17. Hypothesis Tests for Means

17.1. Example.

A health care worker conjectures that vitamin supplements given to expectant mothers will increase the birth weights of the resulting newborns. To test this, she randomly selects 100 expectant mothers and then randomly assigns them to one of two, equally sized groups. Expectant mothers in Group A receive supplements starting in the first trimester of their pregnancy, while expectant mothers in Group B receive a sugar pill. The researcher records the birth weights of the children for all 100 mothers.

Results. *In Group A, there were 51 children for whom the average birth weight was 3.55kg with a standard deviation of 0.62kg.*
In Group B, there were 50 children with an average birth weight of 3.39kg and a standard deviation of 0.22.
Research Question. *Is this significant evidence of an increase in birth weights?*
Note that there are two possible conclusions:
The supplements had the desired outcome of increasing birth weights,
OR
Small sample sizes and random differences in subjects created the appearance of a difference in outcomes where none really exists.
In hypothesis testing, the word "significant" is a technical term with a particular meaning that relates to these two contrary conclusions. We will return to this problem shortly and choose one of the above conclusions and reject the other.

17.2. General situation for hypothesis testing.

You will usually be testing to see if an experiment has generated sufficient evidence to support a conjecture. Your conjecture is called the

Alternative Hypothesis

Typically we will be testing to see if the mean in the experimental group differs from the mean in a control group. Sometimes the research will have access to retrospective historical data and then general population might serve as the control group.

- Control Population
- Mean μ_C not known, but is estimated by the sample mean of the control group $\bar{x}_C$.
- Not subject to the experimental treatment.
- "Null" since nothing is done to the group.

- Experimental Population
- Mean μ_E *not* known but is estimated by the sample mean $\bar{x}_E$ of the experimental group.
- Subject to the experimental treatment.

The untreated control population is generally the status quo. In principle, we could do a census and find the true mean μ_C for the control population. The treated experimental population is only an imagined future population since the experimental treatment has not yet actually been applied to the entire population.

If we actually knew μ_E (the true mean of the experimental population) there would be no need for a statistical test. Instead, all we know is an estimate of the true mean which is based on our sample. Since there are inherent uncertainties in sampling, we will need a way to decide if our results are due to the random error implicit in sampling or actually represent a consequence of the treatment.
Based on the sample mean, we want to decide whether

$$\mu_E = \mu_C \quad \text{or} \quad \begin{cases} \mu_E \neq \mu_C \\ \mu_E < \mu_C \\ \mu_E > \mu_C \end{cases}$$

If we decide $\mu_E = \mu_C$ then the experimental treatment made no difference. This is called the *Null Hypothesis*.

$$H_0 : \ \mu_E = \mu_C$$

Recall that μ_C is the mean of the control (untreated) population and μ_E is the mean of the experimental (treated) population. Instead of knowing the actual, census value for these two means, you will know estimates of $\bar{x}_c$ and $\bar{x}_E$.
Depending on the character of the experimenter's conjecture, the *Alternative Hypothesis* could be any one of the following:

$$H_A : \mu_E > \mu_C$$
$$H_A : \mu_E < \mu_C$$
$$H_A : \mu_E \neq \mu_C$$

The experimenter's conjecture will dictate which one of the above will be the alternative hypothesis. Only one of the above can be the alternative hypothesis.

Hypothesis tests are designed to give a systematic way to choose between the null and the alternative hypotheses. This is accomplished by controlling the likelihood of making an error. There are two kinds of error which are possible.

	H_0 *true*	H_0 *false*
Accept H_0	*OK*	*Type II Error*
Reject H_0	*Type I Error*	*OK*

Hypothesis tests are very conservative. You will change from the status quo (reject H_0) only if the evidence is overwhelming. In other words

You want to make sure you do *not* reject H_0 when H_0 is really true.

or, in terms of Type I and II error:

You want the chances of Type I Error to be as small as possible.

The *significance level* is the maximum chance of Type I error that the researcher is willing to accept. The statistical tests will produce an "observed" significance level, also called a p-value. If the p-value is less than the pre-set significance level, then the researcher rejects the null hypothesis. Otherwise, the researcher accepts the null hypothesis.
The researcher is technically free to decide what constitutes an unacceptably high chance of Type I Error, i.e., an unacceptably high p-value. However, there are some accepted standards. The researcher who deviates from those standards needs to be prepared to justify why the deviation is appropriate. The usual nomenclature is:

$p-$value $\leq 1\%$	Highly Significant results
$1\% < p-$value $\leq 5\%$	Significant results
$5\% < p-$value	Results are not significant

With this background, we can now return to the original example which started this section.

17.3. Example.

A health care worker conjectures that vitamin supplements given to expectant mothers will increase the birth weights of the resulting newborns. To test this, she randomly selects 100 expectant mothers and then randomly assigns them to one of two, equally sized groups. Expectant mothers in Group A receive supplements starting in the first trimester of their pregnancy, while expectant mothers in Group B receive a sugar pill. The researcher records the birth weights of the children for all 100 mothers.

Results. *In Group A, there were 51 children for whom the average birth weight was 3.55kg with a standard deviation of 0.62kg.*
In Group B, there were 50 children with an average birth weight of 3.39kg and a standard deviation of 0.22.

Research Question. *Does this provide evidence of a difference in birth weights between the treated and untreated groups, using a significance level of 4%?*

Solution.

Step 1. First make a list of all the relevant variables.

	Experimental	Control
sample mean	3.55	3.39
standard deviation	0.62	0.22
sample size n	51	50

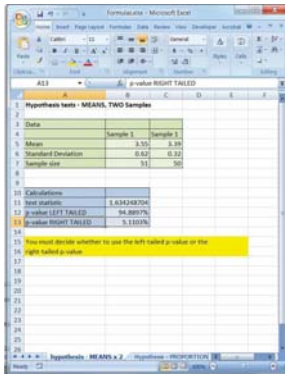
Remember that μ_C is the mean of the control population and μ_E is the mean of the experimental population.

Step 2. Next write down the null and alternative hypotheses:

$$H_0 : \mu_E = \mu_C$$
$$H_A : \mu_E > \mu_C$$

We use "$\mu_E > \mu_C$" since we conjecture that the treatment increases birth weights, and hence improves the health of the newborns.

Step 3. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled `hypothesis-MEANS x 2`. You should use `Sample 1` to record the experimental data and `Sample 2` to record the control data.



Step 4. The spreadsheet calculates two p-values: a "left-tailed" and a "right-tailed" value. If you glance back at the alternative hypothesis

$$H_A : \mu_E > \mu_C$$

you can see that the inequality is like an arrowhead that points to

the right. This indicates that you should use the right-tailed p-value–assuming that `Sample 1` is the experimental data and `Sample 2` is the control data.

Step 4. Thus, the p-value for this data is 5.11%. Since this exceeds the pre-set significance level of 4%, we are unable to reject the null hypothesis and must, therefore, accept the null hypothesis. Thus, at a significance level of 4%, we believe the null hypothesis. ∎

Question. Suppose that the significance level had been 5%? In this case, we would have **rejected** the null hypothesis, since the p-value is *less than* 5%. In this case, we would believe the *alternative hypothesis*, that the supplements do increase the birth weights. The p-value, 4.14%, is the chance that belief is wrong.

Solution Template

Step 1. Make a dictionary assigning values to each of the variables:

	Experimental	Control
sample mean	$\bar{x}_E$	$\bar{x}_C$
standard deviation	s_E	s_C
sample size	n_E	n_C
significance level	α	

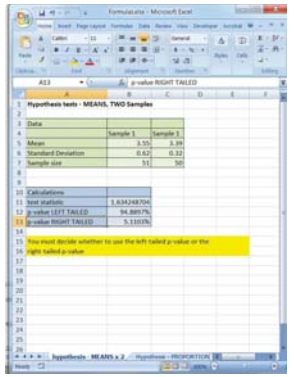
Step 2. Write down the null and alternative hypotheses. The null hypothesis will always be:

$$H_0: \ \mu_E = \mu_C$$

while the alternative hypothesis will be one of the following:

$$H_A: \ \mu_E < \mu_C \quad \textit{(a left tailed test)}$$
$$H_A: \ \mu_E > \mu_C \quad \textit{(a right tailed test)}$$
$$H_A: \ \mu_E \neq \mu_C \quad \textit{(a two tailed test)}$$

Step 3. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled `hypothesis-MEANS x 2`. You should use `Sample 1` to record the experimental data and `Sample 2` to record the control data.



Step 4. Use the form of the alternative hypothesis to select the appropriate p-value from the spreadsheet. If the pre-set significance level is *larger* than the observed p-value, then you can *reject* the null hypothesis, and the p-value represents the probability of a Type I Error. Otherwise, you *accept* the null hypothesis.

End of Solution Template

The spreadsheet calculates the value of a *test statistic*

$$\frac{\bar{x}_E - \bar{x}_C}{\sqrt{\frac{s_E^2}{n_E} + \frac{s_C^2}{n_C}}} \tag{1}$$

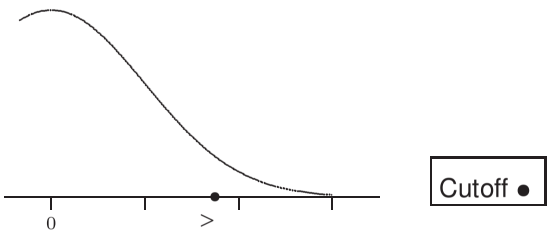
which is, in turn, used to calculate the p-value.
If it is reasonable to assume that the experimental and control populations have the same standard deviation, then the appropriate test statistic would use the pooled sample standard deviation s_{pooled}:

$$\frac{\bar{x}_E - \bar{x}_C}{s_{pooled}\sqrt{\frac{1}{n_E} + \frac{1}{n_C}}}$$

The spreadsheets for this course make the more general assumption that the experimental and control populations have different standard deviations and so formula (1) above.

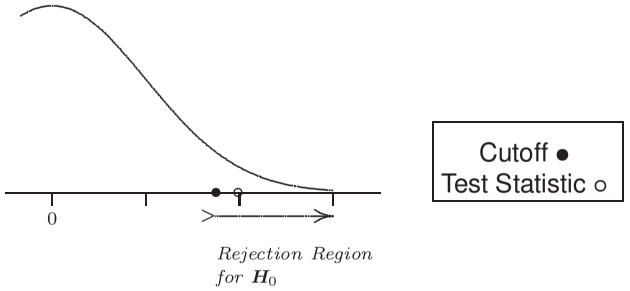
This test statistic is generally assumed to be normally distributed if the sample size is larger than 30; otherwise it has something called a Student's t-distribution (not covered in this course). The value of the test statistic determines the decision that you make. The rationale for the terminology "left-tailed" and "right-tailed" tests derives from the associated graphs of the distribution of the test statistic.

Suppose, for example, the pre-set significance level is 5%. One could do an inside-out problem to find the z-value associated with 5% of the area in the right-hand tail. If you then plotted that value on a number line, you'd get something like the following:



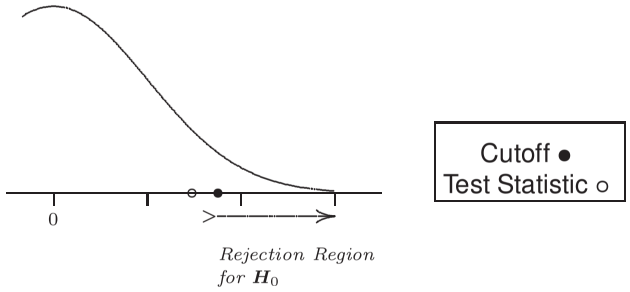
Notice that the area to the right of the solid dot ● is then 5%, or whatever the pre-set significance level happened to be.

Next, plot the calculated test statistic as a hollow dot (○). You might get the following



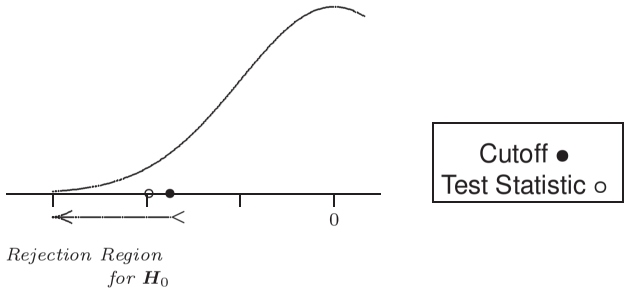
In this case, the test statistic is to the right of the calculated z-value for the pre-set significance level. The area to the right of the test statistic is *less* than the area to the right of the solid dot. Hence, we'd *accept* the null hypothesis.

If, in contrast, you got the following graph

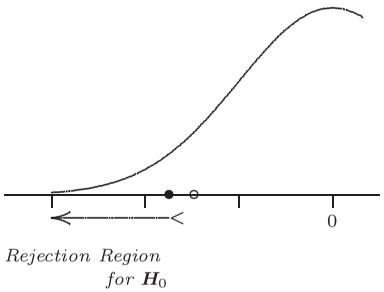


then you'd accept the null hypothesis.

A similar picture for left-tailed tests might be



in which case we'd reject the null hypothesis, or



in which case we'd accept the null hypothesis.
The point of the graphs is to illustrate how the spreadsheet calculations interact with the normal tables.

17.4. Example.

Surgery, planned or unplanned, subjects patients to great psychological stress. A counselor in a surgery ward suspects that patients who receive pre-operative briefings explaining to them exactly what to expect following surgery will experience less stress. To test this hypothesis, the researcher selects a random sample of 80 patients and randomly divides them into two treatment groups, each of size 40. In Group A, the patients receive a pre-operative briefing, while Group B patients receive no briefing.
Following surgery, the researcher administered the "State Anxiety Inventory" (SAI) was given to each patient, although five patients in Group A and two in Group B dropped out of the study. SAI scores range from 20 to 80 with lower scores indicating less anxiety. In Group A, the average anxiety score was 60.5 with a standard deviation of 5.6. In Group B, , the average anxiety score was 63.5, with a standard deviation of 8.3. Is this significant evidence (at the $\alpha = 4\%$ level) that pre-operative briefings reduce patient anxiety following surgery?

Solution.
Step 1. First make a list of all the relevant variables.

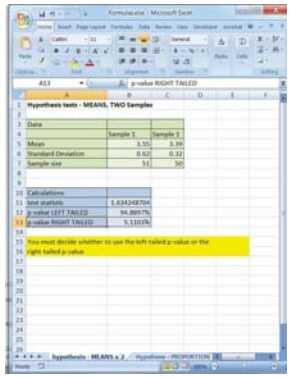
	Experimental	Control
sample mean	60.5	63.5
standard deviation	5.6	8.3
sample size	35	38
significance level	4%	

Step 2. Next write down the null and alternative hypotheses:

$$H_0 : \mu_E = \mu_C$$
$$H_A : \mu_E < \mu_C$$

We use "$<$" since we conjecture that anxiety levels will be *less* in the experimental, treated group.

Step 3. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled `hypothesis-MEANS x 2`. You should use `Sample 1` to record the experimental data and `Sample 2` to record the control data.



Step 4. Since this is a left-tailed test (we have a less-than sign in the alternative hypothesis), we can use the p-value from the spreadsheet is 3.42. Since this is *less than* the pre-set significance level of 4%, we reject the null hypothesis and believe the alternative, i.e., we believe that the briefings reduce anxiety. The chance of Type I Error is 3.42%, the

p-value.