

HYPOTHESIS TESTING

Concepts of Hypothesis Testing

The best way to present the concepts of hypothesis testing is through an example.

Suppose a hospital supervisor believes that the average time it takes medical technicians to complete a certain task has changed. Previously, it took, on the average 2 minutes ($\mu=2$) to complete a certain task with a standard deviation of $\sigma=.2$. So, now the supervisor believes that it $\mu \neq 2$. How might we decide if, based on a sample of 100 observations, there is evidence to show the mean is significantly different than 2 minutes?

Suppose we find a 95% confidence interval for μ —the true time it takes a technician to complete the task. This will be a Z confidence interval as follows:

$$\bar{x} \pm 1.96 \frac{.2}{\sqrt{100}} \equiv \bar{x} \pm .0392. \text{ If we take our sample and find } \bar{x} = 2.2 \text{ then we are}$$

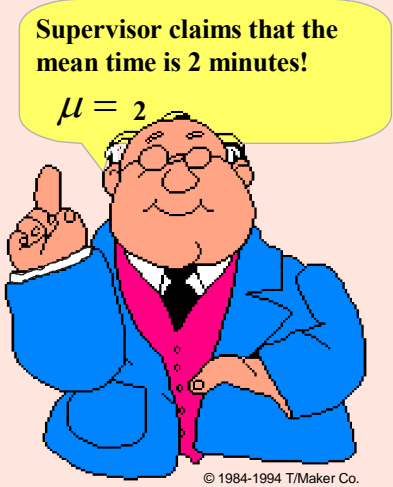
95% confident $2.1608 < \mu < 2.2392$. Since $\mu=2$ is not inside the confidence interval we reject the supervisor's claim and say the data **does provide evidence** to show that the mean time is significantly different from 2 minutes and the **supervisor's claim is false**.

What we just did here is a test of the hypothesis $\mu=2$.

Figure 9.1 What is a Hypothesis

What is a Hypothesis?

- A hypothesis is a claim (assumption) about the population parameter
 - Examples of parameters are population mean or proportion
 - The parameter must be identified before analysis



Supervisor claims that the mean time is 2 minutes!
 $\mu = 2$

© 1984-1994 T/Maker Co.

Every hypothesis testing problem will have a **null hypothesis** that we denote by H_0 and an **alternative hypothesis** that we denote by H_1 .

For our example we are testing $H_0: \mu=2$ vs. $H_1: \mu \neq 2$.

The Null Hypothesis, H_0

- States the assumption (numerical) to be tested
 - e.g.: The mean time technicians take to complete the task is 2 minutes ($H_0 : \mu=2$)
- Is always about a population parameter not about a sample statistic.
- Begins with the assumption that the null hypothesis is true
 - Similar to the notion of innocent until proven guilty
- Refers to the status quo
- May or may not be rejected

The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis
 - e.g.: The true mean time is significantly different from 2 minutes ($H_1 : \mu \neq 2$)
- Challenges the status quo
- Is generally the hypothesis that is believed (or needed to be proven) to be true by the researcher

Based on the confidence interval approach,

Our test statistic is \bar{x} and our decision rule is *reject H_0 if* $\left| \frac{\bar{x} - 2}{\sigma / \sqrt{n}} \right| > z_{\alpha/2}$

Using the confidence interval we decided to reject the H_0 if our null hypothesis value of 2 was not inside our confidence interval. The equivalence of this decision in testing is called our **decision rule**. The decision rule (also called the critical region) tells us when to reject the null hypothesis. We have a 5% error of making the wrong decision of rejecting the null hypothesis when the null hypothesis is true, we call this our **level of significance α** .

Clearly, when doing any type of inference, we are making statements and decisions about a population based on a sample. There will always be chances of errors. ***What are the different types of errors we can make when doing hypothesis testing?***

There are two types of error, Type I error and Type II error.

•Type I Error

- Rejects a true null hypothesis
- Has serious consequences

The probability of Type I Error is

- Called **level of significance α**
- Typical values are .01, .05, .10
- Set by researcher at the beginning

••Type II Error

- Fails to reject a false null hypothesis
- The probability of Type II Error is β
- The **power** of the test is $1-\beta$

Before proceeding with the discussion on how to decide on a level of significance and how to determine errors and power, let's summarize the terminology and the errors.

Summary of Hypothesis-Testing Terminology

Null Hypothesis (H_0): A maintained hypothesis that is held to be true unless sufficient evidence to the contrary is obtained.

Alternative Hypothesis (H_1): A hypothesis against which the null hypothesis is tested and which will be held to be true if the null is held false.

Simple Hypothesis: A hypothesis that specifies a single value for a population parameter of interest.

Composite Hypothesis: A hypothesis that specifies a range of values for a population parameter.

One-Sided Alternative: An alternative hypothesis involving all possible values of a population parameter on either one side or the other of (that is, either greater than or less than) the value specified by a simple null hypothesis.

Two-Sided Alternative: An alternative hypothesis involving all possible values of a population parameter other than the value specified by a simple null hypothesis.

Hypothesis Test Decisions: A decision rule is formulated, leading the investigator to either accept or reject the null hypothesis on the basis of sample evidence.

(Decisions or Decision Rules are often called the Critical Region of the Test and tell you when to reject a null hypothesis)

Type I Error: The rejection of a true null hypothesis.

Type II Error: The acceptance of a false null hypothesis.

Significance Level: The probability of rejecting a null hypothesis that is true. This probability is sometimes expressed as a percentage, so a test of significance level α is referred to as a $100\alpha\%$ -level test.

Power: The probability of rejecting a null hypothesis that is false.

Table 9.1 States of Nature and Decisions on Null Hypothesis, with Probabilities of Making the Decisions, given the States of Nature

Decisions on Null Hypothesis	Null Hypothesis is True	Null Hypothesis is False
Do Not Reject (Accept)	Correct Decision Probability = $1 - \alpha$	Type II error Probability = β
Reject	Type I error Probability = α <i>α is called the significance level</i>	Correct decision Probability = $1 - \beta$ <i>$(1 - \beta)$ is called power</i>

To understand the decisions and the errors, one way of thinking of hypothesis testing is to equate it to the jury trial in which there is a guilty and a not guilty verdict. This is summarized in Figure 9.2

Figure 9.2 Hypothesis Testing Analogy to Jury Trial Verdicts

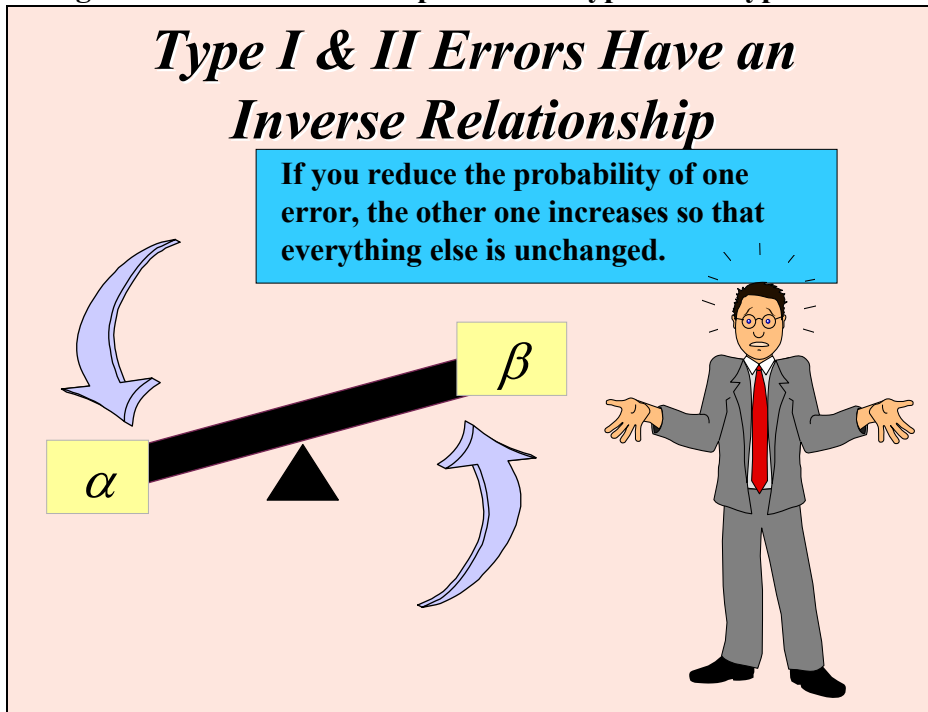
Result Probabilities

H_0 : Innocent

		Jury Trial		Hypothesis Test	
		The Truth		The Truth	
Verdict	Innocent	Guilty	Decision	H_0 True	H_0 False
Innocent	Correct	Error	Do Not Reject H_0	$1 - \alpha$	Type II Error (β)
Guilty	Error	Correct	Reject H_0	Type I Error (α)	Power ($1 - \beta$)

Of course, we would like to have a procedure that minimizes the errors, the problem that arises is that as we minimize the probability of Type I error, the probability of Type II error increases.

Figure 9.3 The Relationship Between Type I and Type II Error

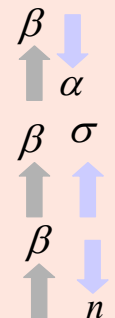


What factors effect the errors and how do we between the two errors.

Figure 9.3 Factors Effecting Type II Error

Factors Affecting Type II Error

- True value of population parameter
 - β Increases when the difference between hypothesized parameter and its true value decrease
- Significance level
 - β Increases when α decreases
- Population standard deviation
 - β Increases when σ increases
- Sample size
 - β Increases when n decreases



How to Choose between Type I and Type II Errors

- Choice depends on the cost of the errors
- Choose smaller Type I Error when the cost of rejecting the maintained hypothesis is high
 - A criminal trial: convicting an innocent person
 - The Exxon Valdez: causing an oil tanker to sink
- Choose larger Type I Error when you have an interest in changing the status quo
 - A decision in a startup company about a new piece of software
 - A decision about unequal pay for a covered group

Always bear in mind, that the Type I error is considered the worse error. In most problems, we fix the level of significance and go with the test that maximizes the power for this level of significance.

Tests of the Mean of a Normal Distribution: Population Variance Known

The initial approach we will take to hypothesis testing involves ***decision rules based on a test statistic and critical values.*** We will see that once a test statistic has been defined, the critical regions are similar from hypothesis testing situation to hypothesis testing situation. The key to all our problems, will be to correctly set up our null and alternative hypothesis and to define the test statistics.

Figure 9.4 Critical Value Approach to Hypothesis Testing

***Critical Values
Approach to Testing***

- Convert sample statistic (e.g.: \bar{X}) to test statistic (e.g.: Z , t or F –statistic)
- Obtain critical value(s) for a specified α from a table or computer
 - If the test statistic falls in the critical region, reject H_0
 - Otherwise do not reject H_0

With this approach, all Z , T , F and χ^2 tests will have the same critical regions, once the data has been converted in to these test statistics.

When sampling from a Normal Population with an unknown mean and a known variance, we saw how we can take a confidence interval approach to the 2-sided testing problem. Let's formalize that idea and it is no surprise that the test statistic for making decisions about the population mean is as follows:

Figure 9.5 Z Test Statistic for a Normal Population with Known Variance

***Z Test Statistic for the Mean
of a Normal Population
(σ Known)***

- Assumptions
 - Population is normally distributed
 - If not normal, requires large samples
- z test statistic

$$z = \frac{\bar{x} - \mu}{\sigma_x} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \quad (9.1)$$

When do we reject the null hypothesis for one-sided hypothesis testing?

If we have a one-sided alternative hypothesis for which the alternative is $\mu > \mu_0$ then it seems logical to reject H_0 if the sample mean is significantly larger than μ_0 . How much is significantly

larger? If the sample mean is more than Z standard errors above μ_0 and what determines this critical Z value is the level of significance α . For the one-sided alternative $\mu < \mu_0$ we would reject the null hypothesis if the sample mean were more than Z standard errors below μ_0 . Hence, we have the following decision rules and they should make sense.

One-Sided Tests for the Mean of a Normal Population*: Population Variance Known

Given that we have a random sample of n observations from a normal population with mean μ and known variance σ^2 . If the observed sample mean is \bar{X} , and Z the test statistic defined by 9.1 with $\mu = \mu_0$, then a test with significance level α

$$H_0: \mu = \mu_0 \text{ or } H_0: \mu \leq \mu_0 \text{ vs. } H_1: \mu > \mu_0$$

is obtained from the decision rule

$$\text{Reject } H_0 \text{ if } Z > Z_\alpha \tag{9.2}$$

$$H_0: \mu = \mu_0 \text{ or } H_0: \mu \geq \mu_0 \text{ vs. } H_1: \mu < \mu_0$$

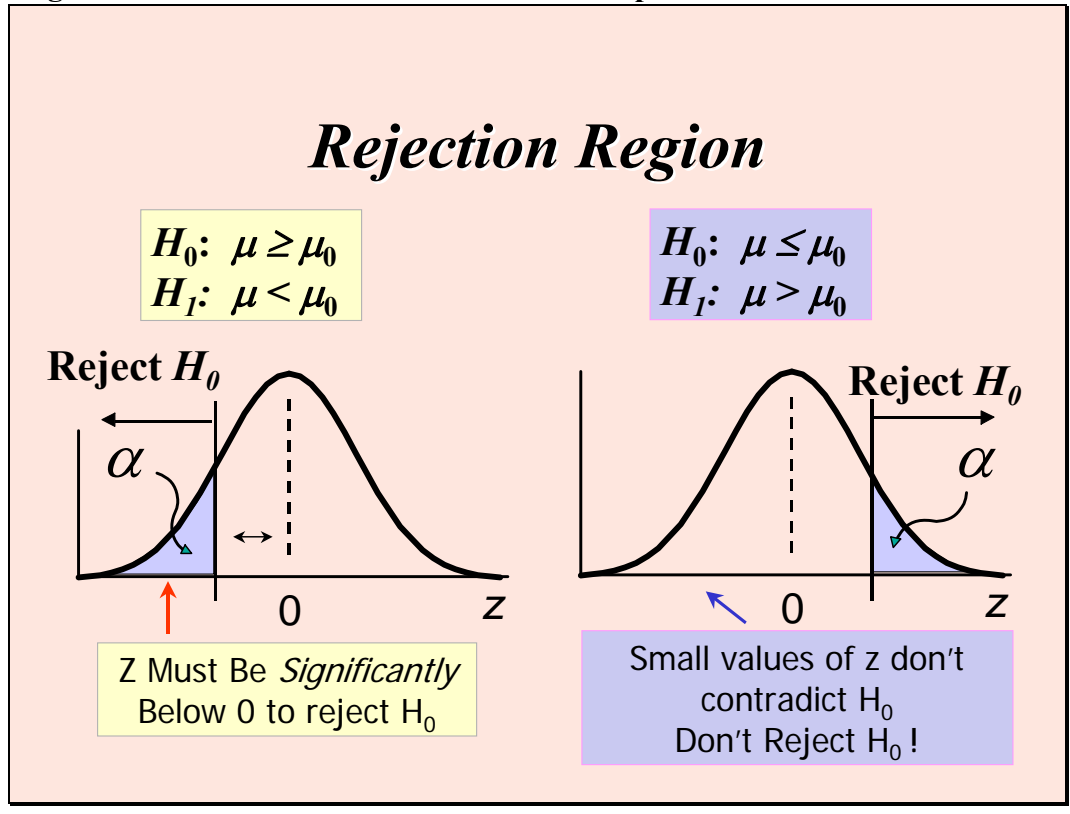
$$\text{Reject } H_0 \text{ if } Z < -Z_\alpha \tag{9.3}$$

where Z_α is the number for which $P(Z > Z_\alpha) = \alpha$ when Z is the standard normal random variable.

*This is also the test for a non-normal population with known variance when the sample size n is large (i.e. the Central Limit Theorem will apply).

In Figure 9.6, we can see exactly what the rejection region for this test looks like.

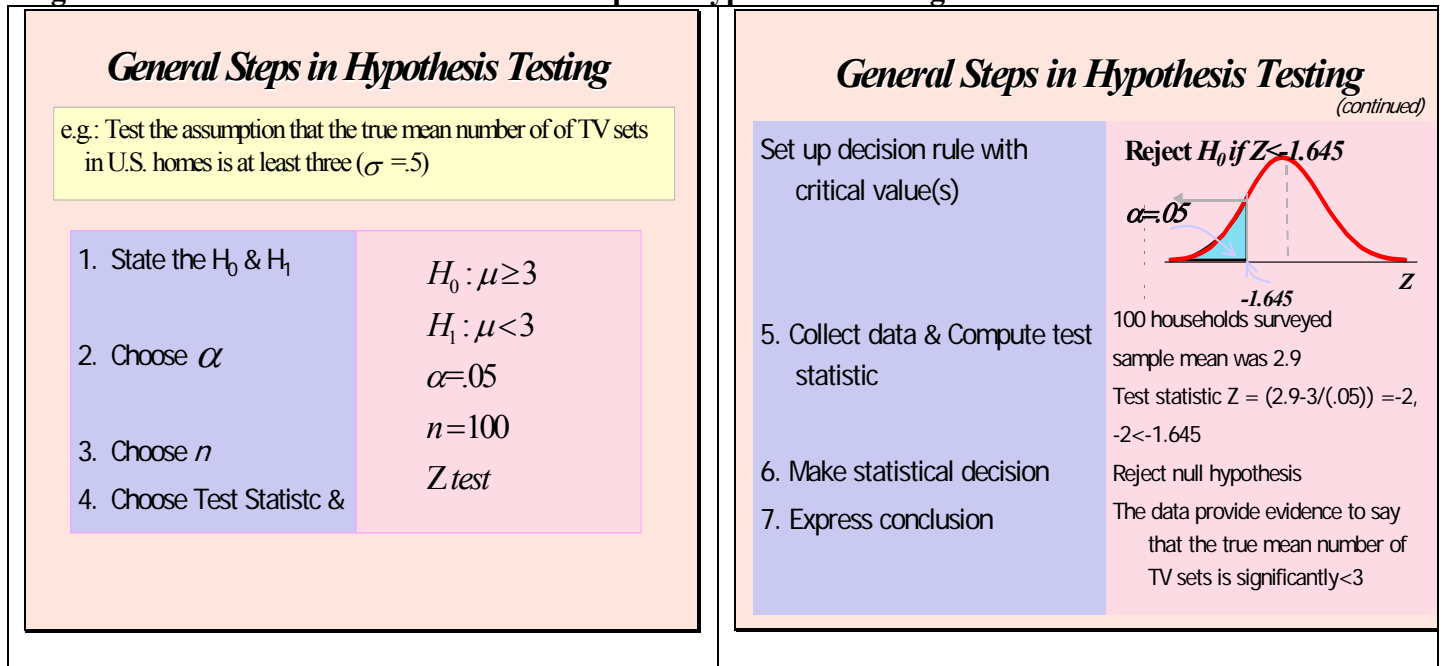
Figure 9.6 One-Sided Z Test for a Normal Population with Known Variance



When doing hypothesis testing there is a correct order and way of doing it so that your answer and decision is objective and correct.

Through the following example we can see the 7 steps of hypothesis testing in the proper order.

Figure 9.7 One-Sided Z Test & General Steps of Hypothesis Testing



Before proceeding with more examples, let's make sure it is clear that ***we understand how tests vary based on their levels of significance.*** In the previous example, our test statistic $Z = -2 < -1.645$ and we rejected H_0 for $\alpha = .05$. However, had we tested the hypotheses at $\alpha = .01$, our decision rule would have been to reject H_0 if $Z < -2.33$. In this case, our test statistic is not < -2.33 , and so we would not reject H_0 . This is important to note for several reasons. First, we see that as we make the ***level of significance bigger, it is easier for our data to reject H_0 .*** We also see why it is important to choose the level of significance in step 2, because our decisions can be different based on what level of significance we choose. We try to use a little bigger level of significance wherever possible, because then our tests have more power. Moreover, if the level of significance is too small, we'll never be able to reject a null hypothesis and we'll never be able to prove anything.

After we do the next example, we'll discuss the **p-value** of a test. The p-value is the smallest level of significance for which the data you collected, will reject H_0 . It should be clear, that for the previous example, since we rejected at $\alpha = .05$ and we did not reject at $\alpha = .01$. that the p-value for this test is between these two numbers. We'll find that exact p-value for this example and discuss it's meaning a little later on in the lesson.

First, let's do another one-sided problem. A cereal company wants to be certain that on the average, its cereal boxes contain at most 368 grams of cereal. The question under consideration, is based on a random sample of 25 cereal boxes, does the data provide evidence

to show the true mean amount of cereal per box is significantly > 368 grams at level of significance .05? Following our 7 steps of hypothesis testing we have:

1. $H_0: \mu \leq 368$ $H_1: \mu > 368$
2. $\alpha=.05$
3. $n=25$

4. **Reject H_0 if $Z = \left(\frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \right) > z_{\alpha} = 1.645$**

5. *When the data was gathered, the sample mean was $\bar{x} = 372.5$, so our test statistic*

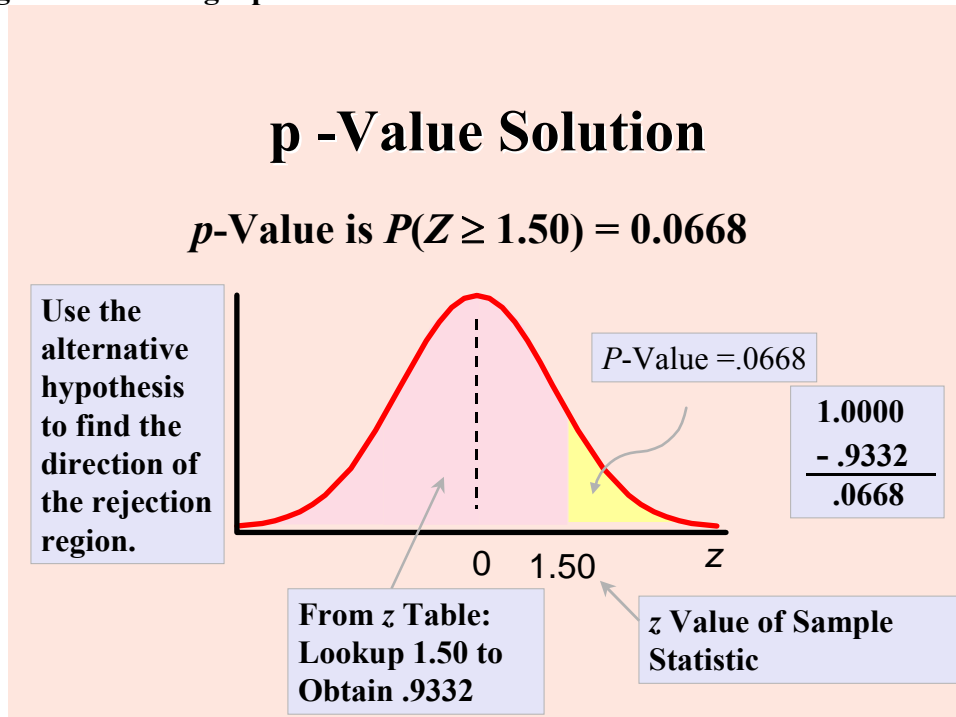
$$Z = \left(\frac{372.5 - 368}{15 / \sqrt{25}} \right) = 1.5$$

6. $1.5 < 1.645$, therefore we cannot reject H_0
7. *The data do not provide evidence to show that the true mean amount of cereal per box is significantly greater than 368 grams.*

What is the p-value of this test?

In the previous example our test statistic was 1.5 and that was not big enough to reject H_0 at $\alpha=.05$. So, we need a bigger level of significance in order for this data to reject the null hypothesis. How big?

Figure 9.8 Finding a p-value for a Z-Test



So for the cereal problem we would need a level of significance of .0668 in order for this data to reject H_0 .

Since, generally, the p-value is the smallest value of α for which an H_0 can be rejected then the following are universal decision rules which hold for every test:

Hypothesis Testing Using the p-value

•Compare the p-value with α

–If p-value $\geq \alpha$ do not reject H_0 –If p-value $\leq \alpha$ reject H_0 **(9.4)**

Interpretation of the p-value for one-sided tests

The probability value or p-value is the smallest significance level at which the null hypothesis can be rejected. Consider a random sample of size n observations from a population that has a normal distribution with unknown mean μ and *known* standard deviation. We are asked to test the null hypothesis

$$H_0: \mu = \mu_0 \text{ or } H_0: \mu \leq \mu_0 \text{ vs. } H_1: \mu > \mu_0$$

The p-value for the test is

$$\text{p-value} = P(Z > Z_p) \qquad (9.5)$$

$$H_0: \mu = \mu_0 \text{ or } H_0: \mu \geq \mu_0 \text{ vs. } H_1: \mu < \mu_0$$

The p-value for the test is

$$\text{p-value} = P(Z < Z_p) \qquad (9.6)$$

where Z_p is the value of your test statistic for the data you collected.

The p-value is regularly computed by most statistical computer programs and provides more information about the test, based on the observed sample mean. Thus it is a popular tool for many statistical applications.

For our cereal example, $Z_p = 1.5$ and according to 9.5, $p\text{-value} = P(Z > 1.5) = .0668$, which it was. So, returning to the TV set example, $Z_p = -2$ and according to formula (9.6), the $p\text{-value} = P(Z < -2) = .0228$. So, the smallest level of significance for which this data will reject H_0 is .0228. Since $\alpha = .05 > .0228$, we reject at that level of significance, but since $\alpha = .01 < .0228$ we could not reject at that level of significance. Just as we said would be the case based on the test statistic and the critical values.

All the ideas for hypothesis testing have now been set down. If we wish to do a two-sided test for a mean, we still have the same test statistic Z defined by (9.1). However, our decision rule must consider Z being too large or too small.

Two-Sided Alternative Hypothesis A Test of the Mean of a Normal Distribution Against Two-Sided Alternative: σ Known

The appropriate procedure for testing, at significance level α , the null hypothesis

$$H_0: \mu = \mu_0 \text{ vs. } H_1: \mu \neq \mu_0$$

is obtained from the decision rule

$$\text{Reject } H_0 \text{ if } |Z| > Z_{\alpha/2} \text{ or equivalently if } Z > Z_{\alpha/2} \text{ or } Z < -Z_{\alpha/2} \quad (9.7)$$

Where Z is the test statistic defined by 9.1 with $\mu = \mu_0$

In addition the p-values can also be computed by noting that the corresponding tail probability would be doubled to reflect a p-value that refers to the sum of the upper and lower tail probabilities for the positive and negative values of Z . The p-value for the two tailed test is

$$\text{p-value} = 2P(Z > Z_{p/2}) \quad (9.8)$$

where $Z_{p/2}$ is the absolute value of your test statistic for the data you collected

The decision rule (9.7) is equivalent to

$$\text{reject } H_0 \text{ if } \mu_0 \notin (1-\alpha)100\% \text{ confidence interval for } \mu. \quad (9.9)$$

Recall the example in Section 9.1 where we wanted to see if the time it took medical technicians to complete a certain task was significantly different from 2 minutes. We proceeded taking a confidence interval approach and found that with 95% confidence $2.1608 < \mu < 2.2392$. According to (9.9) since $2 \notin (2.1608, 2.2392)$ we reject H_0 . Let's use decision rule 9.7 and the 7 steps of hypothesis testing to come to the same conclusion.

1. $H_0: \mu = 2$ vs. $H_1: \mu \neq 2$.

2. $\alpha = .05$

3. $n = 100$

4. Reject H_0 if $|Z| > Z_{\alpha/2} = 1.96$

5. For $n = 100$, $\sigma = .2$ and $\bar{x} = 2.2$ $|Z| = \frac{2.2 - 2}{.2 / \sqrt{100}} = 10$

6. $10 > 1.96$ therefore Reject H_0

7. The data do provide evidence to show the mean time is significantly different from 2 minutes.

The p-value here is according to (9.8) $2P(Z > 10) \cong 0$. Our test statistic is so big, it will reject H_0 at any level of significance. Using the hypothesis testing approach we really see how strongly this data is saying the mean is significantly different from 2.

Tests of the Mean of a Normal Distribution: Population Variance Unknown

In this section we will deal with a Normal Population when the variance is unknown. Essentially, we proceed exactly as we did in Section 9.2, except instead of a Z test statistic we have the following T Test statistic.

Figure 9.9 t Test Statistic for a Normal Population with an Unknown Variance

t Test Statistic for the Mean of a Normal Population :σ Unknown

- Assumption
 - Population is normally distributed
 - If not normal, requires a large sample
- t test statistic with $n-1$ degrees of freedom

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}} \quad (9.10)$$

Tests for the Mean of a Normal Population: Population Variance Unknown

Given that we have a random sample of n observations from a normal population with mean μ and unknown variance σ^2 . If the observed sample mean is \bar{X} and the observed sample variance is S^2 , and t is the test statistic defined by 9.10 with $\mu = \mu_0$, then a test with significance level α

$$H_0: \mu = \mu_0 \text{ or } H_0: \mu \leq \mu_0 \text{ vs. } H_1: \mu > \mu_0$$

is obtained from the decision rule **Reject H_0 if $t > t_{n-1, \alpha}$** (9.11)

$$H_0: \mu = \mu_0 \text{ or } H_0: \mu \geq \mu_0 \text{ vs. } H_1: \mu < \mu_0$$

Reject H_0 if $t < -t_{n-1, \alpha}$ (9.12)

where $t_{n-1, \alpha}$ is the student t value for $n - 1$ degrees of freedom and upper tail probability α .

$$H_0: \mu = \mu_0 \text{ vs. } H_1: \mu \neq \mu_0$$

is obtained from the decision rule

Reject H_0 if $|t| > t_{n-1, \alpha/2}$ or equivalently if $t > t_{n-1, \alpha/2}$ or $t < -t_{n-1, \alpha/2}$ (9.13)

The decision rule (9.13) is equivalent to

reject H_0 if $\mu_0 \notin (1-\alpha)100\%$ confidence interval for μ . (9.14)

where $t_{n-1, \alpha/2}$ is the student t value for $n - 1$ degrees of freedom and upper tail probability $\alpha/2$.

The p -values for these tests are computed in the same way as we did for tests with known variance except that the student t value is substituted for the normal Z value.

Example) Suppose that the financial administrator of a hospital claims that the mean malpractice settlement in a certain area is \$50,000 and settlement amounts are known to be normally distributed. Test this claim at level of significance .05 based on a sample of 36 claims. Proceeding with the 7 steps of hypothesis testing for this example (leaving the data in thousands of dollars):

1. $H_0: \mu = 50$ vs. $H_1: \mu \neq 50$

2. $\alpha=.05$
3. $n=36$
4. Reject H_0 if $|t| > t_{35, .025} = 2.0301$

5. Collect the data and compute the statistics $\bar{x} = 52.2$ $s = 6$ $|t| = \frac{\left| \frac{\bar{x} - 50}{\frac{s}{\sqrt{36}}} \right| = 2.2$

6. $2.2 > 2.0301$ therefore Reject H_0 .
7. The data provide evidence to show that the administrator's claim ($\mu=50$) is false.

Would our answer change at level of significance .01?

If we tested at $\alpha=.01$ we would not have rejected the null hypothesis because then $t_{35, .005} = 2.7238 > 2.2$. Remember, the larger the level of significance, the easier it is to reject H_0 .

At this time, it is important to know that ***ALL Z-TESTS HAVE THE SAME CRITICAL REGIONS AND ALL t-tests HAVE THE SAME CRITICAL REGIONS***

Regardless of what hypothesis you are testing or what parameter(s) make up your null and alternative hypothesis, all Z-tests and all t-tests have the same critical regions. The key is to define your test statistics for each situation. The test statistic depends on the populations and parameters you are testing about. Table 9.2 summarizes the universal critical regions for all one-sided and 2-sided Z-tests and t-tests.

Table 9.2 Universal Critical Regions for Z-tests and t-tests

Critical Region for an upper (lower) tail one-side z test:	Reject H_0 if $Z > Z_\alpha$ (Reject H_0 if $Z < -Z_\alpha$) where Z_α is the value for which $P(Z > Z_\alpha) = \alpha$
Critical Region for an two-sided z test	Reject H_0 if $Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$ ($ Z > Z_{\alpha/2}$) where $Z_{\alpha/2}$ is the value for which $P(Z > Z_{\alpha/2}) = \alpha/2$
Critical Region for an upper (lower) tail one-side t test:	Reject H_0 if $t > t_\alpha$ (Reject H_0 if $t < -t_\alpha$) where t_α is the value for which $P(t > t_\alpha) = \alpha$
Critical Region for an two-sided t test	Reject H_0 if $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$ ($ t > t_{\alpha/2}$) where $t_{\alpha/2}$ is the value for which $P(t > t_{\alpha/2}) = \alpha/2$

Tests for the Population Proportion (Large Samples)

If we look at qualitative data and we wish to do hypothesis testing for the parameter π =proportion of successes in a population, we would use the following test statistic:

Figure 9.10 Z Test Statistic for a Population Proportion

Z Test Statistic for the Population Proportion (Large Sample Sizes)

- Assumption
 - Assuming a random sample of n observations from a population, that has a proportion π whose members possess a particular attribute.
 - If $n\pi(1-\pi) > 9$ and the sample proportion is p
- z test statistic is

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad (9.15)$$

To test one-sided hypotheses

$$H_0: \pi = \pi_0 \text{ or } H_0: \pi \leq \pi_0 \text{ vs. } H_1: \pi > \pi_0$$

$$H_0: \pi = \pi_0 \text{ or } H_0: \pi \geq \pi_0 \text{ vs. } H_1: \pi < \pi_0$$

or the two-sided hypothesis

$$H_0: \pi = \pi_0 \text{ vs. } H_1: \pi \neq \pi_0$$

The decision rules are given in *Table 9.2, using the test statistic (9.15) when $\pi = \pi_0$.*

For all of these tests the **p-value** is the smallest significance level at which the null hypothesis can be rejected as defined by (9.5), (9.6) and (9.8)

Example) A marketing company claims that it receives at least a 4% response from its mailings. Test to see if based on a sample of 1000 the response rate is significantly less than 4% at $\alpha = .05$ significance level. From the sample of 1000, there were 34 responses to the mailing.

We proceed with the 7 steps of hypothesis testing just as we have with all our problems.

First check $n\pi(1-\pi) > 9$. $n\pi(1-\pi) = 1000(.034)(.966) = 38.4 > 9$

1. $H_0: \pi \geq .04$ vs. $H_1: \pi < .04$
2. $\alpha = .05$
3. $n = 1000$
4. Reject H_0 if $Z < -1.645$
5. $n = 1000$ $x = 34$ $p = 34/1000 = .034$

$$z = \frac{.034 - .04}{\sqrt{\frac{.04(1 - .04)}{1000}}} = -.968$$

6. $-.968 > -1.645$ therefore **Do Not Reject H_0**

7. **The data do not provide evidence to show that the marketing company's claim is incorrect.**

From (9.6) p-value = $P(Z < -.968) = .1665$. This is a fairly large p-value and we cannot reject the null hypothesis for any reasonable level of significance.

Tests of the Variance of a Normal Distribution

Suppose X_1, X_2, \dots, X_n i.i.d. $N(\mu, \sigma^2)$ then we already showed that $S^2 = \left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \right]$

is an unbiased estimator for σ^2 and the statistic $\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(v=n-1)}$

moreover, we developed confidence intervals for the population variance based on this statistic. The tests for variance follow the same ideas as those for means and proportions. The key is to define the test statistic and state the general decision rule based on that test statistic.

Recall, in the Chapter 8 notes, we found a 99% confidence interval for the population variance for the amount of life insurance that workers take out. The confidence interval was based on a sample of $n=100$ with the following statistics computed from the sample: $\bar{x} = 80$ and $s = 15$

$$(\$000). \quad \chi_{n-1, \alpha/2}^2 = \chi_{99, .025}^2 = \mathbf{128.4219} \quad \chi_{n-1, 1-\alpha/2}^2 = \chi_{99, .975}^2 = \mathbf{73.6611} \quad \frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2}$$

$$= \frac{99(225)}{128.4219} < \sigma^2 < \frac{99(225)}{73.611} \quad \text{So, with 95\% certainty } \mathbf{173.45 < \sigma^2 < 303}$$

Using similar reasoning as we did in Section 9.1, it seems reasonable that if we wanted to test $H_0: \sigma^2 = 200$ vs. $H_1: \sigma^2 \neq 200$, we could not reject H_0 since 200 is in the confidence interval and hence could possibly be the population variance. Continuing as we did in all the previous sections, the following tests for population variances should make perfect sense.

Figure 9.11 χ^2 Test Statistic for a Population Variance

χ^2 Test Statistic for the Population Variance

- Assumption
 - Population is normally distributed
 - If not normal, requires a large sample
- χ^2 test statistic with $n-1$ degrees of freedom

$$\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \quad (9.16)$$

Tests of Variance of a Normal Population

Given a random sample of n observations from a normally distributed population with variance σ^2 . If we observe the sample variance s^2 , and χ_{n-1}^2 is the test statistic defined by 9.16 with $\sigma^2 = \sigma_0^2$ then the following tests have significance level α .

$$\mathbf{H_0: \sigma^2 = \sigma_0^2 \quad \text{or} \quad H_0: \sigma^2 \leq \sigma_0^2} \quad \text{vs.} \quad \mathbf{H_1: \sigma^2 > \sigma_0^2}$$

$$\text{Reject } H_0 \text{ if } \chi^2 > \chi_{n-1, \alpha}^2 \quad (9.17)$$

$$\mathbf{H_0: \sigma^2 = \sigma_0^2 \quad \text{or} \quad H_0: \sigma^2 \geq \sigma_0^2} \quad \text{vs.} \quad \mathbf{H_1: \sigma^2 < \sigma_0^2}$$

$$\text{Reject } H_0 \text{ if } \chi^2 < \chi_{n-1, \alpha}^2 \quad (9.18)$$

$$\mathbf{H_0: \sigma^2 = \sigma_0^2} \quad \text{vs.} \quad \mathbf{H_1: \sigma^2 \neq \sigma_0^2}$$

$$\text{Reject } H_0 \text{ if } \chi^2 > \chi_{n-1, \alpha/2}^2 \quad \text{or} \quad \chi^2 < \chi_{n-1, \alpha/2}^2 \quad (9.19)$$

Where χ_{n-1}^2 is a chi-square random variable and $P(\chi_{n-1}^2 > \chi_{n-1, \alpha}^2) = \alpha$

The decision rule (9.19) is equivalent to
 reject H_0 if $\sigma_0^2 \notin (1-\alpha)100\%$ confidence interval for σ^2 . (9.20)

Tests for the Difference Between Two Population Means

Two Means, Matched Pairs

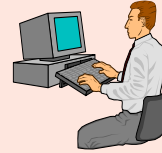
Tests of the Difference Between Population Means: Matched Pairs

Suppose that there is a random sample of n matched pairs of observations from normal distributions with means μ_X and μ_Y . That is, let x_1, x_2, \dots, x_n denotes the values of the observations from the population with mean μ_X ; and y_1, y_2, \dots, y_n the matched sampled values from the population with the mean μ_Y . Let \bar{d} and s_d denote the observed sample mean and standard deviation for the n differences $d_i = x_i - y_i$. If the population distribution of the differences is assumed to be normal, and $\mu_d = \mu_X - \mu_Y$, then this just ***reduces to the one-sample t test given (9.11) – (9.14)***

(If the population variances were known, we would use the one-sample Z-test on the differences.)

Paired Sample t Test: Example

You work in the finance department. Is the new financial package faster ($\alpha=0.05$ level)? You collect the following data entry times:



User	Current Leader (1)	New Software (2)	Difference D_i
C.B.	9.98 Seconds	9.88 Seconds	.10
T.F.	9.88	9.86	.02
M.H.	9.84	9.75	.09
R.K.	9.99	9.80	.19
M.O.	9.94	9.87	.07
D.S.	9.84	9.84	.00
S.S.	9.86	9.87	-.01
C.T.	10.12	9.86	.26
K.T.	9.90	9.83	.07
S.Z.	9.91	9.86	.05

$$\bar{D} = \frac{\sum D_i}{n} = .084$$

$$S_D = \sqrt{\frac{\sum (D_i - \bar{D})^2}{n-1}} = .0844$$

Clearly from the data set-up we can see that this just reduces to a one-sample problem.

Paired Sample t Test: Example Solution

Is the new financial package faster (0.05 level)?

$$H_0: \mu_D \leq 0$$

$$H_1: \mu_D > 0$$

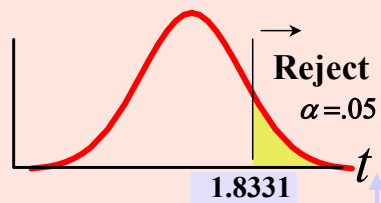
$$\alpha = .05 \quad \bar{D} = .084$$

$$\text{Critical Value} = 1.8331$$

$$df = n - 1 = 9$$

Test Statistic

$$t = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} = \frac{.084 - 0}{.08444 / \sqrt{10}} = 3.15$$



Decision: Reject H_0

t Stat. in the rejection zone.

Conclusion: The new software package is faster.

Two Means, Independent Samples, Known Population Variances

Now we will consider the case where we have independent random samples from two normally distributed populations. The first population has mean μ_x and variance σ_x^2 and we obtain a random sample of size n_x . The second population has mean μ_y and variance σ_y^2 and we obtain a random sample of size n_y . **This situation is a straightforward generalization of the one population problem.**

Figure 9.12 Z Test Statistic for the Difference in Means for 2 Independent Populations

Independent Samples Z Test Statistic (Variances Known)

- Assumptions
 - Samples are randomly and independently drawn from normal distributions
 - Samples are randomly and independently drawn from non-normal distributions but the sample sizes are large
 - Population variances are known
- Test statistic

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \quad (9.21)$$

Tests for Difference Between Population Means: Independent Samples (Known Variances)
Suppose that we have independent random samples of n_x and n_y observations from normal distributions with means μ_x and μ_y and variances σ_x^2 and σ_y^2 . If the observed sample

means are \bar{X} and \bar{Y} and the test statistic is the Z test statistic defined by (9.21) when $\mu_x - \mu_y = D_0$ then the following tests have significance level α .

(i) $H_0: \mu_x - \mu_y = D_0$ or $H_0: \mu_x - \mu_y \leq D_0$ vs. $H_1: \mu_x - \mu_y > D_0$

(ii) $H_0: \mu_x - \mu_y = D_0$ or $H_0: \mu_x - \mu_y \geq D_0$ vs. $H_1: \mu_x - \mu_y < D_0$

(iii) $H_0: \mu_x - \mu_y = D_0$ vs. $H_1: \mu_x - \mu_y \neq D_0$

The decision rules are the Z-test decision rules given in Table 9.2

The decision rule for the two-sided case (iii) is equivalent to reject H_0 if $D_0 \notin (1 - \alpha)100\%$ confidence interval for $\mu_x - \mu_y$.

For all of these tests the **p-value** is the smallest significance level at which the null hypothesis can be rejected as defined by (9.5), (9.6) and (9.8)

If the sample sizes are **large ($n > 100$) and the population variances are unknown** then a good approximation at significance level α can be made if the population variances are replaced by the sample variances. In addition the central limit leads to good approximations even if the populations are not normally distributed.

Two Means, Independent Populations, Unknown Variances Assumed to be Equal

In those cases where the population variance is not known and sample sizes are under 100 we need to use the student's t distribution. There are some theoretical problems when we use the student t distribution for differences between sample means. However, these problems can be solved using the procedure that follows if we can assume that the population variances are equal. This assumption is realistic in many cases where we are comparing groups. In Section 9.8 we will present a procedure for testing the equality of variances from two normal populations.

Recall in Chapter 8 we introduced the **pooled variance estimator** as an estimator of the common variance of the two populations.

The *pooled estimator of the equal population variance* is a weighted average of the two sample variances and is defined as

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)} \quad (9.22)$$

Figure 9.13 t Test Statistic for the Difference in Means for 2 Independent Normal Populations with Equal Variances

Independent Samples t Test Statistic (Variances Unknown but Equal)

- Assumptions
 - Samples are randomly and independently drawn from normal distributions
 - Population variances are unknown but assumed equal
- Test statistic is a t with $n_x + n_y - 2$ degrees of freedom

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} \quad (9.23)$$

Tests for Difference Between Population Means: Independent Samples (Unknown Equal Variances)

Suppose that we have independent random samples of n_x and n_y observations from normal distributions with means μ_x and μ_y and unknown variances σ_x^2 and σ_y^2 . If the observed sample

means are \bar{X} and \bar{Y} and the observed sample variances are s_x^2 and s_y^2 and **the test statistic is the t test statistic defined by (9.23) when $\mu_x - \mu_y = D_0$ with s_p^2 defined by (9.22)**

then the following tests have significance level α .

(i) $H_0: \mu_x - \mu_y = D_0$ or $H_0: \mu_x - \mu_y \leq D_0$ vs. $H_1: \mu_x - \mu_y > D_0$

(ii) $H_0: \mu_x - \mu_y = D_0$ or $H_0: \mu_x - \mu_y \geq D_0$ vs. $H_1: \mu_x - \mu_y < D_0$

(iii) $H_0: \mu_x - \mu_y = D_0$ vs. $H_1: \mu_x - \mu_y \neq D_0$

The decision rules are given in Table 9.2, using the t-test decision rules.

The decision rule for the two-sided case (iii) is equivalent to reject H_0 if $D_0 \notin (1-\alpha)100\%$ confidence interval for $\mu_x - \mu_y$.

For all of these tests the **p-value** is the smallest significance level at which the null hypothesis can be rejected as defined by (9.5), (9.6) and (9.8) with t substituted for Z.

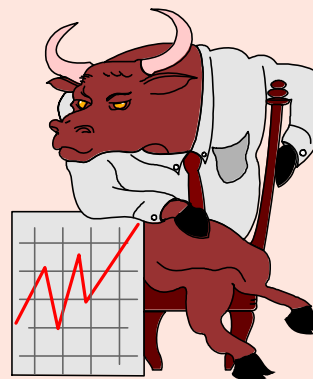
Figure 9.14 Pooled Variance t Test Example: Stock Exchange Problem

Pooled Variance t Test: Example

You are a financial analyst for Charles Schwab. You want to compare dividend yields between stocks listed on the NYSE & NASDAQ. You collect the following data:

	<u>NYSE</u>	<u>NASDAQ</u>
Number	21	25
Mean	3.27	2.53
Std dev	1.30	1.16

Is there a difference in the variances between the NYSE & NASDAQ at the $\alpha = 0.05$ level?



© 1984-1994 T/Maker Co.

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)} = \frac{(20)(1.3)^2 + (24)(1.16)^2}{44} = 1.15$$

Solution

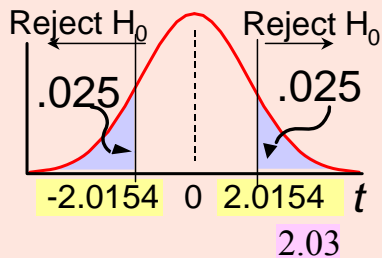
$H_0: \mu_x - \mu_y = 0$ i.e. $(\mu_x = \mu_y)$

$H_1: \mu_x - \mu_y \neq 0$ i.e. $(\mu_x \neq \mu_y)$

$\alpha = 0.05$

$df = 21 + 25 - 2 = 44$

Critical Value(s):



Test Statistic:

$$t = \frac{3.27 - 2.53}{\sqrt{1.510 \left(\frac{1}{21} + \frac{1}{25} \right)}} = 2.03$$

Decision:

Reject at $\alpha = 0.05$

Conclusion:

There is evidence of a difference in means.

Tests for the Difference Between Two Population Proportions (Large Samples)

Next we will develop procedures for comparing two population proportions. We will consider a standard model with a random sample of n_x observations with a proportion p_x "successes" and an independent random sample of n_y observations from a population with a proportion p_y "successes".

In Section 8.7 we saw that for large samples, proportions can be approximated as normally distributed random variables and as a result

$$Z = \frac{(p_x - p_y) - (\pi_x - \pi_y)}{\sqrt{\frac{\pi_x(1 - \pi_x)}{n_x} + \frac{\pi_y(1 - \pi_y)}{n_y}}}$$

Figure 9.14 Z Test Statistic for the Difference in Proportions for 2 Independent Populations

Independent Samples Z Test Statistic For the Difference in Proportions (Large Samples)

- Assumptions
 - Samples are randomly and independently drawn
 - If $n_x(1-\pi_x) > 9$ and the sample proportion is p_x and $n_y(1-\pi_y) > 9$ and the sample proportion is p_y
 - Let p_0 be an estimate of π when $\pi_x = \pi_y = \pi$ and be defined by
$$p_0 = \frac{n_x p_x + n_y p_y}{n_x + n_y}$$

- Test statistic
$$Z = \frac{(p_x - p_y) - (\pi_x - \pi_y)}{\sqrt{\frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y}}} \quad (9.24)$$

Testing the Equality of Two Population Proportions (Large Samples)

Given independent random samples of size n_x and n_y with proportion successes p_x and p_y .

The following tests have significance level α :

(i) $H_0: \pi_x - \pi_y = 0$ or $H_0: \pi_x - \pi_y \leq 0$ vs. $H_1: \pi_x - \pi_y > 0$

(ii) $H_0: \pi_x - \pi_y = 0$ or $H_0: \pi_x - \pi_y \geq 0$ vs. $H_1: \pi_x - \pi_y < 0$

(iii) $H_0: \pi_x - \pi_y = 0$ vs. $H_1: \pi_x - \pi_y \neq 0$

The decision rules are given in Table 9.2, using the test statistic (9.24) when $\pi_x - \pi_y = 0$

For all of these tests the **p-value** is the smallest significance level at which the null hypothesis can be rejected as defined by (9.5), (9.6) and (9.8)

So once again, we have been reduced to another Z test.

Testing for the Equality of the Variances Between Two Normally Distributed Populations

The F Distribution

Given that we have two independent random samples with n_x and n_y observations from two normal populations with variances σ_x^2 and σ_y^2 . If the sample variances are s_x^2 and s_y^2 then the random variable

$$F = \frac{s_x^2 / \sigma_x^2}{s_y^2 / \sigma_y^2} \quad (9.25)$$

has an F distribution with numerator degrees of freedom $(n_x - 1)$ and denominator degrees of freedom $(n_y - 1)$.

An F distribution with numerator degrees of freedom v_1 and denominator degrees of freedom v_2 will be denoted F_{v_1, v_2} . We denote $F_{v_1, v_2, \alpha}$ the number for which

$$P(F_{v_1, v_2} > F_{v_1, v_2, \alpha}) = \alpha$$

The F Test Statistic

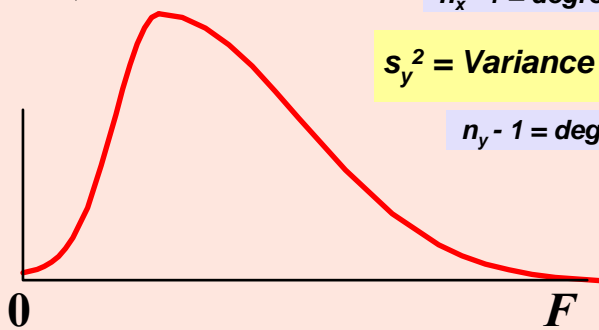
$$F = \frac{s_x^2 / \sigma_x^2}{s_y^2 / \sigma_y^2}$$

$s_x^2 = \text{Variance of Sample } x$

$n_x - 1 = \text{degrees of freedom}$

$s_y^2 = \text{Variance of Sample } y$

$n_y - 1 = \text{degrees of freedom}$



We need to emphasize that this test is quite sensitive to the assumption of normality.

Tests for Equality of Variances from Two Normal Populations

Let s_x^2 and s_y^2 be observed sample variances from independent random samples of size n_x and n_y from normally distributed populations with variances σ_x^2 and σ_y^2 . Use s_x^2 to denote the larger variance, then the following tests have significance level α :

(i) $H_0: \sigma_x^2 = \sigma_y^2$ or $H_0: \sigma_x^2 \leq \sigma_y^2$ vs. $H_1: \sigma_x^2 > \sigma_y^2$

the decision rule is **Reject H_0 if $F = \frac{S_x^2}{S_y^2} > F_{n_x-1, n_y-1, \alpha}$** (9.26)

(ii) $H_0: \sigma_x^2 = \sigma_y^2$ or vs. $H_1: \sigma_x^2 \neq \sigma_y^2$

the decision rule is **Reject H_0 if $F = \frac{S_x^2}{S_y^2} > F_{n_x-1, n_y-1, \alpha/2}$** (9.27)

where s_x^2 is the larger of the two sample variances. Since either sample variance could be larger this rule is actually based on a two tailed test and hence we use $\alpha/2$ as the upper tail probability.

Here, F_{n_x-1, n_y-1} is the number for which $P(F_{n_x-1, n_y-1} > F_{n_x-1, n_y-1, \alpha}) = \alpha$

where F_{n_x-1, n_y-1} has an F distribution with $(n_x - 1)$ numerator degrees of freedom and $(n_y - 1)$ denominator degrees of freedom.

For all of these tests a p-value is defined as the smallest significance level at which the null hypothesis can be rejected. Because of the complexity of the F distribution critical values are computed for only a few special cases. Thus **p-values will be typically computed using a statistical package such as Minitab or by using Excel with PH stat**

Let's return to the Stock Exchange Example where the data is given in Figure 9.14. We already tested for equal means, assuming the variances were equal. The right thing to do is to test this assumption first. Let's test to see if there is a significant difference in the variances.

F Test: Example Solution

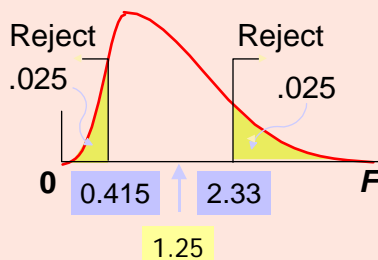
$$H_0: \sigma_x^2 = \sigma_y^2$$

$$H_1: \sigma_x^2 \neq \sigma_y^2$$

$$\alpha = .05$$

$$Df_1 = 20 \quad df_2 = 24$$

Critical value(s):



Test Statistic:

$$F = \frac{s_x^2}{s_y^2} = \frac{(1.3)^2}{(1.16)^2} = 1.25$$

Decision:

Do not reject at $\alpha = 0.05$

Conclusion:

There is no evidence of a difference in variances.

So we were correct to use the pooled variance t Test for this data.

Assessing the Power of a Test

The power function has the following features

1. The farther the parameter is from the hypothesized value in the null hypothesis, the greater is the power of the test-- everything else being equal. This should make perfect sense. If a test is a good test, it should have an easier time doing the right thing if it is farther away from the null hypothesis.
2. The smaller the significance level of the test, the smaller the power, everything else being equal. Thus reducing the probability of Type I error increases the probability of Type II error, but reducing α by 0.01 does not generally increase β by 0.01 -- the changes are not linear.
3. The larger the population variance the lower the power of the test -- everything else being equal. Note that larger sample sizes reduce the variance of the sample mean and thus provide a greater chance that we will reject H_0 when it is not correct.
4. The power of the test at the boundary of H_0 equals α .

Some Comments on Hypothesis Testing

The null hypothesis plays a crucial role in the hypothesis testing framework. In a typical investigation we set the significance level, α , at a small probability value. Then we obtain a random sample and use the data to compute a test statistic. If the test statistic is outside of the acceptance region (depending on the direction of the test) the null hypothesis is rejected and the alternative hypothesis is accepted. When we do reject we have strong evidence --small probability of error -- in favor of the alternative hypothesis. In some cases we may fail to reject a drastically false null hypothesis simply because of limited sample information or because the test has low power. There may be important cases where this outcome is appropriate. ***For example we would not change an existing process that is working effectively unless we had strong evidence that a new process would be clearly even better.*** In other cases, however, the special status of the null hypothesis is neither warranted nor appropriate. In those cases we might consider the costs of making both Type I and Type II errors in a decision process. We might also consider a different specification of the null hypothesis -- noting that rejection of the null provides strong evidence in favor of the alternative. When we have two alternatives we could initially choose either as the null hypothesis. ***However the burden of proof will always be on the alternative hypothesis.***

In specifying a null hypothesis and a testing rule we are defining the rules before we look at the sample data that was generated by a process that includes a random component. ***Thus if we look at the data before defining the null and alternative hypothesis we no longer have the stated probability of error and the concept of "strong evidence" resulting from rejecting the null hypothesis is not valid. For example if we were to decide on the significance level of our test after we have seen the p-values then we cannot interpret our results in probability terms.*** Suppose that an economist compares each of five different income enhancing programs against a standard minimal level using a hypothesis test. After collecting

the data and computing p-values he determines that the null hypothesis -- income not above the standard minimal level-- can be rejected for one of the five programs with a significance level of $\alpha = 0.20$. Clearly this result violates the proper use of hypothesis testing. But we have seen this done.

As statistical computing tools have become more powerful there are a number of new ways of violating the principle of pre-specifying the null hypothesis before seeing the data. The recent popularity of Data Mining introduces new possibilities for abuse. Data mining can provide a description of subsets and differences in a particular large sample of data. However, after seeing the results from a data mining operation analysts may be tempted to define hypothesis tests based on the same data. This clearly violates the principle of defining the hypothesis test before seeing the data. A drug company may screen large numbers of medical treatment cases and discover that five out of 100 drugs have significant effects for the treatment of previously unintended diseases. Such a result might legitimately be used to identify potential research questions for a new research study with new random samples. However, if the original data is then used to test a hypothesis concerning the treatment benefits of the five drugs we have a serious violation of the proper application of hypothesis testing and none of the probabilities of error are correct.