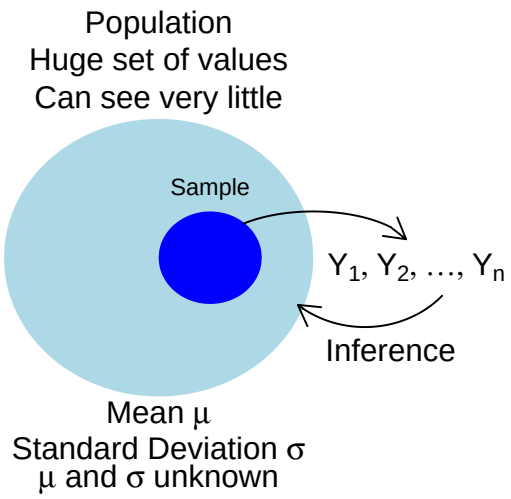
7 Hypothesis Testing in the One-Sample Situation

Suppose that you have identified a population with the goal of estimating the (unknown) **population mean** value, identified by μ. You select a random or representative sample from the population where, for notational convenience, the sample measurements are identified as $Y_1,\ Y_2,\ ...,\ Y_n$, where n is the sample size.



Given the data, our best guess, or estimate, of μ is the sample mean:

$$\bar{Y} = \frac{\sum_i Y_i}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}.$$

There are two main methods for inferences on μ: **confidence intervals** (CI) and **hypothesis tests**. The standard CI and test procedures are based on $\overline{Y}$ and s, the sample standard deviation. I discussed CIs in the last lecture.

Hypothesis Test for μ

Suppose you are interested in checking whether the population mean μ is equal to some prespecified value, say μ_0. This question can be formulated as a two-sided hypothesis test, where you are trying to decide which of two contradictory claims or hypotheses about μ is more reasonable given the observed data. The **null hypothesis**, or the hypothesis under test, is $H_0 : \mu = \mu_0$, whereas the **alternative hypothesis** is $H_A : \mu \neq \mu_0$.

I will explore the ideas behind hypothesis testing later. At this point, I focus on the mechanics behind the test. The steps in carrying out the test are:

1. Set up the null and alternative hypotheses: $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$, where μ_0 is specified by the context of the problem.

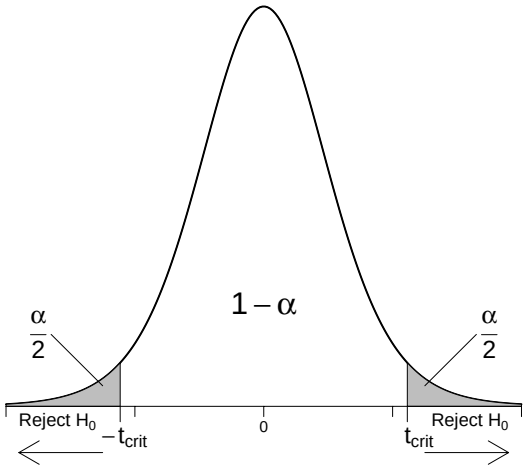
2. Choose the **size** or **significance level** of the test, denoted by α. In practice, α is set to a small value, say, .01 or .05, but theoretically can be any value between 0 and 1.

3. Compute the critical value t_{crit} from the $t-$distribution table with degrees of freedom $df = n - 1$. In terms of percentiles, $t_{crit} = t_{.5\alpha}$.

4. Compute the **test statistic**
$$t_s = \frac{\bar{X} - \mu_0}{SE},$$
where $SE = s/\sqrt{n}$ is the standard error.

5. **Reject** H_0 in favor of H_A (i.e. decide that H_0 is false, based on the data) if $|t_s| > t_{crit}$. Otherwise, do not reject H_0. An equivalent rule is to Reject H_0 if $t_s < -t_{crit}$ or if $t_s > t_{crit}$. I sometimes call the test statistic t_{obs} to emphasize that the computed value depends on the observed data.

The process is represented graphically below. The area under the $t-$probability curve outside $\pm t_{crit}$ is the size of the test, α. One-half α is the area in each tail. You reject H_0 in favor of H_A only if the test statistic is outside $\pm t_{crit}$.



Assumptions for Procedures

I described the classical $t-$test, which assumes that the data are a random sample from the population and that the population frequency curve is normal. The population frequency curve can be

viewed as a "smoothed histogram" created from the population data. You assess the reasonableness of the normality assumption using a stem-and-leaf, histogram, and a boxplot of the sample data. The stem-and-leaf and histogram should resemble a normal curve.

The $t-$test is known as a small sample procedure. For large samples, researchers sometimes use a $z-$test, which is a minor modification of the $t-$method. For the $z-$test, replace t_{crit} with a critical value z_{crit} from a standard normal table. The $z-$critical value can be obtained from the $t-$table using the $df = \infty$ row. The z-test does not require normality, but does require that the sample size n is large. In practice, most researchers just use the $t-$test whether or not n is large – it makes little difference since z and t are very close when n is large.

Example: Age at First Transplant

The ages (in years) at first transplant for a sample of 11 heart transplant patients are as follows: 54 42 51 54 49 56 33 58 54 64 49. The summaries for these data are: $n = 11, \overline{Y} = 51.27$, and $s = 8.26$. Test the hypothesis that the mean age at first transplant is 50. Use $\alpha = .05$. Also, find a 95% CI for the mean age at first transplant.

A good (necessary) first step is to define the population parameter in question, and to write down hypotheses symbolically. These steps help to avoid confusion. Let

$$\mu = \text{mean age at time of first transplant for population of patients.}$$

You are interested in testing $H_0 : \mu = 50$ against $H_A : \mu \neq 50$, so $\mu_0 = 50$.

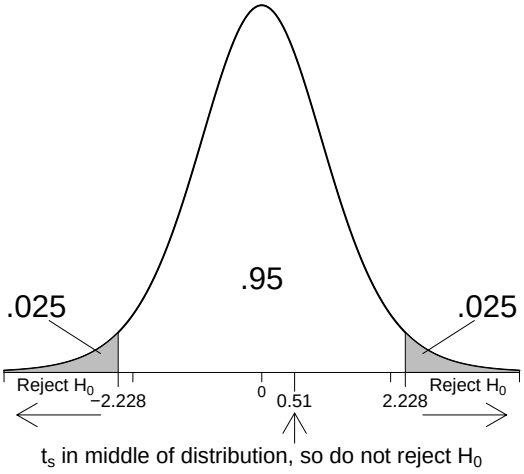
The degrees of freedom are $df = 11 - 1 = 10$. The critical value for a 5% test is $t_{crit} = t_{.025} = 2.228$. (Note $.5\alpha = .5 * .05 = .025$). The same critical value is used with the 95% CI.

Let us first look at the CI calculation. Here $SE = s/\sqrt{n} = 8.26/\sqrt{11} = 2.4904$ and $t_{crit} * SE = 2.228 * 2.4904 = 5.55$. The lower limit on the CI is $51.27 - 5.55 = 45.72$. The upper endpoint is $51.27 + 5.55 = 56.82$. Thus, you are 95% confident that the population mean age at first transplant is between 45.7 and 56.8 years (rounding to 1 decimal place).

For the test,

$$t = \frac{\bar{X} - \mu_0}{SE} = \frac{51.27 - 50}{2.4904} = 0.51.$$

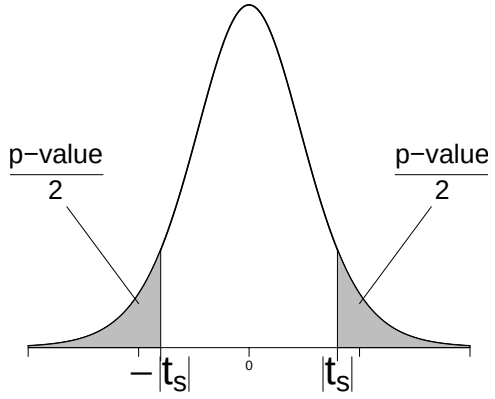
Since $t_{crit} = 2.228$, we do not reject H_0 using a 5% test. Note the placement of t relative to t_{crit} in the picture below. The results of the hypothesis test should not be surprising, since the CI tells you that 50 is a plausible value for the population mean age at transplant. Note: All you can say is that the data *could have* come from a distribution with a mean of 50 – this is not convincing evidence that μ actually *is* 50.



P-values

The **p-value**, or **observed significance level** for the test, provides a measure of plausibility for H_0. Smaller values of the p-value imply that H_0 is less plausible. To compute the p-value for a two-sided test, you

1. Compute the test statistic t_s as above.
2. Evaluate the area under the $t-$probability curve (with $df = n - 1$) outside $\pm|t_s|$.



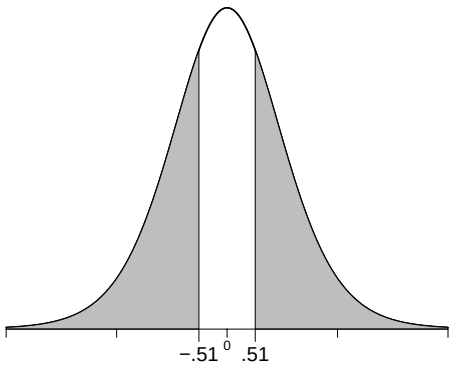
In the picture above, the p-value is the total shaded area, or twice the area in either tail. You can only get bounds on the p-value using the textbook's $t-$table.

Most, if not all, statistical packages, including **Minitab**, summarize hypothesis tests with a p-value, rather than a decision (i.e reject or not reject at a given α level). You can make a decision to reject or not reject H_0 for a size α test based on the p-value as follows - reject H_0 if the p-value is less than or equal to α. This decision is identical to that obtained following the formal rejection procedure given earlier. The reason for this is that the p-value can be interpreted as the smallest value you can set the size of the test and still reject H_0 given the observed data.

There are a lot of terms to keep straight here. α and t_{crit} are constants we choose (actually, one determines the other so we really only choose one, usually α) to set how rigorous evidence against H_0 needs to be. t_s and the p-value (again, one determines the other) are random variables because they are calculated from the random sample. They are the evidence against H_0.

Example: Age at First Transplant

The picture below is used to calculate the p-value. Using hte textbook's table, all we know is that the p-value is greater than .40. (Why?) The exact p-value for the test (generated with JMP-in) is 0.62. For a 5% test, the p-value indicates that you would not reject H_0 (because $.62 > .05$).



Minitab output for the heart transplant problem is given below. Let us look at the output and find all of the summaries we computed. Also, look at the graphical summaries to assess whether the $t-$test and CI are reasonable here.

COMMENTS:

1. The data were entered into the worksheet as a single column (C1) that was labelled **agetran**.
2. To display the data follow the sequence Data > Display Data, and fill in the dialog box.
3. To get the stem and leaf display, follow the sequence Graph > Stem and Leaf ..., then fill in the dialog box.

4. To get a one-sample t-test and CI follow the sequence: STAT > BASIC STATISTICS > 1-sample t... . In the dialog box, select the column to analyze (C1). For the test, you need to check the box for Perform Hypothesis Test and specify the null mean (i.e. μ_0) and the type of test (by clicking on OPTIONS): not equal gives a two-sided test (default), less than gives a lower one-sided test, and greater than gives an upper one-sided test. The results of the test are reported as a p-value. We have only discussed two-sided tests up to now. Click on the Graphs button and select Boxplot of data.

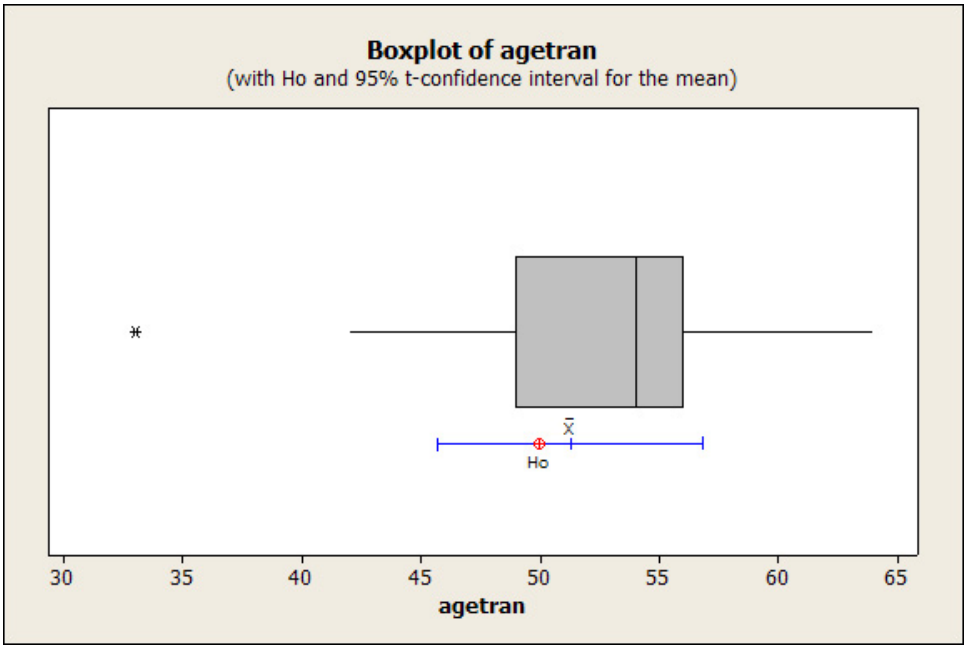
5. I would also follow Stat > Basic Statistics > Display Descriptive Statistics to get a few more summary statistics. The default from the test is a bit limited.

6. If you ask for a test, you will get a corresponding CI. The CI level is set by clicking on Option in the dialog box. If you want a CI but not a test, do not check Perform Hypothesis Test in the main dialog box. A 95% CI is the default.

7. The boxplot will include a CI for the mean.

8. The plots generated with Stat > Basic Statistics > Graphical Summary include a CI for the population mean.

```
Data Display

agetran
   33   42   49   49   51   54   54   54   56   58   64

Stem-and-Leaf Display: agetran

Stem-and-leaf of agetran  N  = 11
Leaf Unit = 1.0

 1   3  3
 1   3
 2   4  2
 4   4  99
(4)  5  1444
 3   5  68
 1   6  4

One-Sample T: agetran

Test of mu = 50 vs not = 50


Variable   N     Mean   StDev  SE Mean         95% CI           T      P
agetran   11  51.2727  8.2594   2.4903  (45.7240, 56.8215)  0.51  0.620

Descriptive Statistics: agetran

Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
agetran   11   0  51.27     2.49   8.26    33.00  49.00   54.00  56.00    64.00
```



Example: Meteorites

One theory of the formation of the solar system states that all solar system meteorites have the same evolutionary history and thus have the same cooling rates. By a delicate analysis based on measurements of phosphide crystal widths and phosphide-nickel content, the cooling rates, in degrees Celsius per million years, were determined for samples taken from meteorites named in the accompanying table after the places they were found.

Suppose that a hypothesis of solar evolution predicted a mean cooling rate of $\mu = .54$ degrees per million year for the Tocopilla meteorite. Do the observed cooling rates support this hypothesis? Test at the 5% level. The boxplot and stem and leaf display (given below) show good symmetry. The assumption of a normal distribution of observations basic to the $t-$test appears to be realistic.

Meteorite	Cooling rates
Walker County	0.69 0.23 0.10 0.03 0.56 0.10 0.01 0.02 0.04 0.22
Uwet	0.21 0.25 0.16 0.23 0.47 1.20 0.29 1.10 0.16
Tocopilla	5.60 2.70 6.20 2.90 1.50 4.00 4.30 3.00 3.60 2.40 6.70 3.80

Let

$$\mu = \text{mean cooling rate over all pieces of the Tocopilla meteorite.}$$

To answer the question of interest, we consider the test of $H_0 : \mu = .54$ against $H_A : \mu \neq .54$. I will explain later why these are the natural hypotheses here. Let us go carry out the test, compute the p-value, and calculate a 95% CI for μ. The sample summaries are $n = 12$, $\overline{Y} = 3.892$, $s = 1.583$. The standard error is $SE_{\overline{Y}} = s/\sqrt{n} = 0.457$.

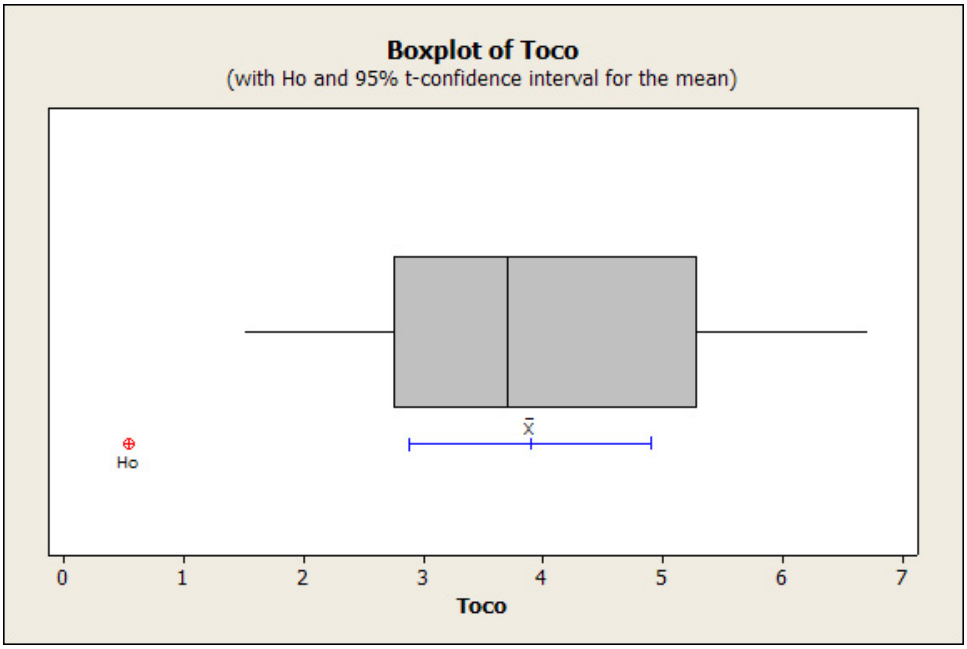
Minitab output for this problem is given below. For a 5% test (i.e. $\alpha = .05$), you would reject H_0 in favor of H_A because the $p - value \leq .05$. The data strongly suggest that $\mu \neq .54$. The 95% CI

says that you are 95% confident that the population mean cooling rate for the Tocopilla meteorite is between 2.89 and 4.90 degrees per million years. Note that the CI gives us a means to assess how different μ is from the hypothesized value of .54.

COMMENTS:

1. The data were entered as a single column in the worksheet, and labelled **Toco**.
2. Remember that you need to specify the null value for the mean (i.e. .54) in the 1-sample t dialog box!
3. I generated a boxplot within the 1-sample t dialog box. A 95% CI for the mean cooling rate is superimposed on the plots.

```
Data Display

Toco
   5.6   2.7   6.2   2.9   1.5   4.0   4.3   3.0   3.6   2.4   6.7   3.8

Stem-and-Leaf Display: Toco

Stem-and-leaf of Toco  N  = 12
Leaf Unit = 0.10

 1   1  5
 2   2  4
 4   2  79
 5   3  0
(2)  3  68
 5   4  03
 3   4
 3   5
 3   5  6
 2   6  2
 1   6  7

One-Sample T: Toco

Test of mu = 0.54 vs not = 0.54


Variable   N     Mean    StDev  SE Mean         95% CI           T      P
Toco      12  3.89167  1.58255  0.45684  (2.88616, 4.89717)  7.34  0.000

Descriptive Statistics: Toco

Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median    Q3  Maximum
Toco      12   0  3.892    0.457  1.583    1.500 2.750   3.700  5.275    6.700
```



The Mechanics of Setting up an Hypothesis Test

When setting up a test you should imagine you are the researcher conducting the experiment. In many studies, the researcher wishes to establish that there has been a change from the **status quo**, or that they have developed a method that produces a **change** (possibly in a specified direction) in the typical response. The researcher sets H_0 to be the **status quo** and H_A to be the **research hypothesis** - the claim the researcher wishes to make. In some studies you define the hypotheses so that H_A is the **take action** hypothesis - rejecting H_0 in favor of H_A leads one to take a radical action.

Some perspective on testing is gained by understanding the mechanics behind the tests. An hypothesis test is a decision process in the face of uncertainty. You are given data and asked which of two contradictory claims about a population parameter, say μ, is more reasonable. Two decisions are possible, but whether you make the correct decision depends on the true state of nature which is unknown to you.

Decision	If H_0 true	If H_A true
Reject H_0 in favor of H_A	Type I error	correct decision
Do not Reject [accept] H_0	correct decision	Type II error

For a given problem, only one of these errors is possible. For example, if H_0 is true you can make a Type I error but not a Type II error. Any reasonable decision rule based on the data that tells us when to reject H_0 and when to not reject H_0 will have a certain probability of making a Type I error if H_0 is true, and a corresponding probability of making a Type II error if H_0 is false and H_A is true. For a given decision rule, define

$$\alpha = \text{Prob(Reject } H_0 \text{ given } H_0 \text{ is true)} = \text{Prob(Type I error)}$$

and

$$\beta = \text{Prob(Do not reject } H_0 \text{ when } H_A \text{ true)} = \text{Prob(Type II error)}.$$

The mathematics behind hypothesis tests allows you to prespecify or control α. For a given α, the tests we use (typically) have the smallest possible value of β. Given the researcher can control α, you set up the hypotheses so that committing a Type I error is more serious than committing a Type II error. The magnitude of α, also called the **size** or **level** of the test, should depend on the seriousness of a Type I error in the given problem. The more serious the consequences of a Type I error, the smaller α should be. In practice α is often set to .10, .05, or .01, with $\alpha = .05$ being the scientific standard. By setting α to be a small value, you reject H_0 in favor of H_A only if the data **convincingly indicate** that H_0 is false.

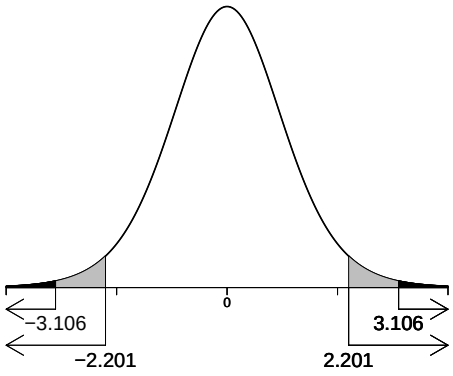
Let us piece together these ideas for the meteorite problem. Evolutionary history predicts $\mu = .54$. A scientist examining the validity of the theory is trying to decide whether $\mu = .54$ or $\mu \neq .54$. Good scientific practice dictates that rejecting another's claim when it is true is more serious than not being able to reject it when it is false. This is consistent with defining $H_0 : \mu = .54$ (the status quo) and $H_A : \mu \neq .54$. To convince yourself, note that the implications of a Type I error would be to claim the evolutionary theory is false when it is true, whereas a Type II error would correspond to not being able to refute the evolutionary theory when it is false. With this setup, the scientist will refute the theory only if the data overwhelmingly suggest that it is false.

The Effect of α on the Rejection Region of a Two-Sided Test

For a size α test, you reject $H_0 : \mu = \mu_0$ if

$$t_s = \frac{\overline{Y} - \mu_0}{SE}$$

satisfies $|t_s| > t_{crit}$.



Rejection Regions for .05 and .01 level tests

The critical value is computed so that the area under the $t-$probability curve (with $df = n-1$) outside $\pm t_{crit}$ is α, with $.5\alpha$ in each tail. Reducing α makes t_{crit} larger. That is, reducing the size of the test makes rejecting H_0 harder because the rejection region is smaller. A pictorial representation is given above for the Tocopilla data, where $\mu_0 = 0.54$, $n = 12$ and $df = 11$. Note that $t_{crit} = 2.201$ and 3.106 for $\alpha = 0.05$ and 0.01, respectively.

The mathematics behind the test presumes that H_0 is true. Given the data, you use

$$t_s = \frac{\bar{Y} - \mu_0}{SE}$$

to measure how far $\overline{Y}$ is from μ_0, relative to the spread in the data given by SE. For t_s to be in the rejection region, $\overline{Y}$ must be significantly above or below μ_0, relative to the spread in the data. To see this, note that rejection rule can be expressed as: **Reject** H_0 if

$$\overline{Y} < \mu_0 - t_{crit}SE \quad \text{or} \quad \overline{Y} > \mu_0 + t_{crit}SE.$$

The rejection rule is sensible because $\overline{Y}$ is our best guess for μ. You would reject $H_0 : \mu = \mu_0$ only if $\overline{Y}$ is so far from μ_0 that you would question the reasonableness of assuming $\mu = \mu_0$. How far $\overline{Y}$ must be from μ_0 before you reject H_0 depends on α (i.e. how willing you are to reject H_0 if it is true), and on the value of SE. For a given sample, reducing α forces $\overline{Y}$ to be further from μ_0 before you reject H_0. For a given value of α and s, increasing n allows smaller differences between $\overline{Y}$ and μ_0 to be **statistically significant** (i.e. lead to rejecting H_0). In problems where small differences between $\overline{Y}$ and μ_0 lead to rejecting H_0, you need to consider whether the observed differences are important.

In essence, the $t-$ distribution provides an objective way to calibrate whether the observed $\overline{Y}$ is typical of what sample means look like when sampling from a normal population where H_0 is true. If all other assumptions are satisfied, and $\overline{Y}$ is inordinately far from μ_0, then our only recourse is to conclude that H_0 must be incorrect.

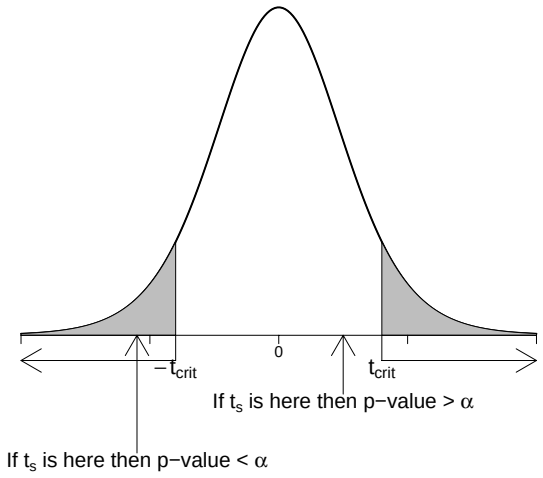
Two-Sided Tests, CI and P-Values

An important relationship among two-sided tests of $H_0 : \mu = \mu_0$, CI, and p-values is that

$$\text{size } \alpha \text{ test rejects } H_0 \Leftrightarrow 100(1-\alpha)\% \text{ CI does not contain } \mu_0 \Leftrightarrow p - value \leq \alpha.$$

$$\text{size } \alpha \text{ test does not reject } H_0 \Leftrightarrow 100(1-\alpha)\% \text{ CI contains } \mu_0 \Leftrightarrow p - value > \alpha.$$

For example, an $\alpha = .05$ test rejects $H_0 \Leftrightarrow 95\%$ CI does not contain $\mu_0 \Leftrightarrow p - value \leq .05$. The picture above illustrates the connection between p-values and rejection regions.



Either a CI or a test can be used to decide the plausibility of the claim that $\mu = \mu_0$. Typically, you use the test to answer the question **is there a difference?** If so, you use the CI to assess **how much of a difference exists**. I believe that scientists place too much emphasis on hypothesis testing. See the discussion below.

Statistical Versus Practical Significance

Suppose in the Tocopilla meteorite example, you rejected $H_0 : \mu = .54$ at the 5% level and found a 95% two-sided CI for μ to be .55 to .58. Although you have sufficient evidence to conclude that the population mean cooling rate μ differs from that suggested by evolutionary theory, the range of plausible values for μ is small and contains only values close to .54. Although you have shown statistical significance here, you need to ask ourselves whether the actual difference between μ and .54 is large enough to be important. The answer to such questions is always problem specific.

Design Issues and Power

An experiment may not be sensitive enough to pick up true differences. For example, in the Tocopilla meteorite example, suppose the true mean cooling rate is $\mu = 1.00$. To have a 50% chance of rejecting $H_0 : \mu = .54$, you would need about $n = 30$ observations. If the true mean is $\mu = .75$, you would need about 140 observations to have a 50% chance of rejecting H_0. In general, the smaller the difference between the true and hypothesized mean (relative to the spread in the population), the more data that is needed to reject H_0. If you have prior information on the expected difference between the true and hypothesized mean, you can design an experiment appropriately by choosing the sample size required to likely reject H_0.

The **power** of a test is the probability of rejecting H_0 when it is false. Equivalently,

power = 1 - Prob(not rejecting H_0 when it is false) = 1- Prob(Type II error).

For a given sample size, the tests I have discussed have maximum power (or smallest probability of a Type II error) among all tests with fixed size α. However, the actual power may be small, so sample size calculations, as briefly highlighted above, are important prior to collecting data. See your local statistician.

One-Sided Tests on μ

There are many studies where a one-sided test is appropriate. The two common scenarios are the **lower one-sided test** $H_0 : \mu = \mu_0$ (or $\mu \geq \mu_0$) versus $H_A : \mu < \mu_0$ and the **upper one-sided test** $H_0 : \mu = \mu_0$ (or $\mu \leq \mu_0$) versus $H_A : \mu > \mu_0$. Regardless of the alternative hypothesis, the tests are based on the t-statistic:

$$t_s = \frac{\overline{Y} - \mu_0}{SE}.$$

For the **upper one-sided test**

1. Compute the critical value t_{crit} such that the area under the t-curve to the **right** of t_{crit} is the desired size α, that is $t_{crit} = t_\alpha$.
2. Reject H_0 if and only if $t_s \geq t_{crit}$.
3. The p-value for the test is the area under the $t-$curve to the **right** of the test statistic t_s.

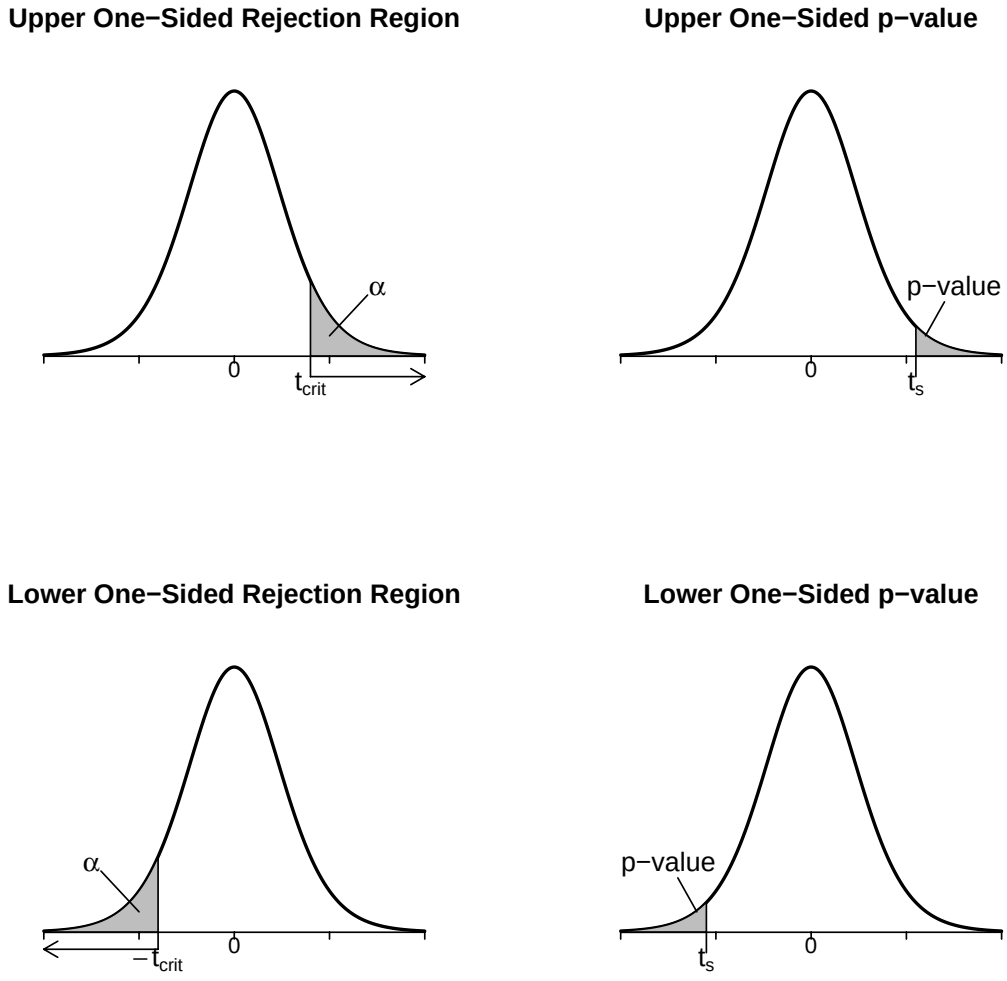
The **upper one-sided test** uses the **upper tail** of the $t-$ distribution for a rejection region. The p-value calculation reflects the form of the rejection region. You will reject H_0 only for large positive values of t_s which require $\overline{Y}$ to be significantly greater than μ_0. Does this make sense?

For the **lower one-sided test**

1. Compute the critical value t_{crit} such that the area under the t-curve to the **right** of t_{crit} is the desired size α, that is $t_{crit} = t_{alpha}$.
2. Reject H_0 if and only if $t_s \leq -t_{crit}$.
3. The p-value for the test is the area under the $t-$curve to the **left** of the test statistic t_s.

The **lower one-sided test** uses the **lower tail** of the $t-$ distribution for a rejection region. The calculation of the rejection region in terms of $-t_{crit}$ is awkward but is necessary for hand calculations because the textbook only give upper tail percentiles. Note that here you will reject H_0 only for large negative values of t_s which require $\overline{Y}$ to be significantly less than μ_0.

Pictures of the rejection region and the p-value evaluation for a lower one-sided test are given on the next page. As with two-sided tests, the p-value can be used to decide between rejecting or not rejecting H_0 for a test with a given size α.



Example: Weights of canned tomatoes

A consumer group suspects that the average weight of canned tomatoes being produced by a large cannery is less than the advertised weight of 20 ounces. To check their conjecture, the group purchases 14 cans of the canner's tomatoes from various grocery stores. The weights of the contents of the cans to the nearest half ounce were as follows: 20.5, 18.5, 20.0, 19.5, 19.5, 21.0, 17.5, 22.5, 20.0, 19.5, 18.5, 20.0, 18.0, 20.5. Do the data confirm the group's suspicions? Test at the 5% level.

Let $\mu =$ the population mean weight for advertised 20 ounce cans of tomatoes produced by the cannery. The company claims that $\mu = 20$, but the consumer group believes that $\mu < 20$. Hence the consumer group wishes to test $H_0 : \mu = 20$ (or $\mu \geq 20$) against $H_A : \mu < 20$. The consumer group will reject H_0 only if the data overwhelmingly suggest that H_0 is false.

You should assess the normality assumption prior to performing the $t-$test. The stem and leaf display and the boxplot suggest that the distribution might be slightly skewed to the left. However, the skewness is not severe and no outliers are present, so the normality assumption is not unreasonable.

Minitab output for the problem is given below. Let us do a hand calculation using the summarized data. The sample size, mean, and standard deviation are 14, 19.679, and 1.295, respectively. The standard error is $SE_{\overline{Y}} = s/\sqrt{n} = .346$. We see that the sample mean is less than 20. But is it sufficiently less than 20 for us to be willing to publicly refute the canner's claim? Let us carry out the test, first using the rejection region approach, and then by evaluating a p-value.

The test statistic is

$$t_s = \frac{\overline{Y} - \mu_0}{SE_{\overline{Y}}} = \frac{19.679 - 20}{.346} = -.93$$

The critical value for a 5% one-sided test is $t_{.05} = 1.771$, so we reject H_0 if $t_s < -1.771$ (you can get that value from Minitab or from the table). The test statistic is not in the rejection region. Using the t-table, the p-value is between .15 and .20. I will draw a picture to illustrate the critical region and p-value calculation. The exact p-value from Minitab is .185, which exceeds .05.

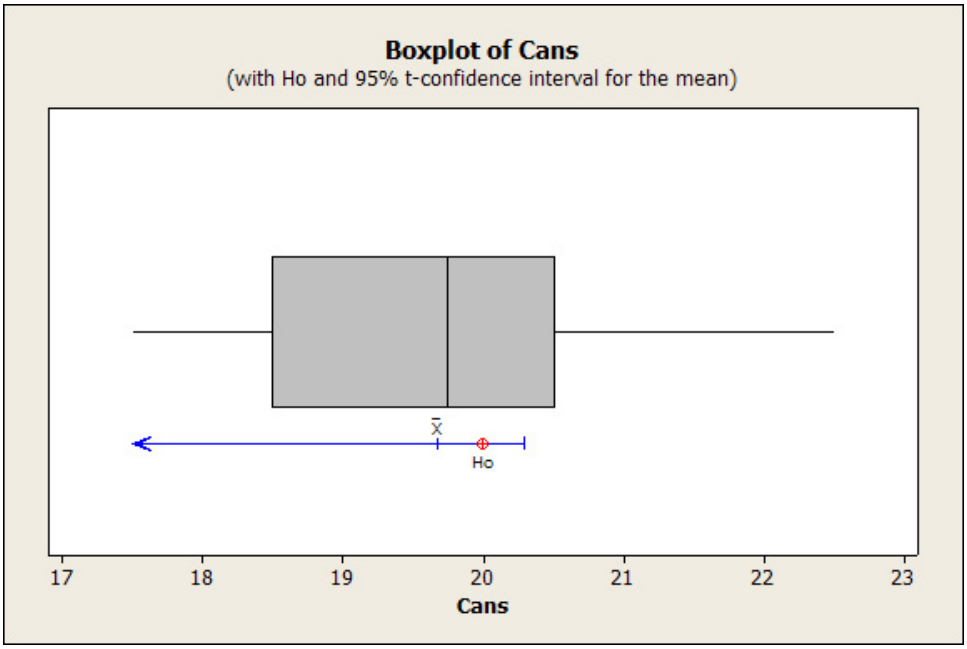
Both approaches lead to the conclusion that we do not have sufficient evidence to reject H_0. That is, we do not have sufficient evidence to question the accuracy of the canner's claim. If you did reject H_0, is there something about how the data were recorded that might make you uncomfortable about your conclusions?

COMMENTS:

1. The data are entered into the first column of the worksheet, which was labelled **cans**.
2. You need to remember to specify the lower one-sided test as an option in the 1 sample t-test dialog box.

```
Descriptive Statistics: Cans

Variable   N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
Cans      14   0  19.679    0.346  1.295   17.500  18.500  19.750  20.500

Variable  Maximum
Cans       22.500

Stem-and-Leaf Display: Cans

Stem-and-leaf of Cans  N  = 14
Leaf Unit = 0.10


 1  17  5
 2  18  0
 4  18  55
 4  19
 7  19  555
 7  20  000
 4  20  55
 2  21  0
 1  21
 1  22
 1  22  5
```

```
One-Sample T: Cans

Test of mu = 20 vs < 20

                                                    95%
                                                  Upper
Variable   N     Mean   StDev  SE Mean    Bound      T      P
Cans      14  19.6786  1.2951   0.3461  20.2915  -0.93  0.185
```



How should you couple a one-sided test with a CI procedure? For a **lower one-sided test**, you are interested only in an **upper bound** on μ. Similarly, with an **upper one-sided test** you are interested in a **lower bound** on μ. Computing these type of bounds maintains the consistency between tests and CI procedures. The general formulas for lower and upper $100(1-\alpha)\%$ confidence bounds on μ are given by

$$\overline{Y} - t_{crit} SE_{\overline{Y}} \qquad \text{and} \qquad \overline{Y} + t_{crit} SE_{\overline{Y}}$$

respectively, where $t_{crit} = t_\alpha$.

In the cannery problem, to get an upper 95% bound on μ, the critical value is the same as we used for the one-sided 5% test: $t_{.05} = 1.771$. The upper bound on μ is

$$\overline{Y} + t_{.05} SE_{\overline{Y}} = 19.679 + 1.771 * .346 = 19.679 + .613 = 20.292.$$

Thus, you are 95% confident that the population mean weight of the canner's 20oz cans of tomatoes is less than or equal to 20.29. As expected, this interval covers 20.

If you are doing a one-sided test in Minitab, it will generate the correct one-sided bound. That is, a lower one-sided test will generate an upper bound, whereas an upper one-sided test generates

a lower bound. If you only wish to compute a one-sided bound without doing a test, you need to specify the direction of the alternative which gives the type of bound you need. An upper bound was generated by Minitab as part of the test we performed earlier. The result agrees with the hand calculation.

Quite a few packages, including only slightly older versions of Minitab, do not directly compute one-sided bounds so you have to fudge a bit. In the cannery problem, to get an upper 95% bound on μ, you take the upper limit from a 90% two-sided confidence limit on μ. The rationale for this is that with the 90% two-sided CI, μ will fall above the upper limit 5% of the time and fall below the lower limit 5% of the time. Thus, you are 95% confident that μ falls below the upper limit of this interval, which gives us our one-sided bound. Here, you are 95% confident that the population mean weight of the canner's 20oz cans of tomatoes is less than or equal to 20.29, which agrees with our hand calculation.

```
One-Sample T: Cans

Variable   N     Mean   StDev  SE Mean         90% CI
Cans      14  19.6786  1.2951   0.3461  (19.0656, 20.2915)
```

The same logic applies if you want to generalize the one-sided confidence bounds to arbitrary confidence levels and to lower one-sided bounds - always double the error rate of the desired one-sided bound to get the error rate of the required two-sided interval! For example, if you want a lower 99% bound on μ (with a 1% error rate), use the lower limit on the 98% two-sided CI for μ (which has a 2% error rate).

Two-Sided Hypothesis Test for p

Suppose you are interested in whether the population proportion p is equal to a prespecified value, say p_0. This question can be formulated as a two-sided hypothesis test. To carry out the test:

1. Define the null hypothesis $H_0 : p = p_0$ and alternative hypothesis $H_A : p \neq p_0$.
2. Choose the size or significance level of the test, denoted by α.
3. Using the standard normal probability table, find the critical value z_{crit} such that the areas under the normal curve to the left and right of z_{crit} are $1 - .5\alpha$ and $.5\alpha$, respectively. That is, $z_{crit} = z_{.5\alpha}$.
4. Compute the test statistic (often to be labeled z_{obs})

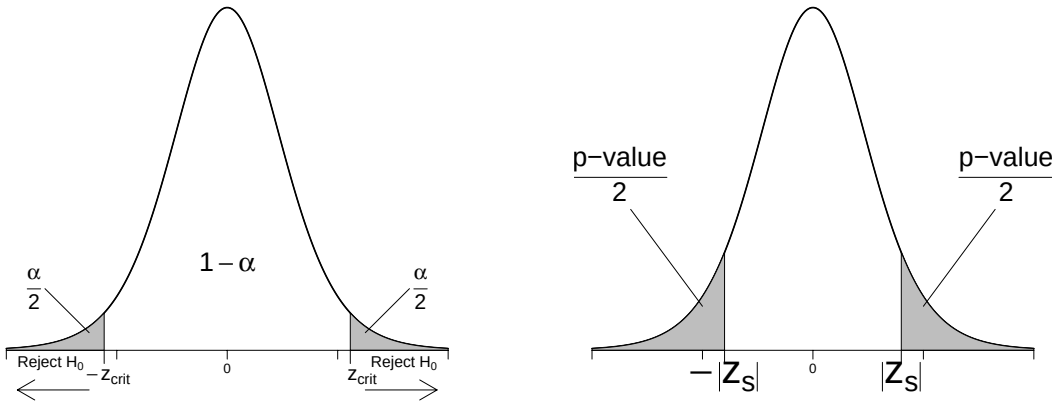
$$z_s = \frac{\hat{p} - p_0}{SE},$$

where the "test standard error" is

$$SE = \sqrt{\frac{p_0(1 - p_0)}{n}}.$$

5. Reject H_0 in favor of H_A if $|z_{obs}| \geq z_{crit}$. Otherwise, do not reject H_0.

The rejection rule is easily understood visually. The area under the normal curve outside $\pm z_{crit}$ is the size α of the test. One-half of α is the area in each tail. You reject H_0 in favor of H_A if the test statistic exceeds $\pm z_{crit}$. This occurs when $\hat{p}$ is significantly different from p_0, as measured by the standardized distance z_{obs} between $\hat{p}$ and p_0.



The P-Value for a Two-Sided Test

To compute the p-value (not to be confused with the value of p!) for a two-sided test:

1. Compute the test statistic z_s.

2. Evaluate the area under the normal probability curve outside $\pm z_s$.

Given the picture above with $z_{obs} > 0$, the p-value is the shaded area under the curve, or twice the area in either tail.

Recall that the null hypothesis for a size α test is rejected if and only if the p-value is less than or equal to α.

Example (Emissions data) Each car in the target population (L.A. county) either has been tampered with (a success) or has not been tampered with (a failure). Let p = the proportion of cars in L.A. county with tampered emissions control devices. You want to test $H_0 : p = .15$ against $H_A : p \neq .15$ (here $p_0 = .15$). The critical value for a two-sided test of size $\alpha = .05$ is $z_{crit} = 1.96$.

The data are a sample of $n = 200$ cars. The sample proportion of cars that have been tampered with is $\hat{p} = 21/200 = .105$. The test statistic is

$$z_s = \frac{.105 - .15}{.025} = -1.78,$$

where the test standard error satisfies

$$SE = \sqrt{\frac{.15 * .85}{200}} = .025.$$

Given that $|z_s| = 1.78 < 1.96$, you have insufficient evidence to reject H_0 at the 5% level. That is, you have insufficient evidence to conclude that the proportion of cars in L.A. county that have been tampered with differs from the statewide proportion.

This decision is reinforced by the p-value calculation. The p-value is the area under the standard normal curve outside ± 1.78. This is about $2 * .0375 = .075$, which exceeds the test size of .05.

REMARK: It is important to recognize that the mechanics of the test on proportions is similar to tests on means, except we use a different test statistic and a different probability table for critical values.

Appropriateness of Test

The z-test is based on a large sample normal approximation, which works better for a given sample size when p_0 is closer to .5. The sample size needed for an accurate approximation increases dramatically the closer p_0 gets to 0 or to 1. Unfortunately, there is no universal agreement as to when the sample size n is "large enough" to apply the test. A simple rule of thumb says that the test is appropriate when $np_0(1 - p_0) \geq 5$.

In the emissions example, $np_0(1 - p_0) = 200 * (.15) * (.85) = 25.5$ exceeds 5, so the normal approximation is appropriate.

Minitab Implementation

This is done precisely as in constructing CIs, covered last week. Follow `Stat > Basic Statistics > 1 Proportion` and enter summarized data. We are using the normal approximation for these calculations. You need to enter p_0 and make the test two-sided under `Options`.

```
Test and CI for One Proportion

Test of p = 0.15 vs p not = 0.15

Sample   X    N  Sample p         95% CI         Z-Value  P-Value
1       21  200  0.105000  (0.062515, 0.147485)    -1.78    0.075
```

My own preference for this particular problem would be to use the exact procedure. What we are doing here is the most common practice, however, and does fit better with procedures we do later. You should confirm that the exact procedure (not using the normal approximation) makes no difference here (because the normal approximation is appropriate).

One-Sided Tests and One-Sided Confidence Bounds

For one-sided tests on proportions, we follow the same general approach adopted with tests on means, except using a different test statistic and table for evaluation of critical values.

For an upper one-sided test $H_0 : p = p_0$ (or $p \leq p_0$) versus $H_A : p > p_0$, you reject H_0 when $\hat{p}$ is significantly greater than p_0, as measured by test statistic

$$z_s = \frac{\hat{p} - p_0}{SE}.$$

In particular, you reject H_0 when $z_s \geq z_{crit}$, where the area under the standard normal curve to the right of z_{crit} is α, the size of the test. That is $z_{crit} = z_\alpha$. The p-value calculation reflects the form of the rejection region, so the p-value for an upper one-sided test is the area under the $z-$curve to the right of z_s. The graphs on page 51 of the notes illustrated all this for the $t-$statistic; the picture here is the same except we now are using a z.

The lower tail of the normal distribution is used for the lower one-sided test $H_0 : p = p_0$ (or $p \geq p_0$) versus $H_A : p < p_0$. Thus, the p-value for this test is the area under the $z-$curve to the left of z_s. Similarly, you reject H_0 when $z_s \leq -z_{crit}$, where z_{crit} is the same critical value used for the upper one-sided test of size α.

Lower and upper one-sided $100(1-\alpha)\%$ confidence bounds for p are

$$\hat{p} - z_{crit}SE \quad \text{and} \quad \hat{p} + z_{crit}SE,$$

respectively, where $z_{crit} = z_\alpha$ is the critical value for a one-sided test of size α and $SE = \sqrt{\hat{p}(1-\hat{p})/n}$ is the "confidence interval" standard error. Recall that upper bounds are used in conjunction with lower one-sided tests and lower bounds are used with upper one-sided tests.

These are large sample tests and confidence bounds, so check whether n is large enough to apply these methods.

Example An article in the April 6, 1983 edition of *The Los Angeles Times* reported on a study of 53 learning impaired youngsters at the Massachusetts General Hospital. The right side of the brain was found to be larger than the left side in 22 of the children. The proportion of the general population with brains having larger right sides is known to be .25. Do the data provide strong evidence for concluding, as the article claims, that the proportion of learning impaired youngsters with brains having larger right sides exceeds the proportion in the general population?

I will answer this question by computing a p-value for a one-sided test. Let p be the population proportion of learning disabled children with brains having larger right sides. I am interested in testing $H_0 : p = .25$ against $H_A : p > .25$ (here $p_0 = .25$).

The proportion of children sampled with brains having larger right sides is $\hat{p} = 22/53 = .415$. The test statistic is

$$z_s = \frac{.415 - .25}{.0595} = 2.78,$$

where the test standard error satisfies

$$SE = \sqrt{\frac{.25 * .75}{53}} = .0595.$$

The p-value for an upper one-sided test is the area under the standard normal curve to the right of 2.78, which is approximately .003. I would reject H_0 in favor of H_A using any of the standard test levels, say .05 or .01. The newspaper's claim is reasonable.

A sensible next step in the analysis would be to compute a lower confidence bound for p. For illustration, consider a 95% bound. The CI standard error is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.415 * .585}{53}} = .0677.$$

The critical value for a one-sided 5% test is $z_{crit} = 1.645$, so a lower 95% bound on p is $.415 - 1.645 * .0677 = .304$. I am 95% confident that the population proportion of learning disabled children with brains having larger right sides is at least .304. Values of p smaller than .304 are not plausible.

You should verify that the sample size is sufficiently large to use the approximate methods in this example.

Minitab does this one sample procedure very easily, and it makes no real difference if you use the normal approximation or the exact procedure (what does that say about the normal approximation?).

```
Test of p = 0.25 vs p > 0.25


                                95%
                              Lower
Sample   X   N  Sample p     Bound  Z-Value  P-Value
1       22  53  0.415094  0.303766     2.78    0.003

Test and CI for One Proportion

Test of p = 0.25 vs p > 0.25


                                95%
                              Lower    Exact
Sample   X   N  Sample p     Bound  P-Value
1       22  53  0.415094  0.300302    0.006
```