Lecture 26

Nancy Pfenning Stats 1000

Chapter 12: More About Confidence Intervals

Recall: Setting up a confidence interval is one way to perform statistical inference: we use a *statistic* measured from the *sample* to construct an interval estimate for the unknown *parameter* for the *population.* We learned in Chapter 10 how to construct a confidence interval for unknown population proportion p based on sample proportion $\hat{p}$, when there was a single categorical variable of interest, such as smoking or not.

In this chapter, we will learn how to construct other confidence intervals:

- for population mean μ based on sample mean $\bar{x}$ when there is one quantitative variable of interest;
- for population mean difference μ_d based on sample mean difference $\bar{d}$ in a matched pairs study when the single set of (quantitative) differences d is the variable of interest;
- for difference between population means $\mu_1 - \mu_2$ based on difference between sample means $\bar{x}_1 - \bar{x}_2$ in a two-sample study.

The latter two situations involve one quantitative variable and an additional categorical variable with two possible values, although we may think of the distribution of differences in the matched-pairs study as a single quantitative variable.

Also discussed in the textbook but not in our course is the method of constructing a confidence interval for the difference between two population proportions $p_1 - p_2$ based on the difference between sample proportions $\hat{p}_1 - \hat{p}_2$. Because such situations involve two categorical variables, they can be handled instead with a chi-square procedure, which will be discussed further in Chapter 15.

The Empirical Rule for normal distributions allowed us to state that in general, the probability is 95% that a normal variable falls withing 2 standard deviations of its mean. Since sample proportion $\hat{p}$ for a large enough sample size n is approximately normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$, we were able to construct an approximate 95% confidence interval for p: $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

In general, an approximate 95% confidence interval for a parameter is the accompanying statistic plus or minus two standard errors; this works well if the statistic's sampling distribution is approximately normal. If we are interested in the unknown population mean μ when there is a single quantitative variable of interest, we use the fact (established in Chapter 9) that sample mean $\bar{x}$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. For a large enough sample size n (say, n at least 30), population standard deviation σ will be fairly well approximated by sample standard deviation s and so our standard error for $\bar{x}$ is $s.e.(\bar{x}) = \frac{s}{\sqrt{n}}$. Also for large n, by virtue of the Central Limit Theorem, the distribution of $\bar{x}$ will be approximately normal, even if the underlying population variable X is not. Thus, for a large sample size n, the Empirical Rule tells us that an approximate 95% confidence interval for population mean μ is

$$\bar{x} \pm 2\frac{s}{\sqrt{n}}$$

Example

> The mean number of credits taken by a sample of 81 statistics students was 15.60 and the standard deviation was 1.8. Construct an approximate 95% confidence interval for the mean number of credits taken by all statistics students; does this interval also have a 95% chance of capturing the mean number of credits taken by all students at the entire university?
>
> $$\bar{x} \pm 2\frac{s}{\sqrt{n}} = 15.60 \pm 2\frac{1.8}{\sqrt{81}} = 15.60 \pm .40 = (15.20, 16.00)$$
>
> This interval applies to statistics students only. Especially because the intro stats courses are 4 credits each instead of the usual 3, these students may average slightly higher credit hours than students in general.

Recall: The Empirical Rule is only roughly accurate; besides, we sometimes may prefer a different level of confidence other than .95. More precise standard normal values for confidence levels .90, .95, .98, and .99 may be obtained the "infinite" row at the bottom of Table A.2. [The row is called "infinite" because t^* multipliers converge to z^* for infinite sample sizes—same as infinite degrees of freedom.] We can summarize the intervals as follows: for a large sample size n, an approximate

90% confidence interval for μ is

$$\bar{x} \pm 1.645 \frac{s}{\sqrt{n}}$$

95% confidence interval for μ is

$$\bar{x} \pm 1.960 \frac{s}{\sqrt{n}}$$

98% confidence interval for μ is

$$\bar{x} \pm 2.326 \frac{s}{\sqrt{n}}$$

99% confidence interval for μ is

$$\bar{x} \pm 2.576 \frac{s}{\sqrt{n}}$$

Example

The mean number of credits taken by a sample of 81 statistics students was 15.60 and the standard deviation was 1.8. Construct a more precise 95% confidence interval for the mean number of credits taken by all statistics students. Then construct a 90% confidence interval.

A 95% confidence interval is

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} = 15.60 \pm 1.96 \frac{1.8}{\sqrt{81}} = 15.60 \pm .39 = (15.21, 15.99)$$

A 90% confidence interval is

$$\bar{x} \pm 1.645 \frac{s}{\sqrt{n}} = 15.60 \pm 1.645 \frac{1.8}{\sqrt{81}} = 15.60 \pm .33 = (15.27, 15.93)$$

Note the trade-off: we obtain a narrower, more precise interval when we make do with a lower level of confidence.

Recall: we learned in Chapter 9 that not all standardized test statistics follow a standard normal (z) curve. In particular, when the sample size n is small, s may be quite different from σ, and the random variable $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ follows a t distribution with $n-1$ degrees of freedom, not a z distribution. Especially for small samples, t has more spread than the standard normal z. It is still symmetric about zero and bell-shaped like the z curve. Table A.2 provides t^* multipliers for constructing 90%, 95%, 98%, or 99% confidence intervals for unknown population mean μ when the sample size is on the small side.

Example

Suppose a sample of only 9 statistics students averaged 15.60 credits, with standard deviation 1.8. Construct 95% and 90% confidence intervals for the mean number of credits taken by all statistics students.

A sample of size $n = 9$ has $df = n - 1 = 9 - 1 = 8$, and so we obtain the correct t^* multipliers from the 8 *df* row of Table A.2:

A 95% confidence interval is

$$\bar{x} \pm 2.31 \frac{s}{\sqrt{n}} = 15.60 \pm 2.31 \frac{1.8}{\sqrt{9}} = 15.60 \pm 1.39 = (14.21, 16.99)$$

A 90% confidence interval is

$$\bar{x} \pm 1.86\frac{s}{\sqrt{n}} = 15.6 \pm 1.86\frac{1.8}{\sqrt{9}} = 15.60 \pm 1.12 = (14.48, 16.72)$$

Not only are the t^* multipliers larger than the z^* multipliers, but we are dividing by the square root of a much smaller sample size n, which results in much wider intervals than we had for the sample of size 81.

Example

Suppose the FAA weighed a random sample of 25 airline passengers during the summer and found their weights to have mean 180, standard deviation 40. Give a 99% confidence interval for the mean summer weight of all airline passengers.

We use the t^* multiplier for the $df = 25 - 1 = 24$ row and the column for confidence level .99. Our 99% confidence interval is

$$180 \pm 2.80\frac{40}{\sqrt{25}} = 180 \pm 22.4 = (147.6, 202.4)$$

Conditions for Using the t Confidence Interval

It is important to remember that if the population X is not normal, then neither is sample mean $\bar{x}$, and so the R.V. $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ does not have a t distribution, and the t^* values from Table A.2 are not necessarily correct. Fortunately, t procedures tend to be robust against non-normality, especially for larger sample sizes n, except for extreme outliers or strong skewness in the population, demonstrated by outliers or skewness in the data. Methods of Chapter 2 are essential now for determining the shape of the population. [Note: there is no way of rescuing data that has been obtained through a poor design. All of our theory requires a simple random sample, taken from a population that is at least 10 times the sample size.]

Thus, t^* values will produce an accurate confidence interval if the sample size is large or if the sample size is small but the data show no outliers or pronounced skewness. The t^* values will **not** produce an accurate confidence interval if the sample size is small and the data show outliers or skewness.

Example

The sample of 9 students included a part-time student taking only 4 credits. Is our confidence interval necessarily accurate? No, because the sample size is small and there is a very low outlier.

Example

Is our confidence interval for mean weight of airline passengers accurate? Weights tend to follow a normal curve, and anyway the sample size of 25 isn't especially small, so the interval obtained above should be quite accurate.

Lecture 27

Last time we learned to construct a confidence interval for unknown population mean of a quantitative variable, such as credits taken by statistics students or weights of airline passengers. Table A.2 provides t^* multipliers for various sample sizes and four different levels of confidence. The "infinite" row contains z^* multipliers which apply when the sample size is large enough that s is virtually identical to σ.

Matched Pairs t Procedures

One of the Three Basic Principles of Experimental Design is to *control* the effects of confounding variables by comparing several treatments, or treatment to control. One way to do a comparison is a **matched pairs** study, where individuals are matched in pairs (!). Two different treatments may be assigned to each pair, with the assignment randomized (for instance, using coinflips), and outcomes are compared within each pair. Alternatively, the response of an individual *before* treatment is paired with his or her response *after* treatment. Or values of a particular variable may be studied for both members of a pair, eg. comparing earnings of husbands and wives.

Although such studies of a quantitative variable originally include an additional categorical variable (such as whether the subject was given the drug or the placebo, or whether the spouse is male or female), a matched pairs situation reverts to the study of a single quantitative variable, namely the *single sample of differences.* The population mean difference is denoted μ_d, whereas the sample mean difference is denoted $\bar{d}$. Robustness is assessed based on the n pairs of observed differences d, not the $2n$ data values.

Example

A social scientist wants to produce statistical evidence that men earn more than women. She records these salaries for a sample of 11 husband-wife pairs:

Husband	Wife	Difference
28	20	8
28	22	6
31	32	-1
32	10	22
34	25	9
35	29	6
36	32	4
40	27	13
45	40	5
80	70	10
145	0	145

Do the data support the scientist's theory? If robust, construct a 95% confidence interval for the population mean difference μ_D and check if it contains zero. Since the sample size is quite small and the income differences have an obvious outlier (145), we should not use t procedures.

Example

Here are average weekly losses of man-hours due to accidents in 10 individual plants before and after a certain safety program was put into operation. Construct a 95% confidence interval for the mean decrease in weekly man-hours lost due to accidents for all plants after implementing the safety program, and use the interval to decide if the program seems effective.

Before	After	Difference
45	36	9
73	60	13
46	44	2
124	119	5
33	35	-2
57	51	6
83	77	6
34	29	5
26	24	2
17	11	6
		$d = 5.2$
		$s_d = 4.1$

First we can verify with a histogram that the data show no outliers or skewness, and are approximately normal. Next,

we find that a 95% confidence interval for the population mean difference μ_d is $\bar{d} \pm t^* \frac{s}{\sqrt{n}}$ where t^* comes from the $df = 9$ row and the .95 confidence level column. Our confidence interval is

$$5.2 \pm 2.262(\frac{4.1}{\sqrt{10}}) = 5.2 \pm 2.9 = (2.3, 8.1)$$

We are 95% confident that the population mean difference in average weekly man hours lost is between 2.3 and 8.1. (Implicit is the assumption that the plants constitute a random sample of all plants for which such a safety program is intended.) Since the interval is strictly to the right of zero, containing only positive numbers, it suggests that there was a real decrease in mean man-hours lost from before to after. However, the study design is somewhat flawed because time could possibly be a confounding variable. Perhaps because of heightening awareness of safety issues (and increased fear of lawsuits), there was a general decrease in man-hours lost due to accidents during that time period, even in plants that did not implement the safety program. How could we control for this possible confounding variable? By comparing our ten plants to another sample of plants over the same time period which did *not* implement the safety program. Such a design, because it involves samples from two distinct populations, is called a **two-sample** design.

Comparing Two Means

We will use inference to compare the mean responses in two groups, each from a distinct population. This is called a **two-sample** siutation, one of the most common settings in statistical applications. One example would be to compare mean IQ's of male and female seventh-graders—i.e., comparing results in an *observational study.* Another example would be to compare the change in blood pressure for two groups of black men, where one group has been given calcium supplements, the other a placebo—i.e., comparing results in an *experiment.* In general, a two-sample t procedure arises in situations where there is one quantitative variable of interest, plus a categorical variable which has two possible values. The variables in the first example are IQ and gender; in the second example they are blood pressure and whether the subject has been given calcium or a placebo.

Responses in each group must be independent of those in the other; sample sizes may differ. The setting is *not* appropriate for matched pairs, which represent a single population. The following notation is used to describe the two populations and the results of two independent random samples:

	Parameters			Statistics		
Population	R.V.	mean	s.d.	sample size	sample mean	sample s.d.
1	X_1	μ_1	σ_1	n_1	$\bar{x}_1$	s_1
2	X_2	μ_2	σ_2	n_2	$\bar{x}_2$	s_2

Naturally enough, we estimate the parameter $\mu_1 - \mu_2$ with the statistic $\bar{x}_1 - \bar{x}_2$. As one would hope and expect, it turns out that the distribution of the R.V. $\bar{x}_1 - \bar{x}_2$ is centered at $\mu_1 - \mu_2$, providing an *unbiased estimator*. The spread of the distribution is not so intuitive; it can be shown that the standard error of $\bar{x}_1 - \bar{x}_2$ is

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We use the above mean and standard error to standardize $\bar{x}_1 - \bar{x}_2$ to the **two-sample t statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Although this R.V. does *not* have a t distribution per se, it can still be used with t^* values in either or two ways:

Option 1: Approximate

$$df = (\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2 / [\frac{1}{n_1 - 1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2 - 1}(\frac{s_2^2}{n_2})^2]$$

and use the t table. [The computer takes this approach, but for obvious reasons we would rather not, if solving a two-sample problem by hand. Instead, we will use...]

Option 2: (conservative approach) use the smaller of $n_1 - 1$,$n_2 - 1$ as our *df* in the t table:

An approximate confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^* uses the smaller of $n_1 - 1$, $n_2 - 1$ as its *df* and the desired confidence level dictates which column from Table A.2 to use. This interval should be fairly accurate as long as the sample sizes are large, or if small samples show no outliers or skewness.

Example

In random samples of 47 male and 31 female seventh-graders in a Midwest school district, IQ's were found to have the following means and standard deviations:

Group	n	$\bar{x}$	s
Males	47	111	12
Females	31	106	14

1. What shapes are required of the underlying populations to justify use of two-sample t procedures? Any shapes should be acceptable, since the sample sizes of 47 and 31 are reasonably large.
2. Use a two-sample t procedure to give a 90% confidence interval for the difference between mean IQ's, males minus females.

   The 90% confidence interval for $\mu_1 - \mu_2$ is given by

   $$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

   where we take t^* to be the value for the smaller of $47 - 1$, $31 - 1$ *df*, and confidence level .90. We find the t^* value for the 30 row, .90 column to be 1.70 and our 90% confidence interval is

   $$(111 - 106) \pm 1.70\sqrt{\frac{12^2}{47} + \frac{14^2}{31}} = 5 \pm 6 = (-1, 11)$$

3. It is common for boys to score somewhat higher than girls on standardized tests. Does this seem to be the case for all seventh-grade boys and girls in this school district? The interval just barely contains zero, so it is difficult to be sure. Eventually we will learn to carry out a formal test of whether or not two means μ_1 and μ_2 are equal.

Pooled Two-Sample t Procedures

If the samples are coming from populations that have equal variances, we can use a **pooled** procedure. The test statistic can be shown to follow a genuine t distribution, with $n_1 + n_2 - 2$ *df*. This places us further down on the t table than taking the smaller of $n_1 - 1$ and $n_2 - 1$ as our *df*, resulting in slightly narrower confidence intervals. One criterion for use of a pooled procedure is to check that sample standard deviations are close enough to suggest equal population standard deviations, and hence equal variances. We do this by verifying that the larger sample standard deviation is no more than twice the smaller.

Example

Looking at the sample standard deviations for IQs, we note that 14 is not more than twice 12, so a pooled procedure seems appropriate.

There are actually much better criteria for use of a pooled procedure, which are outlined in your textbook. In any case, for our purposes in this course the non-pooled procedure will be considered adequate.

Example

In a previous Example, we explored the sampling distribution of sample mean height, when random samples are taken from a population of women whose mean height is claimed to be 64.5. We noted the sample mean height of surveyed Stats female students, and calculated by hand the probability of observing such a high sample mean, if population mean were really only 64.5. We used this probability to decide whether we were willing to believe that population mean was in fact 64.5, or if the population of female Stats students is actually taller, on average. For this Example, we address the same question by using MINITAB to set up a confidence interval for unknown population mean height, given that population standard deviation is 2.5 (thus, a z procedure is used). When a one-sided alternative is *not* specified, the confidence interval just barely contains 64.5 (it goes down to 64.491), and so we can't quite produce evidence that population mean height of females differs 64.5. If a greater-than alternative *is* specified, then our lower bound is 64.538, which would suggest population mean height is higher than 64.5.

If the standard deviation of 2.5 were not given, we would carry out a t procedure. Again, the confidence interval just barely contains 64.5 with a two-sided alternative, and just barely misses it with a one-sided alternative.

Considering the 95% confidence interval will give results that match up neatly with those of a hypothesis test at the 5% level only in the case of a two-sided alternative.

```
One-Sample Z: HT_female

Test of mu = 64.5 vs mu not = 64.5
The assumed sigma = 2.5

Variable          N      Mean     StDev   SE Mean
HT_female       281    64.783     2.637     0.149

Variable            95.0% CI            Z       P
HT_female     (  64.491,  65.075)     1.90  0.058

One-Sample Z: HT_female
```

```
Test of mu = 64.5 vs mu > 64.5
The assumed sigma = 2.5

Variable          N      Mean     StDev   SE Mean
HT_female       281    64.783     2.637     0.149

Variable      95.0% Lower Bound        Z      P
HT_female                64.538     1.90  0.029

One-Sample T: HT_female

Test of mu = 64.5 vs mu not = 64.5

Variable          N      Mean     StDev   SE Mean
HT_female       281    64.783     2.637     0.157

Variable             95.0% CI            T      P
HT_female     (  64.473,  65.093)     1.80  0.073

One-Sample T: HT_female

Test of mu = 64.5 vs mu > 64.5

Variable          N      Mean     StDev   SE Mean
HT_female       281    64.783     2.637     0.157

Variable      95.0% Lower Bound        T      P
HT_female                64.523     1.80  0.037
```

Exercise: *In a previous Exercise, we explored the sampling distribution of sample mean number selected, when random samples are taken from a population where all numbers between 1 and 20 are equally likely, so population mean is 10.5. We noted the sample mean selection by surveyed Stats students, and calculated by hand the probability of observing such a high sample mean, if population mean were really only 10.5. We used this probability to decide whether we were willing to believe that population mean was in fact 10.5, or if students were rather biased towards higher numbers.* For this Exercise, address the same question by using MINITAB to set up a confidence interval for unknown population mean selection, given that population standard deviation is 5.77. Does your interval contain 10.5? What do you conclude?

Exercise: For this Exercise, address the same question again by using MINITAB to set up a confidence interval for unknown population mean selection, but this time assume population standard deviation is unknown. Does your interval contain 10.5? What do you conclude?

Lecture 28

Chapter 13: More About Significance Tests

Recall: Hypothesis tests are a form of statistical inference: we use a *statistic* measured from the *sample* to decide whether or not the unknown *parameter* for the *population* equals a hypothetical value. We learned in Chapter 11 how to test a hypothesis about an unknown population proportion p based on sample proportion $\hat{p}$, when there was a single categorical variable of interest, such as smoking or not.

In this chapter, we will learn how to perform other hypothesis tests:

- about population mean μ based on sample mean $\bar{x}$ when there is one quantitative variable of interest;

- about population mean difference μ_d based on sample mean difference $\bar{d}$ in a matched pairs study when the single set of (quantitative) differences d is the variable of interest;
- about difference between population means $\mu_1 - \mu_2$ based on difference between sample means $\bar{x}_1 - \bar{x}_2$ in a two-sample study.

Also discussed in the textbook but not in our course is the method of testing hypotheses about the difference between two population proportions $p_1 - p_2$ based on the difference between sample proportions $\hat{p}_1 - \hat{p}_2$. Because such situations involve two categorical variables, they can be handled with a chi-square procedure, which will be discussed further in Chapter 15.

Paralleling the specific steps we learned to test a hypothesis about a single proportion, the following five steps can be taken to test a hypothesis about any unknown parameter:

1. Determine the null and alternative hypotheses.
2. Verify that the necessary data conditions are met; if so, standardize the sample statistic.
3. Find the p-value, which is the probability, assuming the null hypothesis is true, that the test statistic would take a value as high/low/different as the one observed.
4. Decide whether results are statistically significant: reject the null hypothesis if the p-value is "small".
5. State conclusions in context.

Hypothesis Tests About μ or μ_d

If we are interested in the unknown population mean μ when there is a single quantitative variable of interest, we use the fact that sample mean $\bar{x}$ has mean μ and standard error $\frac{s}{\sqrt{n}}$. In order to obtain accurate results for smaller sample sizes, since s may be quite different from σ, our standardized test statistic $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ is taken to follow a t distribution with $n-1$ degrees of freedom, not a z distribution.

We learned to use Table A.2 in Chapter 11 to get a range for the P-value in hypothesis tests about p by surrounding our test statistic z with values z^* from the "infinite" row of the table. The columns correspond to symmetric tail areas of .05 for confidence level .90, tail areas .025 for confidence level .95, tail areas .01 for confidence level .98, and tail areas .005 for confidence level .99.

Now we use Table A.2 to get a range for the P-value in hypothesis tests about the mean by surrounding our test statistic t with values t^* from the $df = n-1$ row of the table. Again, the columns correspond to symmetric tail areas of .05, .025, .01, and .005, respectively.

Using Table A.2, our hypothesis test about μ follows these steps:

1. Set up $H_0 : \mu = \mu_0$ vs. $H_a : \mu \left\{ \begin{array}{c} < \\ > \\ \neq \end{array} \right\} \mu_0$
2. Verify that the sample size is large or the data set shows no outliers or skewness; if so, calculate $t_{\text{statistic}} = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$.
3.
$$\begin{aligned} \text{Get an expression for the P-value} &= P(T_{R.V.} \leq t_{\text{statistic}}) \text{ for } H_a : \mu < \mu_0 \\ &= P(T_{R.V.} \geq t_{\text{statistic}}) \text{ for } H_a : \mu > \mu_0 \\ &= 2P(T_{R.V.} \geq |t_{\text{statistic}}|) \text{ for } H_a : \mu \neq \mu_0 \end{aligned}$$
4. Assess significance by comparing the t statistic to t^* values in Table A.2 and getting a range for the P-value.
5. State conclusions in the context of the particular mean of interest.

Example

I had been going under the assumption that my students averaged 15 credits in a semester, but then I thought that because mine is a 4 credit course, their mean may actually be higher than 15. The mean number of credits taken by a sample of 81 statistics students was 15.6 and the standard deviation was 1.8. Does this provide evidence that statistics students overall average more than 15 credits?

1. $H_0 : \mu = 15$ vs. $H_a : \mu > 15$
2. Since $n = 81$ is large, non-normal shape would not be a problem; calculate $t = \frac{15.6-15}{\frac{1.8}{\sqrt{81}}} = 3$
3. P-value= $P(T \geq 3)$
4. For 80 df, 3 is greater than 2.64, so the P-value is less than .005: results are statistically significant, and we reject H_0.
5. Overall, statistics students average more than 15 credits.

Example

Suppose a sample of 36 statistics students had been taken. How many df should we use from Table A.2? Note that the table does not include exactly 35 df, so we must choose between 30 and 40. Always choose the smaller df, because this makes it slightly more difficult to reject the null hypothesis, which is the safer approach to take. Thus, we would carry out the test using the 30 df row of Table A.2.

Conditions for Using the t Test

Just as with confidence intervals, P-value ranges obtained by comparing the t statistic to t^* values in Table A.2 will produce accurate results if the sample size is large or if the sample size is small but the data show no outliers or pronounced skewness. The table will **not** produce accurate results if the sample size is small and the data show outliers or skewness.

Matched Pairs Hypothesis Tests

When the mean of a quantitative variable is explored via a matched pairs design, hypothesis tests are carried out on the population mean difference μ_d based on the sample mean difference $\bar{d}$.

Example

To test if students' mothers tend to be younger than their fathers, I looked at the difference mother's age minus father's age for a sample of 12 students. This difference had mean $\bar{d} = -1.5$ and standard deviation $s_d = 3$. Is the mean difference significantly less than zero, using $\alpha = .05$ as the cut-off probability?

To test $H_0 : \mu_d = 0$ vs. $H_a : \mu_d < 0$, we check if the distribution seems fairly symmetric and outlier-free (it is) and calculate $t = \frac{-1.5-0}{3/\sqrt{12}} = -1.73$. Because the alternative has the "<" sign, our P-value is $P(T \leq -1.73)$. According to the table, the probability of a T random variable with 11 df being greater than 1.80 is .05; likewise the probability of being less than -1.8 is also .05. The test statistic -1.73 isn't as far out on the tail of the t curve as -1.8, so its tail probability is more than .05. We do not have evidence to reject the null hypothesis at $\alpha = .05$. A sample of 12 age differences was not enough to convince us that mothers tend to be younger. [In fact, another much larger sample was taken, producting a much smaller P-value, and this sample did provide evidence that the mean age difference is negative.]

Hypothesis Tests About the Difference Between Two Means

We can test for equality of the mean responses in two groups, each from a distinct population. This is called a **two-sample** siutation, one of the most common settings in statistical applications. One example would be to compare mean IQ's of male and female seventh-graders—i.e., comparing results in an *observational study.* Another example would be to compare the change in blood pressure for two groups of black men, where one group has been given calcium supplements, the other a placebo—i.e., comparing results in an *experiment.*

As with confidence intervals, we use the following notation:

	Parameters			Statistics		
Population	R.V.	mean	s.d.	sample size	sample mean	sample s.d.
1	X_1	μ_1	σ_1	n_1	$\bar{x}_1$	s_1
2	X_2	μ_2	σ_2	n_2	$\bar{x}_2$	s_2

The null hypothesis is $H_0 : \mu_1 = \mu_2$ [same as $H_0 : \mu_1 - \mu_2 = 0$] and the alternative substitutes the appropriate inequality for "=". We carry out our test using the **two-sample t statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and use the smaller of $n_1 - 1$,$n_2 - 1$ as our *df* in the t table. The approximate P-value is found in the usual way from Table A.2.

Example

In random samples of 47 male and 31 female seventh-graders in a Midwest school district, IQ's were found to have the following means and standard deviations:

Group	n	$\bar{x}$	s
Males	47	111	12
Females	31	106	14

Is the mean male IQ significantly higher than that for the females? Test at level $\alpha = .05$.

We will test $H_0 : \mu_1 - \mu_2 = 0$ vs $H_a : \mu_1 - \mu_2 > 0$. Our two-sample t statistic is

$$t = \frac{(111 - 106) - 0}{\sqrt{\frac{12^2}{47} + \frac{14^2}{31}}} = 1.63$$

For 30 *df*, 1.63 is less than 1.70, so our (one-sided) P-value is greater than .05. There is not quite enough evidence to reject H_0 at the .05 level; the population of boys doesn't necessarily average higher than the population of girls in this district.

Example

Suppose the FAA weighed a random sample of 25 airline passengers during the summer and found their weights to have mean 180, standard deviation 40. Are airline passengers necessarily heavier now than they were in 1995, when mean weight for 16 passengers was 160, with standard deviation 30? Answer this question two ways: first by looking at a 90% confidence interval for the difference in mean weights, then by testing at the .05 level if the mean weight increased. (Note that since the sample sizes aren't especially large, we should first check that the weight distributions do not show obvious outliers or skewness.) We have $n_1 = 16$, $n_2 = 25$; $\bar{x}_1 = 160$, $\bar{x}_2 = 180$; $s_1 = 30$, $s_2 = 40$.

1. We use the t^* multiplier for the $df = 16 - 1 = 15$ row and the column for confidence level .90. Our 90% confidence interval for $\mu_1 - \mu_2$ is

   $$160 - 180 \pm 1.75\sqrt{\frac{30^2}{16} + \frac{40^2}{25}} = -20 \pm 19 = (-39, -1)$$

   The interval contains only negative numbers, and suggests a significant increase in mean weight from 1995 to 2002.

2. We test $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 < 0$ about population mean weight in 1995 minus population mean weight in 2002. The test statistic is $t = \frac{160-180}{\sqrt{\frac{30^2}{16}+\frac{40^2}{25}}} = 1.82$ For 15 df, 1.82 is between 1.75 and 2.13, so the P-value is between .05 and .025. We reject H_0 at the .05 level, and conclude that mean weight has increased significantly. The FAA reached this conclusion in the spring of 2003, and made new restrictions on number of passengers aboard smaller planes based on the fact that people are heavier than they used to be.

Multiple Hypothesis Tests

Example

Verbal SATs have mean 500. An education expert samples verbal SAT scores of 20 students each in 100 schools across the state, and finds that in 4 of those schools, the sample mean verbal SAT is significantly lower than 500, using $\alpha = .05$. Are these schools necessarily inferior in that their students do significantly worse on the verbal SATs? No. First note that 20 indicates the sample size here, and 100 is the number of tests—in other words, we test $H_0 : \mu = 500$ vs $H_a : \mu < 500$ over and over, one hundred times. Remember that if $\alpha = .05$ is used as a cutoff, then 5% of the time in the long run we will reject H_0 even when it is true. Roughly, 5 schools in 100 will produce samples of students with verbal SATs low enough to reject H_0, just by chance in the selection process, even if the mean for all students at those schools is in fact 500.

Example

Kanarek and others studied the relationship between cancer rates and levels of asbestos in the drinking water. After adjusting for age and various demographic variables, *but not smoking*, they found a "strong relationship" between the rate of lung cancer among white males and the concentration of asbestos fibers in the drinking water: p-value<.001. [An increase of 100 times the asbestos concentration results in an increase of 1.05 per 1000 in the lung cancer rate—one additional lung cancer case per year for every 20,000 people.] The investigators tested over 200 relationships...the p-value for lung cancer in white males was by far the smallest one they got. Does asbestos in the drinking water cause lung cancer in white males?

No! When they test hundreds of relationships, sooner or later by chance alone some will end up looking significant. [There are other problems with this study: failing to control for the possible confounding variable of smoking, and calling a relationship "strong" even though it would imply just one additional case of lung cancer for every 20,000 white males.]

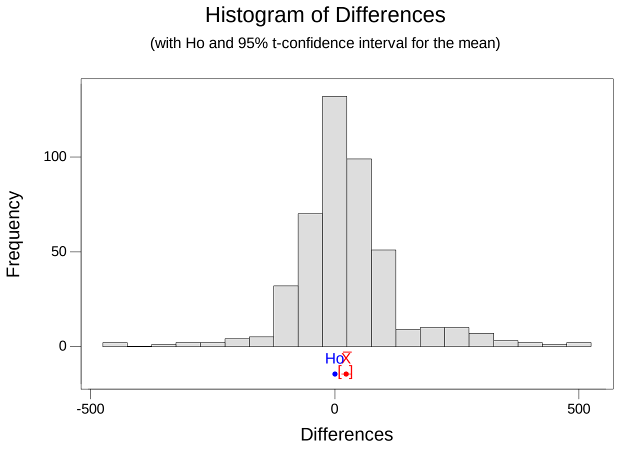
Example

A researcher of ESP tests 500 subjects. Four of them do significantly better (each P-value < .01) than random guessing. Should the researcher conclude that those four have ESP? No! In so many trials, even if each subject is just guessing, chances are that a few of the 500 will do significantly better than guessing (and a few will do significantly worse!). The researcher should proceed with further testing of those four subjects.

In general, we should be aware that many tests run at once will probably produce some significant results by chance alone, even if none of the null hypothese are false.

Example

Do students overall spend more time on the computer than they do watching TV? If so, then when I consider the differences in minutes spent, computer minus TV, for a population of students, I'd hypothesize the mean of the differences to be positive. A paired t procedure based on computer and TV times of several hundred Stats students could be used to test my hypothesis. Since a paired test like this really just involves one quantitative variable—the single sample of differences—an appropriate display would be a histogram of observed differences; note that it is remarkably bell-shaped, but may or may not be centered at zero:



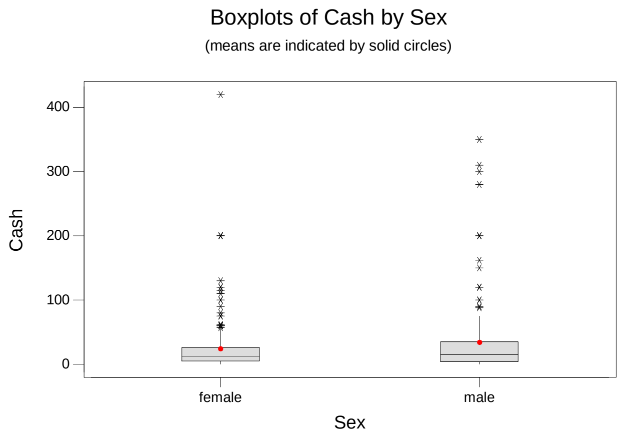
```
Paired T for Compu - TV

                  N      Mean     StDev   SE Mean
Compu           444     81.64     88.61      4.21
TV              444     58.18     70.28      3.34
Difference      444     23.47    110.50      5.24

95% lower bound for mean difference: 14.82
T-Test of mean difference = 0 (vs > 0): T-Value = 4.47  P-Value = 0.000
```

The P-value of 0.000 lets me reject the null hypothesis that the population mean difference is zero, and conclude that it is indeed positive. Apparently students do spend more time overall on the computer than they do watching TV.

Exercise: Find paired data in our survey, such as math and verbal SATs, ages of mothers and fathers, heights of females and their mothers, or heights of males and their fathers. Use MINITAB to test $H_0 : \mu_d = 0$ against an appropriate H_a. State your conclusion in terms of the variable chosen.

Example

Who carries more cash, males or females? Or don't they differ? I can use MINITAB to test the null hypothesis that mean cash carried for populations of females is the same as that for males, vs. a two-sided alternative (I had no preconceptions in advance of one group carrying more money). When comparing values of one quantitative variable for two categorical groups, side-by-side boxplots would be an appropriate display:



```
Two-sample T for Cash

Sex           N      Mean     StDev   SE Mean
female      280      24.0      39.6       2.4
male        159      34.2      58.4       4.6

Difference = mu (female) - mu (male  )
Estimate for difference:  -10.23
95% CI for difference: (-20.47, 0.02)
T-Test of difference = 0 (vs not =): T-Value = -1.97  P-Value = 0.050  DF = 241

* NOTE * N missing = 7
```

The P-value of .05 is on the small side, leading us to conclude that there is a significant difference between males and females in the amount of cash that they carry. Since the difference between sample means, female minus male, was negative, we have reason to believe that overall males carry more cash. The sample mean for females was about $24; for males about $34.

Exercise: Compare values of a quantitative survey variable for two categorical groups, such as males and females or on and off campus students, by testing $H_0 : \mu_1 - \mu_2 = 0$ against an appropriate H_a. State your conclusion in terms of the variable chosen.

Exercise: Read the article **The most important meal**, which reports that in a study of American eight-graders in 96 public schools in San Diego, New Orleans, Minneapolis, and Austin, overweight students were more likely to skip breakfast than students who were not overweight. Unstack the data in our class survey according to gender, then for each gender group test the null hypothesis of equal weights for students who did and did not eat breakfast, according to their survey responses. Make sure to formulate the correct alternative hypothesis.

Exercise: Read **Science lifts 'mummy's curse'** and use the means for Age at death, exposed vs. unexposed, along with the sample sizes n and standard deviations (in parentheses) to test for a significant difference in age at death between those who were and were not exposed to the "mummy's curse". State your conclusions clearly.