Chapter 11

2-Sample Location Problems

Thus far, in Chapters 9 and 10, we have studied inferences about a single population. In contrast, the present chapter is concerned with comparing *two* populations with respect to a measure of centrality, either the population mean or the population median. We assume the following:

1. $X_1, \ldots, X_{n_1} \sim P_1$ and $Y_1, \ldots, Y_{n_2} \sim P_2$ are continuous random variables. The X_i and the Y_j are mutually independent. In particular, there is no natural pairing of X_1 with Y_1, X_2 with Y_2, etc.

2. P_1 has location parameter θ_1 and P_2 has location parameter θ_2. We assume that comparisons of θ_1 and θ_2 are meaningful. For example, we might compare population means, $\theta_1 = \mu_1 = EX_i$ and $\theta_2 = \mu_2 = EY_j$, or population medians, $\theta_1 = q_2(X_i)$ and $\theta_2 = q_2(Y_j)$, but we would not compare the mean of one population and the median of another population. The *shift parameter*, $\Delta = \theta_1 - \theta_2$, measures the difference in population location.

3. We observe random samples $\vec{x} = \{x_1, \ldots, x_{n_1}\}$ and $\vec{y} = \{y_1, \ldots, y_{n_2}\}$, from which we attempt to draw inferences about Δ. Notice that we do *not* assume that $n_1 = n_2$.

The same four questions that we posed at the beginning of Chapter 10 can be asked here. What distinguishes 2-sample problems from 1-sample problems is the number of populations from which the experimental units were drawn. The prototypical case of a 2-sample problem is the case of a treatment population and a control population. We begin by considering two examples.

Example 11.1 A researcher investigated the effect of Alzheimer's disease (AD) on ability to perform a confrontation naming task. She recruited 60 mildly demented AD patients and 60 normal elderly control subjects. The control subjects resembled the AD patients in that the two groups had comparable mean ages, years of education, and (estimated) IQ scores; however, the control subjects were not individually matched to the AD patients. Each person was administered the Boston Naming Test (BNT), on which higher scores represent better performance. For this experiment:

1. An experimental unit is a person.
2. The experimental units belong to one of two populations: AD patients or normal elderly persons.
3. One measurement (score on BNT) is taken on each experimental unit.
4. Let X_i denote the BNT score for AD patient i. Let Y_j denote the BNT score for control subject j. Then $X_1, \ldots, X_{n_1} \sim P_1$, $Y_1, \ldots, Y_{n_2} \sim P_2$, and we are interested in drawing inferences about $\Delta = \theta_1 - \theta_2$. Notice that $\Delta < 0$ if and only if $\theta_1 < \theta_2$. Thus, to document that AD compromises confrontation naming ability, we might test $H_0 : \Delta \geq 0$ against $H_1 : \Delta < 0$.

Example 11.2 A drug is supposed to lower blood pressure. To determine if it does, $n_1 + n_2$ hypertensive patients are recruited to participate in a *double-blind* study. The patients are randomly assigned to a treatment group of n_1 patients and a control group of n_2 patients. Each patient in the treatment group receives the drug for two months; each patient in the control group receives a *placebo* for the same period. Each patient's blood pressure is measured before and after the two month period, and neither the patient nor the technician know to which group the patient was assigned. For this experiment:

1. An experimental unit is a patient.
2. The experimental units belong to one of two populations: hypertensive patients who receive the drug and hypertensive patients who receive the placebo. Notice that there are two populations despite the fact that all $n_1 + n_2$ patients were initially recruited from a single population. *Different treatment protocols create different populations.*
3. Two measurements (blood pressure before and after treatment) are taken on each experimental unit.

4. Let B_{1i} and A_{1i} denote the before and after blood pressures of patient i in the treatment group. Similarly, let B_{2j} and A_{2j} denote the before and after blood pressures of patient j in the control group. Let $X_i = B_{1i} - A_{1i}$, the decrease in blood pressure for patient i in the treatment group, and let $Y_j = B_{2j} - A_{2j}$, the decrease in blood pressure for patient j in the control group. Then $X_1, \ldots, X_{n_1} \sim P_1$, $Y_1, \ldots, Y_{n_2} \sim P_2$, and we are interested in drawing inferences about $\Delta = \theta_1 - \theta_2$. Notice that $\Delta > 0$ if and only if $\theta_1 > \theta_2$, i.e., if the decrease in blood pressure is greater for the treatment group than for the control group. Thus, a drug company required to produce compelling evidence of the drug's efficacy might test $H_0 : \Delta \leq 0$ against $H_1 : \Delta > 0$.

This chapter describes two important approaches to 2-sample location problems:

11.1 If we assume that the data are normally distributed, then we will be interested in inferences about the difference in population means. We will distinguish three cases, corresponding to what is known about the population variances.

11.2 If we assume only that the data are continuously distributed, then we will be interested in inferences about the difference in population medians. We will assume a *shift model*; i.e., we will assume that P_1 and P_2 differ only with respect to location.

The first approach assumes that both populations are normal, but does not assume that their pdfs have the same shape. The second approach assumes that both pdfs have the same shape, but does not assume normality. The case of nonnormal populations with different shapes is more challenging, beyond the scope of this book. If the populations are symmetric, then one need not assume either normality or a shift model.[1]

11.1 The Normal 2-Sample Location Problem

In this section we assume that

$$P_1 = \text{Normal}\left(\mu_1, \sigma_1^2\right) \quad \text{and} \quad P_2 = \text{Normal}\left(\mu_2, \sigma_2^2\right).$$

[1] M. A. Fligner and G. E. Policello (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association*, 76:162–168.

In describing inferential methods for $\Delta = \mu_1 - \mu_2$, we emphasize connections with material in Chapter 9 and Section 10.1. For example, the natural estimator of a single normal population mean μ is the plug-in estimator $\hat{\mu}$, the sample mean, an unbiased, consistent, asymptotically efficient estimator of μ. In precise analogy, the natural estimator of $\Delta = \mu_1 - \mu_2$, the difference in populations means, is $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{X} - \bar{Y}$, the difference in sample means. Because

$$E\hat{\Delta} = E\bar{X} - E\bar{Y} = \mu_1 - \mu_2 = \Delta,$$

$\hat{\Delta}$ is an unbiased estimator of Δ. It is also consistent and asymptotically efficient.

In Chapter 9 and Section 10.1, hypothesis testing and set estimation for a single population mean were based on knowing the distribution of the standardized natural estimator, a random variable of the form

$$\frac{\begin{array}{c}\text{sample}\\ \text{mean}\end{array} \; - \; \begin{array}{c}\text{hypothesized}\\ \text{mean}\end{array}}{\begin{array}{c}\text{standard deviation}\\ \text{of sample mean}\end{array}}.$$

The denominator of this random variable, often called the *standard error*, was either known or estimated, depending on our knowledge of the population variance σ^2. For σ^2 known, we learned that

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \qquad \left\{ \begin{array}{lll} \sim & \text{Normal}(0,1) & \text{if} \quad X_1, \ldots, X_n \sim \text{Normal}\left(\mu_0, \sigma^2\right) \\ \approx & \text{Normal}(0,1) & \text{if} \quad n \text{ large} \end{array} \right\}.$$

For σ^2 unknown and estimated by S^2, we learned that

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \qquad \left\{ \begin{array}{lll} \sim & t(n-1) & \text{if} \quad X_1, \ldots, X_n \sim \text{Normal}\left(\mu_0, \sigma^2\right) \\ \approx & \text{Normal}(0,1) & \text{if} \quad n \text{ large} \end{array} \right\}.$$

These facts allowed us to construct confidence intervals for and test hypotheses about the population mean. The confidence intervals were of the form

$$\left(\begin{array}{c}\text{sample}\\ \text{mean}\end{array} \right) \; \pm \; q \cdot \left(\begin{array}{c}\text{standard}\\ \text{error}\end{array} \right),$$

where the critical value q is the appropriate quantile of the distribution of Z or T. The tests also were based on Z or T, and the significance probabilities were computed using the corresponding distribution.

The logic for drawing inferences about two populations means is identical to the logic for drawing inferences about one population mean—we simply replace "mean" with "difference in means" and base inferences about Δ on the distribution of

$$\frac{\begin{array}{c}\text{sample} \\ \text{difference}\end{array} - \begin{array}{c}\text{hypothesized} \\ \text{difference}\end{array}}{\begin{array}{c}\text{standard deviation} \\ \text{of sample difference}\end{array}} = \frac{\hat{\Delta} - \Delta_0}{\text{standard error}}.$$

Because $X_i \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y_j \sim \text{Normal}(\mu_2, \sigma_2^2)$,

$$\bar{X} \sim \text{Normal}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \bar{Y} \sim \text{Normal}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Because $\bar{X}$ and $\bar{Y}$ are independent, it follows from Theorem 5.2 that

$$\hat{\Delta} = \bar{X} - \bar{Y} \sim \text{Normal}\left(\Delta = \mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

We now distinguish three cases:

1. Both σ_i are known (and possibly unequal). The inferential theory for this case is easy; unfortunately, population variances are rarely known.

2. The σ_i are unknown, but necessarily equal ($\sigma_1 = \sigma_2 = \sigma$). This case should strike the reader as somewhat implausible. If the population variances are not known, then under what circumstances might we reasonably assume that they are equal? Although such circumstances do exist, the primary importance of this case is that the corresponding theory is elementary. Nevertheless, it is important to study this case because the methods derived from the assumption of an unknown common variance are widely used—and abused.

3. The σ_i are unknown and possibly unequal. This is clearly the case of greatest practical importance, but the corresponding theory is somewhat unsatisfying. The problem of drawing inferences when the population variances are unknown and possibly unequal is sufficiently notorious that it has a name: the *Behrens-Fisher problem*.

11.1.1 Known Variances

If $\Delta = \Delta_0$, then

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \text{Normal}(0, 1).$$

Given $\alpha \in (0, 1)$, let q_z denote the $1 - \alpha/2$ quantile of Normal$(0, 1)$. We construct a $(1 - \alpha)$-level confidence interval for Δ by writing

$$\begin{aligned} 1 - \alpha &= P\left(|Z| < q_z\right) \\ &= P\left(|\hat{\Delta} - \Delta| < q_z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \\ &= P\left(\hat{\Delta} - q_z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \Delta < \hat{\Delta} - q_z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right). \end{aligned}$$

The desired confidence interval is

$$\hat{\Delta} \pm q_z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Example 11.3 For the first population, suppose that we know that the population standard deviation is $\sigma_1 = 5$ and that we observe a sample of size $n_1 = 60$ with sample mean $\bar{x} = 7.6$. For the second population, suppose that we know that the population standard deviation is $\sigma_2 = 2.5$ and that we observe a sample of size $n_2 = 15$ with sample mean $\bar{y} = 5.2$. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_z = \texttt{qnorm}(.975) = 1.959964 \doteq 1.96,$$

then

$$(7.6 - 5.2) \pm 1.96 \sqrt{\frac{5^2}{60} + \frac{2.5^2}{15}} \doteq 2.4 \pm 1.79 = (0.61, 4.21).$$

Example 11.4 For the first population, suppose that we know that the population variance is $\sigma_1^2 = 8$ and that we observe a sample of size $n_1 = 10$ with sample mean $\bar{x} = 9.7$. For the second population, suppose that we know that the population variance is $\sigma_2^2 = 96$ and that we observe a sample of size $n_2 = 5$ with sample mean $\bar{y} = 2.6$. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_z = \texttt{qnorm}(.975) = 1.959964 \doteq 1.96,$$

then

$$(9.7 - 2.6) \pm 1.96\sqrt{\frac{8}{10} + \frac{96}{5}} \doteq 7.1 \pm 8.765 = (-1.665, 15.865).$$

To test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, we exploit the fact that $Z \sim \text{Normal}(0, 1)$ under H_0. Let z denote the observed value of Z. Then a natural level-α test is the test that rejects H_0 if and only if

$$\mathbf{p} = P_{\Delta_0}(|Z| \geq |z|) \leq \alpha,$$

which is equivalent to rejecting H_0 if and only if $|z| \geq q_z$. This test is sometimes called the 2-sample z-test.

Example 11.3 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$z = \frac{(7.6 - 5.2) - 0}{\sqrt{5^2/60 + 2.5^2/15}} \doteq 2.629.$$

Because $|2.629| > 1.96$, we reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|Z| \geq |2.629|) = 2 * \texttt{pnorm}(-2.629) \doteq 0.008562.$$

Example 11.4 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$z = \frac{(9.7 - 2.6) - 0}{\sqrt{8/10 + 96/5}} \doteq 1.5876.$$

Because $|1.5876| < 1.96$, we decline to reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|Z| \geq |1.5876|) = 2 * \texttt{pnorm}(-1.5876) \doteq 0.1124.$$

11.1.2 Unknown Common Variance

Now we assume that $\sigma_1 = \sigma_2 = \sigma$, but that the common variance σ^2 is unknown. Because σ^2 is unknown, we must estimate it. Let

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$$

denote the sample variance for the X_i and let

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2$$

denote the sample variance for the Y_j. If we sampled only the first population, then we would use S_1^2 to estimate the first population variance, σ_1^2. Likewise, if we sampled only the second population, then we would use S_2^2 to estimate the second population variance, σ_2^2. Neither is appropriate in the present situation, as S_1^2 does not use the second sample and S_2^2 does not use the first sample. Therefore, we create a weighted average of the separate sample variances,

$$\begin{aligned} S_P^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{1}{n_1 + n_2 - 2} \left[\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2 \right], \end{aligned}$$

the *pooled sample variance*. Then

$$ES_P^2 = \frac{(n_1 - 1)ES_1^2 + (n_2 - 1)ES_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2}{(n_1 - 1) + (n_2 - 1)} = \sigma^2,$$

so the pooled sample variance is an unbiased estimator of a common population variance. It is also consistent and asymptotically efficient for estimating a common normal variance.

Instead of

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2}},$$

we now rely on

$$T = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_P^2}}.$$

The following result allows us to construct confidence intervals and test hypotheses about the shift parameter $\Delta = \mu_1 - \mu_2$.

Theorem 11.1 *If $\Delta = \Delta_0$, then $T \sim t(n_1 + n_2 - 2)$.*

Given $\alpha \in (0, 1)$, let q_t denote the $1 - \alpha/2$ quantile of $t(n_1 + n_2 - 2)$. Exploiting Theorem 11.1, a $(1 - \alpha)$-level confidence interval for Δ is

$$\hat{\Delta} \pm q_t \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_P^2}.$$

Example 11.3 (continued) Now suppose that, instead of knowing population standard deviations $\sigma_1 = 5$ and $\sigma_2 = 2.5$, we observe sample standard deviations $s_1 = 5$ and $s_2 = 2.5$. The ratio of sample variances, $s_1^2/s_2^2 = 4 \neq 1$, strongly suggests that the population variances are unequal. We proceed under the assumption that $\sigma_1 = \sigma_2$ for the purpose of illustration. The pooled sample variance is

$$S_P^2 = \frac{59 \cdot 5^2 + 14 \cdot 2.5^2}{59 + 14} = 21.40411.$$

To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 73) = 1.992997 \doteq 1.993,$$

then

$$(7.6 - 5.2) \pm 1.993\sqrt{\left(\frac{1}{60} + \frac{1}{15}\right) \cdot 21.40411} \doteq 2.4 \pm 2.66 = (-0.26, 5.06).$$

Example 11.4 (continued) Now suppose that, instead of knowing population variances $\sigma_1^2 = 8$ and $\sigma_2^2 = 96$, we observe sample variances $s_1^2 = 8$ and $s_2^2 = 96$. Again, the ratio of sample variances, $s_2^2/s_1^2 = 12 \neq 1$, strongly suggests that the population variances are unequal. We proceed under the assumption that $\sigma_1 = \sigma_2$ for the purpose of illustration. The pooled sample variance is

$$S_P^2 = \frac{9 \cdot 8 + 4 \cdot 96}{9 + 4} = 35.07692.$$

To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 13) = 2.160369 \doteq 2.16,$$

then

$$(9.7 - 2.6) \pm 2.16\sqrt{\left(\frac{1}{10} + \frac{1}{5}\right) \cdot 35.07692} \doteq 7.1 \pm 7.01 = (0.09, 14.11).$$

To test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, we exploit the fact that $T \sim t(n_1 + n_2 - 2)$ under H_0. Let t denote the observed value of T. Then a natural level-α test is the test that rejects H_0 if and only if

$$\mathbf{p} = P_{\Delta_0}(|T| \geq |t|) \leq \alpha,$$

which is equivalent to rejecting H_0 if and only if $|t| \geq q_t$. This test is called *Student's 2-sample t-test.*

Example 11.3 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t = \frac{(7.6 - 5.2) - 0}{\sqrt{(1/60 + 1/15) \cdot 21.40411}} \doteq 1.797.$$

Because $|1.797| < 1.993$, we decline to reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T| \geq |1.797|) = 2 * \texttt{pt}(-1.797, 73) \doteq 0.0764684.$$

Example 11.4 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t = \frac{(9.7 - 2.6) - 0}{\sqrt{(1/10 + 1/5) \cdot 35.07692}} \doteq 2.19.$$

Because $|2.19| > 2.16$, we reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T| \geq |2.19|) = 2 * \texttt{pt}(-2.19, 13) \doteq 0.04747.$$

11.1.3 Unknown Variances

Now we drop the assumption that $\sigma_1 = \sigma_2$. We must then estimate each population variance separately, σ_1^2 with S_1^2 and σ_2^2 with S_2^2. Instead of

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

we now rely on

$$T_W = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Unfortunately, there is no analogue of Theorem 11.1—the exact distribution of T_W is not known.

The exact distribution of T_W appears to be intractable, but B. L. Welch[2] argued that $T_W \approx t(\nu)$, with

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2/n_2)^2}{n_2 - 1}}.$$

[2]B. L. Welch (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362.

B. L. Welch (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrila*, 34:28–35.

Because σ_1^2 and σ_2^2 are unknown, we estimate ν by

$$\hat{\nu} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

Simulation studies have revealed that the approximation $T_W \approx t(\hat{\nu})$ works well in practice.

Given $\alpha \in (0,1)$, let q_t denote the $1-\alpha/2$ quantile of $t(\hat{\nu})$. Using Welch's approximation, an approximate $(1-\alpha)$-level confidence interval for Δ is

$$\hat{\Delta} \pm q_t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

Example 11.3 (continued) Now we estimate the unknown population variances separately, σ_1^2 by $s_1^2 = 5^2$ and σ_2^2 by $s_2^2 = 2.5^2$. Welch's approximation involves

$$\hat{\nu} = \frac{\left(\frac{5^2}{60} + \frac{2.5^2}{15}\right)^2}{\frac{(5^2/60)^2}{60-1} + \frac{(2.5^2/15)^2}{15-1}} = 45.26027 \doteq 45.26$$

degrees of freedom. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 45.26) \doteq 2.014,$$

then

$$(7.6 - 5.2) \pm 2.014\sqrt{5^2/60 + 2.5^2/15} \doteq 2.4 \pm 1.84 = (0.56, 4.24).$$

Example 11.4 (continued) Now we estimate the unknown population variances separately, σ_1^2 by $s_1^2 = 8$ and σ_2^2 by $s_2^2 = 96$. Welch's approximation involves

$$\hat{\nu} = \frac{\left(\frac{8}{10} + \frac{96}{5}\right)^2}{\frac{(8/10)^2}{10-1} + \frac{(96/5)^2}{5-1}} = 4.336931 \doteq 4.337$$

degrees of freedom. To construct a 0.95-level confidence interval for Δ, we first compute

$$q_t = \texttt{qt}(.975, 4.337) \doteq 2.6934,$$

then

$$(9.7 - 2.6) \pm 2.6934\sqrt{8/10 + 96/5} \doteq 7.1 \pm 13.413 = (-6.313, 20.513).$$

To test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, we exploit the approximation $T_W \approx t(\hat{\nu})$ under H_0. Let t_W denote the observed value of T_W. Then a natural approximate level-α test is the test that rejects H_0 if and only if

$$\mathbf{p} = P_{\Delta_0}(|T_W| \geq |t_W|) \leq \alpha,$$

which is equivalent to rejecting H_0 if and only if $|t_W| \geq q_t$. This test is sometimes called *Welch's approximate t-test*.

Example 11.3 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t_W = \frac{(7.6 - 5.2) - 0}{\sqrt{5^2/60 + 2.5^2/15}} \doteq 2.629.$$

Because $|2.629| > 2.014$, we reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T_W| \geq |2.629|) = 2 * \texttt{pt}(-2.629, 45.26) \doteq 0.011655.$$

Example 11.4 (continued) To test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$, we compute

$$t_W = \frac{(9.7 - 2.6) - 0}{\sqrt{8/10 + 96/5}} \doteq 1.4257.$$

Because $|1.4257| < 2.6934$, we decline to reject H_0 at significance level $\alpha = 0.05$. The significance probability is

$$\mathbf{p} = P_{\Delta_0}(|T_W| \geq |1.4257|) = 2 * \texttt{pt}(-1.4257, 4.337) \doteq 0.2218.$$

Examples 11.3 and 11.4 were carefully constructed to reveal the sensitivity of Student's 2-sample t-test to the assumption of equal population variances. Welch's approximation is good enough that we can use it to benchmark Student's test when variances are unequal. In Example 11.3, Welch's approximate t-test produced a significance probability of $\mathbf{p} \doteq 0.012$, leading us to reject the null hypothesis at $\alpha = 0.05$. Student's 2-sample t-test produced a misleading significance probability of $\mathbf{p} \doteq 0.076$, leading us to commit a Type II error. In Example 11.4, Welch's approximate t-test

produced a significance probability of $\mathbf{p} \doteq 0.222$, leading us to retain the null hypothesis at $\alpha = 0.05$. Student's 2-sample t-test produced a misleading significance probability of $\mathbf{p} \doteq 0.047$, leading us to commit a Type I error.

Evidently, Student's 2-sample t-test (and the corresponding procedure for constructing confidence intervals) should not be used unless one is convinced that the population variances are identical. The consequences of using Student's test when the population variances are unequal may be exacerbated when the sample sizes are unequal. In general:

- If $n_1 = n_2$, then $t = t_W$.
- If the population variances are (approximately) equal, then t and t_W tend to be (approximately) equal.
- If the larger sample is drawn from the population with the larger variance, then t will tend to be less than t_W. All else equal, this means that Student's test will tend to produce significance probabilities that are too large.
- If the larger sample is drawn from the population with the smaller variance, then t will tend to be greater than t_W. All else equal, this means that Student's test will tend to produce significance probabilities that are too small.
- If the population variances are (approximately) equal, then $\hat{\nu}$ will be (approximately) $n_1 + n_2 - 2$.
- It will *always* be the case that $\hat{\nu} \leq n_1 + n_2 - 2$. All else equal, this means that Student's test will tend to produce significance probabilities that are too large.

From these observations we draw the following conclusions:

1. If the population variances are unequal, then Student's 2-sample t-test may produce misleading significance probabilities.
2. If the population variances are equal, then Welch's approximate t-test is approximately equivalent to Student's 2-sample t-test. Thus, if one uses Welch's test in the situation for which Student's test is appropriate, one is not likely to be led astray.

3. *Don't use Student's 2-sample t-test!* I remember how shocked I was when I first heard this advice as a first-year graduate student in a course devoted to the theory of hypothesis testing. The instructor, Erich Lehmann, one of the great statisticians of the 20th century and the author of a famous book on hypothesis testing,[3] told us: "If you get just one thing out of this course, I'd like it to be that you should *never* use Student's 2-sample t-test."

11.2 The Case of a General Shift Family

In the preceding section we assumed that $X_i \sim P_1 = \text{Normal}(\mu_1, \sigma_1^2)$ and $Y_j \sim \text{Normal}(\mu_2, \sigma_2^2)$. Notice that, if $\sigma_1^2 = \sigma_2^2$ and $\Delta = \mu_1 - \mu_2$, then

$$X_i - \Delta \sim \text{Normal}\left(\mu_1 - \Delta, \sigma_1^2\right) = \text{Normal}\left(\mu_2, \sigma_2^2\right) = P_2.$$

In this case, P_1 and P_2 differ only with respect to location. This observation leads us to introduce the concept of a shift family.

Definition 11.1 *A family of distributions, $\mathcal{P} = \{P_\theta : \theta \in \Re\}$, is a* shift family *if and only if*

$$X \sim P_\theta \quad \textit{entails} \quad X - \theta \sim P_0.$$

If P_0 has a pdf, then a shift family is a family in which all of the pdfs have a common shape, differing only with respect to location.

Example 11.5 Fix $\sigma^2 > 0$. The family $\{\text{Normal}(\theta, \sigma^2) : \theta \in \Re\}$ is a shift family because, if $X \sim \text{Normal}(\theta, \sigma^2)$, then $X - \theta \sim \text{Normal}(0, \sigma^2)$.

Section 11.1.2 concerned the case of a normal shift family. Section 11.1.3 retained the assumption of normality, but, by allowing unequal variances, relaxed the assumption of a shift family. The present section does the reverse, relaxing the assumption of normality but retaining the assumption of a shift family. The 2-sample location problem for a general shift family is the problem of drawing inferences about the shift parameter, Δ, in the following statistical model.

[3] E. L. Lehmann (1959). *Testing Statistical Hypotheses.* John Wiley & Sons, New York.

1. Let $\mathcal{P} = \{P_\theta : \theta \in \Re\}$ denote a shift family in which P_0 has a pdf. Without loss of generality, we assume that θ is the population median of P_θ.

2. Assume that $X_1, \ldots, X_{n_1} \sim P_{\theta_1}$ and $Y_1, \ldots, Y_{n_2} \sim P_{\theta_2}$ are mutually independent. Without loss of generality, we assume that $n_1 \leq n_2$.

3. We observe random samples $\vec{x} = \{x_1, \ldots, x_{n_1}\}$ and $\vec{y} = \{y_1, \ldots, y_{n_2}\}$, from which we attempt to draw inferences about $\Delta = \theta_1 - \theta_2$, the difference in population medians.

As in Sections 10.2 and 10.3, we will proceed by first developing a test for $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$, then using it to derive point and set estimators of Δ.

11.2.1 Hypothesis Testing

We begin by noting that it suffices to test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$. Given such a test, we can test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$ by replacing X_i with $X_i' = X_i - \Delta_0$, then testing $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$.

Because the X_i and Y_j are continuous random variables, they assume $N = n_1 + n_2$ distinct values with probability one. Assuming that the observed x_i and y_j are in fact distinct, the *Wilcoxon rank sum test* is the following procedure:

1. Rank the N observed values from smallest to largest.

2. Let T_x denote the sum of the ranks that correspond to the X_i and let T_y denote the sum of the ranks that correspond to the Y_j. Notice that

$$T_x + T_y = \sum_{k=1}^{N} k = N(N+1)/2$$

and that we can write

$$T_x = \sum_{k=1}^{N} kI_k,$$

where $I_1, \ldots, I_N$ are Bernoulli random variables defined by $I_k = 1$ if rank k corresponds to an X_i and $I_k = 0$ if rank k corresponds to a Y_j.

3. Under $H_0 : \Delta = 0$, the X_i and the Y_j have the same distribution. Hence, for each k, the probability that rank k corresponds to an X_i is n_1/N. Because each k is equally likely to contribute to T_x,

the distribution of T_x must be symmetric. Furthermore, noting that $P(I_k = 1) = n_1/N$ and $EI_k = n_1/N$, we see that

$$\begin{aligned} ET_x &= E\left(\sum_{k=1}^{N} kI_k\right) = \sum_{k=1}^{N} kET_k = \frac{n_1}{N}\sum_{k=1}^{N} k \\ &= \frac{n_1}{N}\frac{N(N+1)}{2} = \frac{n_1(N+1)}{2}. \end{aligned}$$

4. Let t_x denote the observed value of T_x. The Wilcoxon rank sum test rejects $H_0 : \Delta = 0$ at significance level α if and only if t_x differs sufficiently from ET_x, i.e., if and only if

$$\mathbf{p} = P_{H_0}\left(|T_x - ET_x| \geq |t_x - ET_x|\right) \leq \alpha.$$

As in Section 10.3.1, the challenge lies in computing $\mathbf{p}$. Again we consider a simple example that illustrates the nature of this challenge.

Example 11.6 We test $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$ with a significance level of $\alpha = 0.15$. We draw $n_1 = 2$ observations from one population and $n_2 = 4$ observations from the other, obtaining samples $\vec{x} = \{9.1, 8.3\}$ and $\vec{y} = \{11.9, 10.0, 10.5, 11.3\}$. The pooled ranks are $\{2, 1\}$ and $\{6, 3, 4, 5\}$, so $t_x = 2 + 1 = 3$. Because $T_x \geq 1 + 2 = 3$, $T_x \leq 5 + 6 = 11$, and $ET_x = n_1(N+1)/2 = 7$, the significance probability is

$$\mathbf{p} = P_{H_0}\left(|T_+ - 7| \geq |3 - 7|\right) = P_{H_0}\left(T_+ = 3\right) + P_{H_0}\left(T_+ = 11\right).$$

To compute $\mathbf{p}$, we require the pmf of the discrete random variable T_x under the null hypothesis that $\Delta = 0$.

Under $H_0 : \Delta = 0$, each of the $\binom{6}{2}$ possible ways of drawing two ranks from $\{1, \ldots, 6\}$ is equally likely to produce the ranks that correspond to $\{x_1, x_2\}$. These possibilities are

	1	1	1	1	1	2	2	2	2	3	3	3	4	4	5
	2	3	4	5	6	3	4	5	6	4	5	6	5	6	6
t_x	3	4	5	6	7	5	6	7	8	7	8	9	9	10	11

and the pmf of T_x is as follows:

k	3	4	5	6	7	8	9	10	11
$15P(T_x = k)$	1	1	2	2	3	2	2	1	1

Thus,

$$\mathbf{p} = P_{H_0}(T_+ = 3) + P_{H_0}(T_+ = 11) = \frac{1}{15} + \frac{1}{15} \doteq 0.1333.$$

Because $\mathbf{p} \leq \alpha$, we reject $H_0 : \Delta = 0$.

As was the case for the Wilcoxon signed rank test, it is cumbersome to compute significance probabilities for the Wilcoxon rank sum test unless n_1 and n_2 are quite small. As in Section 10.3.1, we consider two ways of approximating $\mathbf{p}$:

1. Simulation.

   Using `R`, it is easy to draw subsets of n_1 ranks from $\{1, \ldots, N\}$ and compute the value of T_x for each subset. The observed proportion of subsets for which

   $$|T_x - n_1(N+1)/2| \geq |t_+ - n_1(N+1)/2|$$

   estimates the true significance probability. (The more subsets that we draw, the more accurate our estimate of $\mathbf{p}$.) I wrote an `R` function, `W2.p.sim`, that implements this procedure. This function is described in Appendix R and can be obtained from the web page for this book.

2. Normal Approximation.

   It turns out that, for n_1 and n_2 sufficiently large, the discrete distribution of T_x under $H_0 : \Delta = 0$ can be approximated by a normal distribution.

   Theorem 11.2 *Suppose that $X_1, \ldots, X_{n_1} \sim P_{\theta_1}$ and $Y_1, \ldots, Y_{n_2} \sim P_{\theta_2}$ satisfy the assumptions of a shift model. Let $\Delta = \theta_1 - \theta_2$ and $N = n_1 + n_2$. Let T_x denote the test statistic for the Wilcoxon rank sum test of $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$. Under H_0, $ET_x = n_1(N+1)/4$, $VarT_x = n_1 n_2 (N+1)^2/12$, and*

   $$P_{H_0}(T_x \leq c) \to P\left(Z \leq \frac{c - ET_x}{\sqrt{VarT_x}}\right)$$

   as $n_1, n_2 \to \infty$, where $Z \sim Normal(0, 1)$.

   I wrote an `R` function, `W2.p.norm`, that uses Theorem 11.2 to compute approximate significance probabilities. This function is described in Appendix R and can be obtained from the web page for this book.

Example 11.6 (continued) Five replications of the simulation procedure `W2.p.sim(2,4,3)` resulted in approximate significance probabilities of 0.132, 0.126, 0.128, 0.136, and 0.135. The normal approximation of 0.105, obtained from `W2.p.norm(2,4,3)`, is surprisingly good.

Example 11.7 Suppose that $n_1 = 20$ and $n_2 = 25$ observations produce $t_x = 50$. Two replications of `W2.p.sim(20,25,400,10000)` resulted in approximate significance probabilities of 0.1797 and 0.1774. The normal approximation, obtained from `W2.p.norm(20,25,400)`, is 0.1741.

Finally, we consider the case of ties in the pooled sample. As in Section 10.3.1, we will estimate the average significance probability that results from all of the plausible rankings of the pooled sample. To do so, we simply modify the simulation procedure described above by subjecting each observation to small random perturbations. Each perturbed pooled sample results in a unique ordering. By choosing the magnitude of the perturbations sufficiently small, we can guarantee that each complete ordering generated from a perturbed sample is consistent with the partial ordering derived from the original sample. I wrote an `R` function, `W2.p.ties`, that implements this procedure. This function is described in Appendix R and can be obtained from the web page for this book.

Example 11.8 To test $H_0 : \Delta = 3$ versus $H_1 : \Delta \neq 3$ at $\alpha = 0.05$, the following x_i and y_j were observed:

$\vec{x}$	6.6	14.7	15.7	11.1	7.0	9.0	9.6	8.2	6.8	7.2
$\vec{y}$	4.2	3.6	2.3	2.4	13.4	1.3	2.0	2.9	8.8	3.8

Replacing x_i with $x_i' = x_i - 3$, the pooled sample has the following ranks:

$\vec{x}'$	3.6	11.7	12.7	8.1	4.0	6.0	6.6	5.2	3.8	4.2
r_k	6/7	18	19	16	10	14	15	13	8/9	11/12
$\vec{y}$	4.2	3.6	2.3	2.4	13.4	1.3	2.0	2.9	8.8	3.8
r_k	11/12	6/7	3	4	20	1	2	5	17	8/9

The test statistic, t_x, might be as small as

$$t_x = 6 + 18 + 19 + 16 + 10 + 14 + 15 + 13 + 8 + 11 = 130$$

or as large as

$$t_x = 7 + 18 + 19 + 16 + 10 + 14 + 15 + 13 + 9 + 12 = 133.$$

We might use `W2.p.sim` to estimate, or `W2.p.norm` to approximate, the significance probabilities associated with $t_x = 130$ and $t_+ = 133$. Alternatively, we might use `W2.p.ties` to estimate the average significance probability associated with all of the possible ways of ranking the pooled sample. The latter approach resulted in the following estimated **p**:

```
W2.p.ties(x,y,3,100000)
[1] 0.05281
```

As $\mathbf{p} > \alpha$, the evidence against $H_0 : \Delta = 3$ is not sufficiently compelling to reject the null hypothesis.

11.2.2 Point Estimation

Following the reasoning that we deployed in Sections 10.2.2 and 10.3.2, we estimate Δ by determining the value of Δ_0 for which the Wilcoxon rank sum test is least inclined to reject $H_0 : \Delta = \Delta_0$ in favor of $H_1 : \Delta \neq \Delta_0$. The key to determining this value is representing the Wilcoxon rank sum test in a slightly different form.

The smallest possible value of T_x is $1 + \cdots + n_1 = n_1(n_1 + 1)/2$. This value of T_x is attained if and only if each Y_j exceeds each X_i, i.e.,

$$W_{yx} = \#\{(X_i, Y_j) : Y_j < X_i\} = 0.$$

Pursuing this insight, we write

$$x_{(1)} < \cdots < x_{(n_1)}$$

and let r_k denote the rank of $x_{(k)}$ in the pooled sample. Then

$$r_k = k + \#\left\{Y_j : Y_j < X_{(k)}\right\}$$

and

$$\begin{aligned} T_x &= \sum_{k=1}^{n_1} r_k = \sum_{k=1}^{n_1} k + \sum_{k=1}^{n_1} \#\left\{Y_j : Y_j < X_{(k)}\right\} \\ &= n_1(n_1+1)/2 + \#\{(X_i, Y_j) : Y_j < X_i\} \\ &= n_1(n_1+1)/2 + W_{yx}. \end{aligned}$$

Similarly, reversing the roles of x and y, we have

$$T_y = n_2(n_2+1)/2 + W_{xy},$$

where

$$W_{xy} = \#\left\{(X_i, Y_j) \,:\, X_i < Y_j\right\}.$$

We can use W_{yx} or W_{xy} instead of T_x or T_y, in which form the Wilcoxon rank sum test is called the Mann-Whitney test.

Recall that we test $H_0 : \Delta = \Delta_0$ versus $H_1 : \Delta \neq \Delta_0$ by testing $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$ using $\vec{x} - \Delta_0$ and $\vec{y}$. We will be least inclined to reject the null hypothesis when

$$T_{x-\Delta_0} = ET_{x-\Delta_0} = n_1\left(n_1 + n_2 + 1\right)/2 = n_1\left(n_1 + 1\right)/2 + n_1 n_2/2$$

and (equivalently)

$$T_y = ET_y = n_2\left(n_1 + n_2 + 1\right)/2 = n_2\left(n_2 + 1\right)/2 + n_1 n_2/2.$$

These conditions are equivalent to requiring that

$$W_{y,x-\Delta_0} = n_1 n_2/2 = W_{x-\Delta_0,y},$$

i.e.,

$$\#\left\{(X_i, Y_j) \,:\, Y_j < X_i - \Delta_0\right\} = \#\left\{(X_i, Y_j) \,:\, X_i - \Delta_0 < Y_j\right\},$$

i.e.,

$$\#\left\{(X_i, Y_j) \,:\, X_i - Y_j > \Delta_0\right\} = \#\left\{(X_i, Y_j) \,:\, X_i - Y_j < \Delta_0\right\}.$$

This condition will be satisfied if and only if Δ_0 is the median of the pairwise differences, $X_i - Y_j$. The estimator $\hat{\Delta} = \text{median}(X_i - Y_j)$ is called the Hodges-Lehmann estimator. I wrote an `R` function, `W2.hl`, that computes Hodges-Lehmann estimates. This function is described in Appendix R and can be obtained from the web page for this book.

Example 11.8 (continued) To estimate the shift parameter, we compute the median of the pairwise differences:

```
> W2.hl(x,y)
[1] 5.2
```

11.2.3 Set Estimation

We construct a confidence interval for Δ by determining the set of Δ_0 for which the Wilcoxon rank sum test does not reject $H_0 : \Delta = \Delta_0$ in favor of $H_1 : \Delta \neq \Delta_0$. From Section 11.2.2, it is evident that this set consists of those Δ_0 for which at least k $X_i - Y_j$ are less than Δ_0 and at least k $X_i - Y_j$ are greater than Δ_0. The quantity k is determined by the level of confidence that we desire. As in Sections 10.2.3 and 10.3.3, not all confidence levels are possible.

As in Section 10.2.3, the fact that we must approximate the distribution of the discrete random variable T_x under $H_0 : \Delta = \Delta_0$ complicates our efforts to construct confidence intervals. Again, we proceed in three steps:

1. Use the normal approximation to guess a reasonable value of k, i.e., a value for which

$$\begin{aligned} P\left(W_{yx} \leq k-1\right) &= P\left(T_x \leq k-1+n_1\left(n_1+1\right)/2\right) \\ &= P\left(T_x \leq k-1+ET_x - n_1 n_2/2\right) \\ &\approx \alpha/2. \end{aligned}$$

Recall that

$$\begin{aligned} & P\left(T_x \leq k-1+ET_x - n_1 n_2/2\right) \\ &= P\left(T_x \leq k-0.5+ET_x - n_1 n_2/2\right) \\ &= P\left(\frac{T_x - ET_x}{\sqrt{\operatorname{Var} T_x}} \leq \frac{k-0.5-n_1 n_2/2}{\sqrt{\operatorname{Var} T_x}}\right) \\ &\approx P\left(Z \leq \frac{k-0.5-n_1 n_2/2}{\sqrt{\operatorname{Var} T_x}}\right), \end{aligned}$$

where $Z \sim \text{Normal}(0,1)$; hence, given $\alpha \in (0,1)$, a reasonable value of k is obtained by solving

$$P\left(Z \leq \frac{k-0.5-n_1 n_2/2}{\sqrt{\operatorname{Var} T_+}}\right) = \frac{\alpha}{2},$$

resulting in

$$k = 0.5 + n_1 n_2/2 - q_z\sqrt{\operatorname{Var} T_+} = 0.5 + n_1 n_2/4 - q_z\sqrt{n_1 n_2(N+1)/12},$$

where $q_z = \texttt{qnorm}(1-\alpha/2)$.

2. Use simulation to estimate the confidence coefficients,

$$1 - 2P\left(W_{yx} \leq k - 1\right) = 1 - 2P\left(T_x \leq k - 1 + n_1\left(n_1 + 1\right)/2\right),$$

associated with several reasonable choices of k.

3. Finalize the choice of k (and thereby the confidence coefficient) and construct the corresponding confidence interval. The lower endpoint of the interval is the kth $X_i - Y_j$; the upper endpoint is the $(n_1 n_2 + 1 - k)$th $X_i - Y_j$.

I implemented these steps in the `R` function `W2.ci`, described in Appendix R and available from the web page for this book. This function returns a 5×4 matrix. The first column contains possible choices of k, the second and third columns contain the lower and upper endpoints of the corresponding confidence interval, and the fourth column contains the estimated confidence coefficients.

Example 11.8 (continued) To construct a confidence interval with confidence coefficient $1 - \alpha \approx 0.90$ for Δ, the population shift parameter, we obtain the following results:

```
> W2.ci(x,y,.1,10000)
      k Lower Upper Coverage
[1,] 27   3.2   7.3   0.9234
[2,] 28   3.4   7.2   0.9131
[3,] 29   3.4   7.0   0.8973
[4,] 30   3.6   6.9   0.8805
[5,] 31   3.7   6.9   0.8592
```

The estimated confidence coefficient for $k = 29$ (which happens to be the value of k produced by the normal approximation) nearly equals 0.90, so the desired confidence interval is $(3.4, 7.0)$.

11.3 Case Study: Etruscan versus Italian Head Breadth

In a collection of essays on the origin of the Etruscan empire, N.A. Barnicott and D.R. Brothwell compared measurements on ancient and modern bones.[4]

[4]N.A. Barnicott and D.R. Brothwell (1959). The evaluation of metrical data in the comparison of ancient and modern bones. In *Medical Biology and Etruscan Origins*, edited by G.E.W. Wolstenholme and C.M. O'Connor, Little, Brown & Company, p. 136.

141	148	132	138	154	142	150	146	155	158	150	140
147	148	144	150	149	145	149	158	143	141	144	144
126	140	144	142	141	140	145	135	147	146	141	136
140	146	142	137	148	154	137	139	143	140	131	143
141	149	148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138	142	149
142	137	134	144	146	147	140	142	140	137	152	145
133	138	130	138	134	127	128	138	136	131	126	120
124	132	132	125	139	127	133	136	121	131	125	130
129	125	136	131	132	127	129	132	116	134	125	128
139	132	130	132	128	139	135	133	128	130	130	143
144	137	140	136	135	126	139	131	133	138	133	137
140	130	137	134	130	148	135	138	135	138		

Table 11.1: Maximum breadth (in millimeters) of 84 skulls of Etruscan males (top) and 70 skulls of modern Italian males (bottom).

Measurements of the maximum breadth of 84 Etruscan skulls and 70 modern Italian skulls were subsequently reproduced as Data Set 155 in *A Handbook of Small Data Sets* and are displayed in Table 11.1. We use these data to explore the difference (if any) between Etruscan and modern Italian males with respect to head breadth. In the discussion that follows, x will denote Etruscans and y will denote modern Italians.

We begin by asking if it is reasonable to assume that maximum skull breadth is normally distributed. Normal probability plots of our two samples are displayed in Figure 11.1. The linearity of these plots conveys the distinct impression of normality. Kernel density estimates constructed from the two samples are superimposed in Figure 11.2, created by the following `R` commands:

```
> plot(density(x),type="l",xlim=c(100,180),
+ xlab="Maximum Skull Breadth",
+ main="Kernel Density Estimates")
> lines(density(y),type="l")
```

Not only do the kernel density estimates reinforce our impression of normality, they also suggest that the two populations have comparable variances. (The ratio of sample variances is $s_1^2/s_2^2 = 1.07819$.) The difference

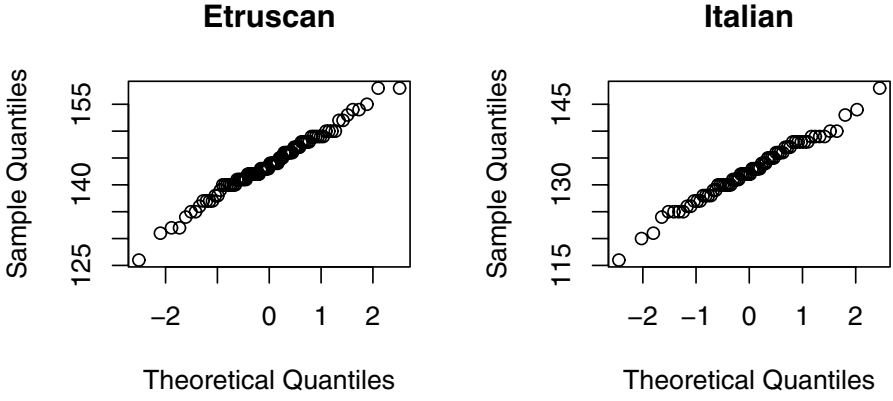

Figure 11.1: Normal probability plots of two samples of maximum skull breadth.

is maximum breadth between Etruscan and modern Italian skulls is nicely summarized by a shift parameter.

Now we construct a probability model. This is a 2-sample location problem in which an experimental unit is a skull. The skulls were drawn from two populations, Etruscan males and modern Italian males, and one measurement (maximum breadth) was made on each experimental unit. Let X_i denote the maximum breadth of Etruscan skull i and let Y_j denote the maximum breadth of Italian skull j. We assume that the X_i and Y_j are independent, with $X_i \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y_j \sim \text{Normal}(\mu_2, \sigma_2^2)$. Notice that, although the sample variances are nearly equal, we do not assume that the population variances are identical. Instead, we will use Welch's approximation to construct an approximate 0.95-level confidence interval for $\Delta = \mu_1 - \mu_2$.

Because the confidence coefficient $1 - \alpha = 0.95$, $\alpha = 0.05$. The desired confidence interval is of the form

$$\hat{\Delta} \pm q\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where q is the $1 - \alpha/2 = 0.975$ quantile of a t distribution with $\hat{\nu}$ degrees of freedom. We can easily compute these quantities in R. To compute $\hat{\Delta}$, the estimated shift parameter:

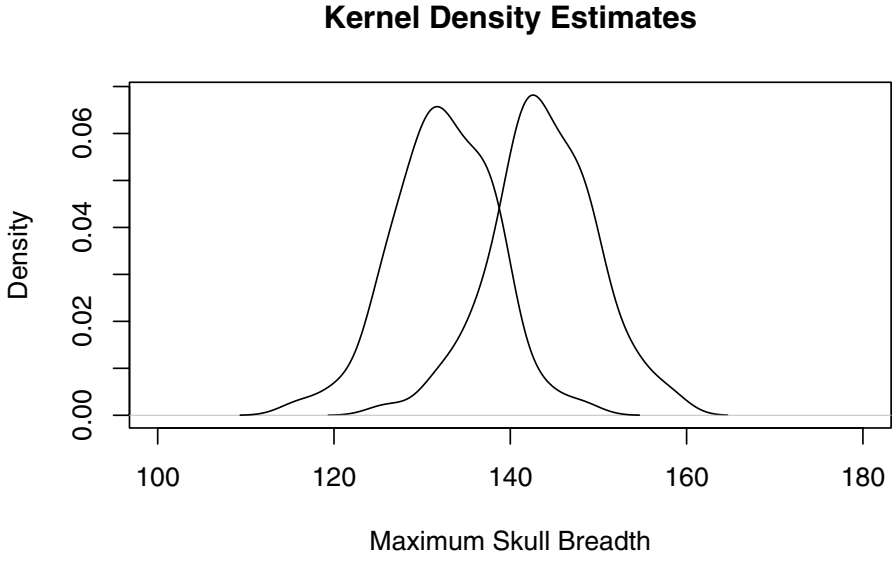

Figure 11.2: Kernel density estimates constructed from two samples of maximum skull breadth. The sample mean for the Etruscan skulls is $\bar{x} \doteq 143.8$; the sample mean for the modern Italian skulls is $\bar{y} \doteq 132.4$.

```
> Delta <- mean(x)-mean(y)
```

To compute the standard error:

```
> n1 <- length(x)
> n2 <- length(y)
> v1 <- var(x)/n1
> v2 <- var(y)/n2
> se <- sqrt(v1+v2)
```

To compute $\hat{\nu}$, the estimated degrees of freedom:

```
> nu <- (v1+v2)^2/(v1^2/(n1-1)+v2^2/(n2-1))
```

To compute q, the desired quantile:

```
> q <- qt(.975,df=nu)
```

Finally, to compute the lower and upper endpoints of the desired confidence interval:

```
> lower <- Delta-q*se
> upper <- Delta+q*se
```

These calculations result in a 0.95-level confidence interval for $\Delta = \mu_1 - \mu_2$ of $(9.459782, 13.20212)$, so that we can be fairly confident that the maximum breadth of Etruscan male skulls is, on average, roughly a centimeter greater than the maximum breadth of modern Italian male skulls.

11.4 Exercises

Problem Set A

1. We have been using various mathematical symbols in our study of 1- and 2-sample location problems. Each of the symbols listed below is used to represent a real number. State which of the following statements applies to each symbol:

   i. The real number represented by this symbol is an unknown population parameter.

   ii. The real number represented by this symbol is calculated from the observed data.

   iii. The real number represented by this symbol is specified by the experimenter.

   Here are the symbols:

   $$\mu \quad \mu_0 \quad \bar{x} \quad s^2 \quad t \quad \alpha \quad \Delta \quad \Delta_0 \quad \mathbf{p} \quad \hat{\nu}$$

2. Assume that $X_1, \ldots, X_{10} \sim \text{Normal}(\mu_1, \sigma_1^2)$ and that $Y_1, \ldots, Y_{20} \sim \text{Normal}(\mu_2, \sigma_2^2)$. None of the population parameters are known. Let $\Delta = \mu_1 - \mu_2$. To test $H_0 : \Delta \geq 0$ versus $H_1 : \Delta < 0$ at significance level $\alpha = 0.05$, we observe samples $\vec{x}$ and $\vec{y}$.

   (a) What test should be used in this situation? If we observe $\vec{x}$ and $\vec{y}$ that result in $\bar{x} = -0.82$, $s_1 = 4.09$, $\bar{y} = 1.39$, and $s_2 = 1.22$, then what is the value of the test statistic?

   (b) If we observe $\vec{x}$ and $\vec{y}$ that result in $s_1 = 4.09$, $s_2 = 1.22$, and a test statistic value of 1.76, then which of the following `R` expressions best approximates the significance probability?

i. `2*pnorm(-1.76)`
ii. `pt(-1.76,df=28)`
iii. `pt(1.76,df=10)`
iv. `pt(-1.76,df=10)`
v. `2*pt(1.76,df=28)`

(c) True of False: if we observe $\vec{x}$ and $\vec{y}$ that result in a significance probability of $\mathbf{p} = 0.96$, then we should reject the null hypothesis.

3. In Section 11.3 we assumed that $P_1 = \text{Normal}(\mu_1, \sigma_1^2)$ and $P_2 = \text{Normal}(\mu_2, \sigma_2^2)$, where P_1 is the distribution of the maximum breadths of Etruscan male skulls and P_2 is the distribution of the maximum breadths of modern Italian male skulls. We then constructed a 0.95-level confidence interval for $\Delta = \mu_1 - \mu_2$, the difference in population means. In this exercise we replace the assumption of normality with the assumption that P_1 and P_2 belong to a shift family.

(a) Which of the following statements is correct? Explain.

i. Assuming that both P_1 and P_2 are normal is *stronger* than assuming that P_1 and P_2 belong to a shift family; i.e., normality is a special case of a shift family.
ii. Assuming that both P_1 and P_2 are normal is *weaker* than assuming that P_1 and P_2 belong to a shift family; i.e., a shift family is a special case of normality.
iii. Neither assumption is stronger than the other, i.e., it is possible for P_1 and P_2 to be normal without belonging to a shift family *and* it is possile for P_1 and P_2 to belong to a shift family without being normal.

(b) Assume that P_1 and P_2 belong to a shift family, with population medians θ_1 and θ_2. Let $\Delta = \theta_1 - \theta_2$. Construct a confidence interval for Δ that has a confidence coefficient of approximately 0.95.

Problem Set B Each of the following scenarios can be modelled as a 1- or 2-sample location problem. For 1-sample problems, let X_i denote the random variables of interest and let $\mu = EX_i$. For 2-sample problems, let X_i and Y_j denote the random variables of interest; let $\mu_1 = EX_i$, $\mu_2 = EY_j$, and $\Delta = \mu_1 - \mu_2$. For each scenario, you should answer/do the following:

(a) What is the experimental unit?

(b) From how many populations were the experimental units drawn? Identify the population(s). How many units were drawn from each population? Is this a 1- or a 2-sample problem?

(c) How many measurements were taken on each experimental unit? Identify them.

(d) Define the parameter(s) of interest for this problem. For 1-sample problems, this should be μ; for 2-sample problems, this should be Δ.

(e) State appropriate null and alternative hypotheses.

Here are the scenarios:

1. A mathematics/education concentrator theorizes that learning mathematics and statistics is sometimes impeded by the widespread use of odd symbols like α, χ, and ω. She reasons that, if her theory is correct, then students who belong to sororities and fraternities—who she presumes are more familiar with Greek letters—should have an easier time learning the mathematical subjects that use such symbols. To investigate, she obtains a list of all William & Mary students who are enrolled in Math 111 (calculus) and a list of all William & Mary students who belong to a sorority or fraternity. She uses this information to choose (at random) 20 calculus students who belong to a sorority or fraternity and 20 calculus students who do not. She persuades each of these students to take a calculus quiz, specially designed to use lots of Greek letters. How might she use the resulting data to test her theory? (Respond to (a)–(e) above.)

2. Umberto theorizes that living with a dog diminishes depression in the elderly, here defined as more than 70 years of age. To investigate his theory, he recruits 15 single elderly men who own dogs and 15 single elderly men who do not own any pets. The Hamilton instrument for measuring depressive tendency is administered to each subject. High scores indicate depression. How might Umberto use the resulting data to test his theory? (Respond to (a)–(e) above. Especially (d).)

3. Irmina, a professional massage/physical therapist and ski instructor, decides to moonlight as an areobics instructor. Her supervisor recommends that she begin each class with 10 minutes of static stretching, but Irmina believes that static stretching is detrimental to athletic

performance. She devises an experiment, for which she recruits 20 aerobics students, that consists of two protocols. In protocol S, a participant walks for 5 minutes, then does 10 minutes of static stretches of the hamstring, quadricep, and calf muscles, then rides a stationary bike for 30 minutes. Protocol D replaces static stretches with dynamic stretches. Each bike is equipped with a heart monitor and the ability to measure watts of power expended. To equalize level of exertion, each participant is asked to maintain a constant training heart rate calculated using the Karvonen formula[5] with an intensity of 0.80. The study participants perform protocol D one week and protocol S the following week. Irmina records the number of watts expended during each 30-minute ride. How might she use the resulting data to persuade her supervisor that dynamic stretching is superior to static stretching? (Respond to (a)–(e) above.)

4. The William & Mary women's tennis team uses championship balls in their matches and less expensive practice balls in their team practices. The players have formed a strong impression that the practice balls do not wear as well as the championship balls, i.e., that the practice balls lose their bounce more quickly than the championship balls. To investigate this perception, Nina and Delphine conceive the following experiment. Before one practice, the team opens new cans of championship balls and practice balls, which they then use for that day's practice. After practice, Nina and Delphine randomly select 10 of the used championship balls and 10 of the used practice balls. They drop each ball from a height of 2 meters and measure the height of its first bounce. How might Nina and Delphine test the team's impression that practice balls do not wear as well as championship balls? (Respond to (a)–(e) above.)

5. A political scientist theorizes that women tend to be more opposed to military intervention than do men. To investigate this theory, he devises an instrument on which a subject responds to several recent U.S. military interventions on a 5-point Likert scale (1="strongly support,"...,5="strongly oppose"). A subject's score on this instrument is the sum of his/her individual responses. The scientist randomly selects 50 married couples in which neither spouse has a registered party affiliation and administers the instrument to each of the 100 individu-

[5]Training Heart Rate = Resting Heart Rate + Intensity $\times$ (220 $-$ Age - Resting Heart Rate).

als so selected. How might he use his results to determine if his theory is correct? (Respond to (a)–(e) above.)

6. A shoe company claims that wearing its racing flats will typically improve one's time in a 10K road race by more than 30 seconds. A running magazine sponsors an event to test this claim. It arranges for 120 runners to enter two road races, held two weeks apart on the same course. For the second race, each of these runners is supplied with the new racing flat. How might the race results be used to determine the validity of the shoe company's claim? (Respond to (a)–(e) above.)

7. Susan theorizes that impregnating wood with an IGR (insect growth regulator) will reduce wood consumption by termites. To investigate this theory, she impregnates 60 wood blocks with a solvent containing the IGR and 60 wood blocks with just the solvent. Each block is weighed, then placed in a separate container with 100 ravenous termites. After two weeks, she removes the blocks and weighs them again to determine how much wood has been consumed. How might Susan use her results to determine if her theory is correct? (Respond to (a)–(e) above.)

8. To investigate the effect of swing dancing on cardiovascular fitness, an exercise physiologist recruits 20 couples enrolled in introductory swing dance classes. Each class meets once a week for ten weeks. Participants are encouraged to go out dancing on at least two additional occasions each week. In general, lower resting pulses are associated with greater cardiovascular fitness. Accordingly, each participant's resting pulse is measured at the beginning and at the end of the ten-week class. How might the resulting data be used to determine if swing dancing improves cardiovascular fitness? (Respond to (a)–(e) above.)

9. It is thought that Alzheimer's disease (AD) impairs short-term memory more than it impairs long-term memory. To test this theory, a psychologist studied 60 mildly demented AD patients and 60 normal elderly control subjects. Each subject was administered a short-term and a long-term memory task. On each task, high scores are better than low scores. How might the psychologist use the resulting task scores to determine if the theory is correct? (Respond to (a)–(e) above.)

10. According to an article in *Newsweek* (May 10, 2004, page 89), recent "studies have shown consistently that women are better than men at

reading and responding to subtle cues about mood and temperament." Some psychologists believe that such differences can be explained in part by biological differences between male and female brains. One such psychologist conducts a study in which day-old babies are shown three human faces and three mechanical objects. The time that the baby stares at each face/object is recorded. Of interest is how much time the baby spends staring at faces versus how much time the baby spends staring at objects. The psychologist's theory predicts that this comparison will differ by sex, with female babies preferring faces to objects to a greater extent than do male babies. How might the psychologist use his results to determine if his theory is correct? (Respond to (a)–(e) above.)

Problem Set C In the early 1960s, the Western Collaborative Group Study investigated the relation between behavior and risk of coronary heart disease in middle-aged men. Type A behavior is characterized by urgency, aggression and ambition; Type B behavior is noncompetitive, more relaxed and less hurried. The following data are the cholesterol measurements of 20 heavy men of each behavior type.[6] We consider whether or not they provide evidence that heavy Type A men have higher cholesterol levels than heavy Type B men.

Cholesterol Levels for Heavy Type A Men									
233	291	312	250	246	197	268	224	239	239
254	276	234	181	248	252	202	218	212	325

Cholesterol Levels for Heavy Type B Men									
344	185	263	246	224	212	188	250	148	169
226	175	242	252	153	183	137	202	194	213

1. Respond to (a)–(e) in Problem Set B.

2. Does it seem reasonable to assume that the samples $\vec{x}$ and $\vec{y}$, the observed values of $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, were drawn from normal distributions? Why or why not?

[6] S. Selvin (1991). *Statistical Analysis of Epidemiological Data*. Oxford University Press, New York, Table 2.1. These data appear as Data Set 47 in *A Handbook of Small Data Sets*. These 40 men were the heaviest in the study. Each weighed at least 225 pounds.

3. Assume that the X_i and the Y_j are normally distributed.

   (a) Test the null hypothesis derived above using Welch's approximate t-test. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

   (b) Construct a (2-sided) confidence interval for Δ with a confidence coefficient of approximately 0.90.

4. Let P_1 denote the distribution of the X_i and let P_2 denote the distribution of the Y_j. Does it seem reasonable to assume that P_1 and P_2 belong to a shift family, i.e., that there exists a real number Δ such that $X_i' = X_i - \Delta \sim P_2$? Why or why not?

5. Assume that P_1 and P_2 do belong to a shift family. Let θ_1 and θ_2 denote the population medians of P_1 and P_2 and let $\Delta = \theta_1 - \theta_2$.

   (a) Modify the null hypothesis derived above by replacing $\Delta = \mu_1 - \mu_2$ with $\Delta = \theta_1 - \theta_2$. Use Wilcoxon's rank sum test to test the modified hypothesis. What is the significance probability? If we adopt a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

   (b) Construct a (2-sided) confidence interval for Δ with a confidence coefficient of approximately 0.90.

Problem Set D Researchers obtained the following measurements of urinary β-thromboglobulin excretion in 12 diabetic patients and 12 normal control subjects.[7]

Normal	4.1	6.3	7.8	8.5	8.9	10.4
	11.5	12.0	13.8	17.6	24.3	37.2
Diabetic	11.5	12.1	16.1	17.8	24.0	28.8
	33.9	40.7	51.3	56.2	61.7	69.2

1. Do these measurements appear to be samples from symmetric distributions? Why or why not?

[7] B. A. van Oost, B. Veldhayzen, A. P. M. Timmermans, and J. J. Sixma (1983). Increased urinary β-thromboglobulin excretion in diabetes assayed with a modified RIA kit-technique. *Thrombosis and Haemostasis*, 9:18–20. These data appear as Data Set 313 in *A Handbook of Small Data Sets*.

2. Both samples of positive real numbers appear to be drawn from distributions that are skewed to the right; i.e., the upper tail of the distribution is longer than the lower tail of the distribution. Often, such distributions can be symmetrized by applying a suitable data transformation. Two popular candidates are:

   (a) The natural logarithm: $u_i = \log(x_i)$ and $v_j = \log(y_j)$.

   (b) The square root: $u_i = \sqrt{x_i}$ and $v_j = \sqrt{y_j}$.

   Investigate the effect of each of these transformations on the above measurements. Do the transformed measurements appear to be samples from symmetric distributions? Which transformation do you prefer?

3. Do the transformed measurements appear to be samples from normal distributions? Why or why not?

4. The researchers claimed that diabetic patients have increased urinary β-thromboglobulin excretion. Assuming that the transformed measurements are samples from normal distributions, how convincing do you find the evidence for their claim?

Problem Set E

1. Chemistry lab partners Arlen and Stuart collaborated on an experiment in which they measured the melting points of 20 specimens of two types of sealing wax. Twelve of the specimens were of one type (A); eight were of the other type (B). Each student then used Welch's approximate t-test to test the null hypothesis of no difference in mean melting point between the two methods:

   - Arlen applied Welch's approximate t-test to the original melting points, which were measured in degrees Fahrenheit.
   - Stuart first converted each melting point to degrees Celsius (by subtracting 32, then multiplying by 5/9), then applied Welch's approximate t-test to the converted melting points.

   Comment on the potential differences between these two analyses. In particular, is it *True* or *False* that (ignoring round-off error) Arlen and Stuart will obtain identical significance probabilities? Please justify your comments.

2. A graduate student in ornithology would like to determine if created marshes differ from natural marshes in their appeal to avian communities. He plans to observe $n_1 = 9$ natural marshes and $n_2 = 9$ created marshes, counting the number of red-winged blackbirds per acre that inhabit each marsh. His thesis committee wants to know how much he thinks he will be able to learn from this experiment.

   Let X_i denote the number of blackbirds per acre in natural marsh i and let Y_j denote the number of blackbirds per acre in created marsh j. In order to respond to his committee, the student makes the simplifying assumptions that $X_i \sim \text{Normal}(\mu_1, \sigma^2)$ and $Y_j \sim \text{Normal}(\mu_2, \sigma^2)$. He estimates that $\text{iqr}(X_i) = \text{iqr}(Y_j) = 10$. Calculate L, the length of the 0.90-level confidence interval for $\Delta = \mu_1 - \mu_2$ that he can expect to construct.

3. A film buff has formed the vague impression that movies tend to be longer than they used to be. Are they really longer? Or do they just *seem* longer? To investigate, he randomly samples U.S. feature films made in 1956 and U.S. feature films made in 1996, obtaining the data displayed in Table 11.2. Do these data provide convincing evidence that 1996 movies are longer than 1956 movies? Compute a significance probability that may be used to encourage or discourage the film buff's impression. Explain how this number should be interpreted. How did you obtain it? Identify and defend any assumptions that you made in your calculations.

Year	Title	Minutes
1956	*Accused of Murder*	74
	Away All Boats	114
	Baby Doll	114
	The Bold and the Brave	87
	Come Next Spring	92
	The Flaming Teen-Age	55
	Gun Girls	67
	Helen of Troy	118
	The Houston Story	79
	Patterns	83
	The Price of Fear	79
	The Revolt of Mamie Stover	92
	Written on the Wind	99
	The Young Guns	87
1996	*$40,000*	70
	Barb Wire	98
	Breathing Room	90
	Daddy's Girl	95
	Ed's Next Move	88
	From Dusk to Dawn	108
	Galgameth	110
	The Glass Cage	96
	Kissing a Dream	91
	Love & Sex etc.	88
	Love is All There Is	120
	Making the Rules	96
	Spirit Lost	90
	Work	90

Table 11.2: Running times of 14 feature films from 1956 and 14 feature films from 1996.