gap between two adjacent rectangles in the histogram unless there is an unobserved possible value of the variable between the corresponding adjacent observed values. For this example there are no gaps; but, there is a gap in the histogram of Figure 8.

In this histogram we are using an interval of values on the number line to indicate a single value of the variable. For example, the rectangle centered over 1 in the histogram of Figure 6 represents the relative frequency of a student having 1 sibling; but its base extends from .5 to 1.5 on the number line. Because it is impossible for the number of siblings to be strictly between 0 and 1 or strictly between 1 and 2, we are identifying the entire interval from .5 to 1.5 on the number line with the actual value of 1. This identification of an interval of values with the possible value at the center of the interval eliminates gaps in the histogram that would incorrectly suggest the presence of unobserved, possible values.

Figure 6. Histogram for number of siblings.

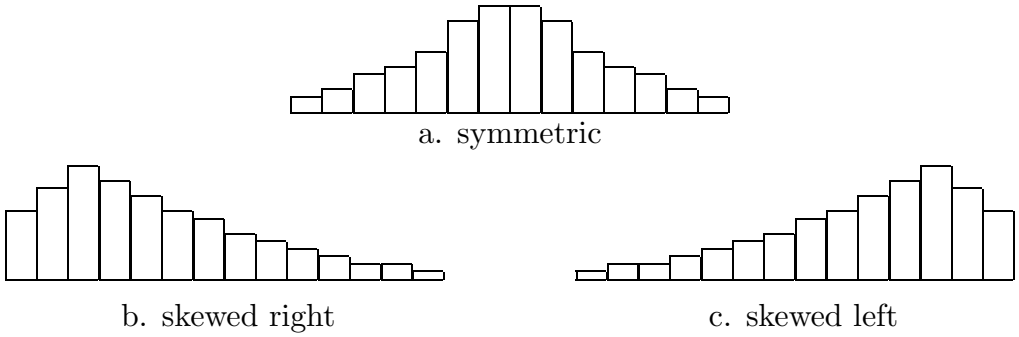


The histogram for the distribution of the number of siblings for the Stat 214 example in Figure 6 has a mound shaped appearance with a single peak over the values 1 and 2, indicating that the most common number of siblings for a student in this group is either 1 or 2. In fact, 31.3% of the students in this group have one sibling and 31.3% have two siblings. It is relatively unusual for a student in this group to be an only child (3%) or to have 5 or more siblings (7.5%).

The histogram of Figure 6, or the associated distribution, is not symmetric. That is, the histogram (distribution) is not the same on the left side (smaller values) of the peak over the values 1 and 2 as it is on the right side (larger values). This histogram or distribution is said to be skewed to the right. The concept of a distribution being skewed to the right is often explained by saying that the right "tail" of the distribution is "longer" than the left "tail". That is, the area in the histogram is more spread out along the number line on the right than it is on the left. For this example, the smallest 25% of the observed values are zeros and ones while the largest 25% of the observed values include values ranging from three to seven. In the present example we might say that there is essentially no left tail in the distribution.

The number of siblings histogram and the histograms for the next three examples discussed below are examples of a very common type of histogram (distribution) which is mound shaped and has a single peak. This type of distribution arises when there is a single value (or a few adjacent values) which occurs with highest relative frequency, causing the histogram to have a single peak at this location, and when the relative frequencies of the other values taper off (decrease) as we move away from the location of the peak. Three examples of common mound shaped distributions with a single peak are provided in Figure 7. The **symmetric** distribution is such that the histogram has two mirror image halves. The **skewed** distributions are more spread out along the number line on one side (the direction of the skewness) than they are on the other side.

Figure 7. Mound shaped histograms with a single peak.



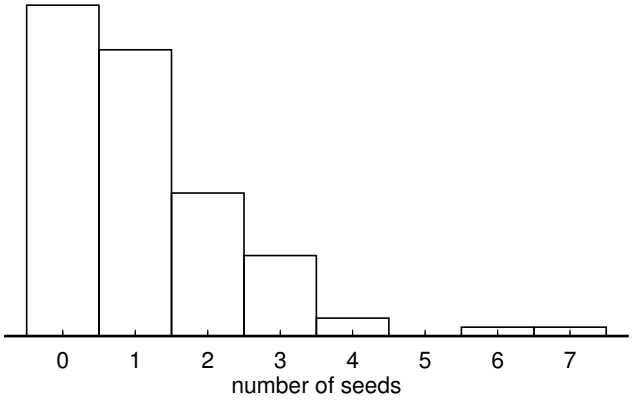
A distribution with a single peak is said to be unimodal, indicating that it has a single mode. The formal definition of a **mode** is a value which occurs with highest frequency. In practice, if two adjacent values are both modes, as are 1 and 2 in the number of siblings example, then we would still say that the distribution is unimodal. Some distributions are bimodal (or multimodal) in the sense of having two distinct modes which are separated by an interval of values with lower relative frequencies. The degree of cloudiness example below provides an example of an extreme version of a bimodal distribution. A more common situation when a bimodal distribution might arise is when the sample under study is a mixture of two subgroups (say males and females) with distinct and well separated modes.

Example. Weed seeds. C. W. Leggatt counted the number of seeds of the weed *potentilla* found in 98 quarter–ounce batches of the grass *Phleum praetense*. This example is taken from Snedecor and Cochran, *Statistical Methods*, Iowa State, (1980), 198; the original source is C. W. Leggatt, *Comptes rendus de l'association international d'essais de semences*, **5** (1935), 27. The 98 observed numbers of weed seeds, which varied from 0 to 7, are summarized in the relative frequency distribution of Table 6 and the histogram of Figure 8. In this example a unit is a batch of grass and the number of seeds in a batch is a discrete quantitative variable with possible values of $0, 1, 2, \ldots$. The distribution of the number of weed seeds is mound shaped with a single peak at zero and it is skewed to the right. The majority of these batches of grass have a small number of weed seeds; but, there are a few batches with relatively high numbers of weed seeds.

Table 6. Weed seed relative frequency distribution.

number of seeds	frequency	relative frequency
0	37	.3776
1	32	.3265
2	16	.1633
3	9	.0918
4	2	.0204
5	0	.0000
6	1	.0102
7	1	.0102
total	98	1.0000

Figure 8. Histogram for number of weed seeds.

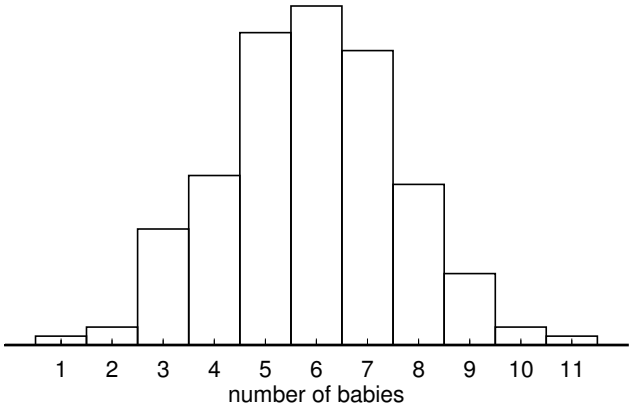


Example. Vole reproduction. An investigation was conducted to study reproduction in laboratory colonies of voles. This example is taken from Devore and Peck, *Statistics*, (1997), 33; the original reference is the article "Reproduction in laboratory colonies of voles", *Oikos*, (1983), 184. The data summarized in Table 7 and Figure 9 are the numbers of babies in 170 litters born to voles in a particular laboratory. In this example a unit is a litter of voles and the number of babies in a vole litter is a discrete quantitative variable with possible values of $1, 2, 3, \ldots$. In this example we see that the distribution of the number of vole babies is mound shaped with a single peak at 6 and it is reasonably symmetric. For these vole litters the majority of the litters have around 6 babies. There are a few litters with relatively small numbers of babies and there are a few with relatively large numbers of babies.

Table 7. Vole baby relative frequency distribution.

number of babies	frequency	relative frequency
1	1	.0059
2	2	.0118
3	13	.0765
4	19	.1118
5	35	.2059
6	38	.2235
7	33	.1941
8	18	.1059
9	8	.0471
10	2	.0118
11	1	.0059
total	170	1.0002

Figure 9. Histogram for number of vole babies.

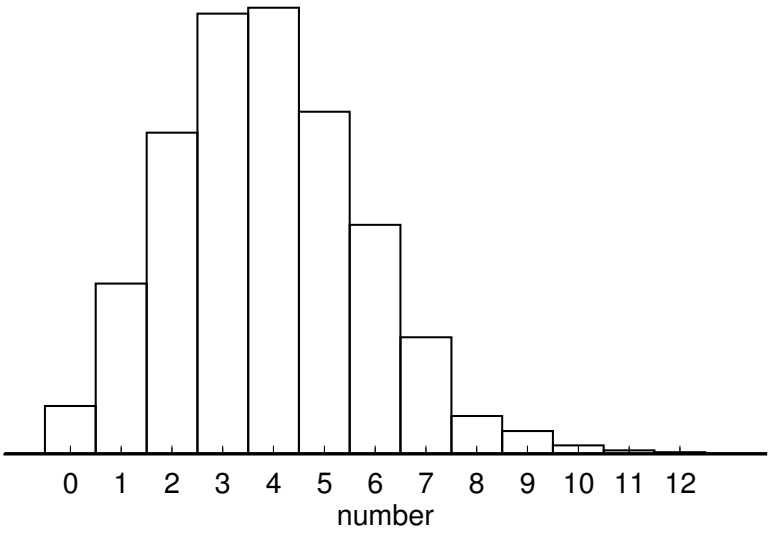


Example. Radioactive disintegrations. This example is taken from Feller, *An Introduction to Probability Theory and its Applications, vol.1*, Wiley, (1957), 149 and Cramér, *Mathematical Methods of Statistics*, Princeton, (1945). In a famous experiment by Rutherford, Chadwick, and Ellis (*Radiations from Radioactive Substances*, Cambridge, 1920) a radioactive substance was observed during 2608 consecutive time intervals of length 7.5 seconds each. In this example a unit is a 7.5 second time interval and the number of particles reaching a counter during the time period is a discrete quantitative variable with possible values of $0, 1, 2, \ldots$. The distribution of the number of radioactive disintegrations is summarized in Table 8 and Figure 10. In this example we see that the distribution of the number of particles per time interval is mound shaped with a single peak around 3 and 4. This distribution is reasonably symmetric but there is some skewness to the right.

Table 8. Radioactive disintegrations relative frequency distribution.

number	frequency	relative frequency
0	57	.0219
1	203	.0778
2	383	.1469
3	525	.2013
4	532	.2040
5	408	.1564
6	273	.1047
7	139	.0533
8	45	.0173
9	27	.0104
10	10	.0038
11	4	.0015
12	2	.0008
total	2608	1.0001

Figure 10. Histogram for radioactive disintegrations.



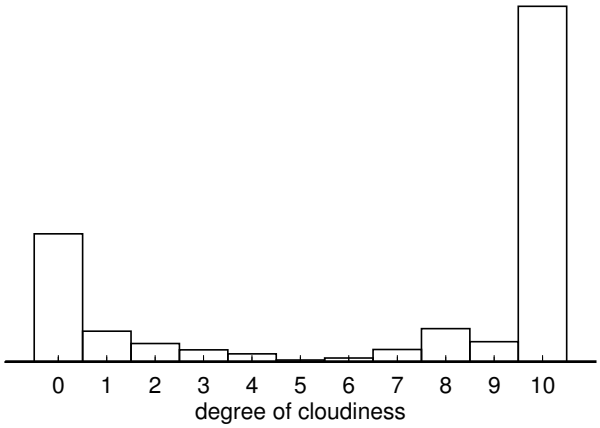
Example. Degree of cloudiness at Breslau. This example is taken from P.R. Rider (1927), *J. Amer. Statist. Assoc.* **22**, 202–208. The estimated degree of cloudiness at Breslau for days during the decade 1876–1885 is summarized in Table 9 and Figure 11. Zero degrees of cloudiness corresponds to an entirely clear day and 10 degrees of cloudiness corresponds to an entirely overcast day. This measurement of degree of cloudiness is essentially a ranking on a scale from 0 to 10 and this variable is properly viewed as being an ordinal qualitative variable. However, as long as we are careful with our interpretation

of its numerical values it is reasonable to treat the degree of cloudiness as a discrete quantitative variable. The distribution of the degree of cloudiness is "U"–shaped with a peak at each of the extremes and relatively low relative frequencies in the middle of the range. More properly, we might say that there is a primary peak (mode) at 10 and a smaller secondary peak (mode) at 0. This "U"–shape indicates that for most of these days it was either entirely clear or nearly clear or it was entirely overcast or nearly overcast. There were relatively few days when the degree of cloudiness was in the middle of the range. The most common value was 10, entirely overcast (57.19%), and the second most common value was 0, entirely clear (20.56%).

Table 9. Degree of cloudiness at Breslau relative frequency distribution.

degree of cloudiness	frequency	relative frequency
0	751	.2056
1	179	.0490
2	107	.0293
3	69	.0189
4	46	.0126
5	9	.0025
6	21	.0057
7	71	.0194
8	194	.0531
9	117	.0320
10	2089	.5719
total	3653	1.0000

Figure 11. Histogram for degree of cloudiness at Breslau.



2.4 Describing continuous quantitative data

There is a fundamental difference between summarizing and describing the distribution of a discrete quantitative variable and summarizing and describing the distribution of a continuous quantitative variable. Since a continuous quantitative variable has an infinite number of possible values, it is not possible to list all of these values. Therefore, some changes to the tabular and graphical summaries used for discrete variables are required.

In practice, the observed values of a continuous quantitative variable are discretized, *i.e.,* the values are rounded so that they can be written down. Therefore, there is really no difference between summarizing the distribution of a continuous variable and summarizing the distribution of a discrete variable with a large number of possible values. In either case, it may be impossible or undesirable to actually list all of the possible values of the variable within the range of the observed data. Thus, when summarizing the distribution of a continuous variable, we will group the possible values into intervals.

Figure 12. Stem and leaf histogram for weight.

In this stem and leaf histogram the stem represents tens and the leaf represents ones. (pounds)

stem	leaf
9	56
10	355
11	0000000255567
12	000555
13	00003455555
14	000000055
15	00556
16	0255
17	000005
18	0055
19	000
20	5

To make this discussion more concrete, consider the weights of the students in the Stat 214 example. We can group the possible weights into the intervals: 90–100, 100–110, ..., 200–210. We will need to adopt an endpoint convention so that each possible weight belongs to only one of these intervals. We will adopt the endpoint convention of including the left (lower) endpoint and excluding the right (upper) endpoint. Under this convention the interval 90–100 includes 90 but excludes 100. A stem and leaf histogram of the weights of the Stat 214 students is given in Figure 12. The **stem and leaf histogram** is an easily constructed version of a **frequency histogram** (a histogram based on frequencies instead of relative frequencies). The stem and leaf histogram uses the numbers themselves to form

the rectangles of the histogram. The stem indicates the interval of values while the leaves provide the "rectangle." For the weight data the actual weight of a student is decomposed into tens (the stem) and ones (the leaf). For example, the first weight is 95 pounds which is decomposed as 95 = 9 tens + 5 ones. Therefore, the weight 95 appears as a 5 leaf in the leaves of the 9 stem.

Notice that the leaves in the stem and leaf histogram of Figure 12 are arranged in increasing order within each stem. Having ordered leaves in the stem and leaf histogram makes certain subsequent tasks easier. For example, having ordered leaves makes it easier to change the stems (change the intervals used to group the values) if this is necessary to obtain a more informative stem and leaf histogram. Furthermore, we can easily determine certain summary statistics directly from a stem and leaf histogram with ordered leaves. (We will discuss summary statistics in Chapter 3.) When constructing a stem and leaf histogram from unordered data, the best way to get ordered leaves is to first form a preliminary stem and leaf histogram with unordered leaves and then revise it to get ordered leaves.

Once the stem and leaf histogram is formed it is easy to construct a frequency distribution, a relative frequency distribution, and a formal relative frequency histogram, if these are desired. By counting the numbers of leaves corresponding to each stem in the stem and leaf histogram we can easily form the corresponding frequency and relative frequency distributions and the formal (relative frequency) histogram.

The weight distribution histogram of Figure 12 has an asymmetric mound shape peaking in the 110–120 pound range and showing skewness to the right. There is a lot of variability in the weights of these students. The majority of the students have weights in the 110–150 pound range; but, weights in the 150–190 pound range are also fairly common.

The appearance of two peaks, one in the 11 (110 pound) stem and one in the 13 (130 pound) stem is probably due to the way these students rounded their weights; therefore, it seems reasonable to say that this distribution has a single peak. Notice that, in general, the appearance of a stem and leaf histogram or a formal histogram for a continuous variable depends on the choice of the intervals used in its construction and minor features such as multiple local modes which are not very far apart might disappear if the intervals were shifted slightly.

Since this weight distribution corresponds to a group consisting of both females and males, we might expect to see two separate peaks; one located at the center of the female weight distribution and another located at the center of the male weight distribution. However, there does not appear to be much evidence of this. Separate stem and leaf histograms for the weight distributions of the females and the males are given in Figure 13. Some care is required in comparing these two stem and leaf (frequency) histograms due to the disparate sample sizes. There are 51 female weights but only 16 male weights.

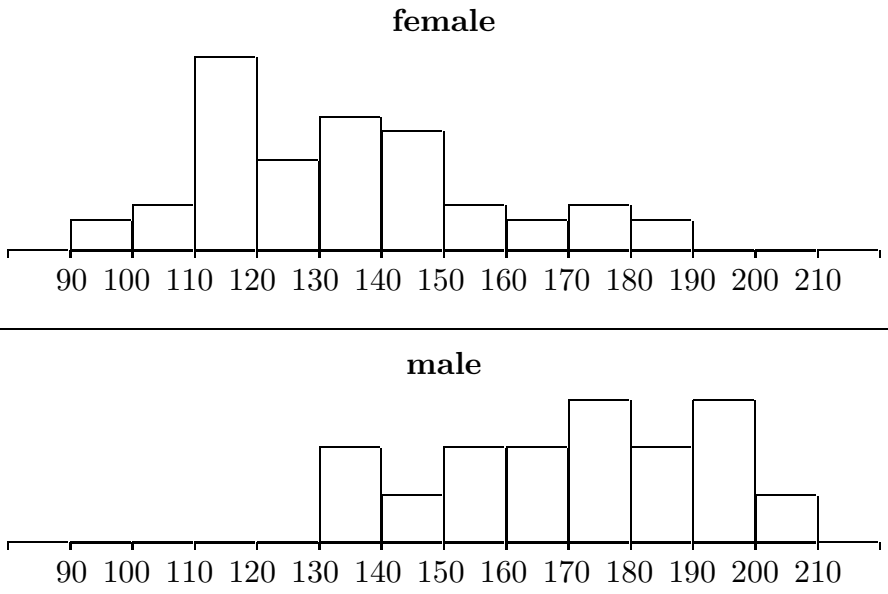
The peak in the female weight distribution stem and leaf histogram appears to be much more pronounced than the peak in the male weight distribution stem and leaf histogram. However, the formal (relative frequency) histograms exhibited in Figure 14 show that this difference in peakedness is not so large. The female weight distribution is skewed to the right with one peak in the 110–150 pound range. The male weight distribution is much more uniform without strong evidence of skewness.

Figure 13. Stem and leaf histograms for weight, by sex.

In these stem and leaf histograms the stem represents tens and the leaf represents ones. (pounds)

Female	**Male**
9 56	9
10 355	10
11 0000000255567	11
12 000555	12
13 000345555	13 05
14 00000055	14 0
15 005	15 56
16 05	16 25
17 000	17 005
18 05	18 05
19	19 000
20	20 5

Figure 14. Histograms for weight, by sex.



The stem and leaf histograms and formal histograms we formed for the weight distribution were based on intervals of length 10 (10 pounds), *e.g.*, 90–100, 100–110, *etc.* Notice that we chose this interval length when we chose to use the last digit of the weight of a student as the leaf and the remaining digits of the weight as the stem in the stem and leaf histogram. In some situations using the last digit of the variable value as the leaf may yield inappropriate intervals. Consider the stem and leaf histogram for the height distribution for the Stat 214 example given in Figure 15. Clearly the majority of the heights are in the 60's, but the shape of the distribution is not clear from this stem and leaf histogram. The intervals used here are too long causing the stem and leaf histogram to be too compressed along the number line to give a useful indication of the shape of the distribution.

Figure 15. Stem and leaf histogram for height.

In this stem and leaf histogram the stem represents tens and the leaf represents ones. (inches)

```
5  9
6  00111122222233333444444445555666666666667777888888899 9
7  000112222245
```

We can refine the stem and leaf histogram by changing the lengths of the intervals into which the data are grouped. This refinement can be viewed as a splitting of the stems of the stem and leaf histogram. To avoid distortion we need to subdivide the intervals (split the stems) so that each of the resulting intervals is of the same length. We can easily do this by either splitting the stems once, yielding 2 intervals of length 5 for each stem instead of 1 interval of length 10, or by splitting the stems five times, yielding 5 intervals of length 2. To demonstrate this splitting of stems stem and leaf histograms of the height distribution with stems split in these fashions are provided in Figures 16 and 17, respectively. In this particular case the stem and leaf histogram of Figure 17 (stems split into five) seems to provide the most informative display of the shape of the height distribution. The height distribution is reasonably symmetric with a single peak at the 66–67 interval.

Figure 16. Stem and leaf histogram for height with stems split into two.

In this stem and leaf histogram the stem represents tens and the leaf represents ones. (inches)

Low stem leaves: 0,1,2,3,4
High stem leaves: 5,6,7,8,9

```
5
5  9
6  00111122222233333444444444
6  55556666666666677778888888999
7  00011222224
7  5
```

Figure 17. Stem and leaf histogram for height with stems split into five.

In this stem and leaf histogram the stem represents tens and the leaf represents ones. (inches)

First stem leaves: 0,1
Second stem leaves: 2,3
Third stem leaves: 4,5
Fourth stem leaves: 6,7
Fifth stem leaves: 8,9

```
5  9
6  001111
6  2222233333
6  444444445555
6  666666666667777
6  8888888999
7  00011
7  22222
7  45
```

To complete our discussion of stem and leaf histograms consider the hypothetical example, with values between -3.9 and 3.9, of Figure 18. The first thing you should notice is that there is a -0 stem and a 0 stem. The negative 0 stem corresponds to the interval from -1 to 0 (not including -1) and the positive 0 stem corresponds to the interval from 0 to 1 (not including 1). If there were any zero observations we could place half of them with each of the zero stems. Notice also that the leaves for the negative stems decrease from left to right so that as we read through the histogram (going from left to right) the values increase from the minimum -3.9 to the maximum 3.9.

Figure 18. A stem and leaf histogram for hypothetical data with negative values and positive values.

-3	98664
-2	8773210
-1	7664432110
-0	9776555442211
0	11222344467
1	1222345578
2	34466789
3	445569

2.5 Summary

In this chapter we discussed tabular and graphical methods for summarizing the distribution of a variable X, *i.e.*, methods for summarizing the way in which the possible values of X are distributed among the units in the sample. The basic idea underlying these summaries is that of using relative frequencies (proportions or percentages) to show how the total relative frequency of one (100%) is partitioned into relative frequencies for each of the possible values of X.

A relative frequency distribution is a table listing the possible values of X and the associated relative frequencies with which these values occurred in the sample. For a qualitative variable or a discrete quantitative variable it is usually possible to tabulate all of the possible values and their relative frequencies. For a discrete quantitative variable with many possible values or a continuous quantitative variable, there are generally too many possible values to list each individually and it is necessary to group the possible values into intervals and then tabulate the relative frequencies for each of these intervals of values.

A graphical representation of the distribution of X is based on the identification of area with relative frequency. Thus, a graphical representation provides a decomposition of a region of area one, representing the total relative frequency of one (100%), into subregions of area equal to the relative frequencies of each of the possible values (or intervals of values) of X. For qualitative variables we emphasized the bar graph with rectangular regions for each value of X. For quantitative variables we used a histogram which is basically a bar graph with the bars suitably arranged along the number line to indicate the relative locations of the values of X. We also discussed stem and leaf histograms which are easily constructed raw frequency histograms; and we noted that stem and leaf histograms should be converted to proper relative frequency histograms before making comparisons of two or more distributions.

The representation of the distribution of a quantitative variable via a histogram allows us to discuss the shape of the distribution. We discussed some basic shapes with most of our emphasis on the distinction between skewed and symmetric mound shaped distributions. For skewed distributions we defined the terms skewed left and skewed right.

2.6 Exercises

For each of the examples in Section 1.2 (excluding those already treated in this chapter): construct suitable tabular and graphical summaries of the distribution(s) and discuss the distribution of the variable(s).

Notes:

For the examples with two or more groups (DiMaggio and Mantle, Guatemalan cholesterol, gear tooth strength), compare and contrast the distributions of the variable for the two (or more) groups.

For the paired data examples (wooly–bear cocoons, homophone confusions) find the differences for each pair of data values and describe the distribution of the differences.

Chapter 3
Descriptive Statistics II: Numerical Summary Values

3.1 Numerical summary values for quantitative data

For many purposes a few well–chosen numerical summary values (statistics) will suffice as a description of the distribution of a quantitative variable. A **statistic** is a numerical characteristic of a sample. More formally, a statistic is a numerical quantity computed from the values of a variable, or variables, corresponding to the units in a sample. Thus a statistic serves to quantify some interesting aspect of the distribution of a variable in a sample. Summary statistics are particularly useful for comparing and contrasting the distribution of a variable for two different samples.

If we plan to use a small number of summary statistics to characterize a distribution or to compare two distributions, then we first need to decide which aspects of the distribution are of primary interest. If the distributions of interest are essentially mound shaped with a single peak (unimodal), then there are three aspects of the distribution which are often of primary interest. The first aspect of the distribution is its location on the number line. Generally, when speaking of the location of a distribution we are referring to the location of the "center" of the distribution. The location of the center of a symmetric, mound shaped distribution is clearly the point of symmetry. There is some ambiguity in specifying the location of the center of an asymmetric, mound shaped distribution and we shall see that there are at least two standard ways to quantify location in this context. The second aspect of the distribution is the amount of variability or dispersion in the distribution. Roughly speaking, we would say that a distribution exhibits low variability if the observed values tend to be close together on the number line and exhibits high variability if the observed values tend to be more spread out in some sense. The third aspect is the shape of the distribution and in particular the degree of skewness in the distribution.

As a starting point consider the **minimum** (smallest observed value) and **maximum** (largest observed value) as statistics. We know that all of the data values lie between the minimum and the maximum, therefore, the minimum and the maximum provide a crude quantification of location and variability. In particular, we know that all of the values of the variable are restricted to the interval from the minimum to the maximum; however, the minimum and the maximum alone tell us nothing about how the data values are distributed within this interval. If the distribution is reasonably symmetric and mound shaped, then the **midrange**, defined as the average of the minimum and the maximum, may provide a suitable quantification of the location of the center of the distribution. The median and mean, which are defined below, are generally better measures of the center of a distribution.

The **range**, defined as the distance from the minimum to the maximum can be used to quantify the amount of variability in the distribution. Note that the range is the positive number obtained by subtracting the minimum from the maximum. When comparing two distributions the distribution with the larger range will generally have more variability than the distribution with the smaller range; however, the range is very sensitive to extreme observations so that one or a few unusually large or small values can lead to a very large range.

We will now consider an approach to the quantification of the shape, location, and variability of a distribution based on the division of the histogram of the distribution into sections of equal area. This is equivalent to dividing the data into groups, each containing the same number of values. We will first use a division of the histogram into halves. We will then use a division of the histogram into fourths.

The median is used to quantify the location of the center of the distribution. In terms of area, the **median** is the number (point on the number line) with the property that the area in the histogram to the left of the median is equal to the area to the right of the median. Here and in the sequel we will use a lower case n to denote the sample size, *i.e.*, n will denote the number of units in the sample. In terms of the n observations, the **median** is the number with the property that at least $n/2$ of the observed values are less than or equal to the median and at least $n/2$ of the observed values are greater than or equal to the median.

A simple procedure for finding the median, which is easily generalized to fractions other that 1/2, is outlined below.

1. Arrange the data (observations) in increasing order from the smallest (obs. no. 1) to the largest (obs. no. n). Be sure to include all n values in this list, including repeats if there are any.
2. Compute the quantity $n/2$.

3a. If $n/2$ is not a whole number, round it up to the next largest integer. The observation at the location indicated by the rounded–up value in the ordered listing of the data is the median.

3b. If $n/2$ is a whole number, then we need to average two values to get the median. The two observations to be averaged are obs. no. $n/2$ and the next observation (obs. no. $n/2 + 1$) in the ordered listing of the data. Find these two observations and average them to get the median.

We can use the distance between the minimum and the median and the distance between the median and the maximum to quantify the amount of skewness in the distribution. The distance between the minimum and the median is the range of the lower (left) half of the distribution, and the distance between the median and the maximum is the range of the upper (right) half of the distribution. If the distribution is symmetric, then

these two distances (median – minimum) and (maximum – median) will be equal. If the distribution is skewed, then we would expect to observe a larger range (indicating more variability) for the half of the distribution in the direction of the skewness. Thus if the distribution is skewed to the left, then we would expect (median – minimum) to be greater than (maximum – median). On the other hand, if the distribution is skewed to the right, then we would expect (maximum – median) to be greater than (median – minimum).

Example. Weed seeds (revisited). Recall that this example is concerned with the number of weed seeds found in $n = 98$ quarter–ounce batches of grass. Since $98/2 = 49$, the median for this example is the average of observations 49 and 50. Referring to Table 6 of Chapter 2 we find that the minimum number of weed seeds is 0, the maximum is 7, and the median is 1, since observations 49 and 50 are each 1. The range for this distribution is $7 - 0 = 7$. Notice that the range of the right half of this distribution (maximum – median) $= 7 - 1 = 6$ is much larger than the range of the left half (median – minimum) $= 1 - 0 = 1$ confirming our observation that this distribution is strongly skewed to the right.

Example. Vole reproduction (revisited). Recall that this example is concerned with the number of babies in $n = 170$ litters of voles. Since $170/2 = 85$, the median for this example is the average of observations 85 and 86. Referring to Table 7 of Chapter 2 we find that the minimum number of babies is 1, the maximum is 11, and the median is 6, since observations 85 and 86 are each 6. The range for this distribution is $11 - 1 = 10$. Notice that the range of the right half of this distribution (maximum – median) $= 11 - 6 = 5$ is equal to the range of the left half (median – minimum) $= 6 - 1 = 5$ confirming our observation that this distribution is symmetric.

A more detailed quantification of the shape and variability of a distribution can be obtained from a division of the distribution into fourths. In order to divide a distribution into fourths, we need to specify three numbers or points on the number line. These statistics are called **quartiles**, since they divide the distribution into quarters. In terms of area, the **first quartile**, denoted by Q_1 (read this as Q sub one), is the number (point on the number line) with the property that the area in the histogram to the left of Q_1 is equal to one fourth and the area to the right of Q_1 is equal to three fourths. The **second quartile**, denoted by Q_2, is the median. The **third quartile**, denoted by Q_3, is the number (point on the number line) with the property that the area in the histogram to the left of Q_3 is equal to three fourths and the area to the right of Q_3 is equal to one fourth. In terms of the n observations, Q_1 is the number with the property that at least $n/4$ of the observed values are less than or equal to Q_1 and at least $3n/4$ of the observed values are greater than or equal to Q_1. Similarly, Q_3 is the number with the property that at least $3n/4$ of the observed values are less than or equal to Q_3 and at least $n/4$ of the observed values are greater than or equal to Q_3.

The method for finding the median given above is readily modified for finding the first and third quartiles. For Q_1, we simply replace $n/2$ by $n/4$ and replace the words 'the median' by Q_1. To find Q_3, use exactly the same method but count down from the largest value instead of counting up from the smallest value. Some calculators and computer programs use variations of the methods given above for finding Q_1 and Q_3. These variations may give slightly different values for Q_1 and Q_3.

Example. Weed seeds (revisited). Since $98/4 = 24.5$, the quartiles Q_1 and Q_3 for this example are the observations located at position 25 counting up for Q_1 and counting down for Q_3. Referring to Table 6 of Chapter 2 we find that $Q_1 = 0$ and $Q_3 = 2$. Notice that the range of the lower three–fourths of this distribution, Q_3 – minimum, is 2 while the range of the upper fourth, maximum – Q_3 is 5. This indicates that 75% (a large proportion) of the batches of grass have relatively few weed seeds, and the skewness in this distribution is due to the high amount of variability among the numbers of weed seeds in the 25% of the batches with between 2 and 7 weed seeds.

Previously we introduced the range as a measure of variability. An alternative measure of variability is provided by the interquartile range. The **interquartile range** (IQR) is the distance between the first quartile Q_1 and the third quartile Q_3, *i.e.*, the interquartile range is the positive number obtained by subtracting Q_1 from Q_3. Notice that the **interquartile range** is the range of the middle half of the distribution. The interquartile range is less sensitive to the presence of a few extreme observations in the data than is the range. For example, if there are one or two unusually large or unusually small values, then these values may have the effect of making the range much larger than it would be if these unusual values were not present. In such a situation, we might argue that the range is too large to be deemed an appropriate overall measure of the variability of the distribution. The interquartile range is not affected by a few unusual values, since it only depends on the middle half of the data. We could use the range of a larger part of the middle of the distribution, say the middle 75% or 90%, as a compromise between the range and the interquartile range.

The five summary statistics: the minimum (min), the first quartile (Q_1), the median (med), the third quartile (Q_3), and the maximum (max), constitute the **five number summary** of the distribution. Each of these five statistics provides a quantification of a particular aspect of the distribution. They quantify where the distribution begins, where the first quarter of the distribution ends, and so on. Furthermore, the distances between these five statistics can be used to quantify the shape (skewness) of the distribution.

The four distances: (Q_1 – min), (med – Q_1), (Q_3 – med), and (max – Q_3), are the ranges of the first, second, third, and fourth quarters of the distribution, respectively. These distances can be used to quantify the amount of variability in the corresponding parts of the distribution. Comparisons of appropriate pairs of these distances provide

indications of certain aspects of the shape of the distribution. The relationship between (med – Q_1) and (Q_3 – med) can be used to quantify the shape (skewness) of the middle half of the distribution. Since (Q_1 – min) and (max – Q_3) are the lengths of the tails (lower and upper fourths) of the distribution, the relationship between these numbers can be used to quantify skewness in the tails of the distribution.

Example. Cholesterol levels in Guatemalans. This example is taken from Devore and Peck, *Statistics*, 3 ed., (1997), Duxbury, p. 23. The original source is "The Blood Viscosity of Various Socioeconomic Groups in Guatemala" in *The American Journal of Clinical Nutrition*, Nov., 1964, 303–307. The Institute of Nutrition of Central America and Panama measured the serum total cholesterol levels for a group of 49 adult, low–income rural Guatemalans and for a group of 45 adult, high–income urban Guatemalans. The serum total cholesterol levels (in mg/dL) are provided in Table 1 and stem and leaf histograms are given in Figure 1.

Table 1. Guatemalan cholesterol data.

Rural group cholesterol levels (in mg/dL).

95	108	108	114	115	124	129	129	131	131
135	136	136	139	140	142	142	143	143	144
144	145	146	148	152	152	155	157	158	158
162	165	166	171	172	173	174	175	180	181
189	192	194	197	204	220	223	226	231	

Urban group cholesterol levels (in mg/dL).

133	134	155	170	175	179	181	184	188	189
190	196	197	199	200	200	201	201	204	205
205	205	206	214	217	222	222	227	227	228
234	234	236	239	241	242	244	249	252	273
279	284	284	284	330					

Before we compute any summary statistics consider the stem and leaf histograms in Figure 1. Based on these histograms we can see that both of these cholesterol level distributions are basically mound shaped with some skewness to the right. In the rural group there are four individuals with somewhat high cholesterol levels (220 or more); there is a gap of 16 separating the cholesterol levels of these individuals from the rest of the rural group. It is this group of four observations which causes the rural distribution to appear skewed to the right. The urban group has similar slightly unusual groups of cholesterol levels; one group having somewhat low levels and one having somewhat high levels. There is one unusually large value (330) in the urban group that we might consider an outlier, since there is a gap of 46 between 330 and the next largest value. (An outlier is an observation that is widely separated from the majority of a distribution.) We will need to

consider the implications of this outlier in our analysis of this example. It is also apparent that the people in the urban group tend to have higher cholesterol levels than the people in the rural group. There appears to be more variability among the cholesterol levels for the urban group. With the urban outlier there appears to be much more variability in the cholesterol levels of the urban group, and without it there appears to be slightly more variability in the urban group cholesterol levels. If we ignore the outlier, the urban group distribution appears to be essentially symmetric.

Figure 1. Guatemalan cholesterol stem and leaf histograms.

The stem represents tens and the leaf represents ones. (mg/dL)

Rural		Urban	
9	5	9	
10	88	10	
11	45	11	
12	499	12	
13	115669	13	34
14	0223344568	14	
15	225788	15	5
16	256	16	
17	12345	17	059
18	019	18	1489
19	247	19	0679
20	4	20	001145556
21		21	47
22	036	22	22778
23	1	23	4469
24		24	1249
25		25	2
26		26	
27		27	39
28		28	444
29		29	
30		30	
31		31	
32		32	
33		33	0

The five number summaries and the associated distances based on them are provided, for the rural group, for the entire urban group, and for the urban group omitting 330, in Table 2. The steps involving in computing the medians and quartiles, for the rural group

and the entire urban group, are outlined below. For the rural group there are $n = 49$ observations so that

(1) $49/2 = 24.5$, thus the median 152 is obs. no. 25, corresponding to the first 2 leaf in the 15 stem.

(2) $49/4 = 12.25$, thus the first and third quartiles are $Q_1 = 136$, the 13^{th} observation counting up, corresponding to the second 6 leaf in the 13 stem, and $Q_3 = 174$, the 13^{th} observation counting down, corresponding to the second 4 leaf in the 17 stem.

For the urban group there are $n = 45$ observations so that

(1) $45/2 = 22.5$, thus the median 206 is obs. no. 23, corresponding to the 6 leaf in the 20 stem.

(2) $45/4 = 11.25$, thus the first and third quartiles are $Q_1 = 196$, the 12^{th} observation counting up, corresponding to the 6 leaf in the 19 stem, and $Q_3 = 239$, the 12^{th} observation counting down, corresponding to the 9 leaf in the 23 stem.

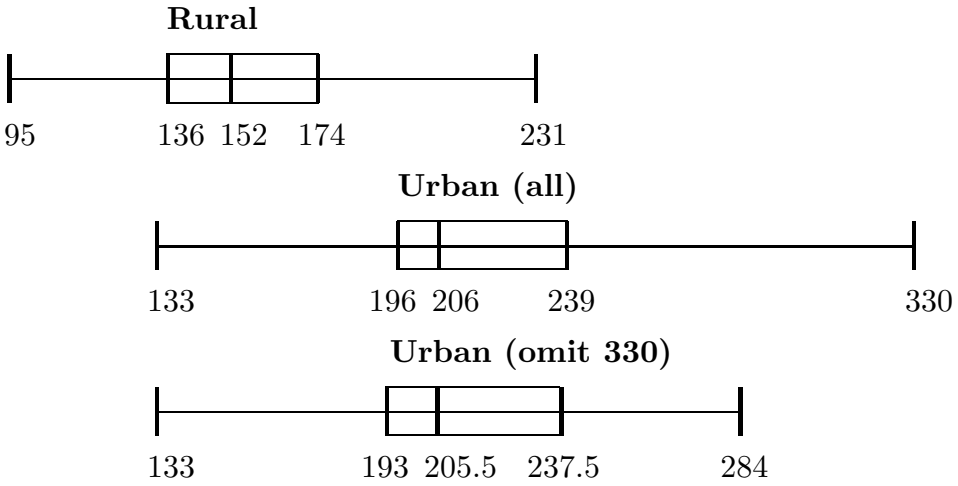
Table 2. Five number summaries with distances.

Rural group. (mg/dL) n=49					
min:	95				
Q_1:	136	Q_1- min:	41		
med:	152	med - Q_1:	16	med - min:	57
Q_3:	174	Q_3- med:	22		
max:	231	max - Q_3:	57	max - med:	79
Urban group (all). (mg/dL) n=45					
min:	133				
Q_1:	196	Q_1- min:	63		
med:	206	med - Q_1:	10	med - min:	73
Q_3:	239	Q_3- med:	33		
max:	330	max - Q_3:	91	max - med:	124
Urban group (omit 330). (mg/dL) n=44					
min:	133				
Q_1:	193	Q_1- min:	60		
med:	205.5	med - Q_1:	12.5	med - min:	72.5
Q_3:	237.5	Q_3- med:	32		
max:	284	max - Q_3:	46.5	max - med:	78.5

Before we continue with our discussion of this example we will introduce a simple graphical display corresponding to the information in Table 2. We can use the five number summary values to form a simple graphical representation of a distribution known as a

box plot or a box and whiskers plot. A box plot does not convey as much information as a stem and leaf histogram but it does give a useful graphical impression of the shape of the distribution as well as its location and variability. Simple box plots for the Guatemalan cholesterol example are provided in Figure 2.

Figure 2. Box plots for cholesterol level.



Notice that each box plot has five vertical marks indicating the locations of the five number summary values. The box which extends from the first quartile to the third quartile and is divided into two parts by the median gives an impression of the distribution of the values in the middle half of the distribution. In particular, a glance at this box indicates whether the middle half of the distribution is skewed or symmetric and indicates the magnitude of the interquartile range (the length of the box). The line segments (whiskers) which extend from the ends of the box to the extreme values (the minimum and the maximum) give an impression of the distribution of the values in the tails of the distribution. The relative lengths of the whiskers indicate the contribution of the tails of the distribution to the symmetry or skewness of the distribution.

Returning to the cholesterol example first consider the shapes of the cholesterol distributions. We can use the distances, based on the five number summary, given in Table 2 to quantify the degree of skewness in these distributions. Comparing the distances for the rural group we find that max – med = 79 > 57 = med – min, Q_3 – med = 22 > 16 = med – Q_1, and Max – Q_3 = 57 > 41 = Q_1 – min. All of these comparisons support our contention that the cholesterol distribution for the rural group is skewed right. For the urban group, including the outlier, we have max – med = 124 > 73 = med – min, Q_3 – med = 33 > 10 = med – Q_1, and Max – Q_3 = 91 > 63 = Q_1 – min. All of these comparisons support our contention that the cholesterol distribution for the urban group is skewed right. If we omit the outlier (330) from the urban group we find that max –

med = 78.5 is only slightly larger than med – min = 72.5 suggesting that without the outlier the cholesterol distribution for the urban group is reasonably symmetric. Without the outlier the middle half of the distribution is still somewhat skewed right, since Q_3 – med = 32 > 12.5 = med – Q_1; but, the range of the left tail (lower fourth) Q_1 – min = 60 is now larger than the range of the right tail (upper fourth) Max – Q_3 = 46.5.

The fact that the median 152 for the rural group is much smaller than the median 206 (with the outlier) or 205.5 (without the outlier) of the urban group supports our contention that the people in the urban group tend to have higher cholesterol levels than the people in the rural group.

With the outlier the range 197 for the urban group is much larger than the range 137 for the rural group. If we omit the outlier, then the range for the urban group is 151 which is still larger than 137 but not by so much. On the other hand, if we consider the interquartile ranges, 38 for the rural group and 43 (44.5 without the outlier) for the urban group, we find that there is a similar amount of variability in the middle halves of these distributions. Hence, our contention that there is much more variability among the cholesterol levels of the urban Guatemalans depends very heavily on the cholesterol level of one individual. Whether we include this individual or not, we are justified in claiming that there is more variability among the cholesterol levels of the urban Guatemalans.

Based on our analysis of these cholesterol level distributions we might propose several hypotheses or conjectures about why these distributions differ as they do. First we might conjecture that the rural Guatemalans are probably more physically active and eat food which is lower in fat than the urban Guatemalans. This would cause the rural Guatemalans to tend to have lower cholesterol levels. Second, we might argue that there is less variability in the cholesterol levels of the rural Guatemalans because their lifestyles and eating habits are probably quite similar.

The approach that we have been using to form summary statistics is to select a single representative value from the observed values of the variable (or the average of two adjacent observed values) to quantify a particular aspect of the distribution. We have also considered statistics that are distances between two such representative values.

An alternative approach to forming a summary statistic is to combine all of the observed values to get a suitable statistic. The first statistic of this type that we consider is the mean. The **mean**, which is the simple arithmetic average of the n data values, is used to quantify the location of the center of the distribution. You could compute the mean by adding all n data values together and dividing this sum by n; however, it is better to use a calculator or a computer.

The sample mean is often denoted by the symbol $\overline{X}$ (read this as X bar). This is a convenient place for us to introduce some standard notation. It is standard practice to use a letter, such as X, to denote a variable and the values of the variable. You are

free to choose a letter with mnemonic value instead of the generic letter X; however, you should not use S or Z as these letters are reserved for special uses. If X denotes the variable of interest, then we will use $\overline{X}$ to denote the mean of the distribution of X. If we used a different letter, say Y, to denote the variable, then we would use $\overline{Y}$ to denote the corresponding mean. We will use function notation to denote the other statistics we defined above. That is, if X denotes the variable, then $\min(X)$, $Q_1(X)$, $\text{med}(X)$, $Q_3(X)$, $\max(X)$, $\text{range}(X)$, and $\text{IQR}(X)$ denote the minimum, the first quartile, the median, the third quartile, the maximum, the range, and the interquartile range, respectively. You should read these symbols as follows: read $\min(X)$ as the minimum of X, $Q_1(X)$ as the first quartile of X, and so on.

Recall that the median is the number (point on the number line) with the property that the area in the histogram to the left of the median is equal to the area to the right of the median. The mean is the number (point on the number line) where the histogram would balance. To understand what we mean by the balance point, imagine the histogram as being cut out of a piece of cardboard. The mean is located at the point along the number line side of this cutout where the histogram cutout would balance. These geometric characterizations of the mean and the median imply that when the distribution is symmetric the mean will be equal to the median. Furthermore, if the distribution is skewed to the right, then the mean (the balance point) will be larger than the median (to the right of the median). Similarly, if the distribution is skewed to the left, then the mean (the balance point) will be smaller than the median (to the left of the median).

The primary use of the mean, like the median, is to quantify the location of the center of a distribution and to compare the locations (centers) of two distributions. Since both the mean and the median can be used to quantify the location of the center of a distribution, it seems reasonable to ask which is more appropriate. If the distribution is approximately symmetric, then the mean and the median will be approximately equal. On the other hand, if the distribution is not symmetric, then the median is likely to provide a better indication of the center of the distribution. For example, if the distribution is strongly skewed to the right, then the mean may be much larger than the median and the mean may not be a good indication of the center of the distribution. For a specific problem it is a good idea to mark the locations of the mean and the median on a histogram of the distribution and consider which seems more reasonable as an indicator of the center of the distribution.

If the mean $\overline{X}$ is deemed suitable as a measure of the center of the distribution of X, then the deviations $(X - \overline{X})$ of the observed values of X from their mean $\overline{X}$ contain information about the amount of variability in the distribution. If there is little variability (the observed values of X are close together and they are close to the mean $\overline{X}$), then the deviations $(X - \overline{X})$ will tend to be small in magnitude (absolute value). On the other

hand, if there is a lot of variability (at least some of the observed values of X are far apart and they are not all close to the mean $\overline{X}$), then the deviations $(X-\overline{X})$ will tend to be large in magnitude. It is this observation which suggests that a summary statistic based on the distances between each of the observed values of the variable and their mean can be used to measure the variability in the distribution. The standard deviation is one such statistic. The **standard deviation** is the square root of the "average" of the squared deviations of the observed values of the variable from their mean. A formula for the standard deviation is given below; however, you should not use this formula to compute the standard deviation. Instead you should use a calculator or a computer to compute the standard deviation. In symbols, the **standard deviation** of the distribution of the variable X, denoted by S_X (read this as S sub X), is

$$S_X = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}}$$

In this formula the capital Greek letter sigma, Σ, represents the statement "the sum of ", and $(X-\overline{X})^2$ denotes the square of the distance from the observed value X to the mean $\overline{X}$. Therefore, the expression under the square root sign in the formula is the "average" of the squared deviations of the observed values of the variable from their mean as mentioned above. The reason for the square root is so that the standard deviation of X and the variable X are in the same units of measurement.

The standard deviation is positive, unless there is no variability at all in the data. That is, unless all of the observations are exactly the same, the standard deviation is a positive number. The standard deviation is a very widely used measure of variability. Unfortunately, the standard deviation does not have a simple, direct interpretation. The important thing to remember is that larger values of the standard deviation indicate that there is more variability in the data. A closely related measure of variability is the **variance** which is simply the square of the standard deviation, *i.e.*, the variance of the distribution of X is $S_X^2 = \sum(X-\overline{X})^2/(n-1)$.

There are quotation marks around the word average in the definition of the standard deviation because we divided by $n-1$ even though there are n squared deviations in the average. The reason for this is that, in a sense, there are only $n-1$ individual pieces of information contained in the collection of n deviations from the mean. It is readily verified that the sum of the deviations from the mean (not the sum of their squares) is equal to zero, *i.e.*, $\Sigma(X-\overline{X}) = 0$. This is the algebraic version of the fact that the mean is the balance point of the distribution. Because of this fact, if we know the values of any $n-1$ of the deviations, then we can determine the value of the remaining deviation. This is the sense in which there are only $n-1$ individual pieces of information contained in the collection of n deviations from the mean; and is the reason that we divide by $n-1$.

The means, medians, standard deviations, ranges, and interquartile ranges for the Guatemalan cholesterol level distributions are given in Table 3. Because all three of these distributions are somewhat skewed to the right, we find that in all three cases the mean is larger than the median. Notice the effects of excluding the outlier from the urban group on these statistics. First consider the mean and the median; excluding this outlier has essentially no effect on the median but has an appreciable effect on the mean. This illustrates the sensitivity of the mean to extreme observations. Next consider the three measures of variability. As we noted above, excluding the outlier has a large effect on the range but little effect on the interquartile range. As with the mean, excluding the outlier has an appreciable effect on the standard deviation. This illustrates that, like the mean, the standard deviation is also sensitive to extreme observations.

In this example, if we base our comparisons of the location and the amount of variability in these distributions on the mean and standard deviation we reach essentially the same conclusions as we did when using the five number summary.

Table 3. Summary statistics for the cholesterol example.

group	mean	median	std. dev.	range	IQR
rural	157.02	152	31.75	137	38
urban (all)	216.87	206	39.92	197	43
urban (omit 330)	214.30	205.5	36.42	151	42

Example. EPA mileage values for subcompact cars. Table 4 contains the EPA mileage values and some related information for 56 subcompact car model/engine combinations. This information was obtained from the June 2000 edition of the model year 2000 fuel economy guide provided on the DOT/EPA web site `www.fueleconomy.gov`. If there were two or more listings for the same car model/engine combination, then only one value was included. In particular, if mileage values were provided for a particular car model/engine combination with both automatic and manual transmissions, then only the mileage value for the manual transmission was included. The car models listed in the EPA fuel economy guide are grouped into size classes based on the combined passenger and cargo volume of the car. For example, subcompact cars have combined volumes between 85 and 99 cubic feet and compact cars have combined volumes between 100 and 109 cubic feet. For this example we will consider the two mileage values (city and highway) as response variables. The other variables in Table 4 might serve as potentially interesting explanatory variables. In this example a car model/engine combination is a unit and we have a pair of responses, city mileage and highway mileage, for each car model. We will first consider the distributions of city and highway mileage values separately, ignoring the fact that we have pairs of mileage values for each model.

Table 4. Model year 2000 subcompact car EPA mileage values.

city denotes city mileage in miles per gallon
hiwy denotes highway mileage in miles per gallon
trans denotes transmission type (automatic or manual) and number of gears
displ denotes engine displacement in liters
cyl denotes number of cylinders
drv denotes front, rear, or all wheel drive

manufacturer	model	city	hiwy	trans	displ	cyl	drv
Acura	Integra	25	31	(M5)	1.8	4	F
Acura	Integra(DOHC/VTEC)	25	30	(M5)	1.8	4	F
Bentley	Azure	11	16	(A4)	6.8	8	R
Bentley	Continental SC	11	16	(A4)	6.8	8	R
Bentley	Continental T	11	16	(A4)	6.8	8	R
BMW	323CI	20	29	(M5)	2.5	6	R
BMW	328CI	21	29	(M5)	2.8	6	R
Chevrolet	Camaro	19	30	(M5)	3.8	6	R
Chevrolet	Camaro	18	27	(M6)	5.7	8	R
Chevrolet	Cavalier	24	34	(M5)	2.2	4	F
Chevrolet	Cavalier	23	33	(M5)	2.4	4	F
Chevrolet	Metro	39	46	(M5)	1	3	F
Chevrolet	Metro	36	42	(M5)	1.3	4	F
Ferrari	Ferrari 456 MGT/MGTA	10	16	(M6)	5.5	12	R
Ford	Escort ZX2	25	33	(M5)	2	4	F
Ford	Mustang	20	29	(M5)	3.8	6	R
Ford	Mustang	17	25	(M5)	4.6	8	R
Ford	Mustang(4 Valve)	17	24	(M5)	4.6	8	R
Honda	Civic	32	37	(M5)	1.6	4	F
Honda	Civic(VTEC)	30	35	(M5)	1.6	4	F
Honda	Civic(DOHC/VTEC)	26	31	(M5)	1.6	4	F
Honda	Prelude	22	27	(M5)	2.2	4	F
Hyundai	Tiburon	23	32	(M5)	2	4	F
Jaguar	XK8	18	25	(A5)	4	8	R
Jaguar	XKR	16	23	(A5)	4	8	R
Lexus	SC 300/SC 400	19	23	(A4)	3	6	R
Lexus	SC 300/SC 400	18	25	(A5)	4	8	R
Mercedes–Benz	CLK320	21	29	(A5)	3.2	6	R
Mercedes–Benz	CLK430	18	25	(A5)	4.3	8	R
Mitsubishi	Eclipse	23	31	(M5)	2.4	4	F
Mitsubishi	Eclipse	20	28	(M5)	3	6	F
Mitsubishi	Mirage	33	40	(M5)	1.5	4	F
Mitsubishi	Mirage	28	36	(M5)	1.8	4	F

This table is continued on the next page.

Table 4. Model year 2000 subcompact car EPA mileage values (continued from the preceding page).

manufacturer	model	city	hiwy	trans	displ	cyl	drv
Pontiac	Firebird/TransAm	19	30	(M5)	3.8	6	R
Pontiac	Firebird/TransAm	18	27	(M6)	5.7	8	R
Pontiac	Sunfire	24	34	(M5)	2.2	4	F
Pontiac	Sunfire	23	33	(M5)	2.4	4	F
Rolls–Royce	Corniche	11	16	(A4)	6.8	8	R
Saab	Saab 9-3 Conv.	22	29	(M5)	2	4	F
Saab	Saab 9-3 Viggen Conv.	20	29	(M5)	2.3	4	F
Saturn	SC	28	40	(M5)	1.9	4	F
Saturn	SC(DOHC)	27	38	(M5)	1.9	4	F
Subaru	Impreza AWD	23	29	(M5)	2.2	4	A
Subaru	Impreza AWD	21	28	(M5)	2.5	4	A
Suzuki	Esteem	30	37	(M5)	1.6	4	F
Suzuki	Esteem	28	35	(M5)	1.8	4	F
Suzuki	Swift	36	42	(M5)	1.3	4	F
Toyota	Solara Conv.	23	30	(A4)	2.2	4	F
Toyota	Solara Conv.	19	26	(A4)	3	6	F
Toyota	Celica	28	34	(M5)	1.8	4	F
Toyota	Celica	23	32	(M6)	1.8	4	F
Volkswagen	Cabrio	24	31	(M5)	2	4	F
Volkswagen	New Beetle	25	31	(M5)	1.8	4	F
Volkswagen	New Beetle	24	31	(M5)	2	4	F
Volvo	C70 Conv.	20	26	(M5)	2.3	5	F
Volvo	C70 Conv.	19	26	(A4)	2.4	5	F

The stem and leaf histograms of Figure 3 summarize the distributions of the EPA city and highway gas mileage values for the $n = 56$ model year 2000 subcompact car models. Notice that each of these distributions includes five unusually low mileage values. Five car models have city mileage values of 10 or 11 mpg and five car models have highway mileage values of 16 mpg. It turns out that the five car models with the lowest city mileage values are also the five car models with the lowest highway mileage values. In both distributions there is a large separation between the five low mileage values and the mileage values of the 51 other subcompact car models. Before we proceed with our examination of this example we need to look at the original data, including all relevant information about the car models, to see why these five car models have such low mileage values.

Figure 3. Stem and leaf histograms for model year 2000 subcompact car EPA mileage values.

The stem represents tens and the leaf represents ones. (mpg)

City	Highway
1 01111	
1	
1	
1 677	1 66666
1 8888899999	1
2 00000111	2
2 223333333	2 33
2 44445555	2 45555
2 67	2 666777
2 8888	2 889999999
3 00	3 0000111111
3 23	3 22333
3	3 44455
3 66	3 677
3 9	3 8
	4 00
	4 22
	4
	4 6

From Table 4 we find that the five subcompact car models with the lowest city and highway mileage values are: three Bentley models, one Ferrari model, and one Rolls–Royce model. This group of car models contains four ultra–luxury models and one high performance sports car. Since these five car models do not fit in with the usual conception of a subcompact car, we will remove them from the data. Thus the remainder of this discussion is restricted to the collection of $n = 51$ subcompact car models remaining after removing the five car models discussed above.

We will first make some observations based on these stem and leaf histograms. The city and highway mileage distributions both appear to be skewed to the right. This indicates that, for both the city and highway mileage values, there tends to be more variability among the larger mileage values than among the lower mileage values. Each of these mileage histograms has a single peak. The peak in the city mileage histogram is located near the lower end of the distribution while the peak in the highway mileage distribution is more centrally located. The locations of these peaks and the mound shapes of these distributions indicate that, for subcompact cars, the car mileage values tend to be clustered around the low 20's and the highway mileage values tend to be clustered around the upper

20's and lower 30's. As we would expect, the highway mileage distribution is located higher on the number line than is the city mileage distribution indicating that these subcompact cars tend to get higher mileage on the highway than they do in the city.

Table 5. Subcompact car EPA mileage summary statistics, excluding the five unusual car models.

statistic	**city**	**highway**
n	51	51
min	16	23
Q_1	19	27
med	23	30
Q_3	26	34
max	39	46
range	23	23
IQR	7	7
Q_1- min	3	4
med $-Q_1$	4	3
Q_3- med	3	4
max $-Q_3$	13	12
mean	23.53	31.12
std dev	5.27	5.18

We will now quantify and expand on our observations about the subcompact car mileage distributions. Relevant summary statistics are given in Table 5. In the discussion below, we will use C to denote the city mileage of a subcompact car model and H to denote the highway mileage of a subcompact car model.

First consider the shapes of the subcompact car mileage distributions. For the city mileage distribution we see that: $\max(C) - \text{med}(C) = 16 > 7 = \text{med}(C) - \min(C)$, $\max(C) - Q_3(C) = 13 > 3 = Q_1(C) - \min(C)$, and $\overline{C} = 23.53 > 23 = \text{med}(C)$. All of these comparisons support our contention that the city mileage distribution is skewed to the right. Notice that $Q_3(C) - \text{med}(C) = 3$ which is approximately equal to $\text{med}(C) - Q_1(C) = 4$; this suggests that the middle half of this distribution is reasonably symmetric. For the highway mileage distribution we see that: $\max(H) - \text{med}(H) = 16 > 7 = \text{med}(H) - \min(H)$, $\max(H) - Q_3(H) = 12 > 4 = Q_1(H) - \min(H)$, and $\overline{H} = 31.12 > 30 = \text{med}(H)$. All of these comparisons support our contention that the highway mileage distribution is skewed to the right. Notice that $Q_3(H) - \text{med}(H) = 4$ which is approximately equal to $\text{med}(H) - Q_1(H) = 3$; this suggests that the middle half of this distribution is also reasonably symmetric.

Next consider the locations of the subcompact car mileage distributions. The median city mileage $\text{med}(C) = 23$ is less than the median highway mileage $\text{med}(H) = 30$ and the mean city mileage $\overline{C} = 23.53$ is less than the mean highway mileage $\overline{H} = 31.12$. Both of these comparisons support our contention that the city mileages of subcompact cars tend to be lower than the highway mileages of subcompact cars. Notice that there is some overlap of the city mileages and the highway mileages indicating that some subcompact cars have city mileage values that are higher than the highway mileage values of some subcompact cars and *vice versa*.

Finally consider the variability in these subcompact car mileage distributions. The facts that: $\text{range}(C) = 23 = \text{range}(H)$, $\text{IQrange}(C) = 7 = \text{IQrange}(H)$, and $S_C = 5.27$ is approximately equal to $S_H = 5.18$, all support the contention that the variability in subcompact car city mileage values is about the same as the variability in subcompact car highway mileage values.

In our comparison of the city and highway mileages for subcompact cars, we ignored the fact that we actually have pairs of city and highway mileage values for each of the 51 car models. If we want to know how the highway mileage of a subcompact car model relates to its city mileage, then we need to base our comparison on the paired city and highway mileages. One way to do this is to consider the difference between the highway mileage and the city mileage for a car model. For each car model we will determine this difference value by subtracting its city mileage value from its highway mileage value. The highway minus city mileage differences for the $n = 51$ subcompact car models are given in Table 6. The 51 difference values in Table 6 are listed (reading across a row and then going to the next row) in the same order as the 51 city and highway mileage values are listed in Table 4, skipping the five unusual car models. A stem and leaf histogram for these differences is given in Figure 4 and the difference summary statistics are given in Table 7.

In the discussion below, we will use D to denote the difference $D = H - C$ between the highway mileage of a subcompact car model and its city mileage. From the stem and leaf histogram the shape of the subcompact car mileage difference distribution appears to be mound shaped and slightly skewed to the right. The facts: $\max(D) - \text{med}(D) = 5 > 3 = \text{med}(D) - \min(D)$, $\max(D) - Q_3(D) = 3 > 2 = Q_1(D) - \min(D)$, $Q_3(D) - \text{med}(D) = 2 > 1 = \text{med}(D) - Q_1(D)$, and $\overline{D} = 7.59 > 7 = \text{med}(D)$, all support our contention that the mileage difference distribution is slightly skewed to the right. This distribution has a single peak (mode) at 7 indicating that for these subcompact car models it is most common for the highway mileage value to exceed the city mileage value by 7 mpg. For the majority of these subcompact car models the mileage difference is fairly close to 7 mpg. However, there are a few car models for which this mileage difference is a good bit larger. For example, the car model with the largest highway minus city mileage difference is the Saturn SC model without DOHC for which the mileage difference is 12 mpg.

Table 6. Model year 2000 subcompact car EPA mileage differences (highway - city,) excluding the five unusual car models.

6	5	9	8	11	9	10	10	7	6
8	9	7	8	5	5	5	5	9	7
7	4	7	8	7	8	8	7	8	11
9	10	10	7	9	12	11	6	7	7
7	6	7	7	6	9	7	6	7	6
7									

Figure 4. Stem and leaf histogram for subcompact car EPA mileage differences (highway - city), excluding the five unusual car models.

The stem represents ones and the leaf represents tenths (mpg).

4	0
5	00000
6	0000000
7	0000000000000000
8	0000000
9	0000000
10	0000
11	000
12	0

Table 7. Subcompact car EPA mileage difference (highway - city) summary statistics, excluding the five unusual car models.

min =	4	Q_1−min =	2
Q_1 =	6	med−Q_1 =	1
med =	7	Q_3−med =	2
Q_3 =	9	max −Q_3 =	3
max =	12		
range =	8	mean =	7.59
IQR =	3	std dev =	1.80

Using the mean mileage difference $\overline{D} = 7.59$, we conclude that, on the average, the highway mileage of a subcompact car model is 7.59 mpg larger than its city mileage. Based on the median mileage difference $\text{med}(D) = 7$, we would conclude that the highway mileage of a subcompact car model is 7 mpg larger than its city mileage; with half of the models having a difference less than 7 and half having a difference larger than 7.

In this example, the difference between the highway and city mileage means, $\overline{H} - \overline{C} = 31.12 - 23.53 = 7.59$, is equal to the mean mileage difference, $\overline{D} = 7.59$; and the difference between the highway and city mileage medians, $\text{med}(H) - \text{med}(C) = 30 - 23 = 7$, is equal to the median mileage difference, $\text{med}(D) = 7$. In paired data situations like this the difference of the two means is always equal to the mean of the differences. On the other hand, the difference between the two medians does not always equal the median of the differences.

It is interesting to note that the five car models which we excluded as outliers due to their unusually low city and highway mileage values would not be unusual in terms of their mileage differences (one 6 and four 5's).

3.2 Modified box plots.

In this section we will consider a modified box plot designed to provide more information about the tails of the distribution. In the modified box plot a more complex method, which provides an indication of extreme observations, is used to construct the whiskers.

The simple box plot defined in Section 3.1 has a box, which extends from the first quartile Q_1 to the third quartile Q_3 divided into two parts by a line at the median, representing the middle of the distribution, and two whiskers, extending from the ends of the box to the most extreme values (the minimum and the maximum), representing the tails of the distribution.

Observations located near the ends of a distribution are said to be extreme. The whiskers in a simple box plot indicate the range of the lower and upper tails and the locations of the most extreme values but do not provide details about the behavior of observations near the extremes of the distribution. An extreme observation which is widely separated from the majority of the observations is said to be an **outlier**. Outliers deserve special consideration, since they may represent interesting exceptional cases or they may represent errors made in recording the data. Note that, depending on the spacing of the observations, extreme observations may or may not be considered outliers. Outliers are easy to spot in a stem and leaf histogram but not in a simple box plot.

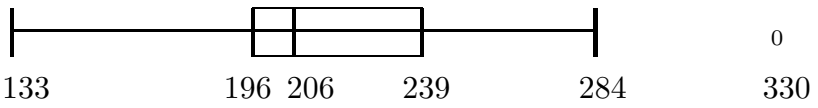
Before we can construct a modified box plot we need to quantify what we mean by an extreme observation. We will use multiples of the interquartile range (IQR) to distinguish between two types of extreme observations. First notice that the IQR is the range of the middle half of the data and the length of the box in the box plot. An observation which is much more than the IQR below the first quartile Q_1 or much more than the IQR above the third quartile Q_3 might reasonably be classified as an extreme observation. We will classify observations which are more than $1.5 \times \text{IQR}$ but less than $3 \times \text{IQR}$ below the first quartile or above the third quartile as somewhat extreme. That is, an observation between $Q_1 - 3 \times \text{IQR}$ and $Q_1 - 1.5 \times \text{IQR}$ or between $Q_3 + 1.5 \times \text{IQR}$ and $Q_3 + 3 \times \text{IQR}$ is said to be

somewhat extreme. We will classify observations which are more than $3 \times \text{IQR}$ below the first quartile or above the third quartile as very extreme. That is, an observation below $Q_1 - 3 \times \text{IQR}$ or above $Q_3 + 3 \times \text{IQR}$ is said to be very extreme.

The quantities $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$ are known as the lower and upper inner fences, and the quantities $Q_1 - 3 \times \text{IQR}$ and $Q_3 + 3 \times \text{IQR}$ are known as the lower and upper outer fences. To construct a modified box plot we first find the five number summary values and the lower and upper fences. To construct the upper whisker we first draw a line from the upper end of the box, Q_3, extending to the largest observation which is less than the upper inner fence $Q_3 + 1.5 \times \text{IQR}$. We then indicate observations beyond the upper inner fence using two symbols; one symbol, such as a 0, is used for observations between the upper inner fence and the upper outer fence and another symbol, such as a *, is used for observations beyond the upper outer fence. The lower whisker is constructed in an analogous fashion.

Consider the urban cholesterol level distribution for the Guatemalan cholesterol example. In this example we have $Q_1 = 196$, $Q_3 = 239$, $\text{IQR} = 43$, $1.5 \times \text{IQR} = 64.5$, and $3 \times \text{IQR} = 129$. The inner fences are $196 - 64.5 = 131.5$ and $239 + 64.5 = 303.5$. There are no observed cholesterol levels below the lower inner fence but there is one observed cholesterol level, 330 mg/dL, above the upper inner fence. The largest cholesterol level between Q_3 and the upper inner fence is 284 mg/dL. The outer fences are 67 and 368. There are no observed cholesterol levels outside the outer fences. The modified box plot for this cholesterol level distribution is given in Figure 5. In this example we would say that the one cholesterol level (330) marked as somewhat extreme is an outlier, since it is fairly widely separated from the other values.

Figure 5. Modified box plot for (all) urban cholesterol levels.



3.3 Numerical measures of relative position

There are many situations when we might wish to quantify the position of a particular value of a variable relative to a sample of values. For example, when presented with the results of a standardized test, we would like to know where our score stands relative to the scores of everyone else who took the test. We will discuss two different ways to quantify the relative position of a particular value of a variable.

The first measure of the relative position of a particular value X is the percentile rank of X which quantifies the location of X in an ordered listing of all of the values in the sample. The **percentile rank** of a particular value X is the percentage of the values in

the sample that are less than or equal to the particular value X. More specifically, if m of the n observed values in the sample are less than or equal to the particular value, then the percentile rank of the particular value is $(m/n)100\%$. Reports of scores on standardized tests often include the actual score and its percentile rank. The percentile rank of an individual's test score indicates how the individual performed on the test relative to the group by providing the percentage of the group that scored no higher than the individual.

Notice that the five number summary values, the minimum, Q_1, the median, Q_3, and the maximum, are the $0^{th}, 25^{th}, 50^{th}, 75^{th}$, and 100^{th} percentiles of the distribution. Therefore, the use of the five number summary values to describe a distribution is an example of the use of selected percentiles to describe a distribution.

Consider the relative standing, in the Guatemalan cholesterol example, of a hypothetical individual with a cholesterol level of 210 mg/dL. Using Table 1 or Figure 1 we find that: The percentile rank of 210 in the rural group is 91.84% ($45/49 = .9184$), since 45 of the 49 rural Guatemalans have cholesterol levels of 210 or less; and, the percentile rank of 210 in the urban group is 51.11% ($23/45 = .5111$), since 23 of the 45 urban Guatemalans have cholesterol levels of 210 or less. Almost all of the rural Guatemalans have cholesterol levels of 210 or less; thus it is clearly unusual for a rural Guatemalan to have a Cholesterol level which is higher than 210. On the other hand, roughly half of the urban Guatemalans have cholesterol levels of 210 or less.

We can also use this percentile rank idea to quantify the difference in location between these cholesterol level distributions. For example, 81.63% of the rural Guatemalans have cholesterol levels of 188 or less, while 80% of the urban Guatemalans have cholesterol levels above 188.

The second measure of the relative position of a particular value X is the Z–score of X which quantifies the location of X relative to the mean $\overline{X}$ of the sample in terms of the standard deviation S_X of the sample. Since the Z–score is based on $\overline{X}$ and S_X, the Z–score is only appropriate when $\overline{X}$ and S_X are appropriate measures of the center and variability in the sample, respectively. We will develop the Z–score in two stages.

First, we need a measure of the location of X relative to the center of the distribution as determined by the mean $\overline{X}$. The deviation, $X - \overline{X}$, of X from the mean $\overline{X}$ is such a measure. The deviation $X - \overline{X}$ is the signed distance from the particular value X to the mean $\overline{X}$. If $X - \overline{X}$ is negative, then X is below (smaller than) the mean. If $X - \overline{X}$ is positive, then X is above (larger than) the mean. In summary, the sign of the deviation $X - \overline{X}$ indicates the location of X relative to the mean $\overline{X}$; and the magnitude of the deviation $|X - \overline{X}|$ is the distance from X to the mean $\overline{X}$, measured in the units of measurement used for the observation X.

Second, we want a measure of the location of X relative to the mean $\overline{X}$ which takes the amount of variability in the data into account. We will obtain such a measure by

using the standard deviation S_X of the sample to standardize the deviation $X - \overline{X}$. Given a particular value X, the sample mean $\overline{X}$, and the sample standard deviation S_X, the **Z–score** corresponding to X is

$$Z = \frac{X - \overline{X}}{S_X}.$$

The sign of the Z–score indicates the location of X relative to the mean $\overline{X}$ and the magnitude of the Z–score is the distance from X to the mean $\overline{X}$ in terms of standard deviation units. For example, if $Z = 2$, then X is two standard deviation units above the mean ($X = \overline{X} + 2S_X$), and, if $Z = -2$, then X is two standard deviation units below the mean ($X = \overline{X} - 2S_X$).

Returning to the Guatemalan cholesterol example and the relative position of an individual with a cholesterol level of 210, let R denote the cholesterol level of a rural Guatemalan and let U denote the cholesterol level of an urban Guatemalan. The rural cholesterol mean is $\overline{R} = 157.02$ mg/dL and the rural cholesterol standard deviation is $S_R = 31.75$ mg/dL. The urban cholesterol mean is $\overline{U} = 216.87$ mg/dL and the urban cholesterol standard deviation is $S_U = 39.92$ mg/dL. The raw deviation of a cholesterol level of 210 from the rural mean is $210 - \overline{R} = 52.98$ mg/dL. Since this quantity is positive, we see that a cholesterol level of 210 mg/dL exceeds the rural mean by 52.98 mg/dL. The raw deviation of a cholesterol level of 210 from the urban mean is $210 - \overline{U} = -6.87$ mg/dL. Since this quantity is negative, we see that a cholesterol level of 210 mg/dL is 6.87 mg/dL below the urban mean.

Standardizing these raw deviations yields a Z–score of $52.98/31.75 = 1.67$ for a rural cholesterol level of 210 mg/dL and a Z–score of $-6.87/39.92 = -.17$ for an urban cholesterol level of 210 mg/dL. Notice that these Z–scores are unitless numbers (number of standard deviation units from the mean) which are directly comparable. Therefore, a rural cholesterol level of 210 mg/dL is 1.67 standard deviation units above the rural mean cholesterol level and an urban cholesterol level of 210 mg/dL is .17 standard deviation units below the urban mean cholesterol level. In terms of standard deviation units, we see that 210 mg/dL is about 10 times as far away from the mean cholesterol level for the rural group as it is for the urban group. In other words, when taking variability into account we find that it is much more unusual for a rural Guatemalan to have a cholesterol level of 210 than it is for an urban Guatemalan to have a cholesterol level of 210.

The remainder of this section is devoted to two interesting results which establish a connection between Z–scores and percentages. The first result, the $68\% - 95\% - 99.7\%$ rule, is an approximate rule not a mathematical fact. Strictly speaking, this rule only applies to distributions that are unimodal (single peaked), mound shaped, and symmetric. A formal statement of this rule is provided below.

The 68%-95%-99.7% rule. *For a distribution that is unimodal (has a single peak), mound shaped, and reasonably symmetric:*
i) Approximately 68% of the observed values will be within one standard deviation unit of the mean. That is, approximately 68% of the observed values will have a Z–score that is between -1 and 1.
ii) Approximately 95% of the observed values will be within two standard deviation units of the mean. That is, approximately 95% of the observed values will have a Z–score that is between -2 and 2.
iii) Approximately 99.7% of the observed values will be within three standard deviation units of the mean. That is, approximately 99.7% of the observed values will have a Z–score that is between -3 and 3. Notice that this indicates that almost all of the observed values will be within three standard deviations of the mean.

When it is applicable, the $68\% - 95\% - 99.7\%$ rule, can be used to determine the relative position of a particular value of a variable based on the corresponding Z–score. Notice that this rule indicates that a fairly large proportion (68%) of the sample will lie within one standard deviation of the mean; a very large proportion (95%) of the sample will lie within two standard deviations of the mean; and, almost all (99.7%) of the sample will lie within three standard deviations of the mean.

The rural and urban cholesterol distributions are both unimodal, mound shaped, and reasonably symmetric. For the rural group we find that 34 of the 49 cholesterol levels (69.39%) are within one standard deviation of the mean; 46 of the 49 cholesterol levels (93.88%) are within two standard deviation of the mean; and, all 49 cholesterol levels (100%) are within three standard deviation of the mean. For the urban group we find that 34 of the 45 cholesterol levels (75.56%) are within one standard deviation of the mean; 42 of the 45 cholesterol levels (93.33%) are within two standard deviation of the mean; and, all 45 cholesterol levels (100%) are within three standard deviation of the mean. Notice that the $68\% - 95\% - 99.7\%$ rule works better for the rural group cholesterol distribution, since it is more symmetric than the urban group cholesterol distribution. We would get better agreement of the urban group cholesterol levels with the $68\% - 95\% - 99.7\%$ rule if we excluded the outlier.

Example. Heights of adult males in the United Kingdom. The heights (in inches) of 8585 adult males born in the United Kingdom (including the whole of Ireland) are summarized in Table 8. This example is taken from Kendall and Stuart, *The Advanced Theory of Statistics, vol.1*, Griffin, (1977), 8. The data are from the *Final Report of the Anthropometric Committee to the British Association*, (1883), 256.

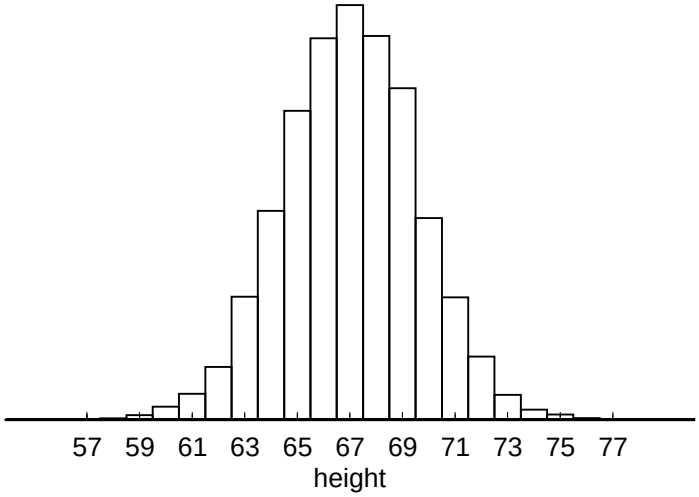
The histogram for the distribution of the 8585 heights of adult males for the United Kingdom height example in Figure 6 is unimodal, mound shaped, and symmetric. The sample mean height for this sample is 67.02 inches and the height standard deviation is

2.57 inches. For this example we find that 5835 of the 8585 heights (67.97%) are within one standard deviation of the mean height; 8307 of the 8585 heights (96.76%) are within two standard deviation of the mean height; and, 8542 of the 8585 heights (99.5%) are within three standard deviation of the mean height. Hence, the $68\% - 95\% - 99.7\%$ rule is quite accurate in its predictions for this UK height example.

Table 8. UK male heights.

height	frequency
57	2
58	4
59	14
60	41
61	83
62	169
63	394
64	669
65	990
66	1223
67	1329
68	1230
69	1063
70	646
71	392
72	202
73	79
74	32
75	16
76	5
77	2
total	8585

Figure 6. Histogram for UK heights.



The second result, Chebyshev's rule, is a mathematical fact that is true for any distribution. Unfortunately, the universal applicability of Chebyshev's rule forces its conclusions to be of more theoretical than practical interest. That is, the conclusions of Chebyshev's rule are valid for any distribution; but, they are often so imprecise that they are of limited practical use.

Chebyshev's rule. *For any distribution:*
i) At least 75% of the observed values will be within two standard deviation units of the mean. That is, at least 75% of the observed values will have a Z–score that is between -2 and 2.
ii) At least 89% of the observed values will be within three standard deviation units of the mean. That is, at least 89% of the observed values will have a Z–score that is between -3 and 3.
iii) In general, given a number $k > 1$*, at least* $[1 - (1/k^2)]100\%$ *of the observed values will be within k standard deviation units of the mean, i.e., at least this percentage of the observed values will have a Z–score that is between -k and k.*

3.4 Summary

In this chapter we introduced numerical summary values (statistics) and discussed the use of such statistics to quantify certain aspects of a distribution and to compare two distributions. Most of our attention focused on the shape of a distribution, the location of the distribution on the number line, and the amount of variability in the distribution. We began by defining the five number summary (minimum, Q_1, median, Q_3, maximum) which partitions the distribution into fourths. We then demonstrated how the five number summary and related statistics, such as the range and interquartile range, can be used

to summarize a distribution and to compare and contrast two distributions. A simple graphical representation of a distribution, the box plot, based on the five number summary was also introduced. We also defined the mean (a measure of location) and the standard deviation (a measure of variability).

Shape (skewness) Comparisons of the distances among the five number summary values can be used to assess and quantify skewness in a distribution as indicated below.

1. To quantify overall skewness in the distribution: compare (median – minimum) to (maximum – median).
2. To quantify skewness in the middle of the distribution: compare (median – Q_1) to (Q_3 – median).
3. To quantify skewness in the tails of the distribution: compare (Q_1 – minimum) to (maximum – Q_3).

Location The median and the mean are used to quantify the location of the center of a distribution on the number line. Recall that the median indicates the point which divides the distribution into halves (the histogram has equal area on each side of the median) while the mean indicates the point at which the distribution balances (the histogram has its center of gravity at the mean). If the distribution is symmetric, then the mean and the median are equal and either will suffice as a measure of the center of the distribution. If the distribution is heavily skewed, then the median is generally preferred over the mean as a measure of the center of the distribution. When comparing two distributions which have more or less the same shape either the median or the mean will suffice for comparing the locations of the distributions. But, when comparing distributions with different shapes the median is generally preferred over the mean for comparing the locations of the distributions.

Variability The range (maximum – minimum), interquartile range ($Q_3 - Q_1$), and standard deviation are used to quantify the variability in a distribution. For each of these statistics a larger value indicates more variability.

In Section 3.3 we discussed the use of percentile ranks and Z–scores to quantify the relative position of a particular value relative to the distribution of a sample. These ideas and in particular the Z–score, which indicates the location of a value relative to the mean in terms of standard deviation units, will reappear when we discuss inference in later chapters.

3.5 Exercises

For each of the examples in Section 1.2 which involve quantitative variables:

1. Determine the following summary statistics: the five number summary (minimum, Q_1, median, Q_3, maximum), the range, the interquartile range, the mean, and the standard deviation. (See the notes below for the examples with two or more groups.)

2. Discuss the distribution of the variable(s) using the summary statistics of question 1 to lend quantitative support to your discussion.

Notes:

For the examples with two or more groups (DiMaggio and Mantle, gear tooth strength), find the indicated summary statistics and compare and contrast the distributions of the variable for the two (or more) groups.

For the paired data examples (wooly–bear cocoons, homophone confusions) find the differences for each pair of data values and find the indicated summary statistics and describe the distribution of the differences.

Chapter 4
Sampling and Experimentation

4.1 Introduction

This chapter serves as a bridge between descriptive statistics and inferential statistics. In the preceding chapters we focused on descriptive methods (descriptive statistics) which are used to explore the distribution of the values of a variable among the units in the sample. In most applications the sample is selected from a well–defined population and the ultimate goal is to use the data to make inferences about the distribution of the values of the variable (or variables) among all of the units in the population. Recall that the **population** is the collection of all of the units that are of interest and the **sample** is a subset of the population for which we have or will obtain data. Thus the purpose of inferential statistics is to use the data, which characterize the sample, to make inferences about the population.

We will concentrate on making inferences about particular aspects or characteristics of the population which can be quantified in terms of population parameters. A **parameter** is a numerical characteristic of the population. Recall that a **statistic** is a numerical characteristic of the sample. Thus, parameters and statistics are analogous quantities which quantify certain aspects of the population (parameter) or the sample (statistic). Since the goal of inference is to characterize certain aspects of the population, the first thing we need to do is to choose a variable that is suitable for inference in the sense that the values of the variable contain information about relevant characteristics of the population. Recall that a **variable** is a characteristic of a unit. Given a suitable variable, we can measure or observe the values of the variable for the units in the sample. These values can then be used to determine the value of a statistic and this statistic can be used to make inferences about the corresponding population parameter. For example, we might use the statistic as an estimate of the corresponding parameter or we might use the statistic to assess the evidence for a particular conjecture about the value of the parameter. Notice that once the sample is obtained and the data are collected we can determine the value of the statistic. On the other hand, we will never know the value of the parameter unless we take a census, *i.e.*, unless the sample is the whole population.

4.2 Sampling

Sampling is the process of obtaining a sample from a population. Our ultimate goal is to use the sample (which we will examine) to make inferences about the population (which we will not examine in its entirety). If the sample is selected from the population in an

appropriate fashion, then we can use the information in the sample to make reliable and quantifiable inferences about the population. When the sample is obtained we will use the distribution of the variable among the units in the sample to make inferences about the distribution of the variable among the units in the population. If the distribution of the variable in the sample was exactly the same as the distribution of the variable in the population, then it would be easy to make inferences about the population; but, this is clearly too much to ask. Therefore we need to determine how to select a sample so that the sample is representative of the population.

The first step in deciding whether a method of choosing a sample will yield a representative sample requires a distinction between two populations. Before we obtain a sample we need to decide exactly which population we are interested in. The **target population** is the collection of all of the units that we want to make inferences about. We then need to determine which population our sample actually comes from. The **sampled population** is the collection of all of the units that *could* be in the sample. Notice that the sampled population is determined by the method used to select the sample.

Ideally the sampling method is chosen so that the sampled population is exactly the same as the target population and we can refer to this collection as the population. In practice, there may be some differences between the target population and the sampled population. When the sampled population is not identical to the target population we cannot be confident that the sample (which comes from the sampled population) will be representative of the target population. Furthermore, we cannot be confident that the statistic (which is based on a sample from the sampled population) will be suitable for inference about the parameter (which corresponds to the target population).

If there is a difference between the sampled population and the target population, in the sense that the distribution of the variable in the sampled population is different from the distribution of the variable in the target population, then a sample (obtained from the sampled population) is said to be **biased** for making inferences about the target population. If we use a biased sample to make inferences about the target population, the resulting inferences will not be appropriate. For example, a statistic based on a biased sample, may provide a suitable estimate of the corresponding parameter in the sampled population; but, it may not provide a suitable estimate of the corresponding parameter in the target population. Therefore, if the sampled population is different from the target population, then we must modify our goals by redefining the target population or we must change the sampled population by modifying our sampling method, since we want these two populations to be the same so that our inferences will be valid for our target population. It may be possible to change the method of obtaining our sample so that all of the units in the target population could be in our sample and these two populations are the same. If it is not possible to change the sampling method, then we must change our

goals by restricting our inferences to the sampled population. In any case, once a sampling method has been chosen, the sampled population is determined and we should restrict our inferences to this sampled population. In conclusion, when making inferences from a sample we must carefully consider the restrictions imposed by the sampling method, since statistical theory can only justify inferences about the sampled population.

Example. Medical malpractice insurance. An insurance company that provides medical malpractice insurance is interested in determining how common it is for a medical doctor to be involved in a malpractice suit. The company plans to obtain a random sample of 500 doctors from the listing in a professional association directory. We will not assume that doctors are required to belong to this association.

In this example a medical doctor is a unit. The implied target population is all medical doctors in the region that is of interest to this insurance company, *e.g.*, if the company is considering offering insurance to all medical doctors with a medical practice in the US, then the target population is all medical doctors with a medical practice in the US. The sampled population is all medical doctors who are listed in the professional association directory. We know that doctors are not required to belong to this association. Furthermore, it may be possible for doctors who do belong to this medical association to not be listed in the current directory. There may also be doctors listed in the directory who are no longer in medical practice or who do not practice in the region of interest. Therefore information based on a random sample from doctors listed in the directory may not be appropriate if the goal is to describe the population of all medical doctors practicing in the region of interest. For example, if medical doctors who have never been sued for malpractice are more likely to be listed than those who have been sued for malpractice, then information based on a random sample from doctors listed in the directory may not be appropriate if the goal is to describe the population of all medical doctors in the region of interest. Therefore, inferences based on a random sample of doctors selected from those listed in this directory should be restricted to only those doctors who are listed in the directory (the sampled population).

Assuming that we have defined a method of selecting a sample so that the sampled population is the same as the target population, we next need to consider exactly how we should select the units that constitute the sample. Since we are assuming that the sampled and target populations are the same, we do not need to worry about the type of bias described above. However, we might introduce bias if we do not select the units for the sample in an appropriate fashion. The approach to sampling that we will adopt is called random sampling. The idea behind random sampling is to eliminate potential bias (intentional or unintentional) from the selection process by using *impersonal random chance* to select the sample. In addition to eliminating bias random sampling also provides the basis for theoretical justification and quantification of inferences based on the sample.

All of the sampling situations we consider can be viewed as being abstractly the same as the simple situation of selecting a sample of balls from a box of balls.

Example. Balls in a box. A box contains a collection of balls. In this situation a unit is a ball and the collection of balls in the box is the population. Our sample will consist of n balls selected from the box.

When a ball is selected we need to "measure" it, *i.e.*, we need to determine the value of the variable for this ball. Before we can do so we need to define a suitable variable. The definition of a variable consists of a description of the variable and an indication of its possible values. For example, suppose that the balls in the box are of various colors; say red, blue, and green. If we are interested in the characteristic color, then we might define the qualitative variable "color of the ball" with possible values of red, blue, and green. On the other hand, suppose that the balls in the box are of various weights. If we are interested in the characteristic weight, then we might define the quantitative variable "weight of the ball" with possible values equal to potential weights in grams.

Given a variable, suitable parameters and statistics are defined to correspond to the variable. With respect to the color of a ball, the proportion of red balls in the box (in the population) is a parameter and the proportion of red balls among the n balls selected from the box (in the sample) is a statistic. With respect to the weight of a ball, the mean weight of the balls in the box (in the population) is a parameter and the mean weight of the n balls selected from the box (in the sample) is a statistic.

The simplest type of random sample is called a simple random sample. A **simple random sample of size n** is a sample of n units selected from the population in such a way that every possible sample of n units has the same chance of being selected. This definition of a simple random sample can be refined to distinguish between two versions of simple random samples. If we require that the possible samples of n units are such that a particular unit can occur at most once in a sample, then we refer to the sample as being a **simple random sample of size n, selected without replacement.** On the other hand, if we allow a particular unit to occur more than once in the sample, then we refer to the sample as a **simple random sample of size n, selected with replacement.**

Example. Balls in a box (revisited). To obtain a **simple random sample of size n, selected without replacement** from the balls in our box, we first mix the balls in the box and select one ball at random (so that each ball in the box has the same chance of being selected). We then remove the selected ball from the box giving us one ball in our random sample. Then we mix the remaining balls in the box and select one ball at random (again so that each ball remaining in the box has the same chance of being selected) and remove the selected ball from the box giving us two balls in our random sample. This process of choosing a ball at random and removing it from the box is continued until n

balls have been selected. These n balls form the simple random sample of size n, selected without replacement.

To obtain a **simple random sample of size n, selected with replacement** from the balls in our box, we first mix the balls in the box and select one ball at random (so that each ball in the box has the same chance of being selected). We then measure or observe the value of the variable for the selected ball giving us the value of the variable for one of the balls in our random sample. Then we return the selected ball to the box, mix the balls in the box, and select one ball at random (again so that each ball in the box has the same chance of being selected) and measure or observe the value of the variable for the selected ball giving us two values of the variable corresponding to the two balls selected for our random sample. This process of choosing a ball, measuring the value of the variable for the ball, and returning the ball to the box is continued until n balls have been selected and measured. These n measurements (values of the variable) correspond to the balls that form the simple random sample of size n, selected with replacement.

If the population from which we wish to select a random sample is not too large, then it is possible to envision actually labeling each unit in the population, placing these labels on a collection of balls, placing these labeled balls in a box, and selecting a simple random sample of these balls as described above. In fact, state lotteries, where a simple random sample of numbers is selected from a collection of allowable numbers (the units), are conducted in this way. If you have ever observed the complicated mechanisms used to select winning lottery numbers, you know that it is difficult to convince people that a method of "drawing balls from a box" yields a proper simple random sample.

We will now discuss a simple alternative to sampling from an actual population of balls based on a sequence of random digits. A **sequence of random digits** is a list of the ten digits $0, 1, 2, \ldots, 9$ with the following two properties.

1. For any given position in the list, each of the ten digits, $0, 1, 2, \ldots, 9$, has the same chance of being in that position.
2. The entries in the list are independent of each other. That is, knowing the values in any part of the list would provide us with no information about any other values in the list beyond the information implied by property 1.

These two properties also hold if we think of the sequence of random digits as a list of two digit numbers $(00, 01, 02, \ldots, 99)$ or as a list of three digit numbers $(000, 001, \ldots, 999)$ or as a list of such numbers with any fixed number of digits.

To use a sequence of random digits to select a simple random sample we first need to assign suitable numerical labels to all of the units in the population. A list of all of the units in the population along with their labels is called a **sampling frame**. To insure that we get a random sample we need to use labels that have the same number of digits. That is, if N is the number of units in the population, then

1. If $N \leq 10$, we should use N of the one digit labels $0, 1, 2, \ldots, 9$.
2. If $11 \leq N \leq 100$, we should use N of the two digit labels $00, 01, 02, \ldots, 99$.
3. If $101 \leq N \leq 1000$, we should use N of the three digit labels $000, 001, \ldots, 999$, and so on.

One place to find a sequence of random digits is in a random number table. A **random number table** is simply a table containing a sequence (list) of random digits. These tables are usually formatted into rows and columns with periodic spaces. This arrangement has no significance, it only serves to make the table easier to use. An alternative to finding and using a random number table is to use a computer program or a calculator to generate random digits or random numbers. We will first discuss the use of a random number table to select a simple random sample.

After suitable labels have been assigned to the units in the population, as discussed above, we then proceed to the random number table to determine the labels that correspond to the units to be included in the sample. We choose a starting point in the table and go through the table (the list) in a systematic fashion reading the appropriate number of digits as we go. For example, we can select some row as our starting point. Then, reading across the row we make a note of the first digit, or digits if the labels have more than one digit, then the second digit or digits, and so on. If we reach the end of the row before obtaining n labels, we simply go on to the next row. If we come upon a digit or digits that was not used as a label, we simply skip it. If we want to sample without replacement, we also skip any label that we have already selected. This process is continued until n labels have been selected from the random number table. The units in the sampling frame that correspond to the n labels we selected from the random number table form the simple random sample.

Table 1. Random digits.

05797	43984	21575	09908	70221	19791	51578	36432	33494	79888
10395	14289	52185	09721	25789	38562	54794	04897	59012	89251
35177	56986	25549	59730	64718	52630	31100	62384	49483	11409
25633	89619	75882	98256	02126	72099	57183	55887	09320	73463
16464	48280	94254	45777	45150	68865	11382	11782	22695	41988

Table 1 contains a small portion of the random number table in the book *A Million Random Digits with 100,000 Normal Deviates*, Rand Corporation, (1955). To illustrate the use of a random number table to select a simple random sample we will select a sample of $n = 10$ students from the students in the Stat 214 example who were registered in section 1. The $N = 36$ students listed on the first page of Table 1 of Chapter 1 form the population. If we use the line numbers in this table as labels, then this page of Table 1

is our sampling frame. Since there are 36 units in this population, we will use two digit labels. Starting at the beginning of the first row of random digits in Table 1 and reading two digits at a time across the first row and the beginning of the second row yields the labels: 05, 79, 74, 39, 84, 21, 57, 50, 99, 08, 70, 22, 11, 97, 91, 51, 57, 83, 64, 32, 33, 49, 47, 98, 88 10, 39, 51, 42, 89, 52, 18, 50, 97, 21, 25. Therefore, if we want a simple random sample of size 10, selected with replacement, then the students listed in lines 5, 21, 8, 22, 11, 32, 33, 10, 18, and 21 form our sample. If we want a simple random sample of size 10, selected without replacement, then we skip the second 21 and the students listed in lines 5, 21, 8, 22, 11, 32, 33, 10, 18, and 25 form our sample.

In this particular example when we read numbers from the random number table we had to skip a lot of pairs of digits since these numbers were not used as labels. In this example $N = 36$ and since $2 \times 36 = 72 < 100$ we can assign two labels to each unit. That is, assign 01 and 37 $(37 = 36 + 1)$ as labels for unit 1 (the student listed in line 1), assign 02 and 38 $(38 = 36 + 2)$ as labels for unit 2, and so on, assigning 36 and 72 as labels for unit 36. If we use these labels we only have to read through part of the first row to get the valid labels 05, 39, 21, 57, 50, 08, 70, 22, 11, 51, 57, and 64. Translating these labels to line numbers (by subtracting 36 if the number is greater than 36) shows that, using this method, the students listed in lines 5, 3, 21, 21, 14, 8, 34, 22, 11, and 15 form our sample of size 10, selected with replacement and the students listed in lines 5, 3, 21, 14, 8, 34, 22, 11, 15, and 28 form our sample of size 10, selected without replacement.

An alternative to finding and using a random number table is to use a computer program or a calculator to generate suitable random numbers. Computer programs will usually provide a list of random numbers with the desired number of digits automatically. Many calculators provide random numbers that are between zero and one. To obtain a random number with the appropriate number of digits from a random number between zero and one, simply read off the appropriate number of digits from the beginning of the number. For example, if a calculator provided the number .12345678 and we needed a three digit number, the number would be 123. You should be aware that the algorithms or methods that computer programs and calculators use to generate a sequence of random numbers vary in their quality. Some of these algorithms are not very successful at generating a valid sequence of random numbers.

When we take a simple random sample, all of the possible samples have the same chance of being selected. There are situations where it is not appropriate for all of the possible samples to have the same chance of being selected. Suppose that there are two or more identifiable subsets of the population (subpopulations). If we obtain a simple random sample from the whole population, then it is possible for the resulting sample to come entirely from one of the subpopulations, or for the sample not to contain any units from one or more of the subpopulations. If we know or suspect that the distribution of

the variable of interest varies among the subpopulations, then a sample which does not contain any units from some of the subpopulations will not be representative of the whole population. Therefore, in a situation like this we should not use a simple random sample to make inferences about the whole population. Instead we should use a more complex kind of random sample. One possibility is to use a sampling method known as **stratified random sampling** which is described below in the context of a simple example.

Suppose we wish to estimate the proportion of all registered voters in the United States who favor a particular candidate in an upcoming presidential election. We might expect there to be differences in the proportion of registered voters who favor this candidate among the various states. For example, we might expect support for this candidate to be particularly strong in his or her home state. Because we are interested in the proportion of all registered voters in the United States who favor this candidate, we want to be sure that all of the states are represented fairly in our sample.

We can use the states to define **strata** (subpopulations), take a random sample from each state (stratum), and then combine these samples to get a sample that is representative of the entire country. This is an example of a stratified random sample. The simplest type of **stratified random sample** is obtained as described in the following three steps.

1. Divide the population into appropriate strata (subpopulations).
2. Obtain a simple random sample within each stratum.
3. Combine these simple random samples to get the stratified random sample from the whole population.

To obtain a representative sample in the opinion poll example, we would need to determine the number of registered voters in each state and select simple random samples of sizes that are proportional to the numbers of registered voters in the states.

4.3 Experimentation

The sampling approach to data collection discussed in the preceding section is often used to perform an observational study. The steps involved in conducting an observational study based on a random sample are summarized below.

1. Obtain a random sample of units from the population of interest.
2. Obtain the data. That is, determine the values of the variable for the units in the sample.
3. Use the data to make inferences about the population. More specifically, use the distribution of the variable in the sample to make inferences about the distribution of the variable in the population from which the sample was taken.

In an **observational study** we obtain a sample of units, observe the values of a variable, and make inferences about the population. The purpose of such an observational

study is to observe the units in the sample and, based on these observations, to make inferences about what we would observe if we examined the entire population. On the other hand, in an **experimental study** we manipulate the units and observe their response to this manipulation. In the experimental context, a particular combination of experimental conditions is known as a **treatment**. The purpose of an experiment is to obtain information about how the units in the population would respond to a treatment; or, to compare the responses of the units to two or more treatments. The response of a unit to a particular treatment is determined by measuring the value of a suitable **response variable**.

The steps involved in conducting a simple experimental study based on a random sample are summarized below.

1. Obtain a random sample of units from the population of interest.
2. Subject the units in the sample to the experimental treatment of interest.
3. Obtain the data. That is, determine the values of the response variable for the units in the sample.
4. Use the data to make inferences about the how the units in the population would respond to the treatment. More specifically, use the distribution of the response variable in the sample to make inferences about the distribution of the response variable in the population from which the sample was taken. In this context it may be easiest to think of the population as the hypothetical population of values of the response variable which would result if all of the units in the population were subjected to the treatment.

We will now discuss the basic ideas of experimentation in more detail in the context of a simple hypothetical experiment. Suppose that a new drug has been developed to reduce the blood pressure of hypertensive patients. The treatment of interest is the administration of the new drug to a hypertensive patient. The change in a patient's blood pressure will be used as the response variable. For this example the plan of the simple experiment described above is summarized in the steps below.

1. Obtain a random sample of n hypertensive patients.
2. Measure the blood pressure of each patient before the new drug is administered.
3. Administer the new drug to each of these patients.
4. After a suitable period of time, measure the blood pressure of each patient.
5. For each patient determine the change in his or her blood pressure by computing the difference between the patient's blood pressure before the drug was administered and the patient's blood pressure after the new drug was administered. This change or difference will serve as the response variable for assessing the effects of the new drug. In this example the relevant population is the hypothetical population of changes in blood pressure that we would observe if all of the hypertensive patients in the population from which the sample was selected had been subjected to this experiment.

Suppose that we actually conducted this experiment. Furthermore, suppose that the data indicate that the hypertensive patients' blood pressures tend to decrease after they are given the new drug, *i.e.*, suppose that the data indicate that most of the patients experienced a meaningful reduction in blood pressure. We can conclude that there is an association between the new drug and a reduction in blood pressure. This association is clear, since the patients (as a group) tended to experience a decrease in blood pressure after they received the new drug. Can we conclude that the new drug caused this decrease in blood pressure? The support for the contention that the new drug caused the decrease in blood pressure is not so clear. In addition to the new drug there may be other factors associated with the observed decrease in blood pressure. For example, the decrease in blood pressure might be explained, in whole or in part, as the physical manifestation of the psychological effect of receiving medication. In other words, we might observe a similar decrease in blood pressure if we administered a placebo to the patients instead of the new drug. It is also possible that some other aspects of the experimental protocol are affecting the patients' blood pressures. The way that this experiment is being conducted does not allow us to separate out the effects of the many possible causes of the decrease in blood pressure. If we hope to establish a cause and effect relationship between taking the new drug and observing a decrease in blood pressure, then we need to use a comparative experiment.

In a **randomized comparative experiment** baseline data is obtained at the same time as the data concerning the treatment of interest. This is done by randomly dividing the available units (patients) into two or more groups and comparing the responses for these groups. In the drug example there is one treatment of interest, administer the new drug. Therefore, in this situation we only need two groups, a control group and a treatment group. The units (patients) in the **control group** do not receive the treatment (do not receive the new drug). The units (patients) in the **treatment group** do receive the treatment (do receive the drug). During the course of the experiment we need to keep all aspects of the experiment, other than the treatment itself, as similar as possible for all of the units in the study. The idea is that, if the only difference between the units in the control group and the units in the treatment group is that the units in the treatment group received the treatment, then any observed differences between the responses of the two groups must be caused by the treatment. In the drug example it would be a good idea to administer a placebo to the patients in the control group, so that they do not know that they did not receive the new drug. It would also be a good idea to "blind" the patients and the people administering the drug or placebo by not telling them which patients are receiving the placebo and which patients are receiving the new drug. The purpose of such blinding is to eliminate intentional or unintentional effects due to patient

or administrative actions which might affect a patient's response. The steps for conducting such a **randomized comparative experiment** are given below.

1. Randomly divide the group of available patients into two groups: a group of n_1 patients to serve as the control group and a group of n_2 patients to serve as the treatment group. These two groups are random samples of sizes n_1 and n_2 from the group of available patients. The samples sizes n_1 and n_2 may be different.
2. Administer the placebo to the patients in the control group and administer the new drug to the patients in the treatment group.
3. Obtain the data. That is, measure the response variable for each of the $n_1 + n_2$ patients in the experiment. For example, we could determine the change (difference) in each patient's blood pressure as measured before and after administration of the placebo or new drug.
4. Compare the responses of the patients in the treatment group to the responses of the patients in the control group and make inferences about the effects of the new drug versus the placebo.

In this example there are two hypothetical populations of changes in blood pressure. The hypothetical population of changes in blood pressure that we would observe if all of the available hypertensive patients were subjected to this experiment and given the placebo and the hypothetical population of changes in blood pressure that we would observe if all of the available hypertensive patients were subjected to this experiment and given the new drug. Notice that, strictly speaking, our inferences in this example only apply to the hypertensive patients who were available for assignment to the groups used in the experiment. If we want to make inferences about a larger population of hypertensive patients, then the group of available patients used in the study should form a random sample from this larger population.

The experiment described above is designed to compare the effects of the new drug to the effects of a placebo. Suppose that we wanted to compare the effects of the new drug to the effects of a standard drug. To make this comparison we could design the experiment with three groups: a control group, a treatment group for the new drug, and a treatment group for the standard drug. If our only goal is to compare the two drugs (treatments), then we could eliminate the placebo control group and run the experiment with the two treatment groups alone.

Example. Cloud seeding. The data referred to in this example are given in Simpson, Olsen, and Eden (1975), *Technometrics*, **17**, 161–166. These data were collected in southern Florida between 1968 and 1972 to determine whether injection of silver iodide into cumulus clouds tends to increase rainfall. Fifty–two days that were deemed suitable for

cloud seeding were randomly divided into two groups of 26 days. An airplane, equipped to inject silver iodide into a target cloud, was flown through the target cloud. For one group of 26 days the device used to inject the silver iodide was loaded and the target cloud was seeded. For the other group of 26 days the device used to inject the silver iodide was not loaded and the target cloud was not seeded. On all 52 days the airplane flew through the target cloud. Furthermore, the pilots and technicians on the plane were not aware of whether the device used to inject the silver iodide was loaded or not. For each day the amount of rainfall (total volume of rain falling from the cloud base), measured in acre–feet, was determined.

In this example there are 52 days that were deemed suitable for cloud seeding. Each of these days is a unit and this group of 52 days is the group of "available units" which were used in the experiment. The response variable is the amount of rainfall measured after the airplane was flown through the cloud. The two relevant hypothetical populations for which inferences could be made in this example are: the collection of 52 rainfall amounts which would have been obtained if the plane had been flown through the cloud but the cloud had not been seeded with silver iodide, and the collection of 52 rainfall amounts which would have been obtained if the plane had been flown through the cloud and the cloud had been seeded with silver iodide.

We can define two population mean rainfall amounts (parameters) corresponding to these populations, *i.e.*, the mean of the 52 rainfall amounts which would have been obtained if the cloud was not seeded on each of the 52 days and the mean of the 52 rainfall amounts which would have been obtained if the cloud was seeded on each of the 52 days. The two sample mean rainfall amounts (statistics) based on the rainfall amounts recorded for each of the two groups of 26 days, *i.e.*, the mean of the 26 rainfall amounts recorded on the 26 days when the cloud was not seeded and the mean of the 26 rainfall amounts recorded on the 26 days when the cloud was seeded, could be used to make inferences about the corresponding population mean rainfall amounts.

Since the 52 days on which this experiment was conducted did not form a random sample from some larger population of days suitable for cloud seeding, we cannot justify extending our inferences beyond these 52 days. We might reasonably argue that our inferences apply to similar days in the area where the study was conducted; but, we cannot use statistical theory to justify extrapolations to days other than these 52.

4.4 Summary

Reliable and quantifiable inferences about a population (about the population distribution of a variable) require careful consideration of the definition of the relevant population and of the method used to obtain the data on which the inferences are based.

A sampling study is conducted by selecting a random sample of units from a population, observing the values of a variable for the units in the sample, and then making inferences or generalizations about the population. More specifically, the distribution of the values of the variable among the units in a random sample is used to make inferences about the distribution of the variable among the units in the population. The first consideration in planning or interpreting the results of a sampling study is the determination of exactly which units could be in the sample. The collection of all units which could be in the random sample is known as the sampled population and this sampled population is the relevant population for inferences based on the sample. The second consideration concerns the proper selection of the units which constitute the sample. We cannot properly quantify inferences unless the sample is a properly selected random sample from the population.

An experimental study differs from a sampling study in that the units used in the experimental study are manipulated and the responses of the units to this experimental manipulation are recorded. For an experimental study the relevant population or populations are hypothetical populations of values of the variable defined by the experimental treatment(s) and corresponding to all of the units available for use in the experiment. That is, the relevant population(s) is the population of values of the variable which would be observed if all of the available units were subjected to the experimental treatment(s). In the context of a comparative experiment we cannot properly quantify inferences unless the units are assigned to the treatments being compared using an appropriate method of random assignment. This random assignment of units to treatments is analogous to the random sampling of a sampling study.

Chapter 5

Inference for a Proportion

5.1 Introduction

A **dichotomous population** is a collection of units which can be divided into two nonoverlapping subcollections corresponding to the two possible values of a dichotomous variable, *e.g.* male or female, dead or alive, pass or fail. It is conventional to refer to one of the two possible values which dichotomize the population as "success" and the other as "failure." These generic labels are not meant to imply that success is good. Rather, we can think of choosing one of the two possible classifications and asking "does the unit belong to the subcollection of units with this classification?" with the two possibilities being yes (a success) and no (a failure). Thus, generically, we can refer to the two subcollections of units which comprise the dichotomous population as the **success group** and the **failure group**. When a unit is selected from the population and the unit is found to be a member of the success group we say that a **success** has occurred. Similarly, when a member of the failure group is selected we say that a **failure** has occurred.

The proportion of units in the population that belong to the success group is the **population success proportion**. This population success proportion is denoted by the lower case letter p. The population success proportion p is a parameter, since it is a numerical characteristic of the population. Notice that the **population failure proportion** $1 - p$ is also a parameter.

The **sample success proportion** or observed proportion of successes in a sample from a dichotomous population is denoted by $\hat{p}$ (read this as p hat). The observed proportion of successes in the sample $\hat{p}$ is a statistic, since it is a numerical characteristic of the sample.

We will consider two forms of inference about the population success proportion p of a dichotomous population. In Section 5.2 we will consider the use of the observed success proportion $\hat{p}$ to estimate the value of the population success proportion p. In Section 5.3 we will consider the use of the observed success proportion $\hat{p}$ to assess the support for conjectures about the value of the population success proportion p.

The approach to inference that we will use here and in other contexts in the sequel is based on the observed value of a statistic and the sampling distribution of the statistic. The **sampling distribution** of a statistic is the distribution of the possible values of the statistic that could be obtained from random samples. We can think of the sampling distribution of a statistic as a theoretical relative frequency distribution for the possible values of the statistic which describes the sample to sample variability in the statistic. The

form of the sampling distribution of a statistic depends on the nature of the population the sample is taken from, the size of the sample, and the method used to select the sample.

The mean and the standard deviation of the sampling distribution are of particular interest. The mean of the sampling distribution indicates whether the statistic is biased as an estimator of the parameter of interest. If the mean of the sampling distribution is equal to the parameter of interest, then the statistic is said to be **unbiased** as an estimator of the parameter. Otherwise, the statistic is said to be **biased** as an estimator of the parameter. To say that a statistic is **unbiased** means that, even though the statistic will overestimate the parameter for some samples and will underestimate the parameter for other samples, it will do so in such a way that, in the long run, the values of the statistic will average to give the correct value of the parameter. When the statistic is **biased** the statistic will tend to consistently overestimate or consistently underestimate the parameter; therefore, in the long run, the values of a biased statistic will not average to give the correct value of the parameter. The standard deviation of the sampling distribution is known as the **standard error** of the statistic. The standard error of the statistic provides a measure of the sample to sample variability in the values of the statistic. The standard error of the statistic can be used to quantify how close we can expect the value of the statistic to be to the value of the parameter.

Note regarding formulae and calculations. Throughout this book selected formulae and intermediate calculations are provided to clarify ideas and definitions. Some readers may find it useful to reproduce these calculations; however, this is not necessary, since a modern statistical calculator or computer statistics program will perform these calculations and provide the desired answer.

5.2 Estimating a proportion

When sampling from a dichotomous population a primary goal is to estimate the population success proportion p, *i.e.*, to estimate the proportion of units in the population success group. The observed proportion of successes in the sample $\hat{p}$ is the obvious estimate of the corresponding population success proportion p.

Clearly there will be some variability from sample to sample in the computed values of the statistic $\hat{p}$. That is, if we took several random samples from the same dichotomous population, we would not expect the computed sample proportions, the $\hat{p}$'s, to be exactly the same. Two questions about $\hat{p}$ as an estimator of p that we might ask are: (1) Can we expect the sample success proportion $\hat{p}$ to be close to the population success proportion p? and (2) Can we quantify how close $\hat{p}$ will be to p? The sampling distribution of $\hat{p}$, which describes the sample to sample variability in $\hat{p}$, can be used to address these questions.

In the introduction to this chapter we mentioned that the sampling distribution of a statistic depends on the way in which the sample is selected, as well as the nature

of the population being sampled. Therefore, before we continue with our discussion of the behavior of $\hat{p}$ as an estimator of p we need to describe a model for sampling from a dichotomous population. This model will be presented in terms of a sequence of n trials. In this context a **trial** is a process of observation or experimentation which results in one of two distinct outcomes (success or failure).

A sequence of n trials is said to constitute a sequence of **n Bernoulli trials with success probability p** if the following conditions are satisfied.

1. There is a common probability of success p for every trial. That is, on every trial the probability that the outcome of the trial will be a success is p.
2. The outcomes of the trials are independent of each other. That is, if we knew the outcome of a particular trial or trials this would provide no additional information about the probability of observing a success (or failure) on any other trial. For example, if we knew that a success (or failure) occurred in the first trial, this would not change the probability of success in any other trial.

The simple examples described below will help to clarify the definition of a sequence of n Bernoulli trials and the connection between sampling from a dichotomous population and Bernoulli trials.

Example. Tossing a fair die. Let a trial consist of tossing a fair (balanced) die and observing the number of dots on the upturned face. Define a success to be the occurrence of a 1, 2, 3, or 4. Since the die is fair, the probability of a success on a single trial is $p = 4/6 = 2/3$. Furthermore, if the die is always tossed in the same fashion, then the outcomes of the trials are independent. Therefore, with success defined as above, tossing the fair die n times yields a sequence of n Bernoulli trials with success probability $p = 2/3$.

Example. Drawing balls from a box. Consider a box containing balls (the population) of which 2/3 are red (successes) and 1/3 are green (failures). Suppose that a simple random sample of size n is selected with replacement from this box. That is, a ball is selected at random, its color is recorded, the ball is returned to the box, the balls in the box are mixed, and this process is repeated until n balls have been selected. Thinking of each selection of a ball as a trial we see that this procedure is abstractly the same as the die tossing procedure described above. That is, the outcomes of the draws are independent, and every time that a ball is drawn the probability of a success (drawing a red ball) is $p = 2/3$. Therefore, selecting a simple random sample of n balls with replacement from this collection of balls can be viewed as observing a sequence of n Bernoulli trials with success probability $p = 2/3$. In general, taking a simple random sample of size n selected with replacement from a population with success proportion p can be viewed as observing a sequence of n Bernoulli trials with success probability p.

Situations like the die tossing example above do not fit into the sample and population setup that we have been using. That is, in the die tossing example there is not a physical population of units from which a sample is obtained. In a situation like this we can think of the outcomes of the n Bernoulli trials (the collection of successes and failures that make up the sequence of outcomes of the n trials) as a sample of values of a variable. The probability model specifies that the probability of success on a single trial is p and the probability of failure is $1-p$. This model describes the population of possible values of the variable. Therefore, we can envision a dichotomous population of values (successes and failures) such that the population success proportion is p; and we can think of the outcome of a single trial as the selection of one value at random from this dichotomous population of values. With this idea in mind, we see that the success probability p of this probability model is a parameter and the observed proportion of successes in the n trials is a statistic.

Returning to our discussion of the sampling distribution of $\hat{p}$ we first present two important properties of this sampling distribution. The observed proportion of successes $\hat{p}$ in a sequence of n Bernoulli trials with success probability p (or equivalently the observed proportion of successes $\hat{p}$ in a simple random sample selected with replacement from a dichotomous population with population success proportion p) has a sampling distribution with the following properties.

1. The mean of the sampling distribution of $\hat{p}$ is the population success probability p. Therefore, $\hat{p}$ is unbiased as an estimator of p.
2. The **population standard error** of $\hat{p}$, denoted by S.E.$(\hat{p})$, is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

The population standard error of $\hat{p}$ depends on n, which will be known, and p, which will be unknown. Notice that the population standard error gets smaller when n gets larger. That is, when sampling from a fixed dichotomous population, the variability in $\hat{p}$ as an estimator of p is smaller for a larger sample size than it is for a smaller sample size. This property reflects the fact that a larger sample provides more information than a smaller sample. The dependence of the population standard error of $\hat{p}$ on the population success probability p is more complicated. The quantity $p(1-p)$ attains its maximum value of $1/4$ when $p = 1/2$ and approaches zero as p approaches zero or one. Therefore, for a fixed sample size n, there will be more variability in $\hat{p}$ as an estimator of p when p is close to $1/2$ than there will be when p is close to zero or one. This behavior reflects the fact that a dichotomous population is most homogeneous when p is near the extremes $p = 0$ and $p = 1$, and is least homogeneous when p is close to $1/2$.

In many sampling situations the sample is not selected with replacement. For example, in an opinion poll we would not allow the same person to respond twice. We will now consider the effects of sampling without replacement.

Example. Drawing balls from a box (revisited). We will now consider how the ball drawing example from above changes when the simple random sample is selected without replacement. As before, let the box containing the balls (the population) be such that 2/3 are red (successes) and 1/3 are green (failures). However, suppose that the simple random sample of n balls is selected without replacement. That is, a ball is selected at random and its color is recorded and this process is repeated, without returning the ball to the box, until n balls have been selected. It is readily verified that the resulting simple random sample of size n selected without replacement cannot be viewed as a sequence of n Bernoulli trials. To see this suppose that the box contains 12 balls of which 8 are red and 4 are green. The probability of selecting a red ball on the first draw, denoted by $P(\text{red first})$, is $P(\text{red first}) = 8/12 = 2/3$. The probability that the second ball drawn is red clearly depends on the color of the first ball that was drawn. If the first ball drawn was red, then $P(\text{red second given red first}) = 7/11$. However, if the first ball drawn was green, then $P(\text{red second given green first}) = 8/11$. Notice that these probabilities are not the same and neither of them is equal to the population success proportion $p = 2/3$. Therefore, when the sample is selected without replacement, the sampling process is not the same as observing a sequence of Bernoulli trials, since the draws are not independent (the probability of drawing a red ball (success) depends on what happened in the earlier draws) and, as a consequence of this lack of independence, the probability of red (success) is not the same on each draw (trial).

The sampling distribution of the observed success proportion $\hat{p}$ is not the same when $\hat{p}$ is based on a sample selected without replacement as it is when $\hat{p}$ is based on a sample selected with replacement. In both sampling situations, the mean of the sampling distribution of $\hat{p}$ is the population success proportion p. Thus $\hat{p}$ is unbiased as an estimator of p whether the sample is selected with or without replacement. On the other hand, the standard error of $\hat{p}$ is not the same when $\hat{p}$ is based on a sample selected without replacement as it is when $\hat{p}$ is based on a sample selected with replacement. (The standard error of $\hat{p}$ is smaller when the sample is selected without replacement than it is when the sample is selected with replacement.) More specifically, unlike the formula for the standard error of $\hat{p}$ when the sample is selected with replacement which does not depend on the size of the population being sampled, when sampling without replacement the standard error of $\hat{p}$ depends on the size of the population. Fortunately, if the size of the population is very large relative to the size of the sample, then, for practical purposes, the probability of

obtaining a success is essentially constant, the outcomes of the draws are essentially independent, and we can use the standard error formula based on the assumption of sampling with replacement even though the sample was selected without replacement.

Remark. *When $\hat{p}$ is computed from a simple random sample of size n selected without replacement from a dichotomous population of size N, the population standard error of $\hat{p}$, S.E.$(\hat{p}) = \sqrt{fp(1-p)/n}$, is smaller than the population standard error for a sample selected with replacement by a factor of $\sqrt{f}$, where $f = (N-n)/(N-1)$. The factor f is known as the finite population correction factor and its effect is most noticeable when N is small relative to n. If N is very large relative to n, then $f \approx 1$ and the two standard errors are essentially equal. Actually, if N is very large relative to n and the data correspond to a simple random sample, then the sampling distribution of $\hat{p}$ is essentially the same whether the sample is selected with or without replacement.*

The sampling distribution of $\hat{p}$ can be represented in tabular form as a probability distribution or in graphical form as a probability histogram. The **probability distribution** of $\hat{p}$ is a theoretical relative frequency distribution which indicates the probability or theoretical relative frequency with which each of the possible values of $\hat{p}$ will occur. The **probability histogram** of $\hat{p}$ is the theoretical relative frequency histogram corresponding to the probabilities (theoretical relative frequencies) in the probability distribution. It is possible to determine the exact sampling distribution of $\hat{p}$, in fact, it is even possible to find a formula which gives the probabilities of each of the possible values of $\hat{p}$. However, for our purposes it is more convenient to work with an approximation to the sampling distribution of $\hat{p}$. (The exact sampling distributions of $\hat{p}$ are discussed in Chapter 4a.) Before we discuss this approximation, which is based on the standard normal distribution, we need to briefly discuss the standard normal distribution.

The normal distribution is widely used as a model for the probability distribution of a continuous variable. We will discuss normal distributions in general in more detail in Chapter 7. Here we will restrict our attention to the standard normal distribution and its use as an approximation to the sampling distribution of $\hat{p}$.

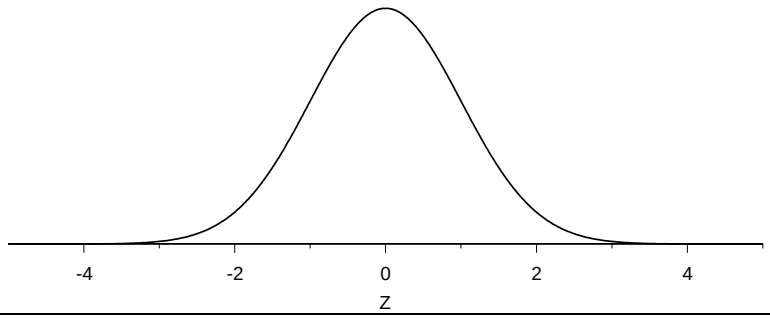
Before we discuss the standard normal distribution we first need to briefly consider the representation of a continuous probability model via a density curve. A **density curve** is a nonnegative curve for which the area under the curve (over the x–axis) is one. We can think of the density curve as a smooth version of a theoretical probability histogram with the rectangles of the histogram replaced by a smooth curve indicating where the tops of the rectangles would be. With a continuous variable it does not make sense to talk about the probability that the variable would take on a particular value, after all if we defined positive probabilities for the infinite collection (continuum) of possible values of the variable these probabilities could not add up to one. It does, however, make sense to

talk about the probability that the variable will take on a value in a specified range. Given two constants $a < b$ the probability that the variable will take on a value in the interval from a to b is equal to the area under the density curve over the interval from a to b on the x–axis. In this fashion the density curve gives the probabilities which a single value of the variable, chosen at random from the infinite population of possible values, will satisfy.

When we use a continuous probability model to approximate the distribution of a discrete statistic, such as $\hat{p}$, we use the area under the density curve, over the appropriate interval on the number line, to approximate the area in the discrete probability histogram over the same interval. The idea here is that, if the density curve of the approximating continuous distribution matches the discrete probability histogram well, then the area under the density curve will provide a good approximation of the corresponding area in the histogram.

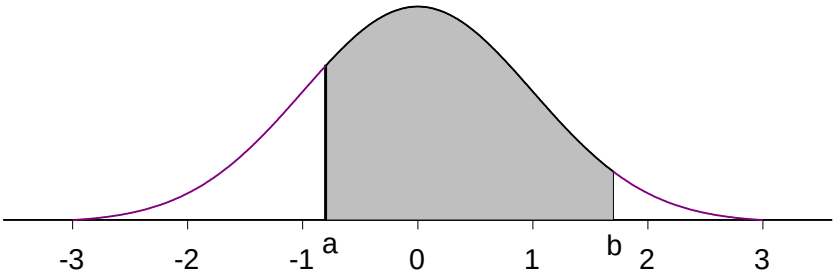
We will now discuss the standard normal distribution which we will use to approximate the sampling distribution of $\hat{p}$. The standard normal distribution can be characterized by its density curve which is the familiar bell shaped curve exhibited in Figure 1. The standard normal distribution and its density curve are symmetric around zero, *i.e.*, if we draw a vertical line through zero in Figure 1, then the two sides of the density curve are mirror images of each other. From Figure 1 it may appear that the standard normal density curve ends at -3 and 3; however, this density curve is actually positive (above the x–axis) for all possible values. The area under the standard normal density curve from -3 to 3 is .9974; thus, there is a small but positive area under the density curve outside of the interval from -3 to 3.

Figure 1. The standard normal density curve.



We will use the upper case letter Z to denote a variable which behaves in accordance with the standard normal distribution and we will refer to such a Z as a standard normal variable. The probability that the standard normal variable Z will take on a value between a and b, denoted by $P(a \leq Z \leq b)$ (read this as the probability that Z is between a and b), is the area under the standard normal density curve over the interval from a to b. A probability of the form $P(a \leq Z \leq b)$ is depicted, for particular values of a and b, as the area of the shaded region in Figure 2.

Figure 2. $P(a \leq Z \leq b)$**, drawn for** $a < 0$ **and** $b > 0$**.**



Computer programs and many calculators can be used to compute standard normal probabilities or equivalently to compute areas under the standard normal density curve. These probabilities can also be calculated using tables of standard normal distribution probabilities. We will not need to perform such calculations here.

The inferential methods we will consider are based on a large sample size normal approximation to the sampling distribution of $\hat{p}$. The normal approximation to the sampling distribution of $\hat{p}$, which is stated formally below, simply says that, for large values of n, the standardized value of $\hat{p}$ obtained by subtracting the population success proportion p from $\hat{p}$ and dividing this difference by the population standard error of $\hat{p}$, behaves in approximate accordance with the standard normal distribution. That is, for large values of n the quantity $(\hat{p}-p)/\text{S.E.}(\hat{p})$ behaves in approximate accordance with the standard normal distribution.

The normal approximation to the sampling distribution of $\hat{\mathbf{p}}$. *Let $\hat{p}$ denote the observed proportion of successes in a sequence of n Bernoulli trials with success probability p (or equivalently the observed proportion of successes in a simple random sample drawn with replacement from a dichotomous population with population success proportion p) and let $a < b$ be two given constants. If n is sufficiently large, then the probability that $(\hat{p}-p)/S.E.(\hat{p})$ is between a and b is approximately equal to the probability that a standard normal variable Z is between a and b. In symbols, using $\approx$ to denote approximate equality, the conclusion from above is that, for sufficiently large values of n,*

$$P\left(a \leq \frac{\hat{p}-p}{S.E.(\hat{p})} \leq b\right) \approx P(a \leq Z \leq b).$$

Remark. *If the population being sampled is very large relative to the size of the sample, then, for practical purposes, this normal approximation to the sampling distribution of $\hat{p}$ may also be applied when $\hat{p}$ is based on a simple random sample selected without replacement.*

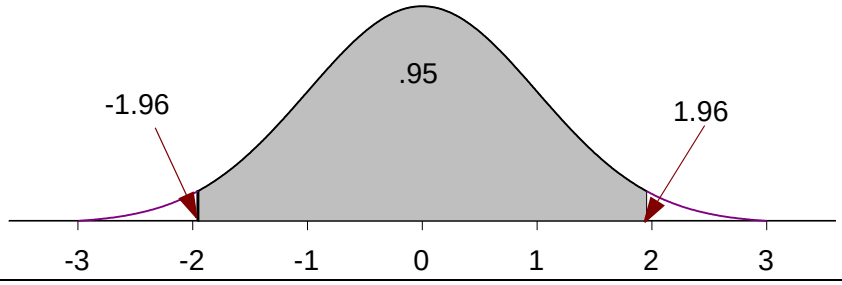
If we apply this approximation with $a = -k$ and $b = k$, then we find that the probability that $\hat{p}$ will take on a value within k population standard error units of p

(within kS.E.$(\hat{p})$ units of p) is approximately equal to the probability that a standard normal variable Z will take on a value between $-k$ and k, *i.e.*,

$$P\big(|\hat{p} - p| \leq k\text{S.E.}(\hat{p})\big) = P\big(p - k\text{S.E.}(\hat{p}) \leq \hat{p} \leq p + k\text{S.E.}(\hat{p})\big) \approx P(-k \leq Z \leq k).$$

The most commonly used choice of the constant k in this probability statement is $k = 1.96$. The probability that the standard normal variable takes on a value between -1.96 and 1.96 is equal to .95, *i.e.*, $P(-1.96 \leq Z \leq 1.96) = .95$; therefore, the probability that $\hat{p}$ will take on a value within 1.96 population standard error units of p is approximately .95. This probability is indicated graphically as the shaded region of area .95 in Figure 3. Two other common choices of the constant k in this probability statement are $k = 1.645$ and $k = 2.576$, which give probabilities (areas) of .90 and .99, respectively.

Figure 3. $P(-1.96 \leq Z \leq 1.96) = .95$



We now return to our discussion of estimating the population success proportion. The following discussion is under the assumption that the data come from a simple random sample of size n drawn with replacement from a dichotomous population with population success proportion p or equivalently that the data correspond to a sequence of n Bernoulli trials with success probability p. For practical purposes, the confidence interval estimates described below are also applicable when the data come from a simple random sample of size n drawn without replacement, provided the population is very large.

Remark. *The basic ideas underlying the inferential methods discussed in this chapter can be used to formulate confidence intervals and hypothesis tests when the data correspond to more complex types of random samples. However, the inferential methods discussed in this chapter are not appropriate for most national opinion polls and sample surveys which rely on complex stratified and/or cluster sampling.*

The observed proportion of successes in our sample $\hat{p}$ provides a single number estimate of the population success probability p. We can think of $\hat{p}$ as our "best guess" of the value of p. From the sampling distribution of $\hat{p}$ we know that $\hat{p}$ is unbiased as an estimator of p; therefore, on the average in the long run (under repeated sampling) we know that $\hat{p}$ provides a good estimate of the unknown parameter p. This unbiasedness, however, does

not guarantee that the observed value of $\hat{p}$, based on a single sample, will be close to the true, unknown value of p.

Instead of reporting a single estimate of the unknown population success proportion p it would be more useful to report a range or interval of plausible values for p. In particular, given the data we would like to be able to say, with a reasonable level of confidence, that the true value of p is between two particular values. A **confidence interval estimate of** p consists of two parts. There is an interval of plausible values for p and a corresponding level of confidence. The **confidence level** indicates our confidence that the unknown p actually belongs to the corresponding interval. We will adopt the usual convention of using a confidence level of 95%. A 95% confidence interval estimate of p is an interval of plausible values for p constructed using a method which guarantees that 95% of such intervals will actually contain the unknown proportion p. That is, a 95% confidence interval is an interval constructed using a method of generating such intervals with the property that this method will work, in the sense of generating an interval that contains p, for 95% of all possible samples.

The starting point for using the normal approximation to the sampling distribution of $\hat{p}$ to construct a 95% confidence interval estimate of p is the approximate probability statement

$$P\big[|\hat{p} - p| \leq 1.96\text{S.E.}(\hat{p})\big] \approx .95.$$

This probability statement indicates that the probability that the statistic $\hat{p}$ is within $1.96\text{S.E.}(\hat{p})$ units of the parameter p is approximately .95. In other words, when we take a simple random sample and compute $\hat{p}$ the value we get will be within $1.96\text{S.E.}(\hat{p})$ of the true p approximately 95% of the time. This is equivalent to saying that the probability that the parameter p is within $1.96\text{S.E.}(\hat{p})$ units of the statistic $\hat{p}$ is approximately .95, which is exactly the type of statement we are looking for. Unfortunately, this interval of values is not computable, since it involves the population standard error $\text{S.E.}(\hat{p})$ which depends on the unknown parameter p and is, therefore, also unknown.

We will consider two methods of forming a confidence interval for p. For ease of notation and greater generality we will let C denote the desired confidence level and k the corresponding cutoff point for the standard normal distribution, *i.e.*, C and k are chosen such that $P(-k \leq Z \leq k) = C$, where Z denotes a standard normal variable. (Some common choices of C and k are: $C = .95$ and $k = 1.96$ for 95% confidence, $C = .90$ and $k = 1.645$ for 90% confidence, and $C = .99$ and $k = 2.576$ for 99% confidence.) In terms of these symbols the starting point for using the normal approximation to the sampling distribution of $\hat{p}$ to construct a confidence interval estimate of p is the approximate probability statement

$$P\big[|\hat{p} - p| \leq k\text{S.E.}(\hat{p})\big] \approx C.$$

The first confidence interval estimate we consider is **the Wilson interval**. This interval estimate is obtained by re–expressing the basic inequality

$$|\hat{p} - p| \leq k\text{S.E.}(\hat{p})$$

as an interval of values for p. The **Wilson confidence interval** is given by

$$\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \leq p \leq \tilde{p}_k + \text{M.E.}(\tilde{p}_k),$$

(read $\tilde{p}_k$ as p tilde sub k) where

$$\tilde{p}_k = \frac{n\hat{p} + \frac{k^2}{2}}{n + k^2}$$

determines the center of the interval, and the **margin of error of** $\tilde{p}_k$

$$\text{M.E.}(\tilde{p}_k) = \sqrt{\tilde{p}_k^2 - \frac{n\hat{p}^2}{n + k^2}}$$

determines the length of the interval.

If we use $k = 1.96$ in these expressions, then we can claim that we are 95% confident that the population success proportion p is between $\tilde{p}_k - \text{M.E.}(\tilde{p}_k)$ and $\tilde{p}_k + \text{M.E.}(\tilde{p}_k)$. There is some chance for confusion about what this statement actually means. The important thing to remember is that it is the statistic $\tilde{p}_k$ and the margin of error $\text{M.E.}(\tilde{p}_k)$ that vary from sample to sample. The population proportion p is a fixed, unknown parameter which does not vary. Therefore, the 95% confidence level applies to the method used to generate the confidence interval estimate. That is, the method (obtain a simple random sample and compute the numbers $\tilde{p} - \text{M.E.}(\tilde{p})$ and $\tilde{p} + \text{M.E.}(\tilde{p})$) used to generate the limits of the confidence interval is such that 95% of the time it will yield a pair of confidence interval limits which bracket the population success proportion p. Therefore, when we obtain a sample, compute the confidence interval, and say that we are 95% confident that this interval contains p what we mean is that we feel "pretty good" about claiming that p is in this interval, since the method works for 95% of all possible samples and so it probably worked for our sample.

Derivation of the Wilson interval. Since $|\hat{p} - p| \geq 0$ and $\text{S.E.}(\hat{p}) = \sqrt{p(1-p)/n}$, we can square each side of the basic inequality to get the equivalent inequality

$$(\hat{p} - p)^2 \leq \frac{k^2}{n}(p - p^2).$$

Straightforward algebra allows us to re–express this inequality as the following quadratic inequality in p

$$(n + k^2)p^2 - 2(n\hat{p} + \frac{k^2}{2})p + n\hat{p}^2 \leq 0.$$

Treating this inequality as an equality and solving for p gives the two values

$$\tilde{p}_k \pm \text{M.E.}(\tilde{p}_k),$$

$$\text{where} \quad \tilde{p}_k = \frac{n\hat{p} + \frac{k^2}{2}}{n + k^2} \quad \text{and} \quad \text{M.E.}(\tilde{p}_k) = \sqrt{\tilde{p}_k^2 - \frac{n\hat{p}^2}{n + k^2}}.$$

Thus the original probability statement

$$P\big[|\hat{p} - p| \le k\text{S.E.}(\hat{p})\big] \approx C.$$

is equivalent to the probability statement

$$P\big[\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \le p \le \tilde{p}_k + \text{M.E.}(\tilde{p}_k)\big] \approx C.$$

The endpoints of this interval, which are functions of $n, \hat{p}$, and k, are computable. Therefore, the Wilson confidence interval is given by

$$\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \le p \le \tilde{p}_k + \text{M.E.}(\tilde{p}_k).$$

It is somewhat tedious to compute the Wilson interval by hand; but it is easy to program a calculator or computer to do the computations. An easy to compute approximation (the **Agresti–Coull interval**) to the Wilson 95% confidence interval is described after the following examples.

Example. Insects in an apple orchard. The manager of a large apple orchard is concerned with the presence of a particular insect pest in the apple trees in the orchard. An insecticide that controls this particular insect pest is available. However, application of this insecticide is rather expensive. It has been determined that the cost of applying the insecticide is not economically justifiable unless more than 20% of the apple trees in the orchard are infested. The manager has decided to assess the extent of infestation in the orchard by examining a simple random sample of 200 apple trees. In this example a unit an apple tree and the target population is all of the apple trees in this orchard. We will assume that the simple random sample is selected from all of the apple trees in the orchard so that the sampled population is the same as the target population. We will also assume that the 200 trees in the sample form a small proportion of all of the trees in the entire orchard so that we do not need to worry about whether the sample is chosen with or without replacement. An appropriate dichotomous variable is whether an apple tree is infested with possible values of yes (the tree is infested) and no (the tree is not infested). Since we are interested in the extent of the infestation we will view a tree that is infested

as a success. Thus, the population success proportion p is the proportion of all of the apple trees in this orchard that are infested.

Two (related) questions of interest in this situation are:
(1) What proportion of all of the trees in this orchard are infested? (What is p?)
(2) Is there sufficient evidence to justify the application of the insecticide? (Is $p > .20$?)
We will consider four hypothetical outcomes for this scenario to demonstrate how a 95% confidence interval estimate can be used to address these questions.

Case 1. Suppose that 35 of the 200 apple trees in the sample are infested so that $\hat{p} = .175$. In this case we know that 17.5% of the 200 trees in the sample are infested and we can conjecture that a similar proportion of all of the trees in the entire orchard are infested. However, we need a confidence interval estimate to get a handle on which values of the population success proportion p are plausible when we observe 17.5% infested trees in a sample of size 200. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .1811$ and a 95% confidence interval ranging from .1286 to .2336. Thus we can conclude that we are 95% confident that between 12.86% and 23.36% of all of the trees in this orchard are infested. Notice that this confidence interval does not exclude the possibility that more than 20% of the trees in the entire orchard are infested, since the upper limit of the confidence interval 23.36% is greater than 20%. In other words, even though less than 20% of the trees in the sample were infested when we take sampling variability into account we find that it is possible that more than 20% (as high as 23.36%) of the trees in the entire orchard are infested.

Case 2. Suppose that 26 of the 200 apple trees in the sample are infested so that $\hat{p} = .13$. In this case we know that 13% of the 200 trees in the sample are infested. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .1370$ and a 95% confidence interval ranging from .0903 to .1837. Thus we can conclude that we are 95% confident that between 9.03% and 18.37% of all of the trees in this orchard are infested. In this case the entire confidence interval is below 20% excluding the possibility that more than 20% of the trees in the entire orchard are infested. Therefore, in this case we have sufficient evidence to conclude that less than 20% of the trees in the entire orchard are infested, *i.e.*, that $p < .20$.

Case 3. Suppose that 45 of the 200 apple trees in the sample are infested so that $\hat{p} = .225$. In this case we know that 22.5% of the 200 trees in the sample are infested. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .2302$ and a 95% confidence interval ranging from .1726 to .2877. Thus we can conclude that we are 95% confident that between 17.26% and 28.77% of all of the trees in this orchard are infested. Notice that this confidence interval does not exclude the possibility that less than 20% of the trees in the entire orchard are infested, since the lower limit of the confidence interval 17.26% is less than 20%. In other words, even though mores than 20% of the trees in the sample were infested

when we take sampling variability into account we find that it is possible that less than 20% of the trees in the entire orchard are infested.

Case 4. Suppose that 54 of the 200 apple trees in the sample are infested so that $\hat{p} = .27$. In this case we know that 27% of the 200 trees in the sample are infested. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .2743$ and a 95% confidence interval ranging from .2132 to .3354. Thus we can conclude that we are 95% confident that between 21.32% and 33.54% of all of the trees in this orchard are infested. In this case the entire confidence interval is above 20% excluding the possibility that less than 20% of the trees in the entire orchard are infested. Therefore, in this case we have sufficient evidence to conclude that more than 20% of the trees in the entire orchard are infested, *i.e.*, that $p > .20$.

Example. Opinions about a change in tax law. Consider a public opinion poll conducted to assess the support for a proposed change in state tax law among the taxpayers in a particular metropolitan area. The target population is the group of approximately 200,000 taxpayers in the particular metropolitan area. Suppose that the opinion poll is conducted as follows: first a simple random sample of 100 taxpayers in the metropolitan area is obtained, restricting the sample to taxpayers who have telephones, then these 100 taxpayers are contacted by telephone and each person is asked to respond to the question "Do you favor or oppose the proposed change in state tax law?" In this example we will define a unit to be an individual taxpayer in this metropolitan area. (Note that, technically, a unit is a household, since more than one taxpayer may share the same telephone number.) The variable is the response of the taxpayer to the indicated question with possible values of: "I favor the change," "I oppose the change," and "I do not have an opinion regarding the change." We will dichotomize this variable (and the population) by recording the responses as either "I favor the change" or "I do not favor the change." Notice that in this example the target and sampled populations are not the same. Since there might well be a relationship between having a telephone and opinion about the proposed tax law change, we will restrict our attention to the sampled population of all taxpayers in this metropolitan area who have telephones. The parameter of interest is the proportion p of all taxpayers in this metropolitan area who have a telephone who favor the proposed tax law change at the time of the survey. In this example the random sample would be selected without replacement. However, since the size of the population, approximately 200,000, is much larger than the sample size $n = 100$, we can use the confidence interval estimation procedure as described above.

Suppose that the poll was conducted and 55 of the 100 taxpayers in the sample responded that they favor the tax law change. The observed proportion who favor the change is thus $\hat{p} = .55$, *i.e.*, 55% of the 100 taxpayers in the sample favored the change. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .5482$ and a 95% confidence interval ranging from .4524 to .6438. Therefore, we are 95% confident that the actual proportion

of taxpayers in this metropolitan area (who have telephones) who favored the proposed change in state tax law at the time of the survey is between 45.24% and 64.38%. Notice that this confidence interval contains values for p that are both less than .5 and greater than .5. Therefore, based on this outcome of the opinion poll there is not sufficient evidence to conclude that more than half of the taxpayers in this metropolitan area (who have telephones) favored the proposed tax law change at the time of this poll.

Now suppose that the poll was conducted and 64 of the 100 taxpayers in the sample responded that they favor the tax law change. The observed proportion who favor the change is thus $\hat{p} = .64$, *i.e.*, 64% of the 100 taxpayers in the sample favored the change. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .6348$ and a 95% confidence interval ranging from .5424 to .7273. Therefore, we are 95% confident that the actual proportion of taxpayers in this metropolitan area (who have telephones) who favored the proposed change in state tax law at the time of the survey is between 54.24% and 72.73%. In this case all of the values for p in the confidence interval are greater than .5. Therefore, based on this outcome of the opinion poll there is sufficient evidence to conclude that more than half of the taxpayers in this metropolitan area (who have telephones) favored the proposed tax law change at the time of this poll. However, we would also note that, based on this confidence interval, the actual percentage in favor of the change might be as small as 54.24%.

In the preceding analysis of this opinion poll example we dichotomized the responses to the question by combining the "oppose" and "no opinion" responses as "do not favor". As an alternative we might prefer to restrict our attention to only those people who are willing to express a definite opinion by restricting our inference to the subpopulation of taxpayers who would have been willing to respond "I favor the change" or "I oppose the change" at the time of the survey. Thus we redefine the population success proportion p as the proportion of all taxpayers in this metropolitan area (who have a telephone and have reached an opinion at the time of the survey) who favor the proposed tax law change at the time of the survey. To implement this approach we simply ignore the part of the sample for which the respondents did not express an opinion, redefine the sample size as n^* the number who responded "favor" or "oppose", and compute the confidence interval conditional on the reduced sample size n^*. For example, if $n = 100$ and if 64 taxpayers favor the change, 26 taxpayers oppose the change, and 10 taxpayers have no opinion, then we restrict our attention to the $n* = 64 + 26 = 90$ taxpayers who expressed an opinion. For this sample the observed proportion who favor the change is $\hat{p} = 64/90 = .7111$. Using the Wilson method with $k = 1.96$ (and $n* = 90$) we get $\tilde{p}_k = .7025$ and a 95% confidence interval ranging from .6104 to .7946. Therefore, we are 95% confident that the actual proportion of taxpayers in this metropolitan area (who have telephones and have reached

an opinion) who favored the proposed tax law change at the time of this poll is between 61.04% and 79.46%.

Remark. *When a confidence interval for a proportion p is based on a simple random sample selected with replacement or a simple random sample selected without replacement from a much larger population the precision of the confidence interval as an estimate of p depends on the absolute size of the sample not the size of the sample relative to the size of the population. For example, if a simple random sample of size $n = 200$ yields $\hat{p} = .65$ (and $k = 1.96$), then $\tilde{p}_k = .6472$, M.E.$(\tilde{p}_k) = .0655$, and we are 95% confident that p is between $.6472 - .0655 = .5817$ and $.6472 + .0655 = .7127$. Any sample of size $n = 200$ for which $\hat{p} = .65$ yields this confidence interval; which has length $2(.0655) = .1310$. Thus if we were sampling from a population of 200,000 or a population of 2,000,000 and if we obtained $\hat{p} = .65$, we would get the same confidence interval. Hence the precision of the confidence interval, as measured by its length, depends on the sample size but does not depend on what fraction of the population was sampled.*

We will now consider a simpler method for computing a confidence interval for p. This confidence interval estimate, known as the **Wald interval**, is in widespread use and many calculators and computer programs will compute it. Unfortunately, this confidence interval estimate has some undesirable properties and we **do not recommend** its use.

As we noted above (for $C = .95$ and $k = 1.96$), the probability statement

$$P\big[|\hat{p} - p| \le k\text{S.E.}(\hat{p})\big] \approx C$$

is equivalent to the probability statement

$$P\big[\hat{p} - k\text{S.E.}(\hat{p}) \le p \le \hat{p} + k\text{S.E.}(\hat{p})\big] \approx C,$$

but this interval of values is not computable, since the population standard error $\text{S.E.}(\hat{p}) = \sqrt{p(1-p)/n}$ depends on the unknown parameter p. The **Wald interval** is obtained by replacing the unknown population standard error by an estimated standard error. The **estimated standard error of** $\hat{p}$

$$\widehat{\text{S.E.}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is obtained by replacing the unknown parameter p in the population standard error by the observable statistic $\hat{p}$. The **Wald confidence interval** is given by

$$\hat{p} - \text{M.E.}(\hat{p}) \le p \le \hat{p} + \text{M.E.}(\hat{p}), \quad \text{where} \quad \text{M.E.}(\hat{p}) = k\widehat{\text{S.E.}}(\hat{p})$$

is the **margin of error of** $\hat{p}$.

Notice that the Wald interval is centered at $\hat{p}$ and its length is determined by the margin of error of $\hat{p}$,

$$\text{M.E.}(\hat{p}) = k\widehat{\text{S.E.}}(\hat{p}).$$

As with the Wilson interval, if we use $k = 1.96$ in these expressions, then we can claim that we are 95% confident that the population success proportion p is between $\hat{p} - \text{M.E.}(\hat{p})$ and $\hat{p} + \text{M.E.}(\hat{p})$. With the same interpretation of "95% confident" as before.

We will now discuss the "undesirable properties" of this interval estimate and the reason we do not recommend it. Even though $\hat{p}$ performs well as a single number estimate of p the Wald confidence interval estimate, based on $\hat{p}$ and $\widehat{\text{S.E.}}(\hat{p})$, does not perform well. When we say that we are 95% confident that the population success proportion p is between $\hat{p} - \text{M.E.}(\hat{p})$ and $\hat{p} + \text{M.E.}(\hat{p})$ we realize that our indicated 95% confidence level is actually an approximation to the true confidence level. For this confidence interval estimate the indicated 95% confidence level differs from the actual confidence level because of the two approximations used to construct this interval, *i.e.*, because of our use of the normal approximation and our use of the estimated standard error. We would hope, at least for reasonably large values of n, that the difference between the indicated 95% confidence level of our interval estimate and its actual confidence level would be small. Unfortunately, this is not necessarily the case and the actual confidence level of this confidence interval estimate may be quite different from the indicated 95%. In particular, the actual confidence level of this 95% confidence interval estimate may be much smaller than 95%. Furthermore, this discrepancy between the indicated 95% confidence level and the actual confidence level is not necessarily negligible even when the sample size n is quite large.

On the other hand, the Wilson confidence interval estimate, based on $\tilde{p}_k$ and $\text{M.E.}(\tilde{p}_k)$, only requires one approximation (the normal approximation) and for this reason it performs better than the Wald confidence interval.

We will now describe the easy to compute approximation (the **Agresti–Coull interval**) to the Wilson 95% confidence interval. If we add 4 artificial observations to the data, 2 success and 2 failures, and then compute the Wald 95% confidence interval, it turns out that we obtain a reasonably accurate approximation of the Wilson 95% confidence interval. More formally, the **Agresti–Coull interval** is obtained by replacing the estimator $\hat{p}$ and its margin of error in the Wald interval by the alternate estimator $\tilde{p}$ (read this as p tilde) and its margin of error. The estimator $\tilde{p}$ is obtained by adding 2 successes and 2 failures to the data, *i.e.*,

$$\tilde{p} = \frac{\text{the number of successes plus 2}}{\text{the number of observations plus 4}} = \frac{n\hat{p} + 2}{n + 4}.$$

The corresponding 95% **margin of error of** $\tilde{p}$ is

$$\text{M.E.}(\tilde{p}) = 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}},$$

which is analogous to the margin of error of $\hat{p}$ with $\tilde{p}$ in place of $\hat{p}$ and $n+4$ in place of the actual sample size n. The **Agresti–Coull 95% confidence interval** estimate of p, is given by

$$\tilde{p} - \text{M.E.}(\tilde{p}) \le p \le \tilde{p} + \text{M.E.}(\tilde{p}),$$

where $\tilde{p}$ and $\text{M.E.}(\tilde{p})$ are as defined above. The reason that this works is that the value of k for a 95% confidence interval is $k = 1.96$ which implies that $k^2/2 \approx 2$ and $k^2 \approx 4$ so that $\tilde{p} \approx \tilde{p}_k$ (for $k = 1.96$) and $\text{M.E.}(\tilde{p}) \approx \text{M.E.}(\tilde{p}_k)$ (for $k = 1.96$). If you have a calculator or computer program which computes the Wald interval, then you can use this "add 2 successes / add 4 observations" trick to approximate the Wilson 95% confidence interval.

5.3 Testing for a proportion

The hypothesis testing procedures discussed in this section are based on the normal approximation to the sampling distribution of $\hat{p}$. Hence we will continue to assume that the data form a simple random sample of size n, selected with replacement, from a dichotomous population with population success proportion p, or equivalently, that the data correspond to the outcomes of a sequence of n Bernoulli trials with success probability p. As before if the population is very large, then these methods can also be used when the data form a simple random sample of size n, selected without replacement.

A **hypothesis** (statistical hypothesis) is a conjecture about the nature of the population. When the population is dichotomous, a hypothesis is a conjecture about the value of the population success proportion p.

A **hypothesis test** (test of significance) is a formal procedure for deciding between two complementary hypotheses. These hypotheses are known as the null hypothesis (H_0 for short) and the research (or alternative) hypothesis (H_1 for short). The research hypothesis is the hypothesis of primary interest, since the testing procedure is designed to address the question: "Do the data support the research hypothesis?" The null hypothesis is defined as the negation of the research hypothesis. The test begins by tentatively assuming that the null hypothesis is true (the research hypothesis is false). The data are then examined to determine whether the null hypothesis can be rejected in favor of the research hypothesis. The probability of observing data as unusual (surprising) or more unusual as that actually observed under the tentative assumption that the null hypothesis is true is computed. This probability is known as the P–value of the test. (The P in P–value indicates that it is a probability it does not refer to the population success proportion p.) A small P–value

indicates that the observed data would be unusual (surprising) if the null hypothesis was actually true. Thus if the P–value is small enough, then the null hypothesis is judged untenable and the test rejects the null hypothesis in favor of the research (alternative) hypothesis. On the other hand, a large (not small) P–value indicates that the observed data would not be unusual (not surprising) if the null hypothesis was actually true. Thus if the P–value is large (not small enough), then the null hypothesis is judged tenable and the test fails to reject the null hypothesis.

There is a strong similarity between the reasoning used for a hypothesis test and the reasoning used in the trial of a defendant in a court of law. In a trial the defendant is presumed innocent (tentatively assumed to be innocent) and this tentative assumption is not rejected unless sufficient evidence is provided to make this tentative assumption untenable. In this situation the research hypothesis states that the defendant is guilty and the null hypothesis states that the defendant is not guilty (is innocent). The P–value of a hypothesis test is analogous to a quantification of the weight of the evidence that the defendant is guilty with small values indicating that the evidence is unlikely under the assumption that the defendant is innocent.

Example. Insects in an apple orchard (revisited). Recall that the manager of a large apple orchard examined a simple random sample of 200 apple trees to gauge the extent of insect infestation in the orchard. The manager has determined that applying the insecticide is not economically justifiable unless more than 20% of the apple trees in the orchard are infested. Since the manager does not want to apply the insecticide unless there is evidence that it is needed, the question of interest here is: "Is there sufficient evidence to justify application of the insecticide?" In terms of the population success proportion p (the proportion of all of the apple trees in this orchard that are infested) the research hypothesis is $H_1 : p > .20$ (more than 20% of all the trees in the orchard are infested); and the null hypothesis is $H_0 : p \leq .20$ (no more than 20% of all the trees in the orchard are infested). A test of the null hypothesis $H_0 : p \leq .20$ versus the research hypothesis $H_1 : p > .20$ begins by tentatively assuming that no more than 20% of all the trees in the orchard are infested. Under this tentative assumption it would be surprising to observe a proportion of infested trees in the sample $\hat{p}$ that was much larger than .20. Thus the test should reject $H_0 : p \leq .20$ in favor of $H_1 : p > .20$ if the observed value of $\hat{p}$ is sufficiently large relative to .20.

Case 1. Suppose that 52 of the 200 apple trees in the sample are infested so that $\hat{p} = .26$. In this case we know that 26% of the 200 trees in the sample are infested and we need to decide whether this suggests that the proportion of all the trees in the orchard that are infested p exceeds .20. More specifically, we need to determine whether observing 52 or more infested trees in a simple random sample of 200 trees would be surprising if in fact no more than 20% of all the trees in the orchard were infested. Assuming that

exactly 20% of all the trees in the orchard are infested, we find that the probability of observing 52 or more infested trees in a sample of 200 trees (seeing $\hat{p} \geq .26$), is .0169 (this is the P–value of the test). In other words, if no more than 20% of all the trees in the orchard were infested, then a simple random sample of 200 trees would give $\hat{p} \geq .26$ about 1.69% of the time. Therefore, observing 52 infested trees in a sample of 200 would be very surprising if no more than 20% of all the trees in the orchard were infested and we have sufficient evidence to reject the null hypothesis $H_0 : p \leq .20$ in favor of the research hypothesis $H_1 : p > .20$. In the case we would conclude that there is sufficient evidence to contend that more than 20% of all the trees in the orchard are infested and, in this sense, application of the insecticide is justifiable.

Case 2. Next suppose that 45 of the 200 apple trees in the sample are infested so that $\hat{p} = .225$. Assuming that exactly 20% of all the trees in the orchard are infested, we find that the probability of observing 45 or more infested trees in a sample of 200 trees (seeing $\hat{p} \geq .225$), is .1884 (this is the P–value of the test). In other words, if no more than 20% of all the trees in the orchard were infested, then a simple random sample of 200 trees would give $\hat{p} \geq .225$ about 18.84% of the time. Therefore, observing 45 infested trees in a sample of 200 would not be very surprising if no more than 20% of all the trees in the orchard were infested and we do not have sufficient evidence to reject the null hypothesis $H_0 : p \leq .20$ in favor of the research hypothesis $H_1 : p > .20$. In the case we would conclude that there is not sufficient evidence to contend that more than 20% of all the trees in the orchard are infested and, in this sense, application of the insecticide not is justifiable.

The research hypothesis in the apple orchard example is a directional hypothesis of the form $H_1 : p > p_0$, where $p_0 = .20$. We will now discuss the details of a hypothesis test for a directional research hypothesis of this form. For the test procedure to be valid the specified value p_0 and the direction of the research hypothesis must be motivated from subject matter knowledge before looking at the data that are to be used to perform the test.

Let p_0 denote the hypothesized value (with $0 < p_0 < 1$) which we wish to compare with p. The research hypothesis states that p is greater than p_0 ($H_1 : p > p_0$). The null hypothesis is the negation of $H_1 : p > p_0$ which states that p is no greater than p_0 ($H_0 : p \leq p_0$). The research hypothesis $H_1 : p > p_0$ specifies that the population is one of the dichotomous populations for which the population success proportion p is greater than p_0. The null hypothesis $H_0 : p \leq p_0$ specifies that the population is one of the dichotomous populations for which the population success proportion p is no greater than p_0. Notice that this competing pair of hypotheses provides a decomposition of all possible dichotomous populations into the collection of dichotomous populations where $p > p_0$ and the research hypothesis is true and the collection of dichotomous populations where $p \leq p_0$ and the null hypothesis is true. Our goal is to use the data to decide which of these two

collections of dichotomous populations contains the actual population we are sampling from.

Since a hypothesis test begins by tentatively assuming that the null hypothesis is true, we need to decide what constitutes evidence against the null hypothesis $H_0 : p \leq p_0$ and in favor of the research hypothesis $H_1 : p > p_0$. The relationship between the observed proportion of successes in the sample $\hat{p}$ and the hypothesized value p_0 will be used to assess the strength of the evidence in favor of the research hypothesis. Generally, we would expect to observe larger values of $\hat{p}$ more often when the research hypothesis $H_1 : p > p_0$ is true than when the null hypothesis $H_0 : p \leq p_0$ is true. In particular, we can view the observation of a value of $\hat{p}$ that is sufficiently large relative to p_0 as constituting evidence against the null hypothesis $H_0 : p \leq p_0$ and in favor of the research hypothesis $H_1 : p > p_0$. Deciding whether the observed value of $\hat{p}$ is "sufficiently large relative to p_0" is based on the corresponding P–value, which is defined below.

The P–value for testing the null hypothesis $H_0 : p \leq p_0$ versus the research hypothesis $H_1 : p > p_0$ is the probability of observing a value of $\hat{p}$ as large or larger than the value of $\hat{p}$ that we actually do observe. The P–value quantifies the consistency of the observed data with the null hypothesis and may be interpreted as a, somewhat indirect, measure of the strength of the evidence in the data in favor of the research hypothesis and against the null hypothesis. Because the P–value is computed under the assumption that the null hypothesis is true (and the research hypothesis is false), the smaller the P–value is, the less consistent the observed data are with the null hypothesis. Therefore, since one of the hypotheses must be true, when we observe a small P–value we can conclude that the research hypothesis is more consistent with the observed data than is the null hypothesis.

The P–value is computed under the assumption that the research hypothesis $H_1 : p > p_0$ is false and the null hypothesis $H_0 : p \leq p_0$ is true. Because the null hypothesis only specifies that $p \leq p_0$, we need to choose a particular value of p (that is no larger than p_0) in order to compute the P–value. It is most appropriate to use $p = p_0$ for this computation. (Recall that in the apple orchard example we used $p_0 = .20$ to compute the P–value.) Using $p = p_0$, which defines the boundary between $p \leq p_0$, where the null hypothesis is true, and $p > p_0$, where the research hypothesis is true, provides some protection against incorrectly rejecting $H_0 : p \leq p_0$.

To compute the P–value we need to know how much variability there is in the sampling distribution of $\hat{p}$ when $p = p_0$. When $p = p_0$ the standard error of $\hat{p}$, which provides a suitable measure of the variability in $\hat{p}$, is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

To use the normal approximation to the sampling distribution of $\hat{p}$ to compute the P–value we first need to determine the calculated Z statistic or Z score corresponding to the observed value of $\hat{p}$. This calculated Z statistic, denoted by Z_{calc}, is

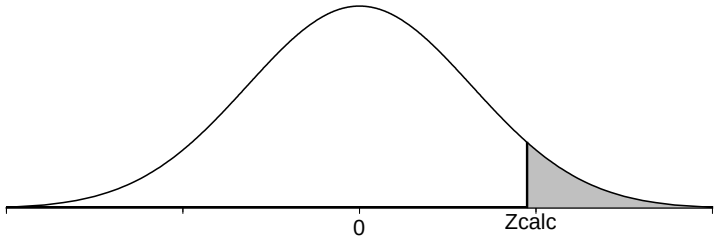
$$Z_{calc} = \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})},$$

where the standard error S.E.$(\hat{p})$ is as defined above. Recall that the P–value for testing the null hypothesis $H_0 : p \leq p_0$ versus the research hypothesis $H_1 : p > p_0$ is the probability of observing a value of $\hat{p}$ as large or larger than the value of $\hat{p}$ that we actually do observe, computed assuming that $p = p_0$. Using the normal approximation, this P–value is equal to the probability that a standard normal variable takes on a value at least as large as Z_{calc}. This P–value is

$$P\text{–value} = P(Z \geq Z_{calc}),$$

where Z denotes a standard normal variable, *i.e.*, this P–value is the area under the standard normal density curve to the right of Z_{calc}, as shown in Figure 4. Notice that the P value (the area to the right of Z_{calc}) is small when Z_{calc} is far to the right of zero which is equivalent to $\hat{p}$ being far to the right of p_0.

Figure 4. P–value for $H_0 : p \leq p_0$ versus $H_1 : p > p_0$.



Once the P–value has been computed we need to decide whether the P–value is small enough to justify rejecting the null hypothesis in favor of the research hypothesis. In the apple orchard example we argued that observing 52 infested trees in a sample of 200 would be very surprising if no more than 20% of all the trees in the orchard were infested, since the corresponding P–value of .0169 was very small. We also argued that observing 45 infested trees in a sample of 200 would not be very surprising if no more than 20% of all the trees in the orchard were infested, since the corresponding P–value of .1884 is fairly large. Deciding whether a P–value is small enough to reject a null hypothesis requires a subjective judgment by the investigator in the context of the problem at hand.

The following general remarks regarding the use of P–values to assess the evidence against a null hypothesis and in favor of a research hypothesis apply to hypothesis tests in general, not just hypothesis tests for a proportion.

One approach to hypothesis testing is to use a fixed cutoff value to decide whether the P–value is "large" or "small". The most common application of this approach is to conclude that there is sufficient evidence to reject the null hypothesis in favor of the research hypothesis only when the P–value is less than .05. When a fixed cutoff value like .05 (5%) is used to decide whether to reject the null hypothesis in favor of the research hypothesis this cutoff value is known as the **significance level** of the test. Hence, if we adopt the rule of rejecting the null hypothesis in favor of the research hypothesis only when the P–value is less than .05, then we are performing a hypothesis test at the 5% level of significance. In accordance with this terminology, the P–value is also known as the **observed significance level** of the test and if the P–value is less than the prescribed significance level, then the results are said to be **statistically significant**.

To perform a hypothesis test at the 5% level of significance we compute the appropriate P–value and compare it to the fixed significance level .05. If the P–value is less than .05, then we conclude that there is sufficient evidence, at the 5% level of significance, to reject the null hypothesis H_0 in favor of the research hypothesis H_1, *i.e.*, if the P–value **is less than** .05, then the data **do** support H_1. If the P–value is not less than .05, then we conclude that there is not sufficient evidence, at the 5% level of significance, to reject the null hypothesis H_0 in favor of the research hypothesis H_1, *i.e.*, if the P–value **is not less than** .05, then the data **do not** support H_1.

Instead of, or in addition to, using a fixed significance level like 5% we can use the P–value as a measure of the evidence (in the data) against the null hypothesis H_0 and in favor of the research hypothesis H_1. Some guidelines for deciding how strong the evidence is in favor of the research hypothesis H_1 are given below.

Guidelines for interpreting a P–value:

1. If the P–value is greater than .10, there is no evidence in favor of H_1.
2. If the P–value is between .05 and .10, there is suggestive but very weak evidence in favor of H_1.
3. If the P–value is between .04 and .05, there is weak evidence in favor of H_1.
4. If the P–value is between .02 and .04, there is moderately strong evidence in favor of H_1.
5. If the P–value is between .01 and .02, there is strong evidence in favor of H_1.
6. If the P–value is less than .01, there is very strong evidence in favor of H_1.

Whether you choose to use a fixed significance level or the preceding guidelines based on the P–value you should always report the P–value since this allows someone else to interpret the evidence in favor of H_1 using their personal preferences regarding the size of a P–value.

In the U.S. legal system there is a similar set of guidelines for assessing the level of proof or weight of the evidence against the null hypothesis of innocence and in favor of the research hypothesis of guilt. The weakest level of proof is "the preponderance of the evidence" (this is similar to a reasonably small P–value), the next level of proof is "clear and convincing evidence" (this is similar to a small P–value), and the highest level of proof is "beyond a reasonable doubt" (this is similar to a very small P–value).

We now return to our discussion for the particular research hypothesis $H_1 : p > p_0$. The steps for performing a hypothesis test for

$$H_0 : p \leq p_0 \quad \text{versus} \quad H_1 : p > p_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(Z \geq Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = (\hat{p} - p_0)/\text{S.E.}(\hat{p})$, and $\text{S.E.}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$. This P–value is the area under the standard normal density curve to the right of Z_{calc}, as shown in Figure 4.

2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p > p_0$ over $H_0 : p \leq p_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportion p is greater than p_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p > p_0$ over $H_0 : p \leq p_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportion p is greater than p_0.

The procedure for testing the null hypothesis $H_0 : p \leq p_0$ versus the research hypothesis $H_1 : p > p_0$ given above is readily modified for testing the null hypothesis $H_0 : p \geq p_0$ versus the research hypothesis $H_1 : p < p_0$. The essential modification is to change the direction of the inequality in the definition of the P–value. Consider a situation where the research hypothesis specifies that the population success proportion p is less than the particular, hypothesized value p_0, *i.e.*, consider a situation where the research hypothesis is $H_1 : p < p_0$ and the null hypothesis is $H_0 : p \geq p_0$. For these hypotheses values of the observed success proportion $\hat{p}$ that are sufficiently small relative to p_0 provide evidence in favor of the research hypothesis $H_1 : p < p_0$ and against the null hypothesis $H_0 : p \geq p_0$. Therefore, the P–value for testing $H_0 : p \geq p_0$ versus $H_1 : p < p_0$ is the probability of observing a value of $\hat{p}$ as small or smaller than the value actually observed. As before, the P–value is computed under the assumption that $p = p_0$. The calculated Z statistic Z_{calc} is defined as before; however, in this situation the P–value is the area under the standard

normal density curve to the left of Z_{calc}, since values of $\hat{p}$ that are small relative to p_0 constitute evidence in favor of the research hypothesis.

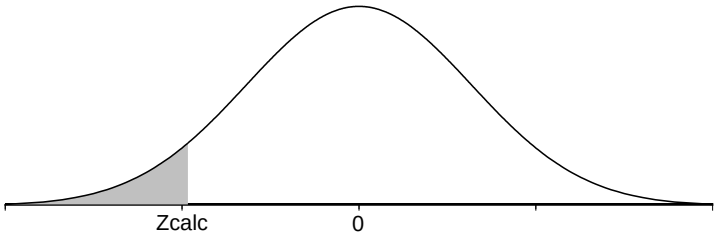
The steps for performing a hypothesis test for

$$H_0 : p \geq p_0 \quad \text{versus} \quad H_1 : p < p_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(Z \leq Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = (\hat{p} - p_0)/\text{S.E.}(\hat{p})$, and $\text{S.E.}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$. This P–value is the area under the standard normal density curve to the left of Z_{calc} as shown in Figure 5.

Figure 5. P–value for $H_0 : p \geq p_0$ versus $H_1 : p < p_0$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p < p_0$ over $H_0 : p \geq p_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportion p is less than p_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p < p_0$ over $H_0 : p \geq p_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportion p is less than p_0.

Example. Acceptance sampling for electronic devices. A large retailer receives a shipment of 10,000 electronic devices from a supplier. The supplier guarantees that no more than 6% of these devices are defective. In fact, if more than 6% of the devices in the shipment are defective, then the supplier will allow the retailer to return the entire shipment, provided this is done with 10 days of receiving the shipment. Therefore, the retailer needs to decide between accepting the shipment and returning the shipment to the supplier. This decision will be based on the information provided by examining a simple random sample of electronic devices selected from the shipment.

In this example one of these electronic devices is a unit and the collection of 10,000 units constituting the shipment is the population. Notice that, in this example, the target population and the sampled population are the same (each is the shipment of 10,000

devices). A suitable variable for the indicated objective is whether an electronic device is defective with the two possible values: yes (it is defective) and no (it is not defective). A relevant parameter is the proportion p of defective devices in the shipment of 10,000 devices. The corresponding statistic $\hat{p}$ is the proportion of defective devices in the sample of devices that is examined.

The boundary between the null and research hypotheses is clearly $p_0 = .06$, since we need to decide whether the population proportion of defective devices p exceeds .06. Assuming that the supplier is trustworthy, it would seem to be a reasonable business practice to accept the shipment of electronic devices unless we find sufficient evidence, by examining the sample of devices, to conclude that more than 6% of the devices in the shipment are defective. Hence, we will use a hypothesis test to determine whether there is sufficient evidence to conclude that the population defective proportion p exceeds .06. More formally, our research hypothesis is $H_1 : p > .06$ and our null hypothesis is $H_0 : p \leq .06$.

To continue with this example we need to know the sample size n and the results of the examination of the sample of electronic devices. Suppose that the simple random sample contains $n = 200$ electronic devices. For a sample of size $n = 200$ the standard error of $\hat{p}$ for testing a hypothesis with $p_0 = .06$ is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{(.06)(.94)}{200}} = .0168.$$

Case 1. Suppose that 16 of the 200 devices in the sample are defective so that $\hat{p} = .08$. In this case we know that 8% of the 200 devices in the sample are defective and we need to decide whether this suggests that more than 6% of all the devices in the shipment are defective. The calculated Z statistic is

$$Z_{calc} = \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})} = \frac{.08 - .06}{.0168} = 1.1910$$

and the P–value is

$$P\text{–value} = P(Z \geq Z_{calc}) = P(Z \geq 1.1910) = .1168.$$

Since this P–value is large there is not sufficient evidence to reject the null hypothesis $p \leq .06$ in favor of the research hypothesis $p > .06$. Therefore, if we observe 16 defective devices in a random sample of $n = 200$ devices, then we should accept the shipment of devices, since there is not sufficient evidence to conclude that more than 6% of the shipment of 10,000 devices is defective.

Case 2. Now suppose that 20 of the 200 devices in the sample are defective so that $\hat{p} = .10$. In this case

$$Z_{calc} = \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})} = \frac{.10 - .06}{.0168} = 2.3820$$

and the P–value is

$$P\text{–value} = P(Z \geq Z_{calc}) = P(Z \geq 2.3820) = .0086.$$

This P–value is very small indicating that we have strong evidence against the null hypothesis $p \leq .06$ and in favor of the research hypothesis $p > .06$. Therefore, if we observe 20 defective devices in a random sample of $n = 200$ devices, then we are justified in returning the shipment of devices, since there is strong evidence that more than 6% of the shipment of 10,000 devices is defective.

In both of the cases described above, in addition to the conclusion of the hypothesis test the retailer might also wonder exactly what proportion of devices in the shipment of 10,000 devices are defective. We can use a 95% confidence interval estimate of p to answer this question.

In the first case there are 16 defective devices in the sample of $n = 200$ giving an observed proportion of defective devices of $\hat{p} = .08$. The confidence interval estimate is based on $\tilde{p}_k = .0879$ and the 95% margin of error M.E.$(\tilde{p}_k) = .0381$. Therefore, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is between $.0879 - .0381 = .0498$ and $.0879 + .0381 = .1260$. As expected, since we did not reject the tentative assumption that $p \leq .06$, we see that this confidence interval includes proportions that are both less than .06 and greater than .06.

In the second case there are 20 defective devices in the sample of $n = 200$ giving $\hat{p} = .10$, $\tilde{p}_k = .1075$, and the 95% margin of error M.E.$(\tilde{p}_k) = .0419$. Therefore, in this case we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is between $.1075 - .0419 = .0656$ and $.1075 + .0419 = .1494$. As expected, since we did reject the tentative assumption that $p \leq .06$, we see that all of the values in this confidence interval are greater than .06. Notice that in this case the P–value .0086 is quite small indicating that there is very strong evidence that the proportion of defective devices in the shipment is larger than .06. However, from the 95% confidence interval estimate of p we find that this proportion of defective devices might actually be as small as .0656, which is not much larger than .06. Thus, the small P–value indicates strong evidence that p is greater than .06 but it does not necessarily indicate that p is a lot larger than .06. Of course the 95% confidence interval estimate also indicates that p may be as large as .1494 which is a good bit larger than .06.

The scenario in the acceptance sampling example where there is strong evidence that $p > .06$ (P–value .0086) but the lower limit of the 95% confidence interval .0656 is not

much larger than .06 highlights the need for a confidence interval to estimate the value of p in addition to a hypothesis test to clarify the practical importance of the result of the test. Bear in mind that a hypothesis test addresses a very formal distinction between two complementary hypotheses and that in some situations the results may be statistically significant (in the sense that the P–value is small) but of little practical significance (in the sense that p is not very different from p_0).

Example. Machine parts. The current production process used to manufacture a particular machine part is known (from past experience) to produce parts which are unacceptable, in the sense that they require further machining, 35% of the time. A new production process has been developed with the hope that it will reduce the chance of producing unacceptable parts. Suppose that 200 parts are produced using the new production process and that 54 of these parts are found to be unacceptable.

In this example we have a sequence of 200 dichotomous trials, where a trial consists of producing a part with the new production process and determining whether it is unacceptable. In this example p denotes the probability that a part produced using the new production process will be unacceptable. We will model these 200 trials as a sequence of $n = 200$ Bernoulli trials with population success probability p. This assumption is reasonable provided: (1) the probability that a part is unacceptable is essentially constant from part to part; and, (2) whether a specific part is unacceptable or not has no effect on the probability that any other part is unacceptable.

In this example the boundary between the null and research hypotheses is clearly $p_0 = .35$. Since these data were collected to determine if the new production process is better than the old process, we want to know whether there is sufficient evidence to conclude that less than 35% of the parts produced using the new production process would be unacceptable. Thus our research hypothesis is $H_1 : p < .35$ and our null hypothesis is $H_0 : p \geq .35$. Since 54 of the 200 parts in our sample are unacceptable we know that $\hat{p} = .27$ and we need to determine whether this is small enough to suggest that the corresponding population probability p is also less than .35. For a sample of size $n = 200$ the standard error of $\hat{p}$ for testing a hypothesis with $p_0 = .35$ is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{(.35)(.65)}{200}} = .0337.$$

The calculated Z statistic is

$$Z_{calc} = \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})} = \frac{.27 - .35}{.0337} = -2.3739$$

and the P–value is

$$P\text{–value} = P(Z \leq Z_{calc}) = P(Z \leq -2.3739) = .0088.$$

Since this P–value is very small, there is sufficient evidence to reject the null hypothesis $p \geq .35$ in favor of the research hypothesis $p < .35$. Hence, based on this sample of 200 parts there is very strong evidence that the new production process is superior in the sense that the probability of producing an unacceptable part is less than .35.

Clearly this conclusion should be accompanied by an estimate of how much smaller this probability is likely to be. Observing 54 unacceptable parts in the sample of $n = 200$ gives $\hat{p} = .27$, $\tilde{p}_k = .2743$, and the 95% margin of error M.E.$(\tilde{p}_k) = .0611$. Therefore, we are 95% confident that the probability of a part produced using the new production process being unacceptable is between $.2743 - .0611 = .2132$ and $.2743 + .0611 = .3354$. As expected, since we did reject the tentative assumption that $p \geq .35$, we see that all of the values in this confidence interval are less than .35. The P–value .0088 is quite small indicating that there is very strong evidence that the probability of producing an unacceptable part is less than .35. However, from the 95% confidence interval estimate of p we find that this probability might actually be as large as .3354 which is not much smaller than .35. Of course the 95% confidence interval estimate also indicates that p may be as small as .2132 which is a good bit smaller than .35.

The hypothesis tests we have discussed thus far are only appropriate when we have enough *a priori* information, *i.e.*, information that does not depend on the data to be used for the hypothesis test, to postulate that the population success proportion p is on one side of a particular value p_0. That is, we have only considered situations where the research hypothesis is directional in the sense of specifying either that $p > p_0$ or that $p < p_0$. In some situations we will not have enough *a priori* information to allow us to choose the appropriate directional research hypothesis. Instead, we might only conjecture that the population success proportion p is different from some particular value p_0. In a situation like this our research hypothesis specifies that the population success proportion p is different from p_0, *i.e.*, $H_1 : p \neq p_0$ and the corresponding null hypothesis specifies that p is exactly equal to p_0, *i.e.*, $H_0 : p = p_0$. As we will see in the inheritance model considered below, when testing to see whether p is equal to a specified value p_0 the null hypothesis $H_0 : p = p_0$ often corresponds to the validity of a particular theory or model and the research hypothesis or alternative hypothesis specifies that the theory is invalid.

In order to decide between the null hypothesis $H_0 : p = p_0$ and the research hypothesis $H_1 : p \neq p_0$, we need to decide whether the observed success proportion $\hat{p}$ supports the null hypothesis by being "close to p_0", or supports the research hypothesis by being "far away from p_0". In this situation the P–value is the probability that the observed success proportion $\hat{p}$ would be as far or farther away from p_0 in either direction as is the value that we actually observe. In other words, the P–value corresponds to large values of the distance $|\hat{p} - p_0|$ (the absolute value of the difference between $\hat{p}$ and p_0). The P–value is computed under the assumption that $p = p_0$ so that the null hypothesis is true. In this

situation the calculated Z statistic Z_{calc} is the absolute value of the Z statistic that would be used for testing a directional hypothesis. That is, the calculated Z statistic is

$$Z_{calc} = \left| \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})} \right|.$$

In terms of this Z statistic the P–value is the probability that the absolute value of a standard normal variable Z would take on a value as large or larger than Z_{calc} assuming that $p = p_0$. This probability is the sum of the area under the standard normal density curve to the left of $-Z_{calc}$ and the area under the standard normal density curve to the right of Z_{calc}. We need to add these two areas (probabilities) since we are finding the probability that the observed success proportion $\hat{p}$ would be as far or farther away from p_0 in either direction as is the value that we actually observe, when $p = p_0$.

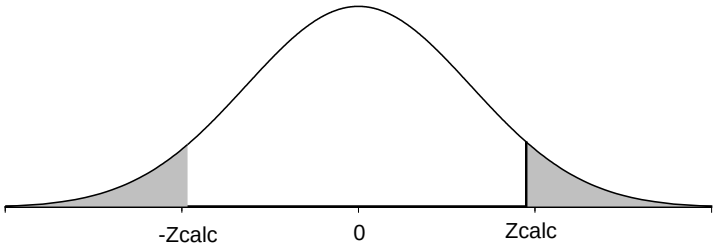
The steps for performing a hypothesis test for

$$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(|Z| \geq Z_{calc}) = P(Z \leq -Z_{calc}) + P(Z \geq Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = |(\hat{p} - p_0)/\text{S.E.}(\hat{p})|$, and $\text{S.E.}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$. This P–value is the sum of the area under the standard normal density curve to the left of $-Z_{calc}$ and the area under the standard normal density curve to the right of Z_{calc} as shown in Figure 6.

Figure 6. P–value for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p \neq p_0$ over $H_0 : p = p_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportion p is different from p_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p \neq p_0$ over $H_0 : p = p_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportion p is different from p_0.

Example. Inheritance in peas (flower color). In his investigations, during the years 1856 to 1868, of the chromosomal theory of inheritance Gregor Mendel performed a series of experiments on ordinary garden peas. One characteristic of garden peas that Mendel studied was the color of the flowers (red or white). When Mendel crossed a plant with red flowers with a plant with white flowers, the resulting offspring all had red flowers. But when he crossed two of these first generation plants, he observed plants with white as well as red flowers. We will use the results of one of Mendel's experiments to test a simple model for inheritance of flower color. Mendel observed 929 pea plants arising from a cross of two of these first generation plants. Of these 929 plants he found 705 plants with red flowers and 224 plants with white flowers.

The gene which determines the color of the flower occurs in two forms (alleles). Let R denote the allele for red flowers (which is dominant) and r denote the allele for white flowers (which is recessive). When two plants are crossed the offspring receives one allele from each parent, thus there are four possible genotypes (ordered combinations) RR, Rr, rR, and rr. The three genotypes RR, Rr, and rR, which include the dominant R allele, will yield red flowers while the fourth genotype rr will yield white flowers. If a red flowered RR genotype parent is crossed with a white flowered rr genotype parent, then all of the offspring will have genotype Rr and will produce red flowers. If two of these first generation Rr genotype plants are crossed, each of the four possible genotypes RR, Rr, rR, and rr is equally likely and plants with white as well as red flowers will occur. Under this simple model for inheritance, with each of the four genotypes having the same probability of occurring (and with each plant possessing only one genotype), the probability that a plant will have red flowers is $p = 3/4$ and the probability that a plant will have white flowers is $1 - p = 1/4$. In other words, this model for inheritance of flower color says that we would expect to see red flowers $3/4$ of the time and white flowers $1/4$ of the time.

We can test the validity of this model by testing the null hypothesis $H_0 : p = 3/4$ versus the alternative hypothesis $H_1 : p \neq 3/4$. Notice that the model is valid under the null hypothesis and the model is not valid under the alternative hypothesis. Mendel observed 705 plants with red flowers out of the $n = 929$ plants giving an observed proportion of plants with red flowers of $\hat{p} = 705/929 = .7589$. The standard error of $\hat{p}$, computed under the assumption that $p = p_0 = 3/4$, is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{(.75)(.25)}{929}} = .0142$$

and the calculated Z statistic is $Z_{calc} = .6251$ giving a P–value of

$$P\text{–value} = P(|Z| \geq Z_{calc}) = P(|Z| \geq .6251) = .5319.$$

This P–value is quite large and we are not able to reject the null hypothesis; therefore, we conclude that the observed data are consistent with Mendel's model. Technically, we should say that the data are not inconsistent with the model in the sense that we cannot reject the hypothesis that $p = 3/4$. In this example, the 95% confidence interval estimate of p ranges from .7303 to .7853.

5.4 Directional confidence bounds

In our discussion of hypothesis testing we considered directional research hypotheses of the form $p > p_0$ and $p < p_0$ as well as nondirectional research hypotheses of the form $p \neq p_0$. However, in our discussion of 95% confidence intervals for p we only considered confidence intervals of the form

$$\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \leq p \leq \tilde{p}_k + \text{M.E.}(\tilde{p}_k).$$

A 95% confidence interval of this form consists of a lower bound $\tilde{p}_k - \text{M.E.}(\tilde{p}_k)$ for p and an upper bound $\tilde{p}_k + \text{M.E.}(\tilde{p}_k)$ for p, thereby giving a range of plausible values for p. In a situation where we have enough *a priori* information to justify a directional research hypothesis we might argue that it would be more appropriate to determine a 95% confidence bound (a lower bound or an upper bound) for p instead of a range of values.

For example, in the acceptance sampling example we might argue that we are less concerned with how large p might be than with how small it might be. Therefore, we might be satisfied with an estimate of the smallest value of p which would be consistent with the data, *i.e.*, we might only need a 95% confidence lower bound for p.

We will now show how a 90% confidence interval for p can be used to provide a 95% confidence lower (or upper) bound for p. The cutoff point k for the margin of error for a 90% confidence interval for p is $k = 1.645$. Three relevant probabilities associated with the 90% confidence interval with lower limit L and upper limit U are:

$$P[L \leq p \leq U] = .90, \quad P[p < L] = .05, \quad \text{and} \quad P[U < p] = .05.$$

Combining the probability that p is between L and U and the probability that p is greater than U we see that

$$P[p > L] = .90 + .05 = .95.$$

In other words, 95% of the time the computed value of the lower limit L of a 90% confidence interval for p will be less than p. Therefore, the lower limit L of a 90% confidence interval for p can be used as a 95% confidence lower bound for p. An analogous argument shows that the upper limit U of a 90% confidence interval for p can be used as a 95% confidence upper bound for p.

Example. Acceptance sampling for electronic devices (revisited). If there are 16 defective devices in a sample of $n = 200$, then the observed proportion of defective devices is $\hat{p} = .08$ and taking $k = 1.645$ gives $\tilde{p}_k = .0856$ and a 90% margin of error of M.E.$(\tilde{p}_k) = .0318$. Therefore, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is at least $.0856 - .0318 = .0538$, which allows for the possibility that $p < .06$.

If there are 20 defective devices in a sample of $n = 200$, then the observed proportion of defective devices is $\hat{p} = .10$ and taking $k = 1.645$ gives $\tilde{p}_k = .1053$ and a 90% margin of error of M.E.$(\tilde{p}_k) = .0351$. Therefore, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is at least $.1053 - .0351 = .0702$, which supports the conclusion that $p > .06$.

5.5 Summary

Basic inferential methods (confidence intervals and hypothesis tests) were introduced in this chapter in the context of making inferences about a population proportion p. These inferential methods are based on the sampling distribution of a statistic (the sample proportion $\hat{p}$ in this chapter) which describes the sample to sample variability in the statistic as an estimator of the corresponding parameter.

The specific inferential methods introduced in this chapter involve the use of the observed proportion of successes $\hat{p}$ in a random sample to make inferences about the corresponding population success proportion p. In particular, we discussed confidence interval estimates of p and formal tests of hypotheses about p. These inferences about p are based on a normal approximation to the sampling distribution of $\hat{p}$ and require certain assumptions about the random sample. Strictly speaking, the inferential methods discussed in this chapter are not appropriate unless these assumptions are valid. The requisite assumptions are that the sample is a simple random sample selected with replacement or equivalently that the sample corresponds to a sequence of Bernoulli trials. We also noted that this approximation works well for a simple random sample selected without replacement provided the population being sampled is very large. The sampling distribution of $\hat{p}$ is the theoretical probability distribution of $\hat{p}$ which indicates how $\hat{p}$ behaves as an estimator of p. Under the assumptions described above, the sampling distribution of $\hat{p}$ indicates that $\hat{p}$ is unbiased as an estimator of p ($\hat{p}$ neither consistently overestimates p nor consistently underestimates p) and provides a measure of the variability in $\hat{p}$ as an estimator of p (the population standard error of $\hat{p}$, S.E.$(\hat{p}) = \sqrt{p(1-p)/n}$). The normal approximation allows us to compute probabilities concerning $\hat{p}$ by re–expressing these probabilities in terms of the standardized variable $Z = (\hat{p} - p)/\text{S.E.}(\hat{p})$ and using the standard normal distribution to compute the probabilities.

A 95% confidence interval estimate of p is an interval of plausible values for p constructed using a method which guarantees that 95% of such intervals will actually contain the unknown proportion p. That is, a 95% confidence interval is an interval constructed using a method of generating such intervals with the property that this method will work, in the sense of generating an interval that contains p, for 95% of all possible samples. We recommended the Wilson interval as a confidence interval estimate of p. For a confidence level C (usually .95) and the corresponding standard normal cutoff point k ($k = 1.96$ when $C = .95$) the Wilson interval is of the form

$$\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \le p \le \tilde{p}_k + \text{M.E.}(\tilde{p}_k),$$

where

$$\tilde{p}_k = \frac{n\hat{p} + \frac{k^2}{2}}{n + k^2}$$

determines the center of the interval, and the **margin of error of** $\tilde{p}_k$

$$\text{M.E.}(\tilde{p}_k) = \sqrt{\tilde{p}_k^2 - \frac{n\hat{p}^2}{n + k^2}}$$

determines the length of the interval. We also discussed a simple but less accurate confidence interval for p (the Wald interval) and a simple approximation to the 95% Wilson interval (the Agresti–Coull interval). The 95% Agresti–Coull interval is of the form

$$\tilde{p} - \text{M.E.}(\tilde{p}) \le p \le \tilde{p} + \text{M.E.}(\tilde{p}),$$

where

$$\tilde{p} = \frac{\text{the number of successes plus 2}}{\text{the number of observations plus 4}} = \frac{n\hat{p} + 2}{n + 4}$$

and

$$\text{M.E.}(\tilde{p}) = 1.96\sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}.$$

The "add 2 successes / add 4 observations" trick used in the Agresti–Coull interval can be used in a computer program or calculator implementation of the Wald interval to approximate the 95% Wilson interval.

A hypothesis test is used to compare two competing, complementary hypotheses (the null hypothesis H_0 and the research or alternative hypothesis H_1) about p by tentatively assuming that H_0 is true and examining the evidence, which is quantified by the appropriate P–value, against H_0 and in favor of H_1. Since the P–value quantifies evidence against H_0 and in favor of H_1, a small P–value constitutes evidence in favor of H_1. Guidelines for interpreting a P–value are given on page 99.

If there is sufficient *a priori* information to specify a directional hypothesis of the form $H_1 : p > p_0$ or $H_1 : p < p_0$, then we can perform a hypothesis test to address the respective questions "Is there sufficient evidence to conclude that $p > p_0$?" or "Is there sufficient evidence to conclude that $p < p_0$?" The null hypotheses for these research hypotheses are their negations $H_0 : p \leq p_0$ and $H_0 : p \geq p_0$, respectively. The hypothesis test proceeds by tentatively assuming that the null hypothesis H_0 is true and checking to see if there is sufficient evidence (a small enough P–value) to reject this tentative assumption in favor of the research hypothesis H_1. The P–values for these directional hypothesis tests are based on the observed value of the Z–statistic $Z_{calc} = (\hat{p} - p_0)/\text{S.E.}(\hat{p})$, where $\text{S.E.}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$ is the standard error for testing. For $H_1 : p > p_0$ large values of $\hat{p}$, relative to p_0, favor H_1 over H_0 and the P–value is the probability that $Z \geq Z_{calc}$. For $H_1 : p < p_0$ we look for small values of $\hat{p}$, relative to p_0, and the P–value is the probability that $Z \leq Z_{calc}$.

For situations where there is not enough *a priori* information to specify a directional hypothesis we considered a hypothesis test for the null hypothesis $H_0 : p = p_0$ versus the alternative hypothesis $H_1 : p \neq p_0$. Again we tentatively assume that H_0 is true and check to see if there is sufficient evidence (a small enough P–value) to reject this tentative assumption in favor of H_1. In this situation the hypothesis test addresses the question "Are the data consistent with $p = p_0$ or is there sufficient evidence to conclude that $p \neq p_0$?" For this non–directional hypothesis test we take the absolute value when computing the observed value of the Z–statistic $Z_{calc} = |(\hat{p} - p_0)|/\text{S.E.}(\hat{p})$, since values of $\hat{p}$ which are far away from p_0 in either direction support $p \neq p_0$ over $p = p_0$. Thus the P–value for this hypothesis test is the probability that $|Z| \geq Z_{calc}$.

For all of these hypothesis tests, the P–value is computed under the assumption that H_0 is true, and the P–value is the probability of observing a value of $\hat{p}$ that is as extreme or more extreme, relative to p_0, than the value we actually observed, under the assumption that H_0 is true (in particular $p = p_0$). In this statement the definition of extreme (large, small, or far from in either direction) depends on the form of H_1.

5.6 Exercises

For each of the following examples:

a) Define the relevant population success proportion or probability. Be sure to indicate the corresponding population.

b) Using the information provided, formulate an appropriate research hypothesis about the population success proportion and briefly explain why your hypothesis is appropriate.

c) Perform a hypothesis test to determine whether the data support your research hypothesis. Provide the P–value and briefly summarize your conclusion in the context of the example.

d) Construct a 95% confidence interval for the success proportion and interpret it in context of the example.

1. A company which provides telephone based support for its products has found that 20% of the users of this service file complaints about the quality of the service they receive. Recently this company retrained its support personnel with the hope of reducing the percentage of users who file complaints. A random sample of 150 customers who used the telephone support after the support personnel had been retrained revealed that 20 customers were not satisfied with the quality of support they received.

2. A manufacturer has found that 15% of the items produced at its old manufacturing facility fail to pass final inspection and must be remanufactured before they can be sold. This manufacturer has recently opened a new manufacturing facility and wants to determine whether the items produced at the new facility are more or less likely to fail inspection and require remanufacture. A random sample of 200 items is selected from a large batch of items produced at the new facility and of these 42 fail inspection and require remanufacturing.

3. A supplier of vegetable seeds has a large number of bean seeds left over from last season and is trying to decide whether these seeds are suitable for sale for the current season. This supplier normally advertises that more than 85% of its bean seeds will germinate. A random sample of 200 of the leftover beans seeds was selected and of these 200 seeds 181 germinated.

Chapter 6
Comparing Two Proportions

6.1 Introduction

In this chapter we consider inferential methods for comparing two population proportions p_1 and p_2. More specifically, we consider methods for making inferences about the difference $p_1 - p_2$ between two population proportions p_1 and p_2. The inferential methods for a single proportion p discussed in Chapter 5 are based on a large sample size normal approximation to the sampling distribution of $\hat{p}$. The inferential methods we will discuss in this chapter are based on an analogous large sample size normal approximation to the sampling distribution of $\hat{p}_1 - \hat{p}_2$. Sections 6.2 and 6.3 deal with inferential methods appropriate when the data consist of independent random samples. The modifications needed for dependent (paired) samples are discussed in Section 6.4.

6.2 Estimation for two proportions (independent samples)

In some applications there are two actual physical dichotomous populations so that p_1 denotes the population success proportion for population one and p_2 denotes the population success proportion for population two. In other applications, such as randomized comparative experiments p_1 and p_2 denote hypothetical population success probabilities corresponding to two treatments. We will assume that the data correspond to two independent sequences of Bernoulli trials: a sequence of n_1 Bernoulli trials with population success probability p_1 and an independent sequence of n_2 Bernoulli trials with population success probability p_2. The assumption that these are independent sequences of Bernoulli trials means that the outcomes of all $n_1 + n_2$ trials are independent. When sampling from physical populations these assumptions are equivalent to assuming that the data consist of two independent simple random samples (of sizes n_1 and n_2) selected with replacement from dichotomous populations with population success proportions p_1 and p_2. In this context the assumption of independence basically means that the method used to select the random sample from the first population is not influenced by the method used to select the random sample from the second population, and *vice versa*.

The observed success proportions $\hat{p}_1$ and $\hat{p}_2$ are the obvious estimates of the two population success proportions p_1 and p_2; and the difference $\hat{p}_1 - \hat{p}_2$ between these observed success proportions is the obvious estimate of difference $p_1 - p_2$ between the two population success proportions. The behavior of $\hat{p}_1 - \hat{p}_2$ as an estimator of $p_1 - p_2$ can be determined from its sampling distribution. As you might expect, since $\hat{p}_1$ and $\hat{p}_2$ are unbiased estimators of p_1 and p_2, $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of $p_1 - p_2$. Thus the sampling

distribution of $\hat{p}_1 - \hat{p}_2$ has mean equal to $p_1 - p_2$. The standard deviation of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is the **population standard error** of $\hat{p}_1 - \hat{p}_2$

$$\text{S.E.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Notice that the population variance $\text{var}(\hat{p}_1 - \hat{p}_2)$ (the square of $\text{S.E.}(\hat{p}_1 - \hat{p}_2)$) is equal to the sum of the population variance of $\hat{p}_1$ and the population variance of $\hat{p}_2$. This property is a consequence of our assumption that the random samples are independent. This expression for the standard error of the difference between two sample success proportions is not appropriate if the random samples are not independent.

As was the case for the sampling distribution of a single sample proportion, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is not the same when $\hat{p}_1$ and $\hat{p}_2$ are based on samples selected without replacement as it is when $\hat{p}_1$ and $\hat{p}_2$ are based on samples selected with replacement. In both sampling situations, the mean of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$. Thus $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of $p_1 - p_2$, whether the samples are selected with or without replacement. On the other hand, as with a single proportion, the standard error of $\hat{p}_1 - \hat{p}_2$ is smaller when the samples are selected without replacement. This implies that, strictly speaking, the confidence interval estimates of $p_1 - p_2$ given below, which are based on the assumption that the samples are selected with replacement, are not appropriate when the samples are selected without replacement. However, if the sizes of the two populations are both very large relative to the sizes of the samples, then, for practical purposes, we can ignore the fact that the samples were selected without replacement. Hence, when we have samples selected without replacement and we know that the populations are very large, it is not unreasonable to compute a confidence interval estimate of $p_1 - p_2$ as if the samples were selected with replacement.

Remark. *When $\hat{p}_1$ and $\hat{p}_2$ are computed from independent simple random samples of sizes n_1 and n_2 selected without replacement from dichotomous populations of sizes N_1 and N_2, the population standard error of $\hat{p}_1 - \hat{p}_2$*

$$\textit{S.E.}(\hat{p}_1 - \hat{p}_2) = \sqrt{f_1 \frac{p_1(1-p_1)}{n_1} + f_2 \frac{p_2(1-p_2)}{n_2}},$$

is smaller than the population standard error for independent samples selected with replacement. In this situation there are two finite population correction factors $f_1 = (N_1 - n_1)/(N_1 - 1)$ and $f_2 = (N_2 - n_2)/(N_2 - 1)$ and the effect on the standard error is most noticeable when one or both of the $N's$ is small relative to the corresponding n. If N_1 and N_2 are both very large relative to the respective n_1 and n_2, then $f_1 \approx 1$, $f_2 \approx 1$, and the two standard errors are essentially equal.

We will consider inferential methods based on a large sample size normal approximation to the sampling distribution of $\hat{p}_1 - \hat{p}_2$. This normal approximation is analogous to the normal approximation to the sampling distribution of $\hat{p}$ of Section 5.2. In the present context the normal approximation simply says that, when both n_1 and n_2 are large, the standardized value of $\hat{p}_1 - \hat{p}_2$, obtained by subtracting the population difference $p_1 - p_2$ and dividing by the population standard error of $\hat{p}_1 - \hat{p}_2$, behaves in approximate accordance with the standard normal distribution. For completeness, a formal statement of this normal approximation is given below.

The normal approximation to the sampling distribution of $\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2$. *Let $\hat{p}_1$ denote the observed proportion of successes in a sequence of n_1 Bernoulli trials with success probability p_1 (or equivalently the observed proportion of successes in a simple random sample drawn with replacement from a dichotomous population with population success proportion p_1). Let $\hat{p}_2$ denote the observed proportion of successes in a sequence of n_2 Bernoulli trials with success probability p_2 (or equivalently the observed proportion of successes in a simple random sample drawn with replacement from a dichotomous population with population success proportion p_2). Assume that these two sequences of Bernoulli trials (or random samples) are independent. Finally let $a < b$ be two given constants and*

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

If n_1 and n_2 are sufficiently large, then the probability that

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{S.E.(\hat{p}_1 - \hat{p}_2)}$$

is between a and b is approximately equal to the probability that a standard normal variable Z is between a and b. In symbols, using $\approx$ to denote approximate equality, the conclusion from above is that, for sufficiently large values of n_1 and n_2,

$$P\left(a \le \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{S.E.(\hat{p}_1 - \hat{p}_2)} \le b\right) \approx P(a \le Z \le b).$$

Remark. *If the two populations being sampled are very large relative to the sizes of the samples, then, for practical purposes, this normal approximation to the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may also be applied when $\hat{p}_1$ and $\hat{p}_2$ are based on independent simple random samples selected without replacement.*

The starting point for using this normal approximation to construct a 95% confidence interval estimate of the difference $p_1 - p_2$ between the two population success proportions is the approximate probability statement

$$P\big(|(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)| \le 1.96\text{S.E.}(\hat{p}_1 - \hat{p}_2)\big) \approx .95.$$

This probability statement indicates that the probability that the actual difference $p_1 - p_2$ is within 1.96S.E.$(\hat{p}_1 - \hat{p}_2)$ units of the observed difference $\hat{p}_1 - \hat{p}_2$ is approximately .95. As was the case with the analogous interval for one proportion, this interval is not computable, since it involves the population standard error S.E.$(\hat{p}_1 - \hat{p}_2)$ which depends on the unknown parameters p_1 and p_2 and is therefore also unknown.

The method we used to derive the Wilson confidence interval for a single proportion will not work in the present context. Therefore, in the present context we will consider a confidence interval estimate of the difference $p_1 - p_2$ based on the estimated difference $\hat{p}_1 - \hat{p}_2$ and the estimated standard error of $\hat{p}_1 - \hat{p}_2$

$$\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

We will refer to this estimated standard error as **the standard error for estimation**. The margin of error of $\hat{p}_1 - \hat{p}_2$ is obtained by multiplying this estimated standard error by a suitable constant k. (Recall that: for a 95% confidence level $k = 1.96$, for a 90% confidence level $k = 1.645$, and for a 99% confidence level $k = 2.576$.) The 95% **margin of error of** $\hat{p}_1 - \hat{p}_2$ is

$$\text{M.E.}(\hat{p}_1 - \hat{p}_2) = 1.96\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2)$$

and the interval from $(\hat{p}_1 - \hat{p}_2) - \text{M.E.}(\hat{p}_1 - \hat{p}_2)$ to $(\hat{p}_1 - \hat{p}_2) + \text{M.E.}(\hat{p}_1 - \hat{p}_2)$ is a 95% confidence interval estimate of the difference $p_1 - p_2$. Thus we can claim that we are 95% confident that the difference $p_1 - p_2$ between the population success proportions is between

$$(\hat{p}_1 - \hat{p}_2) - \text{M.E.}(\hat{p}_1 - \hat{p}_2) \quad \text{and} \quad (\hat{p}_1 - \hat{p}_2) + \text{M.E.}(\hat{p}_1 - \hat{p}_2).$$

Recall that it is the estimate $\hat{p}_1 - \hat{p}_2$ and the margin of error M.E.$(\hat{p}_1 - \hat{p}_2)$ which vary from sample to sample. Therefore, the 95% confidence level applies to the method used to generate the confidence interval estimate.

Example. Rural versus urban voter preferences. Suppose that a polling organization has separate listings of all the registered voters in a large rural district and a large urban district and wishes to compare the proportions of voters in these districts who favor a proposition which is to appear on an upcoming election ballot. Let p_1 denote the proportion of all registered voters in the rural district who favor the proposition at the time of the poll and let p_2 denote the proportion of all registered voters in the urban district who favor the proposition at the time of the poll. (In terms of the box of balls analogy of Chapter 5, we now have two boxes of balls with p_1 denoting the proportion of red balls in box one and p_2 denoting the proportion of red balls in box two.)

The most obvious way to obtain independent random samples in this scenario is to: (1) randomly generate a set of n_1 labels for the rural district, contact the corresponding voters, and compute the estimate $\hat{p}_1$ for the voters in the rural district; and, (2) randomly generate a set of n_2 labels for the urban district, contact the corresponding voters, and compute the estimate $\hat{p}_2$ for the voters in the urban district. (Select a simple random sample of balls from box one and compute $\hat{p}_1$ and, independently, select a simple random sample of balls from box two and compute $\hat{p}_2$.) Assuming that simple random samples are selected (with replacement or from large populations) this method clearly yields independent samples and the confidence interval method described above is valid.

Now suppose that we do not have separate listings of the rural voters and the urban voter but instead have a single listing of all registered voters in a large district which includes both rural and urban voters. In this situation we could randomly generate a set of n labels for the entire district, contact the corresponding voters, and in addition to determining whether the voter favors the proposition also determine whether the voter lives in a rural or urban area. We could then partition the simple random sample of n voters into the subsample of n_1 voters who live in a rural area and the subsample of n_2 voters who live in an urban area. (This is like labeling the balls in box one with a one, labeling the balls in box two with a two, then combining the balls in a single box, selecting a simple random sample of n balls from this box, and dividing it to get a sample of n_1 balls from box one and a sample of n_2 balls from box two.) This approach yields independent random samples but, technically (based on the formal definition), these random samples are not simple random samples, since the sample sizes n_1 and n_2 were not selected in advance. Actually this is not a problem, since it is readily verified that the samples can be viewed as independent sequences of Bernoulli trials (exactly if selection is with replacement and approximately if selection is without replacement from a large population and both subpopulations are also large). Therefore, the confidence interval method described above is also valid when this alternate method of forming independent random samples by partitioning a simple random sample is used.

Example. An opinion poll. The purpose of this example is to demonstrate the application of a 95% confidence interval for $p_1 - p_2$. To make the numbers more realistic we will use numbers from a *New York Times*/CBS News poll conducted September 9–13, 2005. To place this in context note that hurricane Katrina made landfall on September 1, 2005. Like all such national polls this poll was not based on a simple random sample; it employed a complex random sampling method involving stratification and clustering.

Suppose that a listing of telephone numbers for a well–defined population of adults in the U.S. was used to select a simple random sample of $n = 1,167$ adults. When asked "Are you white, black, Asian, or some other race?" 877 of these 1,167 adults chose white and 211 chose black. Therefore, we have independent simple random samples of size $n_1 = 877$

(from the subpopulation of white adults) and $n_2 = 211$ (from the subpopulation of black adults).

First consider the responses to question 10: "Do you think George W. Bush has the same priorities for the country as you have, or not?" Let p_1 denote the proportion of all white adults in this population who would respond "has the same priorities" and let p_2 denote the proportion of all black adults in this population who would respond "has the same priorities". Of the $n_1 = 877$ whites 360 responded "has the same priorities" giving $\hat{p}_1 = .4105$ while 27 of the $n_2 = 211$ blacks responded "has the same priorities" giving $\hat{p}_2 = .1280$. These data clearly suggest that the population proportion p_1 is greater than the population proportion p_2, since 41.05% of the whites responded "has the same priorities" while only 12.80% of the blacks responded this way. In this situation we are 95% confident that $p_1 - p_2$ is between .2269 and .3381. Since this entire interval is positive we can conclude that we are 95% confident that the population proportion of whites who would have responded "has the same priorities" if all had been asked exceeds the analogous population proportion for blacks by at least .2269 and perhaps as much as .3381. In other words, we are 95% confident that the percentage of all whites who would have responded "has the same priorities" exceeds the corresponding proportion for blacks by between 22.69 and 33.81 percentage points.

Next consider the responses to question 14: "Do you think Congress has the same priorities for the country as you have, or not?" Let p_1 denote the proportion of all white adults in this population who would respond "has the same priorities" and let p_2 denote the proportion of all black adults in this population who would respond "has the same priorities". Of the $n_1 = 877$ whites 252 responded "has the same priorities" giving $\hat{p}_1 = .2873$ while 51 of the $n_2 = 211$ blacks responded "has the same priorities" giving $\hat{p}_2 = .2417$. In this case it is not clear whether the data suggest that the population proportion p_1 is greater than the population proportion p_2, since the sample proportions are reasonably similar. In this situation we are 95% confident that $p_1 - p_2$ is between –.0194 and .1107. Since the lower limit of this interval is negative (suggesting $p_1 < p_2$) and the upper limit of this interval is positive (suggesting $p_1 > p_2$) we cannot exclude the possibility that the population proportions p_1 and p_2 are the same.

Finally consider the responses to question 62: "As a result of the recent increase in gas prices, have you cut back on household spending on other things?" Let p_1 denote the proportion of all white adults in this population who would respond yes and let p_2 denote the proportion of all black adults in this population who would respond yes. Of the $n_1 = 877$ whites 517 responded yes giving $\hat{p}_1 = .5895$ while 158 of the $n_2 = 211$ blacks responded yes giving $\hat{p}_2 = .7588$. These data clearly suggest that the population proportion p_1 is less than the population proportion p_2, since only 58.95% of the whites responded yes while 75.88% of the blacks responded yes. In this situation we are 95% confident that

$p_1 - p_2$ is between -.2263 and -.0923 (or equivalently that $p_2 - p_1$ is between .0923 and .2263). Since this entire interval (for $p_1 - p_2$) is negative we can conclude that we are 95% confident that the population proportion of whites who would have responded yes if all had been asked is less than the analogous population proportion for blacks by at least .0923 and perhaps as much as .2263. In other words, we are 95% confident that the percentage of all blacks who would have responded yes exceeds the corresponding percentage for whites by between 9.23 and 22.63 percentage points.

Another common application of this confidence interval for the difference between two population proportions is for randomized comparative experiments. Consider a randomized comparative experiment where $N = n_1 + n_2$ available units are randomly assigned to receive one of two treatments (with n_1 units assigned to treatment 1 and the remaining n_2 units assigned to treatment 2). We can imagine two hypothetical populations of responses and two population success proportions corresponding to the two treatments. The first hypothetical population is the collection of responses (S or F), corresponding to all N available units, which we would observe if all N available units were subjected to treatment 1 and p_1 is the proportion of successes among these units. The second hypothetical population and population success proportion p_2 are defined similarly to correspond to the responses we would observe if all N available units were subjected to treatment 2.

The model corresponding to the assumptions we made to justify the confidence interval for $p_1 - p_2$ treats the data as if they constitute independent simple random samples selected with replacement from these two hypothetical populations. In terms of balls in a box, this means that we are assuming that we have independent simple random samples selected with replacement from two separate boxes of balls, with each box containing N balls. Clearly this model is not appropriate for this application; a more appropriate model treats the data as two dependent random samples selected without replacement from a single box of N balls. Fortunately, even though the underlying assumptions are not valid for this application the method still works reasonably well. Before we describe why it is helpful to consider a specific example.

Example. Leading questions. The wording of questions in surveys can have a major impact on the responses elicited. The effect of wording of questions was investigated in Schuman and Presser, Attitude measurement and the gun control paradox, *Public Opinion Quarterly*, **41** winter 1977–1978, 427–438. Two groups of adults were used to estimate the difference in response to the following two versions of a question regarding gun control.

1. Would you favor or oppose a law which would require a person to obtain a police permit before he could buy a gun?

2. Would you favor a law which would require a person to obtain a police permit before he could buy a gun, or do you think that such a law would interfere too much with the right of citizens to own guns?

We might expect the second version of the question, with the added remark about the right of citizens to own guns, to lead to less responses in favor of requiring a permit.

This study was conducted in 1976. The researchers began with a group of 1263 adults which had been obtained by a random sampling method for a survey conducted by the Survey Research Center of the University of Michigan. These 1263 adults were randomly divided into two groups with 642 adults in the first group and 621 adults in the second group. The adults in the first group were asked to respond to the first version of the gun control question and the adults in the second group were asked to respond to the second version of the gun control question. Twenty–seven adults in the first group and 36 adults in the second group would not respond to the question. Therefore, we will restrict our attention to the 1200 adults who were willing to respond to a question about gun control, and we will use the $n_1 = 615$ adults in the first group and the $n_2 = 585$ adults in the second group who responded to the question as our samples.

In this randomized comparative experiment the group of available units is the group of 1200 adults who were willing to respond to a question about gun control in 1976. Let p_1 denote the proportion of these 1200 adults who would respond "favor" (in 1976) if all 1200 were asked the first question and let p_2 denote the proportion of these 1200 adults who would respond "favor" (in 1976) if all 1200 were asked the second question. Our goal is to estimate the difference $p_1 - p_2$ between these proportions. When the study was conducted 463 of the 615 adults in the first group responded "favor" and 403 of the 585 adults in the second group responded "favor". The observed proportions of adults who respond "favor" are $\hat{p}_1 = .7528$ and $\hat{p}_2 = .6889$ giving a difference of $\hat{p}_1 - \hat{p}_2 = .0639$. The standard error is $\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2) = .02586$ and the margin of error is $\text{M.E.}(\hat{p}_1 - \hat{p}_2) = .0507$; therefore, we are 95% confident that the difference $p_1 - p_2$ is between $.0639 - .0507 = .0132$ and $.0639 + .0507 = .1146$. That is, we are 95% confident that modifying the first question about gun control by adding the comment about the right of citizens to own guns lowers the probability that an individual adult (from this group of 1200 adults) would respond "favor" (in 1976) by at least .0132 and at most .1146.

In summary, we estimate that, in 1976, about 75.28% of these 1200 adults would respond "favor" if asked the first question and we estimate that, if these same people had instead been asked the second question with the comment about the right of citizens to own guns, then we would see a reduction of this percentage in the range of 1.32 to 11.46 percentage points. Thus we find sufficient evidence to conclude that the added comment has the anticipated effect of lowering the percentage who would respond "favor"; note, however, that this reduction might be as small as 1.32 percentage points, as large as 11.46

percentage points, or anywhere within this range. As we noted above these 1200 adults can be viewed as a random sample from the population of adults sampled by the University of Michigan researchers who would respond to a question about gun control, thus it is reasonable to claim that this inference applies to this entire population of adults (in 1976) not just these 1200.

Returning to our discussion of the validity of the assumptions for a randomized comparative experiment we will now expand on the single box of N balls model for this situation. Imagine a box containing N balls and suppose that each ball is marked with two values, one indicating the response to treatment 1 and the other indicating the response to treatment 2. Randomly assigning n_1 units to treatment 1 and observing their response to the treatment is like selecting a simple random sample of n_1 balls without replacement from this box of N balls and observing the values corresponding to treatment 1 on these balls. Once these n_1 balls have been selected for treatment 1 there are only n_2 balls left in the box and we have no choice in our selection of the balls for treatment 2. Thus we cannot view these as independent samples. Furthermore, in this application both of the sample sizes n_1 and n_2 are usually large relative to the number of available units N (often each is approximately half of N) and we should not ignore the fact that we are sampling without replacement.

The fact that the samples are selected without replacement causes the formula we are using for the standard error of $\hat{p}_1 - \hat{p}_2$ to overstate the amount of variability in $\hat{p}_1 - \hat{p}_2$ and as a result this causes the estimate of the standard error used to construct the confidence interval to be too large which makes the confidence interval longer than it should be.

We will discuss the dependence of these samples in the context of the leading question example but the same basic argument applies to randomized comparative experiments in general. We might argue that an individual with strong feelings (pro or con) about gun control would probably respond the same way (favor or oppose) whether the individual was asked the first or second question. If by the luck of the draw many individuals who are strongly supportive of gun control happen to be assigned to the group asked the first question, then there will be fewer such individuals to be assigned to the group asked the second question. This suggests that random assignments which tend to make $\hat{p}_1$ larger (smaller) tend at the same time to make $\hat{p}_2$ smaller (larger). Therefore, in this context we expect negative association between $\hat{p}_1$ and $\hat{p}_2$ so that assignments which give large (small) values of $\hat{p}_1$ tend to give small (large) values of $\hat{p}_2$.

This type of dependence (negative association between $\hat{p}_1$ and $\hat{p}_2$) causes the formula we are using for the standard error of $\hat{p}_1 - \hat{p}_2$ to understate the amount of variability in $\hat{p}_1 - \hat{p}_2$ and as a result this causes the estimate of the standard error used to construct the confidence interval to be too small which makes the confidence interval shorter than it should be.

Fortunately, provided that n_1 and n_2 are reasonably large, the effects of these two violations of the underlying assumptions tend to cancel each other and the confidence interval based on the assumptions of independent simple random samples selected with replacement work reasonably well for randomized comparative experiments.

Remark. *The use of one of the confidence limits of a 90% confidence interval as a 95% confidence bound discussed in Section 5.4 can also be used in the present context. Thus, we can find an upper or lower 95% confidence bound for $p_1 - p_2$ by selecting the appropriate confidence limit from a 90% confidence interval estimate of $p_1 - p_2$.*

6.3 Testing hypotheses about two proportions (independent samples)

In this section we will consider hypothesis tests for hypotheses relating two population success proportions p_1 and p_2. The tests we consider are based on the same normal approximation to the sampling distribution of $\hat{p}_1 - \hat{p}_2$ that we used for confidence estimation. Thus we will assume that the data on which the hypothesis test is based correspond to two independent simple random samples of sizes n_1 and n_2, selected with replacement, from dichotomous populations with population success proportions p_1 and p_2, or equivalently, that the data correspond to the outcomes of two independent sequences of n_1 and n_2 Bernoulli trials with success probabilities p_1 and p_2. However, as with confidence estimation, for practical purposes, we do not need to worry about whether the samples are selected with or without replacement, provided both of the populations are very large; and, these tests are also applicable to randomized comparative experiments.

Many hypotheses about the relationship between the population proportions p_1 and p_2 can be expressed as hypotheses about the relationship between $p_1 - p_2$ and zero, *e.g.*, $p_1 > p_2$ is equivalent to $p_1 - p_2 > 0$. Therefore, we will consider tests which are based on a suitably standardized value of the difference $\hat{p}_1 - \hat{p}_2$ between the observed success proportions.

The P–value for a hypothesis about the relationship between a single proportion p and a hypothesized value p_0 is computed under the assumption that $p = p_0$, therefore, we used $p = p_0$ in the standard error of $\hat{p}$ for the Z–statistic of the test. The P–value for a hypothesis about the relationship between p_1 and p_2 is computed under the assumption that $p_1 = p_2$, therefore, we need to determine a suitable standard error of $\hat{p}_1 - \hat{p}_2$ (the standard error for testing) under this assumption. Notice that $p_1 = p_2$ ($p_1 - p_2 = 0$) specifies a common value for p_1 and p_2 but does not specify what this common value is, *e.g.*, we might have $p_1 = p_2 = .5$ or $p_1 = p_2 = .1$. When $p_1 = p_2$, $\hat{p}_1$ and $\hat{p}_2$ are estimates of the same population success proportion. This suggests that we can pool or combine the information in the two random samples to obtain a pooled estimate, $\hat{p}$, of this common population success proportion. This pooled estimate $\hat{p}$ can then be used to get an

estimate of S.E.$(\hat{p}_1 - \hat{p}_2)$ that is suitable for use in the hypothesis test. If we let p denote the common population success proportion under the assumption that $p_1 = p_2$, then the population standard error of $\hat{p}_1 - \hat{p}_2$ simplifies to

$$\text{S.E.}(\hat{p}_1 - \hat{p}_2) = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Replacing p in this population standard error by the pooled estimate $\hat{p}$ gives **the standard error for testing**

$$\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

where

$$\hat{p} = \frac{\text{the total number of successes in both samples}}{\text{the total number of observations in both samples}} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

When testing $H_0 : p_1 \leq p_2$ versus $H_1 : p_1 > p_2$ values of $\hat{p}_1 - \hat{p}_2$ which are sufficiently larger than zero provide evidence against the null hypothesis $H_0 : p_1 \leq p_2$ and in favor of the research hypothesis $H_1 : p_1 > p_2$. Thus large (positive) values of

$$Z_{calc} = \frac{\hat{p}_1 - \hat{p}_2}{\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2)},$$

where $\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2)$ denotes the standard error for testing, favor the research hypothesis and the P–value is the probability that a standard normal variable takes on a value at least as large as Z_{calc}, *i.e.*, the P–value is the area under the standard normal density curve to the right of Z_{calc}.

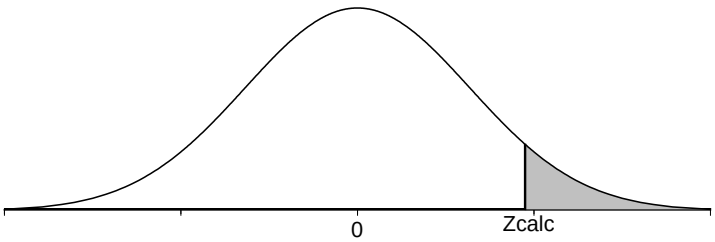
The steps for performing a hypothesis test for

$$H_0 : p_1 \leq p_2 \quad \text{versus} \quad H_1 : p_1 > p_2$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(Z \geq Z_{calc})$, where Z denotes a standard normal variable and Z_{calc} is as defined above. This P–value is the area under the standard normal density curve to the right of Z_{calc} as shown in Figure 1.

Figure 1. P–value for $H_0 : p_1 \leq p_2$ versus $H_1 : p_1 > p_2$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p_1 > p_2$ over $H_0 : p_1 \leq p_2$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population proportion p_1 is greater than the population success proportion p_2.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p_1 > p_2$ over $H_0 : p_1 \leq p_2$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population proportion p_1 is greater than the population success proportion p_2.

When testing $H_0 : p_1 \geq p_2$ versus $H_1 : p_1 < p_2$ values of $\hat{p}_1 - \hat{p}_2$ which are sufficiently smaller than zero provide evidence against the null hypothesis $H_0 : p_1 \geq p_2$ and in favor of the research hypothesis $H_1 : p_1 < p_2$. Thus sufficiently negative values of Z_{calc} (as defined above) favor the research hypothesis and the P–value is the probability that a standard normal variable takes on a value no larger than Z_{calc}, *i.e.*, the P–value is the area under the standard normal density curve to the left of Z_{calc}.

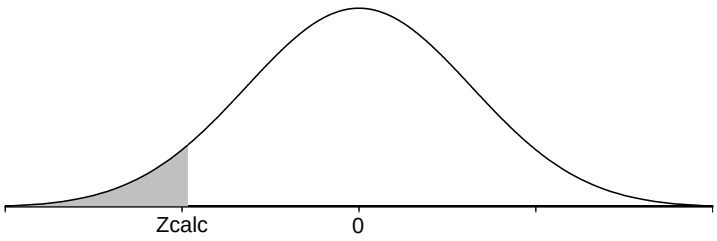
The steps for performing a hypothesis test for

$$H_0 : p_1 \geq p_2 \quad \text{versus} \quad H_1 : p_1 < p_2$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(Z \leq Z_{calc})$, where Z denotes a standard normal variable and Z_{calc} is as defined above. This P–value is the area under the standard normal density curve to the left of Z_{calc} as shown in Figure 2.

Figure 2. P–value for $H_0 : p_1 \geq p_2$ versus $H_1 : p_1 < p_2$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p_1 < p_2$ over $H_0 : p_1 \geq p_2$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population proportion p_1 is less than the population success proportion p_2.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p_1 < p_2$ over $H_0 : p_1 \geq p_2$. That

is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population proportion p_1 is less than the population success proportion p_2.

When testing $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$ values of $\hat{p}_1 - \hat{p}_2$ which are sufficiently far away from zero in either direction provide evidence against the null hypothesis $H_0 : p_1 = p_2$ and in favor of the research hypothesis $H_1 : p_1 \neq p_2$. Thus sufficiently large values of the absolute value of Z_{calc} (as defined above) favor the research hypothesis and the P–value is the probability that a standard normal variable takes on a value below $-|Z_{calc}|$ or above $|Z_{calc}|$, *i.e.*, the P–value is the combined area under the standard normal density curve to the left of $-|Z_{calc}|$ and to the right of $|Z_{calc}|$.

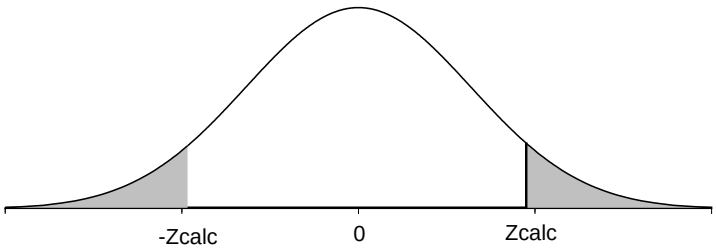
The steps for performing a hypothesis test for

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 \neq p_2$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(|Z| \geq |Z_{calc}|) = P(Z \leq -|Z_{calc}|) + P(Z \geq |Z_{calc}|)$, where Z denotes a standard normal variable and Z_{calc} is as defined above. This P–value is the combined area under the standard normal density curve to the left of $-|Z_{calc}|$ and to the right of $|Z_{calc}|$ as shown in Figure 3.

Figure 3. P–value for $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p_1 \neq p_2$ over $H_0 : p_1 = p_2$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportions p_1 and p_2 are different.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p_1 \neq p_2$ over $H_0 : p_1 = p_2$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportions p_1 and p_2 are different.

Example. An HIV vaccine trial. This example is based on a study described in Flynn *et al.*, Placebo–controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to

prevent HIV–1 infection, *J. of Infect. Dis.*, **191** Mar. 1, 2005, 654–665. A double–blind randomized trial was conducted to investigate the effect of an rgp120 vaccine among men who have sex with men and among women at high risk for heterosexual transmission of type 1 HIV. A group of 5403 volunteers (5095 men and 308 women) was randomly divided into two groups (a control group ($n_1 = 1805$) and a vaccine group ($n_2 = 3598$)). Each volunteer received 7 injections of either placebo or vaccine over a 30 month period. These individuals were tracked for a period of 3 years to see whether they developed HIV–1.

We can envision two hypothetical populations based on this group of 5403 individuals and these two experimental treatments. Since these 5403 volunteers do not form a random sample from some well defined population of people at high risk for developing HIV–1 we should restrict our inferences to these 5403 volunteers. Let p_1 denote the proportion of this group of 5403 volunteers who would develop HIV–1 within 3 years if all 5403 volunteers were given the placebo. Let p_2 denote the proportion of this group of 5403 volunteers who would develop HIV–1 within 3 years if all 5403 volunteers were given the vaccine. We can also think of these proportions as the probabilities that one of these 5403 volunteers would develop HIV–1 within 3 years if he or she was treated with the placebo (p_1) or if he or she was treated with the vaccine (p_2). In terms of these parameters our research hypothesis is $H_1 : p_1 > p_2$ (the vaccine reduces the risk of developing HIV–1) and our null hypothesis is $H_0 : p_1 \leq p_2$ (the vaccine does not reduce the risk of developing HIV–1).

By the end of the 3 years, 126 of the 1805 individuals treated with the placebo developed HIV–1 while 241 of the 3598 individuals treated with the vaccine developed HIV–1. The observed proportions are $\hat{p}_1 = .0698$ and $\hat{p}_2 = .0670$, and the difference is $\hat{p}_1 - \hat{p}_2 = .0028$. The fact that this difference is positive ($\hat{p}_1$ is greater than $\hat{p}_2$) shows that there is some evidence in favor of the research hypothesis $p_1 > p_2$. We need to determine whether observing a difference of .0028, with samples of size $n_1 = 1805$ and $n_2 = 3598$, is sufficiently surprising under the assumption that $p_1 \leq p_2$ to allow us to reject this null hypothesis as untenable. When we use the standard error for testing to standardize this difference we get $Z_{calc} = .3892$. The corresponding P–value$= P(Z \geq .38929) = .3486$ is quite large. In words, this means that (for these sample sizes) if the null hypothesis was true (p_1 was actually no greater than p_2), then we would observe a difference this far above zero about 34.86% of the time. In other words, for the volunteers used in this study, these data do not provide enough evidence to allow us to claim that this vaccine is better than a placebo.

Example. Scotland coronary prevention study. This example is based on the West of Scotland Coronary Prevention Study as described in Shepherd *et al.*, Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia, *New England Journal of Medicine*, **333** Nov. 16, 1995, 1994–1307, and Ford *et al.*, Long–term follow–up of the West of Scotland coronary prevention study, *New England Journal of Medicine*,

357 Oct. 11, 2007, 1477–1486. The primary goal of this study was to determine whether the administration of pravastatin to middle–aged men with high cholesterol levels and no history of myocardial infarction over a period of five years reduces the risk of coronary events. In this context a coronary event is defined as a nonfatal myocardial infarction or death from coronary heart disease. A group of 6595 men, aged 45 to 64 years, with high plasma cholesterol levels (mean 272 mg/dl) was randomly divided into two groups (a control group and a treatment group). The 3302 men in the treatment group received 40 mg of pravastatin daily while the 3293 men in the control group received a placebo. All of the men were given smoking cessation and dietary advice throughout the study.

We can envision two hypothetical populations based on this group of 6595 men and these two experimental treatments. Since these 6595 men do not form a random sample from some well defined population of middle–aged men with high cholesterol levels we should restrict our inferences to these 6595 men. However, the investigators examined these men to see if extrapolations beyond this group may be reasonable and they concluded that: "The subjects in this study were representative of the general population in terms of socioeconomic status and risk factors (Table 1). Their plasma cholesterol levels were in the highest quartile of the range found in the British population. A number had evidence of minor vascular disease, and in order to make the findings of the trial applicable to typical middle-aged men with hypercholesterolemia, they were not excluded."

Let p_1 denote the proportion of this group of 6595 men who would experience a cardiac event (as defined above) if all 6595 men were subjected to the five year pravastatin treatment. Let p_2 denote the proportion of this group of 6595 men who would experience a cardiac event if all 6595 men were subjected to the five year placebo treatment. We can also think of these proportions as the probabilities that one of these 6595 men would have a cardiac event within five years if he was treated with pravastatin (p_1) or if he was treated with placebo (p_2). In terms of these parameters our research hypothesis is $H_1 : p_1 < p_2$ (pravastatin reduces the risk of a coronary event) and our null hypothesis is $H_0 : p_1 \geq p_2$ (pravastatin does not reduce the risk of a coronary event).

By the end of this five year trial, 174 of the 3302 men treated with pravastatin had experienced a cardiac event and 248 of the 3293 men treated with a placebo had experienced a cardiac event. The observed proportions are $\hat{p}_1 = .0527$ and $\hat{p}_2 = .0753$, and the difference is $\hat{p}_1 - \hat{p}_2 = -.0226$. The fact that this difference is negative ($\hat{p}_1$ is less than $\hat{p}_2$) shows that there is some evidence in favor of the research hypothesis $p_1 < p_2$. We need to determine whether observing a difference of $-.0226$, with samples of size $n_1 = 3302$ and $n_2 = 3293$, is sufficiently surprising under the assumption that $p_1 \geq p_2$ to allow us to reject this null hypothesis as untenable. When we use the standard error for testing to standardize this difference we get $Z_{calc} = -3.7523$. The corresponding P–value= $P(Z \leq -3.7523)$ is less than .0001 (approximately 8.8×10^{-5}). In words, this means that (for these sample sizes) if

the null hypothesis was true (p_1 was actually no less than p_2), then we would almost never (less than .01% of the time) observe a difference this far below zero. Therefore, these data provide very strong evidence in favor of the research hypothesis that pravastatin reduces the probability of a cardiac event in the sense that the probability that one of these 6595 men would have a cardiac event within five years would be lower if he was treated with pravastatin than if he was treated with placebo.

In addition to this conclusion that pravastatin reduces the probability of a cardiac event we can construct a confidence interval to quantify the practical importance of this reduction. In this example we are 95% confident that $p_1 - p_2$ is between -.0344 and -.0108 ($p_2 - p_1$ is between .0108 and .0344).

In summary, for these 6595 men, we have very strong evidence (P–value $< .0001$) that pravastatin reduces the risk of a cardiac event (versus placebo). We estimate that about 7.53% of these men would have a cardiac event if they all were treated with a placebo, and we are 95% confident that if they all were treated with pravastatin we would see a 1.08 to 3.44 percentage point reduction in this percentage. Since we are dealing with small percentages it is instructive to note that a reduction from 7.53% ($\hat{p}_2$) to 5.27% ($\hat{p}_1$) is a 30% reduction ($(7.53 - 5.27)/7.53 = .3001$) in the risk of a man having a cardiac event.

A follow–up to this study tracked the men used in this trial for ten additional years to assess the long term effects of treatment with pravastatin. At the end of the five year trial, treatment with pravastatin or placebo ceased, and the patients returned to the care of their primary care physicians. Five years after the conclusion of the trial 38.7% of the original pravastatin group and 35.2% of the original placebo group were being treated with statin drugs. The purpose of the follow–up study was to assess long–term effects regardless of treatment received after the initial trial period.

For this part of the study, let p_3 denote the proportion of this group of 6595 men who would experience a cardiac event within 15 years of the beginning of the initial trial if all 6595 men were subjected to the five year pravastatin treatment. Let p_4 denote the analogous proportion if all the men were subjected to the placebo treatment. In terms of these parameters our research hypothesis is $H_1 : p_3 < p_4$ (pravastatin reduces the long–term risk of a coronary event) and our null hypothesis is $H_0 : p_3 \geq p_4$ (pravastatin does not reduce the long–term risk of a coronary event).

By the end of the 15 year period, 390 of the 3302 men treated with pravastatin had experienced a cardiac event and 509 of the 3293 men treated with a placebo had experienced a cardiac event. The observed proportions are $\hat{p}_3 = .1181$ and $\hat{p}_4 = .1546$, and the difference is $\hat{p}_3 - \hat{p}_4 = -.0365$. The fact that this difference is negative ($\hat{p}_3$ is less than $\hat{p}_4$) shows that there is some evidence in favor of the research hypothesis $p_3 < p_4$. Since the sample sizes for this test are the same as for the test above and since the difference in this case is more extreme than before, we know that the P–value will be even smaller.

In this case, when we use the standard error for testing to standardize this difference we get $Z_{calc} = -4.3136$. The corresponding P–value= $P(Z \leq -4.3146)$ is less than .0001 (approximately 8.0×10^{-6}). Therefore, these data provide very strong evidence in favor of the research hypothesis that the five year pravastatin treatment reduces the probability of a cardiac event in the long–term in the sense that the probability that one of these 6595 men would have a cardiac event within 15 years would be lower if he was treated with pravastatin than if he was treated with placebo. In this case we are 95% confident that p_4 exceeds p_3 by at least .0199 and perhaps as much as .0530. Here we would estimate that about 15.46% of these men would have a cardiac event within 15 years if they were all given the five year placebo treatment and we are 95% confident that the five year pravastatin treatment would reduce this percentage by between 1.99 and 5.30 percentage points.

6.4 Inference for two proportions (paired samples)

The inferential methods for comparing two population success proportions p_1 and p_2 we have considered thus far require independent estimates $\hat{p}_1$ and $\hat{p}_2$. We will now show how these methods can be modified when $\hat{p}_1$ and $\hat{p}_2$ are dependent.

In some situations each unit in the first sample is paired with a corresponding unit in the second sample. The units which form a pair may be the same unit measured at two times or measured under two treatments; or the units which form a pair may be distinct units which are matched on the basis of characteristics believed to be related to the response of interest.

Consider the problem of assessing the effect of a debate between two candidates (A and B) in an upcoming election on voter opinion. Let p_1 denote the population proportion of voters who favor candidate A on the day before the debate and let p_2 denote the population proportion of all voters who favor candidate A on the day after the debate. Instead of selecting two independent simple random samples of voters, we could select a single simple random sample of voters and get responses (whether the voter favors candidate A) for each of these voter one day before the debate and one day after the debate.

Suppose that we wish to compare two methods of training workers to perform a complex task. Let p_1 denote the probability that a worker could perform this task satisfactorily if the worker was trained using the first method and let p_2 denote the probability that a worker could perform this task satisfactorily if the worker was trained using the second method. Instead of randomly assigning workers to two groups, we could use preliminary information about the ability of the workers to perform this task to form matched pairs of workers (each having essentially the same ability). For each pair we could randomly assign one member to be trained using the first method and the other to be trained using the second method. Then we could determine whether each worker could successfully perform the task.

In both of the situations described above the data consist of n ordered pairs of responses (response 1, response 2). Letting S denote a success and F denote a failure, the four possible response pairs are: (S,S), (S,F), (F,S), and (F,F). The probability model for these responses shown in Table 1 is determined by the corresponding population probabilities p_{SS}, p_{SF}, p_{FS}, and p_{FF}. Notice that these four probabilities must sum to one.

Table 1. Probability model for paired dichotomous responses

response 1	response 2	probability
S	S	p_{SS}
S	F	p_{SF}
F	S	p_{FS}
F	F	p_{FF}

The probability that the first response is a success is $p_1 = p_{SS} + p_{SF}$, the probability that the second response is a success is $p_2 = p_{SS} + p_{FS}$, and the difference is $p_1 - p_2 = p_{SF} - p_{FS}$. Therefore, the probabilities p_{SF} and p_{FS} of the outcomes SF and FS where the responses are different determine the difference between the first and second response probabilities. When $\hat{p}_1$ and $\hat{p}_2$ are computed from a random sample of n paired responses, $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_1 - \hat{p}_2$ are unbiased estimators of p_1, p_2, and $p_1 - p_2$. In this situation the **population standard error** of $\hat{p}_1 - \hat{p}_2$

$$\text{S.E.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_{SF} + p_{FS} - (p_{SF} - p_{FS})^2}{n}},$$

depends on the sample size n and the two probabilities p_{SF} and p_{FS}. When n is large, the standardized value of $\hat{p}_1 - \hat{p}_2$, obtained by subtracting the population difference $p_1 - p_2$ and dividing by this population standard error of $\hat{p}_1 - \hat{p}_2$, behaves in approximate accordance with the standard normal distribution.

Given a simple random sample of n response pairs we can use the observed proportions of (S,F) and (F,S) pairs $\hat{p}_{SF}$ and $\hat{p}_{FS}$ to estimate the standard error of $\hat{p}_1 - \hat{p}_2$.

For confidence estimation the estimated standard error of $\hat{p}_1 - \hat{p}_2$ is

$$\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{SF} + \hat{p}_{FS} - (\hat{p}_{SF} - \hat{p}_{FS})^2}{n}}.$$

The 95% **margin of error of** $\hat{p}_1 - \hat{p}_2$ is

$$\text{M.E.}(\hat{p}_1 - \hat{p}_2) = 1.96\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2)$$

and the interval from $(\hat{p}_1 - \hat{p}_2) - \text{M.E.}(\hat{p}_1 - \hat{p}_2)$ to $(\hat{p}_1 - \hat{p}_2) + \text{M.E.}(\hat{p}_1 - \hat{p}_2)$ is a 95% confidence interval estimate of the difference $p_1 - p_2$.

When computing the P–value for a hypothesis test we will assume that $p_1 = p_2$ which is equivalent to assuming that $p_{SF} = p_{FS}$. Under this assumption the population standard error of $\hat{p}_1 - \hat{p}_2$ simplifies to

$$\text{S.E.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_{SF} + p_{FS}}{n}}.$$

Thus for hypothesis testing the estimated standard error of $\hat{p}_1 - \hat{p}_2$ is

$$\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{SF} + \hat{p}_{FS}}{n}}.$$

The Z–statistic for this situation is

$$Z_{calc} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\hat{p}_{SF} + \hat{p}_{FS})/n}} = \frac{n_{SF} - n_{FS}}{\sqrt{n_{SF} + n_{FS}}},$$

where n_{SF} and n_{FS} are the respective frequencies of (S,F) and (F,S) pairs. Notice that this test statistic only depends on the frequencies n_{SF} and n_{FS}, it does not depend on the sample size n.

Example. Instant coffee purchases. This example is based on a study described in Grover and Srinivasan, A simultaneous approach to market segmentation and market structuring, *J. of Marketing Research*, **24** May 1987, 139–153. The authors selected a simple random sample of households from the 4657 households constituting the 1981 MRCA market research panel. The data summarized in Table 2 correspond to a simple random sample of $n = 541$ households selected from the subpopulation of the MRCA households that purchased decaffeinated instant coffee at least twice during the one year study period. These purchases are recorded as Sanka or other to indicate the brand of coffee purchased. Let p_1 denote the population proportion of households that chose Sanka on the first purchase and let p_2 denote the population proportion of households that chose Sanka on the second purchase.

Table 2. Instant coffee purchase data

first purchase	second purchase	freq.	rel. freq.
Sanka	Sanka	155	.2865
Sanka	other	49	.0906
other	Sanka	76	.1405
other	other	261	.4824
		541	1.0000

In this sample 37.71% of the first purchases were Sanka and 42.70% of the second purchases were Sanka. Note that $\hat{p}_{SF} = .0906$ and $\hat{p}_{FS} = .1405$ indicating that 9.06% of

the households switched from Sanka to other and 14.05% of the households switched from other to Sanka. In this case $\hat{p}_1 - \hat{p}_2 = .3771 - .4270 = -.0499$, the standard error for estimation is

$$\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{.0906 + .1405 - (.0906 - .1405)^2}{541}} = .02056$$

and the 95% margin of error is M.E.$(\hat{p}_1 - \hat{p}_2) = .0403$. This gives a 95% confidence interval for $p_1 - p_2$ ranging from $-.0499 - .0403 = -.0902$ to $-.0499 + .0403 = -.0096$. Thus we are 95% confident that the proportion of all households in the subpopulation defined above that chose Sanka first is between .0096 and .0902 smaller than the proportion of all households that chose Sanka second. In other words, for this population of decaffeinated instant coffee purchasers, we are 95% confident that the percentage of all households that chose Sanka on the second purchase is .96 to 9.02 percentage points higher than the percentage of all households that chose Sanka on the first purchase.

To demonstrate the method, consider a test of the null hypothesis $H_0 : p_1 = p_2$ (the same proportion purchase Sanka first as second) versus the research hypothesis $H_1 : p_1 \neq p_2$ (the proportions are different). For this test the Z–statistic is

$$Z_{calc} = \frac{n_{SF} - n_{FS}}{\sqrt{n_{SF} + n_{FS}}} = \frac{49 - 76}{\sqrt{49 + 76}} = -2.4150$$

and the P–value is $P(Z \leq -2.415) + P(Z \geq 2.415) = .0157$. Therefore, there is sufficient evidence to conclude that p_1 and p_2 are different.

Another situation where an inference about $p_1 - p_2$ is based on dependent estimates $\hat{p}_1$ and $\hat{p}_2$ arises when a single sample of units is categorized into three or more categories. Suppose that three or more candidates are listed on a ballot and we want to compare the proportion of all voters who favor candidate A, p_A, with the proportion of all voters who favor candidate B, p_B. Let $p_C = 1 - (p_A + p_B)$ denote the proportion of all voters who favor neither A nor B or who have no opinion. The probability model for this situation given in Table 3 is determined by the corresponding population probabilities p_A, p_B, and p_C. Notice that these three probabilities must sum to one.

Table 3. Probability model for trichotomous responses

response	probability
A	p_A
B	p_B
C	p_C

Assuming that the data form a simple random sample of size n; $\hat{p}_A$, $\hat{p}_B$ and $\hat{p}_A - \hat{p}_B$ are unbiased estimators of p_A, p_B, and $p_A - p_B$. In this situation the **population standard error** of $\hat{p}_A - \hat{p}_B$,

$$\text{S.E.}(\hat{p}_A - \hat{p}_B) = \sqrt{\frac{p_A + p_B - (p_A - p_B)^2}{n}},$$

depends on the sample size n and the two probabilities p_A and p_B.

For confidence estimation the estimated standard error of $\hat{p}_A - \hat{p}_B$ is

$$\widehat{\text{S.E.}}(\hat{p}_A - \hat{p}_B) = \sqrt{\frac{\hat{p}_A + \hat{p}_B - (\hat{p}_A - \hat{p}_B)^2}{n}},$$

and the 95% **margin of error of** $\hat{p}_A - \hat{p}_B$ is

$$\text{M.E.}(\hat{p}_A - \hat{p}_B) = 1.96\widehat{\text{S.E.}}(\hat{p}_A - \hat{p}_B).$$

When computing the P–value for a hypothesis test we will assume that $p_A = p_B$. Under this assumption the population standard error of $\hat{p}_A - \hat{p}_B$ simplifies to

$$\text{S.E.}(\hat{p}_A - \hat{p}_B) = \sqrt{\frac{p_A + p_B}{n}}.$$

Thus for hypothesis testing the estimated standard error of $\hat{p}_A - \hat{p}_B$ is

$$\widehat{\text{S.E.}}(\hat{p}_A - \hat{p}_B) = \sqrt{\frac{\hat{p}_A + \hat{p}_B}{n}}.$$

The Z–statistic for this situation is

$$Z_{calc} = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{(\hat{p}_A + \hat{p}_B)/n}} = \frac{n_A - n_B}{\sqrt{n_A + n_B}},$$

where n_A and n_B are the respective frequencies of categories A and B. As in the previous application, this test statistic only depends on the frequencies n_A and n_B, it does not depend on the sample size n.

Example. Opinions about a change in tax law (revisited). Recall that a simple random sample of 100 taxpayers with telephones was selected and each taxpayer was asked "Do you favor or oppose the proposed change in state tax law?". For this population of taxpayers let p_A denote the proportion who would respond "favor", let p_B denote the proportion who would respond "oppose", and let p_C denote the proportion who would respond "no opinion". When we first looked at this example we considered two ways to dichotomize this population so that we could use inferential methods for a single proportion p. First we considered "favor" versus "not favor" for the entire population and inference

about $p = p_A$ (with $1 - p = p_B + p_C$). Then we considered "favor" versus "oppose" for the subpopulation of taxpayers who had an opinion and inference about $p = p_A/(p_A + p_B)$ (with $1 - p = p_B/(p_A + p_B)$). We now have methods for making inferences about the difference $p_A - p_B$ without restricting the population.

As before, suppose that $n = 100$, 64 taxpayers favor the change, 26 taxpayers oppose the change, and 10 taxpayers have no opinion. For this sample we have $\hat{p}_A = .64$ and $\hat{p}_B = .26$, which gives the estimated standard error for estimation of

$$\widehat{\text{S.E.}}(\hat{p}_A - \hat{p}_B) = \sqrt{\frac{.64 + .26 - (.64 - .26)^2}{100}} = .0869$$

and a 95% margin of error of .1703. Therefore, we are 95% confident that $p_A - p_B$ is between $.38 - .1703 = .2097$ and $.38 + .1703 = .5503$. In other words, we are 95% confident that the actual proportion of taxpayers in this metropolitan area (who have telephones) who favored the proposed tax law change at the time of this poll is between 20.97 and 55.03 percentage points higher than the corresponding proportion who opposed the change.

To demonstrate a hypothesis test consider the research hypothesis $H_1 : p_A > p_B$ that a larger proportion of all the taxpayers in this metropolitan area (who have telephones) favored the proposed tax law change at the time of this poll than opposed the change. The estimated standard error for testing is

$$\widehat{\text{S.E.}}(\hat{p}_A - \hat{p}_B) = \sqrt{\frac{.64 + .26}{100}} = .09487,$$

giving $Z_{calc} = .38/.09487 = 4.0055$ with P–value $= P(Z \geq 4.0055) = 3.1 \times 10^{-5}$. Thus there is very strong evidence that a larger proportion of all the taxpayers in this metropolitan area (who have telephones) favored the proposed tax law change at the time of this poll than opposed the change.

Remark. *This hypothesis test is actually equivalent to the conditional test for the research hypothesis $H_1 : p > .5$ for $p = p_A/(p_A + p_B)$ when the population and sample are restricted to the subpopulation and subsample of taxpayers who had an opinion at the time of the poll. That is, if we compute Z_{calc} and the P–value for $n = 90$ and $\hat{p} = 64/90 = .7111$ we get $Z_{calc} = 4.0055$ and P–value $= 3.1 \times 10^{-5}$.*

6.5 Summary

In this chapter we considered the use of the observed difference between two proportions $\hat{p}_1 - \hat{p}_2$ to make inferences about the corresponding population difference $p_1 - p_2$. First we considered the case when the estimates $\hat{p}_1$ and $\hat{p}_2$ are independent. In this case we assumed that $\hat{p}_1$ and $\hat{p}_2$ were computed from independent random samples. Then we

considered the case when the estimates $\hat{p}_1$ and $\hat{p}_2$ are dependent. In this case we considered two situations. First we assumed that $\hat{p}_1$ and $\hat{p}_2$ were computed from a single random sample of paired observations and then we assumed that $\hat{p}_1$ and $\hat{p}_2$ were computed from a single random sample from a population of units with three or more possible categorical values.

The confidence interval estimates of $p_1 - p_2$ and formal tests of hypotheses about $p_1 - p_2$ are based on a normal approximation to the sampling distribution of $\hat{p}_1 - \hat{p}_2$ and require certain assumptions about the random samples. Strictly speaking, the inferential methods discussed in this chapter are not appropriate unless these assumptions are valid.

Independent estimates

For the independent estimates case the requisite assumptions are that the data consist of two independent simple random samples selected with replacement or equivalently two independent sequences of Bernoulli trials. The assumption of independent random samples is very important. We also noted that this approximation works well for independent simple random samples selected without replacement provided both of the populations being sampled are very large. The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is the theoretical probability distribution of $\hat{p}_1 - \hat{p}_2$ which indicates how $\hat{p}_1 - \hat{p}_2$ behaves as an estimator of $p_1 - p_2$. Under the assumptions described above, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ indicates that $\hat{p}_1 - \hat{p}_2$ is unbiased as an estimator of $p_1 - p_2$ ($\hat{p}_1 - \hat{p}_2$ neither consistently overestimates $p_1 - p_2$ nor consistently underestimates $p_1 - p_2$) and provides a measure of the variability in $\hat{p}_1 - \hat{p}_2$ as an estimator of $p_1 - p_2$ (the population standard error of $\hat{p}_1 - \hat{p}_2$, S.E.$(\hat{p}_1 - \hat{p}_2) = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$). The normal approximation allows us to compute probabilities concerning $\hat{p}_1 - \hat{p}_2$ by re–expressing these probabilities in terms of the standardized variable $Z = [(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)]/\text{S.E.}(\hat{p}_1 - \hat{p}_2)$ and using the standard normal distribution to compute the probabilities.

A 95% confidence interval estimate of $p_1 - p_2$ is an interval of plausible values for $p_1 - p_2$ constructed using a method which guarantees that 95% of such intervals will actually contain the unknown difference $p_1 - p_2$ between the population proportions. In the present context a confidence interval for $p_1 - p_2$ may include only negative numbers, only positive numbers, or a mixture of negative and positive numbers, since we are estimating a difference. The 95% confidence interval estimate of $p_1 - p_2$ is formed by adding and subtracting the appropriate margin of error to an estimate of $p_1 - p_2$. The estimate is $\hat{p}_1 - \hat{p}_2$ and the margin of error M.E.$(\hat{p}_1 - \hat{p}_2)$ used to form the 95% confidence interval is

$$\text{M.E.}(\hat{p}_1 - \hat{p}_2) = 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

The 95% confidence interval is the interval from $(\hat{p}_1 - \hat{p}_2) - \text{M.E.}(\hat{p}_1 - \hat{p}_2)$ to $(\hat{p}_1 - \hat{p}_2) + \text{M.E.}(\hat{p}_1 - \hat{p}_2)$. Notice that the margin of error of $\hat{p}_1 - \hat{p}_2$ is a constant multiple (the multiplier is 1.96) of the estimated standard error of $\hat{p}_1 - \hat{p}_2$,

$$\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

We also discussed formal hypothesis tests to compare two competing, complementary hypotheses (the null hypothesis H_0 and the research or alternative hypothesis H_1) about $p_1 - p_2$. Recall that a hypothesis test begins by tentatively assuming that H_0 is true and examining the evidence, which is quantified by the appropriate P–value, against H_0 and in favor of H_1. Since the P–value quantifies evidence against H_0 and in favor of H_1, a small P–value constitutes evidence in favor of H_1. Guidelines for interpreting a P–value are given on page 99.

If there is sufficient *a priori* information to specify a directional hypothesis of the form $H_1 : p_1 - p_2 > 0$ or $H_1 : p_1 - p_2 < 0$, then we can perform a hypothesis test to address the respective questions "Is there sufficient evidence to conclude that $p_1 > p_2$ $(p_1 - p_2 > 0)$?" or "Is there sufficient evidence to conclude that $p_1 < p_2$ $(p_1 - p_2 < 0)$?" The null hypotheses for these research hypotheses are their negations $H_0 : p_1 \leq p_2$ and $H_0 : p_1 \geq p_2$, respectively. The hypothesis test proceeds by tentatively assuming that the null hypothesis H_0 is true and checking to see if there is sufficient evidence (a small enough P–value) to reject this tentative assumption in favor of the research hypothesis H_1. The P–values for these directional hypothesis tests are based on the observed value of the Z–statistic

$$Z_{calc} = \frac{\hat{p}_1 - \hat{p}_2}{\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2)},$$

where, in this testing context, the estimated standard error is

$$\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

with $\hat{p}$ denoting the proportion of successes in the combined sample of all $n_1 + n_2$ units. For $H_1 : p_1 > p_2$ large values of $\hat{p}_1 - \hat{p}_2$, relative to zero, favor H_1 over H_0 and the P–value is the probability that $Z \geq Z_{calc}$. For $H_1 : p_1 < p_2$ we look for small values of $\hat{p}_1 - \hat{p}_2$, relative to zero, and the P–value is the probability that $Z \leq Z_{calc}$.

For situations where there is not enough *a priori* information to specify a directional hypothesis we considered a hypothesis test for the null hypothesis $H_0 : p_1 = p_2$ versus the alternative hypothesis $H_1 : p_1 \neq p_2$. Again we tentatively assume that H_0 is true and check to see if there is sufficient evidence (a small enough P–value) to reject this tentative

assumption in favor of H_1. In this situation the hypothesis test addresses the question "Are the data consistent with $p_1 = p_2$ or is there sufficient evidence to conclude that $p_1 \neq p_2$?" For this non–directional hypothesis test we take the absolute value when computing the observed value of the Z–statistic

$$Z_{calc} = \frac{|\hat{p}_1 - \hat{p}_2|}{\widehat{\text{S.E.}}(\hat{p}_1 - \hat{p}_2)},$$

since values of $\hat{p}_1 - \hat{p}_2$ which are far away from zero in either direction support $p_1 \neq p_2$ over $p_1 = p_2$. Thus the P–value for this hypothesis test is the probability that $|Z| \geq Z_{calc}$.

For all of these hypothesis tests, the P–value is computed under the assumption that H_0 is true. The P–value is the probability of observing a value of $\hat{p}_1 - \hat{p}_2$ that is as extreme or more extreme, relative to zero, than the value we actually observed, under the assumption that H_0 is true (in particular $\hat{p}_1 = \hat{p}_2$). In this statement the definition of extreme (large, small, or far from in either direction) depends on the form of H_1.

Dependent estimates

For the dependent estimates case we first considered the situation when the data consist of a single simple random sample of success/failure pairs (or equivalently success/failure pairs corresponding to paired sequences of Bernoulli trials). We then considered the situation when the data consist of a simple random sample selected with replacement from a population of units with three or more possible categorical values. The details are given in Section 6.4.

6.6 Exercises

Provide a complete analysis for each of the following examples. Be sure to: define relevant population proportions p_1 and p_2; setup and perform a relevant hypothesis test; and, find a confidence interval for $p_1 - p_2$. Provide a complete summary of your findings in the context of the example.

1. Childers and Ferrell (1979) (*Journal of Marketing Research*, **16**, 429–431) conducted a study to investigate the effects of the format of a survey on the response rate for mailed questionnaires. In this context the response rate is the probability that a recipient will return the completed questionnaire. They created two forms of a questionnaire one with questions on both sides of a single sheet of paper and one with questions on one side of each of two sheets of paper. Before reading the remainder of this example answer the following question. Which of these two formats do you believe would result in a higher response rate and why do you believe this? Childers and Ferrell randomly divided a sample of 440 members of the American Marketing Association into two groups of 220.

Of the 220 people who were sent the one sheet version of the questionnaire, 79 returned the questionnaire. Of the 220 people sent the two sheet version of the questionnaire, 66 returned the questionnaire.

2. This example is based on a study of D.M. Barnes (1988), *Science*, **241**, 1029–1030, as described in Moore (1995). This study was conducted to compare two antidepressants as treatments for cocaine addiction. In particular, the researchers wanted to compare the effects of the antidepressant desipramine with the effects of lithium (a standard treatment for cocaine addiction). A group of 48 chronic cocaine users was randomly divided into two groups of 24. One group was treated with desipramine and the other was treated with lithium. The subjects were tracked for three years and the number of subjects who relapsed into cocaine use during this period was recorded. Ten of the 24 people in the desipramine group relapsed into cocaine use and 18 of the 24 people in the lithium group relapsed.

Chapter 7
Inference for a Mean or Median

7.1 Introduction

There are many situations when we might wish to make inferences about the location of the "center" of the population distribution of a quantitative variable. We will consider methods for making inferences about a population mean or a population median, which are the two most commonly used measures of the center of the distribution of a quantitative variable, in this chapter.

In Chapter 5 we considered inferences about the distribution of a dichotomous variable. Since the distribution of a dichotomous variable is completely determined by the corresponding population success proportion, we found that the sampling distribution of the sample proportion $\hat{p}$ was determined by the sampling method. In general, the distribution of a quantitative variable is not completely determined by a single parameter. Therefore, before we can make inferences about the distribution of a quantitative variable we need to make some assumptions about the distribution of the variable.

A probability model for the distribution of a discrete variable X is a theoretical relative frequency distribution which specifies the probabilities (theoretical relative frequencies) with which each of the possible values of X will occur. In contrast to a relative frequency distribution, which indicates the relative frequencies with which the possible values of X occur in a sample, a probability model or probability distribution specifies the probabilities with which the possible values of X will occur when we observe a single value of X. That is, if we imagine choosing a single value of X at random from all of the possible values of X, then the probability model specifies the probability with which each possible value will be observed. We can represent a discrete probability distribution graphically via a probability histogram (theoretical relative frequency histogram) which is simply a histogram based on the probabilities specified by the probability model.

A probability model for the distribution of a continuous variable X can be represented by a density curve. A **density curve** is a nonnegative curve for which the area under the curve (over the x–axis) is one. We can think of the density curve as a smooth version of a probability histogram with the rectangles of the histogram replaced by a smooth curve indicating where the tops of the rectangles would be. With a continuous variable X it does not make sense to talk about the probability that X would take on a particular value, after all if we defined positive probabilities for the infinite collection (continuum) of possible values of X these probabilities could not add up to one. It does, however, make sense to talk about the probability that X will take on a value in a specified interval or range of values. Given two constants $a < b$ the probability that X takes on a value in the

interval from a to b, denoted by $P(a \leq X \leq b)$, is equal to the area under the density curve over the interval from a to b on the x–axis. Areas of this sort based on the density curve give the probabilities which a single value of X, chosen at random from the infinite population of possible values of X, will satisfy.

Given a probability model for the distribution of a continuous variable X, *i.e.*, given a density curve for the distribution of the continuous variable X, we can define population parameters which characterize relevant aspects of the distribution. For example, we can define the population mean μ as the balance point of the unit mass bounded by the density curve and the number line. We can also think of the population mean as the weighted average of the infinite collection of possible values of X with weights determined by the density curve. We can similarly define the population median M as the point on the number line where a vertical line would divide the area under the density curve into two equal areas (each of size one–half).

7.2 Inference for a population mean

7.2a Introduction

In this section we will consider inference for the mean μ of the population distribution of a continuous variable X. The basic problem we will consider is that of using a random sample of values of the continuous variable X to estimate the corresponding population mean or to test a hypothesis about this population mean.

Before we go further it is instructive to consider some situations where inference about a population mean could be used and the way in which we might interpret a population mean.

In some applications the population mean represents an actual physical constant. Let μ denote the true value of the physical constant (such as the speed of light) which we wish to estimate. Suppose that an experiment has been devised to produce a measurement X of the physical constant μ. A probability model for the distribution of X provides a model for the behavior of an observed value of X by specifying probabilities which X must satisfy. Thus, the probability model provides an explanation of the variability in X as an estimate of μ. It would be unreasonable to expect an observed value of X to be exactly equal to μ; however, if the experiment is carefully planned and executed it would be reasonable to expect the average value of X based on a long series of replications of the experiment to be equal to μ. If this is the case, the population mean of the probability model for X will be equal to the physical constant μ and the standard deviation σ of the probability model will serve as a useful quantification of the variability of the measurement process.

When interest centers on an actual, physical population of units the population mean is the average value of the variable of interest corresponding to all of the units in the

population. Imagine a large population of units, *e.g.*, a population of humans or animals. Let the continuous variable X denote a characteristic of a unit, *e.g.*, X might be some measurement of the size of a unit. For concreteness, let X denote the height of an adult human male and consider the population of all adult human males in the United Kingdom. A probability model for the distribution of X provides a model for the behavior of an observed value of the height X of an adult male selected at random from this population. In this situation a probability model explains the variability among the heights of the adult males in this population. Let μ denote the population mean height of all adult males in the United Kingdom, *i.e.*, let μ be the average height we would get if we averaged the heights of all the adult males in this population. In this context the population mean height is obviously not the "true height" of each of the adult males; however, we can think of the height X of a particular adult male as being equal to the population mean height μ plus or minus an adjustment for this particular male which is due to hereditary, environmental, and other factors. The standard deviation σ of the probability model serves to quantify the variability among this population of heights.

In many applications interest centers on the population mean difference between two paired values. For example, consider a population of individuals with high cholesterol levels and a drug designed to reduce cholesterol levels. Let X_1 denote the cholesterol level of an individual before taking the drug, let X_2 denote this same individual's cholesterol level after being treated with the drug, and let $D = X_1 - X_2$ denote the difference between the two cholesterol levels (the decrease in cholesterol level). A probability model for the distribution of D provides a model for the behavior of an observed value of the difference D for an individual selected at random from this population. In this situation a probability model explains the variability among the differences in cholesterol level due to treatment with the drug for the individuals in this population. The corresponding population mean difference μ is the average difference (decrease) in cholesterol level that we would observe if all of the individuals in this population were treated with this drug. The standard deviation σ of the probability model serves to quantify the variability among the differences for the individuals in this population.

We can envision a probability model for the distribution of a quantitative variable X in terms of a box model. If X has a finite number of possible values, then a probability model specifies the probabilities with which these possible values will occur. If the balls in a box are labeled with the possible values of X and the proportion of balls with each label (value of X) in the box is equal to the probability for that value specified by the probability model, then, according to the probability model, observing a value of X is equivalent to selecting a single ball at random from this box and observing the label on the ball. For a continuous variable X observing the value of X is like selecting a ball at random from a box containing an infinite collection of suitably labeled balls.

Given a probability model for the distribution of X, a collection of n values of X is said to form a random sample of size n if it satisfies the two properties given below.

1. Each value of the variable X that we observe can be viewed as a single value chosen at random from a (usually infinite) population of values which are distributed according to the specified probability model for the distribution of X.
2. The observed values of X are independent. That is, knowing the value of one or more of the observed values of X has no effect on the probabilities associated with other values of X.

In other words, in terms of the box model for the probability distribution of X a random sample of n values of X can be viewed as a collection of n labels corresponding to a simple random sample selected with replacement from a box of suitably labeled balls.

Given a random sample of values of X it seems obvious that the sample mean $\overline{X}$ is an appropriate estimate of the corresponding population mean μ. The sampling distribution of $\overline{X}$, which describes the sample to sample variability in $\overline{X}$, serves as the starting point for our study of the behavior of the sample mean $\overline{X}$ as an estimator of the population mean μ. The exact form of the sampling distribution of $\overline{X}$ depends on the form of the distribution of X. However, the two important properties of the sampling distribution of the sample mean given below are valid regardless of the exact form of the distribution of X.

Let $\overline{X}$ denote the sample mean of a random sample of size n from a population (distribution) with population mean μ and population standard deviation σ. The sampling distribution of the sample mean $\overline{X}$ has the following characteristics.

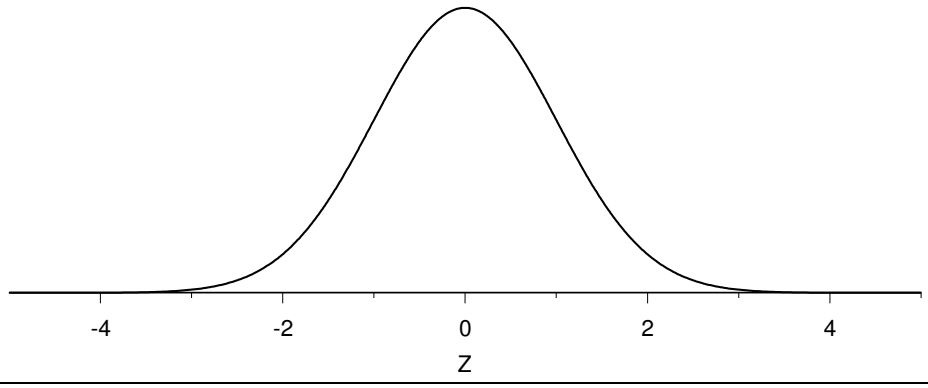
1. The mean of the sampling distribution of $\overline{X}$ is the corresponding population mean μ. This indicates that the sample mean $\overline{X}$ is unbiased as an estimator of the population mean μ. Recall that saying that a statistic is unbiased means that, even though the statistic will overestimate the parameter for some samples and will underestimate the parameter in other samples, it will do so in such a way that, in the long run, the values of the statistic will average to give the correct value of the parameter.
2. The population standard error of the sample mean $\overline{X}$ (the standard deviation of the sampling distribution of $\overline{X}$) is $\text{S.E.}(\overline{X}) = \sigma/\sqrt{n}$. That is, the standard deviation of the sampling distribution of $\overline{X}$ is equal to the standard deviation of the distribution of X divided by the square root of the sample size. Notice that this implies that the sample mean is less variable than a single observation as an estimator of μ; and that if μ and σ are held constant, then the variability in $\overline{X}$ as an estimator of μ decreases as the sample size increases reflecting the fact that a larger sample provides more information than a smaller sample.

Since the form of the sampling distribution of $\overline{X}$ depends on the form of the distribution of X, we will need to make some assumptions about the distribution of X before we can proceed with our discussion of inference for the population mean μ. These assumptions correspond to the choice of a probability model (density curve) to represent the distribution of X. There is an infinite collection of probability models to choose from but we will restrict our attention to a single probability model, the normal probability model, which is appropriate for many situations when the distribution of X is symmetric and mound shaped. This does not imply that all, or even most, distributions of continuous variables are normal distributions. Some of the reasons that we will use the normal distribution as a probability model are: (1) the theory needed for inference has been worked out for the normal model; (2) there are many situations where a normal distribution provides a reasonable model for the distribution of a quantitative variable; (3) even though the inferential methods we discuss are based on the assumption that the distribution of the variable is exactly a normal distribution, it is known that these inferential methods actually perform reasonably well provided the true distribution of the variable is "reasonably similar to a normal distribution"; and, (4) it is often possible to transform or redefine a variable so that its distribution is reasonably modeled by a normal distribution.

7.2b The normal distribution

The normal distribution with mean μ and standard deviation σ can be characterized by its density curve. The density curve for the normal distribution with mean μ and standard deviation σ is the familiar bell shaped curve. The standard normal density curve, which has mean $\mu = 0$ and standard deviation $\sigma = 1$, is shown in Figure 1.

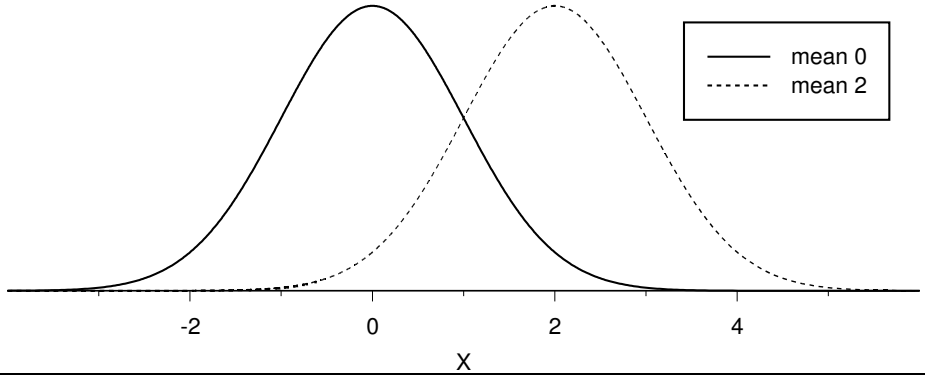
Figure 1. The standard normal density curve.



The normal distribution with mean μ and its density curve are symmetric around μ, *i.e.*, if we draw a vertical line through μ, then the two sides of the density curve are mirror images of each other. Therefore the mean of a normal distribution μ is also the median of the normal distribution. The mean μ locates the normal distribution on the number line so

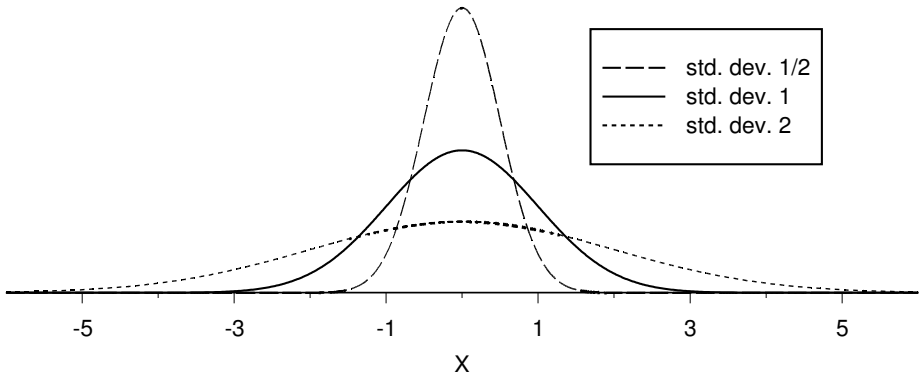
that if we hold σ constant and change the mean μ, the normal distribution is simply shifted along the number line until it is centered at the new mean. In other words, holding σ fixed and changing μ simply relocates the density curve on the number line; it has no effect on the shape of the curve. Figure 2 provides the density curves for normal distributions with respective means $\mu = 0$ and $\mu = 2$ and common standard deviation $\sigma = 1$.

Figure 2. Normal distributions with common standard deviation one and means of zero and two.



The standard deviation σ indicates the amount of variability in the normal distribution. If we hold μ fixed and increase the value of σ, then the normal density curve becomes flatter, while retaining its bell–shape, indicating that there is more variability in the distribution. Similarly, if we hold μ fixed and decrease the value of σ, then the normal density curve becomes more peaked around the mean μ, while retaining its bell–shape, indicating that there is less variability in the distribution. Normal distributions with mean $\mu = 0$ and respective standard deviations $\sigma = .5$, $\sigma = 1$, and $\sigma = 2$ are plotted in Figure 3.

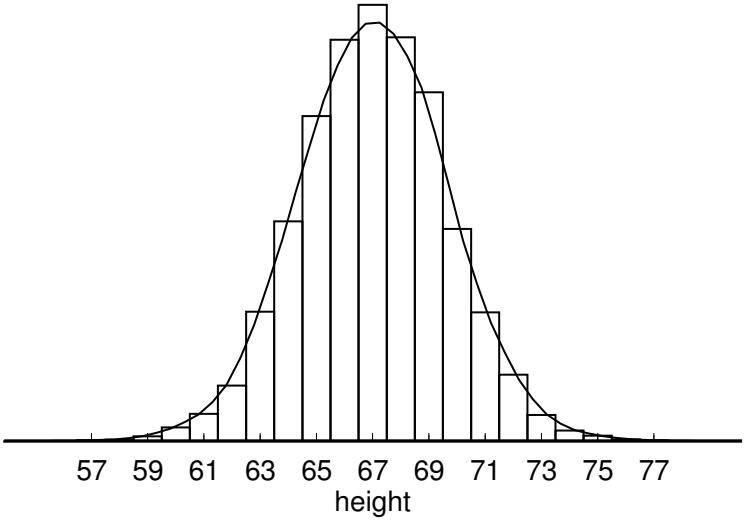
Figure 3. Normal distributions with common mean zero and standard deviations one–half, one, and two.



Example. Heights of adult males. The 8585 heights (in inches) of adult males born in the United Kingdom (including the whole of Ireland) which are summarized in

Table 8 of Section 3.3 provide a good illustration of the fact that normal distributions often provide very good models for populations of physical measurements, such as heights or weights, of individuals. Figure 4 provides a histogram for this height distribution and the density curve for a normal distribution chosen to model these data. You can see that the normal distribution provides a very reasonable model for the heights of adult males born in the United Kingdom.

Figure 4. Histogram and normal density curve for the UK height example.



Computer programs and many calculators can be used to compute normal probabilities or equivalently to compute areas under the normal density curve. These probabilities can also be calculated using tables of standard normal distribution probabilities such as Table 1. Recall that the standard normal distribution is the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The relationship between the standard normal variable Z and the normal variable X, which has mean μ and standard deviation σ, is

$$Z = \frac{X - \mu}{\sigma} \quad \text{or equivalently} \quad X = \mu + Z\sigma.$$

This relationship implies that a probability statement about the normal variable X can be re–expressed as a probability statement about the standard normal variable Z by re–expressing the statement in terms of standard deviation units from the mean. Given two constants $a < b$, observing a value of X between a and b (observing $a \leq X \leq b$) is equivalent to observing a value of $Z = (X - \mu)/\sigma$ between $(a - \mu)/\sigma$ and $(b - \mu)/\sigma$ (observing $(a - \mu)/\sigma \leq (X - \mu)/\sigma \leq (b - \mu)/\sigma$). Furthermore, $Z = (X - \mu)/\sigma$ behaves in accordance with the standard normal distribution so that the probability of observing a value of X between a and b, denoted by $P(a \leq X \leq b)$, is equal to the probability that the standard normal variable Z takes on a value between $(a - \mu)/\sigma$ and $(b - \mu)/\sigma$, *i.e.*,

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right).$$

In terms of areas this probability equality says that the area under the normal density curve with mean μ and standard deviation σ over the interval from a to b is equal to the area under the standard normal density curve over the interval from $(a-\mu)/\sigma$ to $(b-\mu)/\sigma$. Similarly, given constants $c < d$, we have the analogous result that

$$P(c \le Z \le d) = P(\mu + c\sigma \le X \le \mu + d\sigma).$$

Table 1 provides cumulative standard normal probabilities of the form $P(Z \le a)$ for values of a (Z in the table) between 0 and 3.69. Computer programs usually produce cumulative probabilities like these. To use these cumulative probabilities to compute a probability of the form $P(a \le Z \le b)$ note that

$$P(a \le Z \le b) = P(Z \le b) - P(Z \le a)$$

and note that the symmetry of the normal distribution implies that

$$P(Z \le -a) = P(Z \ge a) = 1 - P(Z \le a).$$

Calculators will usually provide probabilities of the form $P(a \le Z \le b)$ directly.

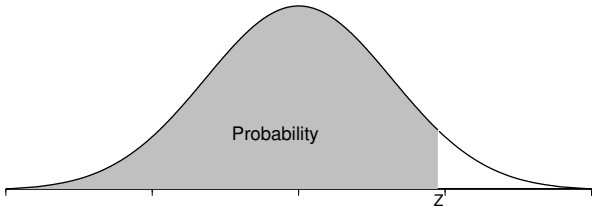


Table 1. Cumulative normal probabilities.

(Areas under the standard normal curve to the left of Z.)

Z	Second decimal place in Z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389

continued on next page

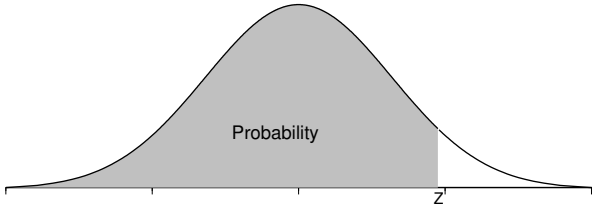


Table 1 (continuation). Cumulative normal probabilities.
(Areas under the standard normal curve to the left of Z.)

Z	Second decimal place in Z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

7.2c Sampling from a normal population

Strictly speaking, the inferential methods based on the Student's t distribution described in Sections 7.2d and 7.2e are only appropriate when the data constitute a random sample from a normal population. However, these methods are known to be generally reasonable even when the underlying population is not exactly a normal population, provided the underlying population distribution is reasonably symmetric and the true density curve has a more or less normal (bell–shaped) appearance. We cannot be sure that an underlying population is normal; however, we can use descriptive methods to look for evidence of possible nonnormality, provided the sample size is reasonably large. The most easily detected and serious evidence of nonnormality you should look for is evidence of extreme skewness or evidence of extreme outlying observations. If there is evidence of extreme skewness or extreme outlying observations, then the inferential methods based on the Student's t distribution should not be used. An alternate approach to inference (for a population median) which may be used when the Student's t methods are inappropriate is discussed in Section 7.3.

Figure 5. Stem and leaf histograms for eight random samples of size 10 from a standard normal distribution.

In these stem and leaf histograms the decimal point is between the stem and the leaves.

(A)		**(B)**		**(C)**		**(D)**	
-2	41	-2	4	-2	1	-2	
-1		-1	3	-1	55	-1	20
-0	20	-0	51	-0	853	-0	6442
0	345569	0	33	0	28	0	145
1		1	112	1	14	1	4
2		2	0	2		2	
(E)		**(F)**		**(G)**		**(H)**	
-2	62	-2		-2		-2	
-1		-1	3	-1		-1	63
-0	42	-0	51	-0	721	-0	77654
0	66	0	11234	0	69	0	68
1	34	1	39	1	00357	1	2
2	24	2		2		2	

Table 2. Five number summaries for the eight random samples of size 10 from Figure 5.

	(A)	**(B)**	**(C)**	**(D)**	**(E)**	**(F)**	**(G)**	**(H)**
min	-2.38	-2.38	-2.14	-1.21	-2.57	-1.29	-.65	-1.62
Q_1	-.13	-.37	-1.29	-.58	-.38	-.08	.08	-.71
median	.34	.30	-.43	-.28	.57	.14	.95	-.56
Q_3	.54	1.14	.67	.32	1.40	.33	1.25	.36
max	.89	2.01	1.35	1.44	2.40	1.93	1.66	1.21

Using a sample to determine whether the underlying population is normal requires some practice. Some indication of the sorts of samples which may arise when the underlying population is normal is provided by the stem and leaf histograms and five number summaries given in Figures 5 and 6 and Tables 2 and 3 for several computer generated random samples from a standard normal distribution. The eight random samples of size 10 of Figure 5 and Table 2 indicate what may happen when a small sample is taken from a population which is normal. Based on these examples it is clear that we should not necessarily view slight skewness (as in A, B, D, F, G, and H) or mild outliers (as in A and E) as evidence of nonnormality. The eight random samples of size 50 of Figure 6 and Table 3 indicate that with a reasonably large sample we can expect to see a reasonably symmetric distribution; but, we may see a few mild outliers as in A, C, and F.

Figure 6. Stem and leaf histograms for eight random samples of size 50 from a standard normal distribution.

In these stem and leaf histograms the decimal point is between the stem and the leaves.

(A)		**(B)**		**(C)**		**(D)**	
-3		-3		-3	2	-3	
-2		-2		-2		-2	86
-2		-2	10	-2	21	-2	10
-1	85	-1	6	-1	66	-1	998
-1	4332110	-1	3100	-1	4322110	-1	4443222
-0	9876665	-0	8877655	-0	97655	-0	9876655
-0	4442211110	-0	42221000	-0	443221111	-0	4421111
0	113344	0	011222333	0	03344	0	011122334
0	555667778	0	55556689	0	566677789	0	566789
1	0134	1	01223	1	011224	1	022
1	5778	1	579	1	55	1	558
2		2	01	2	00	2	1
2	7	2	7	2		2	
(E)		**(F)**		**(G)**		**(H)**	
-2		-2	5	-2		-2	
-2		-2		-2	0	-2	
-1		-1		-1	7666	-1	9666
-1	4320	-1	333111	-1	222110	-1	33220
-0	9875555555	-0	98876655	-0	98766555	-0	87666
-0	32	-0	443333211	-0	4443221	-0	433322211000
0	01223333444	0	0233	0	122223	0	222233334
0	556678889999	0	566666777789	0	6678899	0	5578889
1	0002344	1	022444	1	011	1	11244
1	5567	1	6	1	555699	1	56
2		2	124	2	14	2	4

Table 3. Five number summaries for the eight random samples of size 50 from Figure 6.

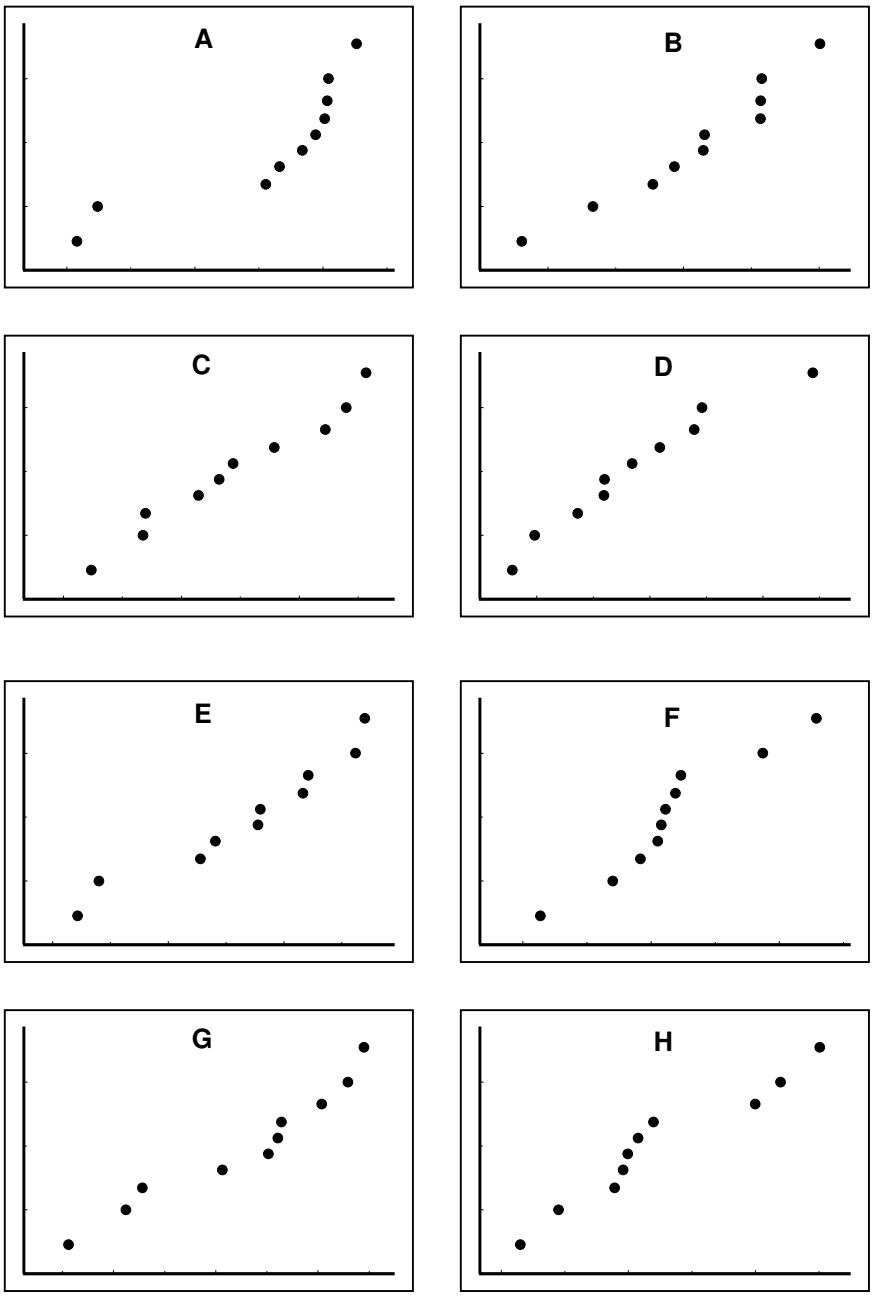
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
min	-1.77	-2.13	-3.20	-2.84	-1.39	-2.47	-1.98	-1.90
Q_1	-.64	-.55	-.91	-1.23	-.48	-.56	-.78	-.59
median	-.06	.11	-.09	-.11	.36	.10	-.13	-.02
Q_3	.67	.78	.72	.54	.92	.71	.87	.67
max	2.75	2.70	1.97	2.10	1.74	2.37	2.44	2.37

Normal probability plots. If you have access to a suitable calculator or computer program, you can construct a normal probability plot as an aid for assessing whether your data are consistent with a sample from a normal distribution. A normal probability plot provides a graphical comparison of the observed sample values with idealized sample values which we might expect to observe if the sample came from a normal distribution. The idealized sample values used in a normal probability plot are known as normal scores. Ideally, we would expect a random sample of size n from a normal distribution to partition the region under the normal density curve into $n + 1$ regions of equal area, with each of these areas being $1/(n + 1)$. The n normal scores, which constitute our idealized random sample, can thus be formed by determining the n values which would partition the area under the normal density curve into $n + 1$ regions each of area $1/(n + 1)$ as suggested above. Once these normal scores are obtained we can plot the ordered normal scores versus the ordered observed data values and examine this normal probability plot looking for evidence of systematic disagreement between the actual sample values and the expected normal scores. If the sample really was a sample from a normal distribution, then we would expect the normal probability plot to approximate a straight line. Therefore, a normal probability plot which differs greatly in appearance from a straight line provides evidence that the sample may not come from a normal distribution.

Figures 7 and 8 provide normal probability plots (normal score versus observed value) for the computer generated random samples from a standard normal distribution of Figures 5 and 6. Some representative examples of normal probability plots for actual data are provided in Figures 9, 10, and 11. (In some of these plots and subsequent normal probability plots the points have been subjected to small random shifts to better indicate points which are coincident or very close together.)

The normal probability plot in Figure 9 is for the height of adult males in the United Kingdom example of Section 7.2b. We noted that the histogram for the distribution of the height of adult males in the United Kingdom given in Figure 4 of this chapter is very well approximated by a normal distribution. The straight line nature of the normal probability plot of Figure 9 indicates that it is quite reasonable to model these heights as forming a random sample from a normal distribution. The normal probability plot in Figure 10 is for the cholesterol levels of the rural Guatemalans from the example in Section 3.1.

Figure 7. Normal probability plots for the for eight random samples of size 10 (from a standard normal distribution) of Figure 5.



The stem and leaf histogram of Figure 1 in Section 3.1 is reasonably symmetric and the fact that the normal probability plot in Figure 10 is reasonably linear indicates that it is reasonable to model the cholesterol levels of these rural Guatemalans as forming a sample from a normal distribution. The normal probability plot in Figure 11 is for the rainfall amounts for the 26 days when the cloud was unseeded in the cloud seeding example of

Section 4.3. The curvature in this normal probability plot indicates that it is not reasonable to model these rainfall amounts as forming a random sample from a normal distribution. The stem and leaf histogram for this example given in Figure 12 shows that this type of curvature in a normal probability plot corresponds to skewness to the right.

Figure 8. Normal probability plots for the for eight random samples of size 50 (from a standard normal distribution) of Figure 6.

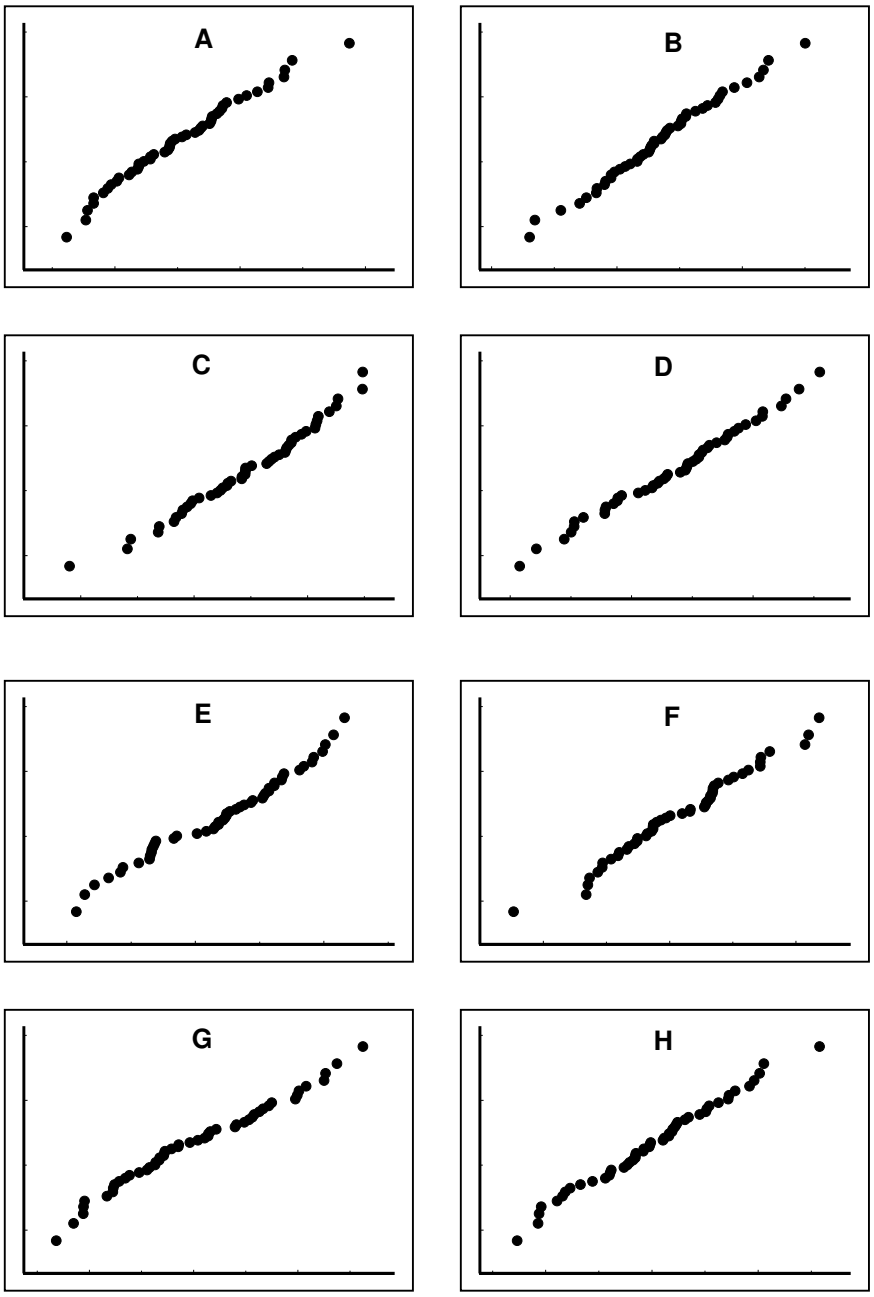


Figure 9. Normal probability plot for the UK height example.

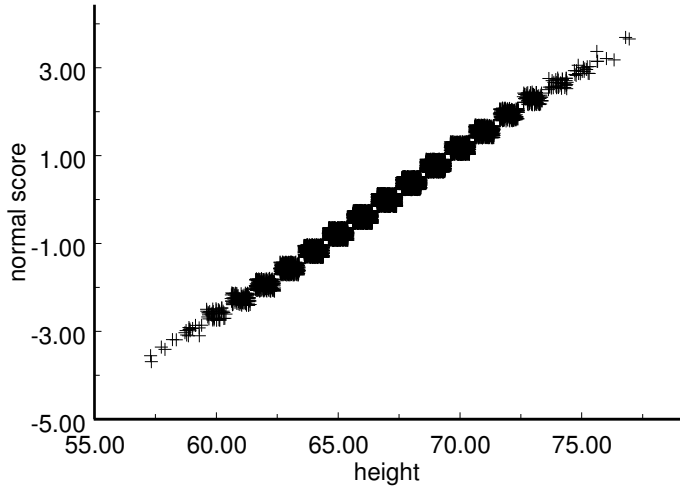


Figure 10. Normal probability plot for rural cholesterol levels.

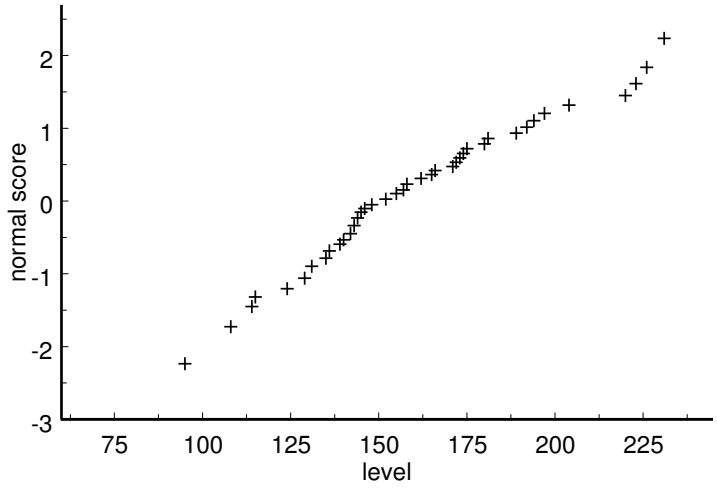


Figure 11. Normal probability plot for unseeded rainfall.

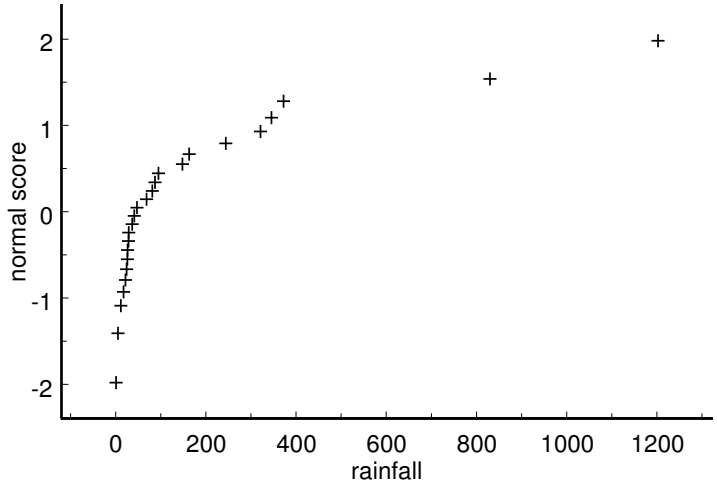


Figure 12. Stem and leaf histogram for unseeded rainfall.

In this stem and leaf histogram the stem represents hundreds and the leaf represents tens. (acrefeet)

stem	leaf
0	000112222223446889
1	56
2	4
3	247
4	
5	
6	
7	
8	3
9	
10	
11	
12	0

7.2d Estimating a normal population mean

The sample mean $\overline{X}$ provides a single number estimate of the population mean μ. We can think of $\overline{X}$ as our "best guess" of the value of μ. As noted above we know that $\overline{X}$ is unbiased as an estimator of μ; therefore, on the average in the long run (under repeated sampling) we know that $\overline{X}$ provides a good estimate of the unknown mean μ. Since this unbiasedness does not guarantee that the observed value of $\overline{X}$, based on a single sample, will be close to the true, unknown value of μ, it would be useful to have a confidence interval estimate of μ.

Recall that when the sample mean corresponds to a random sample from a population distribution with population mean μ and population standard deviation σ the sampling distribution of $\overline{X}$ has mean μ and standard deviation $\text{S.E.}(\overline{X}) = \sigma/\sqrt{n}$ (the population standard error of $\overline{X}$). When the underlying population distribution is normal we can say more about the form of this sampling distribution. If the random sample from which the sample mean $\overline{X}$ is computed is a random sample from a normal population with population mean μ and population standard deviation σ, then the sampling distribution of $\overline{X}$ is a normal distribution with population mean μ and population standard deviation $\text{S.E.}(\overline{X}) = \sigma/\sqrt{n}$. Thus, under these assumptions, the quantity

$$Z = \frac{\overline{X} - \mu}{\text{S.E.}(\overline{X})} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

behaves in accordance with the standard normal distribution.

We know that a standard normal variable Z will take on a value between -1.96 and 1.96 with probability .95 ($P(-1.96 \leq Z \leq 1.96) = .95$). This implies that

$$P\left(-1.96 \leq \frac{\overline{X} - \mu}{\text{S.E.}(\overline{X})} \leq 1.96\right) = .95$$

which is equivalent to

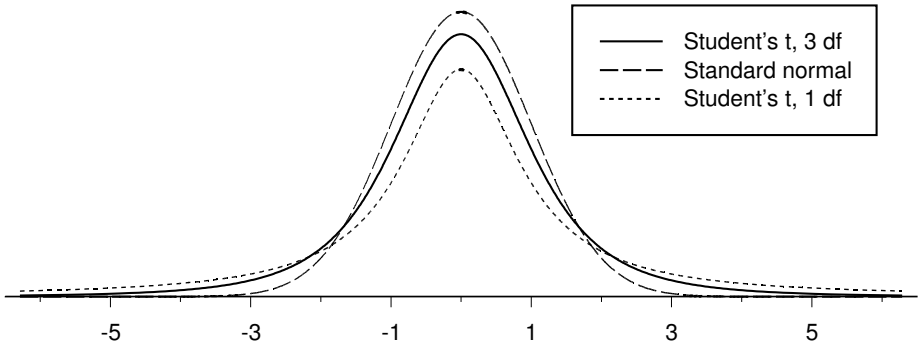
$$P\big(\overline{X} - 1.96\text{S.E.}(\overline{X}) \leq \mu \leq \overline{X} + 1.96\text{S.E.}(\overline{X})\big) = .95.$$

This probability statement says that 95% of the time we will observe a value of $\overline{X}$ such that the population mean μ will be between $\overline{X} - 1.96\text{S.E.}(\overline{X})$ and $\overline{X} + 1.96\text{S.E.}(\overline{X})$. Unfortunately, we cannot use this interval as a confidence interval for μ, since the population standard error $\text{S.E.}(\overline{X}) = \sigma/\sqrt{n}$ depends on the unknown population standard deviation σ and thus is not computable. We can avoid this difficulty by replacing the unknown population standard error $\sigma/\sqrt{n}$ by the sample standard error $\widehat{\text{S.E.}}(\overline{X}) = S/\sqrt{n}$, where S is the sample standard deviation, and basing our confidence interval estimate on the quantity

$$T = \frac{\overline{X} - \mu}{\widehat{\text{S.E.}}(\overline{X})} = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

If the sample mean $\overline{X}$ and the sample standard deviation S are computed from a random sample of size n from a normal population with population mean μ and population standard deviation σ, then the quantity T defined above follows the **Student's t distribution with $n-1$ degrees of freedom**. The Student's t distribution with $n-1$ degrees of freedom is symmetric about zero and has a density curve very similar to that of the standard normal distribution. The main difference between these two distributions is that the Student's t distribution has "heavier" tails than the standard normal distribution. That is, the tails of the Student's t density curve approach the x–axis more slowly than do the tails of the standard normal density curve. As the sample size (and the degrees of freedom) increases the Student's t distribution becomes more similar to the standard normal distribution. In fact, the standard normal distribution is the limiting version of the Student's t distribution in the sense that the Student's t density curve approaches the standard normal density curve when the degrees of freedom increases without bound. The relationship between Student's t distributions and the standard normal distribution is indicated by the plots in Figure 13.

Figure 13. Student's t distributions with 1 and 3 degrees of freedom and standard normal distribution.



Given a constant k such that $P(-k \leq T \leq k) = .95$, where T denotes a Students t variable with $n-1$ degrees of freedom, we have

$$P\left(-k \leq \frac{\overline{X} - \mu}{\widehat{\text{S.E.}}(\overline{X})} \leq k\right) = .95$$

which is equivalent to

$$P\left(\overline{X} - k\widehat{\text{S.E.}}(\overline{X}) \leq \mu \leq \overline{X} + k\widehat{\text{S.E.}}(\overline{X})\right) = .95.$$

The quantity

$$\text{M.E.}(\overline{X}) = k\widehat{\text{S.E.}}(\overline{X}) = \frac{kS}{\sqrt{n}}$$

is the 95% **margin of error** of $\overline{X}$. The preceding probability statement says that 95% of the time we will observe values of $\overline{X}$ and S such that the population mean μ will be between $\overline{X} - \text{M.E.}(\overline{X})$ and $\overline{X} + \text{M.E.}(\overline{X})$. Thus, the interval from $\overline{X} - \text{M.E.}(\overline{X})$ to $\overline{X} + \text{M.E.}(\overline{X})$ is a 95% confidence interval estimate for μ. To compute this confidence interval we need to determine the value of the appropriate margin of error multiplier k. This multiplier depends on the size of the sample so that there is a different multiplier for each sample size. The symmetry of the Student's t distribution and the definition of k imply that k is the 97.5 percentile of the Student's t distribution with $n-1$ degrees of freedom. The 95% margin of error multipliers (k's) based on the Student's t distribution, for several choices of the degrees of freedom (d.f.), are given in Table 4.

This confidence interval estimate may be reported using a statement such as: We are 95% confident that the population mean μ is between $\overline{X} - \text{M.E.}(\overline{X})$ and $\overline{X} + \text{M.E.}(\overline{X})$. Notice that it is the sample mean $\overline{X}$ and the margin of error $\text{M.E.}(\overline{X})$ that vary from sample to sample. The population mean μ is a fixed, unknown parameter that does not

vary. Therefore, the 95% confidence level applies to the method used to generate the confidence interval estimate. That is, the method (obtain a simple random sample and compute the numbers $\overline{X} - \text{M.E.}(\overline{X})$ and $\overline{X} + \text{M.E.}(\overline{X})$ forming the limits of the confidence interval) is such that 95% of the time it will yield a pair of confidence interval limits which bracket the population mean μ.

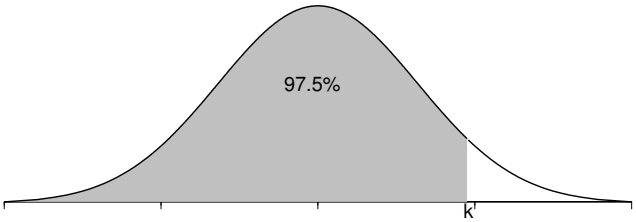


Table 4. 97.5 percentiles of the Student's t distribution.

$d.f.$	k	$d.f.$	k	$d.f.$	k	$d.f.$	k
1	12.706	14	2.145	27	2.052	40	2.021
2	4.303	15	2.131	28	2.048	45	2.014
3	3.182	16	2.120	29	2.045	50	2.008
4	2.776	17	2.110	30	2.042	55	2.004
5	2.571	18	2.101	31	2.040	60	2.000
6	2.447	19	2.093	32	2.037	65	1.997
7	2.365	20	2.086	33	2.034	70	1.994
8	2.306	21	2.080	34	2.032	75	1.992
9	2.262	22	2.074	35	2.030	80	1.989
10	2.228	23	2.069	36	2.028	85	1.988
11	2.201	24	2.064	37	2.026	90	1.986
12	2.179	25	2.060	38	2.024	95	1.985
13	2.160	26	2.056	39	2.023	100	1.982

Example. Newcomb's measurements of the speed of light. In 1882 Simon Newcomb conducted an investigation to measure the speed of light. The essence of Newcomb's experiment was to determine the time it took for light to travel from Fort Myer on the west bank of the Potomac river to a fixed mirror at the foot of the Washington monument 3721 meters away and back. (More details about this and similar examples can be found in Stigler (1977), Do robust estimators work with real data? (with discussion), *Annals of Statistics*, **5**, 1055–1098.) Data from 64 replications of this experiment are provided in Table 5. For ease of handling the times in Table 5 are simplified. The values given in Table 5 are times expressed as billionths of a second in excess of 24.8 millionths of a second, *i.e.*, if a time value from Table 5 is multiplied by 10^{-3} and added to 24.8, the result is the time which Newcomb observed measured in millionths of a second. For example, the first observation in Table 5 is 28 which corresponds to an observed time of 24.828 millionths of a second.

Table 5. Newcomb's time data.

28	22	36	26	28	28	26	24	32	30	27	24	33	21	36	32
31	25	24	25	28	36	27	32	34	30	25	26	26	25	23	21
30	33	29	27	29	28	22	26	27	16	31	29	36	32	28	40
19	37	23	32	29	24	25	27	24	16	29	20	28	27	39	23

Figure 14. Stem and leaf histogram for Newcomb's time data.

In this stem and leaf histogram the stem represents tens of billionths of a second and the leaf represents billionths of a second.

stem	leaf
1	66
1	9
2	011
2	22333
2	4444455555
2	66666777777
2	888888899999
3	00011
3	2222233
3	4
3	66667
3	9
4	0

Figure 15. Normal probability plot for Newcomb's time data.

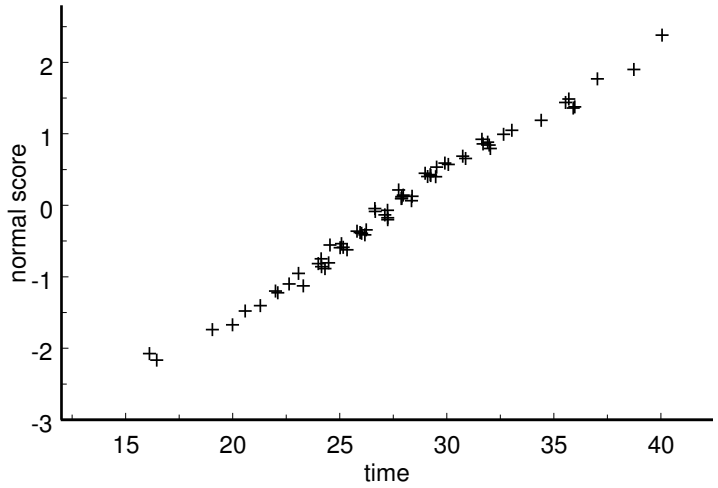


Table 6. Summary statistics for Newcomb's time data.

minimum	16.0	Q_1- minimum	8.5
Q_1	24.5	median $-Q_1$	3.0
median	27.5	Q_3- median	3.5
Q_3	31.0	maximum $-Q_3$	9.0
maximum	40.0		
mean	27.7500	range	24.0
standard deviation	5.0834	IQ range	6.5

Summary statistics for these times are given in Table 6, a stem and leaf histogram is provided in Figure 14, and a normal probability plot is given in Figure 15. Based on the summary statistics, the stem and leaf histogram, and the normal probability plot it seems reasonable to assume that these data form a random sample from a normal population. More specifically, it seems reasonable to assume that these 64 measurements are independent realizations of a normally distributed variable with population mean μ. Actually, the assumption of independence is somewhat questionable, since repetitions of the experiment conducted at nearly the same point in time might exhibit some dependence. We will hope that any such dependence is minor and treat these observations as if they were independent. The population mean μ can be thought of as the "true" time it would take light to make the "trip" as indicated for this experiment if this time could be measured very precisely. We can also view the population mean μ as the long run average of the times which would be obtained if this experiment was repeated over and over. We must keep in mind the fact that Newcomb's method of measurement may introduce a systematic bias which would make the "true" time μ differ from the actual time it would take for light to travel a distance of $7442 = 2 \times 3721$ meters.

The sample mean for these $n = 64$ observations is 27.7500, the sample standard deviation is 5.0834, the standard error of the sample mean is $5.0834/\sqrt{64} = .6354$, and the multiplier for the 95% margin of error, based on the Student's t distribution with $n - 1 = 63$ degrees of freedom, is $k = 1.9983$. Thus the 95% margin of error of the sample mean is $(1.9983)(.6354) = 1.2697$ and the limits for the 95% confidence interval estimate of μ are $27.7500 - 1.2697 = 26.4803$ and $27.7500 + 1.2697 = 29.0197$. Therefore, we are 95% confident that the population mean time for Newcomb's experiment is between 26.4803 and 29.0197 billionths of a second. If we re-express this in terms of the time required for light to travel through the experimental set-up, we find that we are 95% confident that the population mean time measured in this way is between 24.8264803 and 24.8290197 millionths of a second.

For the preceding analysis we have removed two "outliers" from Newcomb's data. The two unusual values are -44 and -2, which correspond to measured times of 24.756 and

24.798 millionths of a second. If these values are added to the stem and leaf histogram of Figure 14, they are clearly inconsistent with the other 64 data values. It seems reasonable to conjecture that something must have gone wrong with the experiment when these unusually small values were obtained. Newcomb chose to omit the observation of -44 and retain the observation of -2. If we consider the 65 observations including the -2, we find that the sample mean is reduced to 27.2923, the sample standard deviation is increased to 6.2493, and the standard error of the mean becomes .7752. Notice that, as we would expect, the presence of this outlier reduces the sample mean (moves the mean toward the outlier) and increases the sample standard deviation (increases the variability in the data). With $n = 65$ observations the multiplier for the 95% margin of error, based on the Student's t distribution with $n - 1 = 64$ degrees of freedom, is $k = 1.9977$, which gives a margin of error of $(1.9977)(.7752) = 1.5486$. Hence, when the unusually small value -2 is included we are 95% confident that the population mean time for Newcomb's experiment is between 25.7437 and 28.8409 billionths of a second. Re-expressing this in terms of the time required for light to travel through the experimental set-up, we find that we are 95% confident that the population mean time measured in this way is between 24.8257437 and 24.8288409 millionths of a second. Including the outlier, -2, has the effect of shifting the confidence interval to the left and making it longer.

Example. Heights of husbands and wives. The data used in this example are part of data set 231 in Hand, Daly, Lunn, McConway, and Ostrowski (1994), *A Handbook of Small Data Sets*, Chapman and Hall, London. The original source is Marsh (1988), *Exploring Data*, Cambridge, UK. A random sample of $n = 169$ married couples was selected from the 1980 OPCS study of the heights and weights of the adult population of Great Britain. The data consist of paired heights (in mm) for these husbands and wives. A few of these paired heights are given in Table 7 and all of the differences (husband's height minus wife's height) are given in Table 8.

Table 7. Husband and wife height data (partial).

couple	husband's height	wife's height	difference
1	1786	1590	196
2	1754	1660	94
3	1755	1590	165
4	1725	1550	175
5	1796	1550	246
...	...	...	...
169	1641	1570	71

Table 8. Husband and wife height differences.

-96	-65	-46	-37	-30	-30	-21	-12	0	0
1	9	13	14	15	19	34	35	35	36
39	40	50	55	56	59	60	60	65	65
66	68	70	70	71	71	73	74	75	75
79	79	81	83	84	84	84	85	85	88
90	94	94	95	96	100	103	103	105	110
110	113	113	114	115	118	119	120	120	123
123	123	123	125	125	125	125	125	128	128
128	129	130	130	130	133	134	135	135	135
135	139	140	141	141	144	144	145	145	145
145	150	151	155	155	155	159	159	160	160
160	160	161	161	164	165	165	165	166	169
170	174	175	175	178	180	181	183	183	188
189	190	190	191	193	194	195	195	195	196
196	196	204	205	209	210	215	219	225	228
228	229	233	235	236	239	241	243	244	246
250	255	258	259	271	276	281	295	303	

Table 9. Summary statistics for height differences.

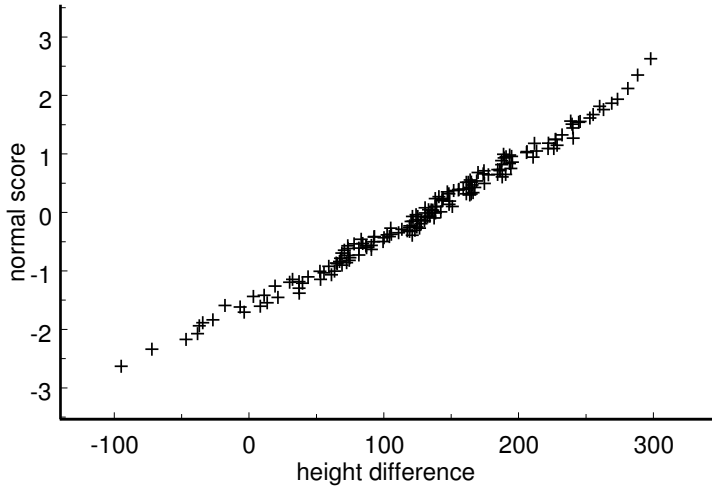
minimum	-96	Q_1- minimum	177
Q_1	81	median $-Q_1$	49
median	130	Q_3- median	51
Q_3	181	maximum $-Q_3$	122
maximum	303		
mean	129.8225	range	399.0
standard deviation	76.0211	IQ range	100.0

Figure 16. Stem and leaf histogram for height differences.

In this stem and leaf histogram the stem represents hundreds and the leaf tens (mm).

stem	leaf
-0	96
-0	4333210
0	0001111333334
0	55556666667777777777888888899999
1	000011111111222222222222222333333333344444444
1	55555556666666666667777788888899999999999
2	000111222233334444
2	55557789
3	0

Figure 17. Normal probability plot for height differences.



Summary statistics for these differences are given in Table 9, a stem and leaf histogram is provided in Figure 16, and a normal probability plot is given in Figure 17. Based on the summary statistics, the stem and leaf histogram, and the normal probability plot it seems reasonable to assume that these differences form a random sample from a normal population of differences. The population mean difference μ_D is the average difference in height corresponding to the population of all married couples in Great Britain in 1980. (Technically, we should restrict this mean to those married couples included in the 1980 COPS study from which the sample was taken.)

The sample mean for these $n = 169$ differences is 129.8225 mm, the sample standard deviation is 76.0211 mm, the standard error of the sample mean is $76.0211/\sqrt{169} = 5.8478$, and the multiplier for the 95% margin of error, based on the Student's t distribution with $n-1 = 168$ degrees of freedom, is $k = 1.9742$. Thus the 95% margin of error of the sample mean is $(1.9742)(5.8478) = 11.5447$ and the limits for the 95% confidence interval estimate of μ_D are $129.8225 - 11.5447 = 118.2778$ and $129.8225 + 11.5447 = 141.3672$. Therefore, we are 95% confident that for the population of all married couples in Great Britain in 1980, on average, the husband's height exceeds the wife's height by at least 118.2778 mm and perhaps as much as 141.3672 mm.

Remark regarding directional confidence bounds. *The use of one of the confidence limits of a 90% confidence interval as a 95% confidence bound discussed in Section 5.4 can also be used in the present context. That is, we can find an upper or lower 95% confidence bound for μ by selecting the appropriate confidence limit from a 90% confidence interval estimate of μ.*

7.2e Tests of hypotheses about a normal population mean

The hypothesis testing procedures discussed in this section are based on the fact that, when $\mu = \mu_0$, the **Student's t test statistic**

$$T_{calc} = \frac{\overline{X} - \mu_0}{\widehat{\text{S.E.}}(\overline{X})} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

follows the Student's t distribution with $n - 1$ degrees of freedom. Recall that, technically, this result requires that the sample from which the sample mean $\overline{X}$ and the sample standard deviation S are computed forms a random sample from a normal distribution with population mean μ. However, these methods are known to be generally reasonable even when the underlying population is not exactly a normal population, provided the underlying population distribution is reasonably symmetric and the true density curve has a more or less normal (bell–shaped) appearance. An alternate approach to inference (for a population median) which may be used when the Student's t methods are inappropriate is discussed in Section 7.3.

Recall that a hypothesis (statistical hypothesis) is a conjecture about the nature of the population. When we considered hypotheses about dichotomous populations we noted that the population was completely determined by the population success proportion p. In the present context of sampling from a normal population two parameters, the mean and the standard deviation, must be specified to completely determine the normal population. A hypothesis about the value of the population mean μ of a normal distribution specifies where the center of this normal distribution, μ, is located on the number line but places no restriction on the population standard deviation.

Even though the logic behind a hypothesis test for a population mean μ is the same as the logic behind a hypothesis test for a population proportion p, we will introduce hypothesis testing for a mean in the context of a simple hypothetical example.

Example. Strength of bricks. Consider a brick manufacturer that has produced a large batch of bricks and wants to determine whether these bricks are suitable for a particular construction project. The specifications for this project require bricks with a mean compressive strength that exceeds 3200 psi. In order to assess the suitability of this batch of bricks the manufacturer will obtain a simple random sample of bricks from this batch and measure their compressive strength. In this example a single brick is a unit, the entire large batch of bricks is the population, and a suitable variable X is the compressive strength of an individual brick (in psi). We will assume that the distribution of X is reasonably modeled by a normal distribution with population mean μ and population standard deviation σ. In this example the population mean μ represents the mean

compressive strength for all of the bricks in this large batch and the population standard deviation quantifies the variability from brick to brick in (measured) compressive strength.

The manufacturer does not want to use these bricks for this construction project unless there is sufficient evidence to claim that the population mean compressive strength for this batch exceeds 3200 psi. Thus, the question of interest here is: "Is there sufficient evidence to justify using this batch of bricks for this project?" In terms of the population mean μ the research hypothesis is $H_1 : \mu > 3200$ (the mean compressive strength for this batch of bricks exceeds 3200 psi); and the null hypothesis is $H_0 : \mu \leq 3200$ (the mean compressive strength for this batch of bricks does not exceed 3200 psi). In other words, the manufacturer will tentatively assume that these bricks are not suitable for this project and will check to see if there is sufficient evidence against this tentative assumption to justify the conclusion that these bricks are suitable for the project.

A test of the null hypothesis $H_0 : \mu \leq 3200$ versus the research hypothesis $H_1 : \mu > 3200$ begins by tentatively assuming that the mean compressive strength for this batch of bricks is no larger than 3200 psi. Under this tentative assumption it would be surprising to observe a sample mean compressive strength $\overline{X}$ that was much larger than 3200. Thus the test should reject $H_0 : \mu \leq 3200$ in favor of $H_1 : \mu > 3200$ if the observed value of $\overline{X}$ is sufficiently large relative to 3200. In order to determine whether $\overline{X}$ is large relative to 3200 we need an estimate of the sample to sample variability in $\overline{X}$. The sample standard error of the sample mean $\widehat{\text{S.E.}}(\overline{X}) = S/\sqrt{n}$, where S denotes the sample standard deviation, provides a suitable measure of the sample to sample variability in $\overline{X}$. Our conclusion will hinge on deciding whether the observed value of the Student's t test statistic

$$T_{calc} = \frac{\overline{X} - 3200}{\widehat{\text{S.E.}}(\overline{X})} = \frac{\overline{X} - 3200}{S/\sqrt{n}}$$

is far enough above zero to make $\mu > 3200$ more tenable than $\mu \leq 3200$. We will base this decision on the probability of observing a value of T as large or larger than the actual calculated value T_{calc} of T, under the assumption that $\mu \leq 3200$. This probability (computed assuming that $\mu = 3200$) is the P–value of the test. We will use the fact that, when $\mu = 3200$, the Student's t statistic T follows the Student's t distribution with $n - 1$ degrees of freedom to calculate the P–value.

First suppose that a simple random sample of $n = 100$ bricks is selected from the large batch. Further suppose that the sample mean compressive strength of these 100 bricks is $\overline{X} = 3481$ psi and the sample standard deviation is $S = 1118.38$. In this case we know that the mean compressive strength of the bricks in the sample $\overline{X} = 3481$ exceeds 3200 and we need to decide whether this suggests that the mean compressive strength of all of the bricks in the batch μ exceeds 3200. In this case the sample standard error of $\overline{X}$ is $\widehat{\text{S.E.}}(\overline{X}) = 1118.38/\sqrt{100} = 111.838$. Thus $\overline{X}$ exceeds 3200 by 281 psi which is

$281/111.838 = 2.5126$ standard error units, *i.e.*, $T_{calc} = 2.5126$. In this case the P–value .0068 is the probability of observing a calculated T value as large or larger than 2.5126 when $\mu \leq 3200$. That is, if $\mu \leq 3200$ and we perform this experiment with $n = 100$, then we would expect to see a value of $\overline{X}$ that is 2.5126 standard error units above the hypothesized value 3200 (2.5126 times $\widehat{\text{S.E.}}(\overline{X})$ psi above 3200) about 0.68% of the time. Therefore, observing values of $\overline{X}$ and S in a sample of size $n = 100$ which yield a value of T_{calc} as large or larger than 2.5126 would be very surprising if the true mean compressive strength for the batch μ was not larger than 3200 and we have sufficient evidence to reject the null hypothesis $H_0 : \mu \leq 3200$ in favor of the research hypothesis $H_1 : \mu > 3200$. In the case we would conclude that there is sufficient evidence to contend that the mean compressive strength for this batch of bricks exceeds 3200 psi, *i.e.*, these bricks are suitable for the construction project.

Now suppose that a simple random sample of $n = 100$ bricks is selected from the large batch and the sample mean compressive strength of these 100 bricks is $\overline{X} = 3481$ psi as before. However, suppose that in this case the sample standard deviation is $S = 2329.3$. As before we know that the mean compressive strength of the bricks in the sample $\overline{X} = 3481$ exceeds 3200 and we need to decide whether this suggests that the mean compressive strength of all of the bricks in the batch μ exceeds 3200. In this case the sample standard error of $\overline{X}$ is $\widehat{\text{S.E.}}(\overline{X}) = 2329.3/\sqrt{100} = 232.93$. Thus $\overline{X}$ exceeds 3200 by 281 psi which is $281/232.93 = 1.2064$ standard error units, *i.e.*, $T_{calc} = 1.2064$. In this case the P–value .1153 is the probability of observing a calculated T value as large or larger than 1.2064 when $\mu \leq 3200$. That is, if $\mu \leq 3200$ and we perform this experiment with $n = 100$, then we would expect to see a value of $\overline{X}$ that is 1.2064 standard error units above the hypothesized value 3200 (1.2064 times $\widehat{\text{S.E.}}(\overline{X})$ psi above 3200) about 11.53% of the time. Therefore, observing values of $\overline{X}$ and S in a sample of size $n = 100$ which yield a value of T_{calc} as large or larger than 1.2064 would not be very surprising if the true mean compressive strength for the batch μ was not larger than 3200 and we do not have sufficient evidence to reject the null hypothesis $H_0 : \mu \leq 3200$ in favor of the research hypothesis $H_1 : \mu > 3200$. In the case we would conclude that there is not sufficient evidence to contend that the mean compressive strength for this batch of bricks exceeds 3200 psi, *i.e.*, these bricks are not suitable for the construction project.

The research hypothesis in the brick example is a directional hypothesis of the form $H_1 : \mu > \mu_0$, where $\mu_0 = 3200$. We will now discuss the details of a hypothesis test for a directional research hypothesis of this form. For the test procedure to be valid the specified value μ_0 and the direction of the research hypothesis must be motivated from subject matter knowledge before looking at the data that are to be used to perform the test.

Let μ_0 denote the hypothesized value which we wish to compare with μ. The research hypothesis states that μ is greater than μ_0; in symbols we will indicate this research hypothesis by writing

$$H_1 : \mu > \mu_0.$$

The null hypothesis is the negation of $H_1 : \mu > \mu_0$ which states that μ is not greater than μ_0; in symbols we will indicate this null hypothesis by writing

$$H_0 : \mu \leq \mu_0.$$

The research hypothesis $H_1 : \mu > \mu_0$ specifies that the normal distribution is one of the normal distributions for which the population mean μ is greater than μ_0. The null hypothesis $H_0 : \mu \leq \mu_0$ specifies that the normal distribution is one of the normal distributions for which the population mean μ is at most μ_0. The population standard deviation σ is not restricted by either hypothesis. Assuming that the population standard deviation is the same regardless of which hypothesis is true, this competing pair of hypotheses provides a decomposition of all possible normal distributions with this σ into the collection of normal distributions where $\mu > \mu_0$ and the research hypothesis is true and the collection of normal distributions where $\mu \leq \mu_0$ and the null hypothesis is true. Our goal is to use the data to decide which of these two collections of normal distributions contains the normal distribution we are actually sampling from.

Since a hypothesis test begins by tentatively assuming that the null hypothesis is true, we need to decide what constitutes evidence against the null hypothesis $H_0 : \mu \leq \mu_0$ and in favor of the research hypothesis $H_1 : \mu > \mu_0$. We will assume that the unknown population standard deviation σ is fixed regardless of the value of μ. The difference $\overline{X} - \mu_0$ between the sample mean $\overline{X}$ and the hypothesized value μ_0, expressed in standard error units, will be used to assess the strength of the evidence in favor of the research hypothesis. Generally, we would expect to observe larger values of $\overline{X}$ more often when the research hypothesis $H_1 : \mu > \mu_0$ is true than when the null hypothesis $H_0 : \mu \leq \mu_0$ is true. In particular, we can view the observation of a value of $\overline{X}$ that is sufficiently large relative to μ_0 as constituting evidence against the null hypothesis $H_0 : \mu \leq \mu_0$ and in favor of the research hypothesis $H_1 : \mu > \mu_0$. To decide whether the observed value of $\overline{X}$ is "sufficiently large relative to μ_0" we need to take the variability in the data into account. We can do this by basing our decision on the corresponding Student's t test statistic,

$$T_{calc} = \frac{\overline{X} - \mu_0}{\widehat{\text{S.E.}}(\overline{X})} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

instead of $\overline{X}$ alone. Notice that values of T_{calc} which are large relative to zero correspond to values of $\overline{X}$ which are large relative to μ_0. Deciding whether the observed value of T_{calc}

is sufficiently large relative to zero to allow rejection of H_0 is based on the corresponding P–value, which is defined below.

The P–value for testing the null hypothesis $H_0 : \mu \leq \mu_0$ versus the research hypothesis $H_1 : \mu > \mu_0$ is the probability of observing a value of a Student's t variable T as large or larger than the calculated value T_{calc} that we actually do observe, *i.e.,* P–value $= P(T \geq T_{calc})$, where T denotes a Student's t variable with $n-1$ degrees of freedom. This P–value is computed under the assumption that the research hypothesis $H_1 : \mu > \mu_0$ is false and the null hypothesis $H_0 : \mu \leq \mu_0$ is true. Because the null hypothesis only specifies that $\mu \leq \mu_0$, we need to choose a particular value of μ (that is no larger than μ_0) in order to compute the P–value. It is most appropriate to use $\mu = \mu_0$ for this computation. Using $\mu = \mu_0$, which defines the boundary between $\mu \leq \mu_0$, where the null hypothesis is true, and $\mu > \mu_0$, where the research hypothesis is true, provides some protection against incorrectly rejecting $H_0 : \mu \leq \mu_0$.

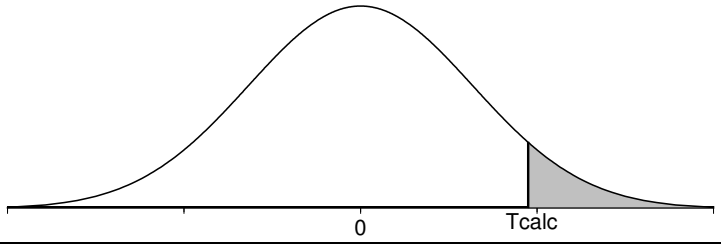
The steps for performing a hypothesis test for

$$H_0 : \mu \leq \mu_0 \text{ versus } H_1 : \mu > \mu_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(T \geq T_{calc})$, where T denotes a Student's t variable with $n-1$ degrees of freedom and $T_{calc} = (\overline{X} - \mu_0)/\widehat{\text{S.E.}}(\overline{X})$ as described above. This P–value is the area to the right of T_{calc} under the density curve for the Student's t distribution with $n-1$ degrees of freedom as indicated in Figure 18.

Figure 18. P–value for $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : \mu > \mu_0$ over $H_0 : \mu \leq \mu_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population mean μ is greater than μ_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : \mu > \mu_0$ over $H_0 : \mu \leq \mu_0$. That

is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population mean μ is greater than μ_0.

The procedure for testing the null hypothesis $H_0 : \mu \leq \mu_0$ versus the research hypothesis $H_1 : \mu > \mu_0$ given above is readily modified for testing the null hypothesis $H_0 : \mu \geq \mu_0$ versus the research hypothesis $H_1 : \mu < \mu_0$. The essential modification is to change the direction of the inequality in the definition of the P–value. Consider a situation where the research hypothesis specifies that the population mean μ is less than the particular, hypothesized value μ_0. For these hypotheses values of the sample mean $\overline{X}$ that are sufficiently small relative to μ_0 provide evidence in favor of the research hypothesis $H_1 : \mu < \mu_0$ and against the null hypothesis $H_0 : \mu \geq \mu_0$. Therefore, the appropriate P–value is the probability of observing a value of a Student's t variable T as small or smaller than the value actually observed. As before, the P–value is computed under the assumption that $\mu = \mu_0$. The calculated t statistic T_{calc} is defined as before; however, in this situation the P–value is the area under the density curve of the Student's t distribution with $n-1$ degrees of freedom to the left of T_{calc}, since values of $\overline{X}$ that are small relative to μ_0 constitute evidence in favor of the research hypothesis.

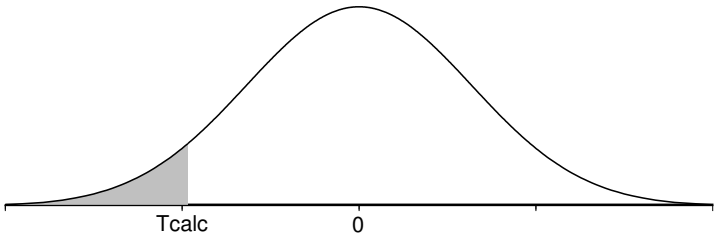
The steps for performing a hypothesis test for

$$H_0 : \mu \geq \mu_0 \text{ versus } H_1 : \mu < \mu_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(T \leq T_{calc})$, where T denotes a Student's t variable with $n-1$ degrees of freedom and $T_{calc} = (\overline{X} - \mu_0)/\widehat{\text{S.E.}}(\overline{X})$ as described above. This P–value is the area to the left of T_{calc} under the density curve for the Student's t distribution with $n-1$ degrees of freedom as indicated in Figure 19.

Figure 19. P–value for $H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : \mu < \mu_0$ over $H_0 : \mu \geq \mu_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population mean μ is less than μ_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : \mu < \mu_0$ over $H_0 : \mu \geq \mu_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population mean μ is less than μ_0.

Example. Brain changes in response to experience. A study by Rosenzweig, Bennett, and Diamond, described in an article in *Scientific American* (1964), was conducted to examine the effects of psychological environment on the anatomy of the brain. The units for this study came from a strain of genetically pure rats. A pair of rats was selected at random from each of 12 litters of rats; one of these rats was placed in group A and the other in group B. Each animal in group A lived with eleven others in a large cage, furnished with playthings which were changed daily. Each animal in group B lived in isolation, with no toys. Both groups of rats were provided with as much food and drink as they desired. After a month, the rats were killed and dissected. One variable which was measured was the weight (in milligrams) of the cortex of the rat. The cortex is the "thinking" part of the brain. The question we wish to address here is whether there is evidence in favor of the contention that the cortex of a rat raised in the more stimulating environment of group A will tend to be larger than the cortex of a rat raised in the less stimulating environment of group B.

The researchers conducted this experiment five times. Data from one of these experiments are given in Table 10. There are sets of three values for each of twelve pairs of littermates in this table: the weight of the cortex of the rat in group A, the weight of the cortex of the rat in group B, and the difference between these two weights (A weight minus B weight); all of these values are measured in milligrams.

Table 10. Rat cortex weight data.

pair	group A	group B	difference	pair	group A	group B	difference
1	690	668	22	7	720	665	55
2	701	667	34	8	718	689	29
3	685	647	38	9	718	642	76
4	751	693	58	10	696	673	23
5	647	635	12	11	658	675	-17
6	647	644	3	12	680	641	39

Summary statistics for the differences in cortex weights (A weight minus B weight) are given in Table 11, a stem and leaf histogram is provided in Figure 20, and a normal probability plot is given in Figure 21. Based on the summary statistics and the stem and leaf histogram there is some evidence that this distribution is slightly skewed to the left; however, the normal probability plot is reasonably linear and it seems reasonable to

assume that these data form a random sample from a normal population. More specifically, it seems reasonable to assume that these 12 differences are independent realizations of a normally distributed variable with population mean μ_D. We can think of this population mean difference μ_D as the average difference that would be obtained if this experiment was conducted using all possible littermate pairs from this strain of genetically pure rats.

Table 11. Summary statistics for the rat cortex weight differences.

minimum	-17.0	Q_1- minimum	34.0
Q_1	17.0	median $-Q_1$	14.5
median	31.5	Q_3- median	15.5
Q_3	47.0	maximum $-Q_3$	29.0
maximum	76.0		
mean	31.0000	range	93
standard deviation	25.3162	IQ range	30

Figure 20. Stem and leaf histogram for the rat cortex weight difference data.

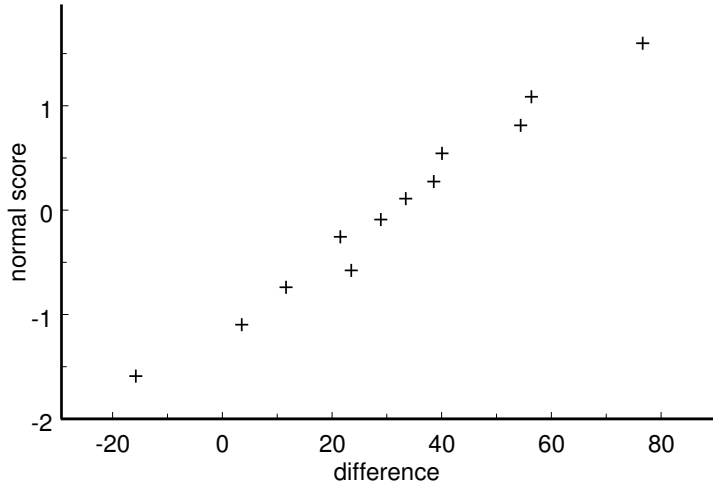
In this stem and leaf histogram the stem represents tens and the leaf represents ones. (milligrams)

stem	leaf
-1	7
-0	
0	3
1	2
2	239
3	489
4	
5	58
6	
7	6

For this example the research hypothesis can be formalized as $H_1 : \mu_D > 0$. This research hypothesis specifies that the population mean difference, μ_D, between the cortex weight of a stimulated rat (a rat raised in a stimulating environment like that of group A) and the cortex weight of a unstimulated rat (a rat raised in a non–stimulating environment like that of group B) exceeds zero, *i.e.*, for this population of pairs of rats, on average, the cortex weight of a stimulated rat would be higher than the cortex weight of an unstimulated rat. The observed value of the Student's t statistic is $T_{calc} = 4.2418$ with 11 degrees of freedom, which gives a P–value of .0007. Notice that this P–value is the probability that a Student's t variable with 11 degrees of freedom will be at least as large as $T_{calc} = 4.2418$

when $\mu_D = 0$. Since this P–value is quite small, there is very strong evidence that the population mean difference μ_D is greater than zero. Hence, the data do support the contention that for this population of pairs of rats, on average, the cortex weight of a stimulated rat would be higher than the cortex weight of an unstimulated rat.

Figure 21. Normal probability plot for weight differences.



To get a feel for the size of this population mean difference we can compute a 95% confidence interval estimate for μ_D. The sample mean difference is $\overline{D} = 31$, the sample standard error of the mean is $\widehat{\text{S.E.}}(\overline{D}) = 7.3082$, and the multiplier for the 95% margin of error, based on the Student's t distribution with 11 degrees of freedom, is $k = 2.201$. Thus the 95% margin of error for $\overline{D}$ is $(2.201)(7.3082) = 16.0853$, and we are 95% confident that the population mean difference μ_D is between 14.9147 and 47.0853 milligrams. In other words, we are 95% confident that for this population of pairs of rats, on average, the cortex weight of the stimulated rat would exceed the cortex weight of the unstimulated rat by at least 14.9147 mg but by no more than 47.0853 mg.

The hypothesis tests we have discuss thus far are only appropriate when we have enough *a priori* information, *i.e.*, information that does not depend on the data to be used for the hypothesis test, to postulate that the population mean μ is on one side of a particular value μ_0. That is, we have only considered situations where the research hypothesis is directional. There are situations when we will not have enough *a priori* information to allow us to choose the appropriate directional research hypothesis. Instead, we might only conjecture that the population mean μ is different from some particular value μ_0. In a situation like this our research hypothesis specifies that the population mean μ is different from μ_0, *i.e.*, $H_1 : \mu \neq \mu_0$.

To decide between the null hypothesis $H_0 : \mu = \mu_0$ and the research hypothesis $H_1 : \mu \neq \mu_0$, we need to decide whether the sample mean $\overline{X}$ supports the null hypothesis

by being "close to μ_0", or supports the research hypothesis by being "far away from μ_0". In this situation the P–value is the probability that the sample mean $\overline{X}$ would be as far or farther away from μ_0 in either direction as is the value that we actually observe. In other words, the P–value is the probability that the standardized distance from $\overline{X}$ to μ_0 (the standardized absolute value of the difference between $\overline{X}$ and μ_0) is as large or larger than the actual observed value of this standardized distance. As before, the P–value is computed under the assumption that the null hypothesis is true and $\mu = \mu_0$. In this situation the calculated t statistic T_{calc} is the absolute value of the t statistic that would be used for testing a directional hypothesis. That is, the calculated t statistic is

$$T_{calc} = \left| \frac{\overline{X} - \mu_0}{\widehat{\text{S.E.}}(\overline{X})} \right| .$$

In terms of this t statistic the P–value is the probability that the absolute value of a Students's t variable with $n-1$ degrees of freedom T would take on a value as large or larger that T_{calc}, assuming that $\mu = \mu_0$. This probability is the sum of the area under the appropriate Student's t density curve to the left of $-T_{calc}$ and the area under this Student's t density curve to the right of T_{calc}. We need to add these two areas (probabilities) since we are finding the probability that the sample mean $\overline{X}$ would be as far or farther away from μ_0 in either direction as is the value that we actually observe, when $\mu = \mu_0$.

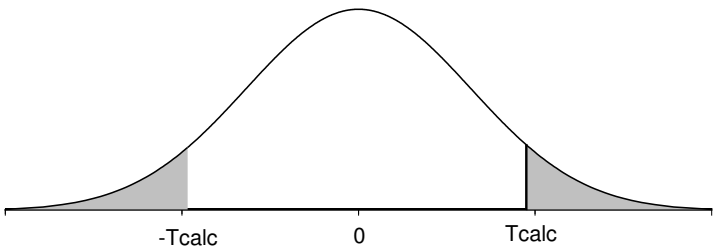
The steps for performing a hypothesis test for

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(|T| \geq T_{calc}) = P(T \leq -T_{calc}) + P(T \geq T_{calc})$, where T denotes a Student's t variable with $n-1$ degrees of freedom and $T_{calc} = |\overline{X} - \mu_0| / \widehat{\text{S.E.}}(\overline{X})$ as described above. This P–value is the sum of the area to the left of $-T_{calc}$ and the area to the right of T_{calc}, where each area is that under the density curve for the Student's t distribution with $n-1$ degrees of freedom over the appropriate region on the x–axis as indicated in Figure 22.

Figure 22. P–value for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : \mu \neq \mu_0$ over $H_0 : \mu = \mu_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population mean μ is not equal to μ_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : \mu \neq \mu_0$ over $H_0 : \mu = \mu_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population mean μ differs from μ_0.

Example. Newcomb's measurements of the speed of light (revisited). The parameter we were estimating in our analysis of Newcomb's measurements of the speed of light was the population mean time μ for light to travel a distance of 7442 meters. Notice that the population mean time μ is actually defined by the process of measurement being used. That is, we can think of μ as the long run average time that we would observe if we replicated Newcomb's experiment a large times. We might wonder how this population mean time μ relates to the "true time" it would take for light to travel a distance of 7442 meters. We don't know exactly what this "true time" is, but we can use a generally accepted, modern measurement of the speed of light to obtain a hypothesized "true time." Stigler (*op. cit.*) used the modern estimate of the speed of light in a vacuum of 299,792.5 km/sec adjusted to give the speed of light in air and converted to a time as measured by Newcomb to obtain a hypothesized "true time" of 33.02. Therefore, our present goal is to determine how the population mean time μ relates to the "true time" $\mu_0 = 33.02$.

We do not have sufficient *a priori* information to specify a directional hypothesis; therefore, we will test the null hypothesis $H_0 : \mu = 33.02$ versus the research hypothesis $H_1 : \mu \neq 33.02$ to determine whether the population mean time μ is equal to or different from the hypothesized "true value" of 33.02. For Newcomb's $n = 64$ measurements the sample mean is $\overline{X} = 27.75$, the sample standard error of the mean is $\widehat{\text{S.E.}}(\overline{X}) = .6354$, and the calculated t statistic is

$$T_{calc} = \left| \frac{27.75 - 33.02}{.6354} \right| = 8.2940.$$

The P–value is less than .0001 indicating that observing a sample mean as far away from the hypothesized "true value" 33.02 (in either direction) as Newcomb did is extremely unlikely if in fact $\mu = 33.02$. We can conclude that the population mean μ corresponding to Newcomb's experiment is almost certainly not equal to the hypothesized value of 33.02. The confidence interval, (26.4803, 29.0197), we computed above suggests that the mean μ that Newcomb was estimating is less than the hypothesized value 33.02.

7.3 Inference for a population median

The inferential methods for a population mean discussed above require at least approximate normality of the population distribution of the variable of interest. In this section we will consider methods for making inferences about a population median which do not require the assumption of a particular form for the population distribution of the variable of interest. That is, the inferential methods considered in this section are applicable for any population distribution for a continuous variable regardless of the shape of the corresponding density curve.

We will begin by discussing a method for testing a hypothesis about a population median. The essence of this method is to re–express the hypothesis about the population median as a hypothesis about a related population proportion and to then use inferential methods for a population proportion to test the hypothesis about the population median.

For a continuous variable X we can think of the population median M as the point on the number line which divides the area under the corresponding population density curve into two equal areas (each of area one–half). The population median M is analogous to the sample median which divides the histogram into two equal areas. Notice that if we observe a single value of X, then the probability that we will observe a value larger than the population median M is 1/2, *i.e.*, $P(X > M) = 1/2$, and similarly, the probability that we will observe a value smaller than M is 1/2, *i.e.*, $P(X < M) = 1/2$.

Let M denote the population median of the distribution of the continuous variable X and consider a hypothesis relating the population median M to a particular, fixed value M_0. We can dichotomize the population of values of X by thinking of the event "observe $X > M_0$" as a success and the event "observe $X < M_0$" as a failure. The population success proportion p corresponding to this dichotomization is $p = P(X > M_0)$, *i.e.*, p is the probability that a single value of X chosen from the infinite population of values of X will be larger than the hypothesized value M_0. The corresponding population failure proportion is $1 - p = P(X < M_0)$.

The three possible relationships between the population median M and the particular value M_0 are readily re–expressed in terms of the corresponding population success probability $p = P(X > M_0)$. If the population median M exceeds the particular value M_0, then, since the area or probability to the right of the population median M is 1/2 and since M_0 is to the left of M on the number line, we must have $p = P(X > M_0) > 1/2 = P(X > M)$. Hence we see that $M > M_0$ is equivalent to $p > 1/2$. Similarly, if M is less than M_0, then M_0 is to the right of M on the number line and we must have $p = P(X > M_0) < 1/2 = P(X > M)$; thus $M < M_0$ is equivalent to $p < 1/2$. Finally, if $M = M_0$, then we must have $p = P(X > M_0) = 1/2 = P(X > M)$; thus $M = M_0$ is equivalent to $p = 1/2$. These relationships allow us to re–express a

hypothesis relating the population median M to the particular value M_0 as a hypothesis about the corresponding population success probability $p = P(X > M_0)$.

For ease of reference the three standard pairs of null and research hypotheses about M are summarized below (recall that $p = P(X > M_0)$)

1. $H_0 : M \leq M_0$ versus $H_1 : M > M_0$ corresponds to $H_0 : p \leq .5$ versus $H_1 : p > .5$.

2. $H_0 : M \geq M_0$ versus $H_1 : M < M_0$ corresponds to $H_0 : p \geq .5$ versus $H_1 : p < .5$.

3. $H_0 : M = M_0$ versus $H_1 : M \neq M_0$ corresponds to $H_0 : p = .5$ versus $H_1 : p \neq .5$.

The dichotomy described above, where $X > M_0$ constitutes a success and $X < M_0$ constitutes a failure, does not allow for the possibility that the continuous variable X is exactly equal to M_0. This is allowable, theoretically, when we are talking about the population distribution of X; however, when we examine the data we may find that one or more of the observations are equal to the hypothesized value M_0. The easiest solution to this potential difficulty is to remove any observations which are exactly equal to M_0 and adjust the sample size accordingly before we perform a hypothesis test. In other words, a hypothesis test for comparing the population median M to the hypothesized value M_0 is based on the observed proportion $\hat{p}$ of successes (values of X which are greater than M_0) relative to n, where n is the number of observations which are not equal to M_0. An alternative to discarding values exactly equal to M_0 is to classify half of these values as successes and half as failures and use the original sample size for n.

Example. Darwin's plant height comparison. Charles Darwin conducted an experiment to determine whether cross–fertilized plants tend to be more vigorous than self–fertilized plants. (The plants used in Darwin's study were young corn (*Zea mays*) plants.) (Darwin (1876), *The Effect of Cross– and Self–fertilization in the Vegetable Kingdom*, second edition, John Murray, London) Darwin selected several plants and fertilized several flowers on each plant; a number of flowers were cross–fertilized with pollen taken from distant plants and a number of flowers were self–fertilized with their own pollen. Seeds gathered from these flowers were allowed to ripen and were then placed in wet sand to germinate. Fifteen seedlings from cross–fertilized seeds were selected and fifteen seedlings from self–fertilized seeds were selected. These seedlings were paired (one cross–fertilized and one self–fertilized) and the two seedlings in each pair were planted on opposite sides of the same pot. After a fixed period of time the height of each plant (in inches) was recorded. The raw data and the associated differences are provided in Table 12. The distribution of

the differences is summarized in Table 13 and Figure 23 and a normal probability plot is given in Figure 24.

Table 12. Plant height data.

	plant height		
pair	cross–fertilized	self–fertilized	difference
1	23.500	17.375	6.125
2	12.000	20.375	-8.375
3	21.000	20.000	1.000
4	22.000	20.000	2.000
5	19.125	18.375	0.750
6	21.500	18.625	2.875
7	22.125	18.625	3.500
8	20.375	15.250	5.125
9	18.250	16.500	1.750
10	21.625	18.000	3.625
11	23.250	16.250	7.000
12	21.000	18.000	3.000
13	22.125	12.750	9.375
14	23.000	15.500	7.500
15	12.000	18.000	-6.000

If it is true that cross–fertilized plants tend to be more vigorous than self–fertilized plants, then we would expect a cross–fertilized plant to be taller than a self–fertilized plant of the same age. Therefore, we can formalize this theory as the research hypothesis that the population median of the difference between the height of a cross–fertilized plant and the height of a self–fertilized plant grown in the same pot M_D is greater than zero, *i.e.*, $H_1 : M_D > 0$. We can think of the population as consisting of all of the pairs of seedlings (one cross–fertilized and one self–fertilized) which could have been used in this experiment; and, we can think of the population median difference M_D as the median of the differences between the heights of these pairs of seedlings (height of the cross–fertilized plant minus height of the self–fertilized plant). Notice that this research hypothesis specifies that, for this population of potential pairs of plants, the cross-fertilized plant would be taller than the self–fertilized plant more than half of the time.

Table 13. Summary statistics for the plant height differences.

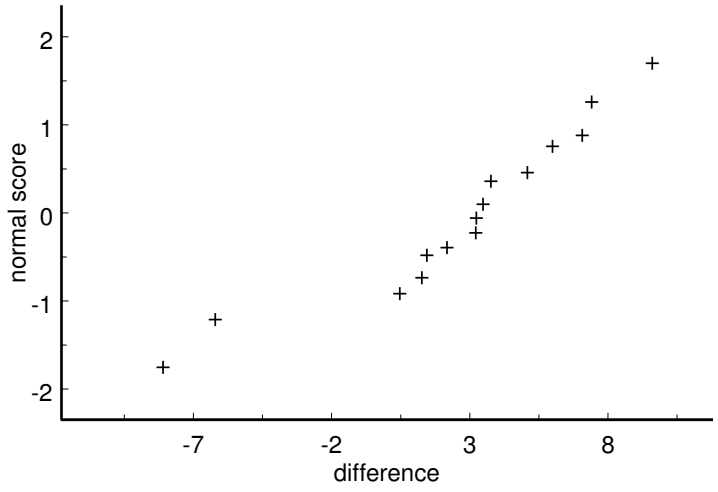
minimum	-8.375	Q_1- minimum	9.375
Q_1	1.000	median $-Q_1$	4.000
median	3.000	Q_3- median	3.125
Q_3	6.125	maximum $-Q_3$	3.250
maximum	9.375		

Figure 23. Stem and leaf histogram for the plant height difference data (rounded).

In this stem and leaf histogram the stem represents tens of inches and the leaf represents inches.

stem	leaf
-0	8
-0	6
-0	
-0	
-0	
0	01
0	2233
0	445
0	67
0	89

Figure 24. Normal probability plot for height differences.



The normality assumption is questionable in this example since the stem and leaf histogram and the normal probability plot show some evidence of extreme skewness to the left in this sample of differences.

Since all but two of the 15 plant height differences are positive, there appears to be strong evidence that the population median height difference is positive. For these data there are no zero differences so we use the actual sample size $n = 15$ to perform our test of $H_0 : M_D \leq 0$ versus $H_1 : M_D > 0$ ($H_0 : p \leq .5$ versus $H_1 : p > .5$, where $p = P(D > 0)$ denotes the proportion of this population of pairs of plants for which the cross–fertilized plant would be taller than the self–fertilized plant). Thirteen of the differences are positive (successes) which gives $\hat{p} = 13/15 = .8667$. The standard error of $\hat{p}$, assuming that $p = .5$,

is $S.E.(\hat{p}) = .1291$, the calculated Z–statistic is $Z_{calc} = 2.8402$, and the P–value is .0023. Therefore, there is very strong evidence that the population median height difference M_D is greater than zero which means that there is very strong evidence in support of the contention that, for this population of pairs of plants, the cross-fertilized plant would be taller than the self–fertilized plant more than half of the time.

We can construct a 95% confidence interval estimate for the population median M by finding the interval of values of M_0 for which a test at the 5% level of significance **does not** lead to the rejection of $H_0 : M = M_0$. Recall that $H_0 : M = M_0$ and $H_1 : M \neq M_0$ are equivalent to $H_0 : p = .5$ and $H_1 : p \neq .5$, where $p = P(X > M_0)$. A test at the 5% level of significance will fail to reject the null hypothesis $H_0 : M = M_0$ if

$$|\hat{p} - .5| \leq 1.96/(2\sqrt{n}),$$

where $\hat{p}$ is the observed proportion of values of X which are greater than M_0. The lower and upper limiting values of $\hat{p}$ in this expression, denoted by $\hat{p}_L$ and $\hat{p}_U$, are

$$\hat{p}_L = .5 - 1.96/(2\sqrt{n}) \quad \text{and} \quad \hat{p}_U = .5 + 1.96/(2\sqrt{n}).$$

The corresponding limits on the value of M_0, denoted by M_L and M_U, are the $\hat{p}_L 100$ percentile and the $\hat{p}_U 100$ percentile of the observed values of X, respectively. We will adopt the rounding convention described below to avoid the need for averaging observed values of X when computing M_L and M_U. First convert $\hat{p}_L$ and $\hat{p}_U$ from proportions to counts by multiplying each by the sample size n. If the count $n\hat{p}_L$ is not a whole number, round it down to the next whole number. If the count $n\hat{p}_U$ is not a whole number, round it up to the next whole number. Finally, find the observed values of X which occur at these locations in an ordered listing of the observed values. As is true when counting to find a sample median or quartile be sure to list any repeated values as many times as they occur. The resulting values for M_L and M_U form the endpoints of our 95% confidence interval estimate for the population median M.

The procedure for calculating the 95% confidence interval estimate for a population median M is summarized below.

1. Arrange the data (observations) in increasing order from smallest (obs. no. 1) to largest (obs. no. n). Be sure to include all n values in this list, including repeats if there are any.
2. Compute the quantity $n\hat{p}_L$ and round it down to the nearest whole number if it is not a whole number. The observation at the location indicated by the rounded–down value in the ordered listing of the data is M_L.

3. Compute the quantity $n\hat{p}_U$ and round it up to the nearest whole number if it is not a whole number. The observation at the location indicated by the rounded–up value in the ordered listing of the data is M_U.
4. Conclude that we are 95% confident that the population median M is between M_L and M_U.

Example. Darwin's plant height comparison (revisited). For this example we have $1.96/(2\sqrt{15}) = .2530$, $\hat{p}_L = .5 - .2530 = .2470$, and $\hat{p}_U = .5 + .2530 = .7530$, giving $n\hat{p}_L = 3.705$ which we round down to 3 and $n\hat{p}_U = 11.295$ which we round up to 12. Observation number 3 is .75 and observation number 12 is 6.125. Therefore, we are 95% confident that the population median height difference for this population of pairs of plants is between .75 inches and 6.125 inches.

7.4 Summary

The majority of this chapter is devoted to inference for the population mean μ of the distribution of a continuous variable X. We began by discussing probability models for the distribution of a continuous variable (density curves) and then introduced the normal distribution which serves as the basis for our inferences about μ (based on the Student's t distribution).

Given data which form a random sample of size n from a population with population mean μ and population standard deviation σ, the sampling distribution of the sample mean $\overline{X}$ has mean μ and the population standard error of $\overline{X}$ is S.E.$(\overline{X}) = \sigma/n$. Thus the sample mean $\overline{X}$ is unbiased as an estimator of the population mean μ and the variability in the sample mean $\overline{X}$ as an estimator of the population mean μ can be quantified by this standard error. If we also assume that the population distribution of X is a normal distribution, *i.e.*, if we assume that the data form a random sample of size n from a normal distribution with population mean μ and population standard deviation σ, then the sampling distribution of $\overline{X}$ is the normal distribution with population mean μ (the same as that of X) and standard deviation S.E.$(\overline{X})$.

Given data which form a random sample of size n from a normal distribution with population mean μ and population standard deviation σ and with sample mean $\overline{X}$ and sample standard deviation $S_X = S$, the quantity

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

follows the Student's t distribution with $n - 1$ degrees of freedom. Therefore, if the assumption that the population distribution of X is normal is reasonable, then we can use the Student's t distribution to make inferences about the population mean μ. It is important to remember the normality assumption needed for the Student's t distribution

and to verify that this assumption is reasonable by examining the data for violations of this assumption. The Student's t methods work reasonably well provided the normality assumption is not totally unreasonable.

The interval from $\overline{X} - kS/\sqrt{n}$ to $\overline{X} + kS/\sqrt{n}$, where k denotes the 97.5 percentile of the Student's t distribution with $n-1$ degrees of freedom, is a 95% confidence interval for μ. We can test a hypothesis relating μ to a specified value μ_0 by using the Student's t test statistic

$$T_{calc} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

to find the appropriate P–value. The P–value for $H_1 : \mu > \mu_0$ is the probability $P(T \geq T_{calc})$; the P–value for $H_1 : \mu < \mu_0$ is the probability $P(T \leq T_{calc})$; and, the P–value for $H_1 : \mu \neq \mu_0$ is the probability $P(|T| \geq |T_{calc}|)$, where T denotes a Student's t variable with $n-1$ degrees of freedom, *i.e.*, these P–values are areas under the density curve of the Student's t distribution with $n-1$ degrees of freedom.

The Student's t inferential methods for a population mean are based on the assumption that the underlying population distribution is reasonably modeled by a normal distribution. When this normality assumption is not tenable we need to consider a method of inference which is applicable under weaker assumptions. One approach to inference about the center of a distribution based on the population median is discussed in Section 7.4. This approach to inference about the population median does not require the assumption of a specific form for the underlying population distribution.

7.5 Exercises

Provide a complete analysis for the following example. Be sure to: define a relevant population mean μ; setup and perform a relevant hypothesis test; and, find a confidence interval for μ. Be sure to include comments regarding the validity of the normality assumption for this example. Provide a complete summary of your findings in the context of the example.

1. An article by Rosner, Willett, and Spiegelman in *Statistics in Medicine* (1989) describes a study conducted to assess the make up of the diets of a population of women. The data used here are as reported in Ott and Longnecker (2002). A sample of $n = 168$ women was obtained and each of these women completed a food frequency questionnaire. The completed questionnaires were then used to determine the percentage of calories from fat in each woman's diet. The values of the variable X, the percentage of calories from fat in a woman's diet, are summarized in the stem and leaf histogram in Figure 25. For display purposes the data have been rounded to the nearest 1 percent. The actual data are given in Table 14. In 2002 the Food and Nutrition Board, a unit of the Institute of Medicine `iom.edu`, recommended that adults should restrict the percentage of calories from fat in their diet to the range from 20% to 35%.

If you wish to analyze these data without entering all $n = 168$ values in your computer or calculator you may use the facts that: the sample mean for these data is 36.91899 and the sample standard deviation is 6.72820.

Figure 25. Stem and leaf histogram for percentage of calories from fat.

In this stem and leaf histogram the stem represents tens and the leaf represents ones. For example, the smallest value is 16% and the largest is 58%.

stem	leaf
1	
1	6
2	014
2	55666677888999999
3	000001111111222222222333333334444444444
3	55555555555556666666666667777778888888888888999999999
4	000000000111111111122222222334444444
4	555666667888
5	00134
5	68

Table 14. Percentage of calories from fat data.

15.92	20.22	20.80	24.06	24.98	25.07	25.53	26.16	26.18	26.27
26.79	27.29	27.54	28.16	28.35	28.76	29.07	29.34	29.34	29.45
29.46	29.55	29.72	29.94	29.99	30.50	30.61	30.71	30.96	30.99
31.02	31.10	31.27	31.61	31.71	31.75	31.97	32.24	32.26	32.28
32.43	32.73	32.86	32.96	33.08	33.11	33.13	33.23	33.32	33.51
33.71	33.72	33.86	33.87	33.95	34.11	34.17	34.25	34.31	34.51
34.86	34.87	34.87	34.89	35.09	35.18	35.18	35.20	35.29	35.32
35.49	35.50	35.56	35.67	35.69	35.71	35.74	35.77	36.07	36.07
36.30	36.32	36.32	36.37	36.58	37.04	37.10	37.14	37.34	37.39
37.47	37.56	37.85	37.88	37.89	38.02	38.04	38.10	38.19	38.21
38.36	38.41	38.45	38.58	38.81	38.88	38.89	38.97	39.13	39.22
39.25	39.40	39.80	39.84	39.95	40.05	40.12	40.25	40.29	40.46
40.48	40.69	41.12	41.15	41.22	41.29	41.30	41.32	41.41	41.42
41.44	41.53	41.64	41.69	41.78	41.89	42.12	42.17	42.20	42.37
42.60	42.98	43.57	43.74	43.79	44.27	44.28	44.33	44.39	44.66
45.06	45.28	45.51	45.82	45.83	46.22	46.38	46.97	47.63	47.83
48.29	49.65	49.86	50.72	53.08	54.05	55.54	57.85		

2. Cox and Snell (1981), *Applied Statistics*, Chapman and Hall, discuss a study conducted to examine changes in blood pressure due to the drug captopril. The original source is MacGregor, Markandu, Roulston, and Jones (1979), Essential hypertension: effect of an oral inhibitor of angiotension–converting enzyme, *British Medical Journal*, **2** 1106–1109. The data given in Table 15 are the blood pressures (in mm Hg) for 15 patients with moderate essential hypertension. The data consist of supine systolic and diastolic blood pressures measured immediately before and two hours after taking 25 mg of the drug captopril. Relevant differences are also provided.

Table 15. Blood pressure data.

patient	**systolic**			**diastolic**		
	before	**after**	**difference**	**before**	**after**	**difference**
1	210	201	9	130	125	5
2	169	165	4	122	121	1
3	187	166	21	124	121	3
4	160	157	3	104	106	-2
5	167	147	20	112	101	11
6	176	145	31	101	85	16
7	185	168	17	121	98	23
8	206	180	26	124	105	19
9	173	147	26	115	103	12
10	146	136	10	102	98	4
11	174	151	23	98	90	8
12	201	168	33	119	98	21
13	198	179	19	106	110	-4
14	148	129	19	107	103	4
15	154	131	23	100	82	18

Chapter 8
Comparing Two Means

8.1 Introduction

In Chapter 7 we considered inferential methods for the location of the center of the population distribution of a single continuous variable. We will now consider extensions of these methods to provide inferential methods for comparing the locations of the population distributions of two continuous variables. More specifically, we will consider methods for making inferences about the difference $\mu_1 - \mu_2$ between two population means μ_1 and μ_2.

First consider a situation where the only difference between the population distributions of two continuous variables, X_1 and X_2, is their location on the number line. In other words, suppose that the density curve for X_2 is identical to the density curve for X_1 except for its location on the number line. We will refer to this assumption about the population distributions of X_1 and X_2 as the **shift assumption**, since this assumption implies that the density curve for X_2 can be obtained by shifting the density curve for X_1 to the right or to the left along the number line. Under this shift assumption the difference, $\mu_1 - \mu_2$, between the two population means completely characterizes the difference between the two population distributions. Notice that under this shift assumption, if there is a positive constant d for which $\mu_1 - \mu_2 = d$ (*i.e.*, $\mu_1 = \mu_2 + d$), indicating that the density curve for X_1 is located d units to the right of the density curve for X_2, then the difference $M_1 - M_2$ between the population medians, M_1 and M_2, is also d (*i.e.*, $M_1 - M_2 = d$ and $M_1 = M_2 + d$). Therefore, under this shift assumption, a comparison of the locations of the two distributions based on the difference between the population means is equivalent to a comparison based on the population medians in the sense that the differences between each of these pairs of parameters is the same.

When the shift assumption is not valid, *i.e.*, when the two population distributions differ in aspects other than a simple shift in location, we must decide which parameter, say the population mean or the population median, is appropriate as a quantification of the location of each distribution and to quantify the difference between the locations of the two distributions. In other words, in the general situation when the shift assumption is not valid the difference between the population means and the difference between the population medians will be different and neither of these differences will completely describe the difference between the two distributions. For example, if the distribution of X_1 is skewed right and the distribution of X_2 is skewed left, it is possible for the population means to be equal while the population medians are different. Hence, when the shift assumption is not valid we must be careful about how we interpret an inference about the difference

between any two particular location parameters, such as the population means, since the distributions differ in aspects other than a simple shift in location.

We will restrict our attention to methods which are appropriate when the data comprise two independent random samples; one random sample (the X_1 values) from a population with population mean μ_1; and, one random sample (the X_2 values) from a population with population mean μ_2. The assumption that these random samples are independent basically means that the method used to select the random sample from the first population is not influenced by the method used to select the random sample from the second population, and *vice versa*.

8.2 Comparing the means of two normal populations

In this section we will assume that the population distribution of X_1 is a normal distribution with population mean μ_1 and population standard deviation σ_1 and the population distribution of X_2 is a normal distribution with population mean μ_2 and population standard deviation σ_2. We will discuss methods for making inferences comparing the locations of these normal distributions as quantified by the difference, $\mu_1 - \mu_2$, between the two population means. As stated in the introduction, we will assume that the data comprise two independent random samples. Let n_1 denote the size of the random sample (of X_1 values) from the normal population with population mean μ_1 and let n_2 denote the size of the random sample (of X_2 values) from the normal population with population mean μ_2. The sample mean $\overline{X}_1$ is the obvious estimate of the corresponding population mean μ_1 and the sample mean $\overline{X}_2$ is the obvious estimate of the corresponding population mean μ_2. Similarly, the difference, $\overline{X}_1 - \overline{X}_2$, between these two sample means is the obvious estimate of the corresponding difference, $\mu_1 - \mu_2$, between the population means. To describe the behavior of $\overline{X}_1 - \overline{X}_2$ as an estimator of $\mu_1 - \mu_2$ we need to know some properties of its sampling distribution.

Some properties of the sampling distribution of $\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2$.

Let $\overline{X}_1$ denote the sample mean of a random sample of size n_1 from a distribution with population mean μ_1 and population standard deviation σ_1 and let $\overline{X}_2$ denote the sample mean of a random sample of size n_2 from a distribution with population mean μ_2 and population standard deviation σ_2. Assume that these two random samples are independent. The sampling distribution of the difference, $\overline{X}_1 - \overline{X}_2$, between these two sample means has the following properties. The first two properties are valid in general and do not depend on the assumption of normal distributions.

1. The mean of the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is the difference, $\mu_1 - \mu_2$, between the corresponding population means. Therefore, just as the sample means $\overline{X}_1$ and

$\overline{X}_2$ are unbiased as estimators of μ_1 and μ_2, respectively, the sample mean difference $\overline{X}_1 - \overline{X}_2$ is unbiased as an estimator of the population mean difference $\mu_1 - \mu_2$.

2. The population standard error of $\overline{X}_1 - \overline{X}_2$ (the standard deviation of the sampling distribution of $\overline{X}_1 - \overline{X}_2$) is

$$\text{S.E.}(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

This expression indicates how the variability of $\overline{X}_1 - \overline{X}_2$ depends on the sample sizes and population standard deviations. Notice that the population variance $\text{var}(\overline{X}_1 - \overline{X}_2)$ (the square of $\text{S.E.}(\overline{X}_1 - \overline{X}_2)$) is equal to the sum of the population variance of $\overline{X}_1$ and the population variance of $\overline{X}_2$. This property is a consequence of our assumption that the random samples are independent; and, this expression for the standard error of the difference between two sample means is not appropriate if the random samples are not independent.

3. If the random samples from which the sample means $\overline{X}_1$ and $\overline{X}_2$ are computed are random samples from **normal distributions** with population means and population standard deviations as given above, then in addition to the two properties above, the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is a **normal distribution** with population mean $\mu_1 - \mu_2$ and population standard deviation $\text{S.E.}(\overline{X}_1 - \overline{X}_2)$.

The choice of the appropriate inferential methods for comparing the two normal population means μ_1 and μ_2 depends on the relationship between the two unknown, population standard deviations σ_1 and σ_2. In particular, the choice of the appropriate estimate of the population standard error of $\overline{X}_1 - \overline{X}_2$ depends on whether the two population standard deviations σ_1 and σ_2 are equal.

Strictly speaking, the inferential methods based on the Student's t distribution described below are only appropriate when the data constitute independent random samples from normal populations. However, these methods are known to be generally reasonable even when the underlying populations are not exactly normal populations, provided the underlying population distributions are reasonably symmetric and the true density curves have a more or less normal (bell–shaped) appearance. We can use descriptive methods to look for evidence of possible nonnormality, provided the sample sizes are reasonably large. As in the one mean situation of Chapter 7, the most easily detected and serious evidence of nonnormality you should look for is evidence of extreme skewness or evidence of extreme outlying observations. If there is evidence of extreme skewness or extreme outlying observations, then inferential methods based on the Student's t distribution should not be

used. An alternate approach to inference, based on ranks, which may be used when the Student's t methods are inappropriate is discussed in Section 8.3.

8.2a Inference when the two population standard deviations are equal

A normal distribution is completely determined by its mean and standard deviation; therefore, in the present context of comparing two normal populations the shift assumption is equivalent to the assumption that the two population standard deviations σ_1 and σ_2 are equal. In other words, if we assume that $\sigma_1 = \sigma_2$, then the only difference between the two normal populations we are comparing is that the normal density curve for X_1 is centered at μ_1 and the normal density curve for X_2 is centered at μ_2.

When the two population standard deviations are equal we can simplify the expression for the population standard error of $\overline{X}_1 - \overline{X}_2$. If we let $\sigma = \sigma_1 = \sigma_2$ denote the common value of the two population standard deviations, then the population standard error of $\overline{X}_1 - \overline{X}_2$ is

$$\text{S.E.}(\overline{X}_1 - \overline{X}_2) = \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

An appropriate estimator of the common standard deviation σ is the **pooled sample standard deviation** which we will denote by S_p. Recall that the sample standard deviation S_X for a single sample of n values of the variable X is the square root of the "average" of the squared deviations of the observed values of X from the sample mean $\overline{X}$,

$$S_X = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}.$$

In the present context, the n_1 values of X_1 have sample mean $\overline{X}_1$ and the n_2 values of X_2 have sample mean $\overline{X}_2$; therefore, the sum of squared deviations in the formula for S_X is replaced by the sum of two such sums of squared deviations, one for each sample. The divisor $n-1$ in the formula for S_X is replaced by the total number of observations $n_1 + n_2$ decreased by 2, *i.e.*, the one sample divisor $n-1$ is replaced by the two sample divisor $n_1 + n_2 - 2$. The resulting formula for the **pooled sample standard deviation** is

$$S_p = \sqrt{\frac{\sum (X_1 - \overline{X}_1)^2 + \sum (X_2 - \overline{X}_2)^2}{n_1 + n_2 - 2}}.$$

This pooled sample standard deviation can also be expressed in terms of the two sample standard deviations S_1 and S_2 as shown below

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}.$$

Strictly speaking, the inferential methods described in this subsection are only valid when the two population standard deviations σ_1 and σ_2 are equal. However, in practice these methods still perform reasonably well provided the two population standard deviations σ_1 and σ_2 are reasonably close to being equal and the two sample sizes n_1 and n_2 are reasonably similar. (This assumption is more critical when the sample sizes are very dissimilar.) A common rule of thumb for assessing the assumption of equal standard deviations says that the assumption of a common population standard deviation is reasonable if the ratio of the sample standard deviations is between 1/2 and 2.

When $\sigma_1 = \sigma_2$, the appropriate **sample standard error** of $\overline{X}_1 - \overline{X}_2$, based on the pooled sample standard deviation S_p, is

$$\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2) = S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

and the corresponding 95% **margin of error** of $\overline{X}_1 - \overline{X}_2$ is

$$\text{M.E.}(\overline{X}_1 - \overline{X}_2) = k\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2) = kS_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where k is the 97.5 percentile of the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom. Thus the interval from $(\overline{X}_1 - \overline{X}_2) - \text{M.E.}(\overline{X}_1 - \overline{X}_2)$ to $(\overline{X}_1 - \overline{X}_2) + \text{M.E.}(\overline{X}_1 - \overline{X}_2)$ is a 95% confidence interval for $\mu_1 - \mu_2$.

Example. Energy consumption. The data used in this example are part of data set 93 in Hand, Daly, Lunn, McConway, and Ostrowski (1994), *A Handbook of Small Data Sets*, Chapman and Hall, London. The original source is two reports issued in 1983 and 1984 by the Open University. A large–scale experiment on energy consumption was conducted in the early 1980's in the Pennyland district of Milton Keynes. A housing development of about 180 houses was built. About half of the houses had a standard level of roof and wall insulation. The others had extra roof and wall insulation (these houses also had double glazing and under–floor insulation). In addition to the differences in level of insulation, many of the houses were designed with passive solar heating features, *e.g.*, southern orientation with most of the windows on the south side. The other houses had a more traditional design. Energy consumption was monitored over several years. Table 1 provides the annual gas consumption (in 1000 kWh) for two independent random samples of houses. One random sample was selected from all of the houses with standard insulation (regardless of design type) and the other was selected from all of the houses with extra insulation (regardless of design type). Summary statistics are given in Table 2, stem and leaf histograms are given in Figure 1, and normal probability plots are provided in Figures 2 and 3.

Table 1. Gas consumption data (1000 kWh) (both designs).

standard insulation						extra insulation						
11.4	13.9	13.9	14.0	15.3	18.0	8.3	11.7	12.7	13.0	13.4	13.6	13.7
18.0	18.1	18.9	19.0	19.0	21.7	13.7	13.8	14.6	15.3	15.6	16.0	18.8

Table 2. Descriptive statistics for gas consumption (both designs).

	standard	**extra**
minimum:	11.40	8.3
Q1:	13.95	13.0
median:	18.00	13.7
Q3:	18.95	15.3
maximum:	21.70	18.8
Q1 - minimum:	2.55	4.7
median - Q1:	4.05	.7
Q3 - median:	.95	2.4
maximum - Q3:	2.75	3.5
mean:	16.7667	13.8714
standard deviation:	2.9959	2.3636
range:	10.3	10.5
IQ range:	5	2.3
sample size:	12	14

Figure 1. Stem and leaf histograms for gas consumption (both designs).

In these stem and leaf histograms the stem represents ones and the leaf represents tenths. (1000 kWh)

standard		**extra**	
		8	3
		9	
		10	
11	4	11	7
12		12	7
13	99	13	046778
14	0	14	6
15	3	15	36
16		16	0
17		17	
18	0019	18	8
19	00		
20			
21	7		

Figure 2. Normal probability plot for gas consumption for houses with standard insulation (both designs).

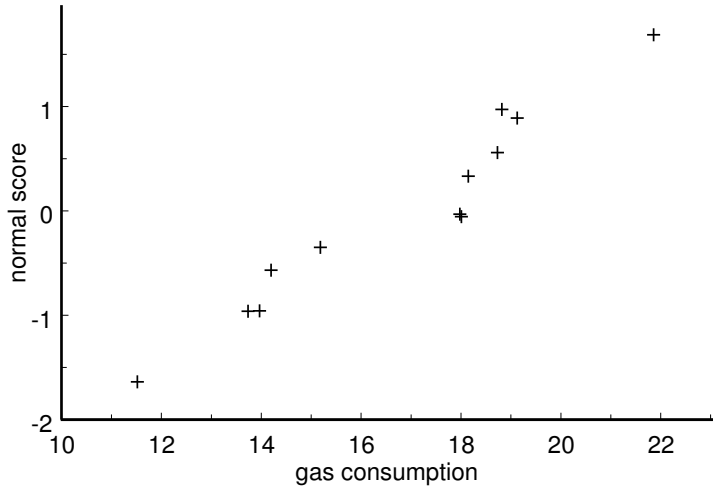
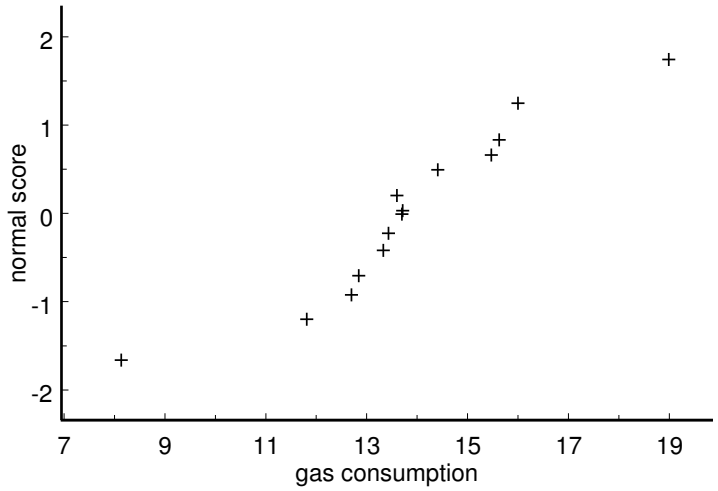


Figure 3. Normal probability plot for gas consumption for houses with extra insulation (both designs).



The summary statistics show some evidence of slight skewness to the left in the distribution for houses with standard insulation. Both stem and leaf histograms appear to be unimodal and reasonably symmetric with mild outliers on both sides. Both normal probability plots are reasonably linear. Thus it seems reasonable to model these data as independent random samples from normal distributions. Furthermore, the two sample standard deviations, 2.9959 and 2.3636, are quite similar; therefore, we can also reasonably assume that the two population standard deviations are equal.

Letting X_1 denote the annual gas consumption for a house with standard insulation and X_2 denote annual gas consumption for a house with extra insulation, we find that the difference in the sample means is $\overline{X}_1 - \overline{X}_2 = 16.7667 - 13.8714 = 3.8953$ (3,895.3 kWh)

suggesting that, among these 180 houses, the mean annual gas consumption for a house with standard insulation μ_1 is approximately 3.8953 thousand kW hours higher than the mean annual gas consumption for a house with extra insulation μ_2. The pooled sample standard deviation $S_p = 2.6720$, the standard error $\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2) = 1.0512$, and the margin of error multiplier $k = 2.064$ for the Student's t distribution with $n_1 + n_2 - 2 = 24$ degrees of freedom yield the 95% confidence interval $(.7258, 5.0648)$ for $\mu_1 - \mu_2$. Thus we are 95% confident that, among these 180 houses, the population mean annual gas consumption for a house with standard insulation is at least 725.8 kW hours and as much as 5,064.8 kW hours higher than the mean annual gas consumption for a house with extra insulation. Note that, technically, this inference is restricted to these 180 houses but we might conjecture that a similar difference would occur for similar houses (with standard and extra insulation) in this same area.

Remark regarding directional confidence bounds. *We can find an upper or lower 95% confidence bound for $\mu_1 - \mu_2$ by selecting the appropriate confidence limit from a 90% confidence interval estimate of $\mu_1 - \mu_2$.*

When $\sigma_1 = \sigma_2$, we can use the **two sample Student's t test statistic**

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2)} = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

based on the standard error computed using the pooled estimate of the common standard deviation, to test hypotheses relating μ_1 to μ_2.

First consider a situation where we want to determine whether there is sufficient evidence to conclude that the population mean μ_1 exceeds the population mean μ_2. Our research hypothesis is the contention that the population mean μ_1 exceeds the population mean μ_2, *i.e.*, $H_1 : \mu_1 > \mu_2$. The corresponding null hypothesis is $H_0 : \mu_1 \leq \mu_2$. Values of $\overline{X}_1 - \overline{X}_2$ which are large relative to zero provide evidence in favor of $H_1 : \mu_1 > \mu_2$, since this hypothesis is equivalent to $H_1 : \mu_1 - \mu_2 > 0$, and against $H_0 : \mu_1 \leq \mu_2$. Since large values of $\overline{X}_1 - \overline{X}_2$ yield large values of the Student's t statistic, we will reject $H_0 : \mu_1 \leq \mu_2$ in favor of $H_1 : \mu_1 > \mu_2$ if the calculated Student's t statistic is sufficiently large. This decision will hinge on the size of the P–value, which is the probability, computed under the assumption that $\mu_1 = \mu_2$, that $\overline{X}_1 - \overline{X}_2$ is as large or larger than the value actually observed and is equal to the probability that a Student's t variable with $n_1 + n_2 - 2$ degrees of freedom is as large or larger than the calculated t value T_{calc}. Notice that this P–value is the area to the right of T_{calc} under the density curve of the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom, since values of $\overline{X}_1 - \overline{X}_2$ that are sufficiently far above zero provide evidence in favor of the research hypothesis.

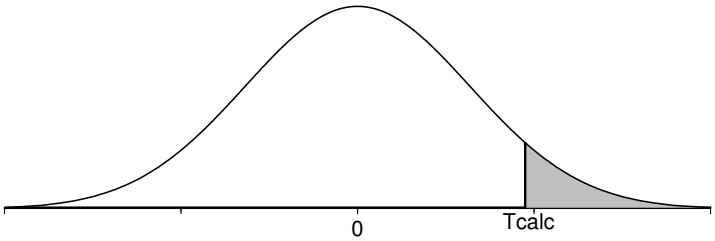
The steps for performing a hypothesis test for

$$H_0 : \mu_1 \leq \mu_2 \text{ versus } H_1 : \mu_1 > \mu_2$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(T \geq T_{calc})$, where T denotes a Student's t variable with $n_1 + n_2 - 2$ degrees of freedom and $T_{calc} = (\overline{X}_1 - \overline{X}_2)/\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2)$ as described above. This P–value is the area to the right of T_{calc} under the density curve for the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom as shown in Figure 4.

Figure 4. P–value for $H_0 : \mu_1 \leq \mu_2$ versus $H_1 : \mu_1 > \mu_2$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : \mu_1 > \mu_2$ over $H_0 : \mu_1 \leq \mu_2$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the first population mean μ_1 is greater than the second population mean μ_2.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : \mu_1 > \mu_2$ over $H_0 : \mu_1 \leq \mu_2$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the first population mean μ_1 is greater than the second population mean μ_2.

The procedure for testing the null hypothesis $H_0 : \mu_1 \leq \mu_2$ versus the research hypothesis $H_1 : \mu_1 > \mu_2$ given above is readily modified for testing the null hypothesis $H_0 : \mu_1 \geq \mu_2$ versus the research hypothesis $H_1 : \mu_1 < \mu_2$. The essential modification is to change the direction of the inequality in the definition of the P–value. Consider a situation where the research hypothesis specifies that the population mean μ_1 is less than the population mean μ_2. Values of $\overline{X}_1 - \overline{X}_2$ that are sufficiently far from 0 in the negative direction provide evidence in favor of the research hypothesis $H_1 : \mu_1 < \mu_2$ and against the null hypothesis $H_0 : \mu_1 \geq \mu_2$. Therefore, the appropriate P–value is the probability of observing a value of $\overline{X}_1 - \overline{X}_2$ as small or smaller than the value actually observed. As before, the P–value is computed under the assumption that $\mu_1 = \mu_2$. The calculated t statistic T_{calc} is defined as before; however, in this situation the P–value is the area to the

left of T_{calc} under the density curve of the Student's t distribution with n_1+n_2-2 degrees of freedom, since values of $\overline{X}_1 - \overline{X}_2$ that are sufficiently far below zero provide evidence in favor of the research hypothesis.

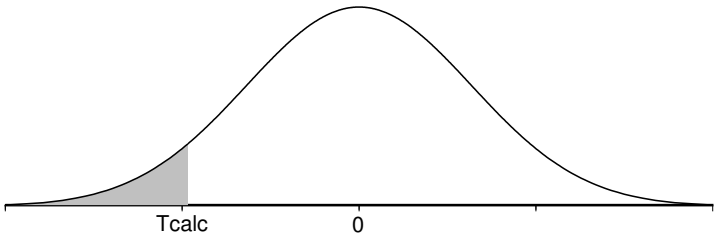
The steps for performing a hypothesis test for

$$H_0 : \mu_1 \geq \mu_2 \text{ versus } H_1 : \mu_1 < \mu_2$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(T \leq T_{calc})$, where T denotes a Student's t variable with $n_1 + n_2 - 2$ degrees of freedom and $T_{calc} = (\overline{X}_1 - \overline{X}_2)/\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2)$ as before. This P–value is the area to the left of T_{calc} under the density curve for the Student's t distribution with n_1+n_2-2 degrees of freedom as shown in Figure 5.

Figure 5. P–value for $H_0 : \mu_1 \geq \mu_2$ versus $H_1 : \mu_1 < \mu_2$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : \mu_1 < \mu_2$ over $H_0 : \mu_1 \geq \mu_2$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the first population mean μ_1 is less than the second population mean μ_2.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : \mu_1 < \mu_2$ over $H_0 : \mu_1 \geq \mu_2$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the first population mean μ_1 is less than the second population mean μ_2.

Example. Energy consumption (revisited). We will now consider the reduction in energy consumption due to extra insulation when the population is restricted to the houses among the 180 houses which have passive solar designs. Table 3 provides the annual gas consumption (in 1000 kWh) for two independent random samples of houses. One random sample was selected from all of the passive solar houses with standard insulation and the other was selected from all of the passive solar houses with extra insulation. Summary statistics are given in Table 4, stem and leaf histograms are given in Figure 6, and normal probability plots are provided in Figures 7 and 8.

Table 3. Gas consumption data (1000 kWh) (passive solar).

standard insulation								extra insulation					
12.3	13.3	13.7	13.8	14.9	15.6	15.9	16.3	10.5	11.3	11.4	12.6	13.0	14.5
16.5	17.2	17.5	17.6	17.8	17.9	18.0	19.9	15.2	15.7	15.7	17.6	19.0	

Table 4. Descriptive statistics for gas consumption (passive solar).

	standard	extra
minimum:	12.30	10.5
Q1:	14.35	11.4
median:	16.40	14.5
Q3:	17.70	15.7
maximum:	19.90	19.0
Q1 - minimum:	2.05	.9
median - Q1:	2.05	3.1
Q3 - median:	1.30	1.2
maximum - Q3:	2.20	3.3
mean:	16.1375	14.2273
standard deviation:	2.0791	2.7225
range:	7.6	8.5
IQ range:	3.35	4.3
sample size:	16	11

Figure 6. Stem and leaf histograms for gas consumption (passive solar).

In these stem and leaf histograms the stem represents ones and the leaf represents tenths. (1000 kWh)

standard	extra
	10 5
	11 34
12 3	12 6
13 378	13 0
14 9	14 5
15 69	15 277
16 35	16
17 25689	17 6
18 0	18
19 9	19 0

Figure 7. Normal probability plot for gas consumption for houses with standard insulation (passive solar).

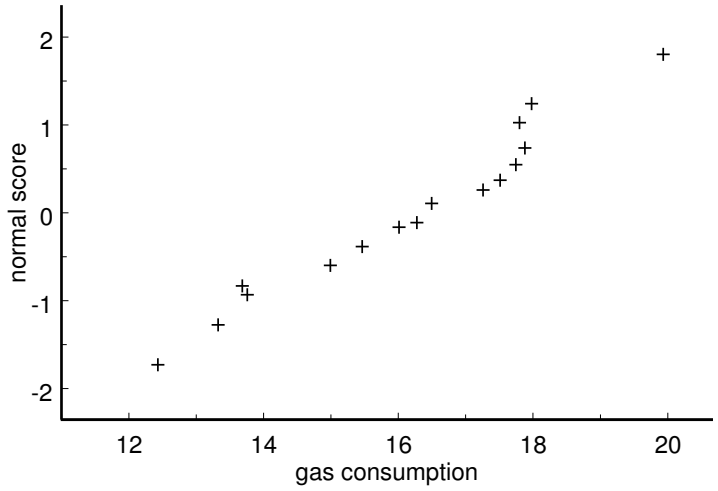
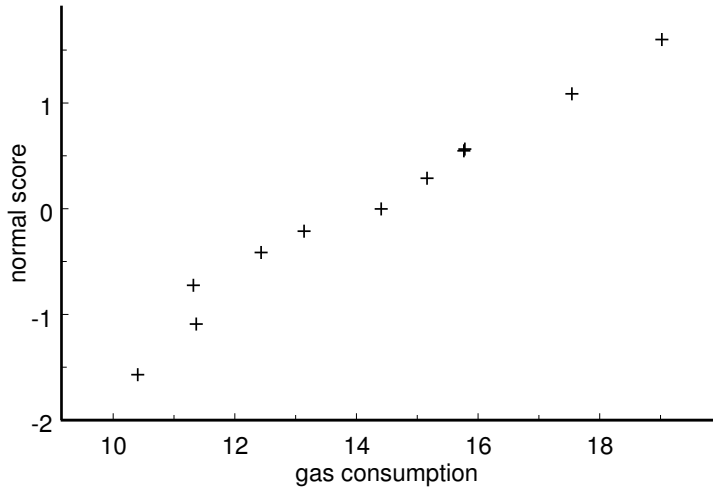


Figure 8. Normal probability plot for gas consumption for houses with extra insulation (passive solar).



In this case both stem and leaf histograms appear to be unimodal and reasonably symmetric. The summary statistics support these claims and both normal probability plots are reasonably linear. Thus it seems reasonable to model these data as independent random samples from normal distributions. The two sample standard deviations, 2.0791 and 2.7225, are quite similar; therefore, we can also reasonably assume that the two population standard deviations are equal.

Let X_1 denote the annual gas consumption for a passive solar house with standard insulation and let X_2 denote annual gas consumption for a passive solar house with extra insulation. Similarly, let μ_1 and μ_2 denote the respective population means for all of the passive solar houses among the 180 houses with standard and extra insulation. The

obvious research hypothesis $H_1 : \mu_1 > \mu_2$ states that among the passive solar houses in this development, on average, the annual gas consumption is lower for a house with extra insulation than it is for a house with standard insulation. For these data the pooled sample standard deviation is $S_p = 2.3576$, the standard error is $\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2) = .9234$, the observed value of the Student's t statistic is $T_{calc} = 2.07$ with 25 degrees of freedom, and the corresponding P–value is .0245. This P–value is reasonably small indicating that there is reasonably strong evidence that μ_1 is greater than μ_2. Therefore, there is reasonably strong evidence that for this population of passive solar houses, on average, the annual gas consumption for a passive solar house with extra insulation is lower than the annual gas consumption for a passive solar house with standard insulation. We can form a 95% confidence interval for $\mu_1 - \mu_2$ to get a feel for the practical importance of this result. Using the margin of error multiplier $k = 2.060$ for the Student's t distribution with 25 degrees of freedom yields the 95% confidence interval $(.0084, 3.8120)$ for $\mu_1 - \mu_2$. Thus we are 95% confident that, among this population of passive solar houses, the population mean annual gas consumption for a house with standard insulation is between 8.4 kW hours and 3,812 kW hours higher than the mean annual gas consumption for a house with extra insulation. Notice that this confidence interval estimate indicates that the difference between these means might be as small as 8.4 kW hours which is not much of a difference. Of course, the confidence interval estimate also allows that the difference in these means might be as large as 3,812 kW hours which is more impressive. In this case, technically, our inferences are restricted to all of the passive solar houses among these 180 houses.

Example. Paspalum grass. This example is taken from Seber (1984), *Multivariate Observations*, Wiley, New York. (The data were provided by Peter Buchanan.) Paspalum grass is a weed which grows in pastures used for grazing farm animals. Scientists at the Mount Albert Research Centre in Auckland conducted a laboratory experiment to determine whether inoculation of paspalum with a fungal infection might be effective in reducing the growth of this weed. The experimenters randomly assigned 48 pots of paspalum to the 8 combinations of treatment (inoculated, not inoculated) and temperature (14, 18, 22, 26 degrees C). For our purposes we will restrict our attention to the 24 pots of plants grown under moderate temperatures (18 or 22 degrees) and we will not distinguish between the two temperatures. Thus we have two samples of size 12. The experimenters measured several characteristics of the paspalum. The response variable we will consider is the fresh weight of the roots (in grams) of the paspalum in a pot. (In this example a pot of paspalum is a unit; the number of plants per pot is not specified.) Table 5 provides the fresh root weights for the 12 pots assigned to each treatment. Summary statistics are given in Table 6, stem and leaf histograms are given in Figure 9, and normal probability plots are provided in Figures 10 and 11.

Table 5. Paspalum root weight (grams).

inoculated						not inoculated					
3.9	4.3	4.9	5.2	6.5	7.6	6.2	8.7	11.0	12.2	12.3	13.1
9.6	10.0	10.1	12.3	13.6	19.7	13.6	14.5	15.4	16.4	16.7	21.8

Table 6. Descriptive statistics for paspalum root weight.

	inoculated	not inoculated
minimum:	3.90	6.20
Q1:	5.05	11.60
median:	8.60	13.35
Q3:	11.20	15.90
maximum:	19.70	21.80
Q1 - minimum:	1.15	5.40
median - Q1:	3.55	1.75
Q3 - median:	2.60	2.55
maximum - Q3:	8.50	5.90
mean:	8.9750	13.4917
standard deviation:	4.6384	4.0230
range:	15.8	15.6
IQ range:	6.15	4.3
sample size:	12	12

Figure 9. Stem and leaf histograms for paspalum root weight.

In these stem and leaf histograms the stem represents tens and the leaf represents ones. The data are rounded. (grams)

inoculated	not inoculated
0 3	
0 445	
0 67	0 6
0 9	0 8
1 00	1 1
1 23	1 2233
1	1 45
1	1 66
1 9	1
	2 1

Figure 10. Normal probability plot for root weight (inoculated).

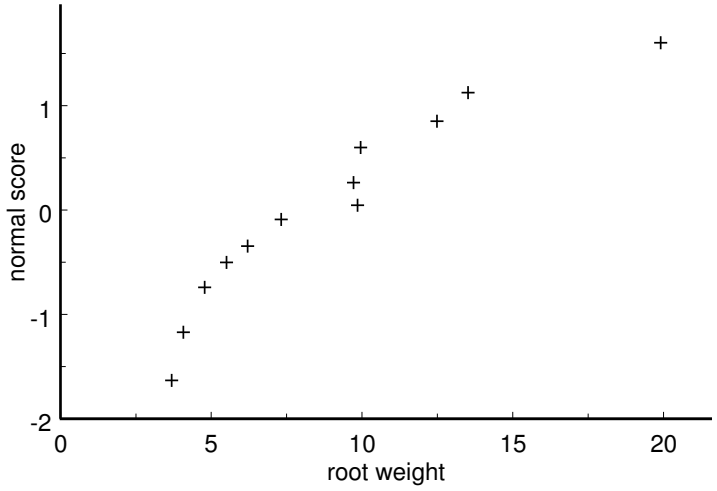
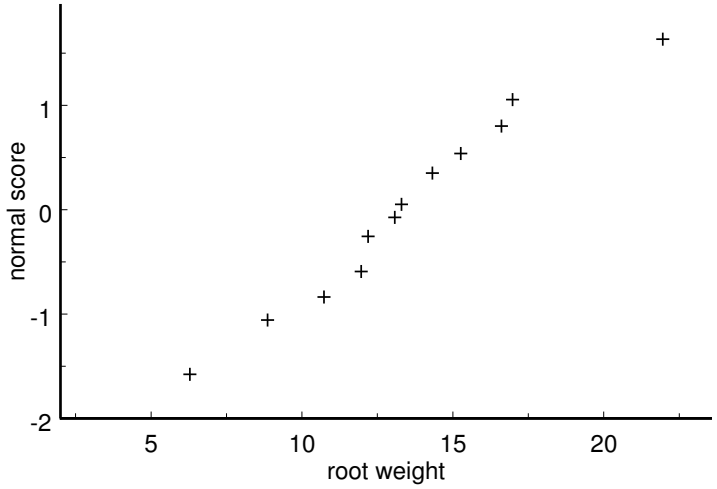


Figure 11. Normal probability plot for root weight (not inoculated).



Let X_1 denote the fresh root weight for a pot of paspalum inoculated with the fungus and let X_2 denote the fresh root weight for a pot of paspalum not inoculated with the fungus. We can think of the corresponding population means μ_1 and μ_2 as the mean fresh root weights we would observe if all 48 of the pots of paspalum had been inoculated (μ_1) or not inoculated (μ_2). We want to determine whether there is sufficient evidence to claim that inoculation with this fungus retards the growth of paspalum in the sense of reducing fresh root weight. In terms of the population means the research hypothesis $H_1 : \mu_1 < \mu_2$ states that, for this collection of 48 pots of paspalum, on average, the fresh root weight would be smaller if the paspalum was inoculated with the fungus than it would be if the paspalum was not inoculated.

Both of the stem and leaf histograms are unimodal and both show some evidence of slight skewness to the right. Each sample contains a mild outlier (19.7 for the inoculated

sample and 21.8 for the not inoculated sample). The summary statistics indicate that it is these outliers which give the impression of skewness to the right. The normal probability plots are reasonably linear suggesting that skewness is not a problem. Thus it seems reasonable to model these data as independent random samples from normal distributions. The two sample standard deviations, 4.6384 and 4.0230, are quite similar; therefore, we can also reasonably assume that the two population standard deviations are equal.

For these data the pooled sample standard deviation is $S_p = 4.3416$, the standard error is $\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2) = 1.7725$, the observed value of the Student's t statistic is $T_{calc} = -2.55$ with 22 degrees of freedom, and the corresponding P–value is .0092. This P–value is very small indicating that there is very strong evidence that μ_1 is less than μ_2. Therefore, there is very strong evidence that for this collection of 48 pots of paspalum, on average, the fresh root weight would be smaller if the paspalum was inoculated with the fungus than it would be if the paspalum was not inoculated.

Using the margin of error multiplier $k = 2.074$ for the Student's t distribution with 22 degrees of freedom yields the 95% confidence interval $(-8.193, -.8410)$ for $\mu_1 - \mu_2$. Thus we are 95% confident that, for this collection of 48 pots of paspalum, the mean fresh root weight we would observe if all 48 of the pots of paspalum had been inoculated is between .8410 grams and 8.1930 grams smaller than the mean fresh root weight we would observe if none of the 48 of the pots of paspalum had been inoculated.

The directional hypothesis tests we discussed above are readily modified for testing a nondirectional hypothesis. To decide between the null hypothesis $H_0 : \mu_1 = \mu_2$ and the research hypothesis $H_1 : \mu_1 \neq \mu_2$, we need to decide whether $\overline{X}_1 - \overline{X}_2$ supports the null hypothesis by being "close to 0", or supports the research hypothesis by being "far away from 0". In this situation the P–value is the probability that $\overline{X}_1 - \overline{X}_2$ would be as far or farther away from 0 in either direction as is the value that we actually observe. In other words, the P–value is the probability that the distance $|\overline{X}_1 - \overline{X}_2|$ between the two sample means (the absolute value of the difference between $\overline{X}_1$ and $\overline{X}_2$) is as large or larger than the actual observed value of this distance. As before, the P–value is computed under the assumption that the null hypothesis is true and $\mu_1 = \mu_2$. In this situation the calculated t statistic T_{calc} is the absolute value of the t statistic that would be used for testing a directional hypothesis. That is, the calculated t statistic is

$$T_{calc} = \frac{|\overline{X}_1 - \overline{X}_2|}{\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2)}.$$

In terms of this t statistic the P–value is the probability that the absolute value of a Students's t variable with $n_1 + n_2 - 2$ degrees of freedom would take on a value as large or larger than T_{calc}, computed assuming that $\mu_1 = \mu_2$. This probability is the sum of the

area under the appropriate Student's t density curve to the left of $-T_{calc}$ and the area under this Student's t density curve to the right of T_{calc}. We need to add these two areas (probabilities) since we are finding the probability that $\overline{X}_1 - \overline{X}_2$ would be as far or farther away from 0 in either direction as is the value that we actually observe, when $\mu_1 = \mu_2$.

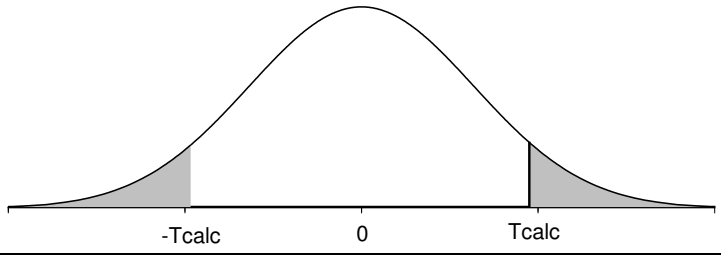
The steps for performing a hypothesis test for

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(|T| \geq T_{calc}) = P(T \leq -T_{calc}) + P(T \geq T_{calc})$, where T denotes a Student's t variable with $n_1 + n_2 - 2$ degrees of freedom and
$$T_{calc} = \frac{|\overline{X}_1 - \overline{X}_2|}{\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2)}.$$
Notice that this calculated t value is the absolute value of the calculated t value we would use for a directional hypothesis. This P–value is the area, under the density curve for the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom, to the left of $-T_{calc}$ plus the area to the right of T_{calc} as shown in Figure 12.

Figure 12. P–value for $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : \mu_1 \neq \mu_2$ over $H_0 : \mu_1 = \mu_2$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the first population mean μ_1 and the second population mean μ_2 are different.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : \mu_1 \neq \mu_2$ over $H_0 : \mu_1 = \mu_2$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population means μ_1 and μ_2 are different.

Example. Fecundity of fruitflies. Sokal, R.R. and Rohlf, F.J. (1969) *Biometry*, W.H. Freeman, p.232, discuss a study conducted to compare the fecundity of three genetic lines of *Drosophila melanogaster*. The data in Table 7 consist of per diem fecundities (number of eggs laid per female per day for the first 14 days of life) for 25 females of three

lines of *Drosophila melanogaster*. Two of these genetic lines were selected for resistance (RS) and susceptibility (SS) to DDT, the third line is a nonselected control (NS). These data can be used to address two questions which were of interest to the investigator. We can use the data for the two selected lines (RS and SS) to determine if there is evidence that the mean fecundity differs for these selected lines. We can then use the data for the control line (NS) to compare the mean fecundity of the control line with that of the two selected lines. For the time being we will use two–sample Student's t tests to address these questions. We consider an alternate approach to this problem in Chapter 12.

Table 7. Fruitfly fecundity data.

resistant RS		**susceptible SS**		**nonselected NS**	
12.8	22.4	38.4	23.1	35.4	22.6
21.6	27.5	32.9	29.4	27.4	40.4
14.8	20.3	48.5	16.0	19.3	34.4
23.1	38.7	20.9	20.1	41.8	30.4
34.6	26.4	11.6	23.3	20.3	14.9
19.7	23.7	22.3	22.9	37.6	51.8
22.6	26.1	30.2	22.5	36.9	33.8
29.6	29.5	33.4	15.1	37.3	37.9
16.4	38.6	26.7	31.0	28.2	29.5
20.3	44.4	39.0	16.9	23.4	42.4
29.3	23.2	12.8	16.1	33.7	36.6
14.9	23.6	14.6	10.8	29.2	47.4
27.3		12.2		41.7	

Figure 13. Stem and leaf histograms for fruitfly fecundity.

In these stem and leaf histograms the stem represents tens and the leaf represents ones. The data are rounded.

resistant (RS)	**susceptible (SS)**	**nonselected (NS)**
1 3	1 1223	1
1 556	1 55667	1 59
2 0002233344	2 0122333	2 033
2 66789	2 79	2 789
3 00	3 0133	3 00444
3 599	3 89	3 577788
4 4	4	4 0222
4	4 8	4 7
5	5	5 2

Let X_{RS} denote the fecundity for an RS female, X_{SS} the fecundity for an SS female, and X_{NS} the fecundity for an NS female; and let μ_{RS}, μ_{SS}, and μ_{NS} denote the corresponding population means. The first question, concerning the relationship between the population mean fecundities μ_{RS} and μ_{SS}, can be addressed via a test of $H_0 : \mu_{RS} = \mu_{SS}$ versus $H_1 : \mu_{RS} \neq \mu_{SS}$. Our approach to the second question will depend on our conclusion for the first. If we decide that there is no difference between the two selected line population mean fecundities ($\mu_{RS} = \mu_{SS}$), then we can combine the data for these two lines and, viewing this as a random sample from a population of selected lines with population mean μ_S, we can test for a difference between the population mean for selected lines and the population mean for the nonselected line by testing $H_0 : \mu_S = \mu_{NS}$ versus $H_1 : \mu_S \neq \mu_{NS}$. On the other hand, if we decide that there is a difference between the population mean fecundities for the two selected lines, then we will need to perform two tests; one for comparing μ_{RS} to μ_{NS} and another for comparing μ_{SS} to μ_{NS}.

Table 8. Descriptive statistics for fruitfly fecundity.

	resistant (RS)	**susceptible (SS)**	**nonselected (NS)**
minimum:	12.8	10.8	14.9
Q1:	20.3	16.0	28.2
median:	23.6	22.5	34.4
Q3:	29.3	30.2	37.9
maximum:	44.4	48.5	51.8
Q1 - minimum:	7.5	5.2	13.3
median - Q1:	3.3	6.5	6.2
Q3 - median:	5.7	7.7	3.5
maximum - Q3:	15.1	18.3	13.9
mean:	25.2560	23.6280	33.3720
standard deviation:	7.7724	9.7685	8.9420
range:	31.6	37.7	36.9
IQ range:	9.0	14.2	9.7
sample size:	25	25	25

The stem and leaf histograms in Figure 13 and the information in Table 8 indicate that the fecundity distributions for the two selected lines (RS and SS) are unimodal with some evidence of skewness to the right; and the fecundity distribution for the nonselected line (NS) is unimodal and reasonably symmetric with slight evidence of skewness to the left in the middle of the distribution. The normal probability plots in Figures 14, 15, and 16 are reasonably linear. Thus it seems reasonable to treat these samples as forming independent random samples from normal populations. The three sample standard deviations are reasonably similar allowing us to also assume a common population standard deviation.

Figure 14. Normal probability plot fruitfly data (resistant line).

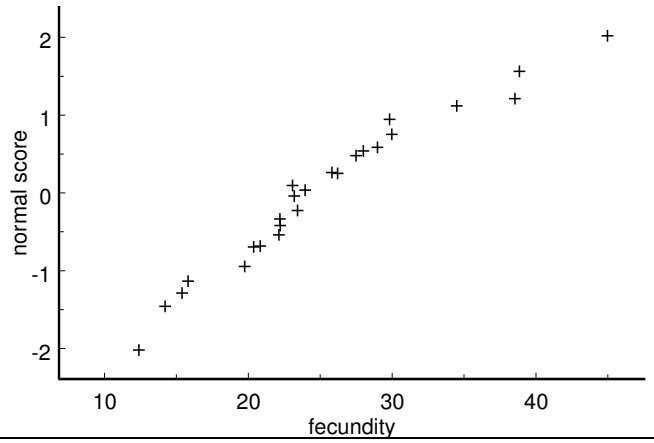


Figure 15. Normal probability plot fruitfly data (susceptible line).

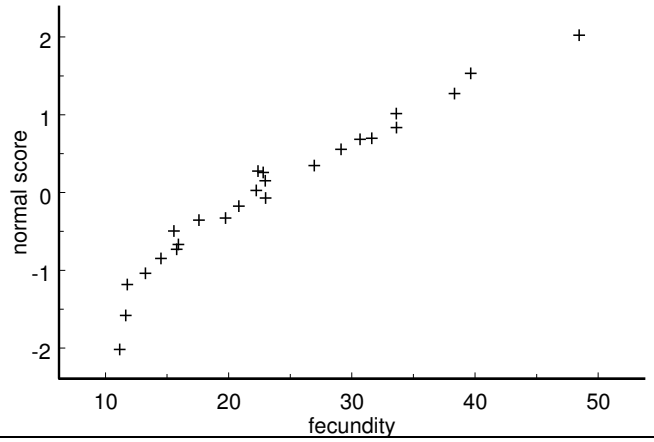
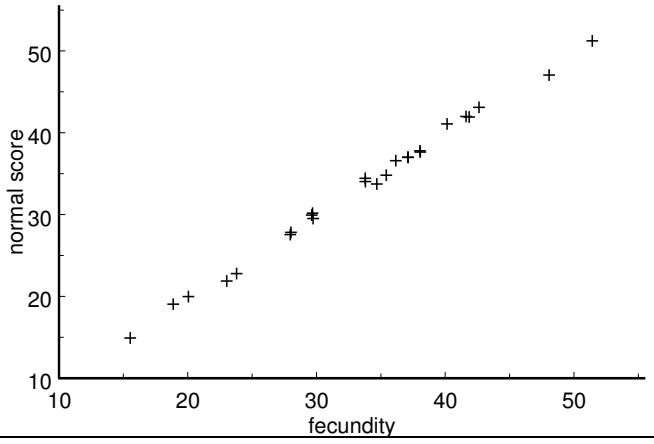


Figure 16. Normal probability plot fruitfly data (nonselected line).



Therefore, we will model the three population distributions as normal distributions with respective population means μ_{RS}, μ_{SS} and μ_{NS} and with common population standard deviation. If we decide to combine the samples from the two selected lines, we will model

the corresponding population distribution as a normal distribution with population mean μ_S and the same common population standard deviation as before.

The difference between the sample mean fecundities for the two selected lines $\overline{X}_{RS} - \overline{X}_{SS} = 1.628$ is small relative to the corresponding standard error $\widehat{\text{S.E.}}(\overline{X}_{RS} - \overline{X}_{SS}) = 2.4967$ suggesting that there is little evidence of a difference between the population means μ_{RS} and μ_{SS}. The observed value of the Student's t statistic for testing $H_0 : \mu_{RS} = \mu_{SS}$ versus $H_1 : \mu_{RS} \neq \mu_{SS}$ is $T_{calc} = .65$ with 48 degrees of freedom, and the corresponding P–value is .5175. This large P–value allows us to conclude that the two population mean fecundities μ_{RS} and μ_{SS} are equal. In light of this conclusion we will now combine the samples for the selected lines as described above and test $H_0 : \mu_S = \mu_{NS}$ versus $H_1 : \mu_S \neq \mu_{NS}$. Recall that μ_S denotes the population mean fecundity for the population of fruitflies obtained by combining the populations for the two selected lines. The difference between the sample mean fecundities for the combined population of selected lines and the nonselected line is $\overline{X}_S - \overline{X}_{NS} = -8.93$ with an associated standard error of $\widehat{\text{S.E.}}(\overline{X}_S - \overline{X}_{NS}) = 2.163$. The observed value of the Student's t statistic for testing $H_0 : \mu_S = \mu_{NS}$ versus $H_1 : \mu_S \neq \mu_{NS}$ is $T_{calc} = -4.13$ with 73 degrees of freedom and a corresponding P–value which is less than .0001. This P–value is quite small indicating that there is very strong evidence that the population mean fecundity for the selected lines μ_S is different from the population mean fecundity μ_{NS} for the nonselected line. The data clearly support the conclusion that the population mean fecundity is higher for the nonselected line, however, technically speaking, we cannot make this conclusion based on the preceding hypothesis test, since we did not have *a priori* reason to justify a directional hypothesis. We can however form a confidence interval for $\mu_{NS} - \mu_S$ and use it to justify this conclusion. In this example, we are 95% confident that $\mu_{NS} - \mu_S$ is between 4.6192 and 13.241. More precisely we are 95% confident that the population mean fecundity (mean number of eggs laid per day for the first 14 days of life) μ_{NS} for the nonselected line exceeds the population mean fecundity μ_S for the selected lines by at least 4.6192 eggs per day and perhaps as much as 13.241 eggs per day. Thus it appears that the population of fruitflies which are either resistant to or susceptible to DDT has lower fecundity on average than the population of fruitflies which are neither resistant nor susceptible to DDT.

Remark regarding the comparison of the difference of two means to a nonzero constant. In some situations we may have enough *a priori* **information to specify a known constant** d **with the goal of comparing the difference** $\mu_1 - \mu_2$ **to this particular constant.** *For example, we might hypothesize that the first population mean* μ_1 *exceeds the second population mean* μ_2 *by more than* $d = 2$ *units, i.e.,* $H_1 : \mu_1 - \mu_2 > 2$ *or* $H_1 : \mu_1 > \mu_2 + 2$. *To test such a hypothesis we simply replace the difference* $\overline{X}_1 - \overline{X}_2$ *by the quantity* $\overline{X}_1 - \overline{X}_2 - d$ *in the formula for* T *and proceed as before. Many computer programs provide an option for testing such a hypothesis.*

8.2b Inference when the two population standard deviations are not equal

In this subsection we will describe an alternate method of inference which can be used when the population standard deviations σ_1 and σ_2 are not equal. Notice that when $\sigma_1 \neq \sigma_2$ the two normal populations are not identical when their population means, μ_1 and μ_2, are equal. Therefore, a statement regarding the difference between two population means does not tell the whole story about the relationship between the corresponding normal populations when the population standard deviations are not equal. This does not indicate that there is anything wrong with comparing population means when the corresponding population standard deviations are unequal. However, it does indicate that the interpretation of a particular difference between two population means is somewhat different when the population standard deviations are different than it is when the population standard deviations are equal.

When the population standard deviations σ_1 and σ_2 are different, the appropriate estimator of the standard error of $\overline{X}_1 - \overline{X}_2$ is based on the two sample standard deviations S_1 and S_2 rather than the pooled sample standard deviation. That is, when $\sigma_1 \neq \sigma_2$ the appropriate **sample standard error** of $\overline{X}_1 - \overline{X}_2$ is

$$\widehat{\text{S.E.}}(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where S_1 is the sample standard error for the sample from the first population (the X_1 values) and S_2 is the sample standard error for the sample from the second population (the X_2 values).

Inference about the relationship between two normal population means when $\sigma_1 \neq \sigma_2$ is based on an approximation to the sampling distribution of the quantity

$$T^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Because the details of this approximation are fairly complicated, you really need an appropriate calculator or computer program to implement this method.

Using this method the 95% **margin of error of** $\overline{X}_1 - \overline{X}_2$ is

$$\text{M.E.}(\overline{X}_1 - \overline{X}_2) = k\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where k is the 97.5 percentile of a Student's t distribution with ν degrees of freedom. The relevant degrees of freedom ν is computed using a complex formula which may yield a value that is not a whole number. An approximate 95% confidence interval for $\mu_1 - \mu_2$

based on this approach is given by the values between $(\overline{X}_1 - \overline{X}_2) - \text{M.E.}(\overline{X}_1 - \overline{X}_2)$ and $(\overline{X}_1 - \overline{X}_2) + \text{M.E.}(\overline{X}_1 - \overline{X}_2)$, where the margin of error is as given above. A suitable calculator or computer program will provide the calculated value of this margin of error or the actual 95% confidence interval values.

To test a hypothesis relating μ_1 to μ_2 using this method we simply replace the Student's t statistic T by the approximate Student's t statistic T^* and compute the P–value using the appropriate degrees of freedom ν. A suitable calculator or computer program will provide the calculated value of the approximate t statistic T^*_{calc} and the associated P–value.

One way to determine whether the assumption of a common population standard deviation is reasonable is to compare the results of the confidence intervals and P–values computed assuming equal standard deviations and not assuming equal standard deviations. If the two methods yield essentially the same conclusions, then the assumption of equal standard deviations is reasonable and the methods based on the pooled estimate of the standard error are appropriate; otherwise, the methods which do not use the pooled estimate of the standard error should be used.

8.3 Inference based on ranks

The inferential methods for comparing two population means discussed above require at least approximate normality of the population distributions of the variables of interest. In this section we will consider methods for making inferences about two population means which do not require the assumption of a particular form for the population distributions of the variables of interest. The methodology we are about to discuss is based on the location shift assumption described in the introduction.

As before we will assume that the data comprise two independent random samples; a random sample of size n_1 from a population of values of a continuous variable X_1 with population mean μ_1 and a random sample of size n_2 from a population of values of a continuous variable X_2 with population mean μ_2. We will also assume that the shift assumption holds meaning that the only difference between these two population distributions is a possible difference in location, *i.e.*, we will assume that the population distributions (density curves) of X_1 and X_2 are identical except for a possible difference between the population means μ_1 and μ_2. We will make no further assumptions about the exact form of this common density curve.

We can look for evidence of a location shift by examining the locations of the n_1 observed values of X_1 relative to the locations of the n_2 observed values of X_2. If there is no location shift, then, by assumption, the population distributions of X_1 and X_2 are identical (and consequently $\mu_1 = \mu_2$) and we would expect the n_1 observed values of X_1 to be randomly dispersed among the n_2 observed values of X_2. On the other hand, if the

density curve for X_1 is located to the right of the density curve for X_2 (the distribution of X_1 is shifted to the right of the distribution of X_2 and consequently $\mu_1 > \mu_2$), then we would expect the observed values of X_1 to tend to be large relative to the observed values of X_2. Similarly, if the density curve for X_1 is located to the left of the density curve for X_2 (the distribution of X_1 is shifted to the left of the distribution of X_2 and consequently $\mu_1 < \mu_2$), then we would expect the observed values of X_1 to tend to be small relative to the observed values of X_2.

We can quantify the locations of the n_1 observed values of X_1 relative to the locations of the n_2 observed values of X_2 by assigning ranks to these $N = n_1 + n_2$ observations. We first combine the n_1 observed values of X_1 with the n_2 observed values of X_2, keeping track of which observations form the X_1 sample and which form the X_2 sample. We then order these $N = n_1 + n_2$ observations from smallest to largest and assign them ranks; the smallest observation having rank 1, the next rank 2, and so on with the largest observation having rank $N = n_1 + n_2$. Finally, we separate these ranks into the group of n_1 ranks of the X_1 sample and the group of n_2 ranks of the X_2 sample.

Let $\overline{R}_1$ and $\overline{R}_2$ denote the respective sample means of the ranks of the X_1 sample and the X_2 sample. Restating the remarks from above in terms of the ranks yields the following. If $\mu_1 = \mu_2$, then we would expect the X_1 ranks to look like a simple random sample of size n_1 selected without replacement from the set of all possible ranks $\{1, 2, \ldots, N\}$ with the remaining n_2 ranks constituting the X_2 ranks; and, we would expect $\overline{R}_1$ and $\overline{R}_2$ to be similar. If $\mu_1 > \mu_2$, then as a group we would expect the X_1 ranks to be large relative to the X_2 ranks and we would expect $\overline{R}_1$ to be large relative to $\overline{R}_2$. If $\mu_1 < \mu_2$, then as a group we would expect the X_1 ranks to be small relative to the X_2 ranks and we would expect $\overline{R}_1$ to be small relative to $\overline{R}_2$. These facts suggest that we can perform a test of a hypothesis relating μ_1 to μ_2 on the basis of the ranks of the two samples instead of the actual data. In particular, we can base a hypothesis test on a suitably standardized version of the difference, $\overline{R}_1 - \overline{R}_2$, between the means of the two sets of ranks. For example, we would view a sufficiently large positive value of $\overline{R}_1 - \overline{R}_2$ as evidence in favor of the research hypothesis that $\mu_1 > \mu_2$.

It is possible to determine the exact sampling distribution of $\overline{R}_1 - \overline{R}_2$; however, using this exact sampling distribution to compute the relevant P–value requires a computer program or an extensive set of tables. The hypothesis test we are about to describe is known as the rank–sum test, the Wilcoxon rank–sum test, and the two–sample Mann–Whitney test. If you have a access to a computer statistics package, check for the availability of this procedure under one of these names. If a computer program is not available, a simple alternative is to use the two sets of ranks (the n_1 ranks of the X_1 sample and the n_2 ranks of the X_2 sample) as input for a two–sample Student's t test as described in Section 8.2a

and below. That is, we can use a suitable calculator or computer program to compute the relevant P–value corresponding to the calculated t statistic

$$T_{calc} = \frac{\overline{R}_1 - \overline{R}_2}{\widehat{\text{S.E.}}(\overline{R}_1 - \overline{R}_2)}$$

for a test of a directional hypothesis and the absolute value of this quantity for a test of a nondirectional hypothesis, where $\widehat{\text{S.E.}}(\overline{R}_1 - \overline{R}_2)$ is computed using the pooled sample standard deviation S_p, based on the ranks, with $n_1 + n_2 - 2$ degrees of freedom. This two–sample t test based on the ranks provides a large sample size (both n_1 and n_2 reasonably large) approximation to the test based on the exact sampling distribution of $\overline{R}_1 - \overline{R}_2$.

Example. This example is provided to clarify the method of ranking and the computations described above. Two artificial samples of sizes $n_1 = 13$ and $n_2 = 13$ are provided. From the stem and leaf histograms given in Figure 17 we see that the shift assumption is reasonable for these data.

Figure 17. Stem and leaf histograms for the hypothetical data.

In these stem and leaf histograms the stem represents tens and the leaf represents ones.

$\mathbf{X}_1$ data	$\mathbf{X}_2$ data
1 01679	
2 1578	2 02469
3 16	3 2478
4 2	4 14
5 1	5 2
	6 1

The ordered data values and corresponding ranks are shown in Table 9. The sample means of these ranks are $\overline{R}_1 = 10.4615$ and $\overline{R}_2 = 16.5385$, the pooled estimated standard deviation is $S_p = 7.1369$, and the estimated standard error of $\overline{R}_1 - \overline{R}_2$ is 2.7993. The calculated t statistic, for a directional hypothesis, is $T_{calc} = -2.1708$ with 24 degrees of freedom. The P–value for $H_1 : \mu_1 \neq \mu_2$ is .0400, the P–value for $H_1 : \mu_1 < \mu_2$ is .0200, and the P–value for $H_1 : \mu_1 > \mu_2$ is .9800.

The Minitab and S–Plus computer programs, which use the exact sampling distribution or a slightly different large sample approximation to this sampling distribution, give P–values for $H_1 : \mu_1 \neq \mu_2$ of .0455 and .0441, respectively, and P–values for $H_1 : \mu_1 < \mu_2$ of .0228 and .0220, respectively. Therefore, at least for this example, it seems that the method we have proposed (using the two–sample Student's t test based on the ranks) and these alternative methods give essentially the same P–values.

Table 9. The ordered data and corresponding ranks.

$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{R}_1$	$\mathbf{R}_2$
10		1			29		14
11		2		31		15	
16		3			32		16
17		4			34		17
19		5		36		18	
	20		6		37		19
21		7			38		20
	22		8		41		21
	24		9	42		22	
25		10			44		23
	26		11	51		24	
27		12			52		25
28		13			61		26

In the preceding discussion we implicitly assumed that the combined data consisted of $N = n_1 + n_2$ distinct values. In practice some observed values may occur more than once in the combined data listing. When there are repetitions or "ties" in the data it is not clear how we should assign the ranks to these tied values. The usual approach is to assign the average of the relevant ranks to all of the observations which are tied at a particular value. An example with hypothetical data is provided below to demonstrate the assignment of ranks when there are ties.

Table 10. The ordered data and corresponding ranks.

$\mathbf{X}_1$	$\mathbf{X}_2$	**tie**	$\mathbf{R}_1$	$\mathbf{R}_2$
5			1	
	6			2
9			3	
10		*	5	
10		*	5	
	10	*		5
	11			7
12			8	
	13			9
14		*	10.5	
	14	*		10.5
17			12	
	18			13

Example. For the hypothetical data in Table 10 with $n_1 = 7$, $n_2 = 6$, there are three observations tied at 10, and there are two observations tied at 14. The ranks corresponding to the three 10's are 4, 5, and 6 which average to 5, thus, we assign each of these observation a rank of 5. Similarly, the ranks corresponding to the two 14's are 10 and 11, thus, we assign each of these observations a rank of 10.5.

Example. Cowbird parasitization of flycatchers. Brown–headed cowbirds search for and lay their eggs in nests built by the willow flycatcher. It is theorized that those flycatchers that recognize but do not vocally react to cowbird calls are more apt to defend their nests and less likely to be found and parasitized by the cowbirds. A study published in *The Condor*, May, 1995, yielded the data regarding 13 active flycatcher nests given in Table 11. Each active flycatcher nest was classified as parasitized (if at least one cowbird egg was present) or not parasitized. Tapes of cowbird songs were played while the flycatcher pairs were sitting in the nest prior to incubation. The vocalization rate (measured as the number of calls per minute) of each flycatcher pair was recorded. According to the theory mentioned above we would expect the vocalization rate to be higher for the parasitized group.

Table 11. Cowbird vocalization data.

parasitized				**not parasitized**			
2.00	1.25	8.50	1.10	1.00	1.00	0	3.25
1.25	3.75	5.50		1.00	.25		

The stem and leaf histograms in Figure 18 both appear to be skewed right and each distribution possesses at least one unusually large value. Therefore, the assumption that the underlying population distributions are normal is not reasonable. However, the assumption that the underlying population distributions differ only in a shift of location is reasonable. As in Table 12, let X_1 denote the vocalization rate for a parasitized flycatcher pair and let X_2 denote the vocalization rate for a non–parasitized flycatcher pair. Furthermore, let μ_1 denote the population mean vocalization rate for the population of all parasitized flycatcher pairs and let μ_2 denote the population mean vocalization rate for the population of all non–parasitized flycatcher pairs. We can formalize the theory from above as the research hypothesis $H_1 : \mu_1 > \mu_2$ indicating that the population mean vocalization rate for the population of all parasitized flycatcher pairs, μ_1, is greater than the population mean vocalization rate for the population of all non–parasitized flycatcher pairs, μ_2.

Figure 18. Stem and leaf histograms for the cowbird data.
In these stem and leaf histograms the stem represents ones and the two digit leaf represents hundredths.

parasitized "X_1" data	**not parasitized "X_2" data**
0	0 00.25
1 10.25.25	1 00.00.00
2 00	2
3 75	3 25
4	
5 50	
6	
7	
8 50	

Table 12. Ordered cowbird data and corresponding ranks.

X_1	X_2	tie	R_1	R_2
	0			1
	.25			2
	1	*		4
	1	*		4
	1	*		4
1.10			6	
1.25		#	7.5	
1.25		#	7.5	
2			9	
	3.25			10
3.75			11	
5.5			12	
8.5			13	

Using the X_1 ranks and the X_2 ranks as the input for a Student's t test yields the calculated t statistic $T_{calc} = 3.3056$ and the P–value .0035. Since this P–value is very small there is strong evidence that the population mean vocalization rate for the population of all parasitized flycatcher pairs, μ_1, is greater than the population mean vocalization rate for the population of all non–parasitized flycatcher pairs, μ_2.

In a situation where we wish to compare the difference $\mu_1 - \mu_2$ to a particular, *a priori* constant value d we first note that a hypothesis relating $\mu_1 - \mu_2$ to d can be re–expressed

as a hypothesis relating $\mu_1 - d$ to μ_2. For example, the three standard research hypotheses have the equivalent forms listed below

$$H_1 : \mu_1 - \mu_2 > d \text{ is equivalent to } H_1 : \mu_1 - d > \mu_2;$$

$$H_1 : \mu_1 - \mu_2 < d \text{ is equivalent to } H_1 : \mu_1 - d < \mu_2; \text{ and}$$

$$H_1 : \mu_1 - \mu_2 \neq d \text{ is equivalent to } H_1 : \mu_1 - d \neq \mu_2.$$

If we shift the random sample of n_1 values of X_1 (with corresponding population mean μ_1) by subtracting the constant d from each X_1 value, we can view the resulting n_1 values of $X_1^* = X_1 - d$ as forming a random sample of size n_1 from a population with population mean $\mu_1* = \mu_1 - d$. Therefore, testing a hypothesis relating $\mu_1 - \mu_2$ to d based on the X_1 sample and the X_2 sample is equivalent to testing the corresponding hypothesis relating $\mu_1^* = \mu_1 - d$ to μ_2 based on the X_1^* sample and the X_2 sample.

We can construct a 95% confidence interval for $\mu_1 - \mu_2$ by finding the interval of values for the difference d for which a test at the 5% level of significance **does not** lead to the rejection of the hypothesis $H_0 : \mu_1 - \mu_2 = d$ (equivalently $H_0 : \mu_1 - d = \mu_2$). Actually finding this interval of values for d is complicated by the fact that the rank based test statistic does not explicitly depend on the actual data values. We need to determine the smallest and largest values (say d_1 and d_2, either of which may be negative) for which the test does not reject $H_0 : \mu_1 - \mu_2 = d$. A simple, but computationally intensive, method of finding this interval of values is based on the $n_1 n_2$ (n_1 times n_2) differences between all possible pairings of the values of X_1 and X_2. By ordering the $n_1 n_2$ differences from smallest to largest it is possible to determine the smallest value of d, say d_1, and the largest value of d, say d_2, which do not lead us to reject $H_0 : \mu_1 - \mu_2 = d$. This determination is based on a large sample size normal approximation to the sampling distribution of $\overline{R}_1$ which states that, when both n_1 and n_2 are reasonably large, the quantity

$$Z = \frac{n_1[\overline{R}_1 - (N+1)/2]}{\sqrt{n_1 n_2 (N+1)/12}}$$

behaves in approximate accordance with the standard normal distribution. This procedure is outlined in the steps given below.

1. Compute the quantity k obtained by rounding

$$\frac{n_1 n_2}{2} - 1.96\sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

to the nearest integer.

2. Compute all $n_1 n_2$ differences $X_1 - X_2$ and order these from smallest to largest including any repeats which occur.
3. The lower limit d_1 for the confidence interval is the difference located at the position k places in from the beginning of the ordered listing (counting up). The upper limit d_2 is the difference located at the position k places in from the end of the ordered listing (counting down).
4. We then conclude that we are 95% confident that the difference $\mu_1 - \mu_2$ is between d_1 and d_2.

Example. Cowbird parasitization of flycatchers (revisited). We will now construct a 95% confidence interval for the difference $\mu_1 - \mu_2$ giving us an estimate of the amount by which the population mean vocalization rate for the population of all parasitized flycatcher pairs, μ_1, exceeds the population mean vocalization rate for the population of all non–parasitized flycatcher pairs, μ_2.

Table 13. The 42 differences $\mathbf{X_1 - X_2}$ for the cowbird data.

		X_2					
		0	.25	1	1	1	3.25
X_1	1.10	1.10	.85	.10	.10	.10	-2.15
	1.25	1.25	1	.25	.25	.25	-2
	1.25	1.25	1	.25	.25	.25	-2
	2	2	1.75	1	1	1	-1.25
	3.75	3.75	3.5	2.75	2.75	2.75	.5
	5.5	5.5	5.25	4.5	4.5	4.5	2.25
	8.5	8.5	8.25	7.5	7.5	7.5	5.25

Table 14. The ordered differences $\mathbf{X_1 - X_2}$.

-2.15	-2	-2	-1.25	.10	.10	.10	.25	.25	.25	.25
.25	.25	.50	.85	1	1	1	1	1	1.10	1.25
1.25	1.75	2	2.25	2.75	2.75	2.75	3.50	3.75	4.50	4.50
4.50	5.25	5.25	5.50	7.50	7.50	7.50	8.25	8.50		

The quantity from step 1 in the confidence interval construction given above is 7.28, which on rounding to the nearest integer gives $k = 7$. Counting up (in Table 14) we find that the seventh difference is .1 and counting down we find that the seventh difference is 5.25. Therefore, we are 95% confident that the population mean vocalization rate for the

population of all parasitized flycatcher pairs, μ_1, exceeds the population mean vocalization rate for the population of all non–parasitized flycatcher pairs, μ_2 by at least .1 and at most 5.25 calls per minute.

8.4 Summary

This chapter is concerned with inference for the difference $\mu_1 - \mu_2$ between two population means. We began by discussing the shift assumption under which the two distributions being compared are identical except for the values of the two population means μ_1 and μ_2. Under this shift assumption an inference about the difference $\mu_1 - \mu_2$ completely characterizes the difference between the two distributions. The majority of this chapter is devoted to inference for the difference between the means of two normal distributions.

Given independent random samples, of size n_1 and n_2, the sampling distribution of the difference $\overline{X}_1 - \overline{X}_2$ between the two sample means has population mean $\mu_1 - \mu_2$ and the population standard error of $\overline{X}_1 - \overline{X}_2$ is S.E.$(\overline{X}_1 - \overline{X}_2) = \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$. Thus the difference $\overline{X}_1 - \overline{X}_2$ is unbiased as an estimator of $\mu_1 - \mu_2$ and the variability of $\overline{X}_1 - \overline{X}_2$ as an estimator of $\mu_1 - \mu_2$ can be quantified using this standard error. If we also assume that the two population distributions are normal distributions, *i.e.*, if we assume that the data form independent random samples from normal distributions, then the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is the normal distribution with population mean $\mu_1 - \mu_2$ and population standard deviation S.E.$(\overline{X}_1 - \overline{X}_2)$.

Given independent random samples from normal distributions with population means μ_1 and μ_2 and with common population standard deviation σ, the quantity

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{(1/n_1) + (1/n_2)}},$$

where S_p denotes the pooled sample standard deviation, follows the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom. Therefore, if the normality and common population standard deviation assumptions are reasonable, then we can use the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom to make inferences about the difference $\mu_1 - \mu_2$.

Under the normality and common population standard deviation assumptions the interval from $(\overline{X}_1 - \overline{X}_2) - kS_p\sqrt{(1/n_1) + (1/n_2)}$ to $(\overline{X}_1 - \overline{X}_2) + kS_p\sqrt{(1/n_1) + (1/n_2)}$, where k denotes the 97.5 percentile of the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom, is a 95% confidence interval for $\mu_1 - \mu_2$. We can test a hypothesis relating $\mu_1 - \mu_2$ to zero by using the Student's t test statistic

$$T_{calc} = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{(1/n_1) + (1/n_2)}}$$

to find the appropriate P–value. The P–value is determined as the appropriate area under the density curve of the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom.

If the normality assumption is reasonable but the assumption of a common population standard deviation is not, then we can use the quantity

$$T^* = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}$$

for inferences about $\mu_1 - \mu_2$. The details of this approach, which is based on a Student's t approximation to the distribution of T^*, are outlined in Section 8.2b.

The Student's t inferential methods for $\mu_1 - \mu_2$ are based on the assumption that the underlying populations are reasonably modeled by normal distributions. When this normality assumption is not tenable we need to consider a method of inference which is applicable under weaker assumptions. If the shift assumption is reasonable, then we can make inferences about $\mu_1 - \mu_2$ based on the ranks of the observations. A Student's t approximation to this rank based approach to inference about $\mu_1 - \mu_2$ is discussed in Section 8.3. This rank based approach to inference does not require the normality assumption but it does require independent samples and the shift assumption.

Chapter 9
Descriptive Statistics for Bivariate Data

9.1 Introduction

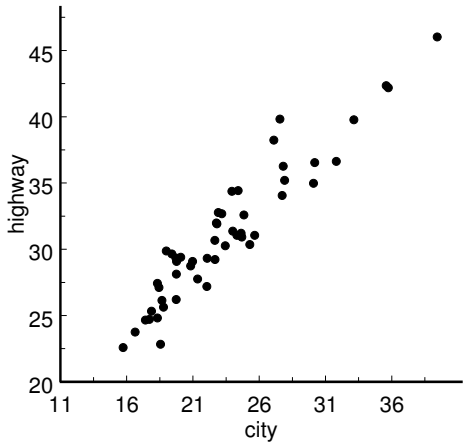
We discussed univariate data description (methods used to explore the distribution of the values of a single variable) in Chapters 2 and 3. In this chapter we will consider bivariate data description. That is, we will discuss descriptive methods used to explore the joint distribution of the pairs of values of a pair of variables. The **joint distribution** of a pair of variables is the way in which the pairs of possible values of these variables are distributed among the units in the group of interest. When we measure or observe pairs of values for a pair of variables, we want to know how the two variables behave together (the joint distribution of the two variables), as well as how each variable behaves individually (the marginal distributions of each variable).

In this chapter we will restrict our attention to bivariate data description for two quantitative variables. We will make a distinction between two types of variables. A **response variable** is a variable that measures the response of a unit to natural or experimental stimuli. A response variable provides us with the measurement or observation that quantifies a relevant characteristic of a unit. An **explanatory variable** is a variable that can be used to explain, in whole or in part, how a unit responds to natural or experimental stimuli. This terminology is clearest in the context of an experimental study. Consider an experiment where a unit is subjected to a treatment (a specific combination of conditions) and the response of the unit to the treatment is recorded. A variable that describes the treatment conditions is called an explanatory variable, since it may be used to explain the outcome of the experiment. A variable that measures the outcome of the experiment is called a response variable, since it measures the response of the unit to the treatment. For example, suppose that we are interested in the relationship between the gas mileage of our car and the speed at which our car is driven. We could perform an experiment by selecting a few speeds and then driving our car at these speeds and calculating the corresponding mileages. In this example the speed at which the car is driven is the explanatory variable and the resulting mileage is the response variable. There are also situations where both of the variables of interest are response variables. For example, in the Stat 214 example we might be interested in the relationship between the height and weight of a student; the height of a student and the weight of a student are both response variables. In this situation we might choose to use one of the response variables to explain or predict the other, *e.g.*, we could view the height of a student as an explanatory variable and use it to explain or predict the weight of a student.

9.2 Association and Correlation

The first step in exploring the relationship between two quantitative variables X and Y is to create a graphical representation of the ordered pairs of values (X, Y) which constitute the data. A **scatterplot** is a graph of the n points with coordinates (X, Y) corresponding to the n pairs of data values. When both of the variables are response variables, the labeling of the variables and the ordering of the coordinates for graphing purposes is essentially arbitrary. However, when one of the variables is a response variable and the other is an explanatory variable, we need to adopt a convention regarding labeling and ordering. We will label the response variable Y and the explanatory variable X and we will use the usual coordinate system where the horizontal axis (the X–axis) indicates the values of the explanatory variable X and the vertical axis (the Y–axis) indicates the values of the response variable Y. With this standard labeling convention, the scatterplot is also called a plot of Y versus X. Some of the scatterplots in this section employ jittering (small random displacements of the coordinates of points) to more clearly indicate points which are very close together.

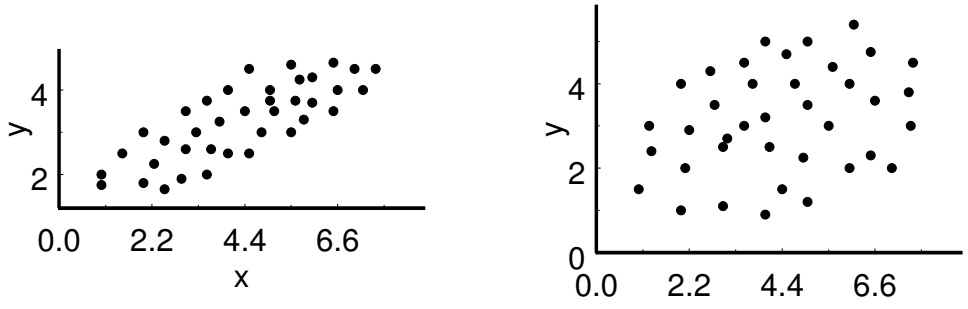
Figure 1. Subcompact car highway mileage versus city mileage.



A scatterplot of the highway EPA mileage of a subcompact car model versus its city EPA mileage, for the $n = 51$ subcompact car models of the example in Section 3.1 (excluding the 5 unusual models), is given in Figure 1. There is an obvious trend or pattern in the subcompact car mileage scatterplot of Figure 1. A subcompact car model with a higher city mileage value tends to also have a higher highway mileage value. This relationship is an example of positive association. We can also see that the trend in this example is more linear than nonlinear. That is, the trend in the subcompact car mileage scatterplot is more like points scattered about a straight line than points scattered about a curved line.

The two plots in Figure 2 illustrate positive linear association. Moving to the right in the X direction we see that the points tend to move upward in the Y direction. That is, as the value of X increases the value of Y tends to increase as well. This linear association (linear trend) is stronger in plot A than it is in plot B. The quantity r provided with these plots is a measure of linear association which will be explained later.

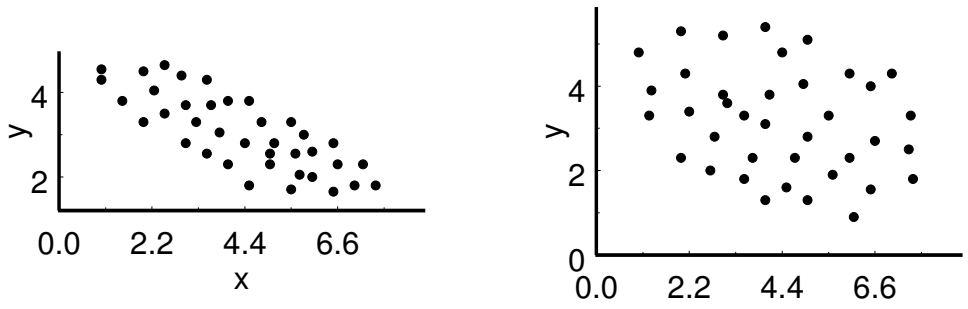
Figure 2. Examples of positive linear association



Plot A $r = .787$ Plot B $r = .295$

The two plots in Figure 3 illustrate negative linear association. Moving to the right in the X direction we see that the points tend to move downward in the Y direction. That is, as the value of X increases the value of Y tends to decrease. Again, this linear association (linear trend) is stronger in plot A than it is in plot B.

Figure 3. Examples of negative linear association.



Plot A $r = -.787$ Plot B $r = -.295$

We might describe the points in a scatterplot as forming a point cloud. A useful heuristic approach to the idea of linear association is provided by picturing an ellipse drawn around the point cloud. By envisioning ellipses drawn around the points in the plots of Figures 2 and 3, we can make the following observations. When there is positive linear association, the long direction (major axis) of the ellipse slopes upward; and when there is negative linear association, the long direction of the ellipse slopes downward. Moreover,

the width of the ellipse in the direction perpendicular to the long direction (the minor axis) indicates the strength of the linear association. That is, a narrower ellipse indicates stronger linear association than does a wider ellipse. Please note that it is the width of the ellipse and not the steepness of the long direction of the ellipse that indicates strength of linear association.

It is difficult, even with a lot of experience, to determine precisely how strong the linear association between two variables is from a scatterplot. Therefore we need to define a numerical summary statistic that can be used to quantify linear association.

We first need to quantify the location of the center of the point cloud in the scatterplot. We will use the two means $\overline{X}$ and $\overline{Y}$ to quantify the center (location) of the point cloud in the X–Y plane. That is, the point with coordinates $(\overline{X}, \overline{Y})$ will serve as our quantification of the center of the point cloud (the center of the ellipse around the data).

To motivate the statistic that we will use to quantify linear association we need to describe the notions of positive and negative linear association relative to the point $(\overline{X}, \overline{Y})$. If X and Y are positively linearly associated, then when X is less than its mean $\overline{X}$ the corresponding value of Y will also tend to be less than its mean $\overline{Y}$; and, when X is greater than its mean $\overline{X}$ the corresponding value of Y will also tend to be greater than its mean $\overline{Y}$. Therefore, when X and Y are positively linearly associated the product $(X - \overline{X})(Y - \overline{Y})$ will tend to be positive. On the other hand, if X and Y are negatively linearly associated, then when X is less than its mean $\overline{X}$ the corresponding value of Y will tend to be greater than its mean $\overline{Y}$; and when X is greater than its mean $\overline{X}$ the corresponding value of Y will tend to be less than its mean $\overline{Y}$. Therefore, when X and Y are negatively linearly associated the product $(X - \overline{X})(Y - \overline{Y})$ will tend to be negative. This observation suggests that an average of these products of deviations from the mean, $(X - \overline{X})(Y - \overline{Y})$, averaging over all n such products, can be used to determine whether there is positive or negative linear association.

If an average of the sort described above is to be useful for measuring the strength of the linear association between X and Y, then we must standardize these deviations from the mean. Therefore, the statistic that we will use to quantify the linear association between X and Y is actually an "average" of the products of the standardized deviations of the observations from their means (Z–scores). This "average" of n values is computed by dividing a sum of n terms by $n - 1$, just as we divided by $n - 1$ in the definition of the standard deviation. Linear association is also known as **linear correlation** or simply **correlation**; and the statistic that we will use to quantify correlation is called the correlation coefficient. The **correlation coefficient** (Pearson correlation coefficient), denoted by the lower case letter r, is defined by the formula

$$r = \sum \left(\frac{X - \overline{X}}{S_X} \right) \left(\frac{Y - \overline{Y}}{S_Y} \right) / (n - 1) \ .$$

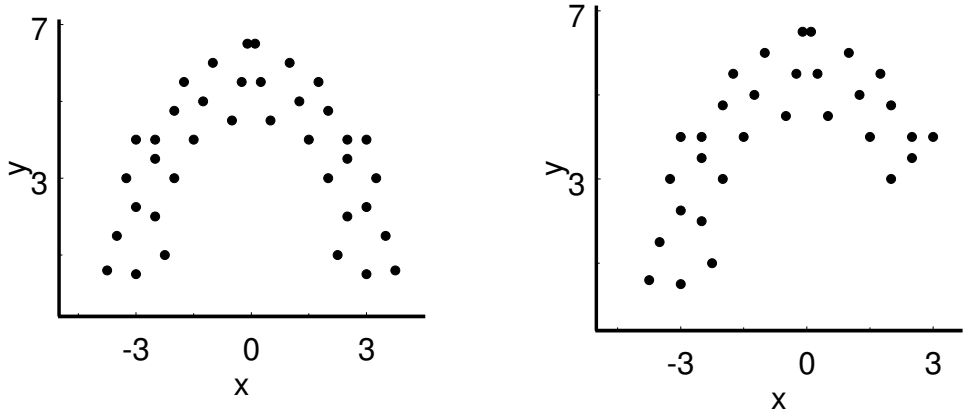
In words, the **correlation coefficient** r is the "average" of the products of the pairs of standardized deviations (Z–scores) of the observed X and Y values from their means. This formula for r is not meant to be used for computation. You should use a calculator or a computer to calculate the correlation coefficient r.

The correlation coefficient is a unitless number that is always between -1 and 1. The sign of r indicates the direction of the correlation between X and Y. A positive r indicates positive correlation and a negative r indicates negative correlation. If $r = 1$, then the variables X and Y are perfectly positively correlated in the sense that the points lie exactly on a line with positive slope. If $r = -1$, then the variables X and Y are perfectly negatively correlated in the sense that the points lie exactly on a line with negative slope. If $r = 0$, then the variables are uncorrelated, *i.e.*, there is no linear correlation between X and Y.

The magnitude of r indicates the strength of the correlation between X and Y. The closer r is to one in absolute value the stronger is the correlation between X and Y. The correlation coefficients for the plots in Figures 2 and 3 are provided below the plots. The correlation coefficient for the highway and city mileage values for the 51 subcompact car models plotted in Figure 1 is $r = .9407$ indicating that there is a strong positive correlation between the city and highway mileage values of a subcompact car.

In many situations the relationship between two variables may involve nonlinear association. The plots in Figure 4 illustrate two versions of nonlinear association. In both plots, as the value of X increases the value of Y tends to increase at first and then to decrease. In plot A of Figure 4 there is no linear association between X and Y (in this plot the ellipse would either be a circle or the long direction of the ellipse would be exactly vertical) and the correlation coefficient is zero. In plot B of Figure 4 there is a positive linear component to the nonlinear association between X and Y (the ellipse would slope upward) and the correlation coefficient is positive.

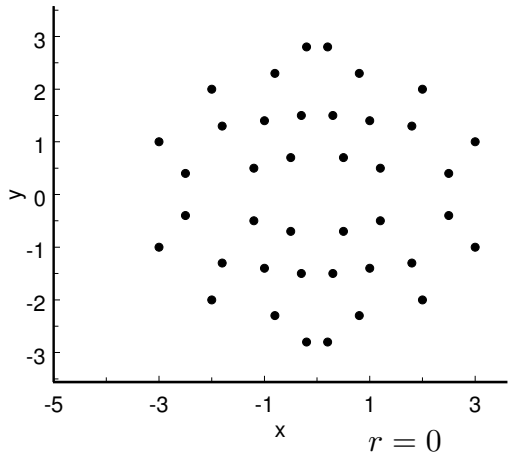
Figure 4. Examples of nonlinear association.



Plot A $r = 0$ Plot B $r = .505$

Plot A of Figure 4 illustrates a situation where there is association between X and Y but there is no correlation (no linear association). The plot of Figure 5 illustrates a situation where there is no association at all between X and Y. When there is no association, the points in the scatterplot appear to be randomly scattered about with no evidence of a trend, linear or nonlinear, and the correlation coefficient is zero. The correlation coefficient is also zero when the long direction of the ellipse around the point cloud is horizontal or vertical.

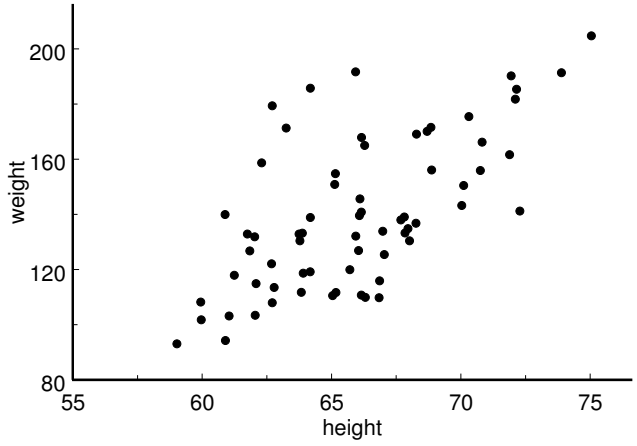
Figure 5. An example of no association.



$r = 0$

Example. Weights and heights for the Stat 214 example. The scatterplot in Figure 6 shows the relationship between the weights (in pounds) and heights (in inches) of the $n = 67$ students in the Stat 214 example of Chapter 1.

Figure 6. Stat 214 weight and height example.



This scatterplot of weight versus height shows positive linear association between these variables. The correlation coefficient is $r = .6375$ which indicates a moderate positive

correlation between the weight of a Stat 214 student and his or her height. This means that there is some tendency for a student who is heavier than average to also be taller than average; and similarly, for a student who is lighter than average to also be shorter than average.

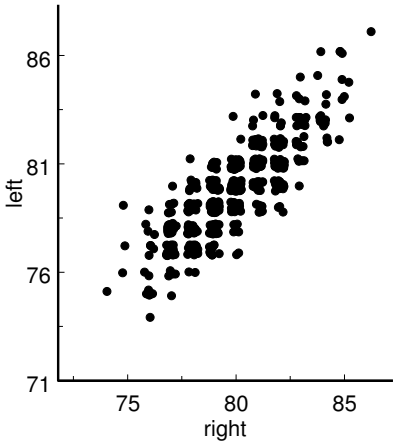
Example. Bee forewing vein length. This example is taken from Sokal and Rohlf, *Biometry*, (1969). The data are from Phillips, *Cornell Exp. Sta. Mem* **121**, (1929). These data consist of right and left forewing vein lengths (in mm times 50) for a sample of 500 worker bees. We would expect larger bees to tend to have larger wings on both sides of the body, and wing vein lengths should reflect this positive association. Thus the purpose of this example is to assess the evidence for this type of symmetry in worker bees. The data are summarized in the form of a joint frequency distribution with appended marginal frequency distributions in Table 1, and a plot of the data is provided in Figure 7. There is strong positive correlation between right and left forewing vein lengths for these bees. The correlation coefficient $r = .8372$, which quantifies the strength of linear association between these measurements, provides a measure of developmental homeostasis (physiological stability) for worker bees.

Table 1. Bee forewing vein length data.

This table provides the joint and marginal frequency distributions of the right and left forewing vein lengths (in mm times 50) for 500 worker bees.

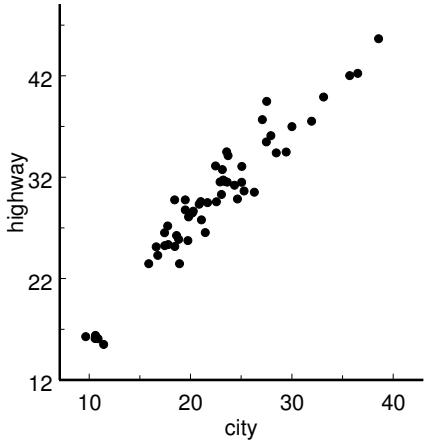
	left vein length														
right vein length	74	75	76	77	78	79	80	81	82	83	84	85	86	87	row freq.
86														1	1
85									1	1	2	2	3		9
84									3	7	2	1	1		14
83							1	4	6	16	3	1			31
82						3	13	19	12	5	2				54
81						8	18	22	19	3	1				71
80				2	6	23	46	23	1	1					102
79				10	17	34	25	8							94
78			2	14	19	12	11	1							59
77		1	4	19	14	6	1								45
76	1	5	2	4	3	1									16
75			1	1		1									3
74		1													1
column freq.	1	7	9	50	59	88	115	77	42	33	10	4	4	1	500

Figure 7. Bee forewing vein length example.



When examining a scatterplot we may find one or more unusual pairs of data values. That is, we may find that there is a point in the plot that is widely separated from the majority of the points in the plot. If the relationship between the coordinates of the unusual point agree with the overall linear pattern of the other points, then the unusual point will have the effect of strengthening the linear association between X and Y. Such an unusual point lengthens and narrows the ellipse and causes the magnitude of the correlation coefficient to increase ($|r|$ gets larger). If the relationship between the coordinates of the unusual point does not agree with the overall linear pattern of the other points, then the unusual point will have the effect of weakening the linear association between X and Y. Such an unusual point makes the ellipse wider and causes the magnitude of the correlation coefficient to decrease ($|r|$ gets smaller).

Figure 8. Subcompact car highway mileage versus city mileage (all 56 models).



The scatterplot of the highway EPA mileage of a subcompact car model versus its city EPA mileage in Figure 8 includes the values for all $n = 56$ subcompact car models

(including the 5 unusual models). Here we see that the points corresponding to the five unusual car models are separated from the other points but agree with the overall linear trend. In this example including these five unusual car models increases the correlation coefficient from .9407 to .9596.

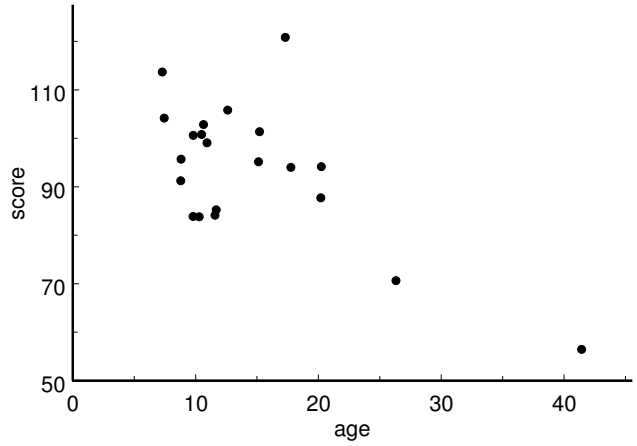
When the data include an unusual point it is a good idea to verify that the data were recorded correctly, since an error might produce an unusual point. Assuming that no data error exists, it might be a good idea to compute the correlation coefficient twice, once with all of the data and once with the unusual point not included. If there is a substantial difference between these two correlation coefficients, then appropriate comments can be added to your discussion of the problem. Another possible reason for an unusual point is the lack of enough relevant data. That is, the separation between the unusual point and the others may be due to the omission of data the inclusion of which would eliminate the separation. Therefore, a substantial difference between the two correlation coefficients (computed with and without the unusual point) might also warrant the collection of additional data for further investigation.

Example. Age at first word and Gesell test scores. This example is concerned with the relationship between the age at which a child begins to use words and the score the child attains on a test of mental ability given at a later age. The data given in Table 2 are the age at which a child spoke its first word and the score that the child attained on the Gesell adaptive test (the test was administered at a much later age). The data used in this example are from Mickey, Dunn, and Clark, *Comput. Biomed. Res.*, (1967) as reported in Rousseeuw and Leroy, *Robust Regression and Outlier Detection*, (1987). There are two response variables in this example: the age (in months) at which the child spoke its first word and the child's score on the Gesell adaptive test.

Table 2. Age at first word and Gesell score data.

child number	age at first word	Gesell score	child number	age at first word	Gesell score
1	15	95	12	9	96
2	26	71	13	10	83
3	10	83	14	11	84
4	9	91	15	11	102
5	15	102	16	10	100
6	20	87	17	12	105
7	18	93	18	42	57
8	11	100	19	17	121
9	8	104	20	11	86
10	20	94	21	10	100
11	7	113			

Figure 9. Plot of Gesell score versus age at first word.



The scatterplot given in Figure 9 shows negative linear association between these two variables. The correlation coefficient is $r = -.6403$ which indicates a moderate negative correlation between the age at which the child spoke its first word and the score that the child received on the Gesell adaptive test. This means that there is some tendency for a child who speaks its first word earlier to score higher on the Gesell test than a child who speaks its first word later.

An examination of the scatterplot for these data reveals that there are at least two unusual points in this data set.

The point (17, 121), corresponding to child number 19, is unusual in the sense that this point is separated from the other points and this pair of values does not agree with the overall negative linear trend in the data. This child spoke its first word at the age of 17 months and scored 121 on the Gesell adaptive test. These values do not fit into the overall pattern of the data for the other 20 children. Judging from the overall pattern we would expect a child that spoke its first word at age 17 months to have a Gesell score in the neighborhood of 90. Therefore, the score for this child appears to be unusually high. One possible explanation for this is that there might have been an error in recording the values for this child. This possibility was checked and it was determined that no error had been made and these are the correct values for this child. There is no justification for removing this child from the study; however, it is instructive to note that if we compute the correlation coefficient for the other 20 children omitting child number 19 we get $r = -.7561$. Thus we see that this single child (single pair of values) has a fairly large influence on the magnitude of the correlation coefficient.

The point (42, 57), corresponding to child number 18, is unusual in the sense that there is a large separation between this point and the other points in the scatterplot; but, this pair of values does agree with the overall negative linear trend in the data. These characteristics cause this pair of values to have a large influence on the linear trend in

the data. This child spoke its first word at the age of 42 months and scored 57 on the Gesell adaptive test. The age at first spoken word of 42 months is very large relative to the ages at first spoken word for the other 20 children in this group. Because of the large separation between this child's age at first spoken word and the ages at first spoken word for the other children in this group, the Gesell adaptive test score of this child will exert a large influence on the overall pattern in the data. In this example the Gesell test score for this child is such that it agrees with and strengthens the overall pattern in the data. If we compute the correlation coefficient for the 20 other children omitting child number 18, we get $r = -.3349$. Therefore, without child number 18 there is a fairly small negative correlation between the age at which a child spoke its first word and the score that the child received on the Gesell adaptive test. As with child 19 we find that this single child 18 (single pair of values) has a fairly large influence on the magnitude of the correlation coefficient.

The two unusual points discussed above demonstrate the two types of unusual points we might find in a correlation problem. At this point we will examine the present example in more detail. Notice that the point (26,71), corresponding to child number 2, is also somewhat separated from the other points in the scatterplot. We see that there are two children, child number 18 and child number 2, who spoke their first words later than the majority of the children. If we omit these two children and recompute the correlation coefficient for the 18 other children we get $r = -.0340$. This shows that the evidence for a negative correlation between age at first word and Gesell test score is very highly dependent on the presence of these two children. It would be a good idea to obtain some more data so that we could determine whether these two children really are unusual or whether we simply do not have much information about children who are late in speaking their first word.

9.3 Regression

Regression analysis is used to study the dependence of a response variable on an explanatory variable. It may be helpful to think of the explanatory variable X as a measurement of an input to a system and the response variable Y as a measurement of the output of the system. If there was an exact linear functional relationship between X and Y, then the response variable Y (the output) could be expressed as a linear function of the explanatory variable X (the input). That is, if there was an exact linear relationship, then there would exist constants a and b such that, for a given value of the explanatory variable X, the corresponding value of the response variable Y could be expressed as $Y = a + bX$. If this was the case, then the points in the scatterplot would lie on the line determined by the equation $Y = a + bX$.

In practice, the linear relationship between X and Y will not be exact and the points (corresponding to the observed values of X and Y) in the scatterplot will not lie exactly on a line. Therefore, assuming that there is a linear relationship between X and Y, we want to determine a line (a linear equation relating X and Y) which adequately summarizes the linear relationship between X and Y. Another way to say this is that we want to determine the line which best fits the data. Of course we first need to decide what we mean by saying that a line fits best. Therefore, we will first define a measure of how well a line fits the data.

The measure of the quality of the fit of a line to the data that we will use is based on the vertical deviations of the observed values of the response variable Y from the corresponding values on the proposed line. The motivation for basing the measure of quality of fit on vertical deviations is that we are using the fixed values of the explanatory variable X to explain the variability in the response variable Y and variation in Y is in the vertical direction. If Y is the response variable value for a particular value of X and $\hat{Y}$ (read this as Y hat) is the value that we would have observed if the relationship was exactly linear, then the deviation $Y - \hat{Y}$ is the signed distance from the point we observed (X, Y) to the point $(X, \hat{Y})$ on the line having the same X coordinate. The deviation $Y - \hat{Y}$ is positive when the point (X, Y) is above the line and negative when the point (X, Y) is below the line. Notice that this deviation is the signed vertical distance from the line to the point (X, Y) along the vertical line through the point X on the X–axis.

If the line fits the data well, then we would expect the points in the scatterplot to be close to the line. That is, we would expect the deviations $Y - \hat{Y}$ to be small in magnitude. We would also expect a line that fits well to pass through the "middle" of the point cloud. That is, we would expect the signs of the deviations $Y - \hat{Y}$ to vary between positive and negative with no particular pattern.

The quantity that we will use to summarize the quality of fit of a line to the data is the sum of the squared deviations of the observed Y values from the $\hat{Y}$ values predicted by the line. In symbols, this quantity is $\sum(Y - \hat{Y})^2$. When comparing the fit of two lines to the data we would conclude that the line for which this sum of squared deviations is smaller provides a better fit to the data.

The **least squares regression line** is the line which yields the best fit in the sense of minimizing the sum of squared deviations of the points from the line. That is, among all possible lines, the least squares regression line is the line which yields the smallest possible sum of squared deviations. It is a mathematical fact that the least squares regression line is the line that passes through the point $(\overline{X}, \overline{Y})$ and has slope b given by the formula

$$b = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sum(X - \overline{X})^2} \ .$$

This formula for the slope of the least squares regression line is provided to show you that there is such a formula and is not meant to be used for computation. You should use a calculator or a computer to calculate the least squares regression line slope b.

If you refer back to the definition of the correlation coefficient, r, you will see that the formula for r is symmetric in X and Y. That is, interchanging the labels assigned to the two response variables has no effect on the value of r. On the other hand, the formula for the slope of the least squares regression line is clearly not symmetric in X and Y. This asymmetry reflects the fact that, in the regression context, the roles of the explanatory variable X and the response variable Y are not interchangeable.

Let $(X, \hat{Y})$ denote the coordinates of a point on the least squares regression line. The definition of the least squares regression line given above and the definition of the slope of a line imply that

$$b = \frac{\hat{Y} - \overline{Y}}{X - \overline{X}} \ .$$

Straightforward manipulation of this expression (first multiply both sides by $(X - \overline{X})$, then add $\overline{Y}$ to both sides) yields the equation

$$\hat{Y} = \overline{Y} + b(X - \overline{X})$$

for the least squares regression line. This equation is called **the mean and slope form of the equation of the least squares regression line**, since it depends on the mean $\overline{Y}$ and the slope b. Simple regrouping of terms shows that the least squares regression line equation can also be written as

$$\hat{Y} = a + bX,$$

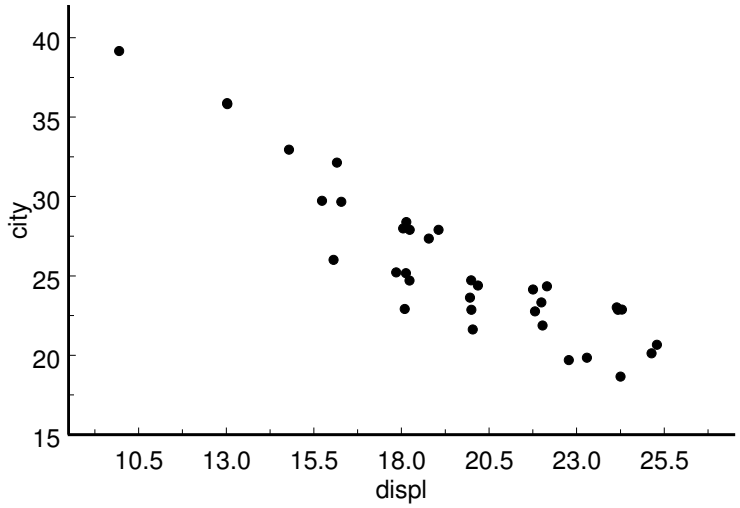
where $a = \overline{Y} - b\overline{X}$ is the y–intercept. You can obtain the value of the y–intercept a of the least squares regression line using a calculator or a computer. This equation is called **the intercept and slope form of the equation of the least squares regression line**, since it depends on the y–intercept a and the slope b. Of these two forms, the intercept and slope form of the equation of the least squares regression line is more convenient for most purposes.

The slope b of the least squares regression line is the (constant) rate of change of $\hat{Y}$ as a function of X, *i.e.*, if we start at a particular point $(X, \hat{Y})$ on the line and move one unit to the right in the X direction, then $\hat{Y}$ moves b units in the Y direction giving the point $(X + 1, \hat{Y} + b)$. If b is positive, the change in $\hat{Y}$ is upward ($\hat{Y}$ increases); and if b is negative, the change in $\hat{Y}$ is downward ($\hat{Y}$ decreases).

Notice that, according to the least squares regression line, $\overline{Y}$ is the response variable value that we would expect to see when $X = \overline{X}$. That is, substituting $X = \overline{X}$ into the least squares regression line equation gives $\hat{Y} = \overline{Y}$. The intercept value, a, is the response value

that we would expect to see, according to the least squares regression line, when $X = 0$. That is, substituting $X = 0$ into the least squares regression line equation gives $\hat{Y} = a$.

Figure 10. Plot of EPA city mileage versus displacement, for the 35 cars with displacements no larger than 2.5 liters.



Example. Subcompact car city mileage and displacement. The displacement (size) of a car engine will clearly have an effect on the gas mileage that the car will obtain. The purpose of this example is to examine the dependence of the EPA city mileage of a subcompact car model on its engine displacement. For this example we will restrict our attention to the $n = 35$ subcompact car models with engine displacements that are no larger than 2.5 liters. For ease of discussion, we will convert the engine displacements from liters to hundreds of cubic centimeters. To convert the engine displacements of Table 4 of Section 3.1 from liters to 100 cc units we simply multiply by ten, since one liter is 1000 cc. The data for this example are also provided in Table 4 of this chapter.

From the scatterplot of city mileage versus engine displacement given in Figure 10 we see that there is strong negative linear association between the city mileage of a subcompact car and its engine displacement (the correlation coefficient is $r = -.9112$). Therefore, for subcompact car models with engine displacements no greater than 2.5 liters it makes sense to use a straight line to summarize and quantify the dependence of city mileage on engine displacement.

Let X denote the engine displacement (in 100 cc's) of a subcompact car model and let Y denote its EPA city mileage (in mpg). Some relevant summary statistics are provided in Table 3. The y–intercept is $a = 49.1590$ mpg and the slope is $b = -1.2020$ mpg/100cc; therefore, the equation of the least squares regression line is $\hat{Y} = 49.1590 - 1.2020X$.

Table 3. Subcompact car EPA city mileage and engine displacement summary statistics, for the 35 car models with displacements no larger than 2.5 liters.

$n =$	35		
$\overline{X} =$	19.4571	$r =$	-.9112
$\overline{Y} =$	25.7714	slope, $b =$	-1.2020
$S_X =$	3.6730	y–intercept, $a =$	49.1590
$S_Y =$	4.8452	$R^2 =$	.8303

The slope -1.2020 is our estimate of the constant rate of change of Y, the subcompact car's city mileage, as a function of X, the car's engine displacement. According to the least squares regression line, for each 100 cc increase (each one unit increase on the 100 cc scale of measurement) in the engine displacement of a subcompact car, we expect to see a decrease (since the slope is negative) of about 1.2020 mpg in the EPA city mileage value. Recalling that the least squares regression line passes through the point $(\overline{X}, \overline{Y})$ on substituting $\overline{X}$ into the equation we see that, according to the least squares regression line, when a subcompact car has an engine displacement of 1945.71 cc its EPA city mileage should be approximately 25.7714 mpg. The nearest engine displacement for which we have data is 1900 cc and the two subcompact car models with a 1900 cc engine displacement have actual EPA city mileages of 27 and 28 mpg; therefore, the least squares regression line prediction for a 1900 cc engine is only slightly lower than these two observed values. The y–intercept is not so easily interpreted. According to the least square regression line, when a subcompact car model has an engine displacement of 0 cc its EPA city mileage should be approximately 49.1590 mpg. Clearly this does not make sense, since an engine displacement of 0 cc is nonsensical; but there is a simple explanation. The linear relationship assumed when we fit the least squares regression line requires a constant rate of change in EPA city mileage as a function of engine displacement. However, there is no reason to expect this relationship to hold for all possible engine displacements. We might expect the rate of change to be different for very small engines than it is for the displacement range for which we have data (1000 cc to 2500 cc). Similarly, we might not expect the linear relationship we determined for the present displacement range to be valid for cars with engines having much larger displacements. Notice that if we use the least squares regression line to predict the EPA city mileage of a car with a large enough engine displacement, then we will get a (nonsensical) negative mileage value, since the least squares regression line will eventually cross the X–axis.

The least squares regression line relationship can be used to determine the value of Y that we would predict or expect to see for a given value of X. If X^* denotes a particular

value of X, then, according to the least squares regression line, we would expect or predict that the corresponding value of Y would be

$$\hat{Y}(X^*) = a + bX^*$$

(read $\hat{Y}(X^*)$ as Y hat of X^*). The **predicted value** $\hat{Y}(X^*)$ is obtained by substituting X^* into the equation of the least squares regression line. Notice that $\hat{Y}(X^*)$ is the second coordinate of the point $(X^*, \hat{Y}(X^*))$ where the vertical line $X = X^*$ through X^* intersects the least squares regression line.

When using the least squares regression line to predict values we need to be aware of the danger of extrapolation. A prediction of a response value corresponding to a value X of the explanatory variable outside of the observed range of X values is called an **extrapolation**. Based on the data we have there is no way to tell whether the linear relationship summarized in the least squares regression line is appropriate for X values outside of the observed range of X values. Therefore, predictions of Y values based on the least squares regression line should be restricted to X values that are within the observed range of X values.

If we compute the predicted value $\hat{Y}$ for an observed value of X (in this context $\hat{Y}$ is called a **fitted value**), then we can use this fitted value to decompose the observed value Y as

$$Y = \hat{Y} + (Y - \hat{Y}) \ .$$

In words, this says that the observed value Y is equal to the fitted value $\hat{Y}$ plus the **residual value** $(Y - \hat{Y})$. Notice that the residual $(Y - \hat{Y})$ is the signed distance from the observed value Y (the point (X, Y)) to the fitted value $\hat{Y}$ (the point $(X, \hat{Y})$) along the vertical line through X. The residual is positive if $Y > \hat{Y}$ (the point is above the line) and negative if $Y < \hat{Y}$ (the point is below the line.) The residual would be zero if the observed value Y (the point (X, Y)) was exactly on the least squares regression line, *i.e.*, if $Y = \hat{Y}$.

The n residuals $(Y - \hat{Y})$ corresponding to the n observed data values can be used to assess the quality of fit of the least squares regression line to the data. If there was an exact linear relationship between X and Y, then the points would lie exactly on the least squares regression line and the residuals would all be zero. Therefore, when the least squares regression line fits the data well, all of the n residuals should be reasonably small in magnitude, *i.e.*, all of the points should be reasonably close to the line. A residual that is large in magnitude in one example may not be large in another example; therefore, it is important to assess the size of the residuals relative to the amount of variability in the observed Y values. If there is a systematic nonlinear pattern in the data, then we would expect to see a systematic pattern of lack of fit of the least squares regression line to the data points. Therefore, a systematic pattern in the residuals would provide evidence

Table 4. Subcompact car city mileage and displacement data, fitted, and residual values.

displacement X	city mileage Y	fitted value $\hat{Y}$	residual $Y - \hat{Y}$
10	39	37.1390	1.8610
13	36	33.5330	2.4670
13	36	33.5330	2.4670
15	33	31.1289	1.8711
16	30	29.9269	.0731
16	32	29.9269	2.0731
16	26	29.9269	-3.9269
16	30	29.9269	.0731
18	25	27.5229	-2.5229
18	25	27.5229	-2.5229
18	28	27.5229	.4771
18	28	27.5229	.4771
18	28	27.5229	.4771
18	23	27.5229	-4.5229
18	25	27.5229	-2.5229
19	27	26.3209	.6791
19	28	26.3209	1.6791
20	25	25.1189	-.1189
20	23	25.1189	-2.1189
20	22	25.1189	-3.1189
20	24	25.1189	-1.1189
20	24	25.1189	-1.1189
20	24	22.7149	1.2851
20	22	22.7149	-.7149
20	24	22.7149	1.2851
20	23	22.7149	.2851
20	23	22.7149	.2851
23	20	21.5129	-1.5129
23	20	21.5129	-1.5129
24	23	20.3109	2.6891
24	23	20.3109	2.6891
24	23	20.3109	2.6891
24	19	20.3109	-1.3109
25	20	19.1089	.8911
25	21	19.1089	1.8911

of systematic lack of fit of the least squares regression line to the data. Finally, if the variability of the observed Y values depends on the corresponding values of X, *e.g.*, the Y values corresponding to small values of X might exhibit less variability than the Y values

corresponding to large values of X, then the least squares regression line may fit better for some intervals of X values than for other intervals of X values. This sort of behavior may be detectable from the relationship between the residual values and the corresponding X values.

The fitted and residual values for the subcompact car city mileage and displacement example are given in Table 4. In this example, all of the residuals except two are less than 3.2 in magnitude. The two large residual values are: -4.5229 corresponding to the Toyota Celica model with a 1800 cc engine and a city mileage of 23 mpg, and -3.9269 corresponding to the Honda Civic DOHC/VTEC model with a 1600 cc engine and a city mileage of 26 mpg. Because these two largest residuals are negative, we see that the observed city mileage values for these two car models are substantially smaller than we would expect them to be according to the least squares regression line. There is no obvious pattern in the signs of the residuals in this example. Hence, we can conclude that the least squares regression line provides a reasonable fit to the subcompact car mileage and displacement data, and that the least squares regression line is suitable as a summary and quantification of the dependence of the EPA city mileage of a subcompact car on its engine displacement.

The least squares regression line uses the values of the explanatory variable X to explain or account for the variability in the observed values of Y. Therefore, a measure of the amount of the variability in the observed Y values that is explained, or accounted for, by the least squares regression line can be used to quantify how well the least squares regression line explains the relationship between the X and Y data values.

The sum of the squares of the residuals $\sum(Y - \hat{Y})^2$ can be viewed as a measure of the variability in the data, taking the values of the explanatory variable X and the least squares regression line into account, since the fitted values depend on the X values and the least squares regression line. The sum of the squares of the deviations of the observed Y values from their mean $\sum(Y - \overline{Y})^2$ can be viewed as a measure of the variability in the observed Y values, ignoring the corresponding X values. Therefore, the difference between these two sums of squared deviations provides a measure of the amount of the variability in the observed Y values that is explained, or accounted for, by the least squares regression line. In symbols, the difference that we are referring to as a measure of the amount of the variability in the observed Y values that is accounted for by the least squares regression line is

$$\sum(Y - \overline{Y})^2 - \sum(Y - \hat{Y})^2.$$

The ratio of this difference to the sum of the squared deviations from the mean $\sum(Y - \overline{Y})^2$ is known as **the coefficient of determination** and is denoted by R^2. The coefficient of

determination is

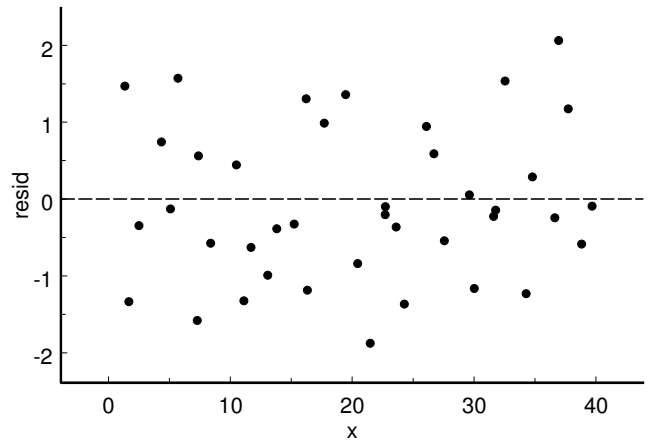
$$R^2 = \frac{\sum(Y - \overline{Y})^2 - \sum(Y - \hat{Y})^2}{\sum(Y - \overline{Y})^2}.$$

Don't worry about having to use this formula to compute R^2. For the least square regression problem that we are considering the coefficient of determination R^2 is equal to the square of the correlation coefficient. The coefficient of determination R^2 is the proportion of the variability in the observed values of Y that is explained, or accounted for, by the least squares regression line. R^2 is a positive number between 0 and 1 and a value of R^2 close to one indicates that the least squares regression line explains the data well.

For the subcompact car mileage and displacement example, the coefficient of determination is $R^2 = .8303$. This means that, through the least squares regression line, the engine displacement of a subcompact car accounts for 83.03% of the variability in the subcompact car mileages. In other words, the least squares regression line based on engine displacement alone is actually quite successful in explaining the city mileage of a subcompact car with an engine displacement between 1 and 2.5 liters.

In addition to an examination of the residual values, we can visually assess the quality of fit of the least squares regression line to the data by plotting the line on the scatterplot of the data. A more formal graphical approach to assessing the quality of fit of the least squares regression line to the data is through a residual plot. The **residual plot** is the plot of the residuals $(Y - \hat{Y})$ versus the observed X values. An ideal residual plot should be such that all of the points are contained in a relatively narrow horizontal band centered at zero; and such that there is no obvious nonlinear pattern inside the band. An example of such an ideal residual plot is provided in Figure 11.

Figure 11. An ideal residual plot.



The residual plots in Figures 12 and 13 illustrate the two problems with the fit of the least squares regression line described earlier. The plot in Figure 12 exhibits evidence of a nonlinear trend in the data. In this example the least squares regression line is too low for small X values, it is too high for middle X values, and it is too low for large X values. We can see evidence of this behavior in the residual plot, since the residuals are positive for small X values, they are mostly negative for middle X values, and they are positive for large X values. The fan shape of the plot in Figure 13 indicates that the variability in the residuals depends on the value of X. In this example, the least squares regression line fits the data better for smaller X values than it does for larger X values. That is, the variability is smaller in the residuals corresponding to small X values than it is for the residuals corresponding to large X values.

Figure 12. A residual plot indicating poor fit.

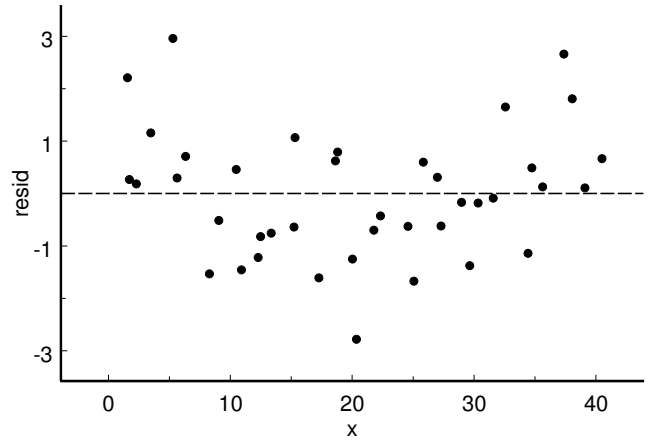
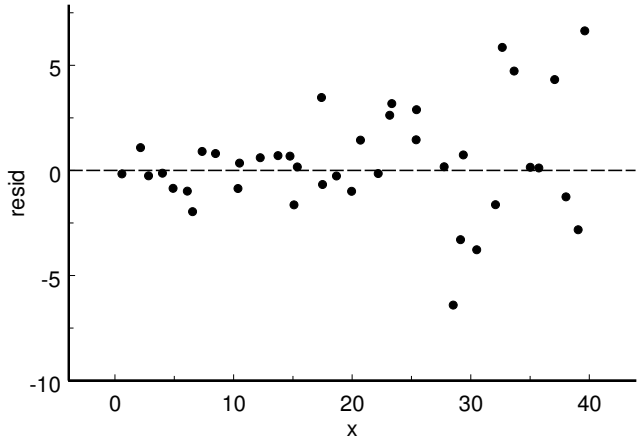


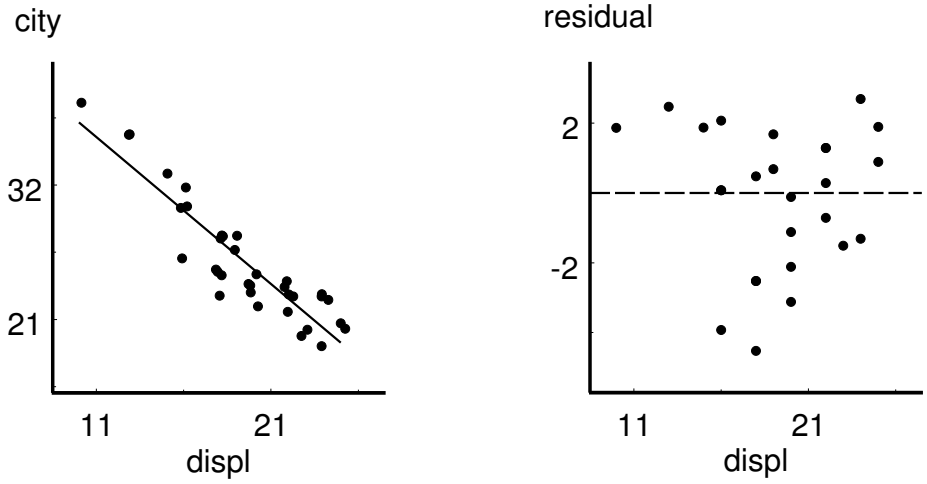
Figure 13. A residual plot indicating nonconstant variability.



When examining a residual plot you should look for obvious patterns that are supported by a reasonable number of points. Try not to let one or two slightly unusual residuals convince you that there is an overall problem. You should also be aware that an uneven distribution of X values (with more points in some X intervals than in other X intervals) may suggest problems that would disappear with the addition of a few more observations. Proper interpretation of residual plots requires practice and a sample size that is large enough to provide reliable information.

The residual plot for the subcompact car city mileage and displacement example is given in Figure 14. This figure also provides a scatterplot of the data and the fitted line. For this example the least squares regression line seems to fit the data reasonably well. However, there is a disturbing aspect of this residual plot. The first three residuals, corresponding to car models with very small engines, are positive and, as a group, these residuals are somewhat separated from the other residuals in this plot. From the scatterplot of Figure 10 we see that these three points are influential points, since the engine displacements of these three car models are small relative to the majority of the engine displacements. Before we discuss this point further, we need to discuss unusual points in the regression context.

Figure 14. Plot with fitted line and residual plot, for the 35 cars with displacements no larger than 2.5 liters.



We have already discussed the effects of unusual points on the correlation coefficient. In the regression context we will distinguish between two types of unusual points. An observation or point is said to be a **bivariate outlier** when the point does not agree with the overall linear trend in the data. A bivariate outlier may be detected visually in the scatterplot or in some examples by the fact that the corresponding residual is rather large relative to the other residuals. A scatterplot with a bivariate outlier is provided in Figure

15. For the particular situation illustrated here, the effect of the bivariate outlier would be to pull the least squares regression line upward and away from the other points in the figure. This would result in an increase in the y–intercept but little change in the slope, since this bivariate outlier is located near the middle of the observed X values.

Figure 15. A scatterplot with a bivariate outlier.

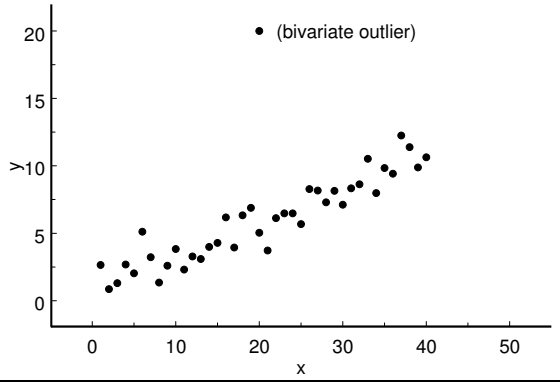
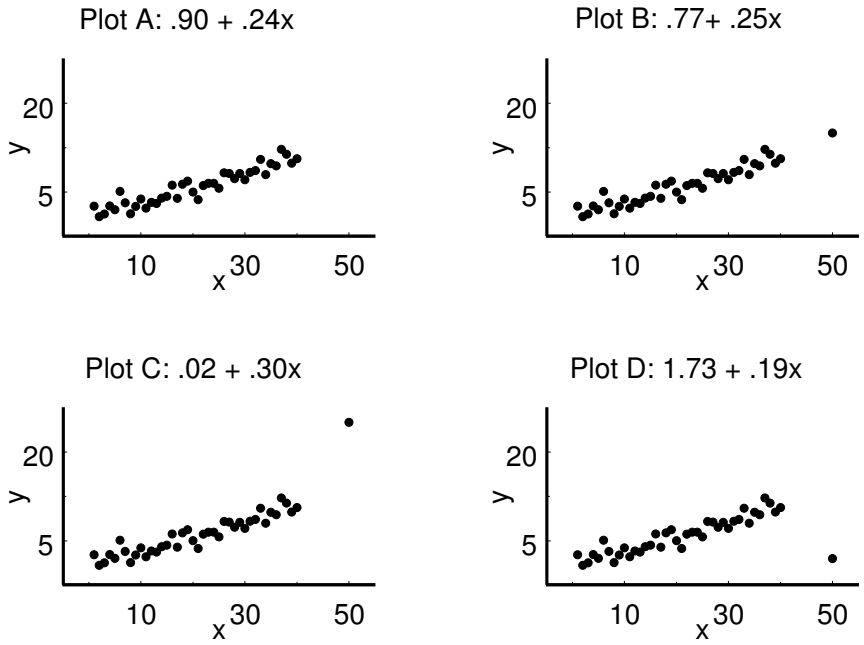


Figure 16. Scatterplots with an influential point.



An observation or point is said to be an **influential point** when the X coordinate of the point is widely separated from the X coordinates of the other points. An influential point may or may not also be a bivariate outlier. The four plots in Figure 16 illustrate the effects of an influential point on the least squares regression line slope (the equations of the regression lines are provided with the plots). The basic data for these plots is given in plot A. A single influential point is added to the data of plot A to yield the other three

plots. The influential point is a bivariate outlier in plots C and D; but it is not a bivariate outlier in plot B. Notice the dramatic effect that this single point has on the slope of the regression line in plots C and D.

We will now return to our discussion of the subcompact car city mileage and displacement example and the three car models with very small engines and somewhat unusual residuals. From the listing in Table 4 we see that there are only three subcompact car models with engine displacements that are less than 1.5 liters. From the plots of Figures 10 and 14 we see that the points corresponding to these car models are influential, since their X values are somewhat separated from the other X values. These three points are not bivariate outliers, since the relationship between the city mileage and engine displacement values for these points agrees with the overall linear trend. It might be interesting to determine how much of an effect these three car models have on the least squares regression line. We are not suggesting that there is necessarily anything wrong with including the three small displacement subcompact car models in the data set, the point here is that there is not much information about small displacement cars and we would like to see just how influential these three observations are. To this end, consider the subcompact car city mileage and displacement data for the 32 car models with engine displacements between 1.5 and 2.5 liters. Summary statistics for these data are given in Table 5 and the data, fitted, and residual values are given in Table 6.

Table 5. Subcompact car EPA city mileage and engine displacement summary statistics, for the 32 car models with displacements between 1.5 and 2.5 liters.

$n =$	32		
$\overline{X} =$	20.1563	$r =$	-.8461
$\overline{Y} =$	24.7188	slope, $b =$	-1.0014
$S_X =$	2.9524	y–intercept, $a =$	44.9030
$S_Y =$	3.4941	$R^2 =$	.7160

The slope of the least squares regression line based on all 35 subcompact car models is -1.2020 mpg/100cc and the slope is -1.0014 mpg/100cc when the three small displacement car models are excluded. As we would expect from the plot of city mileage versus engine displacement, this indicates that the least squares regression line is steeper when the three small displacement car models are included than it is when they are excluded. If we include all 35 car models, then the least squares regression line indicates a decrease in the city mileage of about 1.2020 mpg for each 100cc increase in engine displacement. However, if we exclude the three small displacement car models, then the least squares regression line indicates a decrease in the city mileage of only 1.0014 mpg for each 100 cc increase in engine displacement.

Table 6. Subcompact car city mileage and displacement data, fitted, and residual values, for the 32 car models with displacements between 1.5 and 2.5 liters.

displacement X	city mileage Y	fitted value $\hat{Y}$	residual $Y - \hat{Y}$
15	33	29.8822	3.1178
16	30	28.8808	1.1192
16	32	28.8808	3.1192
16	26	28.8808	-2.8808
16	30	28.8808	1.1192
18	25	26.8780	-1.8780
18	25	26.8780	-1.8780
18	28	26.8780	1.1220
18	28	26.8780	1.1220
18	28	26.8780	1.1220
18	23	26.8780	-3.8780
18	25	26.8780	-1.8780
19	27	25.8766	1.1234
19	28	25.8766	2.1234
20	25	24.8752	.1248
20	23	24.8752	-1.8752
20	22	24.8752	-2.8752
20	24	24.8752	-.8752
20	24	24.8752	-.8752
20	24	22.8724	1.1276
20	22	22.8724	-.8724
20	24	22.8724	1.1276
20	23	22.8724	.1276
20	23	22.8724	.1276
23	20	21.8711	-1.8711
23	20	21.8711	-1.8711
24	23	20.8697	2.1303
24	23	20.8697	2.1303
24	23	20.8697	2.1303
24	19	20.8697	-1.8697
25	20	19.8683	.1317
25	21	19.8683	1.1317

A comparison of the y–intercept values suggests that the effect of the three small displacement car models on the vertical location of the least squares regression line is quite large. The y–intercept of the least squares regression line based on all 35 subcompact

car models is 49.1590 mpg and the y–intercept is 44.9030 mpg when the three small displacement car models are excluded. The difference between these two y–intercepts is the vertical distance between these two lines at the nonsensical displacement value of 0 cc, *i.e.*, the vertical locations of the two the least squares regression lines, at $X = 0$, differ by 4.2560 mpg. Since the slopes of these two lines are different, it would make more sense to compare the vertical locations of these two lines at an X value that is within the observed displacement range. Predicted city mileage values for three representative values of X are given in Table 7. The representative X values are the smallest and largest values common to the two cases and a middle value $X = 20$ that is close to both X means. From these predicted values we see that excluding the three small displacement car models lowers the least squares regression line somewhat in the middle and at the lower end of the observed X range, but raises the line slightly at the upper end of the observed X range.

Table 7. Some representative predicted values.

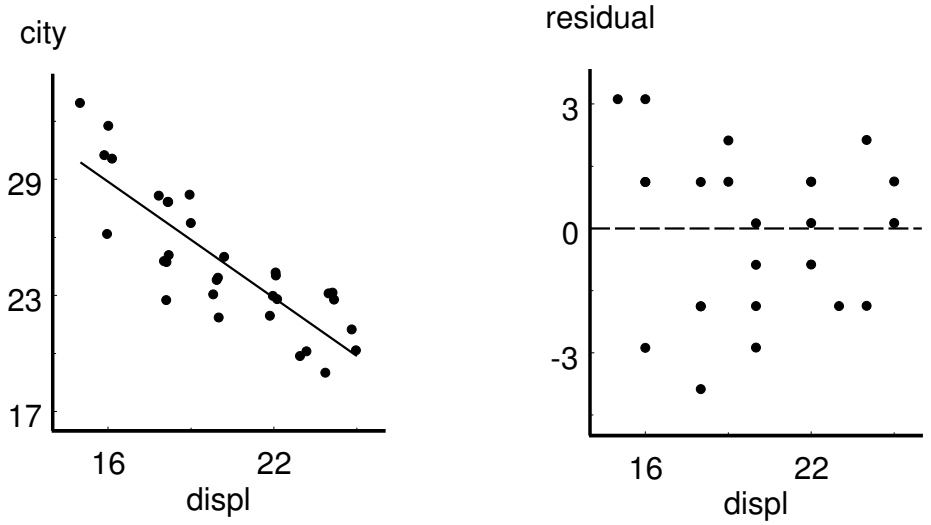
	predicted value	
displacement	including small	excluding small
15	31.1289	29.8822
20	25.1189	24.8752
25	19.1089	19.8683

The residual plot of Figure 17 does not indicate any problems with the fit of the least squares regression line when the three small displacement car models are excluded; it does show some evidence of slightly more variability for car models with smaller engine displacements. When the three small displacement car models are excluded there is only one residual that exceeds 3.2 in magnitude. The Toyota Celica model with a 1800 cc engine and a city mileage of 23 mpg has a residual of -3.8780.

The linear relationship between city mileage and displacement is somewhat weaker when the three small displacement car models are excluded. This is evident from the fact that the correlation coefficient is smaller in magnitude when these three car models are excluded. We also find that the coefficient of determination R^2 is smaller when these three car models are excluded. Since $R^2 = .7160$, we see that even without the three small displacement car models the least squares regression line still accounts for 71.6% of the variation in the subcompact car city mileage values.

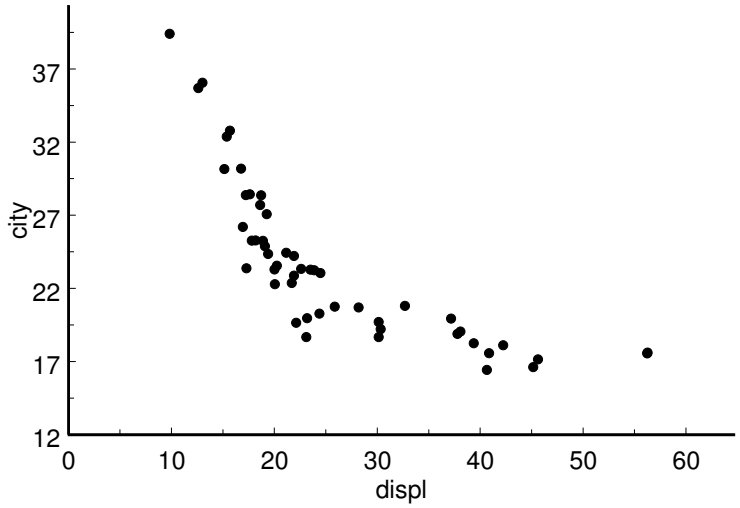
Since relatively few subcompact car models have engine displacements below 1.5 liters (only 9% of the subcompact car models with displacements no larger than 2.5 liters and only 6% of all 51 subcompact car models), we might argue that the three small displacement car models are unusual enough to justify their exclusion from our analysis of the relationship between city mileage and displacement.

Figure 17. Plot with fitted line and residual plot, for the 32 cars with displacements above 1 liter but no larger than 2.5 liters.



Our analysis of the relationship between subcompact car city mileage and engine displacement was initially restricted to car models with displacements no larger than 2.5 liters. You may have wondered why we imposed this restriction. It is instructive to reconsider the relationship between city mileage and displacement when all 51 of the subcompact car models are included. From the scatterplot of Figure 18, which is based on all 51 subcompact car models, we see that the relationship between city mileage and engine displacement is not linear when the car models with large engines are included.

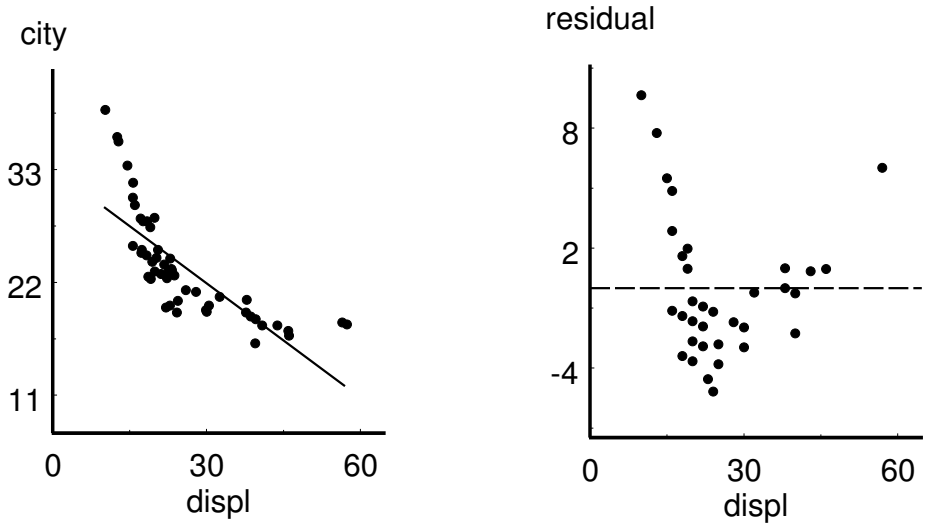
Figure 18. Plot of EPA city mileage versus displacement, for all 51 cars (excluding the 5 unusual cars).



It clearly does not make sense to fit a straight line to these data. However, we will do so to demonstrate what happens in such a situation. We would particularly like to see how this nonlinearity affects the residual plot. The residual plot for this example is given in Figure 19. There is an obvious pattern of systematic lack of fit of the least squares regression line to the data in this residual plot. For small values of X the residuals are positive and decrease as X gets larger. For middle values of X the residuals are negative and they first decrease and then increase as X gets larger. Finally, for large values of X the residuals are positive and they increase as X gets larger. This U–shaped pattern in the residual plot indicates that the least squares regression line is too low for small and large X values; and it is too high for middle X values.

In a situation like this one where there is nonlinear association we need to use more complicated regression methods that allow us to fit a curved line instead of a straight line. It is not particularly difficult to generalize the least squares approach to curved lines; however, this is beyond the scope of this chapter.

Figure 19. Plot with fitted line and residual plot, for all 51 cars (excluding the 5 unusual cars).



Chapter 10
Inference for Bivariate Data

10.1 Inference for Regression

In Chapter 9 we used the least squares regression line to summarize the linear relationship between a response variable Y and an explanatory variable X. We will now consider a more formal (inferential) approach to this problem based on a model for the population distribution of Y. The model we will use is the simple linear regression model which assumes that the population mean response (the population mean of Y) is a linear function of the explanatory variable X.

We will use a simple example to motivate the simple linear regression model and to develop the associated methods of inference. Throughout this discussion we will provide formulae and computations to clarify definitions. You will not need to perform most of these computations, since they can be performed using a suitable calculator or computer statistics program.

Example. Arsenic concentrations. Bencko and Symon (*Env. Res.* 1977) considered the effects of air pollution from a power plant burning coal with a high arsenic content on the health of persons living near the plant. Groups of ten year old boys, each group consisting of 20–27 boys, were selected from ten communities southwest (downwind) of the plant. For each group the response variable Y = average concentration of arsenic in the hair (in parts per million, ppm) was measured and the explanatory variable X = distance of the community from the plant (in kilometers, km) was recorded. The data are given in Table 1.

Table 1. Arsenic Data.

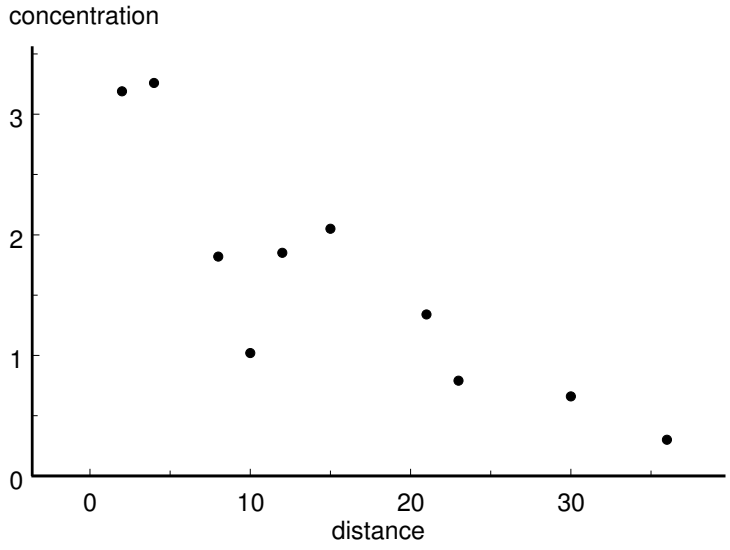
distance	2	4	8	10	12	15	21	23	30	36
arsenic conc.	3.19	3.26	1.82	1.02	1.85	2.05	1.34	0.79	0.66	0.30

First consider a model for the distribution of the response variable Y = arsenic concentration at a particular community located at a distance $X = x$ kilometers downwind from the plant. We will model the observed responses ($Y's$) using normal distributions. More specifically, we will assume that the distribution of the arsenic concentration Y corresponding to a community at a distance of $X = x$ is a normal distribution with population mean $\mu(x)$ and population variance σ^2 (population standard deviation σ). The notation $\mu(x)$ indicates that the population mean response depends on the distance $X = x$ of the

community from the plant. The population variance is assumed to be constant so that the variance of Y is the same regardless of the distance $X = x$.

In the context of this example we would expect the distribution of the arsenic concentration Y to depend on the distance X of the corresponding community from the power plant. In general, we would expect to observe smaller values of Y for communities which are farther away from the plant. The tendency to observe lower arsenic concentrations at communities farther from the plant is supported by the plot of arsenic concentration versus distance in Figure 1.

Figure 1. Plot of arsenic concentration versus distance.



The simple linear regression model assumes that the population mean response $\mu(x)$ is a linear function of the corresponding value $X = x$ of the explanatory variable. The **population simple linear regression line** can be parameterized in the intercept and slope form

$$\mu(x) = \alpha + \beta x,$$

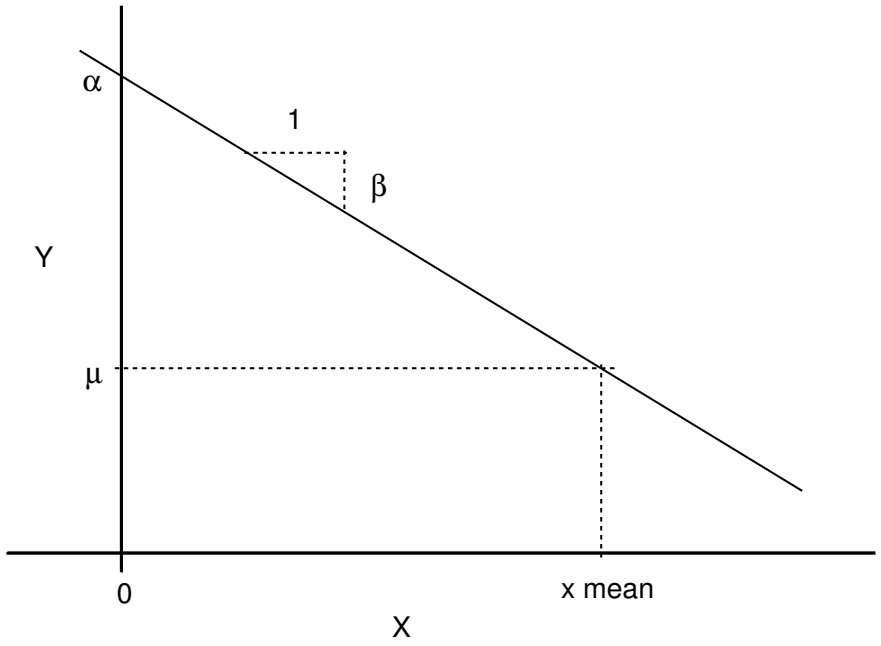
where α is the population intercept and β is the population slope; or, letting $\overline{x}$ denote the mean of the observed values of X, in the mean and slope form

$$\mu(x) = \mu + \beta(x - \overline{x}),$$

where $\mu = \mu(\overline{x})$ denotes the population mean response corresponding to $X = \overline{x}$. The simple linear regression model assumes that there is a constant rate of change, β, in the population mean response, $\mu(x)$, as a function of the explanatory variable x. In many applications this assumption of a constant rate of change will not be appropriate for all possible values of X; however, it may be reasonable if we restrict our attention to a suitable range of X

values. A population regression line, drawn with negative slope, with the parameters α, β, and μ indicated is provided in Figure 2.

Figure 2. A population regression line (drawn with negative slope).



Based on the plot of Figure 1, the assumption of a constant rate of change in the population mean arsenic concentration seems reasonable for distances between 2 and 36 km from the plant. However, we would not necessarily expect this constant rate of change to hold for distances less than 2 km or greater than 36 km. In particular, as the distance from the plant gets very large we would expect the population mean arsenic concentration to decrease more slowly and to eventually stabilize at some background level. Therefore, in this example and in general, we must be careful about making inferences which correspond to extrapolations beyond the range of X values for which we have data.

The estimated or fitted regression line is the least squares regression line introduced in Section 9.3. This fitted regression line passes through the point $(\overline{x}, \overline{Y})$ and has slope b, where

$$b = \frac{\sum(x - \overline{x})(Y - \overline{Y})}{\sum(x - \overline{x})^2} \ .$$

This fitted regression line, which can be expressed as

$$\hat{Y}(x) = a + bx = \overline{Y} + b(x - \overline{x}),$$

is now viewed as an estimate (the least squares estimate) of the population regression line

$$\mu(x) = \alpha + \beta x = \mu + \beta(x - \overline{x})$$

defined above. The least squares estimates of the population slope β, the population mean response $\mu = \mu(\overline{x})$ (corresponding to $X = \overline{x}$), and the population intercept α are the estimated slope $\hat{\beta} = b$, the sample mean response $\hat{\mu} = \overline{Y}$, and the estimated intercept $\hat{\alpha} = a = \overline{Y} - b\overline{x}$, respectively.

The fitted regression line for the arsenic example (graphed in Figure 3) has slope $b = -.07815$ ppm per km which indicates that if the distance of a community from the plant was increased by one km, we would estimate that the population mean arsenic concentration would decrease by .07815 ppm. The fitted line passes through the point $(\overline{x}, \overline{Y}) = (16.1, 1.628)$ which indicates that the estimated mean response for a distance of $X = \overline{x} = 16.1$ km is equal to $\hat{Y}(16.1) = \overline{Y} = 1.628$ ppm. The intercept for this fitted line is $a = 2.8862$ ppm. We can use the residuals and a residual plot, as discussed in Section 9.3, to determine whether this fitted regression line supports the simple linear regression model as an appropriate model for the data at hand. The observed values of the arsenic concentrations Y, the fitted values $\hat{Y}(x)$, and the residual values $Y - \hat{Y}(x)$ are given in Table 2 and the residual plot (a plot of the residuals $Y - \hat{Y}(x)$ versus the distances x) is given in Figure 4.

Table 2. Arsenic data, fitted values, and residuals.

distance X	concentration Y	fitted value $\hat{Y}$	residual $Y - \hat{Y}$
2	3.19	2.7299	0.4601
4	3.26	2.5736	0.6864
8	1.82	2.2610	-0.4410
10	1.02	2.1047	-1.0847
12	1.85	1.9484	-0.0984
15	2.05	1.7140	0.3360
21	1.34	1.2451	0.0949
23	0.79	1.0888	-0.2988
30	0.66	0.5417	0.1183
36	0.30	0.0728	0.2272

The residual plot appears reasonable overall, especially for such a small data set, with little if any evidence that a straight line (constant rate of change) model is not appropriate. There is one residual, -1.0847 for the community 10 km from the plant, which is somewhat large in magnitude indicating that the observed concentration at a distance of 10 kilometers is somewhat smaller than that predicted by the fitted regression line. The magnitude of

this residual is not large enough to cause much concern about the simple linear regression model. We might argue that there is some (slight) evidence of curvature in the residual plot suggesting that the relationship between arsenic concentration and distance is nonlinear; but, again there is not enough evidence to cause much concern. Based on these observations it seems reasonable to use the simple linear regression model for the arsenic example.

Figure 3. Arsenic data with fitted line.

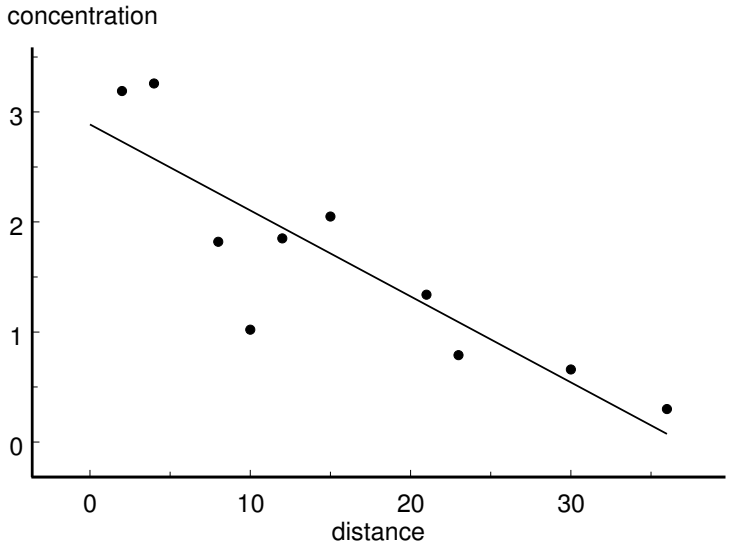
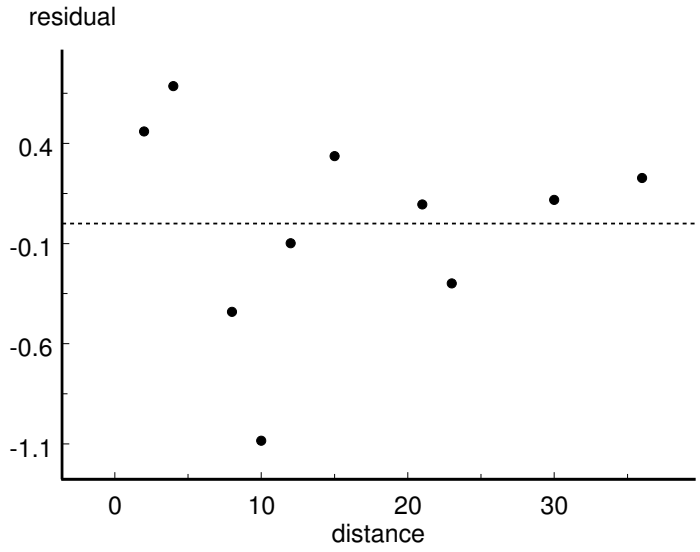


Figure 4. Arsenic example residual plot.



The next step in developing inferential methods for the simple linear regression model is to determine a suitable estimator of the common variance σ^2. We will use a pooled estimator of the common variance based on the residuals. This pooled variance estimator

is analogous to the pooled variance estimator of the two sample problem of Chapter 8. In the regression context the model allows a different mean for each distinct value of X and the fitted values, the $\hat{Y}'s$, provide estimates of these means. Thus the pooled variance estimator for the regression problem is the "average" of the squared residuals

$$S_p^2 = \frac{\sum(Y - \hat{Y})^2}{n-2},$$

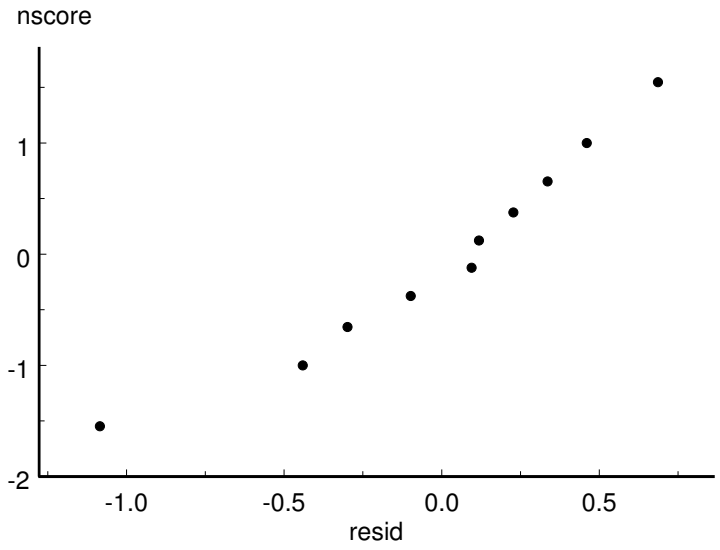
where the sum is over all n observations. The divisor in this pooled variance estimator is $n-2$, since we need two degrees of freedom to estimate the line which determines the fitted values. For the arsenic example the pooled variance estimate is $S_p^2 = .29255$ ($S_p = .54088$) with $n - 2 = 8$ degrees of freedom.

Figure 5. Stem and leaf histogram for arsenic residuals.

In this stem and leaf histogram the stem represents ones and the leaf represents tenths. (ppm)

```
-1  1
-0
-0  431
 0  1123
 0  57
```

Figure 6. Arsenic residuals normal probability plot.



The confidence interval estimates and hypothesis tests we will now develop are based on the pooled variance estimator S_p^2 from above and the Student's t distribution with $n-2$ degrees of freedom. We can examine the residuals to verify that the normality assumption required for these inferential methods is reasonable. The only disturbing feature of the

stem and leaf histogram in Figure 5 for the residuals from the arsenic example is the one large (negative) residual corresponding to $X = 10$ we mentioned above; otherwise, this histogram is consistent with a sample of size ten from a normal distribution. The normal probability plot for the residuals given in Figure 6 also indicates that the normality assumption is reasonable here.

First consider inference for the population slope β. The estimated standard error of the estimated slope b, based on the pooled variance estimator S_p^2, is

$$\widehat{\text{S.E.}}(b) = \frac{S_p}{\sqrt{\sum (x - \overline{x})^2}},$$

and the quantity

$$T = \frac{b - \beta}{\widehat{\text{S.E.}}(b)}$$

follows a Student's t distribution with $n-2$ degrees of freedom. Therefore, we can use the Student's t distribution with $n-2$ degrees of freedom to form a confidence interval for β or to test a hypothesis about β.

For the arsenic example we postulated that the population slope β (the rate of change in arsenic concentration as a function of distance) should be negative, since we expect the arsenic concentration to decrease as the distance from the plant increases. We can address this contention by testing the null hypothesis $H_0 : \beta \geq 0$ versus the research hypothesis $H_1 : \beta < 0$. Under the null hypothesis, with $\beta = 0$, the quantity

$$T = \frac{b - 0}{\widehat{\text{S.E.}}(b)}$$

follows the Student's t distribution with $n - 2 = 8$ degrees of freedom and we can use the Student's t test statistic

$$T_{calc} = \frac{b}{\widehat{\text{S.E.}}(b)}$$

to test $H_0 : \beta \geq 0$ versus $H_1 : \beta < 0$ by rejecting $H_0 : \beta \geq 0$ if T_{calc} is sufficiently far below zero. The P–value for this test is $P(T \leq T_{calc})$, where T denotes a Student's t variable with 8 degrees of freedom. In this example we have $b = -.0782$, $\widehat{\text{S.E.}}(b) = .0161$ and $T_{calc} = -4.85$, which gives a P–value of $P(T \leq -4.85) = .0006$. This P–value is very small providing strong evidence that the population slope β is negative. The 97.5 percentile of the Student's t distribution with 8 degrees of freedom is 2.306 which gives a 95% margin of error for b of M.E.$(b) = 2.306(.0161) = .0371$ and a 95% confidence interval from $-.0782 - .0371 = -.1153$ to $-.0782 + .0371 = -.0411$. Thus we are 95% confident that the population slope β is at least $-.1153$ ppm/km and at most $-.0411$ ppm/km indicating a decrease in the population mean arsenic concentration of at least .0411 ppm

and at most .1153 ppm for each increase of one kilometer in distance from the plant. Note that this interpretation of the slope of the regression line should only be used in the range of distances for which we have data, since we would not necessarily expect this simple linear regression model to hold beyond this range.

The slope provides an estimate of how the population mean response changes as a function of X. We might also want an estimate of the vertical location of the regression line. The population mean response $\mu = \mu(\overline{x})$ (the population mean of the response variable Y at the mean $\overline{x}$ of the explanatory variable values) can be used to indicate the location the population regression line. The sample mean response $\overline{Y}$ provides our estimate of μ. The estimated standard error of the estimated mean response $\overline{Y}$, based on the pooled variance estimator S_p^2, is

$$\widehat{\text{S.E.}}(\overline{Y}) = \frac{S_p}{\sqrt{n}}$$

and the quantity

$$T = \frac{\overline{Y} - \mu}{\widehat{\text{S.E.}}(\overline{Y})}$$

follows a Student's t distribution with $n-2$ degrees of freedom. Therefore, we can use the Student's t distribution with $n-2$ degrees of freedom to form a confidence interval for μ or to test a hypothesis about μ.

To get a feel for the overall population mean arsenic concentration at distances between 2 and 36 km from the plant we can estimate the population mean concentration for a distance of $X = \overline{x} = 16.1$ km, *i.e.*, we can estimate $\mu = \mu(16.1)$. The estimate of the population mean concentration at 16.1 km is $\overline{Y} = 1.628$ and the estimated standard error of $\overline{Y}$ is

$$\widehat{\text{S.E.}}(\overline{Y}) = \frac{S_p}{\sqrt{n}} = .1710.$$

Since there are $n-2 = 8$ degrees of freedom associated with S_p, we know that the quantity

$$T = \frac{\overline{Y} - \mu}{\widehat{\text{S.E.}}(\overline{Y})}$$

follows a Student's t distribution with 8 degrees of freedom. Therefore, the 95% margin of error of $\overline{Y}$ is $\text{M.E.}(\overline{Y}) = 2.306(.1710) = .3943$ and the interval from $1.628 - .3943 = 1.2337$ ppm to $1.628 + .3943 = 2.0223$ ppm is a 95% confidence interval for μ, the population mean arsenic concentration at 16.1 kilometers from the plant. Hence, we are 95% confident that the population mean arsenic concentration for a community 16.1 km from the plant is between 1.2337 and 2.0223 ppm.

The Student's t test statistic obtained from the quantity T above by replacing μ by a specific hypothesized concentration μ_0 could be used to conduct a hypothesis test for

comparing μ with μ_0. Since a relevant μ_0 value is not available, we will not consider such a hypothesis test for the arsenic example.

You may have wondered why we used the population mean response $\mu = \mu(\overline{x})$ instead of the population intercept α to quantify the vertical location of the population regression line. Since the population intercept is the population mean response for $X = 0$ and since, as in the arsenic example, $X = 0$ is often not within the range of the values of the explanatory variable we are interested in, there is often little interest in the value of α except as part of the equation for the fitted regression line. Therefore, it is usually more appropriate to consider inference for μ instead of α.

For a specified value x^* of the explanatory variable (note that this x^* should be in the range of the explanatory variable values for which we have data) we can estimate the corresponding population mean response $\mu(x^*)$ as

$$\hat{Y}(x^*) = a + bx^*$$

or as

$$\hat{Y}(x^*) = \overline{Y} + b(x^* - \overline{x}).$$

The first expression, giving $\hat{Y}(x^*)$ in terms of a and b, is more convenient for computation while the second expression, giving $\hat{Y}(x^*)$ in terms of $\overline{Y}$ and b, allows us to more easily find the estimated standard error of $\hat{Y}(x^*)$ and see how it depends on the location of x^* relative to $\overline{x}$. It can be shown that the estimators $\overline{Y}$ and b are statistically independent and that, because of this independence, we can express the estimated standard error of $\hat{Y}(x^*)$ in terms of the estimated standard errors of $\overline{Y}$ and b. For ease of notation let

$$\widehat{\text{var}}(\overline{Y}) = (\widehat{\text{S.E.}}(\overline{Y}))^2 \text{ and } \widehat{\text{var}}(b) = (\widehat{\text{S.E.}}(b))^2$$

denote the estimated variances of $\overline{Y}$ and b. The estimated standard error of $\hat{Y}(x^*)$ is

$$\widehat{\text{S.E.}}(\hat{Y}(x^*)) = \sqrt{\widehat{\text{var}}(\overline{Y}) + (x^* - \overline{x})^2 \widehat{\text{var}}(b)}.$$

Notice that the $(x^* - \overline{x})^2$ term in this standard error causes the standard error of $\hat{Y}(x^*)$ to increase as the distance between x^* and $\overline{x}$ increases. That is, the variability in $\hat{Y}(x^*)$ as an estimator of $\mu(x^*)$ is smaller for values of x^* close to $\overline{x}$ than it is for values of x^* farther from $\overline{x}$. Some calculators and computer programs will provide the standard errors of $\overline{Y}$ and b but will not provide the standard error of $\hat{Y}(x^*)$, if this is true for your calculator or computer program, you can use the expression above to find S.E.$(\hat{Y}(x^*))$.

We can use the fact that the quantity

$$T = \frac{\hat{Y}(x^*) - \mu(x^*)}{\widehat{\text{S.E.}}(\hat{Y}(x^*))}$$

follows a Student's t distribution with $n-2$ degrees of freedom to form a confidence interval for $\mu(x^*)$ or to test a hypothesis about $\mu(x^*)$. Notice that $\alpha = \mu(0)$ and we can make inferences about the population intercept using the present approach with $x^* = 0$.

Consider the problem of estimating the population mean arsenic concentration $\mu(20)$ for a hypothetical community located 20 km from the power plant. Our estimate of $\mu(20)$ is

$$\hat{Y}(20) = 2.8862 - .07815(20) = 1.3232.$$

In this example

$$\widehat{\text{var}}(\overline{Y}) = .02925 \text{ and } \widehat{\text{var}}(b) = .0002595$$

so that the estimated standard error of $\hat{Y}(20)$ is

$$\widehat{\text{S.E.}}(\hat{Y}(20)) = \sqrt{.02925 + (20 - 16.1)^2(.0002595)} = .1822,$$

which gives a margin of error of M.E.$(\hat{Y}(20)) = (2.306)(.1822) = .4202$. Therefore, we can be 95% confident that the population mean response $\mu(20)$ for a distance of 20 km is between 1.3232 - .4202 = .9030 ppm and 1.3232 + .4202 = 1.7434 ppm.

In some situations instead of estimating the population mean response $\mu(x^*)$ for $X = x^*$ we might wish to predict the actual response $Y(x^*)$ which would be observed if we were to measure Y when $X = x^*$. We can model the actual response value corresponding to $X = x^*$ as $Y(x^*) = \mu(x^*) + \epsilon$, where $\mu(x^*)$ is the corresponding population mean response and ϵ represents a random, normally distributed quantity with mean zero and standard deviation σ. The fitted value $\hat{Y}(x^*)$ which served as our estimate of $\mu(x^*)$ provides a suitable prediction (estimate) of the actual response value $Y(x^*)$ as well, since $\mu(x^*)$ is the mean of the distribution of $Y(x^*)$. However, there is more variability in $\hat{Y}(x^*)$ when it is viewed as a predictor of $Y(x^*)$ than there is when it is viewed as an estimator of $\mu(x^*)$. We can use the standard error of prediction

$$\text{S.E.P.}(\hat{Y}(x^*)) = \sqrt{\widehat{\text{var}}(\overline{Y}) + (x^* - \overline{x})^2\widehat{\text{var}}(b) + S_p^2}$$

to quantify the variability in $\hat{Y}(x^*)$ as a predictor of $Y(x^*)$, and in particular, we can use this standard error of prediction to form an interval estimate of $Y(x^*)$. Notice that the standard error of prediction differs from the standard error for estimating $\mu(x^*)$ by the addition of the term S_p^2 under the square root sign. This added term accounts for the variability in the ϵ of the expression for $Y(x^*)$ given above.

We estimated the population mean arsenic concentration $\mu(20)$ for a hypothetical community located 20 km from the power plant above. Now consider the prediction of the

actual response we would have observed if there was a community 20 km from the plant. Since $S_p^2 = 0.29255$, the standard error for prediction for a distance of 20 km is

$$\text{S.E.P.}(\hat{Y}(20)) = \sqrt{.02925 + (20 - 16.1)^2(.0002595) + .29255} = .5707$$

and the 95% prediction interval for the actual response at 20 km is the interval from 1.3232 - (2.306)(.5707) = 1.3232 - 1.3161 = .0071 ppm to 1.3232 + (2.306)(.5707) = 1.3232 + 1.3161 = 2.6393 ppm. Therefore, with 95% confidence we can predict that if we were to measure the actual arsenic concentration at a community 20 km from the plant we would get a value between .0071 ppm and 2.6393 ppm. Notice that this prediction interval is a good bit longer than the 95% confidence interval (.9030, 1.7434) for the population mean arsenic concentration for a community 20 km from the plant, since the prediction interval takes the variability of the measurement process into consideration.

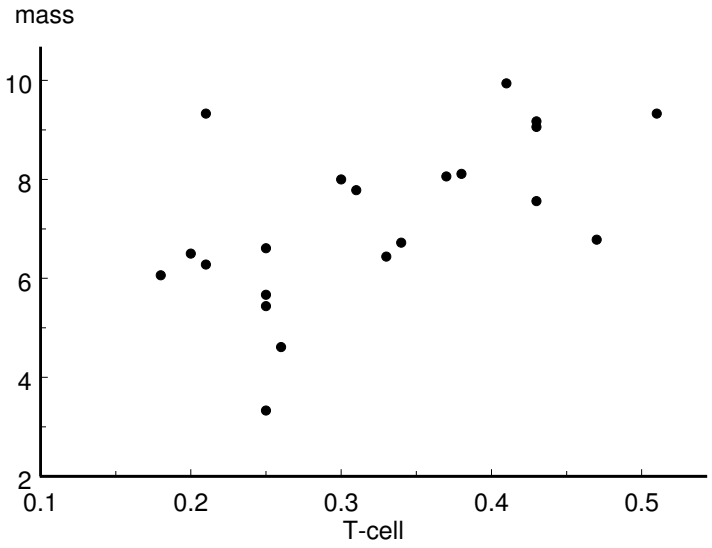
Example. Wheatear weight lifting and health status example. The black wheatear is a small passerine (perching) bird that is resident in Spain and Morroco. The male black wheatear demonstrates an exaggerated sexual display by collecting stones from the ground and placing them in cavities in cliffs, caves, or buildings while the female mate is present. Soler, Martín–Vivaldi, Marín, and Møller, *Behav. Ecol.* **10**, 281–286, (1999) investigated the relationship between such weight lifting and health status for black wheatears. The data in the first two columns of Table 3 (which were read from Figure 1 of this paper) correspond to a sample of $n = 21$ male black wheatears. The two variables are: the bird's T–cell response (in mm) which is a measure of the strength of the bird's immune system; and stone mass (in g) which is the average weight of the stones moved by the bird. The T–cell response is essentially the increase in the thickness of the patagium (wing web) in response to the injection of a lectin. A larger T–cell value indicates a stronger immune system response.

These authors conjectured that male black wheatears signal their current health status to their partners by carrying heavy stones. In particular, they conjectured that birds with stronger immune systems would be expected to carry heavier stones. The plot of stone mass versus T–cell response in Figure 7 shows a reasonably strong linear relationship between stone mass and T–cell response. The authors argued that the T–cell response was very precisely measured; thus, it is reasonable to treat T–cell response as the explanatory variable in a simple linear regression model for stone mass.

Table 3. Wheatear data, fitted values, and residuals.

T-cell response (mm) X	stone mass (g) Y	fitted value $\hat{Y}$	residual $Y - \hat{Y}$
.18	6.06	5.7537	0.3063
.20	6.50	5.9540	0.5460
.21	6.28	6.0542	0.2258
.21	9.33	6.0542	3.2758
.25	3.33	6.4549	-3.1249
.25	5.44	6.4549	-1.0149
.25	5.67	6.4549	-0.7849
.25	6.61	6.4549	0.1551
.26	4.61	6.5551	-1.9451
.30	8.00	6.9558	1.0442
.31	7.78	7.0560	0.7240
.33	6.44	7.2563	-0.8163
.34	6.72	7.3565	-0.6365
.37	8.06	7.6570	0.4030
.38	8.11	7.7572	0.3528
.41	9.94	8.0577	1.8823
.43	7.56	8.2581	-0.6981
.43	9.06	8.2581	0.8019
.43	9.17	8.2581	0.9119
.47	6.78	8.6588	-1.8788
.51	9.33	9.0595	0.2705

Figure 7. Plot of stone mass versus T-cell response.



The equation for the fitted regression line for this example (see Figure 8) is

$$\hat{Y} = 3.9505 + 10.0177X,$$

where Y denotes the stone mass in grams and X denotes the T–cell response in mm.

Figure 8. Wheatear data with fitted line.

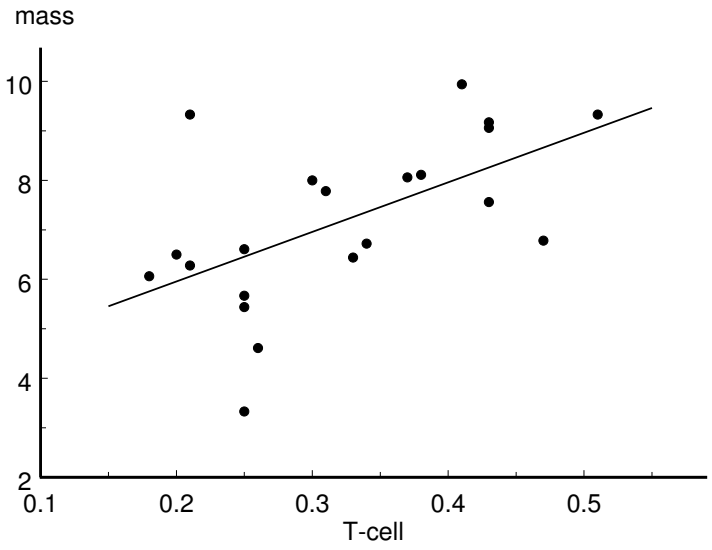
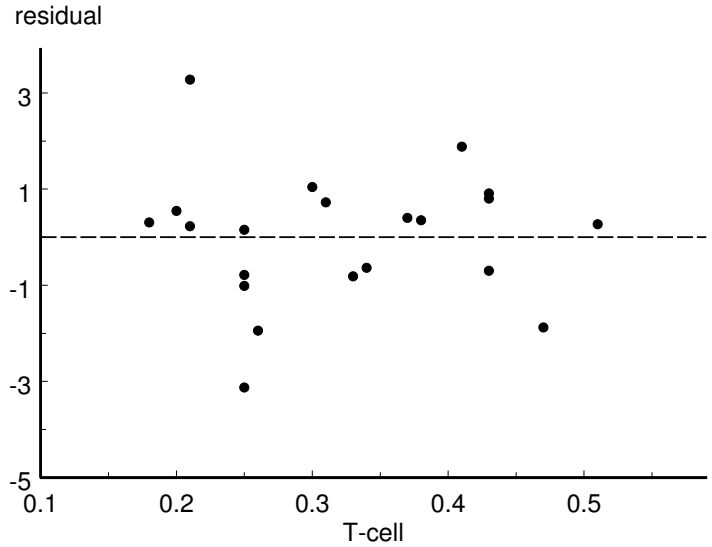


Figure 9. Wheatear example residual plot.



The plot in Figure 8 and the residual plot in Figure 9 show that there are two points which are relatively far away from the fitted line. The observation with $X = .21$ and $Y = 9.33$ has a residual of 3.2758 (see Table 3) and the observation with $X = .25$ and $Y = 3.33$ has a residual of -3.1249. All of the other residuals have magnitudes which are less

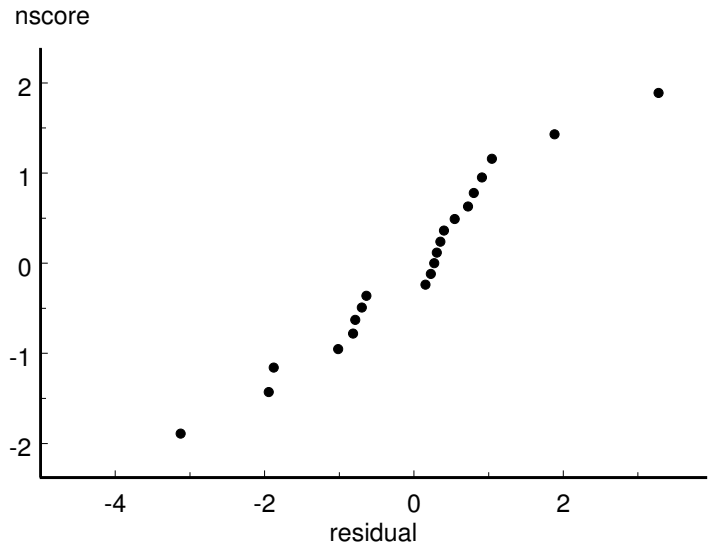
than two. Because of these two mild outliers the coefficient of determination $R^2 = .3330$ is not very large. Notice that even with these two unusual points T–cell response alone still explains 33.3% of the variability in stone mass. If we had data for some other relevant explanatory variables we could fit a more complex regression model which would account for more of the variability in stone mass.

Figure 10. Stem and leaf histogram for wheatear residuals.

In this stem and leaf histogram the stem represents ones and the leaf represents tenths. (grams)

```
-3  1
-2
-1  980
-0  8766
 0  1223345789
 1  08
 2
 3  2
```

Figure 11. Wheatear residuals normal probability plot.



The two mild outliers are apparent in the stem and leaf histogram of Figure 10 and in the normal probability plot of Figure 11. Even with these points there does not seem to be any problem in treating these stone mass data as forming a random sample of size 21 from a normal distribution.

Table 4. Wheatear regression summary statistics.

$n =$	21	error d.f. $=$	19
$\overline{X} =$	.3223	$S_p^2 =$	2.0249
$\overline{Y} =$	7.1800	$\widehat{SE}(\overline{Y}) =$	.3105
y–intercept, $a =$	3.9505	$\widehat{SE}(a) =$	1.0936
slope, $b =$	10.0177	$\widehat{SE}(b) =$	3.2528
$R^2 =$	.3330		

The slope, $b = 10.0177$ g per mm, of the fitted line indicates that if the T–cell response was increased by 1 mm, then we would estimate that the population mean stone mass would increase by 10.0177 grams. Since the T–cell response cannot increase by 1 mm and stay within the range of the data, we might rephrase this by saying that if the T–cell response increased by .1 mm, then we would estimate that the population mean stone mass would increase by 1.00177 grams. We can state, with 95% confidence, that the population slope β is between 3.2096 and 16.8258 g per mm. We can use a test of $H_0 : \beta \leq 0$ versus $H_1 : \beta > 0$ to quantify the evidence in favor of the conjecture that the population mean stone mass is an increasing function of the T–cell response. For this test we have $T_{calc} = 3.08$ giving a P–value of .0031. This provides strong evidence that the population slope is positive so that a stronger T–cell response yields a higher population mean stone mass.

There is some interest here in considering the population mean stone mass for a low T–cell response value (say $X = .25$) and for a high T–cell response value (say $X = .45$). The estimate of the population mean stone mass for $X = .25$ is $\hat{Y}(.25) = 6.4549$ with standard error .3897; and a 95% confidence interval for $\mu(.25)$ goes from 5.6393 to 7.2705 grams. The estimate of the population mean stone mass for $X = .45$ is $\hat{Y}(.45) = 8.4584$ with standard error .5184; and a 95% confidence interval for $\mu(.45)$ goes from 7.3734 to 9.5435 grams.

Chapter 11

Chi–square Tests

11.1 Introduction

In this chapter we will consider the use of chi–square tests (χ^2–tests) to determine whether hypothesized models are consistent with observed data. These tests are based on the χ^2–square statistic which serves as an index of discrepancy between a collection of observed frequencies and a hypothesized collection of expected frequencies. The χ^2–statistic summarizes the differences between the values actually observed and the values we would expect to see if the hypothesized model was correct; with a large χ^2 value indicating that the hypothesized model is not consistent with the observed data. The first step in forming the χ^2–statistic is to find the observed frequencies with which each possible value occurs in the data and the expected frequencies with which each possible value should occur according to the hypothesized model. For each value the difference between the observed frequency and the expected frequency is computed, this difference is then squared and this squared difference is divided by the corresponding expected frequency. These standardized squared differences are then added yielding the χ^2–statistic

$$\chi^2 = \sum \frac{(\text{observed frequency } - \text{ expected frequency})^2}{\text{expected frequency}} ,$$

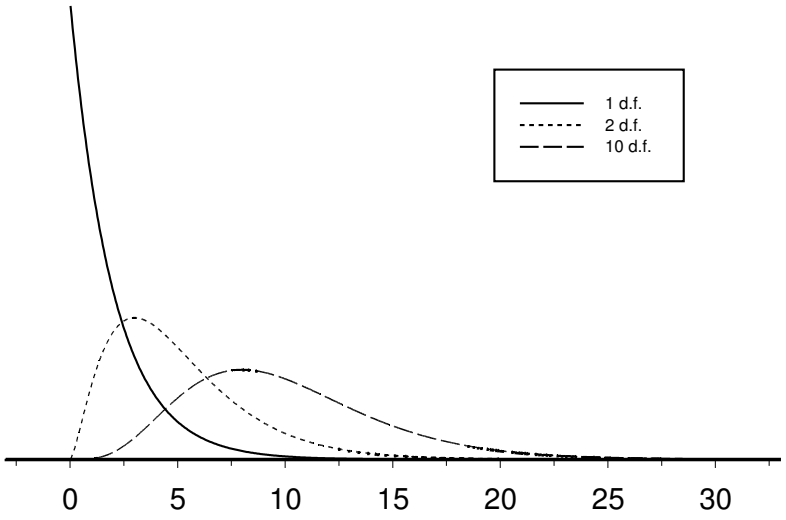
where the sum is over all of the possible values. Large values of this χ^2–statistic indicate evidence that, at least some of, the observed frequencies do not agree with the hypothesized expected frequencies and thus that the hypothesized model may not be correct. That is, large values of the χ^2–statistic indicate that the observed data are not consistent with the hypothesized model.

The χ^2 distributions are skewed to the right with density curves which are positive only for positive values of the variable. Density curves for representative χ^2 distributions are provided in Figure 1. The χ^2 distributions for 1 and 2 degrees of freedom have their mode at zero; for larger degrees of freedom (d.f.) the mode of the χ^2 distribution is located at d.f. $- 2$. Notice that the variability in the χ^2 distribution increases as the degrees of freedom increases. For the χ^2–tests discussed in this chapter a large value of the χ^2–statistic constitutes evidence against the null hypothesis and the P–values for these tests are areas under the appropriate χ^2 density curve to the right of the observed value χ^2_{calc} of the χ^2–statistic.

The χ^2–tests and associated P–values discussed in this chapter are based on large sample approximations which require reasonably large expected frequencies. One rule of

thumb regarding this requirement says that no more than 20% of the expected frequencies should be less than 5 and all of the expected frequencies should be at least 1. If these conditions are not satisfied, you can combine some categories (values of the variable) to increase the expected frequencies which are too small.

Figure 1. Chi–square distribution density curves.



We will consider three different applications of χ^2–tests in this chapter. In Section 11.2 we will consider χ^2–tests for goodness of fit. These goodness of fit tests can be viewed as extensions of the Z–test for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ of Chapter 5 to populations with more than two possible classifications. In Section 11.3 we will consider χ^2–tests for homogeneity. These tests of homogeneity can be viewed as extensions of the Z–test for $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$ of Chapter 6 to two or more populations with two or more possible classifications. Finally in Section 11.4 we will consider χ^2–tests for independence. These tests for independence are used to examine the relationship between two or more classification factors.

Throughout this chapter we will provide details of the computations involved in computing χ^2–statistics. This does not indicate that you should compute these statistics by hand; however, if you choose to do so be sure to avoid rounding at intermediate stages. Some calculators and most statistical programs will compute χ^2–statistics, associated P–values, and other relevant information.

11.2 Chi–square Tests for Goodness of Fit

A χ^2–test for goodness of fit is used to determine whether the outcomes predicted by a hypothesized model are consistent with observed data. The hypothesized model is used to determine the outcomes we would expect to observe and the χ^2–statistic is used to quantify the agreement between the observed outcomes and the expected outcomes. A

small value of the χ^2–statistic indicates that the observed outcomes are in agreement with the outcomes predicted by the hypothesized model (the data are consistent with the model) and a large value indicates inconsistency between the observed data and the hypothesized model.

First consider the χ^2–test for goodness of fit for situations where the hypothesized model completely specifies the probabilities for each of the possible outcomes. More formally, consider a situation where the population units can be categorized into k mutually exclusive and exhaustive classifications and where the model completely specifies the probabilities, $p_1, p_2, \ldots, p_k$, of belonging to these k classifications. The χ^2–test of goodness of fit is used to test the null hypothesis that the k probabilities specified by the model are correct versus the alternative hypothesis that these probabilities are not all correct. The χ^2–test is most easily presented in terms of the observed frequencies (observed counts), $f_1, f_2, \ldots, f_k$, of the k classifications and the hypothetical expected frequencies (expected counts), $F_1, F_2, \ldots, F_k$, predicted by the model. Assuming that the data correspond to a random sample of size n, we can express the expected frequencies in terms of the model probabilities as $F_1 = np_1, F_2 = np_2, \ldots, F_k = np_k$. We will develop the χ^2–test in the context of several examples.

Example. Inheritance in peas (flower color). In Section 5.3 we described a simple Mendelian inheritance model for the color of pea plant flowers arising from crossing two first generation plants. This model hypothesizes that the probability that a plant has red flowers is $p_R = 3/4$ and the probability that a plant has white flowers is $p_W = 1 - p_R = 1/4$. Mendel observed $n = 929$ pea plants arising from a cross of two first generation plants of which 705 plants had red flowers and 224 plants had white flowers. Under the hypothesized model we would expect to see red flowers 3/4 of the time and white flowers 1/4 of the time. Thus, for Mendel's experiment with a total of 929 plants we would expect to see about 696.75 plants with red flowers and about 232.25 plants with white flowers.

We can test the consistency of this model with the data by comparing the observed frequencies of red and white flowered plants to the corresponding expected frequencies. The first step in this comparison is to find the differences between the observed and expected frequencies of plants for each of the two colors. In this example we have differences of $705 - 696.75 = 8.25$ (red) and $224 - 232.25 = -8.25$ (white.) These differences add to zero, since both the observed and expected frequencies add to 929. The second step is to square each difference and standardize it by dividing by the corresponding expected frequency. This standardization gives $68.0625/696.75 = .0977$ (red) and $68.0625/232.25 = .2931$ (white.) Adding these standardized squared differences gives the χ^2–statistic $\chi^2_{calc} = .0977 + .2931 = .3908$. These computations are summarized in Table 1.

Table 1. Pea plant flower color example.

flower color	observed frequency	expected frequency	obs − exp	$(\text{obs}-\text{exp})^2/\text{exp}$
red	705	696.75	8.25	.0977
white	224	232.25	-8.25	.2931
total	929	929		$\chi^2_{calc} = .3908$

A large value of the χ^2–statistic indicates evidence against the null hypothesis that the model is valid $H_0 : p_R = 3/4$ and $p_W = 1/4$ and in favor of the alternative hypothesis that the model is not valid H_1 : it is not true that $p_R = 3/4$ and $p_W = 1/4$. We can determine whether $\chi^2_{calc} = .3908$ is large by computing the relevant P–value. The P–value for this χ^2–test is the probability of observing a value of χ^2 as large or larger than the calculated value $\chi^2_{calc} = .3908$ computed using the appropriate χ^2 distribution. In a situation like the present example, where there are k categories or classifications and the model completely specifies the k corresponding probabilities, the appropriate χ^2 distribution is the χ^2 distribution with $k-1$ degrees of freedom. In this example there are $k = 2$ possible classifications (red or white) and the model completely specifies the two corresponding probabilities ($p_R = 3/4$ and $p_W = 1/4$), so the χ^2 distribution with $k - 1 = 1$ degree of freedom is used to compute the P–value. With $\chi^2_{calc} = .3908$ and one degree of freedom we get the P–value $P(\chi^2 \geq .3908) = .5310$. This P–value is quite large and we are not able to reject the null hypothesis; therefore, we conclude that the observed data are consistent with Mendel's model.

In this example there are $k = 2$ classifications and $p_W = 1 - p_R$, thus the hypotheses specified in terms of p_R and p_W above can be written more simply as $H_0 : p_R = 3/4$ and $H_1 : p_R \neq 3/4$. In section 5.3 we used the normal approximation to perform a Z–test for these hypotheses. The χ^2–test presented above is actually equivalent to this Z–test. To see this equivalence consider the Z–test for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$; for this test we have

$$\begin{aligned} Z^2 &= \frac{(\hat{p}-p_0)^2}{p_0(1-p_0)/n} = \frac{(n\hat{p}-np_0)^2}{np_0(1-p_0)} \\ &= [(1-p_0)+p_0]\frac{(n\hat{p}-np_0)^2}{np_0(1-p_0)} \\ &= \frac{(n\hat{p}-np_0)^2}{np_0} + \frac{(n[1-\hat{p}]-n[1-p_0])^2}{n(1-p_0)} = \chi^2; \end{aligned}$$

and a value of Z which is far away from zero corresponds to a large value of $Z^2 = \chi^2$. Furthermore, it can be shown that the square of a standard normal variable follows the χ^2 distribution with one degree of freedom; and thus these two approaches are equivalent.

The χ^2–test is of more interest when there are three or more classifications, since there is no Z–test in these cases. The χ^2–test for an example with $k = 4$ classifications is developed below.

Example. Inheritance in peas (seed shape and color). We will now consider the Mendelian inheritance model for two independently inherited characteristics. In particular we will consider the characteristics seed shape, with possible shapes of round (R, dominant) and wrinkled (r, recessive), and seed color, with possible colors of yellow (Y, dominant) and green (y, recessive). If a $RRYY$ genotype plant with round yellow seeds is crossed with a $rryy$ genotype plant with wrinkled green seeds, the offspring will all have round yellow seeds and genotype $RrYy$. If two of the resulting $RrYy$ genotype plants with round yellow seeds are crossed, there are 16 equally likely possible genotypes. The nine genotypes $RRYY, RRYy, RRyY, RrYY, RrYy, RryY, rRYY, rRYy, rRyY$ yield round yellow seeds; the three genotypes $rrYY, rrYy, rryY$ yield wrinkled yellow seeds; the three genotypes $RRyy, Rryy, rRyy$ yield round green seeds; and, the single genotype $rryy$ yields wrinkled green seeds. The facts that these 16 possible genotypes are equally likely and each plant possesses only one genotype yield the probability distribution summarized in Table 2.

Table 2. Pea plant seed shape/color distribution.

shape/color	probability
round yellow	9/16
wrinkled yellow	3/16
round green	3/16
wrinkled green	1/16

The results of one of Mendel's experiments regarding seed shape and color, with $n = 556$ plants, are summarized in Table 3. Table 3 also contains the expected frequencies, computed using the distribution of Table 2, and the computations leading to the χ^2 statistic. In this example there are $k = 4$ classifications and the P–value $P(\chi^2 \geq .4700) = .9254$

Table 3. Pea plant seed shape and color example.

shape/color	observed frequency	expected frequency	obs − exp	$(\text{obs} - \text{exp})^2/\text{exp}$
round yellow	315	312.75	2.25	.0162
wrinkled yellow	101	104.25	-3.25	.1013
round green	108	104.25	3.75	.1349
wrinkled green	32	34.75	-2.75	.2176
total	556	556		$\chi^2_{calc} = .4700$

for the χ^2–test is computed using the χ^2 distribution with $k-1=3$ degrees of freedom. In this example, the P–value is quite large and we are not able to reject the null hypothesis; therefore, we conclude that the observed data are consistent with Mendel's model.

In both of the preceding examples the data are consistent with the hypothesized model. The following example, which is also concerned with a Mendelian inheritance model, illustrates a situation where the data are not consistent with the model.

Example. Inheritance in maize (leaf characteristics). This example is taken from Snedecor and Cochran (1980), the original source is Lindstrom (1918) *Cornell Agr. Exp. Sta. mem. 13*. Lindstrom crossed two types of maize (corn) plants and classified the resulting plants into four categories based on the appearance of the leaves. The Mendelian model for this example is analogous to the model of the pea plant seed shape and color example with respective probabilities of $9/16, 3/16, 3/16$, and $1/16$. Thus the model predicts that the four leaf types should occur in a ratio of $9:3:3:1$. The data and the computations are summarized in Table 4.

Table 4. Maize leaf type example.

leaf type	observed frequency	expected frequency	obs − exp	$(\text{obs}-\text{exp})^2/\text{exp}$
green	773	731.813	41.1875	2.3181
golden	231	243.938	-12.9375	0.6862
green–striped	238	243.938	-5.9375	0.1445
green–golden–striped	59	81.313	-22.3125	6.1226
total	1301	1301		$\chi^2_{calc} = 9.2714$

In this example $\chi^2_{calc} = 9.2714$ is large indicating disagreement between the model and the data. The P–value .0259, computed using the χ^2 distribution with 3 degrees of freedom, is small enough to allow us to conclude that Lindstrom's data are not consistent with the Mendelian model which predicts frequencies in the ratio of $9:3:3:1$.

Examination of the four terms we added to get χ^2_{calc} indicates that the green–golden–striped term 6.1226 is large relative to the other terms. Thus the evidence against the model seems to be due to the fact that the observed frequency of green–golden–striped plants 59 is much smaller than the expected frequency 81.313. Lindstrom argued that this discrepancy could be explained by "the weakened condition of the last three classes due to their chlorophyll abnormality". In particular, he noted that the plants in the green–golden–striped class were not very vigorous (did not grow well). This suggests that the evidence against the model may be due to the fact that some of the green–golden–striped plants did not survive long enough to be counted. Therefore, we might wonder whether

our rejection of the 9 : 3 : 3 : 1 model can be attributed to the poor survivorship of the green–golden–striped plants. We will now perform an exploratory analysis to address this question.

First consider the 1242 plants in the first three classifications. According to the model the frequencies for these three classifications should be in the ratio 9 : 3 : 3. The computations for a χ^2 test for this subset of the original data are demonstrated in Table 5. For this subset of the original data we have $\chi^2_{calc} = 2.6914$ with $3 - 1 = 2$ degrees of freedom which gives a P–value of .2604. Therefore, there is evidence that the frequencies in the first three classes are consistent with the predicted ratio of 9 : 3 : 3.

Table 5. Maize leaf type example, 9:3:3 model.

leaf type	observed frequency	expected frequency	obs − exp	$(\text{obs} - \text{exp})^2/\text{exp}$
green	773	745.2	27.8	1.0371
golden	231	248.4	-17.4	1.2188
green–striped	238	248.4	-10.4	0.4354
total	1242	1242		$\chi^2_{calc} = 2.6914$

This test and the fact that the observed frequency of green–golden–striped plants is much smaller than expected suggest that the reason that the original data do not agree with the model may be poor survivorship of the green–golden–striped plants, since the data for the other classes do agree with the model.

In some situations, like the two examples which follow, the hypothesized model does not completely specify the probabilities for the k possible outcomes and it is necessary to estimate these probabilities before performing the χ^2 goodness of fit test.

Example. Radioactive disintegrations. This example is taken from Feller (1957), p. 149 and Cramér (1946) p. 436. In a famous experiment by Rutherford, Chadwick, and Ellis (*Radiations from Radioactive Substances*, Cambridge, 1920) a radioactive substance was observed during 2608 consecutive time intervals of length 7.5 seconds each. The number of particles reaching a counter was recorded for each period giving the results summarized in Table 6.

The Poisson distribution, as discussed in Chapter 4a, provides a plausible model for the number of particles, X, observed in this experiment. Therefore, we will perform a χ^2 goodness of fit test to see whether a Poisson distribution is suitable as a model for the distribution of the observed number of particles in this experiment. The Poisson model places no upper bound on the number of particles which could be observed; so, for this

test, we will use "10 or more particles" as the largest possible "value" of the variable. The Poisson distribution with parameter λ specifies probabilities of the form

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

for $x = 0, 1, 2, \ldots$ Notice that this probability function does not completely specify the

Table 6. Radioactive disintegrations data.

number	observed frequency
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
12	2
total	2608

probabilities of the possible values of X, since there is an unknown parameter λ in the formula. Therefore, to perform the χ^2–test we first need to use the data to estimate λ. Since λ is the mean of the Poisson distribution, we can use the sample mean 3.8704 as an estimate of λ and use the formula from above to determine the expected frequencies. Thus, for $x = 0, 1, \ldots, 9$, we compute the expected frequencies using the formula

$$2608 \left(\frac{3.8704^x}{x!} e^{-3.8704} \right)$$

and we subtract the sum of these expected frequencies from 2608 to find the expected frequency for $X \geq 10$. The observed and expected frequencies and the terms used to calculate the χ^2–statistic are summarized in Table 7. Because we estimated the parameter λ of the hypothesized Poisson distribution we need to reduce the degrees of freedom for the χ^2–test by one. For this example we have $\chi^2_{calc} = 12.8815$ with $k - 2 = 9$ degrees of freedom which gives a P–value of .1680. Since this P–value is not small we can conclude that a Poisson model with $\lambda = 3.8704$ provides a reasonable model for the number of radioactive disintegrations observed in this experiment.

Table 7. Radioactive disintegrations.

number	observed frequency	expected frequency	obs − exp	$(\text{obs} - \text{exp})^2/\text{exp}$
0	57	54.3769	2.6231	0.1265
1	203	210.4604	-7.4604	0.2645
2	383	407.2829	-24.2829	1.4478
3	525	525.4491	-.4491	0.0004
4	532	508.4244	23.5756	1.0932
5	408	393.5610	14.4390	0.5297
6	273	253.8730	19.1270	1.4410
7	139	140.3700	-1.3700	0.0134
8	45	67.9110	-22.9110	7.7294
9	27	29.2047	-2.2047	0.1664
≥ 10	16	17.0865	-1.0865	0.0691
total	2608	2608		$\chi^2_{calc} = 12.8815$

Example. Bacteria counts. This example is taken from Feller (1957), p.153. The original source is T. Matuszewsky, J. Supinska, and J. Neyman (1936), *Zentralblatt für Bakteriologie, Parasitenkunde und Infektionskrankrankheiten*, II Abt., **95**.

Table 8. Bacteria counts data.

number	observed frequency
0	5
1	19
2	26
3	26
4	21
5	13
6	8
total	118

A Petri dish with bacteria colonies was examined under a microscope. The dish was divided into small squares and the number of bacteria colonies, visible as dark spots, was recorded for each square. The data are given in Table 8. If the bacteria colonies were randomly distributed over the Petri dish, without being clustered together, then the Poisson model should hold. The sample mean number of bacteria colonies is 2.9322 and, as in the preceding example, we can use this sample mean to estimate the parameter λ of the Poisson model.

Table 9. Bacteria counts.

number	observed frequency	expected frequency	obs − exp	$(\text{obs} - \text{exp})^2/\text{exp}$
0	5	6.2870	-1.2870	.2635
1	19	18.4347	.5653	.0173
2	26	27.0272	-1.0272	.0390
3	26	26.4164	-.4164	.0066
4	21	19.3645	1.6354	.1381
5	13	11.3562	1.6438	.2380
≥ 6	8	9.1140	-1.1141	.1362
total	118	118		$\chi^2_{calc} = .8386$

The observed and expected frequencies and the terms used to calculate the χ^2–statistic are summarized in Table 9. Again, since we estimated the parameter λ of the hypothesized Poisson distribution, we need to reduce the degrees of freedom for the χ^2–test by one. For this example we have $\chi^2_{calc} = .8386$ with $k - 2 = 5$ degrees of freedom which gives a P–value of .9745. Since this P–value is very large, we can conclude that a Poisson model with $\lambda = 2.9322$ provides a reasonable model for the number of bacteria colonies as observed in this experiment. This indicates that the conjecture that the bacteria colonies are randomly distributed over the Petri dish, without being clustered together, is consistent with the observations.

11.3 Chi–square Tests for Homogeneity

We will now consider χ^2–tests for the homogeneity of two or more population distributions. These tests can be viewed as generalizations of the Z–test of equality of two population proportions of Section 6.2 to allow for more than two populations or more than two possible classifications.

A probability distribution for a qualitative variable with k possible values corresponding to k mutually exclusive classifications can be represented by a collection $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ of k probabilities which sum to one. Given m such probability distributions, we have m collections of probabilities $\mathbf{p}_1 = (p_{11}, p_{12}, \ldots, p_{1k})$, $\mathbf{p}_2 = (p_{21}, p_{22}, \ldots, p_{2k})$, ..., and $\mathbf{p}_m = (p_{m1}, p_{m2}, \ldots, p_{mk})$ as shown in Table 10. The null hypothesis of homogeneity of these m distributions, $H_0 : \mathbf{p}_1 = \mathbf{p}_2 = \cdots = \mathbf{p}_m$, specifies that the probability of observing a unit in a particular classification is the same for all m of the populations, *i.e.*, for each classification $j = 1, 2, \ldots, k$, we have $p_{1j} = p_{2j} = \cdots = p_{mj}$.

Table 10. Notation for m populations and k classifications.

	classification probabilities				
population	1	2	...	k	sum
1	p_{11}	p_{12}	...	p_{1k}	1
2	p_{21}	p_{22}	...	p_{2k}	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
m	p_{m1}	p_{m2}	...	p_{mk}	1

Suppose that m independent random samples of sizes $n_1, n_2, \ldots,$ and n_m are obtained from these m population distributions and let f_{ij} denote the observed frequency of units in classification j for the sample from population i as shown in Table 11. Under the null hypothesis we would expect the m collections of k observed frequencies in each row (sample) of Table 11 to be the same (no difference from row to row).

We can use the combined observed frequencies $F_1, F_2, \ldots, F_k$ in the last line of Table 11, obtained by adding the corresponding frequencies in the respective columns, and the combined sample size $n = n_1 + n_2 + \cdots + n_m$ to form estimates of the frequencies we would expect to observe under the null hypothesis of homogeneity. We first compute the estimates $\hat{p}_1 = F_1/n$, $\hat{p}_2 = F_2/n$, ..., and $\hat{p}_k = F_k/n$ of the assumed common classification probabilities $p_1, p_2, \ldots,$ and p_k and then we multiply this collection of $\hat{p}'s$ by the respective sample sizes to get the expected frequencies for each population (row of the table). The P–value for the resulting χ^2–statistic, which is based on the $m \times k$ observed and expected frequencies, is obtained from the χ^2 distribution with $(m-1)(k-1)$ degrees of freedom.

Table 11. Data for m populations and k classifications.

population	observed frequencies				sample
	1	2	...	k	size
1	f_{11}	f_{12}	...	f_{1k}	n_1
2	f_{21}	f_{22}	...	f_{2k}	n_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
m	f_{m1}	f_{m2}	...	f_{mk}	n_m
combined	F_1	F_2	...	F_k	n

We will first apply this χ^2–test of homogeneity to an example with $m = 2$ dichotomous $(k = 2)$ populations.

Example. Cocaine addiction. This example is based on a study of D.M. Barnes (1988), *Science*, **241**, 1029–1030, as described in Moore (1995). This study was conducted to compare two antidepressants as treatments for cocaine addiction. In particular, the researchers wanted to compare the effects of the antidepressant desipramine with the effects of lithium (a standard treatment for cocaine addiction.) A group of 48 chronic cocaine users was randomly divided into two groups of 24. One group was treated with desipramine and the other was treated with lithium. The subjects were tracked for three years and the number of subjects who relapsed into cocaine use during this period was recorded. The data are summarized as observed frequencies in Table 12.

Table 12. Cocaine example: observed and expect frequencies.

observed frequency				expected frequency			
	relapsed				relapsed		
treatment	yes	no	total	treatment	yes	no	total
desipramine	10	14	24	desipramine	14	10	24
lithium	18	6	24	lithium	14	10	24
combined	28	20	48				

For this example we can view the data as independent random samples of size 24 from dichotomous populations with population success probabilities p_D and p_L, where p_D is the probability that one of these 48 cocaine users would relapse into cocaine use if all 48 users were treated with desipramine and p_L is the analogous probability assuming that all 48 users were treated with lithium. We can use a χ^2–test to test the null hypothesis $H_0 : p_D = p_L$ that the probability of relapse is the same for both treatments versus the alternative hypothesis $H_1 : p_D \neq p_L$ of different probabilities of relapse. Under the null hypothesis we would expect to observe the same relapse proportions under each treatment; furthermore, since 28 of the 48 users suffered a relapse we can use the combined sample relapse proportion $\hat{p} = 28/48$ as our estimate of the common relapse probability we would expect to observe under the null hypothesis. The expected frequencies in Table 12 were computed using this $\hat{p}$ as the estimated common relapse probability and the sample sizes, which are both 24. The differences between the observed and expected frequencies and the four components of the χ^2–statistic are given in Table 13.

The P–value for $\chi^2_{calc} = 192/35 = 5.487$, computed using the χ^2 distribution with $(2-1)(2-1) = 1$ degrees of freedom, is $P(\chi^2 \geq 5.487) = 0.0192$. This small P–value allows us to reject the null hypothesis of homogeneity and conclude that $p_D \neq p_L$ indicating that the probability of relapse is not the same when a user is treated with desipramine as when the user is treated with lithium. Since this example involves two dichotomous populations, we could have used the Z–test of Section 6.2, which is equivalent to the χ^2 test from above

in this situation, to perform this test. More importantly, since we have two dichotomous populations, we can use the Z–interval of Section 6.1 to quantify the size and direction of the difference between p_D and p_L. The sample success proportions are $\hat{p}_D = .4167$ and $\hat{p}_L = .75$ and the 95% confidence interval for $p_L - p_D$ is $(.0708, .5959)$. Hence, we are 95% confident that treating one of these 48 cocaine users with desipramine instead of lithium would reduce the probability of relapse by at least .0708 and as much as .5959.

Table 13. Cocaine example: chi–square computations.

obs - exp			$(\text{obs - exp})^2/\text{ exp}$		
	relapsed			relapsed	
treatment	yes	no	treatment	yes	no
desipramine	-4	4	desipramine	16/14	16/10
lithium	4	-4	lithium	16/14	16/10
				$\chi^2_{calc} = 192/35 = 5.487$	

The next example with $m = 3$ populations and $k = 3$ categories will be used to demonstrate the extension of the χ^2–test of homogeneity to situations with three or more populations and three or more categories.

Example. Attitudes of School Children. This example is based on a study described by Chase and Dummer (1992), *Research Quarterly for Exercise and Sport*, **63**, 418–424, as described in DeGroot and Schervish (2002). This study was conducted to examine the attitudes of school–aged children in Michigan. Three independent random samples of children were obtained. A sample of 149 children from rural areas, a sample of 151 children from suburban areas, and a sample of 178 children from urban areas. Each child was asked which of the following was most important to them: good grades, athletic ability, or popularity. The observed frequencies are given in Table 14 and the expected frequencies, based on the combined probability estimates $247/478 = .5167$, $90/478 = .1883$, and $141/478 = .2950$ and the sample sizes 149, 151, and 178 are given in Table 15.

Table 14. Attitude example: observed frequencies.

sample	good grades	athletic ability	popularity	sample size
rural	57	42	50	149
suburban	87	22	42	151
urban	103	26	49	178
combined	247	90	141	478

Table 15. Attitude example: expected frequencies.

sample	good grades	athletic ability	popularity	sample size
rural	76.9937	28.0544	43.9519	149
suburban	78.0272	28.4310	44.5418	151
urban	91.9791	33.5146	52.5063	178

The differences between the observed and expected frequencies and the nine components of the χ^2–statistic are given in Table 16. The P–value for $\chi^2_{calc} = 18.8276$, computed using the χ^2 distribution with $(3-1)(3-1) = 4$ degrees of freedom, is $P(\chi^2 \geq 18.8276) = 0.0008$. This very small P–value indicates very strong evidence that the attitude distributions (the three probabilities for the three choices given to these children) are not the same for the three areas.

Table 16. Attitude example: chi–square computations.

	obs - exp			(obs - exp)^2/exp		
sample	good grades	athletic ability	popularity	good grades	athletic ability	popularity
rural	-19.9937	13.9456	6.0481	5.19197	6.93225	0.83227
suburban	8.9728	-6.4310	-2.5418	1.03184	1.45466	0.14505
urban	11.0209	-7.5146	-3.5063	1.32053	1.68493	0.23414

The two largest (obs - exp)^2/exp terms, 5.19197 for the rural–good grades category and 6.93225 for the rural–athletic ability category, are much larger than the other terms. This fact and the observed relative frequencies given in Table 17 suggest that the attitude distributions might be the same for the suburban and urban children but different for the rural children.

Table 17. Attitude example: observed relative frequencies.

sample	good grades	athletic ability	popularity
rural	.3826	.2819	.3356
suburban	.5762	.1457	.2781
urban	.5787	.1461	.2753

The χ^2–statistic based on the data for suburban and urban children only is $\chi^2_{calc} = .0034$ with $(2-1)(3-1) = 2$ degrees of freedom, which gives a P–value of .9983 and supports the contention that the attitude distribution is the same for the suburban children as it is for the urban children. Furthermore, if we combine the suburban sample and the urban

sample to form a nonrural sample of size 329, the χ^2–statistic for comparing the rural and nonrural samples is $\chi^2_{calc} = 18.8243$ with $(2-1)(3-1) = 2$ degrees of freedom and the P–value is less than .0001, confirming our conjecture that the attitude distribution for the rural children is not the same as that for the nonrural children.

11.4 Chi–square Tests for Independence

A χ^2–test for independence is used to determine whether two or more qualitative classification factors are independent. In this section we will restrict our attention to crossed classifications of units with respect to two qualitative classification factors. Two classification factors, A and B, are said to be independent, if the conditional probabilities for the levels of factor A (respectively, factor B), obtained by fixing the level of factor B (factor A), are the same regardless of the level at which factor B (factor A) is fixed. To avoid complex notation we will describe independence and develop the χ^2–test for independence in the context of the following example.

Example. Hawaiian blood types. This example uses data from A.E. Mourant, *et al.*, *The Distribution of Blood Groups and Other Polymorphisms*, Oxford University Press, London, 1976. The Blood Bank of Hawaii cross classified 145,057 individuals according to their blood type (A, AB, B, O) and their ethnic group (Hawaiian, Hawaiian–Chinese, Hawaiian–White, White). The frequencies for each of the 16 combinations of the 4 levels of these two qualitative classification factors are given in Table 18.

Table 18. Blood type and ethnic group observed frequencies.

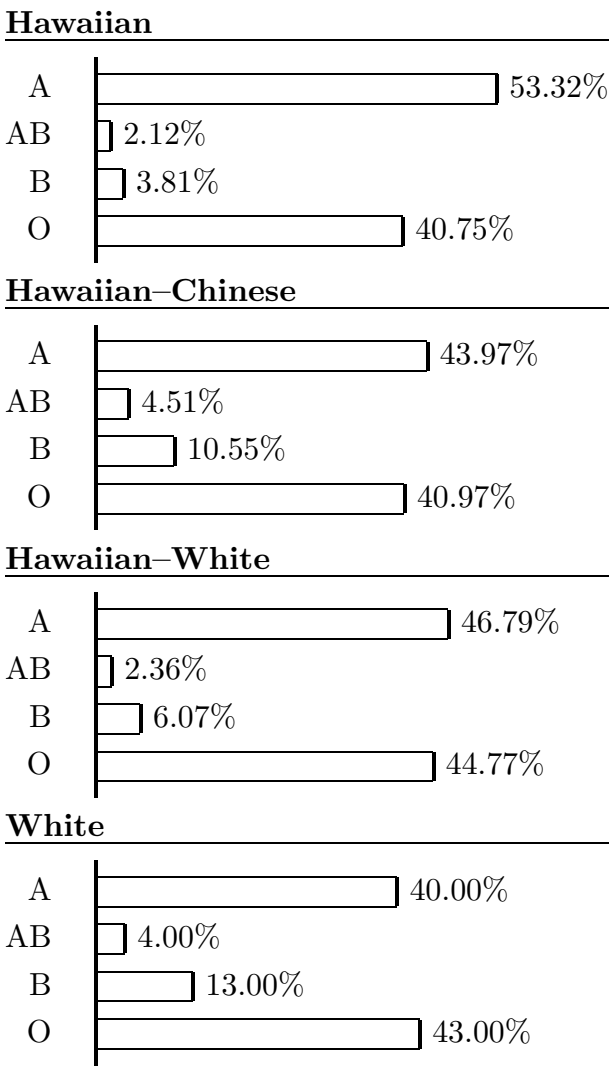
	ethnic group				
blood type	Hawaiian	Hawaiian–Chinese	Hawaiian–White	White	total
A	2490	2368	4671	50008	59537
AB	99	243	236	5001	5579
B	178	568	606	16252	17604
O	1903	2206	4469	53759	62337
total	4670	5385	9982	125020	145057

The question we want to consider here is whether the distribution of blood types is independent of the distribution of the ethnic groups. If the distribution of blood types is the same for each of the four ethnic groups, then classification with respect to blood type is independent of classification with respect to ethnic group. Furthermore, independence of two factors is symmetric so that if the distribution of blood types is independent of the

distribution of ethnic groups, then it also follows that the distribution of ethnic groups is independent of the distribution of blood types.

Under the hypothesis of independence the theoretical conditional distributions of blood type are the same for each ethnic group. The conditional distributions of blood type for each ethnic group summarized in Figure 2 show some evidence that the distributions of blood type are not the same for these ethnic groups indicating dependence between classification with respect to blood type and classification with respect to ethnic group.

Figure 2. Conditional distributions of blood type by ethnic group.



We can compute the expected frequencies for this example the same way we did for the χ^2–tests of homogeneity in Section 11.3. The deviations between the observed and expected frequencies and the 16 terms which are summed to give the χ^2–statistic are given in Table 19. In this example several of the χ^2 terms are large indicating where the hypothesis of independence is not supported by these data. The χ^2–statistic for testing the

independence of blood type and ethnic group is $\chi^2_{calc} = 1078.6036$ with $(4-1)(4-1) = 9$ degrees of freedom and the P–value is less than .0001. Therefore, there is very strong evidence against the null hypothesis of independence. We can conclude that the data collected by the Blood Bank of Hawaii are clearly inconsistent with the hypothesis of independence and that the distribution of blood types is not the same for these four ethnic groups.

Table 19. Hawaiian blood type example chi–square information.

The first number is the deviation (obs - exp) and the number in parentheses is the χ^2 term (obs - exp)^2/exp.

	ethnic group			
blood type	Hawaiian	Hawaiian–Chinese	Hawaiian–White	White
A	573.25	157.79	574	-1305
	(171.45)	(11.265)	(80.419)	(33.191)
AB	-80.61	35.889	-147.9	192.64
	(36.179)	(6.2189)	(56.989)	(7.7177)
B	-388.7	-85.52	-605.4	1079.7
	(266.65)	(11.191)	(302.56)	(76.83)
O	-103.9	-108.2	179.32	32.729
	(5.3783)	(5.055)	(7.4962)	(.0199)

Chapter 12
Comparing Two or More Means

12.1 Introduction

In Chapter 8 we considered methods for making inferences about the relationship between two population distributions based on the relationship between the means of these distributions. In many situations interest centers on the relationship among more than two population distributions. Therefore, in this chapter we consider methods of inference for comparing two or more population distributions based on the relationships among the corresponding population means.

We will restrict our attention to situations where the population distributions (density curves) of $k \geq 2$ continuous variables, $Y_1, Y_2, \ldots$, and Y_k, are identical except for their locations on the number line. This generalizes the **shift assumption** of the two population problem to the $k \geq 2$ population problem. Under this shift assumption inferences for comparing the k population distributions reduce to inferences for comparing the k population means. As in the two population case, when the shift assumption is not valid we must be careful about how we interpret an inference about the relationships among the population means.

We will restrict our attention to methods which are appropriate when the data comprise k independent random samples: a random sample of size n_1 from a population with population mean μ_1 (the Y_1 sample); a random sample of size n_2 from a population with population mean μ_2 (the Y_2 sample); $\ldots$, and a random sample of size n_k from a population with population mean μ_k (the Y_k sample). The assumption that these random samples are independent basically means that the method used to select the random sample from a particular population is not influenced by the method used to select the random sample from any other population.

We will use the following small example to clarify the definitions and computations introduced in this chapter. You should use a suitable calculator or computer program to perform these computations.

Example. Potato leafhopper survival. D. L. Dahlman (M.S. thesis, Iowa State University, 1963) studied the survival and behavioral responses of the potato leafhopper *Empoasca Fabae* (Harris) on synthetic media. The data given in Table 1 are survival times (in days) defined as the number of days until 50% of the insects in a cage were dead. This study was conducted using a completely randomized experiment design with two cages (units) assigned to each of four treatment groups ($n_1 = n_2 = n_3 = n_4 = 2$). That is, the 8 cages were randomly assigned to the 4 treatments so that there were two cages in each treatment group. The treatments consisted of four modifications of the basic 2%

agar synthetic feeding medium. The treatments were a control (2% agar), 2% agar plus fructose, 2% agar plus glucose, and 2% agar plus sucrose, respectively.

Table 1. Potato Leafhopper Data.

treatment	survival time
control	2.3
control	1.7
fructose	2.1
fructose	2.3
glucose	3.0
glucose	2.8
sucrose	3.6
sucrose	4.0

We can define the four population means by imagining what would have happened if all of the eight cages were assigned to a particular treatment group. For example, we can define the control population mean $\mu_1 = \mu_C$ as the mean survival time we would have observed if all 8 cages had been assigned to the control group; we can define the fructose population mean $\mu_2 = \mu_F$ as the mean survival time we would have observed if all 8 cages had been assigned to the fructose group; and so on. The notation we will use in the sequel is summarized in Table 2.

Table 2. Potato Leafhopper Population Means.

treatment:	control	fructose	glucose	sucrose
population mean:	μ_C	μ_F	μ_G	μ_S

12.2 Comparing the means of k normal populations

In this section we consider inferences about the relationships among k normal means. First we discuss the analysis of variance (ANOVA) and the overall F–test; then we consider the sequential use of F–tests for comparing nested models; and finally we discuss simultaneous confidence interval estimates for linear combinations of means.

12.2a Assumptions, notation, and the overall F–test

In order to develop inferential methods we need to make an assumption about the form of the population distributions of the $Y's$. We will assume that the k population distributions are normal distributions with a common population variance σ^2 (common population standard deviation σ). Thus, the population distribution of Y_1 is a normal

distribution with population mean μ_1 and population variance σ^2; the population distribution of Y_2 is a normal distribution with population mean μ_2 and population variance σ^2; ..., and the population distribution of Y_k is a normal distribution with population mean μ_k and population variance σ^2. As stated in the introduction, we will assume that the data comprise k independent random samples. Notice that we are assuming that the k population variances are equal which, together with the normality assumption, implies that the shift assumption is valid.

First consider the question of whether the k means $\mu_1, \ldots, \mu_k$ are all equal. We can address this question by performing a hypothesis test for the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ versus the alternative hypothesis that at least two of the k means are different ($H_1 :$ it is not true that $\mu_1 = \cdots = \mu_k$). Notice that this alternative hypothesis specifies that the k means are not all equal, it does not specify how the means differ and, in particular, it does not specify that there are k distinct means. We will motivate the method used to perform this hypothesis test about the k means as a comparison of two estimators of the common population variance σ^2.

To make the notation clear we will need to use double subscripts on the observations. As indicated in Table 3, we will let Y_{ij} denote the j^{th} observation in the i^{th} group (i^{th} sample), for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$, and we will let $\overline{Y}_i$ denote the sample mean for the i^{th} group.

Table 3. Notation for the k group (sample) problem.

group	data	sample mean	population mean
group 1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	$\overline{Y}_1$	μ_1
group 2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$	$\overline{Y}_2$	μ_2
.	.	.	.
.	.	.	.
.	.	.	.
group k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$	$\overline{Y}_k$	μ_k

The pooled estimator S_p^2 of the common variance σ^2 for the model with k population means $\mu_1, \ldots, \mu_k$ is the natural extension of the pooled variance estimator of the two sample case to k samples. That is, S_p^2 is the sum of the squared deviations of the observations from their respective group sample means divided by the appropriate degrees of freedom which is $n - k = (n_1 - 1) + \cdots + (n_k - 1)$, where $n = n_1 + \cdots + n_k$ is the total number of observations. The numerator of S_p^2, denoted by SS(within the k groups), is the sum of

squares within the k groups (the sum of squared deviations of the observations within each group from their respective group sample mean). In symbols we have

$$S_p^2 = \frac{\text{SS(within the } k \text{ groups)}}{n-k}, \text{ with}$$

$$\begin{aligned}\text{SS(within the } k \text{ groups)} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_i\right)^2 \\ &= \sum_{j=1}^{n_1}\left(Y_{1j}-\overline{Y}_1\right)^2 + \cdots + \sum_{j=1}^{n_k}\left(Y_{kj}-\overline{Y}_k\right)^2.\end{aligned}$$

The pooled variance estimator S_p^2 is a valid (unbiased) estimator of the common variance σ^2 when the k group means $\mu_1, \ldots, \mu_k$ are distinct and also when some or all of the means are equal. The computations for finding S_p^2 described above are illustrated for the potato leafhopper data in Table 4.

Table 4. Potato leafhopper deviations from treatment means.

treatment	observation	treatment mean	deviation from mean	squared deviation from mean
control	2.3	2.0	.3	.09
control	1.7	2.0	-.3	.09
fructose	2.1	2.2	-.1	.01
fructose	2.3	2.2	.1	.01
glucose	3.0	2.9	.1	.01
glucose	2.8	2.9	-.1	.01
sucrose	3.6	3.8	-.2	.04
sucrose	4.0	3.8	.2	.04
				sum of squared deviations = .3
				$S_p^2 = .3/4 = .075$

Under the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ we can view the k random samples as constituting one random sample of size $n = n_1 + \cdots + n_k$ from a normal population with population variance σ^2. Therefore, when H_0 is true we can estimate the common variance σ^2 using the squared deviations of the observations from the overall sample mean $\overline{Y}$ ($\overline{Y}$ is the average of all n observations and in terms of the k sample means, $\overline{Y}_1, \ldots, \overline{Y}_k$, $\overline{Y} = (n_1\overline{Y}_1 + \cdots + n_k\overline{Y}_k)/n$). The variance estimator S_0^2 is the sum of the squared deviations of the observations from the overall sample mean divided by the appropriate degrees of freedom which is $n-1$. The numerator of S_0^2, denoted by SS(about the overall mean),

is the sum of squares about the overall mean (the sum of the squared deviations of the observations from the overall sample mean). In symbols we have

$$S_0^2 = \frac{\text{SS(about the overall mean)}}{n-1}, \text{ with}$$

$$\text{SS(about the overall mean)} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}\right)^2,$$

and we see that S_0^2 is simply the usual one sample estimator of the variance computed ignoring the existence of the k groups.

The variance estimator S_0^2 is a valid (unbiased) estimator of the common variance σ^2 if, and only if, the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ is true. If $H_0 : \mu_1 = \cdots = \mu_k$ is not true, then S_0^2 is positively biased as an estimator of the common variance σ^2, *i.e.*, if H_0 is not true, then S_0^2 tends to systematically overestimate σ^2. The computations for finding S_0^2 described above are illustrated for the potato leafhopper data in the Table 5.

Table 5. Potato leafhopper deviations from overall mean.

treatment	observation	overall mean	deviation from mean	squared deviation from mean
control	2.3	2.725	-.425	.180625
control	1.7	2.725	-1.025	1.050625
fructose	2.1	2.725	-.625	.390625
fructose	2.3	2.725	-.425	.180625
glucose	3.0	2.725	.275	.075625
glucose	2.8	2.725	.075	.005625
sucrose	3.6	2.725	.875	.765625
sucrose	4.0	2.725	1.275	1.625625
				sum of squared deviations = 4.275
				$S_0^2 = 4.275/7 = .6107$

We have defined two estimators S_p^2 and S_0^2 of the common variance σ^2. Both of these estimators are unbiased estimators of σ^2 when $H_0 : \mu_1 = \cdots = \mu_k$ is true. The estimator S_p^2 is unbiased as an estimator of σ^2 even when H_0 is not true; but, S_0^2 is positively biased as an estimator of σ^2 when H_0 is not true. Therefore, we can view an observed value of S_0^2 which is sufficiently large relative to the observed value of S_p^2 as evidence against the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$.

Before we discuss a method for determining whether the observed value of S_0^2 is large relative to S_p^2 we consider a decomposition of the deviation of an observation from the overall mean and a corresponding decomposition of the sum of squares about the overall mean.

The deviation of an observation Y_{ij} from the overall mean $\overline{Y}$ can be expressed as the sum of the deviation of the observation from its group mean $\overline{Y}_i$ and the deviation of its group mean from the overall mean, *i.e.*,

$$Y_{ij} - \overline{Y} = \left(Y_{ij} - \overline{Y}_i\right) + \left(\overline{Y}_i - \overline{Y}\right).$$

Furthermore, it can be shown that, there is a corresponding decomposition of the sum of squares about the overall mean as the sum of the sum of squares within the k groups plus the sum of squares among the k groups, *i.e.*,

$$\text{SS(about the overall mean)} = \text{SS(within the } k \text{ groups)} + \text{SS(among the } k \text{ groups)},$$

where

$$\text{SS(among the } k \text{ groups)} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{k} n_i\left(\overline{Y}_i - \overline{Y}\right)^2.$$

This decomposition is often summarized in a tabular form known as an analysis of variance table or ANOVA table as shown in Table 6.

Table 6. A basic ANOVA table.

source of variation	degrees of freedom	sum of squares
among groups	$k-1$	SS(among the k groups)
within groups	$n-k$	SS(within the k groups)
total	$n-1$	SS(about the overall mean)

Notice that the ANOVA table also indicates the corresponding decomposition of the total degrees of freedom, $n-1$, into the sum of the degrees of freedom among the k groups, $k-1$, and the degrees of freedom within the k groups, $n-k$. You can think of these degrees of freedom as indicating the "amount of information" contained in the corresponding sums of squares. If we use the degrees of freedom to normalize the sum of squares, by dividing the sum of squares by its degrees of freedom, the resulting "average" is known as a mean square, denoted by MS.

From the ANOVA decomposition of the sum of squares about the overall mean we can identify the sum of squares among the k groups, SS(among the k groups), as the term which causes S_0^2 to be positively biased as an estimator of σ^2. Therefore we can determine whether S_0^2 is large relative to S_p^2 by determining whether SS(among the k groups) is large

relative to SS(within the k groups). We will base this determination on the ratio of the mean squares corresponding to these sums of squares. The relevant ratio is the F–statistic

$$F_{calc} = \frac{\text{MS(among the } k \text{ groups)}}{\text{MS(within the } k \text{ groups)}} = \frac{\text{SS(among the } k \text{ groups)}/(k-1)}{\text{SS(within the } k \text{ groups)}/(n-k)}.$$

When the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ is true this F–statistic follows the F distribution with numerator degrees of freedom $k-1$ and denominator degrees of freedom $n-k$. Sufficiently large values of F_{calc} constitute evidence against $H_0 : \mu_1 = \cdots = \mu_k$. The P–value for this hypothesis test is the probability of observing a value of the F–statistic as large or larger than the calculated value F_{calc}, *i.e.*, the P–value is

$$P\text{–value} = P(F \geq F_{calc}),$$

where F denotes a variable which follows the F distribution with $k-1$ and $n-k$ degrees of freedom. (The F distributions are skewed to the right with density curves which are positive only for positive values of the variable.)

The ANOVA for the potato leafhopper example (including mean squares) is given in Table 7. In this example the calculated F–statistic is $F_{calc} = 1.325/.075 = 17.6667$ and the P–value (computed using the F distribution with 3 and 4 degrees of freedom) is $P(F \geq 17.6667) = .0090$. Since the P–value .0090 is very small, we conclude that there is very strong evidence that diet does have an effect on the survival time of potato leafhoppers in the sense that at least two of the four treatment mean survival times are different.

Table 7. Potato leafhopper ANOVA table.

source of variation	degrees of freedom	sum of squares	mean square
among groups	3	3.975	1.325
within groups	4	.300	.075
total	7	4.275	

12.2b F–tests for comparing nested models

The overall F–test, for $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, developed above is too general to be of much use by itself. This overall F–test only allows us to conclude that either the k group means are all equal or they are not all equal. In many situations, like the potato leafhopper example, there is enough subject matter information to formulate more specific

potential restrictions on the k group means. We will now discuss the use of F–tests for sequential comparisons (hypothesis tests) of a nested sequence of candidate models for k group means. We will develop this approach in the context of the potato leafhopper example.

Some natural groupings of the means μ_C, μ_F, μ_G and μ_S of the potato leafhopper example can be formed using the facts that fructose and glucose are 6–carbon sugars while sucrose is a 12–carbon sugar. Consider the following sequence of four nested models for the relationship among these means. These models are nested in the sense that each model in the sequence is a special case (restricted version) of the model that precedes it in the sequence. Thus model (2) is a special case (restricted version) of model (1); model (3) is a special case of model (2); and, model (4) is a special case of model (3).

model (1): The full model with four separate means, μ_C, μ_F, μ_G and μ_S.

model (2): The reduced model with three means, μ_C, μ_S, and the 6-carbon sugar mean μ_6, corresponding to the assumption that there is no difference between the effects of the two 6–carbon sugars in the sense that $\mu_F = \mu_G$.

model (3): The more reduced model with two means, μ_C, and the added sugar mean μ_A, corresponding to the assumption that there is no difference between the effects of the 6–carbon sugars and the 12–carbon sugar in the sense that $\mu_S = \mu_6$.

model (4): The most reduced model with one mean, μ, corresponding to the assumption that there is no difference between the effects of the added sugar diets and the control (no added sugar) diet in the sense that $\mu_C = \mu_A$.

Before we proceed with this example a brief discussion of the hypothesis testing approach to the comparison of a full model with a reduced model is in order. The reduced model is simpler than the full model in the sense that it specifies fewer means. Therefore, unless there is sufficient evidence to the contrary, we would prefer the simpler reduced model over the more complicated full model. This suggests a test of the null hypothesis

H_0 : The restrictions which define the reduced model are valid.
(The reduced model suffices and the full model is not needed.)

versus the alternative hypothesis

H_1 : The restrictions which define the reduced model are not valid.
(The reduced model does not suffice and the full model is needed.)

If we find sufficient evidence to reject the null hypothesis, then we conclude that the full model is needed and we abandon the reduced model. But, if we do not find sufficient

evidence to reject the null hypothesis, then we conclude that we do not need the full model and the simpler reduced model will suffice.

We will now outline the approach we will use for our analysis of the sequence of four models for the potato leafhopper example.

Step 1: We will first consider a hypothesis test for comparing the full model (1) with the reduced model (2). The full model (1) specifies that there are four means μ_C, μ_F, μ_G, and μ_S. Since the reduced model (2) is obtained from model (1) by imposing the restriction that $\mu_F = \mu_G$, our null hypothesis is

$$H_0 : \mu_F = \mu_G$$

and our alternative hypothesis is

$$H_1 : \mu_F \neq \mu_G.$$

Under H_0 there is a common population mean survival time, μ_6, for the two 6–carbon sugar diets and our model only requires the three means μ_C, μ_S, and μ_6. Under H_1 there are two 6–carbon sugar diet means μ_F and μ_G and our model specifies the four means μ_C, μ_F, μ_G, and μ_S.

If we find sufficient evidence to reject H_0, we will conclude that we cannot reduce the four treatment means to three treatment means as indicated, since $\mu_F \neq \mu_G$ and thus we need four treatment means in our model. If this happens we will stop.

If we are not able to reject H_0 we will conclude that there is no difference between the two 6–carbon sugar treatment means ($\mu_F = \mu_G$) and we only need three treatment means in our model μ_C, μ_S, and the 6-carbon sugar mean μ_6. If this happens we will continue by comparing model (2) (which now plays the role of the full model) with the reduced model (3).

Step 2: If our comparison of model (1) and model (2) (step 1) results in the conclusion that we do not need the four means of model (1), then we will consider a test for comparing the current full model (2) with the reduced model (3). Model (2) specifies that there are three means μ_C, μ_S, and μ_6. Since the reduced model (3) is obtained from model (2) by imposing the restriction that $\mu_S = \mu_6$, our null hypothesis is

$$H_0 : \mu_S = \mu_6$$

and our alternative hypothesis is

$$H_1 : \mu_S \neq \mu_6.$$

Under H_0 there is a common population mean survival time, μ_A, for the three added sugar diets and our model only requires the two means μ_C and μ_A. Under H_1 there are two

added sugar diet means, μ_S and μ_6, and our model specifies the three means μ_C, μ_S and μ_6.

If we find sufficient evidence to reject H_0, we will conclude that we cannot reduce the three treatment means to two treatment means as indicated, since $\mu_S \neq \mu_6$, and thus we need three treatment means in our model. If this happens we will stop.

If we are not able to reject H_0 we will conclude that there is no difference between the 6–carbon sugar treatment mean and the sucrose treatment mean ($\mu_S = \mu_6$) and we only need two treatment means in our model μ_C and the added sugar mean μ_A. If this happens we will continue by comparing model (3) (which now plays the role of the full model) with the reduced model (4).

Step 3: If our comparison of model (2) and model (3) (step 2) results in the conclusion that we do not need the three means of model (2), then we will consider a test for comparing the current full model (3) with the reduced model (4). Model (3) specifies that there are two means μ_C and μ_A. Since the reduced model (4) is obtained from model (3) by imposing the restriction that $\mu_C = \mu_A$, our null hypothesis is

$$H_0 : \mu_C = \mu_A$$

and our alternative hypothesis is

$$H_1 : \mu_C \neq \mu_A.$$

Under H_0 there is a common population mean survival time, μ, for all of the diets and our model only requires the one mean μ. Under H_1 there are two means, μ_C and μ_A.

If we find sufficient evidence to reject H_0 we will conclude that we cannot reduce the two treatment means to one treatment mean as indicated since, $\mu_C \neq \mu_A$, and thus we need two treatment means in our model. If this happens we will stop.

If we are not able to reject H_0 we will conclude that there is no difference between the control (no added sugar) treatment mean and the added sugar treatment mean ($\mu_C = \mu_A$) and we only need one treatment mean in our model. If this happens we will stop, since this is the end of our sequence of models.

Now that we have a plan of attack for our comparisons we need to know how to perform an F–test to compare a full model with a reduced model. Consider a full model with a group means and a reduced model with b group means ($b < a$) obtained by restrictions which result in a reduction of the a groups (means) of the full model into the b groups (means) of the reduced model. The sum of squares among the b groups in the reduced model SS(among the b groups) = SS(reduced model) is actually part of the sum of squares among the a groups in the full model SS(among the a groups) = SS(full model). The sum

of squares due to the full model after the reduced model SS(full model | reduced model) is defined as the difference,

$$\begin{aligned}\text{SS(full model | reduced model)} &= \text{SS(full model)} - \text{SS(reduced model)} \\ &= \text{SS(among the } a \text{ groups)} - \text{SS(among the } b \text{ groups)},\end{aligned}$$

between the two model sums of squares. The degrees of freedom for this sum of squares is the corresponding difference, df(full model) − df(reduced model) = $a - b$, between the two model degrees of freedom. Partitioning the sum of squares among the a groups of the full model into the sum of squares among the b groups of the reduced model and the sum of squares for the full model after the reduced model yields the ANOVA of Table 8.

Table 8. ANOVA table for model comparison.

source of variation	degrees of freedom	sum of squares
reduced model	$b-1$	SS(reduced model)
full model after reduced model	$a-b$	SS(full model | reduced model)
within the a groups of the full model	$n-a$	SS(within the a groups)
total	$n-1$	SS(about the overall mean)

The F–test for comparing these models can be viewed as a test of

$$H_0: \text{ the reduced model with } b \text{ group means will suffice}$$

versus

$$H_1: \text{ the full model with } a \text{ group means is needed.}$$

More formally, the null hypothesis specifies that the restrictions which reduce the a means of the full model to the b means of the reduced model are valid. The F–statistic for this comparison is

$$F_{calc} = \frac{\text{MS(full model | reduced model)}}{\text{MS(within the } a \text{ groups of the full model)}}.$$

If the P–value $P(F \geq F_{calc})$, where F denotes an F variable with $a - b$ and $n - a$ degrees of freedom, is small enough, we reject H_0 and conclude that the reduced model with b group means is not appropriate and we need the full model with a group means. If the P–value is not small enough, we fail to reject H_0 and conclude that the reduced model with b group means is appropriate and we do not need the full model with a group means.

We will now use this method to evaluate the sequence of four models proposed above for the potato leafhopper example. The ANOVA's for models (1) and (2) are given in Tables 9 and 10.

Table 9. ANOVA table for model (1).

source of variation	degrees of freedom	sum of squares	mean square
among the 4 groups	3	3.975	1.325
within the 4 groups	4	.300	.075
total	7	4.275	

Table 10. ANOVA table for model (2).

source of variation	degrees of freedom	sum of squares	mean square
among the 3 groups	2	3.485	1.7425
within the 3 groups	5	.790	.1580
total	7	4.275	

The ANOVA for comparing model (1) and model (2) provided in Table 11 can be constructed from the information in the preceding ANOVA tables. The only computation required is to subtract the reduced model among groups sum of squares from the full model among groups sum of squares to get SS(full model | reduced model) $= 3.975 - 3.485 = .49$, with $3 - 2 = 1$ degrees of freedom.

Table 11. ANOVA table for comparing model (1) and model (2).

source of variation	degrees of freedom	sum of squares	mean square
reduced model	2	3.485	1.7425
full model after reduced model	1	.490	.4900
within the 4 groups	4	.300	.0750
total	7	4.275	

The calculated F–statistic for comparing model (1) and model (2) is $F_{calc} = .49/.075 = 6.5333$ with a P–value of .0629. (This P–value is computed using the F distribution with 1 and 4 degrees of freedom.) This P–value is not small enough to allow us to reject the hypothesis that $\mu_F = \mu_G$ so we conclude that the three means (μ_C, μ_S, μ_6) of the reduced model (2) will suffice and we do not need the four means of the full model (1). We now proceed to compare the current full model (2) to the reduced model (3). The ANOVA for model (3) is given in Table 12.

Table 12. ANOVA table for model (3).

source of variation	degrees of freedom	sum of squares	mean square
among the 2 groups	1	1.4017	1.4017
within the 2 groups	6	2.8733	.4789
total	7	4.275	

We can produce the ANOVA for comparing model (2) and model (3) of Table 13 as before. In this case we find that SS(full model | reduced model) $= 3.485 - 1.4017 = 2.0833$, with $2 - 1 = 1$ degrees of freedom.

Table 13. ANOVA table for comparing model (2) and model (3).

source of variation	degrees of freedom	sum of squares	mean square
reduced model	1	1.4017	1.4017
full model after reduced model	1	2.0833	2.0833
within the 3 groups	5	.7900	.1580
total	7	4.275	

The calculated F–statistic for comparing model (2) and model (3) is $F_{calc} = 2.0833/.158 = 13.1854$ with a P–value of .0150. (This P–value is computed using the F distribution with 1 and 5 degrees of freedom.) This P–value is small enough to allow us to reject the hypothesis that $\mu_S = \mu_6$ so we conclude that the three means (μ_C, μ_S, μ_6) of model (2) are needed in the sense that the reduced model (3) with two means (μ_C, μ_A) does not suffice. We will stop at this point and base any further inferences about these diets on the three means of model (2).

Remark: At each stage of our sequential comparison of models for the potato leafhopper example we arrived at the reduced model by combining two groups from the full model which caused the degrees of freedom for the full model after the reduced model to be one in each comparison. It is possible to compare models for which the degrees of freedom for the full model after the reduced model is larger than one. We can demonstrate this by supposing that it had not occurred to us to consider combining the two 6–carbon sugar groups. That is, suppose that our initial comparison had been between model (1), with four separate means, and model (3), with two means (μ_C and μ_A), one for the no added sugar control diet group and one for the added sugar diet group. In this case the reduced model is obtained from the full model by combining the three added sugar groups to get

a single added sugar group and the corresponding null hypothesis is $H_0 : \mu_F = \mu_G = \mu_S$. Thus, in this case the full model has 4 means (3 degrees of freedom), the reduced model has 2 means (1 degree of freedom), and the sum of squares for the full model after the reduced model has $3 - 1 = 2$ degrees of freedom. For this comparison we would have SS(full model | reduced model) $= 3.975 - 1.4017 = 2.5733$, with $3 - 1 = 2$ degrees of freedom, SS(within the 4 groups of the full model) = .3 with 4 degrees of freedom, and a calculated F–statistic of $F_{calc} = (2.5733/2)/(.3/4) = 17.1553$. If we were to perform this test, the P–value would be computed using the F distribution with 2 and 4 degrees of freedom.

12.2c Confidence intervals for linear combinations of means

A linear combination of the k population means $\mu_1, \ldots, \mu_k$ is a quantity of the form

$$\lambda = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k,$$

where the coefficients, $c_1, \ldots, c_k$, are suitably chosen constants. For example, if we take all of the coefficients in this linear combination to be $1/k$, we obtain the average of the means $(\mu_1 + \mu_2 + \cdots + \mu_k)/k$. If we take one coefficient to be 1, a second to be -1, and the others to be 0, we obtain a difference of two means, *e.g.*, taking $c_1 = -1$, $c_2 = 1$ and the other $c_i = 0$, yields $\mu_2 - \mu_1$.

The obvious estimate of the linear combination $\lambda = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k$ is the corresponding linear combination of the sample means

$$\hat{\lambda} = c_1\overline{Y}_1 + c_2\overline{Y}_2 + \cdots + c_k\overline{Y}_k.$$

In the present context of k independent random samples of sizes $n_1, \ldots, n_k$ with a common population variance σ^2, the population standard error of this estimated linear combination is

$$\text{S.E.}(\hat{\lambda}) = \sqrt{\sigma^2\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots \frac{c_k^2}{n_k}\right)}$$

which can be estimated, using the pooled estimator $S_p^2 = \text{MS(within)}$ of the common variance, by the sample standard error

$$\widehat{\text{S.E.}}(\hat{\lambda}) = \sqrt{S_p^2\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots \frac{c_k^2}{n_k}\right)}.$$

A set of confidence intervals is said to form a set of simultaneous 95% confidence intervals if the procedure which yields the set of confidence intervals is such that 95% of the time all of the intervals will contain the corresponding parameters. We can use the

Scheffé method to form simultaneous 95% confidence intervals for linear combinations of k population means. The basic idea of this method is to use a margin of error multiplier which is large enough to insure that the collection of confidence intervals it produces for all possible linear combinations of the k means form a set of simultaneous 95% confidence intervals. The margin of error multiplier for Scheffé's method when there are k means in the model is $\sqrt{(k-1)F_{(k-1,n-k)}(.95)}$, where $F_{(k-1,n-k)}(.95)$ is the 95th percentile of the F distribution with $k-1$ and $n-k$ degrees of freedom. Thus the 95% Scheffé margin of error for $\hat{\lambda} = c_1\overline{Y}_1 + c_2\overline{Y}_2 + \cdots + c_k\overline{Y}_k$ is

$$\text{M.E.}(\hat{\lambda}) = \sqrt{(k-1)[F_{(k-1,n-k)}(.95)]S_p^2\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots \frac{c_k^2}{n_k}\right)}.$$

We now return to our analysis of the potato leafhopper example for which we have concluded that model (2) with the three means μ_C, μ_S, and μ_6 is the appropriate model. We now need to make some sort of inference about the relationship among these three population mean survival times. We will use selected linear combinations and confidence intervals to explore the relationship among these three population mean survival times.

The sample means for the three diet groups are $\overline{Y}_C = 2$ (based on the $n_C = 2$ control diet observations), $\overline{Y}_S = 3.8$ (based on the $n_S = 2$ sucrose diet observations), and $\overline{Y}_6 = 2.55$ (based on the $n_6 = 4$ 6–carbon sugar diet observations). The pooled estimate of the population variance is $S_p^2 = \text{MS(within)} = .158$ with 5 degrees of freedom. For this model we have $k = 3$ means and $n = 8$ observations; therefore, the Scheffé margin of error multiplier is $\sqrt{2(5.7861)} = 3.4018$ (since the 95th percentile of the F distribution with 2 and 5 degrees of freedom is 5.7861).

We will begin our comparisons among the three means by estimating the three pairwise differences, $\mu_S - \mu_C$, $\mu_6 - \mu_C$, and $\mu_S - \mu_6$. First note that given two sample means $\overline{Y}_1$ and $\overline{Y}_2$ based on n_1 and n_2 observations the estimated standard error of $\overline{Y}_1 - \overline{Y}_2$ is

$$\widehat{\text{SE}}(\overline{Y}_1 - \overline{Y}_2) = \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

The estimates of the three pairwise differences and the corresponding standard errors and simultaneous 95% margins of error are given in the Table 14.

Table 14. Estimates of the pairwise differences.

difference	**estimate**	**standard error**	**margin of error**
$\mu_S - \mu_C$	$\overline{Y}_S - \overline{Y}_C = 1.8$	$\sqrt{.158\left(\frac{1}{2} + \frac{1}{2}\right)} = .3975$	$3.4018(.3975) = 1.3522$
$\mu_6 - \mu_C$	$\overline{Y}_6 - \overline{Y}_C = .55$	$\sqrt{.158\left(\frac{1}{2} + \frac{1}{4}\right)} = .3442$	$3.4018(.3442) = 1.1709$
$\mu_S - \mu_6$	$\overline{Y}_S - \overline{Y}_6 = 1.25$	$\sqrt{.158\left(\frac{1}{2} + \frac{1}{4}\right)} = .3442$	$3.4018(.3442) = 1.1709$

Adding and subtracting these margins of error from the corresponding estimates to get confidence intervals we conclude that we are 95% confident that $.4478 \leq \mu_S - \mu_C \leq 3.1522$, $-.6209 \leq \mu_6 - \mu_C \leq 1.7209$, and $.0791 \leq \mu_S - \mu_6 \leq 2.4209$. These confidence intervals suggest that μ_6 and μ_C are not different and that μ_S is larger than both of the other means. Thus, a confidence interval for $\mu_S - (\mu_C + \mu_6)/2$ would be useful for indicating how much larger μ_S is than the average of the other two means. Since this expression is a linear combination of the three means we can add a confidence interval for this combination to our set of confidence intervals and still have simultaneous confidence of 95%. Our estimate of $\mu_S - (\mu_C + \mu_6)/2$ is $\overline{Y}_S - (\overline{Y}_C + \overline{Y}_6)/2 = 1.525$ with standard error

$$\widehat{\text{SE}}\left(\overline{Y}_S - \frac{\overline{Y}_C + \overline{Y}_6}{2}\right) = \sqrt{S_p^2\left[\frac{1}{n_S} + \frac{1}{4n_C} + \frac{1}{4n_6}\right]}$$
$$= \sqrt{.158\left[\frac{1}{2} + \frac{1}{8} + \frac{1}{16}\right]} = .3296$$

and margin of error $3.4018(.3296) = 1.1212$. Thus we are 95% confident that

$$.4038 \leq \mu_S - (\mu_C + \mu_6)/2 \leq 2.6462 \text{ and } -.6209 \leq \mu_6 - \mu_C \leq 1.7209.$$

Based on these confidence intervals we can conclude that there is no difference between the effects of adding a 6–carbon sugar to the diet or using the standard diet with no added sugar in the sense that the data are consistent with the claim that $\mu_C = \mu_6$. On the other hand, we find that adding the 12–carbon sugar sucrose to the potato leafhopper diet increases the mean survival time by something between .4038 and 2.6462 days over the average of the mean survival times corresponding to a diet with no added sugar or with an added 6–carbon sugar, *i.e.*, we can conclude with 95% confidence that μ_S exceeds the average $(\mu_C + \mu_6)/2$ by at least .4038 days and as much as 2.6462 days.

We will now revisit the fruitfly fecundity example of Chapter 8 and consider an analysis for this example using the methods of the present chapter.

Example. Fecundity of fruitflies (revisited). Sokal, R.R. and Rohlf, F.J. (1969) *Biometry*, W.H. Freeman, p.232, discuss a study conducted to compare the fecundity of three genetic lines of *Drosophila melanogaster*. The data, provided in Table 5 of Chapter 8, consist of per diem fecundities (number of eggs laid per female per day for the first 14 days of life) for 25 females of three lines of *Drosophila melanogaster*. Two of these genetic lines were selected for resistance (RS) and susceptibility (SS) to DDT, the third line is a nonselected control (NS). Recall that the investigator wanted to know if there was any evidence that the population mean fecundities for the two selected lines (μ_{RS} and μ_{SS})

were different. The investigator also wanted to know how the population mean fecundity μ_{NS} for the nonselected line related to the mean fecundities of the selected lines.

When we first considered this example, we found that the data was reasonably modeled as consisting of three independent random samples, each of size 25, from normal distributions with respective population mean fecundities μ_{RS}, μ_{SS}, and μ_{NS} and with common population variance σ^2. We can address the investigator's question about the relationship between the mean fecundities of the selected lines using the following sequence of two nested models.

model (1): The full model with three separate means, μ_{RS} for the resistant line, μ_{SS} for the susceptible line, and μ_{NS} for the nonselected line.

model (2): The reduced model with two means, μ_{NS} for the nonselected line and μ_S for the selected lines corresponding to the assumption that there is no difference between the mean fecundities for the two selected lines in the sense that $\mu_{RS} = \mu_{SS}$.

The ANOVA's for models (1) and (2) are given in Tables 15 and 16 and the ANOVA for comparing these models is given in Table 17.

Table 15. ANOVA table for the full model with 3 lines.

source of variation	degrees of freedom	sum of squares	mean square
among the 3 lines	2	1362.2115	681.1057
within the 3 lines	72	5659.0224	78.5975
total	74	7021.2339	

Table 16. ANOVA table for reduced model with 2 lines.

source of variation	degrees of freedom	sum of squares	mean square
among the 2 lines	1	1329.0817	1329.0817
within the 2 lines	73	5692.1522	77.9747
total	74	7021.2339	

Table 17. ANOVA table for comparing the models.

source of variation	degrees of freedom	sum of squares	mean square
2 line model	1	1329.0817	1329.0817
3 line model after 2 line model	1	33.1298	33.1298
within the 3 lines	72	5659.0224	78.5975
total	74	7021.2339	

The calculated F–statistic for comparing model (1) and model (2) is $F_{calc} = 33.1298/78.5975 = .42$ with a P–value of .5182. (This P–value is computed using the F

distribution with 1 and 72 degrees of freedom.) This P–value is quite large indicating that there is no evidence that μ_{RS} is different from μ_{SS}. We can conclude that the three means (μ_{RS}, μ_{SS}, and μ_{NS}) of the full model (1) are not needed and we are justified in adopting the simplified model (2) with mean fecundity μ_{NS} for the nonselected line and mean fecundity μ_S for the selected lines. The remainder of our analysis will be in terms of this reduced model.

Before we proceed with our analysis of this example it is instructive to compare the ANOVA F–test we just used to test the null hypothesis $H_0 : \mu_{RS} = \mu_{SS}$ versus the alternative hypothesis $H_1 : \mu_{RS} \neq \mu_{SS}$ with the t–test we used in Chapter 8 for this same hypothesis test. The ANOVA F–test is equivalent to a t–test based on the difference $\overline{Y}_{RS} - \overline{Y}_{SS} = 1.628$ and the pooled sample variance MS(within) = 78.5975 with 72 degrees of freedom. This pooled sample variance has 72 degrees of freedom, since it is computed using all three of the samples. The t–test we considered in Chapter 8 was based on the difference $\overline{Y}_{RS} - \overline{Y}_{SS} = 1.628$ and the pooled sample variance S_p^2 with 48 degrees of freedom based on the two samples from the selected lines. Thus, these two t–tests differ because they use different estimated standard errors due to the way in which the population variance is estimated. If the assumption of a common variance for all three lines is reasonable, then the ANOVA F–test is better than the t–test of Chapter 8, since it is based on a better (higher degrees of freedom) estimate of the population variance.

Since there are only two means in the reduced model we can use the overall F–test to compare these means. The calculated F–statistic for testing the null hypothesis $H_0 : \mu_{NS} = \mu_S$ is $F_{calc} = 1329.0817/77.9747 = 17.05$ with a P–value that is less than .0001. (This P–value is computed using the F distribution with 1 and 73 degrees of freedom.) This very small P–value indicates that there is very strong evidence that the mean fecundity for the nonselected line μ_{NS} is not the same as the mean fecundity μ_S for the selected lines. This F–test for comparing these two means is equivalent to the t–test we performed in Chapter 8 in these sense that these two tests give the same P–value. In fact, for the present circumstance of comparing two means (using a model with only two means) the square of the Student's t–statistic is equal to the F–statistic. We can form a 95% confidence interval for the difference $\mu_{NS} - \mu_S$ between these mean fecundities to determine which mean is larger and to get an estimate of the size of this difference. In this example, we are 95% confident that $\mu_{NS} - \mu_S$ is between 4.6192 and 13.241. That is, we are 95% confident that the population mean fecundity (mean number of eggs laid per day for the first 14 days of life) μ_{NS} for the nonselected line exceeds the population mean fecundity μ_S for the selected lines by at least 4.6192 eggs per day and perhaps as much as 13.241 eggs per day.

In conclusion, we have found that the distributions of fruitfly fecundity for two selected populations are identical (since we assumed a common variance and since we failed to reject

$\mu_{RS} = \mu_{SS}$); but, the distribution of fruitfly fecundity for the nonselected population differs from the distribution for the selected population by having a larger (by 4.6192 to 13.241 eggs per day) population mean fecundity.

Before we leave this example we will consider one more approach to its analysis. Suppose that we did not have enough *a priori* information to allow use to confidently propose a reasonable sequence of nested models for our analysis. In this situation we could perform an exploratory analysis by using the Scheffé method to form simultaneous 95% confidence intervals for interesting linear combinations of the three population mean fecundities.

We begin our analysis by considering the three pairwise differences between the population mean fecundities. The estimates of the three pairwise differences and the simultaneous 95% confidence intervals are given in the Table 18.

Table 18. Estimates of the pairwise differences.

difference	**estimate**	**confidence interval**
$\mu_{NS} - \mu_{RS}$	8.116	(1.848, 14.384)
$\mu_{NS} - \mu_{SS}$	9.744	(3.476, 16.012)
$\mu_{RS} - \mu_{SS}$	1.628	(-4.640, 7.896)

Based on these simultaneous confidence intervals we can conclude that the population mean fecundities μ_{RS} and μ_{SS} for the selected lines are not different and we can conclude that the population mean fecundity for the nonselected line μ_{NS} is larger than each of the other population mean fecundities. Since we have concluded that the selected line means are not different it would be of interest to also consider a contrast between the nonselected line population mean μ_{NS} and the average $(\mu_{RS} + \mu_{SS})/2$. The estimate of the contrast $\mu_{NS} - (\mu_{RS} + \mu_{SS})/2$ is 8.93 and the Scheffé method gives the confidence interval $(3.5020, 14.3581)$. Thus we can conclude, with 95% confidence that $-4.640 \leq \mu_{RS} - \mu_{SS} \leq 7.896$ and $3.5020 \leq \mu_{NS} - (\mu_{RS} + \mu_{SS})/2 \leq 14.3581$. This allows us to conclude that $\mu_{RS} = \mu_{SS}$ and μ_{NS} exceeds $(\mu_{RS} + \mu_{SS})/2$ by at least 3.5020 and as much as 14.3581 eggs per day.

Chapter 4a
Probability Models

4a.1 Introduction

Chapters 2 and 3 are concerned with data description (descriptive statistics) where a sample of values of the variable X is obtained and the distribution of the observed values of X within the sample is examined. Chapter 4 is concerned with methods of obtaining a sample, by sampling or experimentation, which will allow us to use the information in the sample to make inferences about a population. The majority of the remainder of this book is devoted to methods for using a sample to make inferences about a population (inferential statistics). The basic idea of inferential statistics is to use the distribution of X in a suitably chosen sample to make inferences about the distribution of X in the population (the population distribution of X). In other words, the goal of inferential statistics is to obtain a sample of values of X and use these sample values to make inferences about the process which generates them by making inferences about the (theoretical) probabilities which these values must obey.

To make inferences about the population distribution of X we need to formulate a suitable probability model for the distribution of X. This probability model may be only partially specified with one or more unknown parameters, in which case we will need to estimate this parameter or these parameters in order to make inferences about the population distribution. This chapter provides a general discussion of probability models and presents some specific probability models.

4a.2 Probability models for a variable with a finite number of values

Let $x_1, x_2, \ldots, x_k$ denote the k distinct possible values of the variable X. A probability model for the distribution of X is an assignment of k probabilities $p_1, p_2, \ldots, p_k$ to the k possible values of X. For $i = 1, 2, \ldots, k$, p_i denotes the probability that X will take on the value x_i, in symbols we write $P(X = x_i) = p_i$. We can think of these probabilities as the theoretical relative frequencies with which X will take on these values, according to this probability model. Notice that any collection of k probabilities, each of which is between 0 and 1, which sum to 1 defines a potential probability model for the distribution of X.

We can think of a probability model for the distribution of X in terms of a box model. Imagine a box of balls where each ball is labeled with one of the k possible values of X and where the proportion of balls in the box labeled x_i is p_i, for $i = 1, 2, \ldots, k$. According to the probability model, measuring or observing the value of X is equivalent to selecting a ball at random from this box and observing the label on the ball. If we are formulating a model

for sampling from a physical population of units, then a ball represents a population unit and the balls in the box constitute the population. If X represents the outcome of some process of measurement or experimentation, then a ball represents a particular outcome and the box of balls represents the population of possible outcomes or values for X.

Tabular and graphical representations of probability models or population distributions are analogous to relative frequency distributions and bar graphs or histograms. That is, a probability distribution for a variable with k possible values is a table listing the possible values and the associated probabilities. If X is qualitative, so that the possible values $x_1, \ldots, x_k$ are simply names for k possible categories, then we can use a bar graph with k bars to give a graphical representation of the probability distribution. If X is quantitative, so that the possible values $x_1, \ldots, x_k$ are meaningful numerical values, then we can use a probability histogram to represent the distribution graphically. Notice that a graphical representation of a probability distribution uses area to represent probability.

The probability model (probability distribution) for a dichotomous (two–valued) variable is the Bernoulli model with success probability p. It is conventional to refer to one of the two possible values (outcomes) as a "success" and the other as a "failure." These generic labels are not meant to imply that observing a success is good. Rather, we can think of choosing one of the two possible outcomes and asking, "Did we observe the chosen outcome?", with the two possibilities being yes (a success) and no (a failure). The Bernoulli model with success probability p has two probabilities $P(\text{success}) = p$ and $P(\text{failure}) = q = 1 - p$, where p is between zero and one.

The two examples below provide simple examples of Bernoulli distributions with success probability $p = 2/3$. The first indicates how this distribution applies to selecting a unit at random from a physical population. The second indicates the application to observing the outcome of an experimental trial.

Example. A box containing balls of two types. Consider a box containing balls of which 2/3 are red and 1/3 are green. Define observing a red ball as a success. If we mix the balls thoroughly and then select one ball at random, the probability that we will obtain a red ball is 2/3 and the probability that we will obtain a green ball is 1/3. The corresponding probability distribution specifies the probability of observing a red ball as $P(\text{red}) = 2/3$ and the probability of observing a green ball as $P(\text{green}) = 1/3$. Thus, with success corresponding to red, the color of the ball selected follows the Bernoulli distribution with success probability $p = 2/3$.

Example. Tossing a fair die once. Suppose that a fair (balanced) die is tossed once and the number of dots on the upturned face is observed. Define a success to be the occurrence of a 1, 2, 3, or 4. Since the die is fair, the probability of a success on a single trial is $p = 4/6 = 2/3$. Therefore, with success defined as above, tossing the fair die once yields a Bernoulli variable with success probability $p = 2/3$.

The next example provides an instance where theoretical considerations in the form of a simple Mendelian inheritance model lead to a Bernoulli distribution.

Example. Inheritance in peas (flower color). In his investigations, during the years 1856 to 1868, of the chromosomal theory of inheritance Gregor Mendel performed a series of experiments on ordinary garden peas. One characteristic of garden peas that Mendel studied was the color of the flowers (red or white). When Mendel crossed a plant with red flowers with a plant with white flowers, the resulting offspring all had red flowers. But when he crossed two of these first generation plants, he observed plants with white as well as red flowers.

The gene which determines the color of the flower occurs in two forms (alleles). Let R denote the allele for red flowers (which is dominant) and r denote the allele for white flowers (which is recessive). When two plants are crossed the offspring receives one allele from each parent, thus there are four possible genotypes (combinations) RR, Rr, rR, and rr. The three genotypes RR, Rr, and rR, which include the dominant R allele, will yield red flowers while the fourth genotype rr will yield white flowers. If a red flowered RR genotype parent is crossed with a white flowered rr genotype parent, then all of the offspring will have genotype Rr and will produce red flowers. The basic Mendelian inheritance model assumes that a pair of alleles is formed by randomly choosing one allele from each parent. Under this model, if two of these first generation Rr genotype plants are crossed, each of the four possible genotypes RR, Rr, rR, and rr is equally likely and plants with white as well as red flowers will occur. Under this simple model, with each of the four genotypes having the same probability of occurring, the probability that a plant will have red flowers is $P(\text{red}) = 3/4$ and the probability that a plant will have white flowers is $P(\text{white}) = 1/4$. This Bernoulli distribution for flower color is summarized in Table 1.

Table 1. Pea plant flower color distribution.

flower color	probability
red	3/4
white	1/4

Example. Inheritance in peas (seed shape and color). We will now consider the Mendelian inheritance model for two independently inherited characteristics. In particular we will consider the characteristics seed shape, with possible shapes of round (R, dominant) and wrinkled (r, recessive), and seed color, with possible colors of yellow (Y, dominant) and green (y, recessive). If an $RRYY$ genotype plant with round yellow seeds is crossed with an $rryy$ genotype plant with wrinkled green seeds, the offspring will all have round yellow seeds and genotype $RrYy$. If two of the resulting $RrYy$ genotype plants with round yellow seeds are crossed, there are 16 equally likely possible genotypes. The nine genotypes $RRYY, RRYy, RRyY, RrYY, RrYy, RryY, rRYY, rRYy, rRyY$ yield

round yellow seeds; the three genotypes $rrYY, rrYy, rryY$ yield wrinkled yellow seeds; the three genotypes $RRyy, Rryy, rRyy$ yield round green seeds; and, the single genotype $rryy$ yields wrinkled green seeds. The fact that these 16 possible genotypes are equally likely yields the probability distribution summarized in Table 2.

Table 2. Pea plant seed shape/color distribution.

shape/color	probability
round yellow	9/16
wrinkled yellow	3/16
round green	3/16
wrinkled green	1/16

4a.3 Probability models for discrete quantitative variables

The remainder of our discussion of probability models is restricted to probability models for quantitative variables. Let us begin by considering certain parameters associated with the probability model for a quantitative variable X with possible numerical values $x_1, x_2, \ldots, x_k$ and corresponding probabilities $p_1, p_2, \ldots, p_k$. In this case the possible values are meaningful numerical values, such as counts or measurements of a physical quantity, and can be viewed as points on the number line. Thus we can discuss the shape of a distribution, the locations of representative values such as the population median or quartiles, and quantities which describe the location of and variability in the distribution.

The population mean $\mu = \mu_X$ of X (population mean of the distribution of X) is the location of the balancing point of the probability histogram or, algebraically, the weighted average of the possible values of X with weights given by the corresponding probabilities. More formally, we have

$$\mu_X = x_1 p_1 + \cdots + x_k p_k = \sum_{i=1}^{k} x_i p_i.$$

The population mean is the long run average value of X in the sense that if we observed values of X repeatedly and if these observations occurred with the specified probabilities, then in the long run the average of these values would be equal to μ_X, since x_1 would occur $p_1 \times 100\%$ of the time, x_2 would occur $p_2 \times 100\%$ of the time, and so on. For this reason the population mean of X is often called the expected value of X and denoted $\mu_X = E(X)$ with the understanding that expected value is short for long run average expected value.

The population variance $\sigma^2 = \sigma_X^2$ of X (population variance of the distribution of X) is defined as the weighted average of the squared deviations of the possible values of X

from the mean μ_X with weights given by the corresponding probabilities. More formally, we have

$$\sigma_X^2 = (x_1 - \mu_X)^2 p_1 + \cdots + (x_k - \mu_X)^2 p_k = \sum_{i=1}^{k} (x_i - \mu_X)^2 p_i.$$

The population standard deviation $\sigma = \sigma_X$ of X is the (positive) square root of the population variance.

We can also define population quantiles or percentiles, such as the population median and the population quartiles, as the points on the number line where the probabilities (areas) in the probability histogram to the left and right of the point are equal to the appropriate values.

Example. Tossing a fair die. Suppose we toss a fair (balanced) die once and let X denote the number on the upturned face with possible values of $1, \ldots, 6$. Since the die is balanced each of the six possible values of X is equally likely to appear when we toss the die and the appropriate probability model is the uniform distribution on the set $\{1, \ldots, 6\}$ which assigns probability $1/6$ to each possible outcome. The corresponding population mean and variance are $\mu = 7/2$ and $\sigma^2 = 35/12$; the population median is also $7/2$, since this distribution is symmetric.

In general, the discrete uniform distribution on the integers $1, \ldots, k$ assigns probability $1/k$ to each of the k integers in this set of possible values. The population mean (and median) of this uniform distribution is $\mu = (k+1)/2$ and the population variance is $\sigma^2 = (k-1)(k+1)/12$.

Example. Tossing a fair die, revisited. Now suppose we toss a fair (balanced) die twice and let X denote the sum of the two numbers we obtain. Taking order into account, there are 36 possible outcomes (36 combinations of the two numbers obtained) and, since the die is fair, each of these outcomes is equally likely. Since these 36 combinations are mutually exclusive we can add the probabilities corresponding to each distinct value of X yielding the distribution given in Table 3 and Figure 1. This distribution is symmetric so the population mean (and median) of the sum is $\mu = 7$. It can be shown that the population variance of the sum is $\sigma^2 = 35/6$.

Figure 1. Probability histogram for the sum when a fair die is tossed twice.

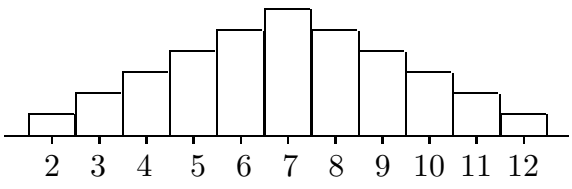


Table 3. Probability distribution for the sum when a fair die is tossed twice.

sum	probability
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

4a.4 Probability models for counts

We will now consider three probability models for counts. The first two models, the binomial and the hypergeometric, can be motivated by considering the distribution of the number of successes in a sample, selected with or without replacement, from a dichotomous population. A dichotomous population is a population of units which can be partitioned into two distinct groups; the population success group and the population failure group.

Suppose that a simple random sample of size n is selected with replacement from a dichotomous population with population success proportion p. We can view the results of this sampling process as forming a sequence of n trials, where a trial is the selection of a unit from the population. These trials possess two important properties.

1. For each trial the probability of success is p.
2. The outcomes of the trials are independent.

A sequence of trials with these properties is known as a sequence of independent Bernoulli trials with success probability p.

The probability of observing a specified sequence of successes and failures as the outcome of a sequence of independent Bernoulli trials is equal to the product of the probabilities of success (S) and failure (F) (as appropriate) on each trial. For example, with three Bernoulli trials the probability of observing the sequence SFS, denoted $P(SFS)$, is equal to the product of the probabilities of the three outcomes, *i.e.*, $P(SFS) = P(S)P(F)P(S)$. In the present context on each trial $P(S) = p$ and $P(F) = 1 - p$ and we have $P(SFS) = p(1-p)p = p^2(1-p)$.

To determine the appropriate probability model for the number of successes X in a sequence of n Bernoulli trials we need to determine the number of sequences of $S's$ and $F's$ which give each possible value of X. First consider the case with $n = 3$ trials. The

eight possible sequences of $S's$ and $F's$ and the associated probabilities and values of X are given in Table 4.

Since the eight sequences in Table 4 are mutually exclusive, we can add the probabilities corresponding to each distinct value of X to get the probability distribution in Table 5. This distribution of the number of successes in three Bernoulli trials is the binomial distribution with parameters $n = 3$ and p.

Table 4. Outcomes of a simple random sample of size 3 selected with replacement.

sequence	probability	X
FFF	$(1-p)^3$	0
SFF	$p(1-p)^2$	1
FSF	$p(1-p)^2$	1
FFS	$p(1-p)^2$	1
SSF	$p^2(1-p)$	2
SFS	$p^2(1-p)$	2
FSS	$p^2(1-p)$	2
SSS	p^3	3

Table 5. Probability distribution for the number of successes in 3 Bernoulli trials. (Binomial distribution)

sum X	probability P(X)
0	$(1-p)^3$
1	$3p(1-p)^2$
2	$3p^2(1-p)$
3	p^3

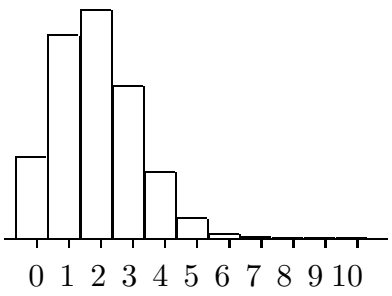
Notice that the probabilities in Table 5 are of the form $cp^x(1-p)^{n-x}$, where x is the number of successes, $n-x$ is the number of failures, and c is the number of ways to choose locations for the x $S's$ in the sequence of n $S's$ and $F's$. The number c, usually denoted by $\binom{n}{x}$ or C_x^n, is the binomial coefficient giving the number of combinations of n things taken x at a time. In general, for $x = 0, 1, \ldots, n$, the probability of observing $X = x$ successes in a sequence of n Bernoulli trials with success probability p is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

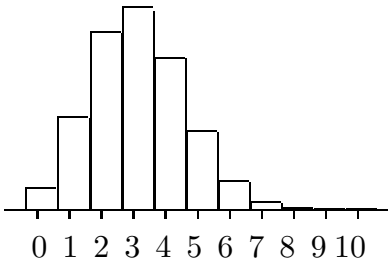
This probability function determines the binomial distribution with parameters n and p. Most statistical software programs and some calculators will compute these probabilities

or the cumulative probabilities $P(X \leq x)$. It can be shown that the mean of this binomial distribution is $\mu = np$ and the variance is $\sigma^2 = npq = np(1-p)$.

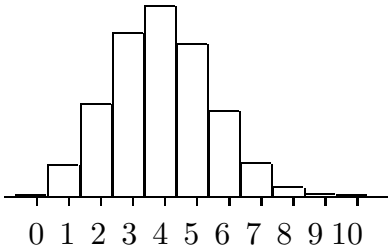
Figure 2. Probability histograms for the number of successes.



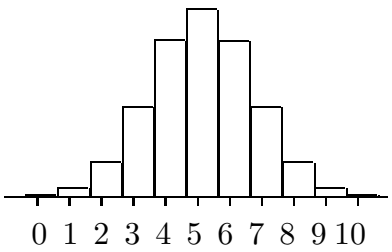
binomial $(n = 10, p = .2, \mu = 2, \sigma^2 = 1.6)$



binomial $(n = 10, p = .3, \mu = 3, \sigma^2 = 2.1)$



binomial $(n = 10, p = .4, \mu = 4, \sigma^2 = 2.4)$



binomial $(n = 10, p = .5, \mu = 5, \sigma^2 = 2.5)$

Several binomial distributions are presented in Figure 2. Notice that the binomial distribution is skewed right when p is small (near zero) and that, for $p < .5$, there is less skewness as p increases with the distribution becoming symmetric when $p = .5$. If we examined binomial distributions with $p > .5$ we would observe the analogous pattern with skewness to the left and with less skewness for p near .5.

Now suppose that a simple random sample of size n is selected without replacement from a dichotomous population with population success proportion p. As before we can view the results of this sampling process as forming a sequence of n trials, where a trial is the selection of a unit from the population. However, it is clear that neither of the two properties of independent Bernoulli trials is satisfied when we sample without replacement.

To motivate the appropriate probability model for the number of successes X in a sample chosen without replacement first consider the case with $n = 3$ trials. In this situation the distribution of X depends on the number of population units which are classified as successes M and the total number of units in the population N. Suppose that there are $N = 20$ population units of which $M = 5$ are successes. As before the probability of observing a specified sequence of $S's$ and $F's$ is equal to the product of the probabilities of the three outcomes; however, the probabilities of success and failure on a particular trial depend on what has happened on the previous trials. For example, for the sequence SFS

we write $P(SFS) = P(S)P(F|S)P(S|SF)$, where $P(F|S)$ is the conditional probability of a failure given that we have observed a success and $P(S|SF)$ is the conditional probability of a success given that we have observed a success and a failure. With $M = 5$ population success units and $N-M = 15$ population failure units we find that $P(S) = 5/20$, $P(F|S) = 15/19$, and $P(S|SF) = 4/18$, so that $P(SFS) = (5/20)(15/19)(4/18)$. The eight possible sequences of $S's$ and $F's$ and the associated probabilities and values of X are given in Table 6.

Table 6. Outcomes of a simple random sample of size 3 selected without replacement, when there are M=5 population success units and N=20 population units.

sequence	probability	X
FFF	$(15 \cdot 14 \cdot 13)/(20 \cdot 19 \cdot 18)$	0
SFF	$(5 \cdot 15 \cdot 14)/(20 \cdot 19 \cdot 18)$	1
FSF	$(15 \cdot 5 \cdot 14)/(20 \cdot 19 \cdot 18)$	1
FFS	$(15 \cdot 14 \cdot 5)/(20 \cdot 19 \cdot 18)$	1
SSF	$(5 \cdot 4 \cdot 15)/(20 \cdot 19 \cdot 18)$	2
SFS	$(5 \cdot 15 \cdot 4)/(20 \cdot 19 \cdot 18)$	2
FSS	$(15 \cdot 5 \cdot 4)/(20 \cdot 19 \cdot 18)$	2
SSS	$(5 \cdot 4 \cdot 3)/(20 \cdot 19 \cdot 18)$	3

Table 7. Probability distribution for the number of successes in a simple random sample of size 3 selected without replacement, when there are M=5 population success units and N=20 population units.

sum X	probability P(X)
0	$(15 \cdot 14 \cdot 13)/(20 \cdot 19 \cdot 18) = .3991$
1	$3(5 \cdot 15 \cdot 14)/(20 \cdot 19 \cdot 18) = .4605$
2	$3(5 \cdot 4 \cdot 15)/(20 \cdot 19 \cdot 18) = .1316$
3	$(5 \cdot 4 \cdot 3)/(20 \cdot 19 \cdot 18) = .0088$

Since the eight sequences in Table 6 are mutually exclusive, we can add the probabilities corresponding to each distinct value of X to get the probability distribution in Table 7. This distribution of the number of successes in a simple random sample of size 3 selected without replacement from a dichotomous population of $N = 20$ population units with $M = 5$ population success units is the hypergeometric distribution with parameters $M = 5$, $N = 20$, and $n = 3$. Since only M population units are classified as successes and only $N - M$ are classified as failures, we must have $x \leq M$ and $n - x \leq N - M$. Subject to

these restrictions the hypergeometric probabilities can be computed using the probability function

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

Most statistical software programs and some calculators will compute these probabilities or the cumulative probabilities $P(X \leq x)$. The hypergeometric distribution with parameters M, N, and n has mean

$$\mu = np = n\left(\frac{M}{N}\right)$$

and variance

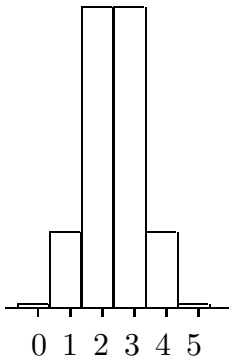
$$\sigma^2 = n\left(\frac{M}{N}\right)\left(\frac{N-M}{N}\right)\left(\frac{N-n}{N-1}\right) = npq\left(\frac{N-n}{N-1}\right),$$

where $p = M/N$ is the population success proportion.

Notice that the mean number of successes $\mu = np$ is the same whether we sample with replacement (the binomial (n, p) mean) or without replacement (the hypergeometric (M, N, n) mean with $p = M/N$). However, the variance of the hypergeometric (without replacement) distribution is smaller by a factor of $f = (N-n)/(N-1)$, *i.e.*, the binomial variance is npq and the hypergeometric variance is $fnpq$. The factor f is known as the finite population correction factor and its effect is most noticeable when N is small relative to n. If N is very large relative to n, then $f \approx 1$ and the two variances are essentially equal. Actually, if N is very large relative to n, then the binomial and hypergeometric distributions are essentially the same. The difference between the binomial and the hypergeometric distributions is illustrated, for $n = 5$ and $p = .5$, by the probability histograms in Figure 3. All of the distributions of Figure 3 are symmetric with mean $\mu = 2.5$, since we have $n = 5$ trials and the population success proportion is $p = .5$. Notice that as the size of the population increases the hypergeometric distributions become more similar to the binomial distribution and, in particular, there is very little difference between the hypergeometric distribution with $M = 50$ and $N = 100$ and the binomial distribution.

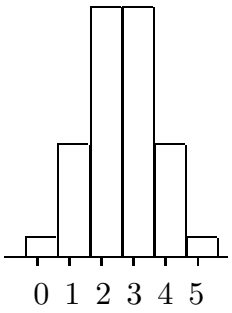
If X denotes the number of successes in a simple random sample of size n selected from a dichotomous population with population success proportion p, then $\hat{p} = X/n$, the proportion of successes in the sample, provides an obvious estimator of the population success proportion p. The distribution of $\hat{p}$ describes the variability (from sample to sample) in $\hat{p}$ as an estimator of p. The distribution of $\hat{p}$ is also known as the sampling distribution of $\hat{p}$. The binomial and hypergeometric distributions can be used to determine the respective distributions of $\hat{p}$ when sampling with replacement and sampling without replacement.

Figure 3. Hypergeometric and binomial probability histograms, with p=.5.



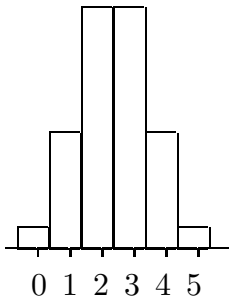
hypergeometric

$(n = 5, M = 5, N = 10, \mu = 2.5, \sigma^2 = .9868)$



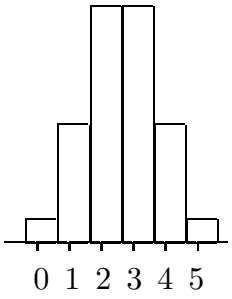
hypergeometric

$(n = 5, M = 20, N = 40, \mu = 2.5, \sigma^2 = 1.1218)$



hypergeometric

$(n = 5, M = 50, N = 100, \mu = 2.5, \sigma^2 = 1.2249)$



binomial

$(n = 5, p = .5, \mu = 2.5, \sigma^2 = 1.25)$

If X denotes the number of successes in a simple random sample of size n selected with replacement from a dichotomous population with population success proportion p, then the possible values of X are $x = 0, 1, \ldots, n$ and the corresponding possible values of $\hat{p}$ are $x/n = 0, 1/n, 2/n, \ldots, 1$. Thus, when the sample is selected with replacement, the probability that $X = x$ is equal to the probability that $\hat{p} = x/n$, *i.e.,* for $x = 0, 1, \ldots, n$,

$$P\left(\hat{p} = \frac{x}{n}\right) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

In this case, the mean of the sampling distribution of $\hat{p}$ is $E(\hat{p}) = p$ $(E(\hat{p}) = E(X)/n)$ and the variance of the sampling distribution of $\hat{p}$ is $\text{var}(\hat{p}) = pq/n$ $(\text{var}(\hat{p}) = \text{var}(X)/n^2)$. Notice that this sampling distribution does not depend on the size of the population.

If X denotes the number of successes in a simple random sample of size n selected without replacement from a dichotomous population with population success proportion p, then the sampling distribution of $\hat{p}$ depends on the size of the population. Let N denote the size of the dichotomous population being sampled and let M denote the number of

population units classified as successes. When the sample is selected without replacement, for $x = 0, 1, \ldots, n$, subject to the restrictions $x \leq M$ and $n - x \leq N - M$,

$$P\left(\hat{p} = \frac{x}{n}\right) = P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

In this case, the mean and variance of the sampling distribution of $\hat{p}$ are

$$E(\hat{p}) = p = \left(\frac{M}{N}\right)$$

and

$$\text{var}(\hat{p}) = \left(\frac{M}{N}\right)\left(\frac{N-M}{N}\right)\left(\frac{N-n}{n(N-1)}\right) = \frac{pq}{n}\left(\frac{N-n}{N-1}\right),$$

where $p = M/N$ is the population success proportion. Note that, as before, $E(\hat{p}) = E(X)/n$ and $\text{var}(\hat{p}) = \text{var}(X)/n^2$.

We will now consider the Poisson distribution which provides a realistic model for counts of "rare events" in many practical settings. Consider a sequence of events occurring randomly in time or space and a count such as the number of radioactive particle emissions per unit time, the number of meteorites that collide with a satellite during a single orbit, the number of defects per unit length of some material, or the number of weed seeds per unit volume in a large batch of wheat seeds. We can picture the time (or location) of each occurrence as a point on the positive part of the number line. Consider the following assumptions about the times (locations) of these occurrences:

1) The probability of exactly one occurrence in a small interval of length t is approximately νt, where $\nu > 0$ is the mean rate at which events occur per unit time (the mean rate of occurrence).

2) The probability of more than one occurrence in a small interval of length t is negligible compared to the probability of exactly one occurrence in a small interval of length t.

3) The numbers of occurrences in non–overlapping intervals are independent in the sense that information concerning the number of events in one interval reveals nothing about the number of events in the other interval.

If we let X denote the number of occurrences in a period of length t, then these three assumptions imply that X follows the Poisson distribution with parameter $\lambda = \nu t$. The possible values of X are $0, 1, \ldots$, with no theoretical upper bound on the value, and for $\lambda > 0$ the Poisson probabilities can be computed using the probability function

$$P(X = x) = \frac{\lambda^x}{x!}e^{-\lambda},$$

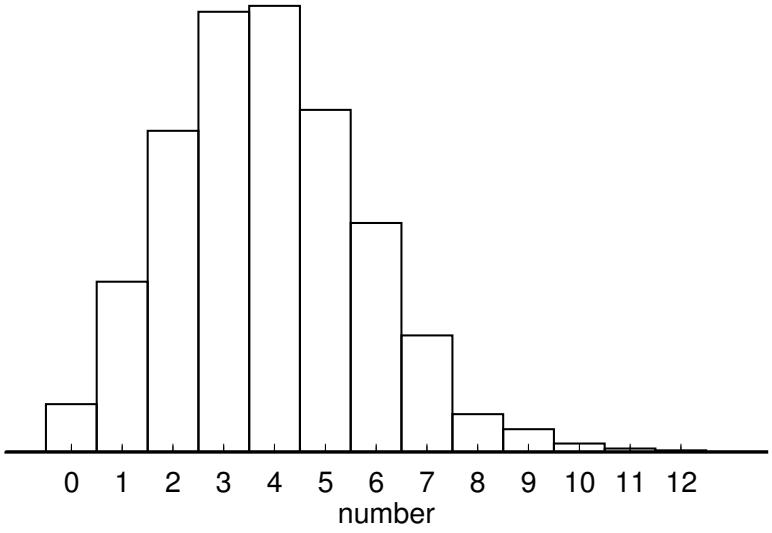
where $e \approx 2.718$ is the base of the natural logarithm and $x! = x(x-1)\cdots 1$ is x factorial. The mean and variance of the Poisson distribution with parameter λ are both equal to λ.

Example. Radioactive disintegrations. This example is taken from Feller (1957), p. 149 and Cramér (1946) p. 436. In a famous experiment by Rutherford, Chadwick, and Ellis (*Radiations from Radioactive Substances*, Cambridge, 1920) a radioactive substance was observed during 2608 consecutive time intervals of length $t = 7.5$ seconds each. The number of particles reaching a counter was recorded for each period. The results are summarized in Figure 4 and Table 8. (In Table 8 the observations greater than or equal to 10 are grouped together. The data actually contain 10 tens, 4 elevens, and 2 twelves.) The last column of Table 8 contains expected relative frequencies (probabilities) computed using a Poisson model with λ estimated from these data. These Poisson probabilities appear to match the observed relative frequencies fairly well. A formal test of the goodness of fit of this Poisson model to these data, which is discussed in Chapter 11, indicates that the model does fit well ($\chi^2 = 12.885$, 9 d.f., P–value .17).

Table 8. Relative frequency distribution for radioactive disintegrations.

number	observed frequency	observed relative frequency	expected relative frequency
0	57	.0219	.0209
1	203	.0778	.0807
2	383	.1469	.1562
3	525	.2013	.2015
4	532	.2040	.1949
5	408	.1564	.1509
6	273	.1047	.0973
7	139	.0533	.0538
8	45	.0173	.0260
9	27	.0104	.0112
≥ 10	16	.0051	.0065
total	2608	.9991	.9999

Figure 4. Histogram for radioactive disintegrations (with ≥ 10 expanded).



Example. Bacteria counts. This example is taken from Feller (1957), p.153. The original source is T. Matuszewsky, J. Supinska, and J. Neyman (1936), *Zentralblatt für Bakteriologie, Parasitenkunde und Infektionskrankrankheiten*, II Abt., **95**. A Petri dish with bacteria colonies was examined under a microscope. The dish was divided into small squares and the number of bacteria colonies, visible as dark spots, was recorded for each square. In this example t is the area the square within which the count is determined and we will take this area to be one. If the bacteria colonies were randomly distributed over the Petri dish, without being clustered together, then the Poisson model should hold. The results for one of several experiments are summarized in Figure 5 and Table 9. The last column of Table 9 contains expected relative frequencies (probabilities) computed using a Poisson model, with λ estimated from these data. In this example the observed relative frequency in the "≥ 6" line is for "exactly 6", but, the expected relative frequency is for all values greater than or equal to 6. These Poisson probabilities appear to match the observed relative frequencies fairly well. Therefore, the evidence supports the contention that the bacteria colonies are randomly distributed over the Petri dish. A formal test of the goodness of fit of this Poisson model to these data, which is discussed in Chapter 11, indicates that the model does fit well ($\chi^2 = .8386$, 5 d.f., P–value .9745).

Figure 5. Histogram for bacteria counts.

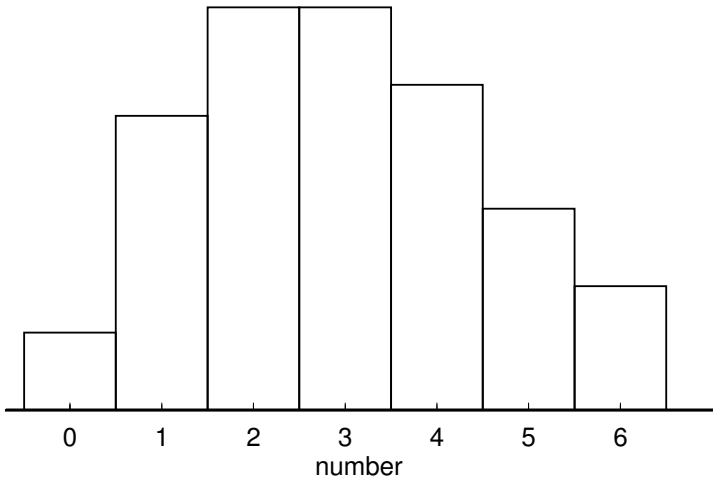


Table 9. Relative frequency distribution for bacteria counts.

number	observed frequency	observed relative frequency	expected relative frequency
0	5	.0424	.0533
1	19	.1610	.1562
2	26	.2203	.2290
3	26	.2203	.2239
4	21	.1780	.1641
5	13	.1102	.0962
≥ 6	8	.0678	.0772
total	118	1.0000	.9999

4a.5 Probability models for continuous quantitative variables

We will now consider probability models for the distribution of a continuous quantitative variable. A probability model for the distribution of a continuous variable X can be represented by a density curve. A **density curve** is a nonnegative curve for which the area under the curve (over the x–axis) is one. We can think of the density curve as a smooth version of a probability histogram with the rectangles of the histogram replaced by a smooth curve indicating where the tops of the rectangles would be. With a continuous variable X it does not make sense to talk about the probability that X would take on a particular value, after all if we defined positive probabilities for the infinite collection (continuum) of possible values of X these probabilities could not add up to one. It does, however, make sense to talk about the probability that X will take on a value in a specified interval or range of values. Given two constants $a < b$ the probability that X takes on a

value in the interval from a to b, denoted by $P(a \leq X \leq b)$, is equal to the area under the density curve over the interval from a to b on the x–axis. Areas of this sort based on the density curve give the probabilities which a single value of X, chosen at random from the infinite population of possible values of X, will satisfy.

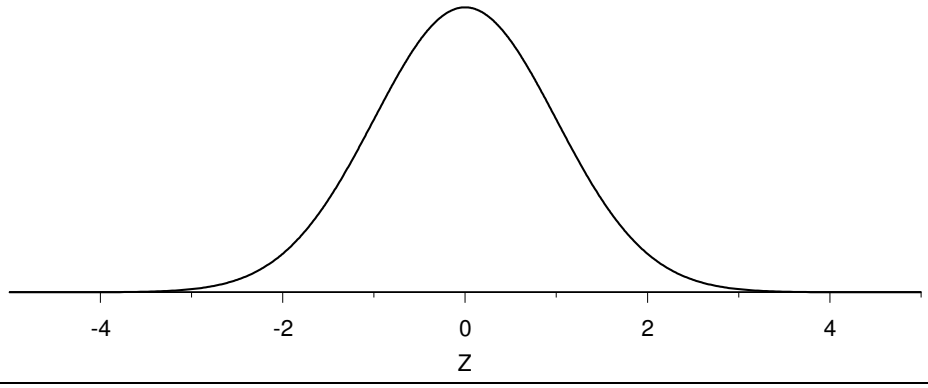
Given a probability model for the distribution of a continuous variable X, *i.e.*, given a density curve for the distribution of the continuous variable X, we can define population parameters which characterize relevant aspects of the distribution. For example, we can define the population mean μ as the balance point of the unit mass bounded by the density curve and the number line. We can also think of the population mean as the weighted average of the infinite collection of possible values of X with weights determined by the density curve. We can similarly define the population median M as the point on the number line where a vertical line would divide the area under the density curve into two equal areas (each of size one–half).

The most widely used continuous probability model is the normal probability model or normal distribution. The normal distribution with mean μ and standard deviation σ can be characterized by its density curve, which is the familiar bell shaped curve. The normal density curve corresponds to the probability density function

$$f(x) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

The standard normal density curve, which has mean $\mu = 0$ and standard deviation $\sigma = 1$, is shown in Figure 6.

Figure 6. The standard normal density curve.



The normal distribution with mean μ and its density curve are symmetric around μ, *i.e.*, if we draw a vertical line through μ, then the two sides of the density curve are mirror images of each other. Therefore the mean of a normal distribution μ is also the median of the normal distribution. The mean μ locates the normal distribution on the number line so that if we hold σ constant and change the mean μ, the normal distribution is simply shifted

along the number line until it is centered at the new mean. In other words, holding σ fixed and changing μ simply relocates the density curve on the number line; it has no effect on the shape of the curve. Figure 7 provides the density curves for normal distributions with respective means $\mu = 0$ and $\mu = 2$ and common standard deviation $\sigma = 1$.

Figure 7. Normal distributions with common standard deviation one and means of zero and two.

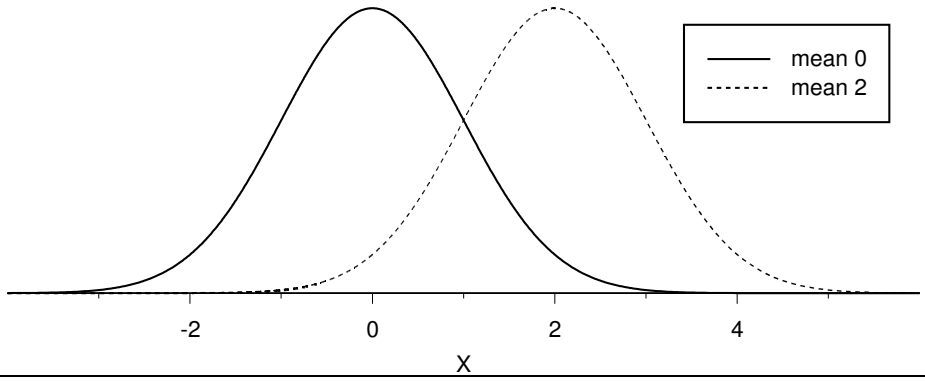
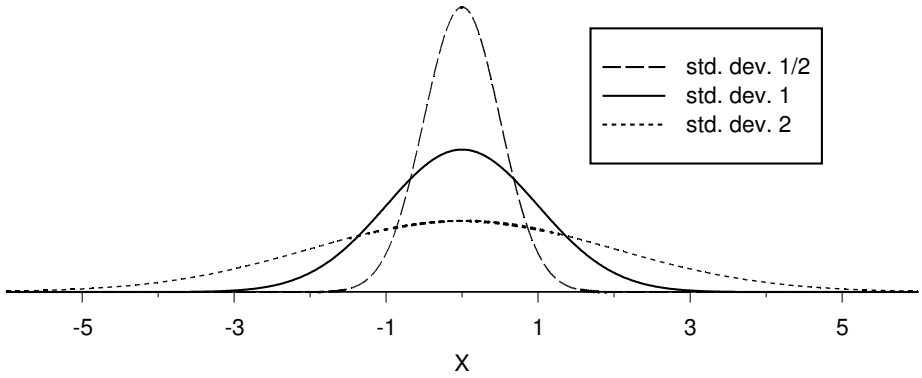


Figure 8. Normal distributions with common mean zero and standard deviations one–half, one, and two.



The standard deviation σ indicates the amount of variability in the normal distribution. If we hold μ fixed and increase the value of σ, then the normal density curve becomes flatter, while retaining its bell–shape, indicating that there is more variability in the distribution. Similarly, if we hold μ fixed and decrease the value of σ, then the normal density curve becomes more peaked around the mean μ, while retaining its bell–shape, indicating that there is less variability in the distribution. Normal distributions with mean $\mu = 0$ and respective standard deviations $\sigma = .5$, $\sigma = 1$, and $\sigma = 2$ are plotted in Figure 8.

Computer programs and many calculators can be used to compute normal probabilities or equivalently to compute areas under the normal density curve. These probabilities can also be calculated using tables of standard normal distribution probabilities such as Table

10 at the end of this chapter. Recall that the standard normal distribution is the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The relationship between the standard normal variable Z and the normal variable X, which has mean μ and standard deviation σ, is

$$Z = \frac{X - \mu}{\sigma} \quad \text{or equivalently} \quad X = \mu + Z\sigma.$$

This relationship implies that a probability statement about the normal variable X can be re–expressed as a probability statement about the standard normal variable Z by re–expressing the statement in terms of standard deviation units from the mean. Given two constants $a < b$, observing a value of X between a and b (observing $a \leq X \leq b$) is equivalent to observing a value of $Z = (X - \mu)/\sigma$ between $(a - \mu)/\sigma$ and $(b - \mu)/\sigma$ (observing $(a - \mu)/\sigma \leq (X - \mu)/\sigma \leq (b - \mu)/\sigma$). Furthermore, $Z = (X - \mu)/\sigma$ behaves in accordance with the standard normal distribution so that the probability of observing a value of X between a and b, denoted by $P(a \leq X \leq b)$, is equal to the probability that the standard normal variable Z takes on a value between $(a - \mu)/\sigma$ and $(b - \mu)/\sigma$, which is denoted by $P[(a - \mu)/\sigma < Z < (b - \mu)/\sigma]$, *i.e.*,

$$P(a < X < b) = P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right).$$

In terms of areas this probability equality says that the area under the normal density curve with mean μ and standard deviation σ over the interval from a to b is equal to the area under the standard normal density curve over the interval from $(a-\mu)/\sigma$ to $(b-\mu)/\sigma$. Similarly, given constants $c < d$, we have the analogous result that

$$P(c < Z < d) = P(\mu + c\sigma < X < \mu + d\sigma).$$

Most tables of the standard normal distribution and many computer programs provide cumulative standard normal probabilities of the form $P(Z \leq a)$ for selected values of a. To use these cumulative probabilities to compute a probability of the form $P(a \leq Z \leq b)$ note that

$$P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$$

and note that the symmetry of the normal distribution implies that

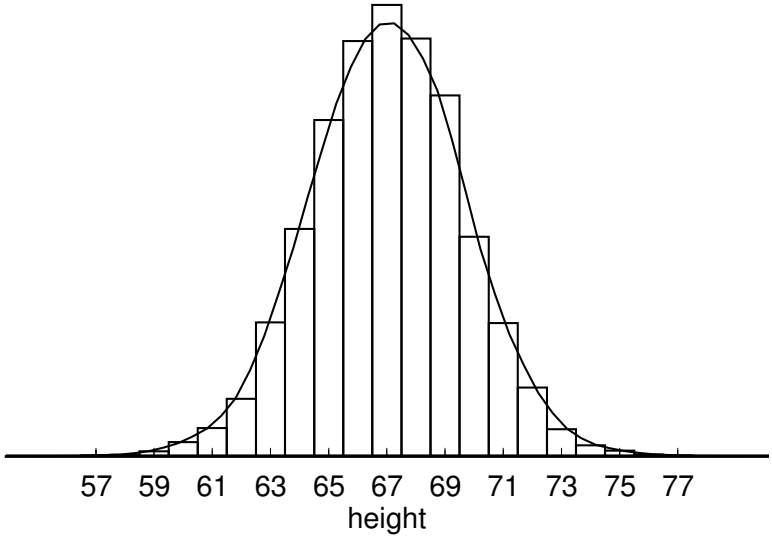
$$P(Z \leq -a) = P(Z \geq a) = 1 - P(Z \leq a).$$

Calculators will usually provide probabilities of the form $P(a \leq Z \leq b)$ directly.

Example. Heights of adult males. Consider the heights (in inches) of adult males born in the United Kingdom (including the whole of Ireland) which are summarized in the Table 8 of Section 3.3.

These height data provide a good illustration of the fact that normal distributions often provide very good models for a population of physical measurements of individuals, such as heights or weights. Figure 9 provides a histogram for this height distribution and the density curve for a normal distribution chosen to model these data. You can see that the normal distribution provides a very reasonable model for the heights of adult males born in the United Kingdom.

Figure 9. Histogram and normal density curve for the UK height example.



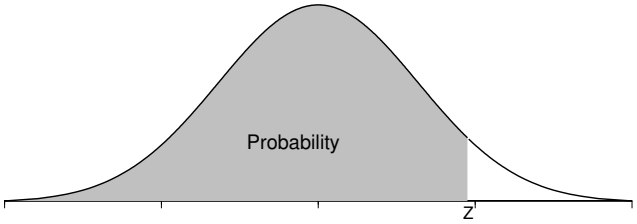


Table 10. Cumulative normal probabilities.
(Areas under the standard normal curve to the left of Z.)

Z	Second decimal place in Z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

continued on next page

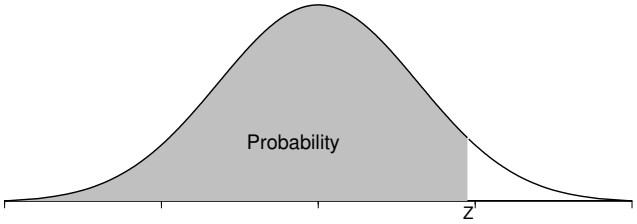


Table 10 (continuation). Cumulative normal probabilities.
(Areas under the standard normal curve to the left of Z.)

Z	Second decimal place in Z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

List of Examples

Index