

**Prof. Matthew Lebo**  
**POL 602**

## **Hypothesis Testing**

As emphasized all along, the objective of statistics is to make inferences about unknown population parameters using information from sample data. Up until now, the only type of inference we have made is the estimate—using sample data to construct point or interval estimates of parameters of interest.

Now I want to introduce another type of inference, *hypothesis testing*, or tests of hypotheses that our theories give us about the particular value of parameters.

The text compares hypothesis testing to the scientific method in general terms. Of course, hypothesis testing is only one part of the overall method, but it is the most important part of a deductive philosophy of science. The basic steps are:

1. Develop a hypothesis (derive from deductive theory) concerning the values of one or more population parameters (e.g. the median income of Republicans is larger than that of Democrats; the population variance of school budgets is greater than 30,000 dollars, etc).
2. Sample the population
3. Evaluate the hypothesis using the observed data

Obviously, for social scientists, the procedure is one of the central objects of our work. I should point out, however, that although we will devote a large proportion of the total class time to what is called “classical,” or “Neyman-Pearson” hypothesis testing, much of the testing that goes on in the discipline is but loosely related to the strict procedures we will discuss. Note that a very general question concerning this topic appeared on the 2001 Methods Qualifying Exam here in our department.

The text introduces four primary elements of a statistical test of a hypothesis.

The first two elements are related to the *research hypothesis* itself. For example, suppose we believe that an increase in federal budget outlay on airport security will decrease the frequency of terrorist incidents. In classical hypothesis testing, we partition this research hypothesis into two elements:

1. The *null hypothesis* ( $H_0$ ); and
2. The *alternative hypothesis* ( $H_a$ ) – aka the research hypothesis.

In our example the research hypothesis, that the increase in spending will decrease the frequency of attacks, is re-cast as the alternative hypothesis. The null hypothesis (or null) is simply the *complement* (not the opposite) of  $H_a$ .

In this case, the null hypothesis might be that increasing spending has no effect on the rate of terrorist strikes. Thus, of the null and alternative hypothesis, *one and only one must be true*.

This distinction is central to classical hypothesis testing. If, after we conclude our analysis, we think that we were correct, we do not say that we have supported our research or alternative hypothesis—instead we say that we have “rejected the null hypothesis.”

How do we test the alternative against the null?

The next tool we need is where all of the estimation and statistical theory we have studied comes in:

### 3. The *test statistic*

The test statistic, like an estimator, is a function of the observed sample data upon which the statistical decision will be made. In particular, if the value of the test statistic for our problem falls in...

### 4. The *rejection region*

then we can reject the null hypothesis. If the computed value falls outside, then we *fail to reject* the null (we never support the null hypothesis).

## Large Sample Hypothesis Tests

Just as the central limit theorem made it easy for us to calculate confidence intervals for large samples, so can it be used in the hypothesis testing of large sample means, proportions, and differences of means and proportions.

Suppose we want to test a set of hypotheses concerning a parameter  $\theta$  based on a random sample  $Y_1, Y_2, \dots, Y_n$ . In large samples, as we have discussed, our estimator  $\hat{\theta}$  has an approximately normal sampling distribution with mean  $\theta$  and standard error  $\sigma_{\hat{\theta}}$ .

If  $\theta_0$  is a specific value of  $\theta$ , we may wish to test  $H_0 : \theta = \theta_0$  against the alternative  $H_a : \theta > \theta_0$ . If  $\hat{\theta}$  is close to  $\theta_0$ , it seems reasonable that we would fail to reject the null that the actual population parameter is equal to  $\theta_0$ .

Values of  $\hat{\theta}$  much larger than  $\theta_0$ , however, lead us to reject the null in favor of the alternative hypothesis. In terms of our four elements, this test looks like:

$$H_0 : \theta = \theta_0$$

$$H_a : \theta > \theta_0$$

Test Statistic:  $\hat{\theta}$

RR:  $\{\hat{\theta} > k\}$  for some  $k$ .

How do we choose  $k$ , i.e., how far is far enough away to reject the null? We find  $k$  by fixing the acceptable type I error probability and using what we know about the sampling distribution to find  $k$ .

Note that a Type I error is rejecting the null hypothesis when it is in fact true – more in a few pages.

If  $H_0 : \theta = \theta_0$  is true, then  $\hat{\theta}$  has an approximately normal sampling distribution with mean  $\theta$  and standard error  $\sigma_{\hat{\theta}}$ . Therefore, for an  $\alpha$  – level test,

$$k = \theta_0 + z_{\alpha} \sigma_{\hat{\theta}}$$

thus the rejection region (where we reject the null in favor of our hypothesis):

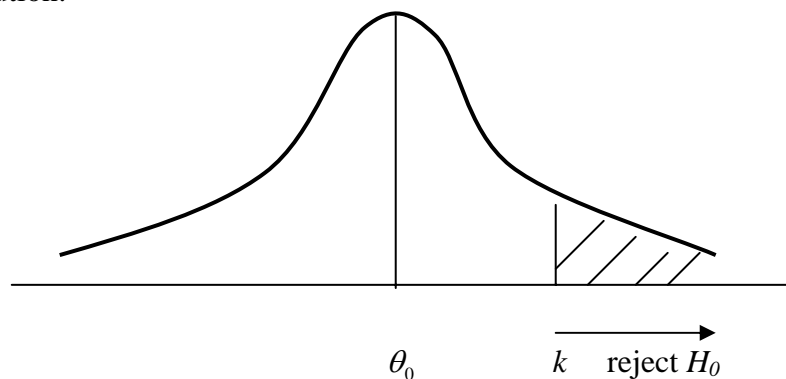
$$RR = \left\{ \hat{\theta} : \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} > z_{\alpha} \right\}$$

From our discussion of point estimates, our test statistic is just a standardized score:

$$Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

or, the estimate of the parameter minus the value of the parameter if the null hypothesis is true, all divided by the standard error of the estimator.

Graphically, we reject the null when our estimate is at least  $k$  away from the mean of the sampling distribution:



Another way to think about this is that, since we are dealing in this example with an unbounded continuous random variable for our estimate, it could in theory take on any value.

However, by picking a type I error rate, we are saying essentially that if it is at least  $k$ , we are comfortable assuming that the actual parameter value is greater than  $\theta_0$ .

In this case, we thus reject the null hypothesis if the test statistic falls far enough into the *upper tail* of the standard normal distribution. Our alternative hypothesis  $H_a : \theta > \theta_0$  is called an *upper tail* alternative and the  $RR = \{\hat{\theta} : z > z_\alpha\}$  is an *upper-tail rejection region*.

It is also possible to test *lower-tail* hypotheses ( $H_a : \theta < \theta_0$ ) as well as *two-tailed* hypotheses ( $H_a : \theta \neq \theta_0$ ). The choice of the correct hypothesis type is given to us by the theory we are investigating.

In sum, for large sample  $\alpha$  – level hypothesis tests:

$$\begin{aligned}
 &H_0 : \theta = \theta_0 \\
 &H_a : \begin{cases} \theta > \theta_0, \text{ upper tail} \\ \theta < \theta_0, \text{ lower tail} \\ \theta \neq \theta_0, \text{ two-tailed} \end{cases} \\
 &\text{Test Statistic: } Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} \\
 &RR : \begin{cases} \{z > z_\alpha\}, \text{ upper} \\ \{z < -z_\alpha\}, \text{ lower} \\ \{|z| > z_{\alpha/2}\}, \text{ two-tailed} \end{cases}
 \end{aligned}$$

Example (10.5): VP of sales complains that salespeople are averaging no more than 15 contracts a week. To challenge this claim, you take a sample of 36 salespeople and find their average number of contracts is 17 with a variance in sample of 9. Using a test at the  $\alpha = .05$  level, does this evidence contradict the VP's claim?

Solution: Our hypothesis is that the VP is wrong and that salespeople are doing better than 15 contracts/week. In other words:

$$H_0 : \mu = 15 \quad H_a : \mu > 15$$

From the point estimator table, we know that when  $n$  is greater than about 30, a good estimator of  $\mu$  is  $\bar{Y}$ , which has a normal sampling distribution with  $\mu_{\bar{Y}} = \mu, \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$ .

Our test statistic is thus:

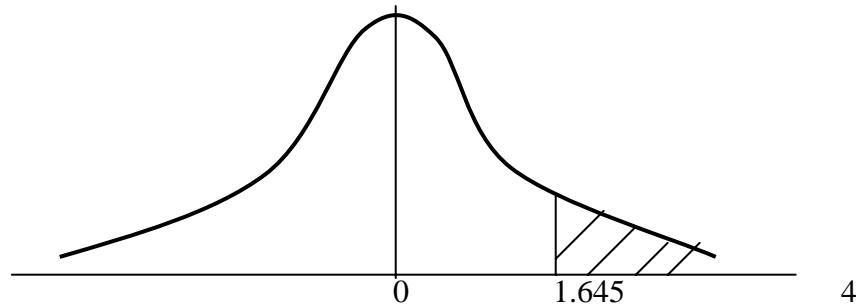
$$Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} = \frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{\sigma / \sqrt{n}}$$

Once again, we must estimate the population standard deviation with the sample data we have. Thus:

$$Z = \frac{17 - 15}{3 / \sqrt{36}} = 4$$

Our rejection region is  $\{z > z_{.05}\} = \{z > 1.645\}$  from the table (backwards). Since  $4 > 1.645$ , we reject the null that the mean number of contracts is 15 in favor of our alternative that the mean is actually larger than 15. Formally, we say that *at the  $\alpha = .05$  level of significance*, the evidence indicates that the VP's claim is incorrect.

Graphically,



Example (10.6): A machine in a factory must be repaired if it produces more than 10% defective items out of all the items produced in a day. We take a random sample of 100 items and find 15 defective; the line supervisor says we need to repair the machine. Does the evidence support this conclusion at the .01 level?

Solution: Assuming that the probability of each item being defective is equal and independent, we can model this process using a Binomial random variable. In this case:

$$H_0 : p = .10 \quad H_a : p > .10$$

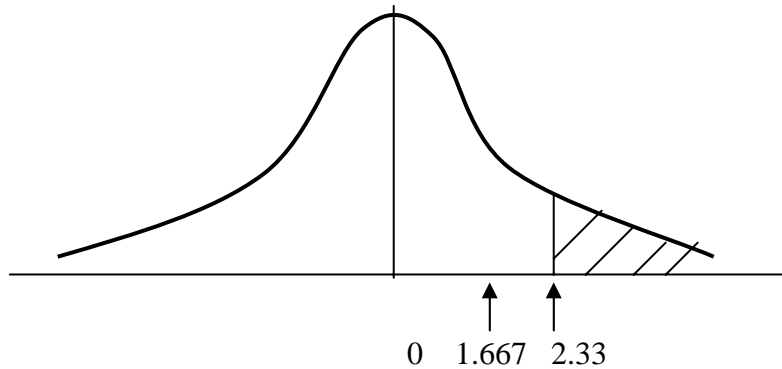
The test statistic is our estimator for a proportion:

$$Z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

notice in this step that we **do not** use the in-sample proportion when calculating the standard error (as we have up until this point). In this case, we use instead the null value of  $p$ , .10:

$$Z = \frac{.15 - .10}{\sqrt{\frac{.10(1-.10)}{100}}} = \frac{5}{3} \approx 1.667$$

Checking our table to find the rejection region, we get  $RR = \{z > 2.33\}$ . Our estimate is thus not in the rejection region. Our evidence does not support the supervisor's conclusion at the .01 level. Again, graphically:



Example (10.7): A psychological study compares the reaction times of men and women to a certain stimulus. A random sample of 50 men had a mean response of 3.6 seconds with a variance of .18. An independent random sample of 50 women had a mean of 3.8 with variance .14. Is there a difference between the times of men and women in the population at the .05 level of significance?

Solution: since we have no *a priori* belief that either sex is faster, we use a two-tailed test with:

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0$$

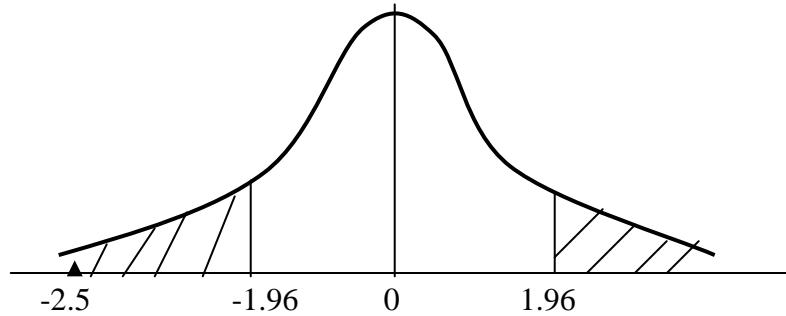
Our test statistic will be the point estimate of the difference of two means, with one simple modification:

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

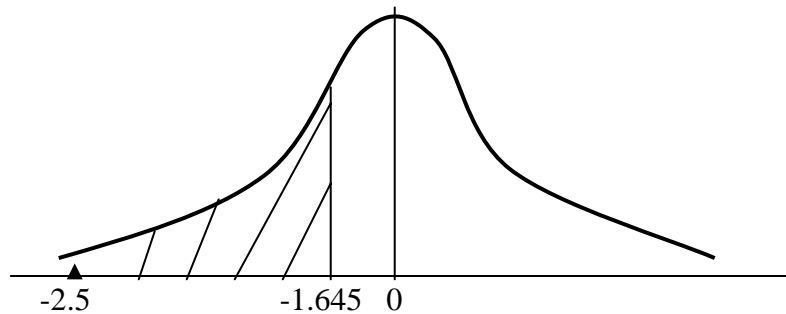
where  $D_0$  is the difference in means in the null case (0 in our example, but I include this for your general use in other situations). Once again using the sample variances for estimates of the population values, we find:

$$Z = \frac{(3.6 - 3.8) - 0}{\sqrt{\frac{.18}{50} + \frac{.14}{50}}} = -2.5$$

What is our rejection region to compare this to? Recall  $RR_{two-tailed} = \{|z| > z_{\alpha/2}\}$ , so in this example,  $z_{\alpha/2} = z_{0.025} = 1.96$ . Thus our test statistic falls within the rejection region and we can say that men are different from women at the .05 level (or at least their reaction times are). Graphically:



Note that if we did have *a priori* knowledge or theory that the times of men should be lower than those of women, our RR would have been from  $-1.645$  to negative infinity. That is, it would have been the same size overall, but larger in the direction we were testing:



A two-tailed test is thus more conservative, as it assumes no prior knowledge. The default setting for most statistical software packages is the two-tailed test. Remember, however, that if your theory leads you to expect a certain direction for your hypothesis, you can use a one-tailed test.

### Error



Since any test statistic is a function of random variables, it is itself a random variable. Because of this, as we would expect, hypothesis testing is a probabilistic process and thus errors can occur.

In particular, what we call a *type I error* (or  $\alpha$ , commonly referred to as the *level* of the test) occurs when we mistakenly reject the null hypothesis when it is, in fact, true.

A *type II error* (or  $\beta$ ) occurs when we fail to reject the null even though it is actually not true (like a “false negative” in biomedical research or clinical medicine).

It is important to note that there is almost always a trade-off between these two error types (a fact often forgotten in the rush to decrease  $\alpha$ ).

These two error types are important to distinguish, and you will be hearing about them again and again in this course and beyond. (The 2005 methods qualifier had a question on them.) Diagrammatically:

	Fail to Reject	Reject
Null is True		Type I Error  $\alpha$
Null is Not True	Type II Error  $\beta$	

Example (10.1 adapted): You work for a political candidate worried about his chances of electoral victory. In order to put his fears to rest, you decide to test the hypothesis that the probability of a given voter supporting him is less than 0.5 (remember in a large population that this is equivalent to the proportion being less than half—i.e. loss at the polls).

What are our formal hypotheses?

$$H_a : p < .5 \quad H_0 : p \geq .5$$

What is a possible test statistic? We choose  $Y$ , the number of “successes” or voters that we sample who favor our candidate (we could also think about this equivalent in terms of the proportion of the sample favoring).

What about the rejection region (RR)? How few voters should be found to favor our candidate before we can safely reject the null hypothesis?

Let us assume that money is tight in campaign headquarters, and we can only afford a telephone sample of 15 voters. If we select  $RR = \{y \leq 2\}$  (or, equivalently for any sample size,  $p \leq .1\bar{3}$ ), what are our associated  $\alpha, \beta$ ? (In other words, if we choose only to reject

the null when 2 or fewer voters surveyed say they will vote for our candidate, what are our probabilities of incorrectly assuming we will lose and incorrectly assuming that we will win?)

$$\begin{aligned}
 \alpha &= P(\text{type I error}) \\
 &= P(\text{rejecting } H_0 \text{ when true}) \\
 &= P(\text{test statistic is in RR when } H_0 \text{ is true}) \\
 &= P(Y \leq 2 \mid p = .5) \\
 &= \sum_{y=0}^2 \binom{15}{y} (.5)^y (.5)^{15-y} \\
 &= \binom{15}{0} (.5)^{15} + \binom{15}{1} (.5)^{15} + \binom{15}{2} (.5)^{15} = .004
 \end{aligned}$$

This seems to be a small risk: our RR will incorrectly reject the null when it is true in only 4 out of 1000 samples given a large number of repeated identical and independent samples.

Of course, as we have observed in the past, in statistics nothing is free. What about type II error in this case?

This is not as easy. Why not?

First, Type II error probabilities are not set by the experimenter.

Second, each hypothesis test has an infinite number of Type II error probabilities – one for each value of the parameter admissible under the alternative hypothesis.

So we need to make an additional assumption – choose one value for the alternative hypothesis – to compute this value.

In this case, we'll use what we think is the worst-case proportion of the population supports our candidate. Let us assume that this  $p = .3$ . Then,

$$\begin{aligned}
 \beta &= P(\text{type II error}) \\
 &= P(\text{failing to reject } H_0 \text{ when false}) \\
 &= P(\text{test statistic is not in RR when } H_0 \text{ is false}) \\
 &= P(Y > 2 \mid p = .3) \\
 &= \sum_{y=3}^{15} \binom{15}{y} (.3)^y (.7)^{15-y} \\
 &= .873
 \end{aligned}$$

In other words, in 873/1000 repeated samples, our test will incorrectly conclude that the proportion of voters in the population supporting our candidate is greater than 0.5 even if it is as low as 0.3! That is, since our RR is so small, our test fails to reject the null that  $p = .5$  (or is less than .5) even when only 3/15 respondents say they will vote for the candidate.

How can we fix this? As I noted before, there is an implicit tradeoff between type I and II errors. We can see this if, in this example, we enlarge the rejection region to decrease our type II error. A more logical choice of RR would be  $\{y \leq .5\}$ , since this in this region the candidate will lose the election (and the null will be rejected).

What is the new type II error, again assuming an actual  $p = .3$ ?

$$\begin{aligned}
 \beta &= P(\text{type II error}) \\
 &= P(\text{failing to reject } H_0 \text{ when false}) \\
 &= P(\text{test statistic is not in RR when } H_0 \text{ is false}) \\
 &= P(Y > 5 \mid p = .3) \\
 &= \sum_{y=6}^{15} \binom{15}{y} (.3)^y (.7)^{15-y} \\
 &= .278
 \end{aligned}$$

which is a more acceptable probability of mistakenly failing to reject the null. What, however, is the effect of this new RR on type I error?

$$\begin{aligned}
 \alpha &= P(\text{type I error}) \\
 &= P(\text{rejecting } H_0 \text{ when true}) \\
 &= P(\text{test statistic is in RR when } H_0 \text{ is true}) \\
 &= P(Y \leq 5 \mid p = .5) \\
 &= \sum_{y=0}^5 \binom{15}{y} (.5)^y (.5)^{15-y} \\
 &= .151
 \end{aligned}$$

which is not so good (this would be generally unacceptable in applied research, although there are no hard and fast rules, as we shall discuss later).

Is there any way out of this tradeoff? The only possible solution is easy to see but hard to implement in many cases: *increase n*.

Only by using more information about our population can we reliably lower both type I and II error simultaneously.

### Type II Error in Large Samples

While finding the type II error can be problematic in many cases, once again the central limit theorem makes it easy in large samples.

Just as in the example above, when calculating type II errors, the researcher must make a “what if” assumption (above, it was that the worst case  $p$  was .3 in the Jones election). In the notation below, theta subscripted by  $a$  denotes this alternative assumption.

The logic is straightforward. Take, for example, an upper-tail test. Recall that the rejection region is:

$$RR = \{\hat{\theta} : \hat{\theta} > k\}$$

and, by definition, that type II error is:

$$\beta = P(\hat{\theta} \text{ not in RR when } H_0 \text{ is false})$$

which is equivalent to:

$$\begin{aligned}\beta &= P(\hat{\theta} \leq k \mid \theta = \theta_a) \\ &= P\left(\frac{\hat{\theta} - \theta_a}{\sigma_{\hat{\theta}}} \leq \frac{k - \theta_a}{\sigma_{\hat{\theta}}} \mid \theta = \theta_a\right)\end{aligned}$$

where the second equation is just the standardized version of the first.

We can thus find the probability of a type II error by using this equation and the standard normal tables. Question: how does sample size effect this equation? [For large sample estimators, as  $n$  increases, the standard error, and thus beta, decreases].

Example (10.8): Let us return to the case of the VP of sales to illustrate this point. Recall that this VP believes the mean number of sales calls to be no greater than 15 per week on average. We took a sample of 36 salespeople and found a mean of 17 with a variance of 9, and tested his hypothesis at the .05 level.

Suppose the VP now wants to know how effective this test would be at detecting a difference in one call per week from the null. That is, he wants to test the null hypothesis  $H_0 : \mu = 15$  against the alternative  $H_a : \mu = 16$ . What is the probability of a type II error in this new test?

The first step is to calculate the rejection region of the .05 level upper-tail test:

$$z = \frac{\bar{y} - \mu_0}{\sigma / \sqrt{n}} > 1.645$$

or

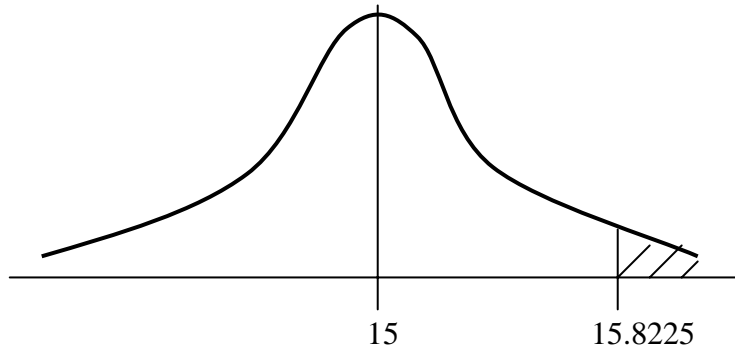
$$\bar{y} > \mu_0 + 1.645 \left( \frac{\sigma}{\sqrt{n}} \right)$$

substituting:

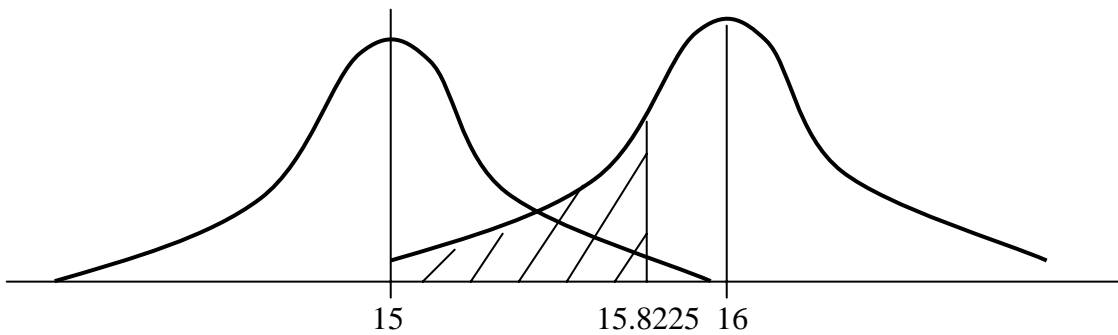
$$\bar{y} > 15 + 1.645 \left( \frac{3}{6} \right)$$

$$\bar{y} > 15.8225$$

This rejection region is thus:



We, however, are interested in the probability that we fail to detect that the actual population mean is 16 when we test at the .05 level. Graphically, then, we are interested in:



which is equal to:

$$\begin{aligned} \beta &= P \left( \frac{\bar{Y} - \mu_a}{\sigma / \sqrt{n}} \leq \frac{k - \mu_a}{\sigma / \sqrt{n}} \right) \\ &= P \left( Z \leq \frac{15.8225 - 16}{3 / \sqrt{36}} \right) \\ &= P(Z \leq -.36) = .3594 \end{aligned}$$

using the table in the regular, forward, direction and using the standard assumption that the sample variance is an adequate estimate for the population variance.

The interpretation of this result is that, if we are interested in the hypothesis that the mean is just one unit greater than the null mean, an  $n$  of 36 is insufficient to detect this difference reliably. What would a better  $n$  be?

The text derives a useful equation to answer this problem for any one-tailed test. The two-tailed case requires additional assumptions and will not be discussed (the experimental design course I believe will cover this).

Suppose we have an upper-tail hypothesis:

$$H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0$$

and we wish to specify acceptable values of  $\alpha$  and  $\beta$ , where  $\beta$  is evaluated for the case when

$$\mu = \mu_a, \mu_a > \mu_0$$

our solution is write out the equations for  $\alpha$  and  $\beta$  and solve them simultaneously for  $n$ . We know that:

$$\alpha = P(Z > z_\alpha) = P\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} > \frac{k - \mu_0}{\sigma/\sqrt{n}} \text{ when } \mu = \mu_0\right)$$

and

$$\beta = P(Z \leq -z_\beta) = P\left(\frac{\bar{Y} - \mu_a}{\sigma/\sqrt{n}} \leq \frac{k - \mu_a}{\sigma/\sqrt{n}} \text{ when } \mu = \mu_a\right)$$

setting these equations equal and solving for  $n$  yields:

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2}$$

which is equally true in the lower-tail case.

Example: (10.9) returning again to the sales VP, suppose he wants to test

$$H_0 : \mu = 15 \quad H_a : \mu = 16$$

with  $\alpha = \beta = .05$ . What is the minimum  $n$  he needs to determine a difference at this level?

$$n = n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2} = \frac{(1.645 + 1.645)^2 (9)}{(16 - 15)^2} = 97.4 \approx 98$$

### *Large Sample CI's and Hypothesis Testing*

Recall that, for large samples, the two-sided confidence interval for a confidence coefficient of  $(1 - \alpha)$  is given by:

$$\hat{\theta} \pm z_{\alpha/2} \sigma_{\hat{\theta}}$$

This should look quite similar to our  $\alpha$ -level hypothesis test, since it also uses the standardized normal variable  $Z$  and the same point estimators and standard errors.

Recall that the two-tailed rejection region for large sample tests is given by:

$$RR = \{ |z| > z_{\alpha/2} \}$$

we can also think of an “acceptance” region (probably a better term would be a fail to reject region) defined as the rest of the real number line:

$$\overline{RR} = \{ -z_{\alpha/2} \leq z \leq z_{\alpha/2} \}$$

This is equivalent to the endpoints of a  $100(1 - \alpha)\%$  confidence interval! In other words, another way to think about a two-sided hypothesis test is: Do not reject the null ( $\theta = \theta_0$ ) in favor of the alternative ( $\theta \neq \theta_0$ ) if  $\theta_0$  lies within a  $100(1 - \alpha)\%$  confidence interval. (note that this logic is the same for one-tailed tests, you just have to consider the one-sided confidence intervals).

Another way to think about this is that, for a given confidence level, a confidence interval gives us a range of probable values for the true parameter (it really doesn't quite do this, as we have discussed, but think of this as a heuristic).

If this interval estimate does not contain the null value, then we can reject the null hypothesis. If, however, the confidence interval does contain the null value, then (since *any* point within the confidence interval could be the true population parameter) we can not reject the null hypothesis.

Suppose you believe that the difference in reaction times between men and women to a particular stimulus is not zero (you don't know what direction, however), and, at the .05 level, your estimation procedures gives you a confidence level of  $[-1.7, 2.3]$ .

Since this interval contains 0, the null value, you cannot reject the null at the .05 level. Suppose recalculating the confidence interval, this time at the .10 level, yields an interval:  $[-.2, 1.4]$ .

The interval is smaller, but you have traded some of your certainty. Now, since 0 is excluded from the interval, you *can* reject the null hypothesis—but at the .10 level, not the .05 level.

This is the logic of the *p-value*. More formally, if  $W$  is a test statistic, then the *p-value* or attained significance level, is the smallest level of significance  $\alpha$  for which the observed data indicate that the null hypothesis should be rejected.

Thus, the smaller the *p-value* becomes, the more compelling the statistical evidence. An added benefit is the ability of the reader of results to determine for herself if the results are statistically significant.

While the author will almost always interpret the results as well (and usually at the .05 level in political science), reporting *p*'s is essential in the presentation of results—it allows you to shift the responsibility of deciding whether the results are statistically significant, at least in part, to the reader.

One additional point to remember, as a reader and an author, that as the sample size increases, the width of confidence intervals should decrease, for the same significance level. Accordingly, while you might be willing to accept a *p* of .05 or even .07 with a *n* of 100, these would not be so impressive with an *n* of 10,000 or even 1,000.

For our purposes, we will compute *p* using the tables of probabilities for the various distributions. I should note, however, that exact *p* is usually reported by statistical software.

Example (problem 10.40 in the text): For a given airline, flights must average 60% occupancy to be profitable. If a sample of 120 days of the same flight (e.g. Atlanta to Dallas) yields a mean of 58% occupancy with a standard deviation of 11%, at what attained level of significance can we conclude that the flight is unprofitable?

Solution: Our hypothesis can be stated:  $H_0 : \mu \geq .6$      $H_a : \mu < .6$ . Our test statistic is

$$z = \frac{\bar{y} - \mu_0}{\sigma / \sqrt{n}} = \frac{.58 - .6}{.11 / \sqrt{120}} = -1.99$$

Using the table in the normal way (the forward direction) we find that:

$$p = P(Z < -1.99) = .0233$$

so we can reject the null hypothesis at any level greater than or equal to .0233.

## Small Sample Hypothesis Tests

Just as in the case of small-sample estimators, small-sample hypothesis tests can not rely on the central limit theorem.

Today we will introduce two small sample tests: a test for the mean of a normally-distributed population, and a test for the difference of means of two normally-distributed populations. As before, we are making the additional assumption of normality so that we can model the sampling distribution of our test statistic with the  $t$  distribution (Student's  $t$ ). In all other ways, these tests closely resemble the large sample analogs.

*Small Sample Test for  $\mu$ :*

Assumption:  $Y_1, Y_2, \dots, Y_n$  constitute a random sample from a normal distribution with

$$E(Y_i) = \mu.$$

$$H_0 : \mu = \mu_0$$

$$H_a : \begin{cases} \mu > \mu_0, \text{ upper tail} \\ \mu < \mu_0, \text{ lower tail} \\ \mu \neq \mu_0, \text{ two-tailed} \end{cases}$$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \mu_0}{S / \sqrt{n}}$$

$$RR : \begin{cases} \{t > t_\alpha\}, \text{ upper} \\ \{t < -t_\alpha\}, \text{ lower} \\ \{|t| > t_{\alpha/2}\}, \text{ two-tailed} \end{cases}$$

where  $t$  has  $\nu = n - 1$  degrees of freedom (Appendix III table 5).

Example (10.13): Same problem as earlier example: 8 shells tested with mean muzzle velocity of 2959 fps and standard deviation of 39.1 fps. The manufacturer claims that the new gunpowder produces an average velocity of 3000 fps. Can we contradict his claim at the .025 level?

Solution: Assuming muzzle velocities are approximately normally distributed,

$$H_0 : \mu = 3000 \quad H_a : \mu < 3000$$

First, we convert our sample results to standard  $t$  units:

$$t = \frac{\bar{y} - \mu_0}{s / \sqrt{n}} = \frac{2959 - 3000}{39.1 / \sqrt{8}} = -2.966$$

What is the rejection region in standard units? From the table, with an alpha of .025, and 7 d.f., we find  $t = -2.365$ . Therefore, we can reject the null hypothesis at the .025 level and conclude that the muzzle velocity is, on average, slower than 3000 fps.

We can also find the  $p$  value associated with this small sample test. Recall that our test statistic is  $-2.966$ . Looking at the table, in the 7 d.f. row, we see that we can reject the null at the .025 level, but not at the .01 level. Using only this table, all we can say is that :

$$.01 \leq p \leq .025$$

We can, however, use a computer program to find true  $p$  values for  $t$  hypothesis tests, either by using a statistical calculator (included in many econometrics packages) or by integrating under the  $t$ 's pdf (a non-trivial exercise for a program with a computer algebra system such as Maple). For example, using STATA 8.0, the command:

**display ttest(7,2.966)**

returns the upper tail version of our hypothesis test (df, t):  $p = .01046307$

#### *Small Sample Difference of Means*

A very common application of the  $t$  distribution is in comparing the means of two normal populations with equal variances. Suppose we have two independent random samples,  $Y_{11}, Y_{12}, \dots, Y_{1n_1}$  and  $Y_{21}, Y_{22}, \dots, Y_{2n_2}$ , drawn from populations with means  $\mu_1, \mu_2$  and common variance  $\sigma^2$ . Our hypothesis tests look like:

$$H_0 : \mu_1 - \mu_2 = D_0$$

$$H_a : \begin{cases} \mu_1 - \mu_2 > D_0, & \text{upper tail} \\ \mu_1 - \mu_2 < D_0, & \text{lower tail} \\ \mu_1 - \mu_2 \neq D_0, & \text{two-tailed} \end{cases}$$

$$\text{Test Statistic: } T = \frac{\bar{Y}_1 - \bar{Y}_2 - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

$$RR : \begin{cases} \{t > t_\alpha\}, & \text{upper} \\ \{t < -t_\alpha\}, & \text{lower} \\ \{|t| > t_{\alpha/2}\}, & \text{two-tailed} \end{cases}$$

where  $P(T > t_\alpha) = \alpha$ ,  $v = n_1 + n_2 - 2$  d.f.

Example (10.14): The assembly line problem revisited:

<i>Standard Procedure</i>	<i>New Procedure</i>
n1 = 9	n2 = 9
$\bar{Y}_1 = 35.22$	$\bar{Y}_2 = 31.56$
$\sum_{i=1}^9 (y_{1i} - \bar{y}_1)^2 = 195.56$	$\sum_{i=1}^9 (y_{2i} - \bar{y}_2)^2 = 160.22$

Is there sufficient evidence to indicate a difference in means at the .05 level?

Solution:

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0$$

Since we have no *a priori* theory about the direction of the difference, we use a two-tailed test. Our first step is to compute the pooled standard deviation (which is equal to a simple average since the n's are equal):

$$s_p = \sqrt{\frac{195.56 + 160.22}{9 + 9 - 2}} = 4.716$$

Now we compute our test statistic:

$$t = \frac{35.22 - 31.56 - 0}{4.716 \sqrt{\frac{1}{9} + \frac{1}{9}}} = 1.65$$

From the table (16 d.f.,  $\alpha/2 = .025$ ), we find that the standard  $k$  (or rejection bound) is 2.120. Since the absolute value of our test statistic is not greater than this value, we fail to reject the null hypothesis at the .05 level (two-tailed).

What is the  $p$  value? From the table, we see that the area under the curve we are interested in lies between  $t_{.10}$  and  $t_{.05}$ . However, recall that we halved the alpha to calculate the two-tailed test. We must then remember to double the one-tailed values. Our  $p$  is thus:  $.1 \leq p \leq .2$ . Using STATA, we find the exact  $p = .1184332$ .

Before moving on, let me point out that the same discussion comparing confidence intervals and hypothesis tests also applies to these small-sample tests. Lastly, you might be concerned about the normality assumption needed to use these  $t$  tests. How normal is normal? Simulations have shown that moderate departures from normality have little

effect on the probability distribution of the test statistic. That is, the tests are said to be *robust*.

### Hypothesis Tests Concerning Variance

Once again we assume that we have a sample from a normal distribution with unknown mean and variance. Our hypothesis tests about the variance are of the form:

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \begin{cases} \sigma^2 > \sigma_0^2, \text{ upper tail} \\ \sigma^2 < \sigma_0^2, \text{ lower tail} \\ \sigma^2 \neq \sigma_0^2, \text{ two-tailed} \end{cases}$$

$$\text{Test Statistic: } \chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

$$RR : \begin{cases} \{ \chi^2 > \chi^2_{\alpha} \}, \text{ upper} \\ \{ \chi^2 < \chi^2_{1-\alpha} \}, \text{ lower} \\ \{ \chi^2 > \chi^2_{\alpha/2} \text{ or } \chi^2 < \chi^2_{1-\alpha/2} \}, \text{ two-tailed} \end{cases}$$

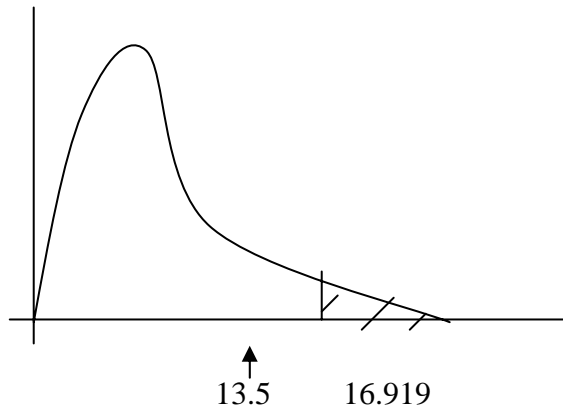
where  $\chi^2$  is chosen so that for  $\nu = n - 1$  d.f.,  $P(\chi^2 > \chi^2_{\alpha}) = \alpha$

Example (10.16): A company produces engine parts that are supposed to have a variance in diameter no larger than .0002 inches squared. A random sample of 10 parts has a sample variance of .0003. At the .05 or 5% level, test the upper tail hypothesis (variance > .0002) against the null that the variance = .0002.

Solution: Assuming normally distributed diameters, we compute our test statistic:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{9(.0003)}{.0002} = 13.5$$

From the table, we find our rejection bound:  $\chi_{.05}^2 = 16.919$  (9 d.f.). Since our test statistic does not fall in the rejection region, we fail to reject the null:



What is the  $p$  of this test? All that the table can tell us is that  $p > .10$ . Using the STATA command: **display chi2tail(9,13.5)** (d.f., value) returns approximately  $p = .14125$ . Remember that if the two-tailed case, you have to double the  $p$  from the table to account for the fact that you halved it earlier (see 10.18 in the text for a good example of this).

### Comparing Variances

What if we wish to test an hypothesis concerning the difference of two variances? If we assume again that the two distributions are normal (with unknown means), we can construct a test statistic from the sample variances such that the simple ratio of them is distributed  $F$ :

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 > \sigma_2^2$$

$$\text{Test Statistic : } F = \frac{S_1^2}{S_2^2}$$

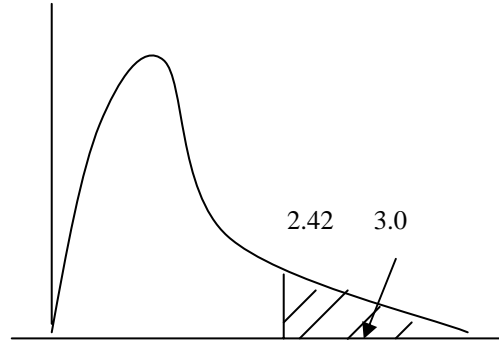
$$\text{Rejection Region: } F > F_\alpha$$

with  $\nu_1 = n_1 - 1$  d.f. in the numerator and  $\nu_2 = n_2 - 1$  d.f. in the denominator (the  $F$  distribution, recall, takes two parameters to identify it).

Note too, that if we wish to test a lower tail hypothesis, we simply reverse the order and test an upper tail hypothesis (the distribution with the *hypothesized larger* variance is always in the numerator).

Example (10.19): We wish to compare the variance of the machine parts created in the earlier example with machine parts from another vendor. In the previous sample of 10, the sample variance was .0003. In the new sample of 20, the variance is .0001. Is the variance of the second sample statistically smaller at the .05 level of significance?

Solution: The null is that the two variances are equal, the alternative is that the first is larger. Our test statistic is simple:  $F = \frac{.0003}{.0001} = 3$ . What is the critical value (where the rejection region starts)? Our  $F$  has 9 numerator and 19 denominator degrees of freedom. From the table, the critical value of  $F_{.05}$  is 2.42. We can thus reject the null at the .05 level:



In terms of finding the  $p$  value of this test, once again the table only tells us that  $p$  is somewhere between .025 and .01. From STATA, we find ( $\text{display Ftail}(9,19,3)$ )  $p = .02096$ .

What if we are interested in a two-tailed hypothesis about variances? Our  $F$  table does not give lower tail values, but it is easy to show that:

$$P(F_b^a > F_{b,\alpha/2}^a) = P\left[\left(F_b^a\right)^{-1} < \left(F_{b,\alpha/2}^a\right)^{-1}\right] = \alpha / 2$$

in other words, the critical value that cuts off the lower end of an  $F_a^b$  distribution can be found by inverting the  $F_{b,\alpha/2}^a$ . What this means practically for us is that we only have to check the  $F$  statistic with the larger variance in the numerator and the smaller in the denominator, remembering to divide alpha by 2 in the two-tailed case.

Example (10.21): 14 men and 10 women are given electric shocks to measure their pain thresholds. The mean for men is 16.2, variance 12.7 and the mean for women is 14.9, variance 26.4. Is there evidence to reject the null hypothesis that the variances for men and women are equal at the .10 level?

Solution: Assuming normality of pain thresholds, compute the test statistic:

$$F = \frac{26.4}{12.7} = 2.079$$

check this against the critical value ( $F = 2.71$ , with 9 numerator and 13 denominator d.f.). We must then fail to reject the null hypothesis that men and women have different variances in pain threshold.

Finally, I should note that this hypothesis test (putting the larger in the numerator) does not actually have a test statistic distributed  $F$  (the ratio is constrained to be  $\geq 1$ ). Nevertheless, the  $F$  tables provide an acceptable approximation. I should also point out that these tests of variance are *not* robust; they are extremely sensitive to non-normality.

### Power of Tests

Before we move on to a new topic, I want to end our discussion of hypothesis testing with a few theoretical issues. The first concerns the “power” of a test. Suppose that  $W$  is a test statistic and  $RR$  is the rejection region for a test of an hypothesis involving the value of parameter  $\theta$ . The power of a test,  $power(\theta)$ , is the probability that the test will lead to the rejection of the null hypothesis when the actual parameter value is  $\theta$ :

$$power(\theta) = P(W \text{ in } RR \text{ when parameter value is } \theta)$$

If we wanted to test a null hypothesis that that  $\theta = \theta_0$ , and we assume that  $\theta_a$  is the value of the parameter under the alternative hypothesis, then the

$$power(\theta_a) = P(\text{Rejecting } H_0 \text{ when } \theta = \theta_a)$$

a brief inspection of this quantity gives an alternative definition of power:

$$power(\theta_a) = 1 - \beta(\theta_a)$$

In other words, power is the probability of *not* making a Type II error—that is, actually rejecting the null hypothesis when it is false, or not overlooking a relationship in your data that really is there.

Power is useful in two areas: research design, and evaluation of the research of others. Since it’s really just another way to talk about Type II error, I won’t go into it in any more detail (those of you taking experimental design will, however). Just be sure you are aware of it.

Homework problems: 10.2, 10.3, 10.7, 10.9, 10.11, 10.13, 10.14, 10.15, 10.17, 10.22, 10.23, 10.27, 10.33, 10.43, 10.44, 10.47, 10.48, 10.51, 10.52, 10.53, 10.55, 10.57, 10.61, 10.63, 10.64, 10.65, 10.66, 10.68, 10.71, 10.73.