NORTHWESTERN UNIVERSITY | SCHOOL OF CONTINUING STUDIES

Handout: Hypothesis Testing
PREDICT 401: Introduction to Statistical Analysis

Hypothesis Testing

Hypothesis testing uses the tools we have just learned to address questions such as:

1) How likely is the difference I observe to be "real" versus simply occurring by chance?
2) Is the treatment group actually different than the control group?
3) Did this sample come from this population?

What is a hypothesis test? It is a statistical test that allows us to determine whether a given hypothesis is true.

Example: My hypothesis is that a reduction in income tax will result in more savings in retirement accounts. We can formally test this hypothesis—which means that we can run a statistical test to determine how likely it is that reducing income tax actually results in more retirement savings. Typically, we gather data from a sample and try to determine how reflective the sample's behavior is of the population.

> Note: In this context, a hypothesis does not necessarily have to reflect your own beliefs. Rather, it is a construct that helps us conduct research. You don't have to believe a hypothesis (taking an aspirin each day will make your hair grow longer), or even really understand it (a daily dose of hemograxophilicane will increase one's meta-cognitive auxiliary functioning) to test it.

In formal hypothesis testing, we speak of two distinct types of hypotheses (both of which are about a population).

Alternative Hypothesis	*Null Hypothesis*
Often called the "research hypothesis"	Opposite of the research hypothesis
This is the one we favor / support	The one we hope to reject
We usually do NOT test this hypothesis	**This is the hypothesis we typically test**
Always has a sign with **inequality** ($\neq$, $>$, or $<$)	Always has a sign with equality: ($=$, $\leq$, or $\geq$)
Notation: H_a	Notation: H_o

For **every** research question, we can create an alternative and null hypothesis. This is **always** true, since **they are opposites**.

Examples:

Hypothesis 1: The incidence of lung cancer in men is different than the incidence of lung cancer in women.

Alternative Hypothesis: lung cancer incidence is different in men and women.
H_a: $LC_{men} \neq LC_{women}$

Null Hypothesis: Lung cancer incidence is the same in men and women.
H_o: $LC_{men} = LC_{women}$

Hypothesis 2: The mean achievement test score of urban students is lower than the mean of suburban students.

Alternative Hypothesis: Suburban students have a higher mean.
H_a: $\mu_{suburban} > \mu_{urban}$

Null Hypothesis: Suburban students do not have a higher mean.
H_o: $\mu_{suburban} \leq \mu_{urban}$

Hypothesis 3: The mean retirement savings for people with a tax cut will be more than the savings for people without a tax cut.

Alternative Hypothesis: The mean savings with tax cut will be higher.
H_a: $\mu_{tax\ cut} > \mu_{no\ tax\ cut}$

Null Hypothesis: The mean savings with tax cut will not be higher.
H_o: $\mu_{tax\ cut} \leq \mu_{no\ tax\ cut}$

"Direction" of Hypotheses
Hypotheses can have two "directional qualities":

1) 2-tailed hypotheses ("non-directional"): For instance: Is the population average amount of TV viewing **equal to** 12 hours? (This is non-directional because the opposite of "**equal**" can be either less than or greater than.)

2) 1-tailed hypotheses ("directional"): For instance: Is the population average amount of TV viewing **greater than or equal to** 12 hours? (This is "directional" because the opposite of **greater than or equal** is less than.)

© 2011 Northwestern University School of Continuing Studies

Direction is important. It profoundly influences the results of our hypothesis test.

Are the three hypotheses above non-directional or directional?

> **IMPORTANT**: We never "accept" or "prove" a hypothesis. We either "reject" or "fail to reject" a hypothesis. Typically, we test a null hypothesis and either reject it or fail to reject it. If we reject the null, then we might put some faith in our alternative hypothesis. If we fail to reject the null, then we might suspect the alternative hypothesis is wrong.

How to test a hypothesis:

1) Establish a null and alternative hypothesis
    a. State the question statistically
    b. State the opposite statistically
    c. Write out the null hypothesis
    d. Write out the alternative hypothesis
2) Choose a significance level (or region of rejection).
    a. You do this because you need to decide when you will reject the null hypothesis in favor of the alternative hypothesis. (The significance level, recall, is α.)
    b. For instance, a 5% significance level means that: **If there is less than a 5% probability that the sample mean (or proportion, difference, etc.) came from my hypothesized population mean, I will reject the null hypothesis (in favor of the alternative hypothesis).**
3) Choose a test statistic (z or t)
    a. Most common: **choose t** because
        i. If the sample were small (<30-ish), we would choose t.
        ii. If we didn't know the standard deviation of the population from which the sample is selected, we would choose t.
        iii. So, you would choose z only if the sample is large and we know the population standard deviation. But when the sample is large, t is pretty much equivalent to z. So, t is always a good bet.
4) Determine whether you much conduct a 1- or 2- tailed test.
5) Identify critical values (the value of t or z you've defined as your threshold)
6) Compute test statistic

    There are 2 test statistics of interest: the t-statistic and the z-statistic. We calculate them from the confidence interval equations you already know.

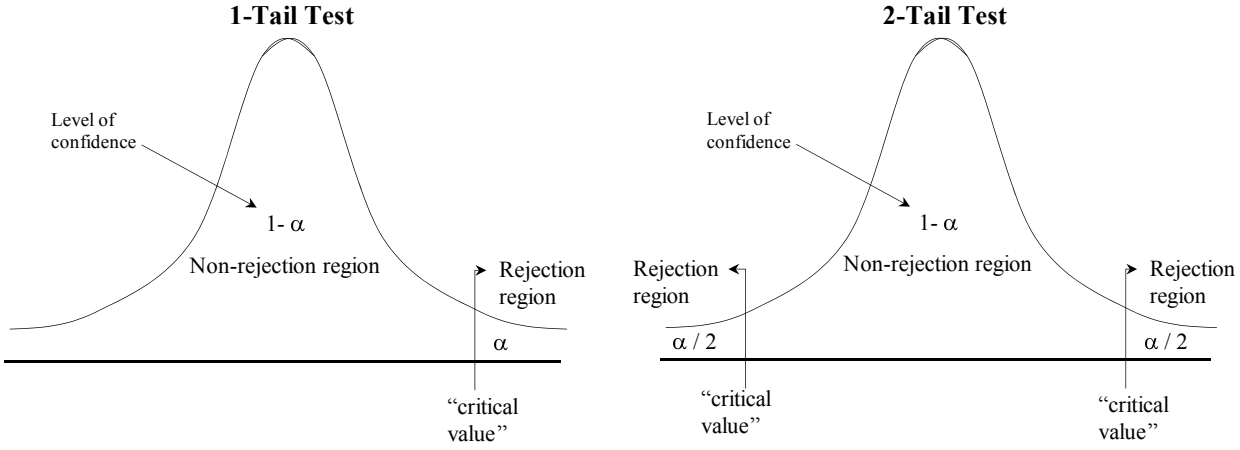
$$z = \frac{(\bar{x} - \mu)}{(\sigma/\sqrt{n})} \qquad t = \frac{(\bar{x} - \mu)}{(s/\sqrt{n})}$$

    where

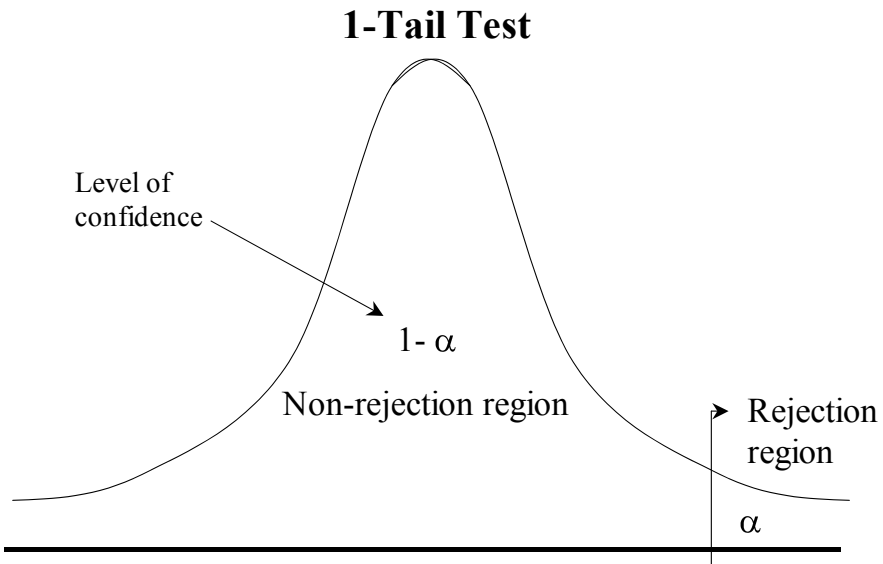
    $\bar{x}$ = sample mean
    μ = hypothesized value of the population we want to test
    σ = population standard deviation
    s = sample standard deviation
    n = sample size

7) Evaluate: determine whether the test statistic is within the acceptable range, as determined by your critical value.

Graphically:



Whether we need to do a 1- or 2-tailed test is very important. Why? Because if we want to be 90% sure of something, and we are doing a 1-tailed test, then all 10% of our allowed error will be in 1 tail. If we do a 2-tailed test, then the 10% will be divided between 2 tails (5% each). And, the values of t are different for .05 and .10.



If α= .01, then	z = 2.33
If α= .05, then	z = 1.65
If α= .10, then	z = 1.28

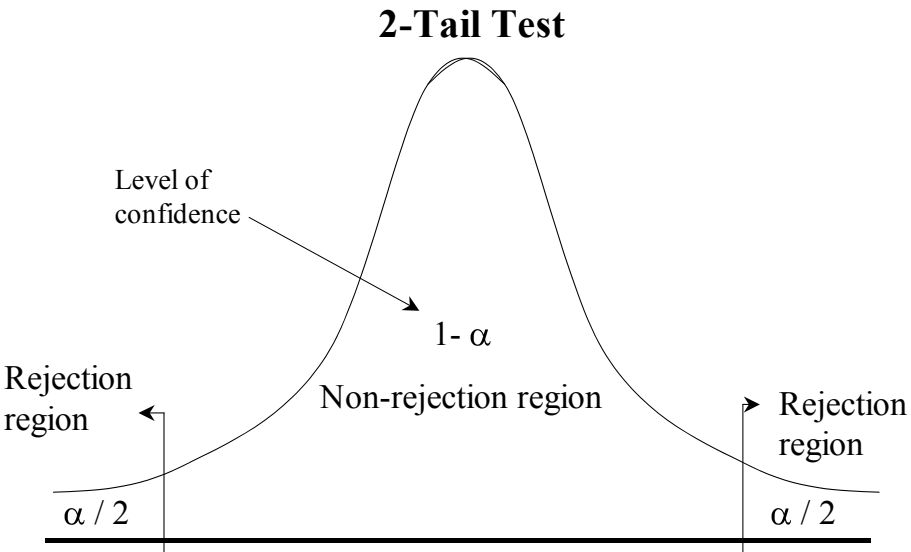
Here, the values of z are shown.

If t were used instead, you would need to know the size of the sample (*n*) in addition to the value of α.

Some values of t-critical

If α =	***n* = 10**	***n* = 61**	***n* = ∞**
0.01	t = 2.82	t = 2.39	t = 2.33
0.05	t = 1.83	t = 1.67	t = 1.64
0.10	t = 1.38	t = 1.30	t = 1.28

Note that this column is almost exactly the same as values of z.



z = -2.57	If α= .01, then	z = 2.57
z = -1.96	If α= .05, then	z = 1.96
z = -1.65	If α= .10, then	z = 1.65

Here, the values of z are shown.

If t were used instead, you would need to know the size of the sample (*n*) in addition to the value of α.

Some values of t-critical

If α =	**then α / 2 =**	***n* = 10**	***n* = 61**	***n* = ∞**
0.01	0.005	t = 3.25	t = 2.66	t = 2.58
0.05	0.25	t = 2.26	t = 2.00	t = 1.96
0.10	0.05	t = 1.83	t = 1.67	t = 1.64

Note that this column is almost exactly the same as values of z.

© 2011 Northwestern University School of Continuing Studies

Example of Hypothesis Testing:
A group of 100 individuals is given a battery of health tests, on which the maximum score is 100. You want to know whether the mean of all 100 (the population) is greater than 70. After administering the test, you look at six scores selected at random. The scores are 62, 92, 75, 68, 83, and 95. Can you be 95% sure that the group mean is greater than 70?

Step 1) State the null and alternative hypotheses:

The research question is: "Is the mean score of the population greater than 70?"

H_o: $\mu \leq 70$ H_a: $\mu > 70$.

Recall: you're going to **test the null** (meaning you'll see if you should reject it).

Step 2) Choose a significance level:

Recall: This is the probability of rejecting the null when it is really true. Here, we want to be 95% certain that the mean score is at least 70. So, $\alpha = .05$. Or, 5% of the time we will reject the null when it is really true. Here, that means that 5% of the time we will say that the population's mean really is greater than 70, when, in fact, it is 70 or less.

Step 3) Choose a test statistic:

A t-statistic is best for 2 reasons:

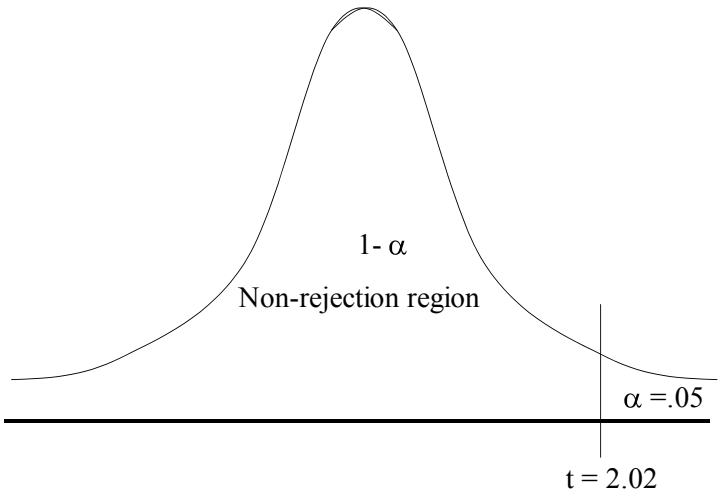
1) We do not know the population standard deviation
2) The sample size is small.

Step 4) Determine if this is a 1 or 2 tailed test:

This is 1-tailed because we are asserting that the mean is greater than some value.

Step 5) Identify Critical Value(s):

Here, t = 2.02 because $\alpha = .05$ and d.f. = 6-1 = 5.



© 2011 Northwestern University School of Continuing Studies

Step 6) Compute test statistic:

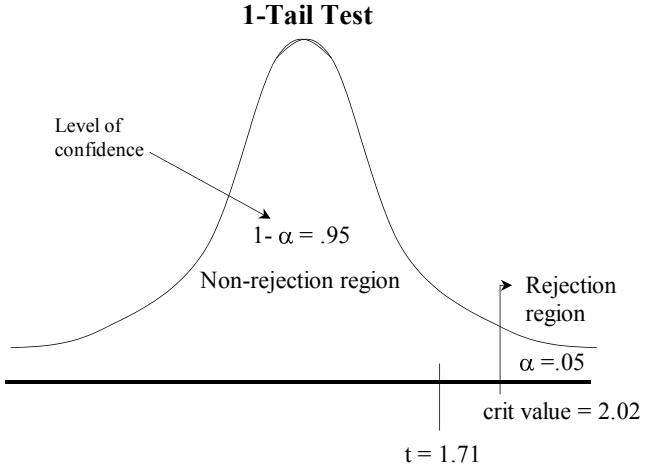
Using the data given, we find:

1) The sample mean is 475/6 = 79.17
2) The sample standard deviation = 13.17
3) Calculate t:

$$t = \frac{(\bar{x} - \mu)}{(s/\sqrt{n})} = (79.17 - 70) / (13.17/\sqrt{6}) = 1.71.$$

Step 7) Evaluate:

Recall our null hypothesis: The mean score is less than 70. Can we reject it? Well, if we are in the region of rejection, then yes. However, look back at our critical value (t = 2.02). The calculated value of t is 1.71. We are in the "non-rejection region" and thus, we **cannot** reject the null. In other words, we cannot reject the statement that the group's mean is less than or equal to 70. So are we 95% sure that that the population mean will be over 70? No.



What is the relationship between hypothesis testing and confidence intervals? Consider the following example:

Imagine you want to improve emergency vehicle response time in Chicago with a new citywide program meant to enhance communication. Before the program, the average response time was 5.00 minutes. To determine whether the program is successful, you take a random sample of 25 post-program response times, which have a mean of 4.5 minutes and a standard deviation of 2.00 minutes. For this determination, you can do a hypothesis test, which makes use of a confidence interval. The 90% confidence interval for the true post-program mean is

$$\bar{x} - (t_{\alpha/2,n-1} \cdot s/\sqrt{n}) \leq TruePopMean \leq \bar{x} + (t_{\alpha/2,n-1} \cdot s/\sqrt{n})$$

$$4.5 - (1.71*0.4) \leq TruePopMean \leq 4.5 + (1.71*0.4)$$

$$3.816 \leq \mu \leq 5.184$$

So, we are 90% confident that the true mean of the population of emergency vehicle response times after the program falls between 3.816 and 5.184 minutes.

Since the average before the program was 5.0, can we conclude that response times have decreased since the program?

Essentially, we are asking:
"Does getting an average of 4.50 minutes from a sample of 25 indicate that response times are lower since the program was instituted, or did I just get this sample mean by chance?"

You could look at the confidence interval, see that 5.0 falls within it, and say "Oh well. We can't conclude that response times are lower."

But, formal hypothesis tests are better methods of determining this (e.g., tests can be more flexible by allowing us to incorporate directionality and—in certain circumstances—formal tests can yield more and better information).

Set up the null: H_o: $\mu \geq 5.0$.
Set up alternative: H_a: $\mu < 5.0$.

If there is less than a 10% probability that this sample came from a population with a mean greater than or equal to 5.0, then I will reject the null hypothesis and feel comfortable in believing that response times have decreased.

© 2011 Northwestern University School of Continuing Studies

Understanding some statistical jargon:
If you test something at $\alpha = .05$ (be 95% confident), and reject the null, we say that your conclusion is statistically significant at the .05 level. Another way you see this is "$p \leq .05$." When you want to be 99% confident, then $\alpha = .01$, then statistical significance is often written "$p \leq .01$." Oftentimes, researchers will say something like "I tested the relationship between ____ and ____, and find a value of ____. This is significant at the .05 level." In a table of results, that value would probably have an asterisk and be labeled "$p \leq .05$."

© 2011 Northwestern University School of Continuing Studies

Review: For a given α, a higher value of the test statistic (typically *t*) means a greater chance that we will reject the null (which says that the sample tested is the same, and maybe lower/higher than the entire population). Specifically, what parameters influence the likelihood of rejecting the null? (Or, what variables are included in the calculation of that test statistic?)

- **The sample mean**. The "farther away" from the hypothesized mean in the null hypothesis, the more likely we are to reject the null. The closer to the hypothesized mean, the less sure we are that the groups are truly different. (*Ceteris paribus*, or, assuming everything else remains the same.)
- **The spread of the sample tested**. *Ceteris paribus*, the greater the spread, the less sure we are of the nature of the true population that is represented by the particular sample. Thus, the less likely we are to reject the null.
- **The number of observations** (people, things, etc.). *Cp*, the greater the number of observations, the more sure we are of the nature of the group that the sample represents. Thus, the greater the number, the more likely we are to reject the null.

© 2011 Northwestern University School of Continuing Studies

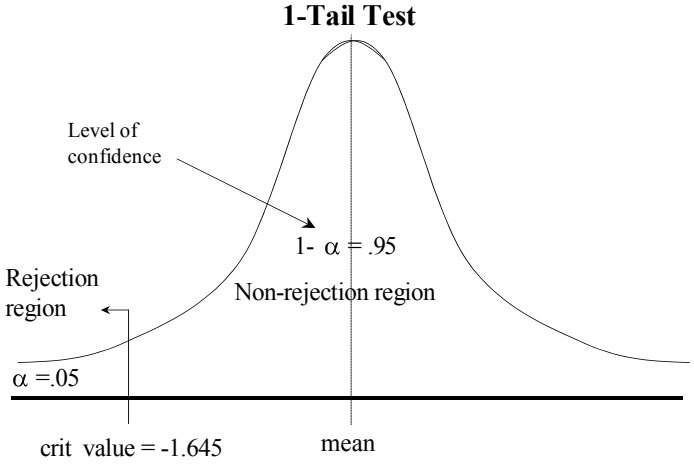
Another example of hypothesis testing:
Suppose that an agency implements a new method of processing welfare applications. Prior to the new method, the mean number of weeks to process an application was four. In a random sample of 250 applicants under the new method, the mean number of weeks is 3.6, with a standard deviation of 1.5. Does the new method result in a decrease in processing time?

Think it through first. You have a mean of 3.6 weeks for 250 people, who have a standard deviation of 1.5 weeks. What do you think? Are you convinced that the new method reduces the time from four weeks? What would convince you more? (Consider the sample mean, the number of people, and the sample standard deviation.)

H_o: $\mu \geq 4$ weeks
H_a: $\mu < 4$ weeks

Decision rule: If there is less than a 5% probability that this sample mean comes from a population with a mean time of 4 weeks (or more), then we reject the null.

This is a 1-tailed test, since direction matters.



(Note: Why is the critical value on the left (negative side)?)

Note that the distribution drawn above is the distribution AS IF the value in the "equal part" of the null were correct. For this particular context, the above distribution has a mean of 4. If the test statistic falls far to the left, then we think that this distribution we've drawn is not a good representation of the population from which the sample was drawn. In other words, we think that the population from which this sample was drawn is not equal to 4.

So, if the test statistic associated with 3.6 is more than 1.645 standard errors away from the hypothesized population mean of 4, then there is less than a 5% probability that this sample mean would have come from a population where the true mean was 4. Then, we would reject the null.

Calculate our test stat: $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ = (3.6 – 4) / (1.5 /√250) = -4.22.

Since this falls in the rejection region (less than -1.645), then we reject the null. We think that it is very unlikely that this group of 250 actually came from a population where the mean was 4.0 weeks. In other words, we know that if we did have a population where the mean was 4 weeks, it would be very, very unlikely for us to randomly select 250 whose applications would get processed in an average of 3.6 weeks and a standard deviation of 1.5 weeks.

Imagine that we only selected a sample of nine people, though, not 250. Would this make it more likely that they would have a mean of 3.6 and standard deviation of 1.5? Absolutely:

© 2011 Northwestern University School of Continuing Studies

Test stat: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ = (3.6 – 4) / (1.5 /√9) = -0.8. See? It is in the ***non-rejection*** region.

In other words, because it is only nine people, we just do not have enough observations to be sure (given the values of all the other variables).

© 2011 Northwestern University School of Continuing Studies

So far, we have run hypothesis tests for one particular situation—when we see a sample mean and we want to know what it says about the population.

But, we can do lots of other things, too. For instance, we can test the hypothesis that the means between two populations is different, based on samples from both populations.

When we test differences in means, we do everything the same, but now the equation for the test statistic is different:

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

Yet another example of hypothesis testing:
Imagine you want to know if the weight of babies of mothers in Medicaid is different than the weight of babies of mothers in special insurance programs. You have collected data on the mean birth weight (in grams) for 1074 babies. You sort the data on whether the mothers were covered by Medicaid, whether they participated in a special insurance program, or whether they were uninsured at the time of birth.

	Mean weight	*SD*	*Sample Size*
Uninsured	3210	626	461
Medicaid	3154	666	504
Special Program	3318	604	109

Ideally, we would like to know whether the insurance program *causes* differences in weight. The first question we should ask is whether there actually *is* a true difference in weight.
Set up alternative and null hypotheses:
Null = H_o: $\mu_{SP} = \mu_{Medicaid}$
Alt = H_a: $\mu_{SP} \neq \mu_{Medicaid}$

Let's say we want to be 95% confident. This is a 2-tailed test, so our critical values are 1.96 and -1.96. If our test statistic falls between these, then we cannot reject the null.

Note that since our null is that the two means are equal, the value of $\mu_1=\mu_2$. So,
$t = (3318-3154) / \sqrt{[(666^2/504)+(604^2/109)]}$
$t = 2.52$.

Since this is in the rejection region, we reject the null hypothesis that the means are equal. Thus, we strongly suspect that those whose mothers on Medicaid are not the same weight as those whose mothers were in special programs.

© 2011 Northwestern University School of Continuing Studies

One last point on hypothesis testing: Type I and Type II Error

Can your hypothesis test lead to an wrong conclusion?

Possibility #1) No—you are infallible. Congratulations.

Possibility #2) Yes—and you should know how to handle and discuss the possibility of errors.

Two types of error can occur in hypothesis testing:

Type I error: You already know about Type I error. It is called α. This means that α of the time (say, 5%) we will reject the null hypothesis when, in fact, it is correct. You, the researcher, control the level of this error typically by determining the level of error you are living to live with (and how informative you want your test to be).

Type II error: This is the opposite. When you make Type II error, you fail to reject the null when you should have.

Visualizing this in a table can help:

		"The Truth"	
		H_o	H_a
Our Action:	Fail to reject H_o	$1 - \alpha$	β (Type II)
	Reject H_o	α (Type 1)	$1 - \beta$

Note: α and β are probabilities, so all values in table are between (and including) 0 and 1.

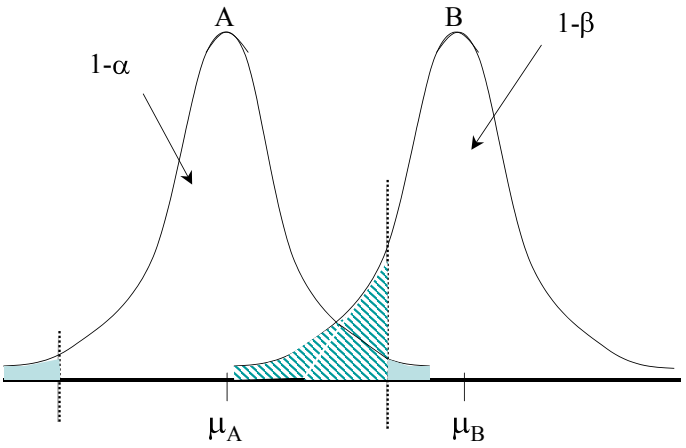
Example:
Imagine we initiate a new state nurse compensation program to increase retention.
We want to know whether the program works, so we set up:
$\mathbf{H_a}$: the program works (retention after the program is greater than before)
$\mathbf{H_o}$: the program doesn't work (retention after the program is equal to or less than retention before the program)
Let $\alpha = .05$.

If the program doesn't work ($\mathbf{H_o}$ is true):
Then 95% (or, 1-α) of the time we will not reject the null. Since the program does not work, we'll be right in not rejecting the null.
But, 5% of the time we will reject the null and say the program did work. Oops.

If the program does work ($\mathbf{H_a}$ is true):
Then 1-β of the time, we will reject the null. This is good, since the null is wrong.
But, β of the time, we won't reject the null, even though the null is wrong. Oh, well.

© 2011 Northwestern University School of Continuing Studies

The graphical version on Type I and Type II error:



Consider just distribution A first:
Imagine we have a null hypothesis that the true population mean is μ_A and we have set up a significance level of α. We can test the null hypothesis, and, as usual, if our sample value falls into the small shaded regions above, then we reject the null. We think that the sample comes from a different population, but α of the time, we will be wrong. If we are wrong, we have made a Type I error. (We have rejected the null when the null was correct.) In other words, we think our sample does not belong to population A in the above diagram when, in fact, it does.

Where does distribution B come into play?
By rejecting the null, we are saying that we think the sample belongs to another population, like B shown above.

If, in fact, the sample does come from population B, then what happens if our sample's test statistic falls into the area above that is shaded with diagonal lines? We would fail to reject the null, right? (Because we are in the non-rejection region for population A, or "inside" the critical values.) Since we are in the non-rejection region, we conclude that our sample **does** belong to population A. But the diagonally shaded area shows us that sometimes we'll be wrong—we can make a Type II error by not rejecting the null when we should have.

Definition of ***Power:***

Power = 1 - β = probability of rejecting H_o when H_a is true.

Note the tradeoff between Type I and Type II error. As Type I error gets small, Type II error gets big. Researchers must decide which one they are most willing to put up with and then set the research parameters appropriately.

If you must err, is it better to make Type I or Type II error?
It really depends on the context, your goals, and lots of other things. If you are researching whether a program has an effect, then you have to ask yourself: "Would I rather say that a program has no effect when it has an effect or say that it has an effect when there's really no effect?"

If the treatment has a big down side (like bad side effects, or a very high cost of implementation), then you might need to be very sure that there is an effect before you implement it. In other words, you want to make sure that you don't accidentally say "okay, let's implement the program" when the truth is that it has no effect. In this case, you'd want to minimize Type I error (or make α very low). So, imagine you are a college administrator and you want to implement a very expensive program you have heard about that can reduce the amount of underage drinking on campus. Using the money for this purpose takes away money from hiring more qualified instructors and is very controversial in the local community. You would really hate to spend highly valuable resources on the program and have it not work.

© 2011 Northwestern University School of Continuing Studies

But, if the potential benefits significantly outweigh the potential detriments, then you might not want to make the mistake of not doing it if there is an effect. Here, you'd want to minimize Type II error and insist on very high power. Imagine that you are in charge of trying to control a very virulent and deadly disease in a large urban area. You might want to make sure that you administer a relatively new (untested) drug if there is even the slightest chance that it will work. So, you make sure that you do not conclude that the drug does not work when it actually does work.

© 2011 Northwestern University School of Continuing Studies