An Introduction to Probability and Statistics

(statistics 325)

J. Calvin Berry
Mathematics Department
University of Louisiana at Lafayette

http://www.ucs.louisiana.edu/~jcb0773/

December 2017 edition (hyperlinked 26 June 2018)

20180626

Table of Contents

Part 1: Introduction and Descriptive Statistics

Part 2: Probability

Part 3: Random Variables

Part 4: Inferential Statistics

1 Introduction toc

1.1 Basic ideas toc

The entity of primary interest is a group

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.*

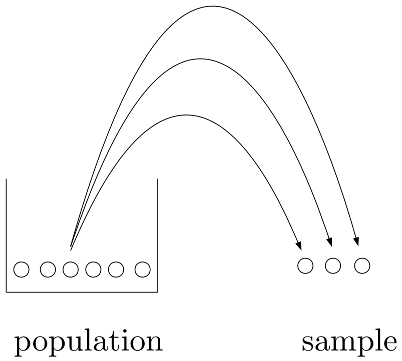
Information about a unit – variables

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.

The population and the sample

The population is the collection of all of the units that we are interested in. The sample is the subset of the population that we will examine. (We will define a sample more precisely when we discuss random sampling.)

Figure 1.1 Population (box of balls) and sample. A ball represents a population unit. The balls removed from the box represent the sample.



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.

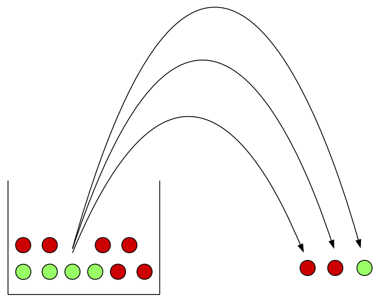
In some applications, such as experimental studies, the population is best viewed as a hypothetical population of values of one or more variables. For example, suppose that we are interested in the effects of an alternative diet on weight gain in some population of experimental animals. We might conduct 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.

Information about a group – parameters and statistics

Recall that a **variable** is a measurable characteristic of an individual unit. One way to characterize a group of units is to examine the values of the variable corresponding to all of the units in the group and determine one or more suitable summary values. For example, given a group of adults, we might compute the average age of the group or the proportion who have full–time jobs. A **parameter** is a numerical characteristic of the population. A **statistic** is a numerical characteristic of the sample. That is, a parameter is a number which characterizes a population and a statistic is a number which characterizes a sample.

An illustration is provided in Figure 1.2 for a population of 10 balls and sample of 3 balls. In this figure the characteristic of interest is the color of the ball and the color red (darker shade) is of particular interest. The parameter is the proportion of red balls in the population, 6/10, and the statistic is the proportion of red balls in the sample, 2/3.

Figure 1.2 Parameter and statistic for a population of 10 balls and sample of 3 balls.



parameter $p = 6/10$ (population proportion red)

statistic $\hat{p} = 2/3$ (sample proportion red)

Example 1.1 NHANES The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. We will use some data from the 2013–2014 NHANES to illustrate the basic ideas we are discussing. For now we will concentrate our attention on some body size measurements for the 5588 adults (age 20 and over) in the 2013–2014 NHANES.

For present purposes we will view this group of $N = 5588$ adults as the population. In the original context of the survey this is a sample. Thus, for our purposes:

1. A unit is an individual adult.

2. The population is the collection of $N = 5588$ adults about whom we have information.

3. The sample is a collection of $n = 50$ individuals (units) which I selected at random from the population of $N = 5588$ adults.

4. We will consider six variables:
The sex of the person (male or female);
The age of the person (years);
The weight of the person (pounds);
The height of the person (inches);
The BMI (body mass index) of the person (kg/m^2); and,
The waist circumference of the person (inches).

Table 1.1 contains the values of the six variables for the $n = 50$ people in the sample. The values in a particular row correspond to an individual (one unit).

Table 1.1 NHANES 2013-2014 simple random sample of $n = 50$.

line	sex	age	weight	height	bmi	waist
1	male	48	285.34	70.1181	40.9	48.8976
2	female	80	172.48	63.5433	30.1	41.6535
3	male	48	186.78	69.8819	26.9	37.5197
4	male	80	167.2	67.5591	25.8	37.0079
5	female	20	156.2	67.7953	23.9	32.2835
6	female	43	196.02	66.1811	31.5	46.8504
7	male	54	200.86	64.8425	33.7	42.5197
8	female	24	153.78	63.4252	26.9	40.7874
9	female	25	122.32	64.1732	20.9	29.0945
10	male	58	142.34	69.0157	21.1	34.252
11	female	74	154.22	63.1496	27.2	37.8346
12	female	74	173.8	61.378	32.5	45.3937
13	male	78	154.22	67.4016	23.9	33.4646
14	female	72	89.1	56.9291	19.4	29.1339
15	male	48	178.2	72.2835	24	36.811
16	male	64	231.66	62.874	41.3	51.7323
17	male	41	164.34	68.0315	25	36.7717
18	female	39	143.88	60.7874	27.4	33.7402
19	male	49	305.14	72.1654	41.3	missing
20	male	73	141.24	70.3937	20.1	37.4409
21	female	67	203.94	62.0079	37.4	44.7244
22	female	26	101.86	59.8031	20.1	29.0157
23	female	73	150.7	65.7087	24.6	38.189
24	male	60	199.32	72.1654	27	40.0394
25	male	40	206.58	67.4409	32	43.2677
26	male	27	181.94	67.3622	28.2	38.622
27	male	62	199.98	67.3622	31.1	43.3858
28	female	71	176.66	64.6063	29.8	42.9134
29	female	80	166.1	62.3228	30.1	39.3701
30	male	39	280.72	71.9291	38.2	52.0472
31	female	48	127.6	61.4173	23.8	31.2598
32	male	46	156.64	67.7953	24	missing
33	female	80	147.4	61.3386	27.6	40.7087
34	female	35	183.04	65.2362	30.3	41.2205
35	female	57	140.36	62.2047	25.6	34.9213

continued below

Table 1.1 continuation of NHANES 2013-2014 simple random sample of $n = 50$.

line	sex	age	weight	height	bmi	waist
36	female	40	170.06	65.9055	27.6	35.748
37	female	56	182.6	59.685	36.1	39.4882
38	male	34	192.5	70.3937	27.4	37.9921
39	male	75	163.9	64.7244	27.6	48.8189
40	female	76	147.84	60.4724	28.5	36.6535
41	male	57	254.32	70.7087	35.8	43.8189
42	male	77	187.22	69.2126	27.5	40.1181
43	male	45	179.3	70.7087	25.3	36.5354
44	male	48	249.48	67.126	39	51.063
45	female	47	214.06	64.6457	36.1	42.5984
46	male	72	205.26	70.3543	29.2	42.8346
47	male	80	127.38	66.4567	20.3	missing
48	male	24	199.32	71.2205	27.7	40.6299
49	male	80	162.8	66.7323	25.8	39.4094
50	male	21	119.24	66.6929	18.9	26.3386

Some parameters and statistics (population and sample means) for the numerical variables in this example are given in Table 1.2. With respect to the categorical variable sex of the person; there are 2919 females and 2669 males in the population and there are 22 females and 28 males in the sample. This gives the (parameter) population percentage female as 52.24% and the (statistic) sample percentage female as 44%.

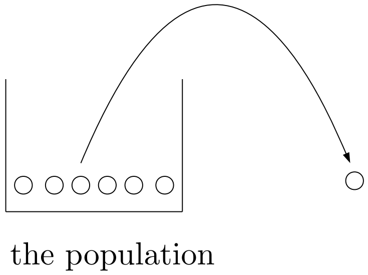
Table 1.2 Population and sample means for the HANES example.

	parameter	**statistic**
variable	**population mean**	**sample mean**
age	49.15	54.70
weight	179.22	177.94
height	65.77	66.11
bmi	29.10	28.53
waist	39.05	39.47

Which direction? Probability versus statistics

Probability theory is used to model the outcomes of random processes. The basic probability situation is illustrated in Figure 1.3. Here we know all there is to know about the characteristics of the balls in the box and we want to make a statement about what will happen when we select a ball at random from the box and examine it. For example, for the box of Figure 1.2, where 60% of the balls in the box are red, if we select one ball at random, there is a 60% chance (probability) that it will be red.

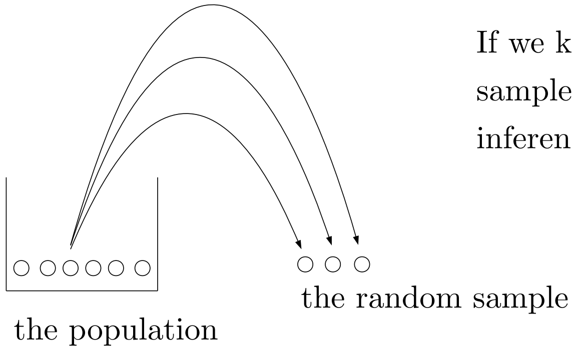
Figure 1.3 Probability: What will happen when we select a ball at random?



If we know all about the balls in the box, then we can assign probabilities to the outcomes we may observe when we select a ball at random.

Statistical theory is used to make inferences from a random sample to a population. The basic statistics situation is illustrated in Figure 1.4. Here we know all there is to know about the characteristics of the balls in the random sample and we want to make a statement about what we would find if we examined all of the balls in the box (the entire population).

Figure 1.4 Statistics: What can we say about the balls in the box?



If we know all about the balls in the random sample, then we can use statistics to make inferences about the balls in the box.

1.2 Sampling

toc

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 second consideration concerns the proper selection of the units which constitute the sample. We will discuss these considerations in more depth in the rest of this section.

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.

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. This scenario was illustrated in Figure 1.1.

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.**

To obtain a **simple random sample of size n** 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 determine 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. If we are sampling with replacement we return the ball to the box before the next draw. If we are sampling without replacement we do not return the ball to the box. We then mix the balls in the box and continue this process of selecting a ball from the box at random until n balls have been selected. These n balls (or the values of the variable for these balls) form the simple random sample of size n.

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. For most purposes it is best to use a computer or calculator to select a simple random sample. The computer will simulate the process of drawing balls at random from a box.

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 so that the sample does not 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.

1.3 Experimentation

toc

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(s) 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.

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.

2 Descriptive Statistics toc

2.1 Tabular and graphical summary toc

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. A variable is chosen to measure some characteristic of the units in the group; therefore, the distribution of a variable contains all of the available information about the characteristic (as measured by that variable) for the group. 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.

Generally speaking, it is hard to tell much about the distribution of a variable by examining the data in raw form. 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 Types of variables

toc

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. If there is an inherent ordering of the possible values of a qualitative variable, then it is said to be **ordinal**. For example the sex (female or male) of a college student is nominal while the classification (freshman, sophomore, junior, senior) is ordinal.

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. 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.

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.

2.3 Qualitative data

toc

Bar graphs are used to summarize the distribution of a qualitative variable. 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 disk, the result is a pie graph or pie chart. In a **pie graph** or pie chart the interior of a disk (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.

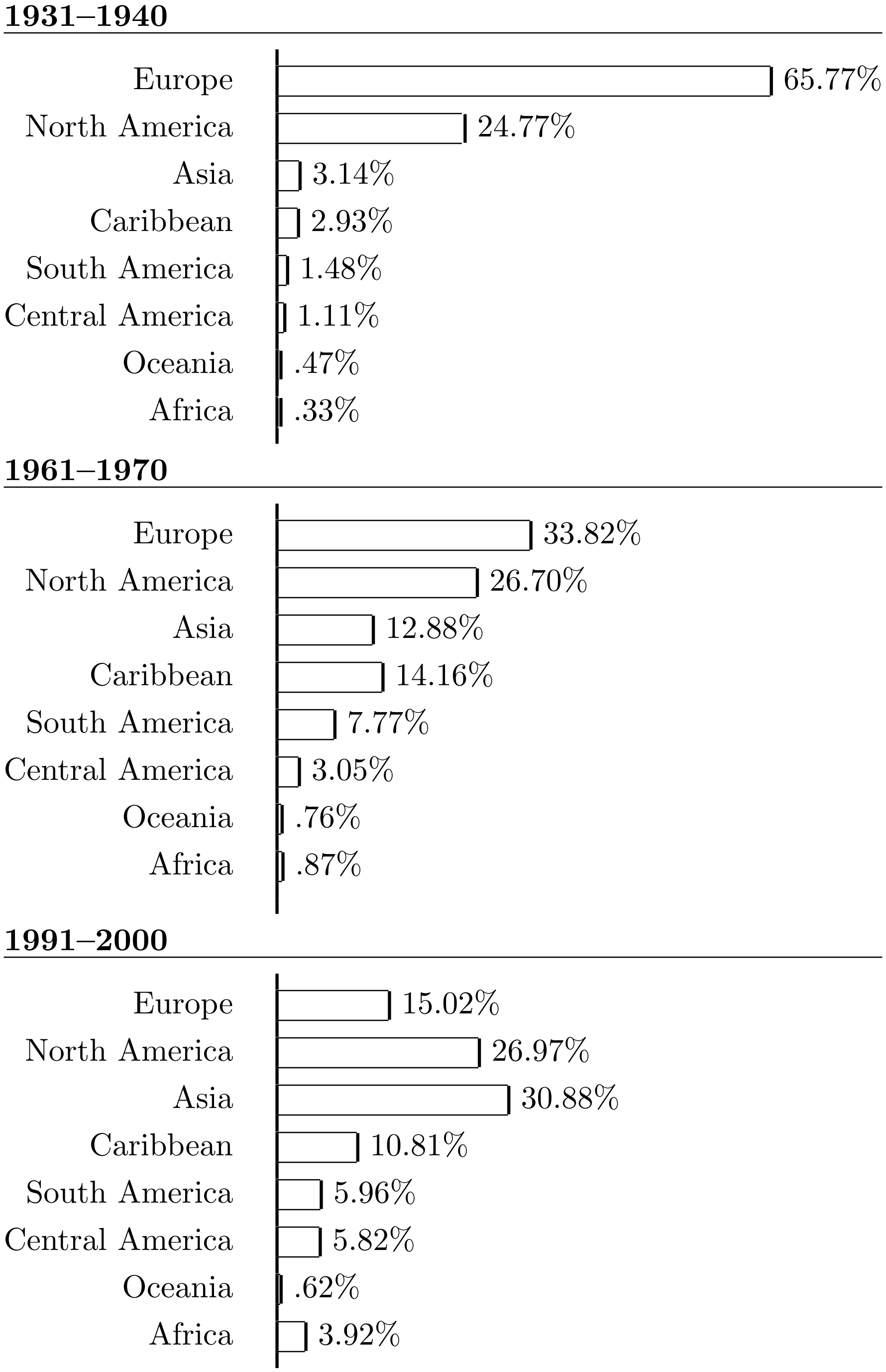
Example 2.1 Immigrants to the United States. The data concerning immigrants admitted to the United States summarized by decade as raw frequency distributions in Table 2.1 were taken from the *2002 Yearbook of Immigration Statistics*, USCIS, (`www.uscis.gov`). Immigrants for whom the country of last residence 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 2.1 have been arranged so that the percentages for the 1931–1940 decade are in decreasing order.

Table 2.1 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

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.

Figure 2.1 Region of last residence for immigrants to USA, by decade.



2.4 Quantitative data

toc

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.

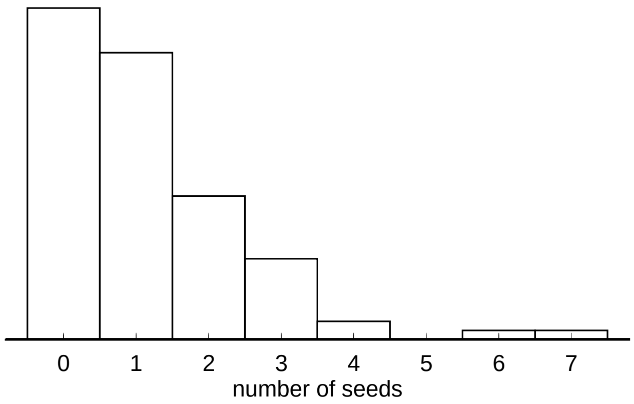
We will use a graphical representation called a histogram to summarize the distribution of a discrete quantitative variable. Like the bar graph we used to represent the distribution of a qualitative variable, the histogram provides a representation of the distribution of a quantitative variable using 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). We will use the following weed seed example to illustrate the methodology.

Example 2.2 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 2.2 and the histogram of Figure 2.2. 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 $0, 1, 2, \ldots$.

Table 2.2 Weed seeds 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 2.2 Histogram for number of weed seeds.



Consider the histogram for the number of weed seeds in a batch of grass of Figure 2.2. 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 is a gap at 5 since none of the batches had exactly 5 weed seeds.

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 2.2 represents the relative frequency that a batch of grass contains exactly 1 weed seed;

but, its base extends from .5 to 1.5 on the number line. Because it is impossible for the number of weed seeds 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.

The histogram for the distribution of the number of weed seeds in Figure 2.2 has a mound shaped appearance with a single peak at zero, indicating that the most common number of weed seeds is zero. In fact, 37.76% of the batches of grass contain no weed seeds. Among the batches that do contain weed seeds we see that 32.65% contain one weed seed and 16.33% contain two. Thus, 86.74% of the 98 batches of grass contain two or fewer weed seeds and 95.92% contain three or fewer weed seeds. In summary, 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.

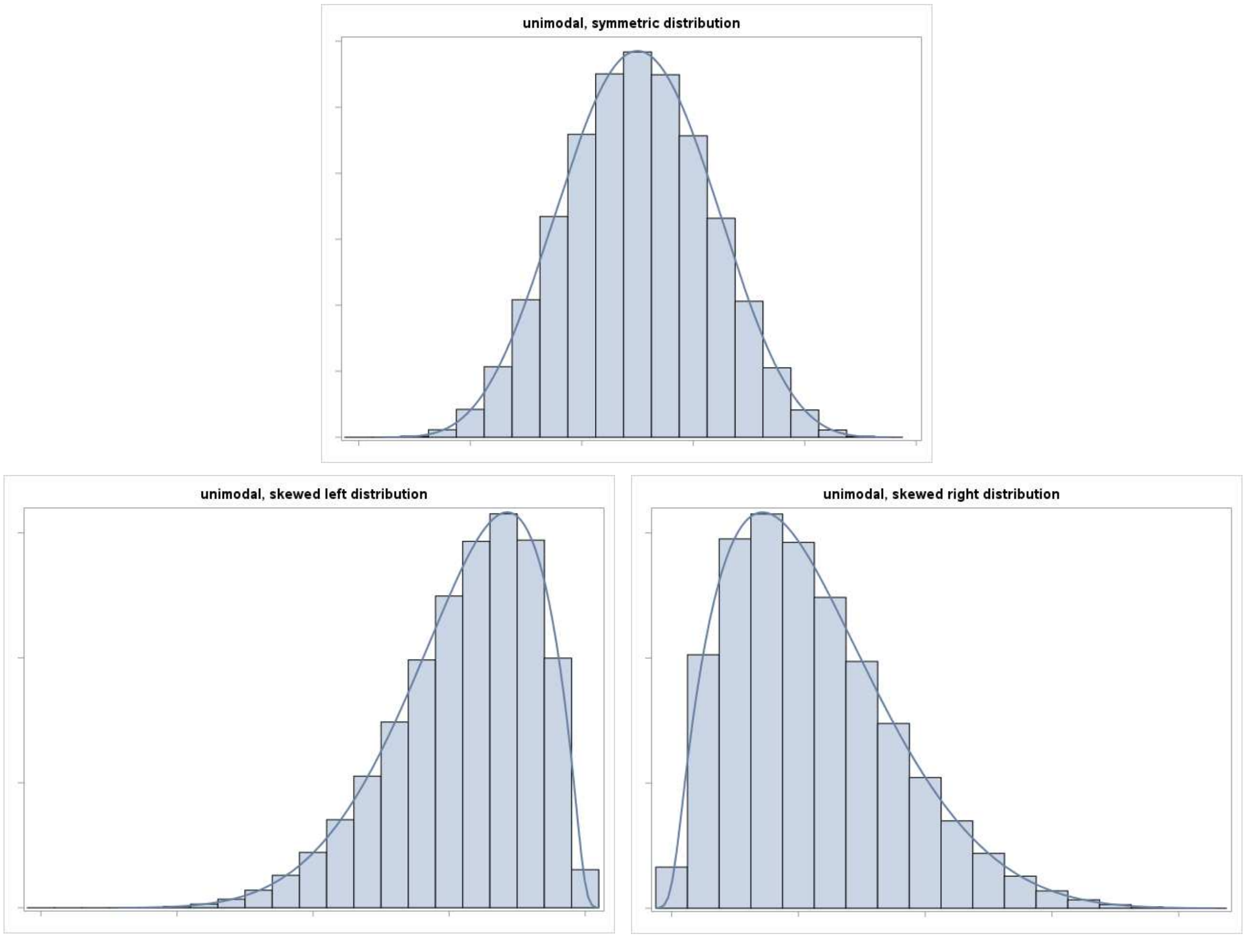
The histogram of Figure 2.2, or the associated distribution, is not symmetric. That is, the histogram (distribution) is not the same on the left side (smaller values) as it is on the right side (larger values). This histogram or distribution is said to be skewed to the right. The concept of a distribution 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 two to seven. In the present example we might say that there is essentially no left tail in the distribution.

The number of weed seeds histogram provides an example of a very common type of histogram (distribution) which is mound shaped and has a single peak. (A distribution with a single peak is said to be unimodal.) 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 mound shaped distributions with a single peak are provided in Figure 2.3. For these illustrations a smooth curve is used to indicate the shape of the histogram. 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 2.3 Distribution shapes – symmetry and skewness

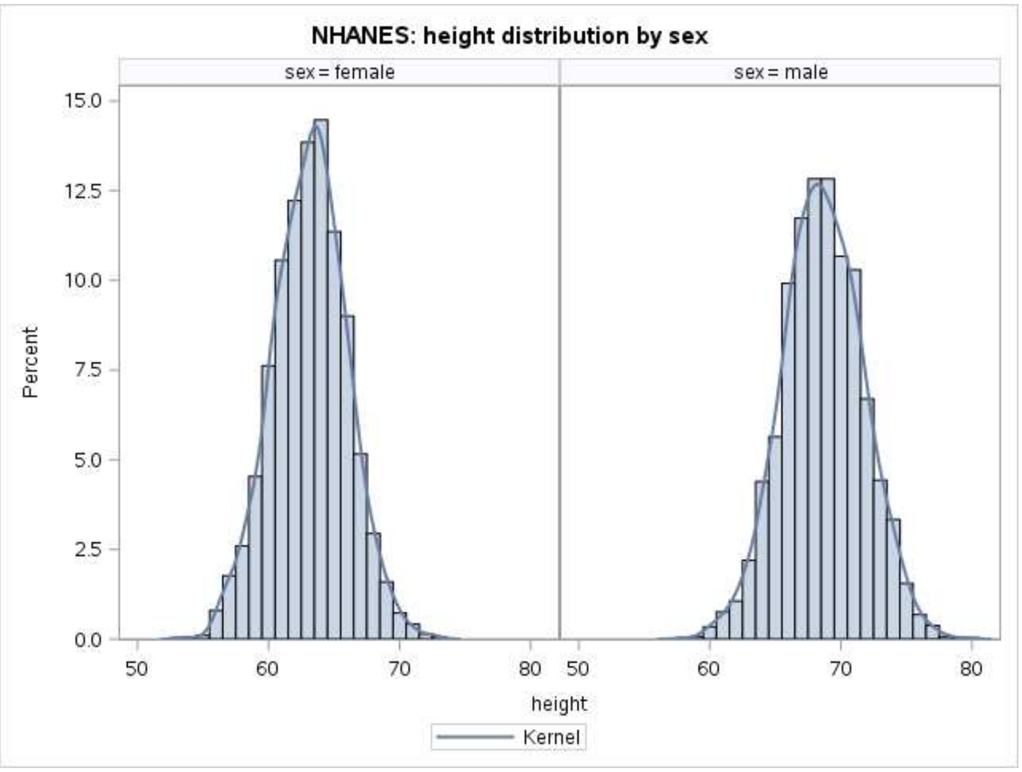


skewed left skewed right

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.

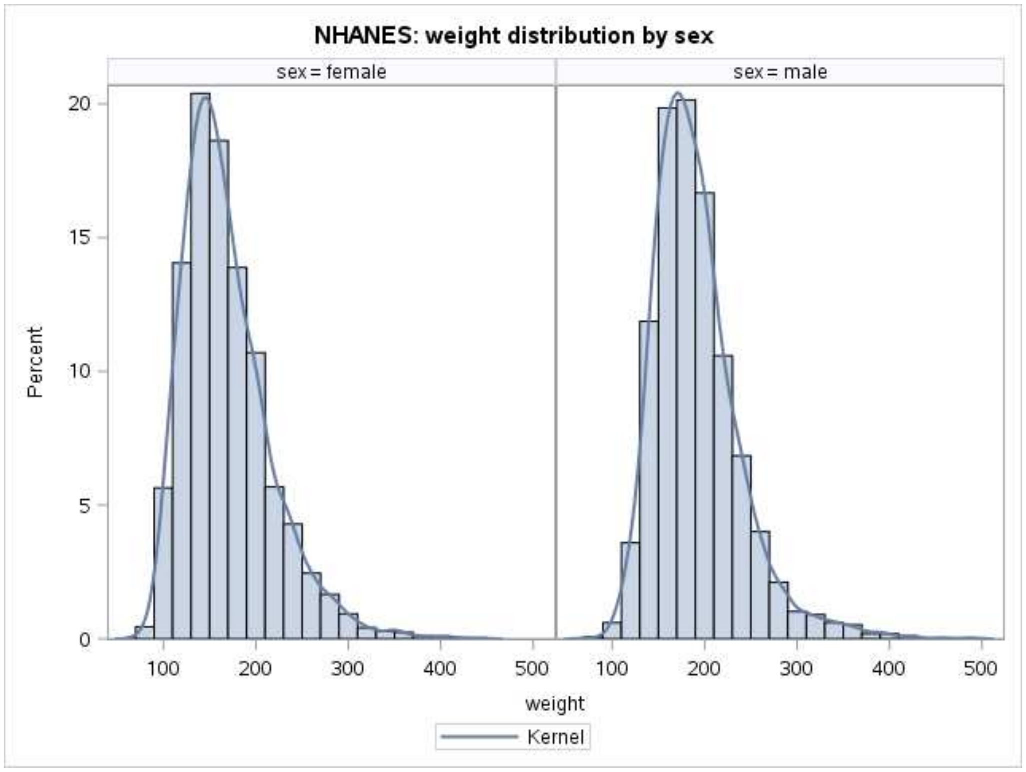
In practice, the observed values of a continuous quantitative variable are discretized, *i.e.,* the values are rounded so that they can be written down. Therefore, there is really no difference between summarizing the distribution of a continuous variable and summarizing the distribution of a discrete 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 2.4 NHANES 2013–2014 adult height distribution histograms.



Example 1.1 NHANES (revisited). We will use some data from the 2013–2014 NHANES to illustrate the basic ideas we are discussing. For this application we will use all of the 5588 adults (age 20 and over) in the 2013–2014 NHANES for whom a height or weight measurement is available. This group forms a random sample from the population of all adults (age 20 and over) in the US at the time of the survey. Histograms for the heights and weights of the adult participants in the 2013–2014 NHANES, grouped by sex, are given in Figures 2.4 and 2.5. A smooth version of each histogram (smooth curve) is also provided. The height distributions are both unimodal and reasonably symmetric. The weight distributions are both unimodal and skewed to the right. We will discuss these distributions in more depth shortly.

Figure 2.5 NHANES 2013–2014 adult weight distribution histograms.



Notice that the data have been grouped into intervals in order to construct these histograms. For the height distributions the intervals are of length one inch. For the weight distributions the intervals are of length 20 pounds. (In the context of histograms these intervals are also know as bins.) For example, in the weight histograms the area of the rectangle centered over 100 is the proportion of the individuals in the group who had a weight between 90 and 110 pounds.

2.5 Numerical summary

toc

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. For example, the female height distribution histogram of Figure 2.4 is more peaked than the male height distribution histogram, which is somewhat flatter or more spread out. This indicates that, for this NHANES data, there is less variability among the heights of the females than there is among the heights of the males. The weight distribution histograms of Figure 2.5 suggest that the variability among the weights of the females is similar to the variability among the weights of the males. 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.

Median computation procedure.
step 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.
step 2. Compute the quantity $n/2$.
step 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.
step 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 2.2 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 2.2 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.

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.

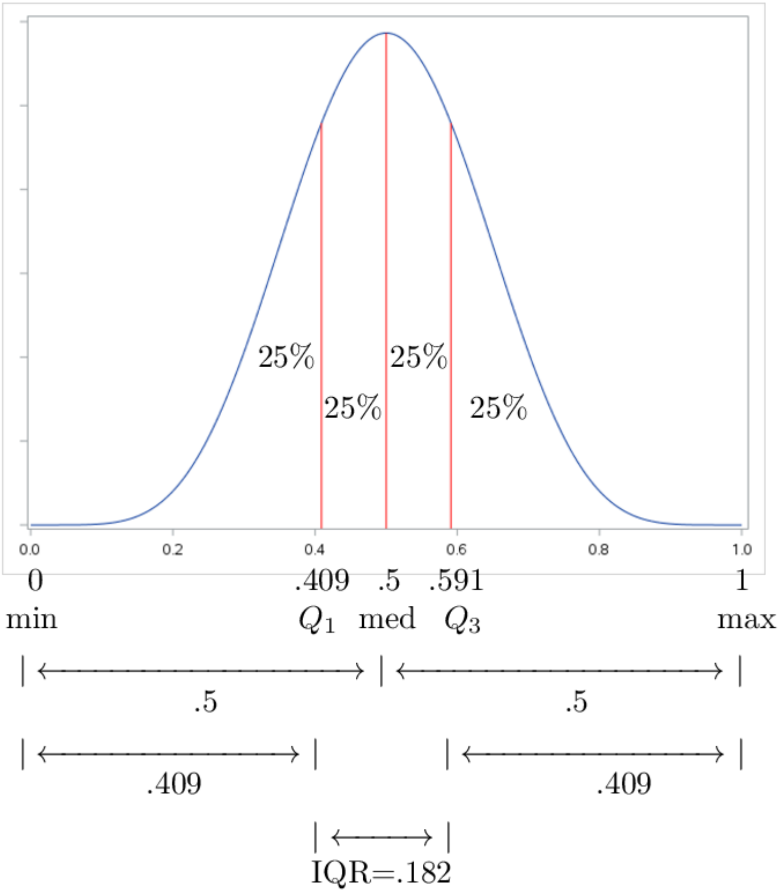
Example 2.1 Weed seeds (revisited). Since $98/4 = 24.5$, the quartiles Q_1 and Q_3 for this example are the observations located at position 25 counting up for Q_1 and counting down for Q_3. Referring to Table 2.2 we find that $Q_1 = 0$ and $Q_3 = 2$. Notice that the range of the lower three–fourths 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.

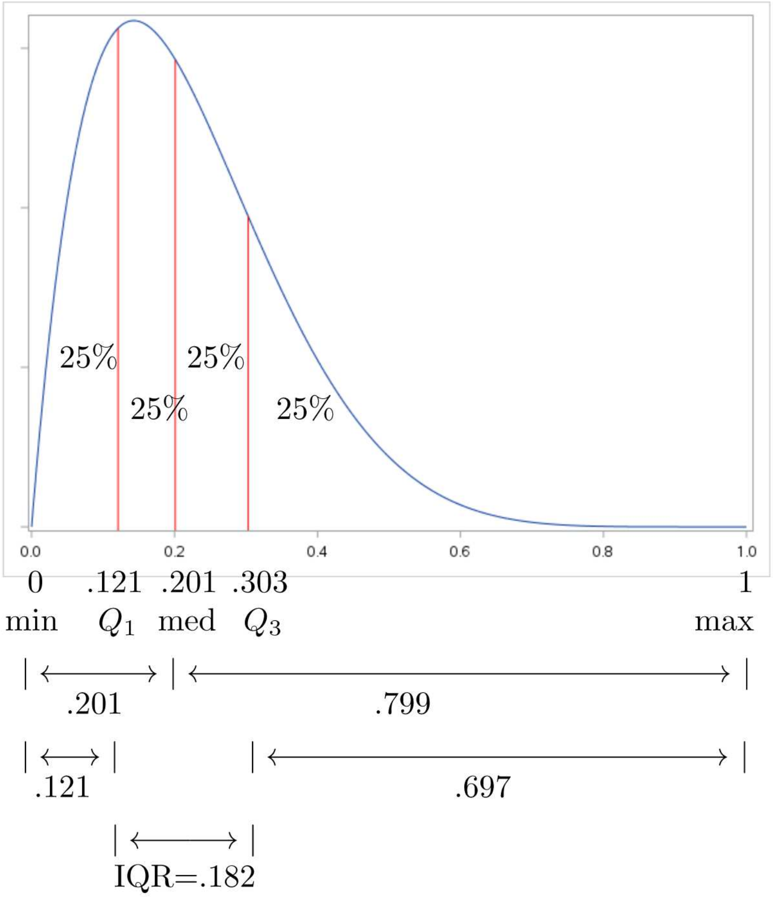
The four distances: (Q_1 – min), (med – Q_1), (Q_3 – med), and (max – Q_3), are the ranges of the first, second, third, and fourth quarters of the distribution, respectively. These distances can be used to quantify the amount of variability in the corresponding parts of the distribution. Comparisons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

Figure 2.6. Mound shaped, single peak, symmetric distribution



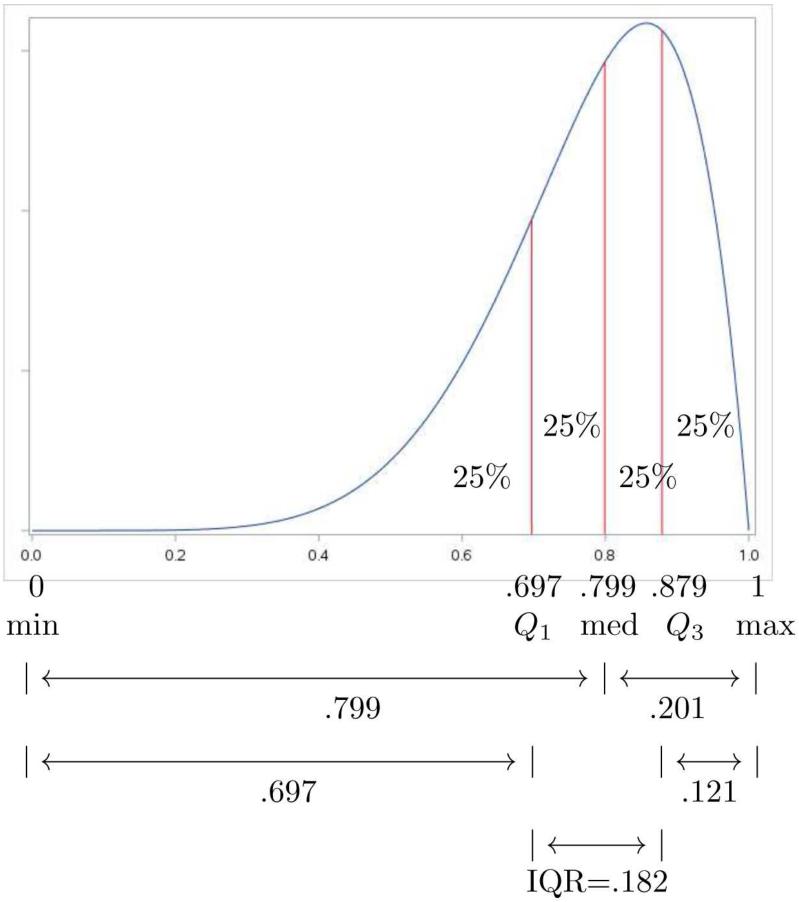
The distribution of Figure 2.6 is mound shaped with a single peak (mode) at .5. This distribution is symmetric. Since this distribution is symmetric we see that the range of the left half of this distribution .5 is equal to the range of the right half; the range of the left tail .409 is equal to the range of the right tail; and, the median .5 is exactly half way between $Q_1 = .409$ and $Q_3 = .591$.

Figure 2.7 Mound shaped, single peak, skewed right distribution



The distribution of Figure 2.7 is mound shaped with a single peak (mode) around .15. This distribution is clearly skewed right. The fact that the range of the right half of this distribution .799 is about 4 times .201 the range of the left half shows extreme skewness to the right. This skewness is mostly due to the fact that the range of the right tail .697 is almost 6 times .121 the range of the left tail. Notice that the middle half of the distribution is reasonably symmetric since the median .201 is about half way between $Q_1 = .121$ and $Q_3 = .303$.

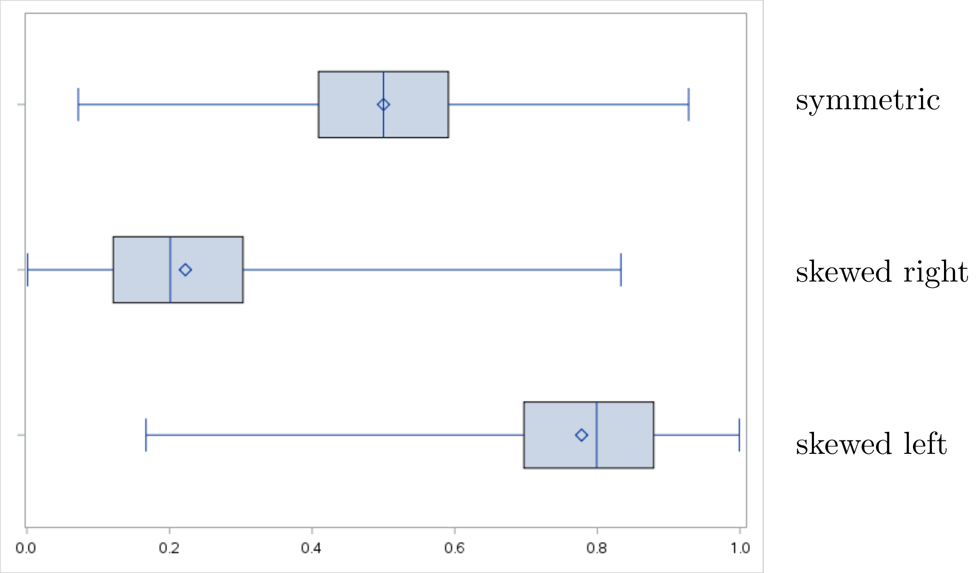
Figure 2.8 Mound shaped, single peak, skewed left distribution



The distribution of Figure 2.8 is mound shaped with a single peak (mode) around .85. This distribution is clearly skewed left. The fact that the range of the left half of this distribution .799 is about 4 times .201 the range of the right half shows extreme skewness to the left. This skewness is mostly due to the fact that the range of the left tail .697 is almost 6 times .121 the range of the right tail. Notice that the middle half of the distribution is reasonably symmetric since the median .799 is about half way between $Q_1 = .697$ and $Q_3 = .879$.

We can use the five number summary values to form a simple graphical representation of a distribution known as a **boxplot** or a box and whiskers plot. A boxplot provides a useful graphical impression of the shape of the distribution as well as its location and variability. Simple boxplots for unimodal mound shaped distributions similar to the distributions of Figures 2.6, 2.7, and 2.8 are provided in Figure 2.9.

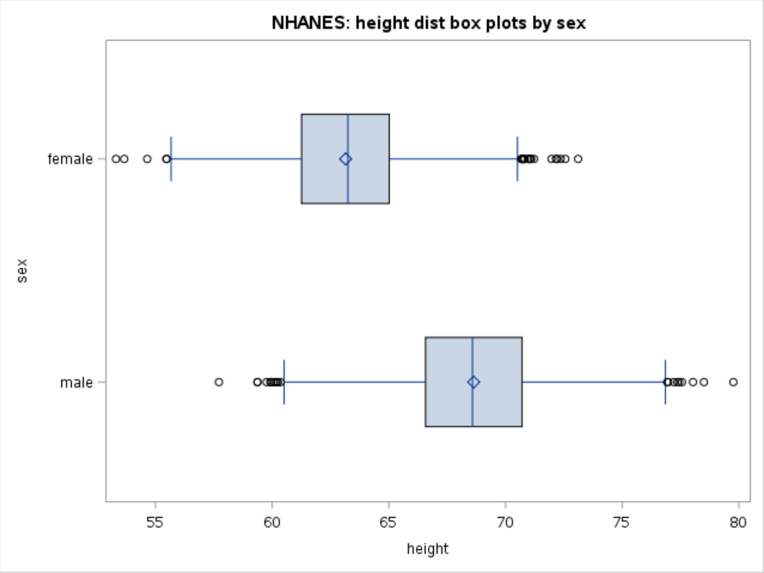
Figure 2.9 Box plots for unimodal mound shaped distributions.



Each boxplot 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.

Example 1.1 NHANES (revisited). Boxplots for the NHANES height and weight distributions are given in Figures 2.10 and 2.11. In these boxplots the whiskers are modified to show extreme observations (indicated by o's in these plots) in the tails of the distribution. Summary information for the NHANES height and weight distributions is given in Tables 2.3 and 2.4.

Figure 2.10 NHANES height distribution boxplots.



reasonably symmetric

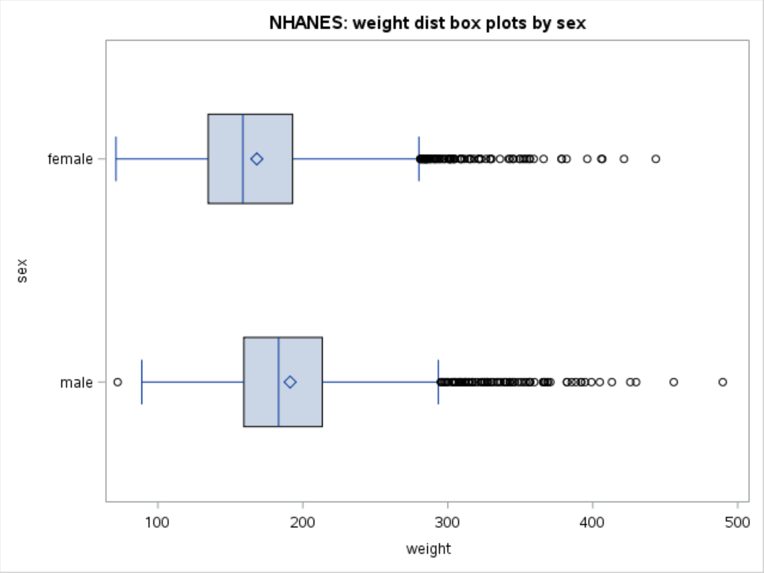
reasonably symmetric

Table 2.3 NHANES height distribution summary information. (height in inches)

	group	
statistic	**females**	**males**
n	2888	2642
location		
mean	63.15	68.63
median	63.25	68.58
variability		
std deviation	2.82	3.06
variance	7.97	9.39
range	19.80	22.05
IQR	3.76	4.13
5 number summary		
min	53.31	57.72
Q_1	61.26	66.57
median	63.25	68.58
Q_3	65.02	70.71
max	73.11	79.76
distances		
Q_1-min	7.95	8.85
med-Q_1	1.99	2.01
Q_3-med	1.77	2.13
max-Q_3	8.09	9.05

As noted earlier, the height distribution histograms for the adult males and the adult females of Figure 2.4 are both unimodal and reasonably symmetric. The boxplots of Figure 2.10 also indicate reasonable symmetry. We will now use the distances from Table 2.3 to quantify this claim of reasonable symmetry. For the adult males we see that the range of the left tail 7.95 is almost the same as 8.09 the range of the right tail. Looking at the middle half of the male height distribution we find that the range of the left side $\text{med} - Q_1 = 1.99$ is only slightly larger than the range of the right side $Q_3 - \text{med} = 1.77$. Similarly, for the adult females the range of the left tail 8.85 is almost the same as 9.05 the range of the right tail; and, $\text{med} - Q_1 = 2.01$ is only slightly smaller than $Q_3 - \text{med} = 2.13$.

Figure 2.11 NHANES weight distribution boxplots.



strongly skewed right

strongly skewed right

Turning to the weight distributions, recall that the weight distribution histograms for the adult males and the adult females of Figure 2.5 are both unimodal and strongly skewed to the right. The boxplots of Figure 2.11 also indicate strong skewness to the right. From each of these boxplots it is clear that in both cases the skewness is due to the extreme variability in the right tail of the distribution, that is, due to high variability among the weights of the heaviest 25% of the group. Note also that in each case the middle half of the distribution is reasonably symmetric. The distances in Table 2.4 readily support these observations. For the male weights $\max - Q_3 = 250.58$ is much larger than $Q_1 - \min = 63.58$, while, relatively speaking, $\text{med} - Q_1 = 23.98$ is only slightly smaller than $Q_3 - \text{med} = 34.32$. Similarly, for the female weights max $- Q_3 = 276.32$ is much larger than $Q_1 -$ min $= 87.12$, while, relatively speaking, med $- Q_1 = 23.98$ is only slightly smaller than $Q_3 -$ med $= 30.14$.

Table 2.4 NHANES weight distribution summary information. (weight in pounds)

	group	
statistic	**females**	**males**
n	2888	2645
location		
mean	168.23	191.21
median	158.62	183.26
variability		
std deviation	47.70	46.80
variance	2275	2190
range	372.46	417.56
IQR	58.30	54.12
5 number summary		
min	71.06	72.16
Q_1	134.64	159.28
median	158.62	183.26
Q_3	192.94	213.40
max	443.52	489.72
distances		
Q_1-min	63.58	87.12
med-Q_1	23.98	23.98
Q_3-med	34.32	30.14
max-Q_3	250.58	276.32

2.6 Quantiles and percentiles in general

toc

We will now provide an extension of the method we used to compute the median and quartiles to allow an arbitrary fraction. Given a proportion p (a fraction between zero and one), the pth **quantile** ($p \times 100th$ **percentile**) of the distribution of X is the value Q_p with the property that if we choose a value of X at random, then X will be less than Q_p with probability p and X will be greater than Q_p with probability $1-p$, *i.e.*, $p \times 100\%$ of the time X will be less than Q_p and $(1-p) \times 100\%$ of the time X will be greater than Q_p. In terms of the histogram of the distribution of X this means that the area in the histogram to the left of Q_p is p and the area to the right of Q_p is $1-p$. Note that the first quartile is the $25th$ percentile, the median is the $50th$ percentile, and the third quartile is the $75th$ percentile.

In terms of a sample of n values, the *pth* quantile is the number Q_p with the property that at least pn of the sample values are less than or equal to Q_p and at least $(1-p)n$ of the sample values are greater than or equal to Q_p. In order to compute a quantile we first need to sort (order) the data. Let $x_1, x_2, \ldots, x_n$ denote the (unordered) data. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the ordered values with $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. These ordered values are know as **order statistics**.

Quantile computation procedure. For $0 < p < 1$, the *pth* quantile Q_p of the n sample values $x_1, \ldots, x_n$ of X is computed as follows. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ denote the sample values ordered from smallest to largest. There are two cases
case 1: If there is an integer k such that $pn = k$, then $Q_p = \frac{x_{(k)}+x_{(k+1)}}{2}$
case 2: If there is an integer k such that $k < pn < k+1$, then $Q_p = x_{(k+1)}$.

2.7 The mean and standard deviation

toc

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).

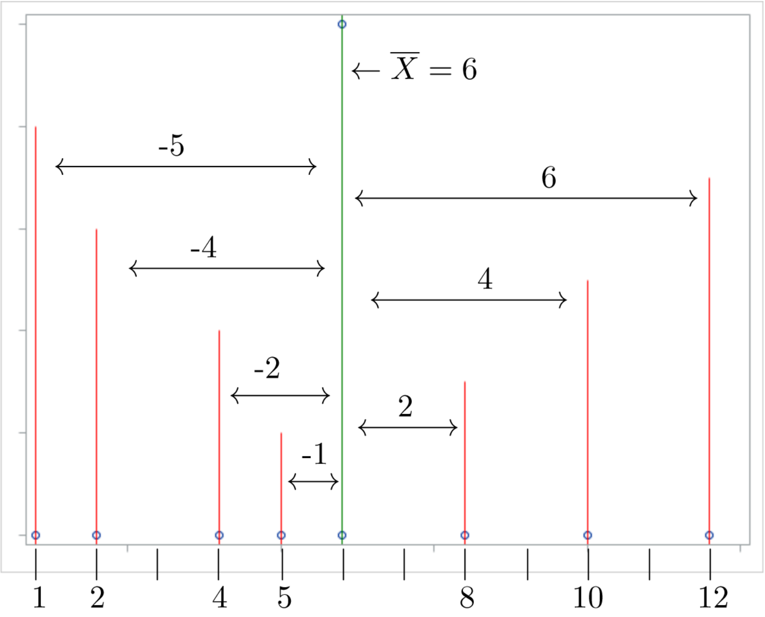
Recall that the median is the number (point on the number line) with the property that the area in the histogram to the left of the median is equal to the area to the right of the median. The mean is the number (point on the number line) where the histogram would balance. To understand what we mean by 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 application 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.

The two measures of variability we discussed earlier, the range and the interquartile range, are distances between two representative values, the minimum and maximum for the range and the first and third quartiles for the interquartile range. We will now discuss a more complex measure of variability which is based on the distances between each of the observations and a single representative value. 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. These deviations from the mean are illustrated, for a small sample of values, in Figure 2.12.

Figure 2.12 An example showing deviations from the mean.
X values: 1, 2, 4, 5, 8, 10, 12; $n = 7$; mean: $\overline{X} = 6$



The most commonly used measure of variability based on these deviation from the mean is the **standard deviation**. The **standard deviation** is the square root of the "average" of the squared deviations of the observed values of the variable from their mean. The formula for the standard deviation given below is not intended for computation purposes; you should use a calculator or a computer to compute the standard deviation. The standard notation for the **sample standard deviation** of the distribution of the variable X is S_X (read this as S sub X). The defining formula 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, as mentioned above, 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. 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 quantity

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

is the **sample variance** of the distribution of the variable X. The calculations corresponding to these formulae are demonstrated below for the sample used in Figure 2.12.

Example 2.3 Computation of the standard deviation. Consider the sample of seven ($n = 7$) X values: 1, 2, 4, 5, 8, 10, 12. The sample mean is

$$\overline{X} = \frac{1+2+4+5+8+10+12}{7} = \frac{42}{7} = 6.$$

The sample variance is

$$\begin{aligned} S_X^2 &= \frac{(1-6)^2 + (2-6)^2 + (4-6)^2 + (5-6)^2 + (8-6)^2 + (10-6)^2 + (12-6)^2}{7-1} \\ &= \frac{(-5)^2 + (-4)^2 + (-2)^2 + (-1)^2 + (2)^2 + (4)^2 + (6)^2}{6} \\ &= \frac{25+16+4+1+4+16+36}{6} = \frac{102}{6} = 17. \end{aligned}$$

The sample standard deviation is

$$S_X = \sqrt{17}.$$

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.

There are quotation marks around the word average in the definition of the sample standard deviation because we divided by $n - 1$ even though there are n squared deviations in the average. When the standard deviation (variance) is computed for the population this divisor is changed to n and a lower case Greek sigma is used instead of an S. That is, the **population standard deviation** is defined as

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

and the **population variance** is σ_X^2.

Example 1.1 NHANES (revisited). We noted earlier that the shapes of the female and male height distributions are very similar and that the shapes of the female and male weight distributions are also very similar. We will now use summary statistics from Tables 2.3 and 2.4 to compare and contrast the locations and the variability in these distributions.

We will first look at the height distributions. As you would expect the males tend to be taller than the females. On average the males are about 5.5 inches taller than the females (male mean height of 68.63 inches versus female mean height of 63.15 inches). Since the height distributions are reasonably symmetric the median heights are very similar to the mean heights. In terms of variability, there is slightly more variability among the heights of the males: male height standard deviation of 3.06 inches verses female height standard deviation of 2.82 inches; male height range of 22.05 inches verses female height range of 19.80 inches; and, male height interquartile range of 4.13 inches verses female height interquartile range of 3.76 inches.

Now we will look at the weight distributions. Again, as you would expect, the males tend to be heavier than the females. On average the males are about 23 pounds heavier than the females (male mean weight of 191.21 pounds versus female mean weight of 168.23 pounds). Since the both weight distributions are strongly skewed right each median weight is smaller than the corresponding mean weight; but, as it turns out, the difference between the median weights is similar to the difference between the mean weights (the median weight difference is about 25 pounds). The weight distributions are quite similar in terms of variability.

2.8 A measure of relative position

toc

Percentiles can be used to quantify the location of a particular value of X relative to a group. Another widely used measure of the relative position of a value within a group is its Z–score. The Z–score of X 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 ($X = \overline{X} - 2S_X$).

Interpretation of a Z–score requires some knowledge of the connection between Z–scores and percentiles. The 68%−95%−99.7% rule given below allows us to associate a percentage with a Z–score. This rule works best for distributions that are unimodal (single peaked), mound shaped, and symmetric. A formal statement of the rule follows.

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.

An aside – Chebyshev's rule

Another connection between Z–scores and percentages is provided by Chebyshev's rule. 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.