Chapter 12

Inference for Two Numerical Populations

12.1 Comparing the Means of Two Populations; Independent Samples

We have two populations. If you want to study them individually, use the methods of Chapter 11. In this section we learn how to compare the populations, using estimation and hypothesis testing.

In this section we assume that we have random samples from the two populations and that the samples are independent. (Independent samples were discussed in Chapter 9.)

We begin with some notation. The first population has mean μ_1, standard deviation σ_1 and variance σ_1^2. The second population has mean μ_2, standard deviation σ_2 and variance σ_2^2.

Of course, the researcher does not know these six numbers, but Nature does.

We begin with the problem of estimation. Our goal is to estimate $\mu_1 - \mu_2$. Our data consist of independent random samples from the two populations.

Denote the data from the first population by: $x_1, x_2, \ldots, x_{n_1}$; and denote the data from the second population by: $y_1, y_2, \ldots, y_{n_2}$.

It is, of course, important to look at the data and think about the purpose of the research. If it seems reasonable scientifically to compare the two populations by comparing their means, then we will proceed with the methods introduced in this section.

We summarize our two sets of data by computing their means and standard deviations, which are denoted by:

$$\bar{X}, S_1, \bar{Y} \text{ and } S_2$$

when we view them as random variables, with observed values:

$$\bar{x}, s_1, \bar{y} \text{ and } s_2.$$

Our point estimate of $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y}$.

There is a CLT for this problem too. First, it shows us how to standardize our estimator:

$$W = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}. \tag{12.1}$$

Second, it states that we can approximate probabilities for W by using the snc and that in the limit as both sample sizes become larger and larger, the approximations are accurate.

First, we need to eliminate the unknown parameters in the denominator of W. Because there are now two unknown parameters where in Chapter 11 there was one, this will require additional care. Second, we will need to decide what to use for our reference curve: the snc of the CLT (and Slutsky) or the t curves of Gosset.

When all the smoke has cleared, statisticians suggest three methods, referred to in my text as Cases 1, 2 and 3. I personally think that Case 2 is scientifically worthless, so we won't cover it. (It is *mathematically* interesting, which is, in my opinion, why books feature it. Me, I put it in the book because I wanted someone else to use my book too.)

We will begin with Case 3; I will follow the popular terminology and call this the large sample approximation method.

12.1.1 Case 3: The Large Sample Approximation

Case 3 makes a lot of sense to the new student of Statistics: simply replace the population variances by their corresponding sample variances. This changes our earlier W to W_3. (The 3 is for Case 3.)

$$W_3 = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}. \qquad (12.2)$$

Case 3 states that we should use the snc as our reference curve. This leads to the following formula for the CI for $\mu_1 - \mu_2$:

$$(\bar{x} - \bar{y}) \pm z\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \qquad (12.3)$$

I will illustrate the use of this formula with an example from a student project.

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

In order to analyze his data, we will assume that we have independent random samples from two populations. Luke's data yielded the following summary statistics:

$$\bar{x} = 179.6, s_1 = 62.1, n_1 = 40, \bar{y} = 166.2, s_2 = 54.2 \text{ and } n_2 = 40.$$

The 95% confidence interval for $(\mu_1 - \mu_2)$ is

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

In words, based on my confidence interval, 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.

Our Case 3 CI is based on two approximations: replacing the population variances with the sample variances and using the snc. It is natural (especially in view of all our work in Chapter 11) to wonder whether the approximation is good. The answer turns out to be surprising. The Case 3 CI works well if both sample sizes are 20 or larger.

Note there is no reference to skewness. I will digress to discuss why skewness, which was so important in Chapter 11, is now unimportant.

First, I will show you the results of a simulation study. Suppose that the pdf is the same for both populations and the pdf is strongly skewed to the right. Thus, the difference of the means that I am trying to estimate is 0. I took independent random samples, both of size 20, from the populations and constructed the 95% CI for Case 3. I did this 1,000 times. My results are: 19 of the CIs are too large and 27 are too small. Thus, 954 (95.4%) are correct! I will discuss this example and this issue in lecture.

Case 3 also yields a hypothesis test. The null hypothesis is: $H_0 : \mu_1 = \mu_2$, and there are three options for the alternative:

$$H_1 : \mu_1 > \mu_2;\ \mu_1 < \mu_2;\ \text{or } \mu_1 \neq \mu_2.$$

Next we note that if H_0 is true, then the quantity $\mu_1 - \mu_2$ in the numerator of W_3 becomes 0. This leads us to our test statistic:

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}. \tag{12.4}$$

Once we collect our data, we calculate z, the observed value of Z, and calculate our approximate P-value using the rule introduced in Chapter 9.

Case 3 will **not** be on the final. We have learned it for two reasons.

- It is very important in practice.
- It helps get us ready for Case 1.

12.1.2 Case 1: Assuming Equal Population Variances

If either (or both) sample size is small (fewer than 20) then two related difficulties arise.

- The sample variance from the small sample size is not a very accurate estimate of its population variance.
- The snc may not give a very accurate approximation to a probability of interest.

The second of these difficulties is not major; just replace the snc by a t curve (although the issue of df is vexing). The first difficulty is more serious.

We are required to make an additional assumption about our problem. We must assume that $\sigma_1^2 = \sigma_2^2$.

Be careful **not** to read too much into this assumption. Just as I stated in Chapter 11, a researcher **never** knows the numerical value of a population variance. I am **not assuming** that I know either σ_1^2 or σ_2^2; I am assuming that these two unknown variances are the same number. (Here is a weak analogy. I meet two people who appear, to me, to be identical twins. Whereas I don't know either person's age, I am willing to assume their ages are the same.)

But still, assuming $\sigma_1^2 = \sigma_2^2$ does seem *big*. (In fact, Case 2 drops this assumption.) Here are some of the reasons why I advocate using Case 1.

1. Whereas we assume the two variances are *exactly equal* to derive the formula below, lots of simulation studies and some math theory suggest that all we really need in practice is for the variances to be close. (Interestingly, the proper way to compare variances is by taking their ratio, not difference.)

2. If the variances are very different (remember my lecture example on the effect of two social policies on length of life) then scientifically comparing the means is not a good way to compare the populations. Thus, Case 2 provides a mathematical solution to a problem that is not interesting to a scientist!

3. We never see the σ's in practice, but we do see the sd's of the two sets of data. It is my *experience* that if the two populations being compared are even the least bit similar, then their sample sd's are close, which provides evidence that their σ's are close.

4. Finally, if your data gives you very different sd's (despite my experiences related above), you can then decide not to use Case 1 and try to figure out some other way to analyze the data (perhaps by defining successes and failures and reverting to Chapter 9). Or, if you are stubborn, you can read about Case 2 in my text (or any of a large number of other texts).

Well, that is enough enrichment/intellectual honesty. Let's now learn about the method.

Because of the Case 1 assumption, the two variances in the denominator of W are the same number; call the common value of σ_1 and σ_2, just plain σ without a subscript.

We need to use our data to estimate σ^2. Our data gives us s_1^2 and s_2^2. One idea is that we could ignore one of these and use the other one to estimate σ^2, but, intuitively, it makes more sense to combine them in some way. The 'obvious' way to combine them is to calculate their mean:

$$(s_1^2 + s_2^2)/2.$$

This obvious way turns out to be 'best' mathematically if $n_1 = n_2$. If the sample sizes differ, however, we can improve on the obvious method. The idea is that we should give more 'weight' or 'emphasis' to the sample variance that is based on more data.

After studying this problem for some time, mathematicians discovered that the 'best' way to combine the sample variances is to compute s_p^2, where the 'p' is for the word 'pooled:'

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{12.5}$$

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 equals the number in the denominator.

4. If $n_1 = n_2$, then $s_p^2 = (s_1^2 + s_2^2)/2$, our obvious combination.

If we replace the unknown σ^2's in W with S_p^2, we get

$$W_1 = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(S_p^2/n_1) + (S_p^2/n_2)}}.$$

Statisticians like to factor S_p^2 out of the square root, giving:

$$W_1 = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{(1/n_1) + (1/n_2)}}.$$

Gosset showed that the appropriate reference curve for W_1 is the t curve with $df = n_1 + n_2 - 2$. There are two ways to remember the formula for df:

1. Add the sample sizes and then remember to subtract 2 (unlike subtracting one as we did in Chapter 11).

2. Add the df from each sample.

The second of these methods is preferred by statisticians and explains why we prefer to think in degrees of freedom rather than sample sizes; **df's add**.

In any event, the formula for the CI is below:

$$(\bar{x} - \bar{y}) \pm t s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

In this formula, t is obtained from the t calculator with df equal to $n_1 + n_2 - 2$, as given above. I will illustrate the use of this formula with two examples of student projects.

The data in this example come from a class project submitted 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 her data, we will assume that we have independent random samples from two populations. Sheryl's data yielded the following summary statistics:

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

We begin our analysis by computing s_p^2.

$$s_p^2 = \frac{4(4.38)^2 + 4(4.87)^2}{5 + 5 - 2} = \frac{4(19.18) + 4(23.72)}{8} = 21.45.$$

Thus, $s_p = \sqrt{21.45} = 4.63$.

The 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$(383.2 - 355.2) \pm 2.306(4.63)\sqrt{1/5 + 1/5} = 28.0 \pm 6.75 = [21.25, 34.75].$$

In words, based on my confidence interval, I conclude that Sheryl's mean time increases by between 21.25 and 34.75 seconds when the weighted pannier is added to her bike.

The data in this next example come from a class project performed by Dawn. A trial consisted of Dawn placing 10 cat treats in front of her cat Bob (no relation). In treatment 1, the treats were chicken-flavored and in treatment 2 they were tuna-flavored. The response is the number of treats Bob eats in 10 minutes. Dawn completed 20 trials, using randomization to assign 10 trials to each flavor.

In order to analyze her data, we will assume that we have independent random samples from two populations. Dawn's data yielded the following summary statistics:

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

We begin our analysis by computing s_p^2.

$$s_p^2 = \frac{9(2.025)^2 + 9(2.079)^2}{10 + 10 - 2} = \frac{9(4.10) + 9(4.32)}{18} = 4.21.$$

Thus, $s_p = \sqrt{4.21} = 2.052$.

The 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$(5.1 - 2.9) \pm 2.101(2.052)\sqrt{1/10 + 1/10} = 2.2 \pm 1.93 = [0.27, 4.13].$$

In words, based on my confidence interval, I conclude that Bob's mean consumption of treats increases by between 0.27 and 4.13 when he is offered chicken rather than tuna flavor.

Case 1 also yields a hypothesis test. The hypotheses are the same in Case 1 as they are in Case 3. The test statistic is:

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{(1/n_1) + (1/n_2)}}.$$

On the assumption the null hypothesis is correct, the sampling distribution of T is the t curve with $df = n_1 + n_2 - 2$.

We find the P-value following **exactly** the same rules we learned in Chapter 11. I will illustrate the ideas with Dawn's data on Bob.

Dawn chose the alternative that $\mu_1 \neq \mu_2$. The observed value of the test statistic is:

$$t = \frac{2.2}{2.025\sqrt{1/10 + 1/10}} = 2.43.$$

The area to the right of 2.43 under the t-curve with 18 df is 0.0129. Thus, the P-value for the third alternative is $2(0.0129) = 0.0258$.

We could also do a hypothesis test for Sheryl's data, but it strikes me as silly. (Before we collect data we **know** that the weight will slow her. It is interesting to estimate **how much** she is slowed, but testing the null of no slowing seems, as I said above, silly.) Anyways, it is below.

$$t = \frac{28.0}{4.63\sqrt{1/5 + 1/5}} = 9.56.$$

Using our online calculator, the approximate P-value for the alternative $>$ is 0.00001. Using a more precise program, I find that the area is 0.000006, or 6 in one million.

12.2 Paired Data

Recall Formula 12.3, the CI for $\mu_1 - \mu_2$ in Case 3. The half-width of the CI is $z\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$. The following often happens in practice. The values of $\bar{x}$ and $\bar{y}$ are sufficiently different so that the results are of practical importance (see Chapter 8), but the CI is so wide that the study is not conclusive. The interval is wide because s_1 and/or s_2 is large. But why is a standard deviation large? It is large because there is a large amount of subject-to-subject variation. Now we could fix this problem by making n_1 and/or n_2 larger, but more subjects in a study costs more money and sometimes runs the ethical risk of exposing more people to an inferior treatment.

There is another approach. We can try to reduce the subject-to-subject variation. Here is the idea. Suppose I have two treatments for a medical condition. Instead of having one group of people receiving one treatment and another group of people receiving the other treatment, why not give each person both treatments? Then we can compare the treatments within each person; thus, if Ralph is in the study we can compare Ralph's responses to the treatments.

Of course, for many studies it is physically impossible to give a person both treatments; in these studies my new idea won't work. But there are studies where it is possible to give a person both treatments.

Consider a study of two treatments for headache pain, call them drug A (treatment 1) and drug B (treatment 2). A trial is to wait until a person has a headache and then have the person take one of the drugs. The response is a the person's subjective assessment of head pain 30 minutes after taking the drug, measured on a scale of 0 (no pain) to 10 (worst pain ever).

There are issues of experimental design that I will discuss briefly. First, the subject should be ignorant of which drug is taken when. Also, it could bias the study if, say, every person took drug A first and then drug B. There are two ways to deal with this possible 'order bias.'

1. For each patient, the order of the drugs—A first or B first— is decided by randomization. The analysis would ignore the actual order for each patient and hope that randomization has reduced the effect of the bias, if indeed it exists.

2. Another idea is to have a **cross-over design**. In this case, one-half of the subjects take drug A first and the other subjects take drug B first. Once selected for study, by whatever method, subjects are assigned to these two groups by randomization. The analysis of the data includes an evaluation as to whether the order is important.

Table 12.1: Hypothetical Data on Headache Relief.

	Subject															
Drug	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9
B	3	0	4	3	5	1	2	2	4	3	4	7	5	6	8	7
Difference	-1	2	-1	0	-1	3	3	3	2	3	3	0	3	2	1	2

In practice, if you really are worried that an order effect might be large, then you should definitely use the cross-over design. If you believe that the order effect is either nonexistent or small, then you will probably opt for randomization because using a cross-over design makes the analysis of the data more difficult. Regarding this last point, we do not have time in these notes to learn how to analyze a cross-over design.

Table 12.1 presents hypothetical data to show the possible impact of this 'subject reuse.' Remember that smaller numbers for pain are better. Following our usual practice—remember that drug A is treatment 1— we will denote the above data in the 'A' row as x's, the data in the 'B' row as y's and the data in the 'Difference' row as d's. Thus, each subject has an x, a y and a $d = x - y$. I obtained the following summary statistics:

$$\bar{x} = 5.500, s_1 = 2.366, \bar{y} = 4.000, s_2 = 2.251, \bar{d} = 1.500 \text{ and } s_d = 1.592,$$

where s_d denotes the standard deviation of the 16 differences. Note that $\bar{d} = \bar{x} - \bar{y} = 5.500 - 4.000 = 1.500$. This is how it should be: the mean of the differences equals the differences of the means.

If we use the population model, then $\mu_1 - \mu_2 = \mu_d$. Note that we cannot use our rules for variances from Chapter 7, because there is no reason to believe that X and Y are independent; in fact, looking at the data makes me virtually certain they are not independent: Regardless of treatment some subjects tend to have low levels of pain and some tend to have high levels of pain.

We can use the methods of Chapter 11 to obtain a CI for μ_d and/or conduct a test of hypotheses. Using our web calculator, we find that the t for the 95% CI with $df = n - 1 = 16 - 1 = 15$ is 2.131. Thus, the 95% CI for μ_d is

$$1.500 \pm 2.131(1.592/\sqrt{16}) = 1.500 \pm 0.848 = [0.652, 2.348].$$

This CI allows us to conclude that B is better than A; on average on drug B pain is between 0.624 and 2.376 units smaller than on drug A.

For a hypotheses test, usually a researcher will select 0 as the special value of interest for μ_d; i.e. one is usually interested in whether the population means are equal. I will make this choice for these data and use the two-sided alternative.

The observed value of the test statistic is

$$t = \frac{1.500}{1.592/\sqrt{16}} = 3.769.$$

With the help of our online calculator, the area under the $t(15)$ curve to the right of 3.769 is 0.0009. Doubling this value, we obtain 0.0018 as the approximate P-value. This very small P-value means we have very strong evidence for the alternative; thus, we would reject the null for any common choice of α.

It is instructive to consider a 'pretend' study. Let's imagine that the researchers used 32 subjects and each subject was assigned one treatment, 16 subjects to each treatment. Each subject recorded a response for just one headache. Now, let's pretend that this new study obtained the values for x's and y's given in Table 12.1. In other words, let's pretend that the data in this table came from independent random samples.

First, we can compute the 95% CI for $\mu_1 - \mu_2$. The value of t in the CI for $df = n_1 + n_2 - 2 = 16 + 16 - 2 = 30$ is 2.042. Also, for the values of s_1 and s_2 given earlier, you can verify that $s_p = 2.309$. Thus, the CI is

$$(5.500 - 4.000) \pm 2.042(2.309)\sqrt{(1/16) + (1/16)} = 1.500 \pm 1.667 = [-0.167, 3.167].$$

This CI includes 0, so it is inconclusive. We cannot decide which drug is better. Note also that the half-width of this interval, 1.667 is almost twice as large as the half-width of the CI obtained with subject reuse, which was 0.848.

We can test the null that the two populations have the same mean versus the two-sided alternative. The observed value of the test statistic is

$$t = \frac{1.500}{2.309\sqrt{(1/16) + (1/16)}} = 1.837.$$

With the help of our online calculator, the area under the $t(30)$ curve to the right of 1.837 is 0.0381. Doubling this value, we obtain 0.0762 as the approximate P-value. This borderline P-value means we have moderate evidence for the alternative, but we would fail to reject the null for $\alpha = 0.05$.

In summary, in this hypothetical example of headaches, subject reuse greatly improved the efficiency of our analyses.

I will do another quick example of subject reuse. As I write this (November, 2009), the Yankees have just won their 27th World Series. There was much talk this season about all the home runs that were hit in the new Yankee Stadium. I decided to investigate the issue of whether it is easier to hit home runs in Yankee Stadium. (Note to baseball fans: I do not claim that this is the best way to study this issue; think of some ways you might improve on this study.)

There were nine Yankee players who played regularly; they all had at least 383 official at-bats and at least 13 home runs. (Of the remaining players, the most official at-bats was 248 and the most home runs was seven.) For each player I calculated d: the number of home runs he hit in home games minus the number he hit in away games. The sorted data are:

$$-13, -2, 3, 5, 6, 6, 8, 9, 10.$$

The small outlier, -13, surprised me; Nick Swisher had a few more at-bats in away games, but not nearly enough to explain his 21 home runs on the road, compared to only 8 at home. I calculated the following summary statistics: $\bar{d} = 3.56$ and $s_d = 7.16$. The 95% CI for μ_d is

$$3.56 \pm 2.306(7.16/\sqrt{9}) = 3.56 \pm 5.50 = [-1.94, 9.06].$$

This interval is inconclusive and very wide. I am surprised at this answer. It is instructive to see what would happen if we drop Nick Swisher's -13 from the analysis. The summary statistics become: $\bar{d} = 5.63$ and $s_d = 3.81$. Notice that dropping one outlier, reduced the standard deviation by 47%, from 7.16 to 3.81! The 95% CI for μ_d is

$$5.63 \pm 2.365(3.81/\sqrt{8}) = 5.63 \pm 3.19 = [2.44, 9.82].$$

This new analysis tells a completely different story than the data including Swisher.

I am **not** advocating throwing out Swisher's data. But you should understand how much one observation can change an analysis. Swisher has played for three teams in the three seasons 2007–2009, including only one with the Yankees. The Yankees could have easily signed a different player than Swisher and I would not be surprised if he is gone from NY by next season. (I was wrong.)

We have seen two examples of 'subject reuse' also called 'subdividing subjects.' The other way that paired data arises is by matching units. One needs to be very careful when matching units because if you do it the wrong way it can invalidate your study.

First, let me give you an improper way to match subjects. Suppose that a researcher wants to perform a study to see which state has taller men in college: Wisconsin or Minnesota. (Yes, this is a silly study.) The researcher plans to select 20 men at random from each population and, thus, will use Case 3 to compare the means.

Now Nature, played by me, enters the picture. Unbeknownst to the researcher, both populations have the same pdf, which I will take to be the normal pdf with $\mu = 69$ and $\sigma = 2.75$ inches. The researcher has just learned about paired data and decides to pair the men after sampling; i.e. the researcher will take the tallest man from each state and form a pair. After forming this first pair, the researcher will take the tallest remaining man from each state and form a pair. And so on, until the shortest man from each state is selected to form a pair. This sounds like a good idea to many researchers, but it is wrong!

I will prove to you that it is wrong by performing a computer simulation. I simulated 10,000 runs of the above study. To be precise, each run consisted of:

1. Selecting independent random samples, both of size 20, from the two populations.

2. Obtaining the sorted data from Wisconsin: $x_{(1)}, x_{(2)}, x_{(3)}, \ldots x_{(20)}$, and Minnesota: $y_{(1)}, y_{(2)}, y_{(3)}, \ldots y_{(20)}$.

3. Calculating the 20 differences of the sorted data:

$$d_1 = x_{(1)} - y_{(1)}, d_2 = x_{(2)} - y_{(2)}, \ldots d_{20} = x_{(20)} - y_{(20)}.$$

4. Calculating $\bar{d}$, s_d and the 95% CI for μ_d: $\bar{d} \pm 2.093(s_d/\sqrt{20})$.

Now remember, I created the populations so that $\mu_d = 0$. Thus, a simulated CI is correct if, and only if, it contains 0. In my 10,000 simulated runs, 6,061 of the '95% CI's' were incorrect! This is amazingly bad! CI's that are supposed to be incorrect 5% of the time in the long run are incorrect almost 61% of the time!

Most researchers, I conjecture, have too much sense of shame to cheat as blatantly as I have described above: sort two independent samples and then form pairs. But many, I suspect cheat almost as badly. My evidence? I found several texts on "Introductory Statistics in Psychology" that advocate the following.

1. Select two independent samples.

2. Measure some feature of each subject that is not the response, but is directly related to the response. (For the height example, they might advocate forming pairs based on weight.)

3. Sort values based on the feature and form pairs based on these sorted values, as I did above for the heights.

4. Analyze as paired data.

This is also cheating. Now the size of the cheat won't be as large as above (almost 61% instead of the advertised 5% above), but it will be sizable and it will be cheating.

So, why do people cheat like this? (I know, I ask the silliest rhetorical questions.) Well, researchers usually like to find differences and if one can proclaim, "I have found a difference and there is only a 5% chance I made an error," it sounds so much better than the more accurate, "I have found a difference, but there might be a 61% chance I made an error."

So, when is it valid to match units? Two situations.

1. For observational studies, you must form pairs for the entire population. Then select a random sample of pairs.

I don't see how you could do this for the height study, unless you measure every man in both states, which would eliminate the need for estimation. But you could do this for some studies. For example, suppose you are interested in comparing the heights of all married men in Wisconsin to the heights of all married women. (At the time of this typing, in Wisconsin legal marriage is between exactly two persons and the two persons must be one of each sex.) A natural way to form pairs is to have each married couple be a pair. Then by sampling couples you automatically sample pairs. This sampling strategy would be better than independent samples if there is a strong enough tendency in society for tall to marry tall and short to marry short. (Weight or age might work better than height, but I really don't know much about how people choose marriage partners.)

2. For experimental studies, you may form pairs however you want as long as you do this before you randomly assign one member of each pair to each treatment.

Table 12.2: Study of the Preview Feature in Tetris.

	Pair									
Treatment	1	2	3	4	5	6	7	8	9	10
1: Preview	106	112	118	102	112	110	130	110	127	138
2: No preview	84	93	86	86	94	88	108	91	79	91
$(1) - (2)$	12	19	32	16	18	22	22	19	48	47

Well, I don't have much else to say, but rather than leave you with a blank page in these notes, I will give one more example.

Let's return to the game of Tetris. We learned in Chapter 6 that the individual trials definitely have memory. But now I want to focus on an entire game as a trial. Many years ago, I enjoyed playing Tetris (I was pretty horrible at every other video game). My game had a feature that allowed you to see or not see the next shape while you are manipulating the current shape. (Seeing was the default.) It seemed to me that selecting the default, preview, option would lead to much higher scores. So, I decided to collect data to investigate this matter.

A game is a trial and the response is the number of lines I completed before the game ended. I decided to perform 20 trials, with 10 on each setting. I was very worried that fatigue or boredom would affect my later scores, so I formed pairs out of consecutive trials: 1 and 2; 3 and 4; and so on. Within each pair I randomly assigned one game to each treatment. My data are in Table 12.2.

Not surprisingly, and obviously from even a quick glance at the data, I was a much better player with the preview option. It is not so clear that pairing was needed; we shall explore this issue below.

I calculated the following summary statistics:

$$\bar{x} = 116.5, s_1 = 11.56, \bar{y} = 90.0, s_2 = 7.77, \bar{d} = 26.5 \text{ and } s_d = 11.87,$$

With $df = 9$, the value needed for the 95% CI is $t = 2.262$. Thus, the 95% CI for μ_d is

$$26.50 \pm 2.262(11.87/\sqrt{10}) = 26.50 \pm 8.49 = [18.01, 34.99].$$

At the 95% confidence level, on average, my ability increased between 18 and 35 lines when I had the preview option on.

For comparison, we will now pretend that the data come from independent random samples. First,

$$s_p^2 = [(11.56)^2 + (7.77)^2]/2 = 97.00.$$

Thus, $s_p = 9.85$. The 95% CI for $\mu_1 - \mu_2$ is

$$(116.5 - 90.0) \pm 2.101(9.85)\sqrt{1/10 + 1/10} = 26.50 \pm 9.25 = [17.25, 35.75].$$

Pairing was effective, but not by much.