Chapter 19

Comparing Two Numerical Response Populations: Independent Samples

Chapter 19 is very much like Chapter 15. The major—and obvious—difference is that in the earlier chapter the response was a dichotomy, but in this chapter the response is a number. If you revisit the material on the four types of studies in Section 15.2 you can see that the fact that the response was a dichotomy is irrelevant. In other words, everything you learned earlier about how the interpretation of an analysis depends on the type of study remains true in this chapter. In particular, for an observational study, if you conclude that numerical populations differ, then you don’t know—based on the statistical analysis—why they differ. On the other hand, for an experimental study, if you conclude that numerical populations differ, then you may conclude that there is a causal link between the treatment and response.

In addition, the meaning of independent random samples for the different types of studies remains the same in the current chapter. There is even an extension of Simpson’s Paradox for a numerical response, but time limitations will prevent me from covering this topic.

It is also true that Chapter 19 builds on the work of Chapters 17 and 18. In particular, recall that in Chapter 17 you learned that the population for a numerical response is a picture and the kind of picture depends on whether the response is a count or a measurement.

19.1 Notation and Assumptions

The researcher has two populations of interest. The methods of Chapters 17 and 18 may be used to study the populations separately. In this chapter, you will learn how to compare the populations.

- Population 1 has mean $\mu_1$, variance $\sigma_1^2$ and standard deviation $\sigma_1$.
- Population 2 has mean $\mu_2$, variance $\sigma_2^2$ and standard deviation $\sigma_2$.

I realize that specifying both the variance and standard deviation is redundant, but it will prove useful to have both for some of the formulas we develop.

We will consider procedures that compare the populations by comparing their means.
• We assume that we will observe $n_1$ i.i.d. random variables from population 1, denoted by:

$$X_1, X_2, X_3, \ldots, X_{n_1}.$$ 

These will be summarized by their mean $\bar{X}$, variance $S_1^2$ and standard deviation $S_1$. The observed values of these various random variables are denoted by:

$$x_1, x_2, x_3, \ldots, x_{n_1}, \bar{x}, s_1^2$$

• We assume that we will observe $n_2$ i.i.d. random variables from population 2, denoted by:

$$Y_1, Y_2, Y_3, \ldots, Y_{n_2}.$$ 

These will be summarized by their mean $\bar{Y}$, variance $S_2^2$ and standard deviation $S_2$. The observed values of these various random variables are denoted by:

$$y_1, y_2, y_3, \ldots, y_{n_2}, \bar{y}, s_2^2$$

• We assume that the two samples are independent.

I apologize for the cumbersome and confusing notation. In particular, in my $\mu$’s, $\sigma^2$’s, $n$’s $S^2$’s, and so on, I use a subscript to denote the population, either 1 or 2; this is very user-friendly. You need to remember, however, that the random variables, data and some summaries from population 1 are denoted by $x$’s and the corresponding notions from population 2 are denoted by $y$’s. There is a long tradition of doing things this way in introductory Statistics. While it is confusing, its one virtue is that it allows you to avoid double subscripts until you take a more advanced Statistics class.

(Enrichment: Here is the problem with double subscripts—well, other than the obvious problem that they sound, and are, complicated. If I write $x_{123}$ does it mean:

• Observation number 123 from one source of data?

• Observation 23 from population 1? or

• Observation 3 from population 12?

This could be made clear with commas; use $x_{1,23}$ for the second answer above and $x_{12,3}$ for the third answer. The only problem is: In my experience, statisticians and mathematicians don’t want to be bothered with commas!)

The methods introduced in this chapter involve comparing the populations by comparing their means. For tests of hypotheses, this translates to the null hypothesis being $\mu_1 = \mu_2$, or, equivalently, $\mu_1 - \mu_2 = 0$. For estimation, $\mu_1 - \mu_2$ is the feature that will be estimated with confidence.

Our point estimator of $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y}$. There is a Central Limit Theorem for this problem, just as there was in Chapter 17. First, it shows us how to standardize our estimator:

$$W = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}. \quad (19.1)$$

496
Second, it states that we can approximate probabilities for \( W \) by using the \( N(0,1) \) curve and that in the limit, as both sample sizes become larger and larger, the approximations are accurate.

In order to obtain formulas for estimation and testing, we need to eliminate the unknown parameters in the denominator of \( W, \sigma^2_1 \) and \( \sigma^2_2 \). We also will need to decide what to use for our reference curve: the \( N(0,1) \) curve of the Central Limit Theorem and Slutsky or one of the t-curves of Gosset.

Statisticians suggest three methods for handling these two issues, which I refer to as Cases 1, 2 and 3. I won’t actually show you Case 3 because I believe that it nearly worthless to a scientist; I will explain why I feel this way.

We will begin with Case 1; I will follow the popular terminology and call this the large sample approximate method.

### 19.2 Case 1: The Slutsky (Large Sample) Approximate Method

This method comes from Slutsky’s Theorem. In Equation [19.1] for \( W \), replace each population variance by its sample variance. The resultant random variable is:

\[
W_1 = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s^2_1}{n_1}\right) + \left(\frac{s^2_2}{n_2}\right)}}.
\]

(Note that I have placed the subscript of ‘1’ on \( W \) to remind you that this is for Case 1.) It can be shown that in the limit, as both sample sizes grow without bound, the \( N(0,1) \) pdf provides accurate probabilities for \( W_1 \). Thus, for finite values of \( n_1 \) and \( n_2 \), the \( N(0,1) \) pdf will be used to obtain approximate probabilities for \( W_1 \). As a general guideline, I recommend using Case 1 only if \( n_1 \geq 30 \) and \( n_2 \geq 30 \).

The usual algebraic manipulation of the ratio that is \( W_1 \) yields the following result.

**Result 19.1 (Slutsky’s approximate confidence interval estimate of \( (\mu_1 - \mu_2) \).)** With the notation and assumptions given in Section [19.1] Slutsky’s approximate confidence interval estimate of \( (\mu_1 - \mu_2) \) is:

\[
(\bar{x} - \bar{y}) \pm z^* \sqrt{\left(\frac{s^2_1}{n_1}\right) + \left(\frac{s^2_2}{n_2}\right)}.
\]

As always in these intervals, the value of \( z^* \) is determined by the desired confidence level and can be found in Table [12.1] on page [296].

Before I give you an example of the use of Formula [19.3], I will tell you about the test of hypotheses for this section.

As I stated earlier in this chapter, the null hypothesis is \( \mu_1 = \mu_2 \) or, equivalently, \( \mu_1 - \mu_2 = 0 \). There are three options for the alternative:

\[
H_1: \mu_1 > \mu_2; \quad H_1: \mu_1 < \mu_2; \quad \text{or} \quad H_1: \mu_1 \neq \mu_2.
\]

I will abbreviate these as >, < and \( \neq \); no confusion should result provided you remember that \( \mu_1 \) is to the left of the math symbol and \( \mu_2 \) is to its right.

497
In order to obtain the formula for the test statistic, I look at Equation 19.2. I want to know how this random variable behaves if the null hypothesis is true. Well, if the null hypothesis is true, then \( \mu_1 - \mu_2 = 0 \). Make this substitution into Equation 19.2 and we get our test statistic \( Z \), given below.

\[
Z = \frac{(\bar{X} - \bar{Y})}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}.
\]  

(19.4)

Given the assumptions of this section, on the additional assumption that the null hypothesis is true, the sampling distribution of \( Z \) is approximated by the N(0,1) curve. The observed value of the test statistic \( Z \) is given by

\[
z = \frac{(\bar{x} - \bar{y})}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}.
\]  

(19.5)

The rules for finding the P-value are given in the following result.

**Result 19.2** In the formulas below, \( z \) is given in Equation 19.5 and areas are computed under the N(0,1) curve.

1. For the alternative \( \mu_1 > \mu_2 \), the approximate P-value equals the area to the right of \( z \). Equivalently, the approximate P-value equals the area to the left of \( -z \).

2. For the alternative \( \mu_1 < \mu_2 \), the approximate P-value equals the area to the left of \( z \). Equivalently, the approximate P-value equals the area to the right of \( -z \).

3. For the alternative \( \mu_1 \neq \mu_2 \), the approximate P-value equals twice the area to the right of \( |z| \). Equivalently, the approximate P-value equals twice the area to the left of \( -|z| \).

In a later section of this chapter I will discuss the quality of the approximations behind Slutsky’s confidence interval and Result 19.2.

I will end this section with illustrations of the estimation and testing methods of this section with a real data set from a student project. Other examples are given in the Practice and Homework Problems.

Luke performed a completely randomized design with a numerical response. A trial consisted of Luke hitting a pitched baseball. In treatment 1, he used an aluminum bat and in treatment 2 he used a wooden bat. The response is the distance, in feet, that the ball traveled. Luke assigned 40 hits to each treatment, by randomization.

Trials that resulted in Luke missing the ball or hitting a foul-tip were ‘done over.’ This made the randomization a bit trickier (details not given), but I believe that Luke made the correct decision in doing this, for the following reason: The purpose of the study is to compare distances the ball travels to see whether wood or aluminum was superior. I can see no reason to blame the bat’s material for a poor swing by Luke. (As I recall, Luke stated in his report that he had very few of these do-overs.)

In order to analyze Luke’s data, we will assume that the data from each treatment are i.i.d. trials from a population and that the two sets of trials are independent. Luke’s data yielded the following summary statistics:

\[ \bar{x} = 179.6, s_1 = 62.1, n_1 = 40, \bar{y} = 166.2, s_2 = 54.2 \text{ and } n_2 = 40. \]
Luke’s two sample standard deviations are similar in value; note that $62.1/54.2 = 1.146$; i.e., the standard deviation with the aluminum bat is about 15% larger than the standard deviation with the wooden bat. I will return to this issue later in this chapter.

Slutsky’s 95% confidence interval estimate of $(\mu_1 - \mu_2)$ (Formula 19.3) is:

$\left(179.6 - 166.2\right) \pm 1.96\sqrt{\frac{(62.1)^2}{40} + \frac{(54.2)^2}{40}} = 13.4 \pm 1.96(13.03) = 13.4 \pm 25.5 = [-12.1, 38.9].$

Note that I have explicitly written the value of the radical, 13.03, because we will need it soon. In words, based on the confidence interval estimate, Luke’s data are inconclusive. The mean with the aluminum bat is between 12.1 feet smaller and 38.9 feet larger than the mean with the wooden bat.

For his test of hypotheses, Luke chose the alternative $> \because$ because the conventional wisdom in baseball is that a ball hit with an aluminum bat travels farther than a ball hit with a wooden bat. Luke’s observed value of the test statistic, Equation 19.5, is

$$z = \frac{(\bar{x} - \bar{y})}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} = \frac{13.4}{13.03} = 1.028.$$  

With the help of


I find that the area under the N(0,1) curve to the right of $z = 1.028$ equals 0.1520. There is evidence in the data in support of Luke’s one-sided alternative, but Luke’s approximate P-value does not meet the accepted standard for being convincing. Note, as an aside, that the approximate P-value for $<$ is 0.8480 and for $\neq$ is 2(0.1520) = 0.3040.

### 19.3 Case 2: Congruent Normal Populations

In the previous section I gave you Slutsky’s (large sample) method for comparing population means via estimation and testing. The natural follow-up is for me to show you how to compare means for the situation in which either or both of the sample sizes is small. I will do this in the current section.

This section is very mathematical, but not in the sense of having lots of algebra. It is mathematical in the sense that I will be presenting methods for a very specific set of mathematical assumptions. In this section, we will assume that the two populations being compared are congruent Normal pdfs.

According to dictionary.com, congruent means:

Coinciding at all points when superimposed.

This implies that the two populations have identical spreads. For example, the N($\mu_1, \sigma$) and N($\mu_2, \sigma$) curves are congruent for all real numbers $\mu_1$ and $\mu_2$ and all positive real numbers $\sigma$. Thus, there are many pairs of Normal pdfs that satisfy the condition of being congruent; and, of course, many that do not.

Recall the definition of a constant treatment effect, given in Definition 5.1 on page 91.
In a clone-enhanced study, suppose that the response on treatment 1 minus the response on treatment 2 equals the nonzero number \( c \) for every unit. In this situation we say that the treatment has a constant treatment effect equal to \( c \).

I argued in Part I that the constant treatment effect, if present, greatly simplifies the interpretation of statistical analyses. In short, I would say that assuming a constant treatment effect is both helpful and elegant.

Suppose we have an experimental comparative study. This means, recall, that there is one superpopulation of units and the two populations we compare represent what would happen if all of the units were assigned to the same treatment. Suppose that population 2 is the \( \text{N}(20,5) \) pdf. If the constant treatment effect is \( c = 4 \), then population 1 is the \( \text{N}(24,5) \) pdf. In general, if there is a constant treatment effect, then the two populations are congruent. In addition, if one population is a Normal curve, then so is the other.

To reiterate: In this section we will assume that both populations are Normal pdfs, with the added condition that they have the same variance. In our earlier notation, we assume that

\[
\sigma_1^2 = \sigma_2^2.
\]

Because these variances are assumed to be the same, for convenience I will drop the subscripts and write \( \sigma^2 \) for the common value of the population variance and \( \sigma \) for the common value of the population standard deviation.

We begin with the random variable \( W \) in Equation 19.1 on page 496. In this equation, replace the two population variances with their common value \( \sigma^2 \), yielding:

\[
W = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{\sigma^2}{n_1}\right) + \left(\frac{\sigma^2}{n_2}\right)}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)}}.
\]

(19.6)

In the last expression, I have moved \( \sigma^2 \) outside the square root sign, which mathemagically makes the exponent disappear! It can be shown that on the assumption of this subsection—Normal pdfs with common variances—the distribution of \( W \) is given exactly by the \( \text{N}(0,1) \) pdf. (I mention this in passing; we won’t have any use for this fact.)

The remaining issue is the removal of the unknown \( \sigma \) from the formula for \( W \). The proof of the best way to estimate \( \sigma \) is complicated, so I won’t give it. In addition, I am unable to show you a brief motivation of the formula; thus, I will just give you the result.

**Definition 19.1 (The pooled variance.)** With the notation of this chapter, our point estimate of \( \sigma^2 \) is:

\[
s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.
\]

(19.7)

We call \( s_p^2 \) the pooled estimate of the variance \( \sigma^2 \). The idea is that we combine, or pool, two estimates of \( \sigma^2 \) (\( s_1^2 \) and \( s_2^2 \)) to obtain a better estimate.

Note the following about this formula for \( s_p^2 \).
1. Each sample variance appears in the numerator.

2. The coefficient of each sample variance is equal to its degrees of freedom.

3. The sum of the coefficients in the numerator equals the number in the denominator. Thus, $s_p^2$ is referred to as a weighted average (mean) of the two sample variances, with weights given by degrees of freedom.

4. If $n_1 = n_2$, then $s_p^2$ reduces to $(s_1^2 + s_2^2)/2$, a natural combination, which is the unweighted average (mean) of the two sample variances.

Below is the main result.

**Result 19.3** Define the random variable $W_2$ as follows:

$$W_2 = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p \sqrt{(1/n_1) + (1/n_2)}}$$

where $S_p^2$ is the random variable that has observed value $s_p^2$ given in Equation 19.7.

Given the assumptions of this subsection, the exact distribution of $W$ is given by the t-curve with $df = n_1 + n_2 - 2$.

The usual algebraic expansion of the ratio in Equation 19.8 yields a confidence interval estimate of $(\mu_1 - \mu_2)$, given below.

**Result 19.4 (The Gosset confidence interval for $(\mu_1 - \mu_2)$)** The Gosset confidence interval for $(\mu_1 - \mu_2)$ is given by:

$$(\bar{x} - \bar{y}) \pm t^* s_p \sqrt{(1/n_1) + (1/n_2)}.$$  \hspace{1cm} (19.9)

The value of $t^*$ depends on the sample sizes and the desired confidence level, as described below.

1. Select the desired confidence level and write it as a decimal; e.g., 0.95 or 0.99.

2. Subtract the desired confidence level from one and call it the error rate. Divide the error rate by two and subtract the result from one; call the final answer $c$; e.g., 0.95 gives $c = 1 - 0.05/2 = 0.975$ and 0.99 gives $c = 1 - 0.01/2 = 0.995$.

3. Go the website


   Enter the value of $n_1 + n_2 - 2$ in the Degrees of freedom box; enter $c$ in the Cumulative probability box; and click on Calculate. The value $t^*$ will appear in the t score box.

For testing, we use the same hypotheses that we used in Section 19.2, reproduced below for convenience. The null hypothesis is:

$$H_0: \mu_1 = \mu_2.$$  

There are three options for the alternative:
In order to obtain the formula for the test statistic, look at Equation \[19.8\]. We want to know how this random variable behaves if the null hypothesis is true. Well, if the null hypothesis is true, then \(\mu_1 - \mu_2 = 0\). Make this substitution into Equation \[19.8\] and we get our test statistic \(T\), given below.

\[
T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{(1/n_1) + (1/n_2)}},
\]

Given the assumptions of this section (congruent normal populations) and the assumption that the null hypothesis is true, the exact distribution of \(T\) is given by the t-curve with \(df = n_1 + n_2 - 2\). The observed value of the test statistic \(T\) is given by

\[
t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{(1/n_1) + (1/n_2)}}. \tag{19.11}
\]

The rules for finding the P-value are given in the following result.

**Result 19.5** In the rules below, \(t\) is given in Equation \[19.11\] and areas are computed under the t-curve with \(df = n_1 + n_2 - 2\).

1. For the alternative \(\mu_1 > \mu_2\), the P-value equals the area to the right of \(t\). Equivalently, the P-value equals the area to the left of \(-t\).

2. For the alternative \(\mu_1 < \mu_2\), the P-value equals the area to the left of \(t\). Equivalently, the P-value equals the area to the right of \(-t\).

3. For the alternative \(\mu_1 \neq \mu_2\), the P-value equals twice the area to the right of \(|t|\). Equivalently, the P-value equals twice the area to the left of \(-|t|\).

I will illustrate the use of these rules with a student project performed by Sheryl. A trial consisted of Sheryl performing a 1.5 mile sprint on her bicycle. In treatment 1, Sheryl loaded her pannier with 20 pounds and in treatment 2 she removed her pannier from her bike. The response is the time, in seconds, Sheryl required to complete the sprint. Sheryl assigned 5 trials to each treatment by randomization.

In order to analyze Sheryl’s data, we will assume that we have independent i.i.d. trials from two normal populations with a common variance. Sheryl’s data yielded the following summary statistics:

\[
\bar{x} = 383.2, s_1 = 4.38, n_1 = 5, \bar{y} = 355.2, s_2 = 4.87, \text{ and } n_2 = 5.
\]

Note that the ratio

\[
s_2/s_1 = 4.87/4.38 = 1.11,
\]

lends some support to the assumption of equal population variances.

We begin our analysis by computing \(s_p^2\).

\[
s_p^2 = \frac{4(4.38)^2 + 4(4.87)^2}{5 + 5 - 2} = \frac{4(19.18) + 4(23.72)}{8} = 21.45.
\]
Because \( n_1 = n_2 \) we could have computed:

\[
s_p^2 = \frac{(4.38)^2 + (4.87)^2}{2} = 21.45.
\]

In any event, \( s_p = \sqrt{21.45} = 4.63 \).

You may verify that for \( df = 5 + 5 - 2 = 8 \) and 95%, \( t^* = 2.306 \). Thus, the 95% confidence interval estimate of \((\mu_1 - \mu_2)\) is

\[
(383.20 - 355.20) \pm 2.306(4.63)\sqrt{1/5 + 1/5} = 28.00 \pm 2.306(2.928) = 28.00 \pm 6.75 = [21.25, 34.75].
\]

In words, I conclude that Sheryl’s mean time for completing her sprint increased by between 21.25 and 34.75 seconds when the weighted pannier is added to her bike.

I will also perform a test of hypotheses for Sheryl’s data. The obvious choice for the alternative is \( > \) because all would agree that adding weight will slow the bicycle. (Sheryl was not biking down a steep hill!)

The observed value of the test statistic is

\[
t = 28.00/2.928 = 9.563.
\]

With the help of the t-curve website, you can verify that the area under the t-curve with \( df = 8 \) to the right of \( t = 9.563 \) is equal to 0.0000, rounded to the nearest ten-thousandth. Minitab gives a more precise answer: 0.0000059, or slightly more than one in two-hundred thousand. In any event, this is a really small P-value!

I believe that performing a test for Sheryl’s data is a bit, well, dumb. It is obvious that adding weight will slow Sheryl’s biking. If you like to prove the obvious—which admittedly statisticians do quite often—then you will often get really small P-values. By contrast, I do think it is very interesting to estimate how much the weight increases her mean time to complete the sprint.

19.4 Case 3: Normal Populations with Different Spread

Case 2, presented above, is a nice piece of mathematics. Mathematically, it is a pretty general result: the populations need to be normal pdfs, but that’s a big family; and the two population variances need to be the same number. With these restrictions the probabilities—confidence levels and P-values—are exact. Let me digress and explain the thought process of a mathematical statistician. If this sounds too arrogant, let me explain my view of the thought process of a mathematical statistician.

Mathematical results have conditions that are necessary for their proof. In Case 2, the conditions are normal populations and congruence (equal variances). Mathematical statisticians relax a condition and then figure out what will happen mathematically. Case 3 is the result of mathematical statisticians weakening the Case 2 assumptions by dropping the assumption of congruence. In other words, in Case 3, the first population is the \( N(\mu_1, \sigma_1) \) pdf and the second population is the \( N(\mu_2, \sigma_2) \) pdf.

Below are the main features of the Case 3 solution.
1. I won’t use Case 3 in these Course Notes. My reason is given below this list.

2. The Case 3 solution does not give exact probabilities. I don’t see this as a major problem, but I believe that it should be mentioned.

3. Like the Case 2 method, the Case 3 method uses a Gosset’s t-curve as its reference, in this case the Gosset’s t-curve is an approximation.

4. The main computational difficulty with Case 3 is that the formula for the degrees of freedom for the approximating t-curve is very complicated. As a result, if one is restricted to using a hand-held calculator, Case 3 is quite a mess. If, however, one takes advantage of living in the information age, the formula for the degrees of freedom is not an issue. Both Minitab and the multi-purpose vassarstats website allow one to obtain a Case 3 answer without calculating the degrees of freedom by hand.

5. This is a key point. Everybody agrees that, in practice, we don’t need the variances to be exactly equal in order for Case 2 to give useful and approximately exact answers. There is some disagreement on how much they can differ before Case 2 answers become seriously deficient. In my opinion they need to differ a great deal—which, for space limitations, I will leave undefined—before I would discard Case 2 in favor of Case 3.

I will share with you two arguments for why I don’t like Case 3. Bear with me please, because these arguments take some time to explain.

First, after you have finished this chapter, including the Practice Problems and Homework, look again at all the real data examples that I have given you. In every case I report the values of \( s_1 \) and \( s_2 \) and note that these values are reasonably close to each other. This has been my experience with real data. Almost always with real data that I have seen, the values of \( s_1 \) and \( s_2 \) have been similar. Each time this happens, it suggests that for the phenomenon being studied, there is, at most, weak evidence of a major difference between \( \sigma_1 \) and \( \sigma_2 \). Of course, one can imagine or manufacture a situation in which \( \sigma_1 \) and \( \sigma_2 \) clearly differ by a great deal, but in my experience these situations often—though not always—are examples of really stupid science! For example, let population 1 be the heights of male college students and let population 2 be the lengths of newborn male humans. (They can’t stand yet, so we use length, but it’s the same measure as height!) I have no doubt that \( \sigma_1 \) is much larger than \( \sigma_2 \); but, really, who is dumb enough to compare these two populations? Does one really need Statistics to know that college men are taller than newborn males?

This leads me to my second reason. When a researcher decides to compare populations by comparing means, then it is almost always the case that one is trying to find the population with the larger [smaller] mean because, if larger [smaller] responses are preferred, that population will be the better population. Let me introduce you to a hypothetical—and quite fanciful—example of what I mean.

Let’s assume that we all agree that, Life is good: to die at age 50 is better than to die at age 40, and so on. Thus, suppose that, as Nature, you can decide between two possible distributions for the length of all persons’ lives. Your two options are both Normal curves; the first population has
mean $\mu_1 = 70$ years and the second population has mean $\mu_2 = 68$ years. As Nature, you determine which population is better and decide that it will be the distribution for all people. What should you decide? Think about it.

Well, shame on you if you said, “Population 1 because it has a larger mean.” You are not fit for the job of Nature! Your decision is too rash. Why do I say this? Because I have not told you the standard deviations of the two populations!

Now, suppose I told you that population 1 is the $N(70,30)$ pdf and that population 2 is the $N(68,1)$ pdf. Now, as Nature, which would you choose? I will now argue that the only sensible choice is population 2.

Indeed, I believe that population 1 would be horrible. It might even have a catastrophic impact on American society! With population 1, 16% of the people would die before age 40 and 16% would live past 100! (You think Social Security has financial problems now; 2.5% of population 1 would live past the age of 130!) By contrast, with population 2, 95% of the people would die between the ages of 66 and 70. (You have no doubt determined that I am old—64 at the time of this typing. Shouldn’t my selfishness kick in and have me opt for population 1? No, for two reasons. First, Nature must be immortal and I am taking the role of Nature. Second, even though I enjoy being 64 much more than I imagined I would four decades ago, I really can’t imagine that 115 will be loads of fun! Ideally, we all become like the Rutger Hauer character at the end of Blade Runner; http://www.youtube.com/watch?v=a_saUN4j7Gw and not like the character he played in The Hitcher—sorry, there is no appropriate link for this movie!)

The moral above is not restricted to Normal populations. If two populations have wildly different variances, then it might be the case that comparing means is not a good idea! Thus, in my mind, Case 3 solves a problem that is mathematically interesting, but that is not important, and indeed might be misleading, to a scientist. Note that this changes the way I want you to view Case 1. As you recall, Case 1 does not require symmetric congruent populations, only large sample sizes. But if—based on data or theory—you suspect that the two populations have very different spreads, think hard about whether you want to compare populations by comparing means.

### 19.5 Miscellaneous Results

This last section before the Computing section briefly introduces some useful ideas and methods.

#### 19.5.1 Accuracy of Case 2 Confidence Levels

In this subsection I will address my decision not to show you Case 3 for Normal pdfs.

I performed three simulation experiments; scan the results in Table 19.1 and then read my description below of the experiments. For each rep, I had Minitab generate independent random samples of sizes $n_1 = n_2 = 20$. The first population is a $N(\mu, \sigma)$ pdf and the second population is a $N(\mu, k\sigma)$ pdf. In the first simulation, $k = 2$; in the second simulation, $k = 4$; and in the third
Table 19.1: Results from three simulation experiments. Each simulation had 10,000 reps, with a rep consisting independent samples of size $n_1 = n_2 = 20$ from two sequences of i.i.d. trials from a Normal pdf. For each sample, the 95% confidence interval estimate of $\mu_1 - \mu_2$ for Case 2 is computed and Nature classifies it as too small, too large or correct.

<table>
<thead>
<tr>
<th>$\sigma_2 / \sigma_1$</th>
<th>Number of Too Small Intervals</th>
<th>Number of Too Large Intervals</th>
<th>Number of Incorrect Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>230</td>
<td>241</td>
<td>471</td>
</tr>
<tr>
<td>4</td>
<td>278</td>
<td>272</td>
<td>550</td>
</tr>
<tr>
<td>8</td>
<td>277</td>
<td>274</td>
<td>551</td>
</tr>
</tbody>
</table>

simulation, $k = 8$. In words, the two Normal pdfs being compared are not congruent. In fact, the three simulations consider the situation in which the second population’s standard deviation is two, four or eight times larger than the first population’s standard deviation. These simulations are valid for any value of $\mu$ and any positive $\sigma$.

The simulation study shows that the actual confidence levels for Case 2 are equal or close to the nominal confidence levels, even though the assumption of congruence is violated. Indeed, for $k = 8$ the two populations are strongly not congruent, yet the Case 2 intervals perform as advertised. In fairness, I must state that if the study is unbalanced, $n_1 \neq n_2$, Case 2 might not perform as well. Moral: Try for a balanced study if possible.

19.5.2 Slutsky; Skewness

Recall that my guide is to use Slutsky’s method, Case 1, if both sample sizes equal or exceed 30: $n_1 \geq 30$ and $n_2 \geq 30$. If you wonder why there is no mention of population skewness, keep reading.

I want to show you an important connection between Case 1 and Case 2. In Case 1, we replace the denominator of $W$ in Equation 19.1 by:

$$\sqrt{(S_1^2 / n_1) + (S_2^2 / n_2)}.$$

In Case 2, we replace the denominator of $W$ by:

$$S_p\sqrt{(1/n_1) + (1/n_2)}.$$

If the study is balanced, $n_1 = n_2$, then

$$S_p^2 = (S_1^2 + S_2^2) / 2.$$ 

Some simple algebra (well, simple if one enjoys algebra; otherwise, it is tedious) shows that in the case of balance (and, hence, replacing $n_2$ by its equal, $n_1$):

$$\sqrt{(S_1^2 / n_1) + (S_2^2 / n_1)} = S_p\sqrt{(1/n_1) + (1/n_1)}.$$
Thus, in the case of balanced studies, the only difference between Cases 1 and 2 is that the former uses the $N(0,1)$ pdf as its reference curve and the latter uses the t-curve with $df = n_1 + n_2 - 2$. Given my restriction on the use of Case 1, then we are comparing a $N(0,1)$ pdf to a t-curve with at least 58 degrees of freedom; these curves are not very different.

Next, I want to look briefly at the issue of skewed populations. Suppose that both populations 1 and 2 are the log-normal pdf with parameters 5 and 1, pictured in Figure 17.8 on page 449. In Table 17.10 you saw that Gosset’s 95% confidence interval estimate of the mean performed very poorly by having too many incorrect intervals for this pdf and $n \leq 320$.

I will give you the results of one simulation experiment to explore how skewness effects our Case 2 inference. Note that while I don’t want to mislead you, this is only one simulation experiment! We simply do not have time for a more in-depth study of this issue.

Each rep of my simulation experiment generated independent random samples of sizes $n_1 = n_2 = 20$ from two populations, both of which are the log-normal pdf with parameters 5 and 1. Because the populations are identical, $\mu_1 = \mu_2$ and a confidence interval for $\mu_1 - \mu_2$ will be correct if, and only if, it includes zero. Here are my results: 189 of the simulated 95% confidence intervals were too large; and 193 of the simulated 95% confidence intervals were too small. Thus, a total of $189 + 193 = 382$ intervals were incorrect; many fewer than the target of 500. Why did this happen? By taking a difference, $\bar{x} - \bar{y}$, the effect of skewness on a balanced study largely disappears. The too few incorrect intervals is the result of the intervals often being too wide, because the skewness effects the individual standard deviations (remember Figure 17.10 on page 455).

My general recommendation is that for a balanced study, Case 2 gives pretty good answers for populations that are not congruent Normal pdfs. The situation for unbalanced studies is much more complicated and I don’t have time to present it to you. (Sorry.)

19.6 Computing

The vassarstats website that we have used previously is very helpful for this chapter.

19.6.1 Comparison of Means

Please go to:

http://vassarstats.net

The left-side of the page lists a number of options; click on t-Tests & Procedures. This takes you to a new set of options; click on the top one, Two-Sample t-Test for Independent or Correlated Samples. This takes you to a new page. In the Setup section, click on Independent Samples. (If you forget to do this, it’s no problem; Independent Samples is the default.) Next, enter the data, by typing or pasting, and click on Calculate.

The above is getting pretty abstract, so let’s try this out with some real data. I will use Dawn’s data on her cat Bob, which you learned about in Chapter 1. I entered:

1 3 4 5 5 6 6 6 7 8

507
(chicken responses) for Sample A and I entered

\[ 0 \hspace{0.5em} 1 \hspace{0.5em} 1 \hspace{0.5em} 2 \hspace{0.5em} 3 \hspace{0.5em} 3 \hspace{0.5em} 3 \hspace{0.5em} 4 \hspace{0.5em} 5 \hspace{0.5em} 7 \]

(tuna responses) for Sample B and clicked on Calculate. I will explain the output presented by vassarstats:

1. Under Data Summary, we find the sample sizes—both 10—and the means, \( \bar{x} = 5.1 \) and \( \bar{y} = 2.9 \). Sadly, the site gives us neither sample standard deviation, but we could obtain them from the entries for SS. (If you don’t remember how, don’t worry.)

2. Under Results, we find the value of \( \bar{x} - \bar{y} = 2.2 \); the observed value of the test statistic, \( t = 2.4 \), for Case 2; and the P-values for the alternative \( > \) and \( \neq \). I know that the one-tailed P-value is for the alternative that agrees with the data. Note that the P-value for \( < \) is one minus the P-value for \( > \).

3. You may safely ignore the information under F-Test for . . . , because we are not covering this topic.

4. You may safely ignore the information under t-Test Assuming Unequal . . . , because we are not covering this topic. If you can’t resist looking, I will note that this is the Case 3 analysis. Note that the Case 3 analysis is nearly identical to the Case 2 analysis.

5. Finally, the bottom section presents the 95% and 99% confidence intervals for the separate means (you learned about this topic in Chapter 17) as well as the Cases 2 and 3 95% and 99% confidence intervals for \( \mu_1 - \mu_2 \).

19.7 Summary

In this chapter, we consider the problem of comparing two populations with numerical responses. We assume that we have i.i.d. random variables from each population and we assume that the two samples are independent. First, I will consider the problem of estimating the difference of population means, \( \mu_1 - \mu_2 \).

Case 1 is Slutsky’s (large sample) approximate method. The confidence interval estimate of \( \mu_1 - \mu_2 \) is given in Formula [19.3], which is reproduced below:

\[
(x - y) \pm z^* \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}.
\]

My advice is that this formula works well provided \( n_1 \geq 30 \) and \( n_2 \geq 30 \). In theory, Slutsky’s confidence interval makes no assumptions about the two populations being compared, but note my remarks in this chapter on the issue of populations with extremely different spreads.

Case 2 assumes that the two populations are congruent Normal pdfs. This case yields Gosset’s confidence interval estimate of \( \mu_1 - \mu_2 \), given in Formula [19.9] and reproduced below:

\[
(x - y) \pm t^* s_p \sqrt{(1/n_1) + (1/n_2)}.
\]
Recall that $s^2_p$ is defined in Equation 19.7 on page 500. If the Case 2 assumptions are true, then the confidence level of this interval is exact. Otherwise, this formula works well for populations that are not Normal curves unless the population variances are very different.

Case 3 is Case 2 without the assumption that the Normal curves are congruent. You are not responsible for this case; indeed, I don’t even show it to you! I argue why this case is rarely useful at best, and potentially misleading at worst.

Next, I will talk about tests of hypotheses for comparing the population means. The null hypothesis is $\mu_1 = \mu_2$, and there are three options for the alternative:

\[ H_1: \mu_1 > \mu_2; \quad H_1: \mu_1 < \mu_2; \quad \text{or} \quad H_1: \mu_1 \neq \mu_2. \]

For Case 1, the observed value of the test statistic is given in Equation 19.5, and is reproduced below:

\[ z = \frac{(\bar{x} - \bar{y})}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}. \]

The rules for using $z$ to find the approximate $P$-value is given in Result 19.2.

For Case 2, the observed value of the test statistic is given in Equation 19.11, and is reproduced below:

\[ t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{(1/n_1) + (1/n_2)}}. \]

The rules for using $t$ to find the $P$-value is given in Result 19.5.

### 19.8 Practice Problems

1. Earlier in these notes I told you about my friends Bert and Walt playing mahjong. I mentioned that the version Bert plays is easier than the version Walt plays. In particular, with Bert’s version the beginning arrangement of tiles is selected at random from arrangements for which it is possible to win. By contrast, Walt’s version begins with a random arrangement of tiles. (It is indisputable that for many arrangements, winning is impossible.) Thus, it is no surprise that Bert has a higher probability of winning than Walt. I want, however, to explore a different question: When they both lose, who does better, Bert or Walt?

Let population 1 denote Bert’s score and let population 2 denote Walt’s score. In Chapter 17, I gave you the following summary statistics:

\[ \bar{x} = 23.93, s_1 = 11.33, n_1 = 71, \bar{y} = 23.05, s_2 = 10.74 \text{ and } n_2 = 216. \]

Making our usual assumption of independent samples from two sequences of i.i.d. trials, perform the following analyses.

(a) Compare the values—by taking a ratio—of the two sample standard deviations. Comment.

(b) Calculate Slutsky’s approximate 95% confidence interval estimate of $\mu_1 - \mu_2$. Comment.
(c) Obtain Slutsky’s approximate P-value for the alternative $\neq$. Comment.

2. I presented Reggie’s study of darts in the Chapter 1 Homework on page 25. Summary statistics for Reggie’s data are below:

$$\bar{x} = 201.53, s_1 = 11.199, \bar{y} = 188.00, s_2 = 15.104 \text{ and } n_1 = n_2 = 15.$$  

Make the usual assumptions of this chapter to analyze Reggie’s data.

(a) Compare the values—by taking a ratio—of the two sample standard deviations. Comment.

(b) Calculate the values of $s_p^2$ and $s_p$.

(c) Calculate the Case 2 (Gosset’s) 95% confidence interval estimate of $\mu_1 - \mu_2$. Comment.

(d) Obtain the Case 2 P-value for the alternative $>$. Comment.

3. In this chapter, I showed you that if two the populations are identical and log-normal with parameters 5 and 1, then Gosset’s confidence interval works reasonably well for $n_1 = n_2 = 20$. This example shows that if the populations are skewed and different, then Gosset might not work so well. In particular, I let population 1 be the exponential pdf with rate equal to 0.1 (mean equal to 10) and I let population 2 be the exponential pdf with rate equal to 0.2 (mean equal to 5). Thus, the true value of $\mu_1 - \mu_2$ is $10 - 5 = 5$. Also, in addition to the two populations being strongly skewed, they have different variances: $\sigma_1^2 = 100$ and $\sigma_2^2 = 25$.

I performed a simulation experiment with 10,000 reps. Each rep consisted of:

- Selecting a random sample of size $n_1 = 20$ from population 1.
- Selecting a random sample of size $n_2 = 20$ from population 2.
- The two samples are independent.
- Gosset’s Case 2 95% confidence interval estimate of $\mu_1 - \mu_2$ is obtained.
- Nature (well, me) determines whether the interval estimate is too large, too small or correct.

I obtained the following results: 107 intervals were too large; and 528 intervals were too small. Comment on these results.

19.9 Solutions to Practice Problems

1. (a) The ratio of the larger to the smaller is

$$\frac{11.33}{10.74} = 1.055.$$  

The sample standard deviations are nearly identical.
(b) The confidence interval is:

\[
(23.93 - 23.05) \pm 1.96 \sqrt{(11.33)^2/71 + (10.74)^2/216} = 0.88 ± 1.96(1.5304) = 0.88 ± 3.00 = [-2.12, 3.88].
\]

The interval is inconclusive and quite wide compared to the point estimate of the difference of means.

(c) The observed value of the test statistic is

\[
z = \frac{0.88}{1.5304} = 0.575.
\]

The area under the N(0,1) pdf to the right of 0.575 is 0.2826. Thus, the approximate P-value for the alternative \( \neq \) is 2(0.2826) = 0.5652. The evidence in support of the alternative is weak.

2. (a) The ratio of the larger to the smaller is

\[
15.104/11.199 = 1.349.
\]

This is a the largest ratio we have seen, but it is still quite small.

(b) Because the study is balanced,

\[
s_p^2 = \frac{(11.199)^2 + (15.104)^2}{2} = \frac{353.548}{2} = 176.774.
\]

Thus, \( s_p = \sqrt{176.774} = 13.296. \)

(c) You can verify that with \( df = 15 + 15 - 2 = 28, t^* = 2.048. \) Thus, Gosset’s 95% confidence interval estimate is

\[
(201.53 - 188.00) ± 2.048(13.296)\sqrt{2/15} = 13.53 ± 2.048(4.855) = 13.53 ± 9.94 = [4.59, 23.47].
\]

This interval indicates that Reggie’s population mean score from 10 feet is between 4.59 and 23.47 points larger than his population mean score from 12 feet.

(d) The observed value of the test statistic is

\[
t = \frac{13.53}{4.855} = 2.787.
\]

The area under the t-curve with \( df = 28 \) to the right of 2.787 is 0.0047; this is the P-value for the alternative >. It is exact if the populations are Normal pdfs; otherwise, it is approximate.

The evidence in support of the alternative is strong.

3. There are two disappointing results. First, the number of incorrect intervals is 107 + 528 = 635 is clearly larger than the target value of 500. Not horribly larger, but quite a bit. Second, I am always disappointed when one type of incorrect interval greatly outnumbers the other type. Following the ideas from one population inference, I would not use Case 2 for a one-sided alternative for this situation.
19.10 Homework Problems

1. Recall Sara’s study of golf, introduced in Chapter 2. Let population 1 denote the distance, in yards, Sara hit the ball with the 3-Wood and let population 2 denote the distance, in yards, Sara hit the ball with the 3-Iron. I presented the following summary statistics in Chapter 2:

\[
\bar{x} = 106.875, \ s_1 = 29.87, \ n_1 = 40, \ \bar{y} = 98.175, \ s_2 = 28.33 \text{ and } n_2 = 40.
\]

Making our usual assumption of independent samples from two sequences of i.i.d. trials, perform the following analyses.

(a) Compare the values—by taking the ratio—of the two sample standard deviations. Comment.

(b) Calculate Gosset’s approximate 95% confidence interval estimate of \( \mu_1 - \mu_2 \). Comment.

(c) Obtain Gosset’s approximate P-value for the alternative \( \neq \). Comment.

2. I introduced you to Dawn’s study of her cat Bob in Chapter 1. Below are summary statistics for Dawn’s data:

\[
\bar{x} = 5.1, \ s_1 = 2.025, \ \bar{y} = 2.9, \ s_2 = 2.079 \text{ and } n_1 = n_2 = 10.
\]

Make the usual assumptions of this chapter to analyze Dawn’s data.

(a) Compare the values—by taking a ratio—of the two sample standard deviations. Comment.

(b) Calculate the values of \( s_p^2 \) and \( s_p \).

(c) Calculate the Case 2 (Gosset’s) 95% confidence interval estimate of \( \mu_1 - \mu_2 \). Comment.

(d) Obtain the Case 2 P-value for the alternative >; for the alternative \( \neq \). Comment.

3. Please refer to Practice Problem 3. I did the same simulation experiment, but this time both populations were exponential with mean equal to 5. Thus, a correct confidence interval estimate will include \( 5 - 5 = 0 \).

I obtained the following results: 225 intervals were too large; and 268 intervals were too small. Comment on these results.