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).

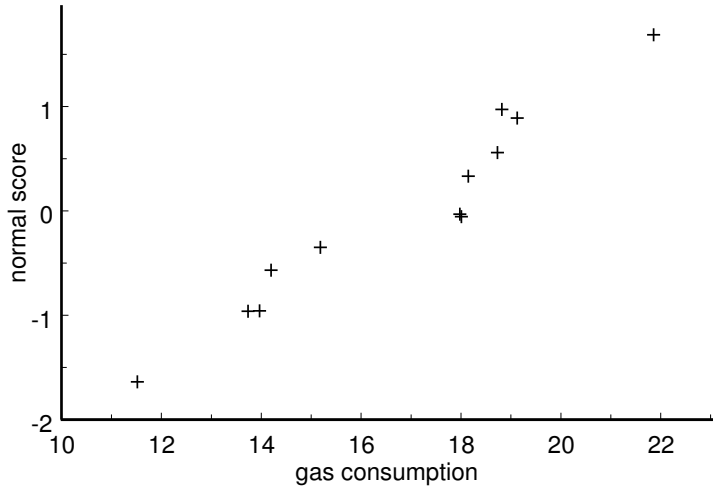
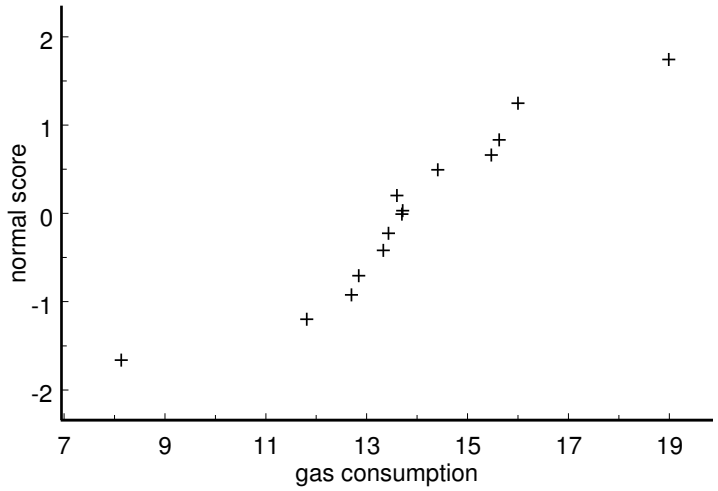


Figure 3. Normal probability plot for gas consumption for houses with extra insulation (both designs).



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.

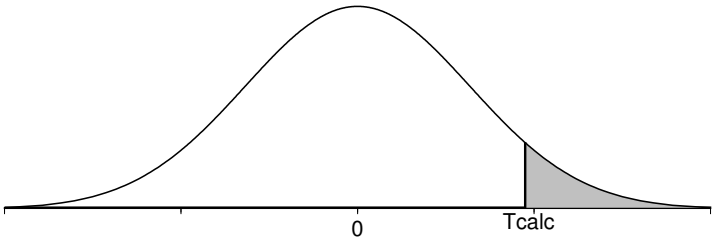
The steps for performing a hypothesis test for

$$H_0 : \mu_1 \leq \mu_2 \text{ versus } H_1 : \mu_1 > \mu_2$$

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.

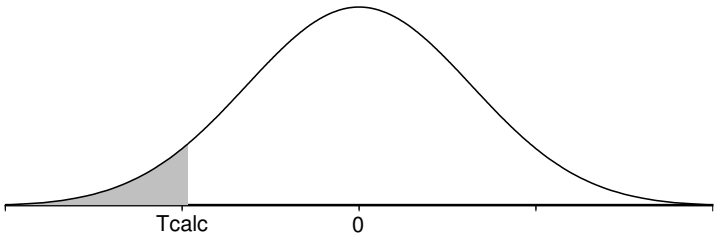
The steps for performing a hypothesis test for

$$H_0 : \mu_1 \geq \mu_2 \text{ versus } H_1 : \mu_1 < \mu_2$$

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).

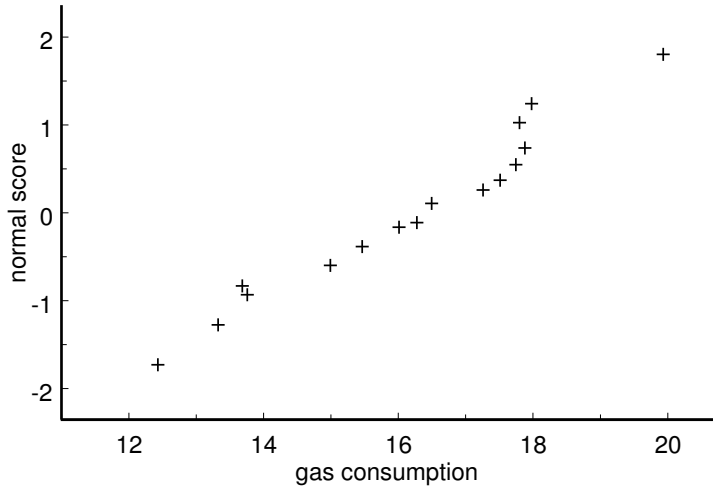
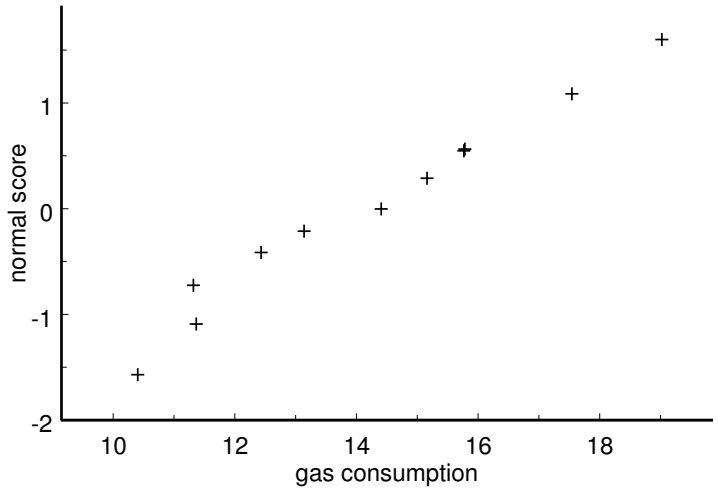


Figure 8. Normal probability plot for gas consumption for houses with extra insulation (passive solar).



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).

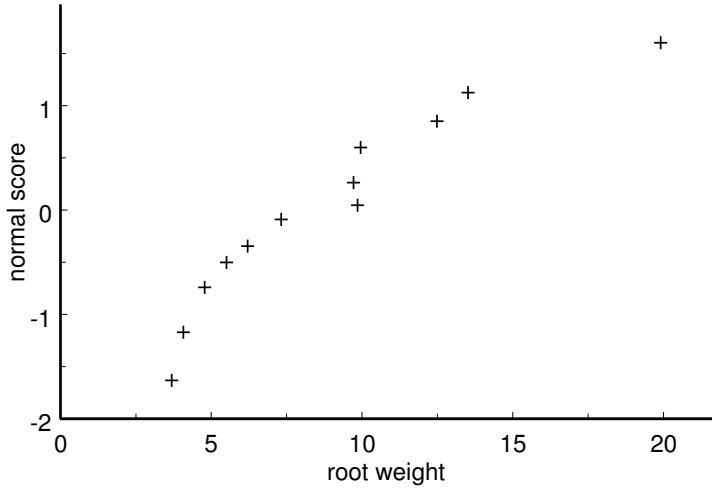
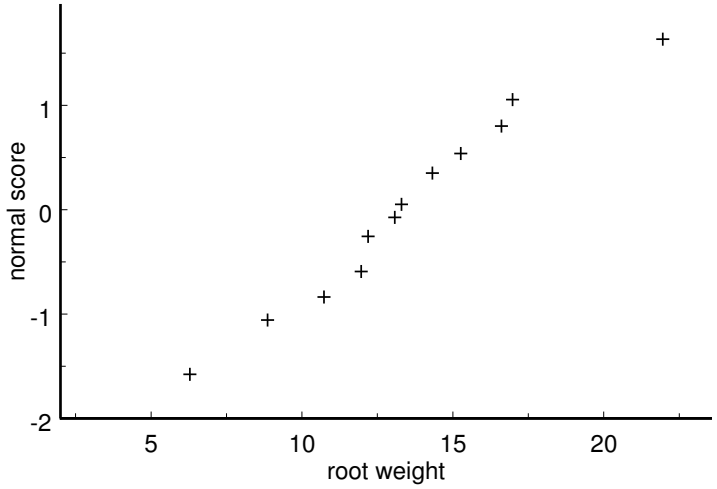


Figure 11. Normal probability plot for root weight (not inoculated).



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 $\mu_1 = \mu_2$.

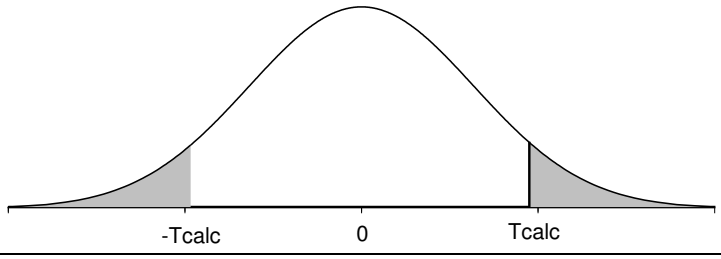
The steps for performing a hypothesis test for

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2$$

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).

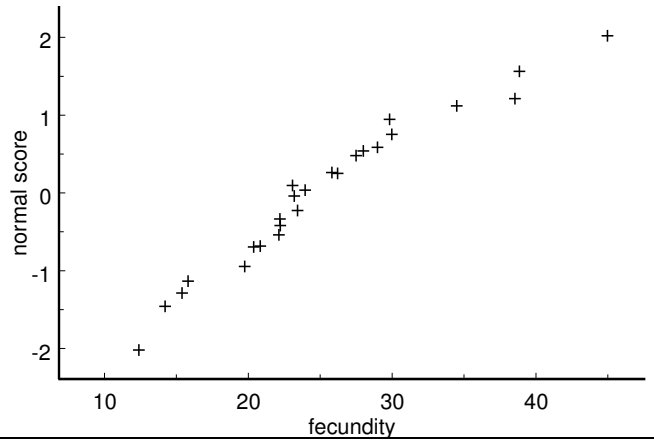


Figure 15. Normal probability plot fruitfly data (susceptible line).

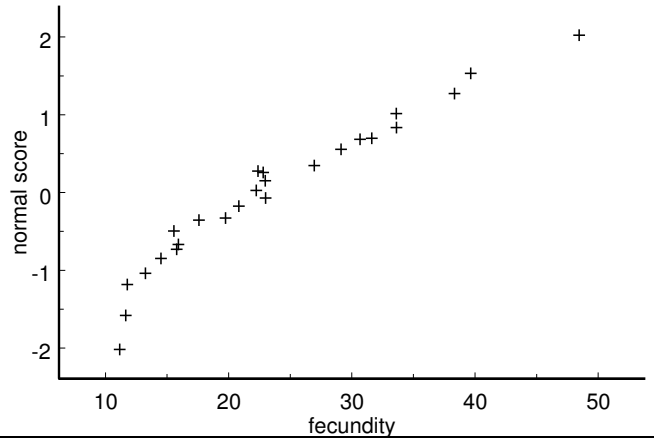
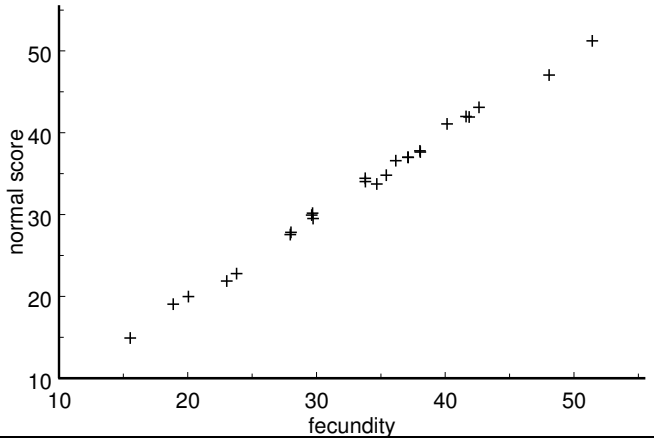


Figure 16. Normal probability plot fruitfly data (nonselected line).



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.