9 Hypothesis Tests

(Ch 9.1-9.3, 9.5-9.9)

Statistical Hypotheses

Statistical hypothesis: a claim about the value of a parameter or population characteristic.

Examples:

- *H:* μ = 75 cents, where μ is the true population average of daily per-student candy+soda expenses in US high schools
- *H:* $p < .10$, where p is the population proportion of defective helmets for a given manufacturer
- If μ_1 and μ_2 denote the true average breaking strengths of two different types of twine, one hypothesis might be the assertion that $\mu_1 - \mu_2 = 0$, or another is the statement $\mu_1 - \mu_2 > 5$

Components of a Hypothesis Test

1. **Formulate the hypothesis to be tested**.
2. **Determine the appropriate test statistic and calculate it** using the sample data.
3. **Comparison of test statistic to critical region** to draw initial conclusions.
4. **Calculation of p-value**.
5. **Conclusion**, written in terms of the original problem.

Components of a Hypothesis Test

1. **Formulate the hypothesis to be tested**.
2. **Determine the appropriate test statistic and calculate it** using the sample data.
3. **Comparison of test statistic to critical region** to draw initial conclusions.
4. **Calculation of p-value**.
5. **Conclusion**, written in terms of the original problem.

1. Null vs Alternative Hypotheses

In any hypothesis-testing problem, there are always two competing hypotheses under consideration:

1. The status quo (null) hypothesis
2. The research (alternative) hypothesis

The objective of **hypothesis testing** is to decide, based on sample information, *if the alternative hypotheses is actually supported by the data.*

We usually do new research to challenge the existing (accepted) beliefs.

1. Null vs Alternative Hypotheses

Is there strong evidence for the alternative?

The burden of proof is placed on those who believe in the alternative claim.

This initially favored claim (H_0) will not be rejected in favor of the alternative claim (H_a or H_1) unless the sample evidence provides significant support for the alternative assertion.

If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of the null hypothesis.

The two possible conclusions: ***1) Reject H_0.***

2) Fail to reject H_0.

1. Null vs Alternative Hypotheses

Why be so committed to the null hypothesis?

- Sometimes we do not want to accept a particular assertion unless (or until) data can show strong support
- Reluctance (cost, time) to change

Example: Suppose a company is considering putting a new type of coating on bearings that it produces.

The true average wear life with the current coating is known to be 1000 hours. With μ denoting the true average life for the new coating, the company would not want to make any (costly) changes unless evidence strongly suggested that μ exceeds 1000.

1. Null vs Alternative Hypotheses

An appropriate problem formulation would involve testing

$$H_0: \mu = 1000 \text{ against } H_a: \mu > 1000.$$

The conclusion that a change is justified is identified with H_a, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced, or "elaborated upon."

1. Null vs Alternative Hypotheses

The alternative to the null hypothesis H_0: $\theta = \theta_0$ will look like one of the following three assertions:

1. H_a: $\theta \neq \theta_0$
2. H_a: $\theta > \theta_0$ (in which case the null hypothesis is $\theta \leq \theta_0$)
3. H_a: $\theta < \theta_0$ (in which case the null hypothesis is $\theta \geq \theta_0$)

- The equality sign is *always* with the null hypothesis.
- The alternate hypothesis is the claim for which we are seeking statistical proof.

Components of a Hypothesis Test

1. **Formulate the hypothesis to be tested**.
2. **Determine the appropriate test statistic and calculate it** using the sample data.
3. **Comparison of test statistic to critical region** to draw initial conclusions.
4. **Calculation of p-value**.
5. **Conclusion**, written in terms of the original problem.

2. Test Statistics

A **test statistic** is a rule, based on sample data, for deciding whether to reject H_0.

The test statistic is a function of the sample data that will be used to make a decision about whether the null hypothesis should be rejected or not.

2. Test Statistics

Example: Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards.

$$H_0: p = .10 \text{ versus } H_a: p < .10$$

Our data is a random sample of $n = 200$ boards from company B.

What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?

2. Test Statistics

Which test statistic is “best”??

There are an infinite number of possible tests that could be devised, so we have to limit this in some way or total statistical madness will ensue!

Choice of a particular test procedure must be based on the probability the test will produce incorrect results.

2. Errors in Hypothesis Testing

Definition

- A **type I error** is when the null hypothesis is rejected, but it is true.
- A **type II error** is not rejecting H_0 when H_0 is false.

This is very similar in spirit to our diagnostic test examples

- False negative test = type I error
- False positive test = type II error

2. Errors in Hypothesis Testing

Definition

- A **type I error** is when the null hypothesis is rejected, but it is true.
- A **type II error** is not rejecting H_0 when H_0 is false.

This is very similar in spirit to our diagnostic test examples

- False negative test = type I error
- False positive test = type II error

How do we apply this to the circuit board problem?

2. Type I errors

Usually: Specify the largest value of α that can be tolerated, and then find a rejection region with that α.

The resulting value of α is often referred to as the **significance level** of the test.

Traditional levels of significance are .10, .05, and .01, though the level in any particular problem will depend on the seriousness of a type I error—

The more serious the type I error, the smaller the significance level should be.

2. Errors in Hypothesis Testing

We can also obtain a smaller value of α -- the probability that the null will be incorrectly rejected – by decreasing the size of the rejection region.

However, this results in a larger value of β for all parameter values consistent with H_a.

No rejection region that will simultaneously make both α and all β's small. A region must be chosen to strike a compromise between α and β.

2. Testing means of a normal population with known σ

Null hypothesis: $H_0 : \mu = \mu_0$

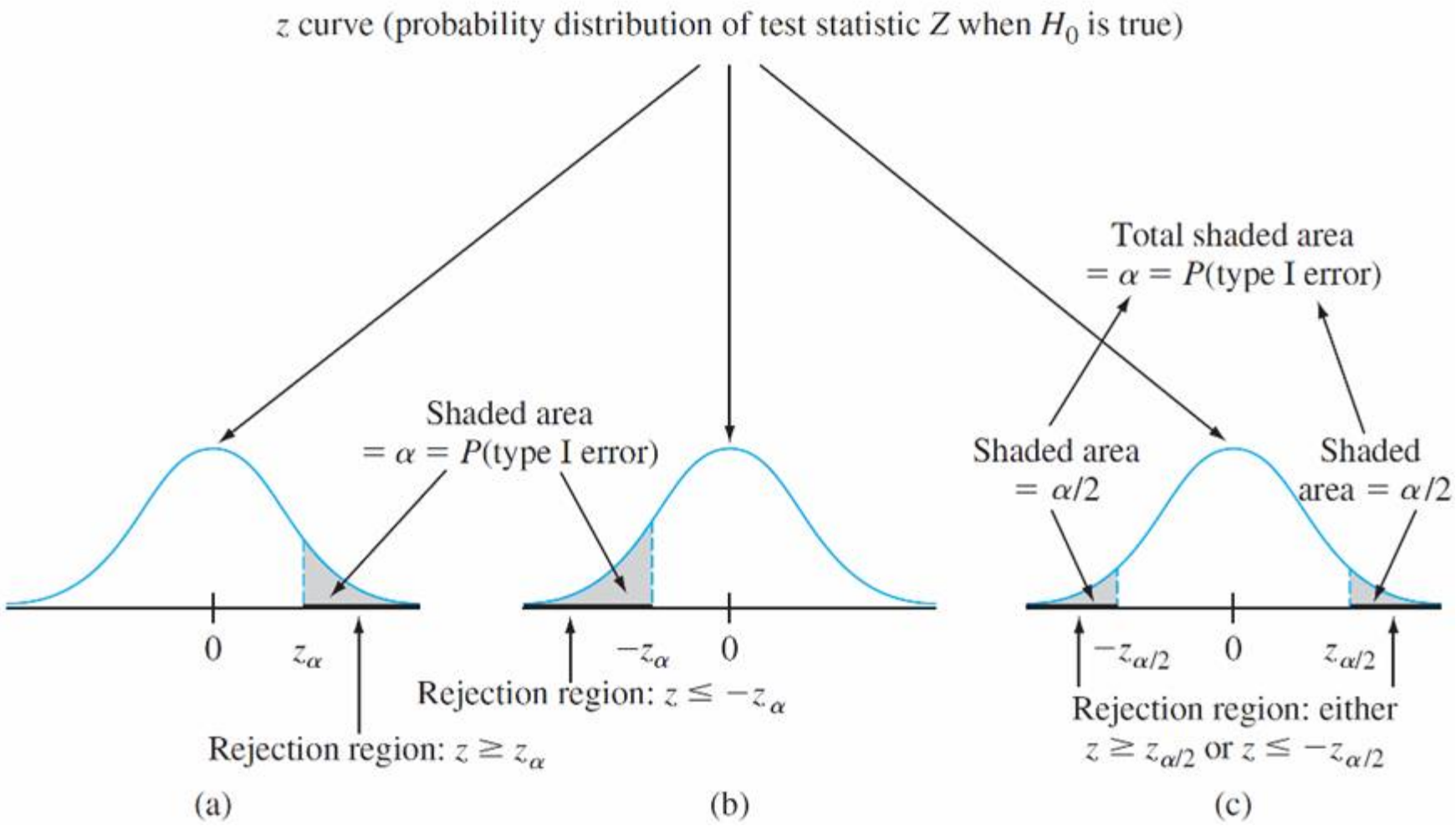
Test statistic value : $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
$H_a: \mu > \mu_0$	$z \geq z_\alpha$ (upper-tailed test)
$H_a: \mu < \mu_0$	$z \leq -z_\alpha$ (lower-tailed test)
$H_a: \mu \neq \mu_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)

Components of a Hypothesis Test

1. **Formulate the hypothesis to be tested**.
2. **Determine the appropriate test statistic and calculate it** using the sample data.
3. **Comparison of test statistic to critical region** to draw initial conclusions.
4. **Calculation of p-value**.
5. **Conclusion**, written in terms of the original problem.

3. Critical region



Rejection regions for *z* tests: (a) upper-tailed test; (b) lower-tailed test; (c) two-tailed test

Example

An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for more than 5 hours (300 minutes) on a single gallon of regular gasoline. (The leading brand lawnmower engine runs for 300 minutes on 1 gallon of gasoline.)

From his stock of engines, the inventor selects a simple random sample of 50 engines for testing. The engines run for an average of 305 minutes. The true standard deviation σ is known and is equal to 30 minutes, and the run times of the engines are normally distributed.

Test hypothesis that the mean run time is more than 300 minutes. Use a 0.05 level of significance.

2. Testing means of a large sample

When the sample size is large, the *z* tests for case I are easily modified to yield valid test procedures without requiring either a normal population distribution or known σ.

Earlier, we used the key result to justify large-sample confidence intervals:

A large *n (>30)* implies that the standardized variable

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

has *approximately* a standard normal distribution.

2. Testing means of a small sample coming from a normal

The One-Sample t Test

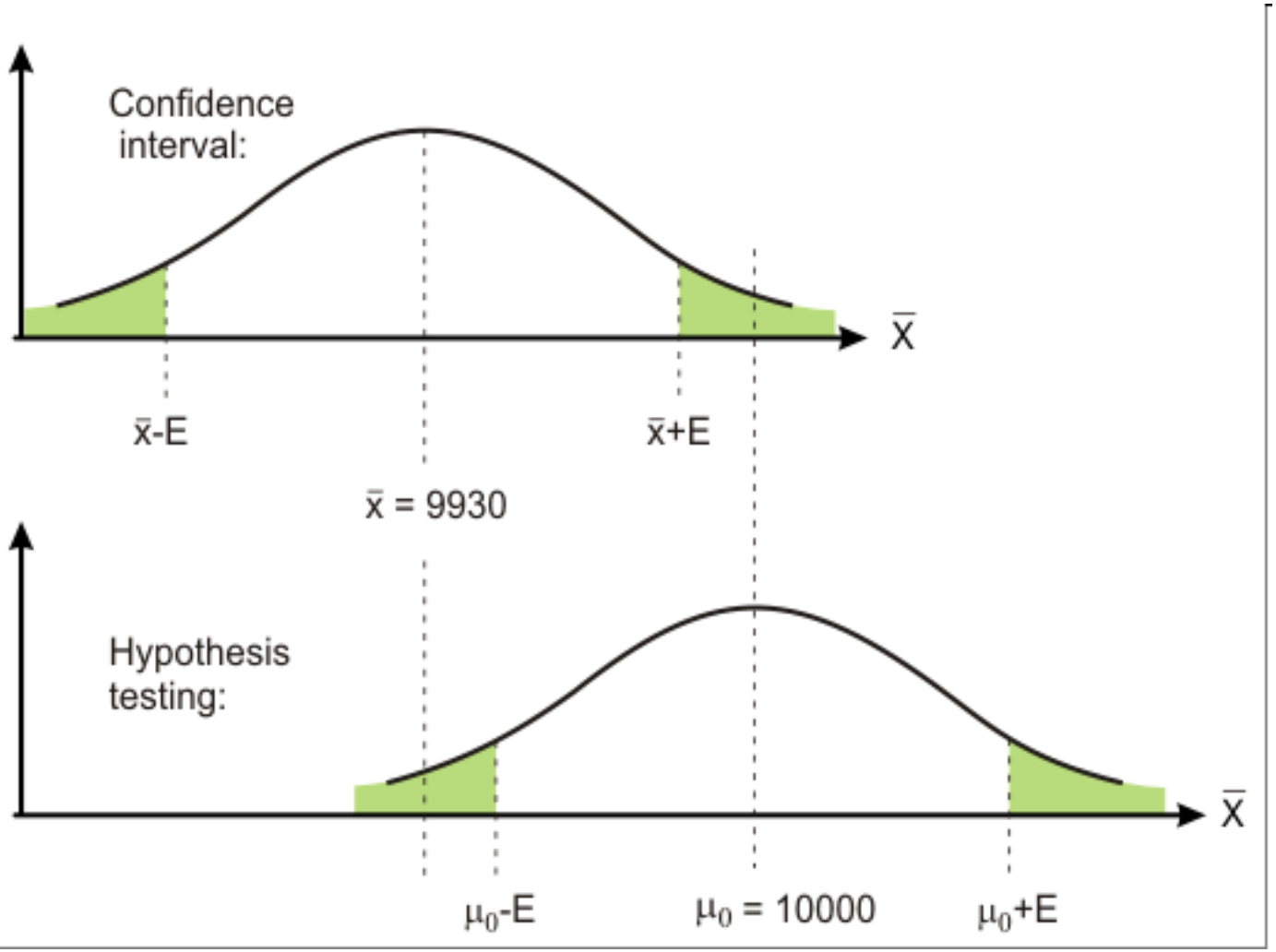
Null hypothesis: H_0: $\mu = \mu_0$

Test statistic value: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	Rejection Region for a Level α Test
H_a: $\mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper-tailed)
H_a: $\mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower-tailed)
H_a: $\mu \neq \mu_0$	either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$ (two-tailed)

CI vs. Hypotheses

Rejection regions have a lot in common with confidence intervals.



Source:

CI vs. Hypotheses

Example: The Brinell scale is a measure of how hard a material is. An engineer hypothesizes that the mean Brinell score of all subcritically annealed ductile iron pieces is not equal to 170. It is known that these scores follow a normal distribution.

The engineer measured the Brinell score of 25 pieces of this type of iron and calculated the sample mean to be 174.52 and the sample standard deviation to be 10.31.

Perform a hypothesis test that the true average Brinell score is not equal to 170, as well as the corresponding confidence interval. Set alpha = 0.01.

Components of a Hypothesis Test

1. **Formulate the hypothesis to be tested**.
2. **Determine the appropriate test statistic and calculate it** using the sample data.
3. **Comparison of test statistic to critical region** to draw initial conclusions.
4. **Calculation of p-value**.
5. **Conclusion**, written in terms of the original problem.

4. *p*-Values

The p-value measures the "extremeness" of the sample.

<u>Definition:</u> The p-value is the probability we would get the sample we have or something more extreme *if the null hypothesis were true.*

So, the smaller the P-value, the more evidence there is in the sample data against the null hypothesis and for the alternative hypothesis.

So what constitutes "sufficiently small" and "extreme enough" to make a decision about the null hypothesis?

4. *p*-Values

The p-value measures the "extremeness" of the sample.

Definition: The p-value is the probability we would get the sample we have or something more extreme *if the null hypothesis were true*.

- This probability is calculated assuming that the null hypothesis is true.
- Beware: The *p*-value is not the probability that H_0 is true, nor is it an error probability!
- The *p*-value is between 0 and 1.

4. *p*-Values

Select a significance level α (as before, the desired *type I error probability*), then α defines the rejection region.

Then the decision rule is:

reject H_0 if P-value $\leq \alpha$

do not reject H_0 if P-value $> \alpha$

Thus if the *p*-value exceeds the chosen significance level, the null hypothesis cannot be rejected at that level.

Note, the *p*-value can be thought of as the smallest significance level at which H_0 can be rejected.

P-Values for *z* Tests

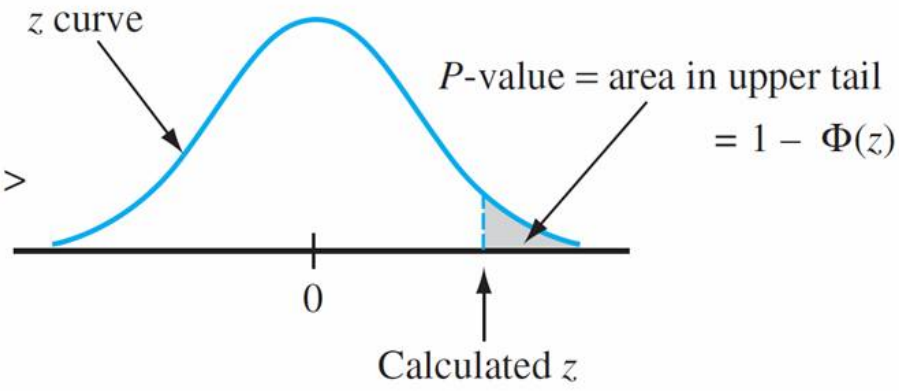
The calculation of the *P*-value depends on whether the test is upper-, lower-, or two-tailed.

$$P\text{-value: } P = \begin{cases} 1 - \Phi(z) & \text{for an upper-tailed } z \text{ test} \\ \Phi(z) & \text{or an lower-tailed } z \text{ test} \\ 2[1 - \Phi(|z|)] & \text{for a two-tailed } z \text{ test} \end{cases}$$

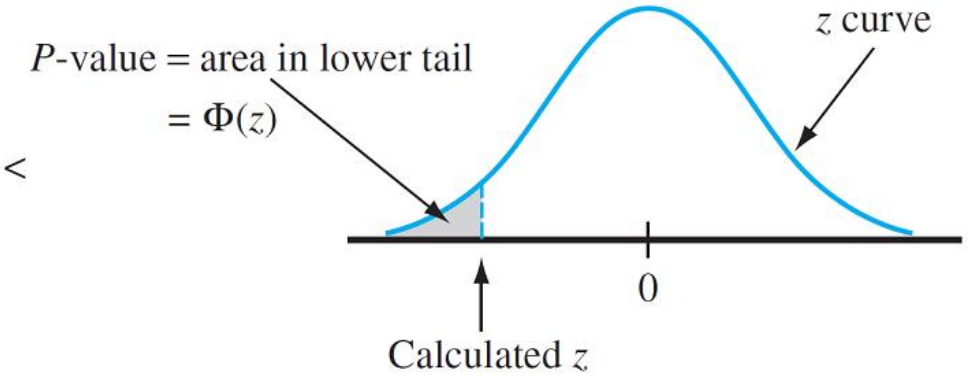
Each of these is the probability of getting a value at least as extreme as what was obtained (assuming H_0 true).

P-Values for *z* Tests

1. **Upper-tailed test**
   H_a contains the inequality >



2. **Lower-tailed test**
   H_a contains the inequality <

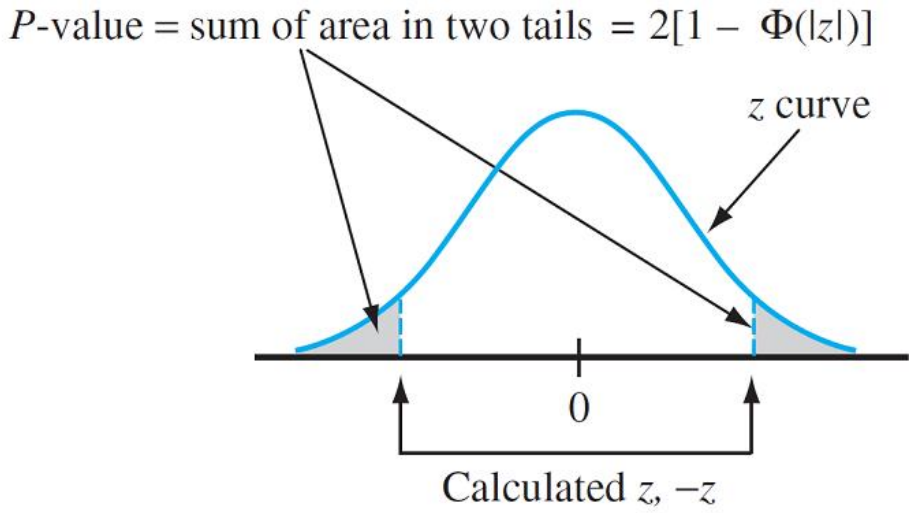


P-Values for *z* Tests

cont' d

3. **Two-tailed test**

   H_a contains the inequality $\neq$



Example

Back to the lawnmower engine example: There, we had

$$H_o: \mu = 300 \quad \text{vs} \quad H_a: \mu > 300$$

and

$$Z = 1.18$$

What is the p-value for this result?

Example

Back to the lawnmower engine example: There, we had

$$H_o: \mu = 300 \quad \text{vs} \quad H_a: \mu > 300$$

and

$$Z = 1.18$$

Assuming our average doesn't change much, what sample size would we need to see a statistically significant result?

Example

Back to the Brinell scale example: There, we had

$$H_o: \mu = 170 \quad \text{vs} \quad H_a: \mu \neq 170$$

and

$$T = 2.19$$

What is the p-value for this result?

Example

Back to the Brinell scale example: There, we had

$$H_o: \mu = 170 \quad \text{vs} \quad H_a: \mu \neq 170$$

and

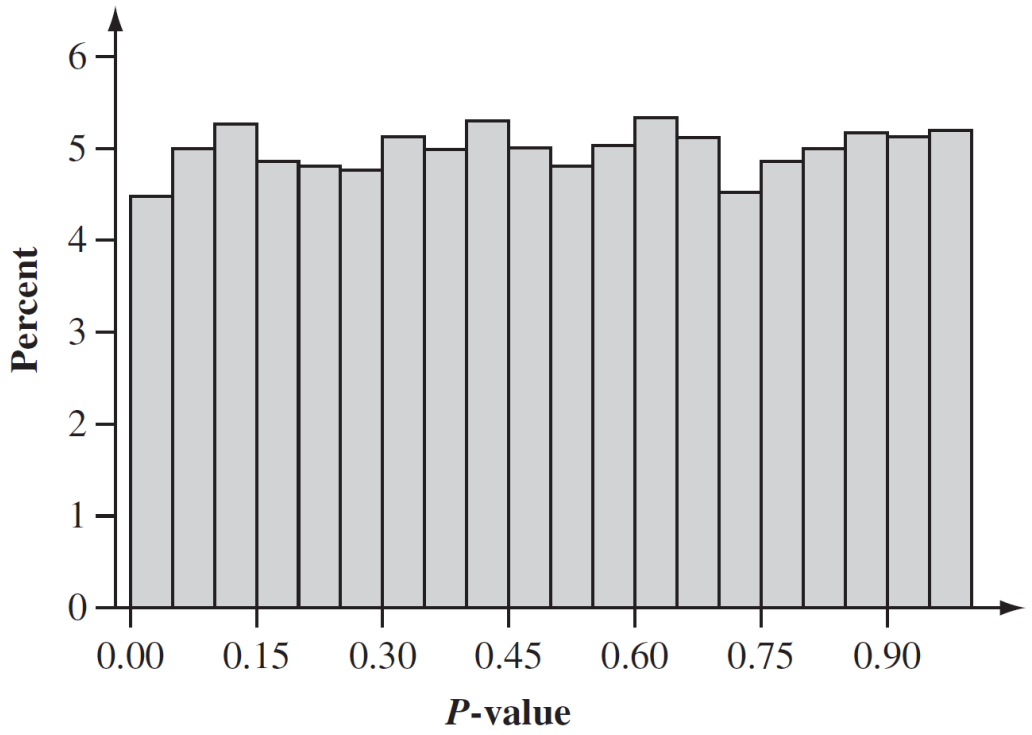
$$T = 2.19$$

What if we had used alpha = 0.05 instead?

Distribution of p-values

Figure below shows a histogram of the 10,000 *P*-values from a simulation experiment under a null $\mu = 20$ (with $n = 4$ and $\sigma = 2$).

When H_0 is true, the probability distribution of the *P*-value is a uniform distribution on the interval from 0 to 1.



Distribution of p-values

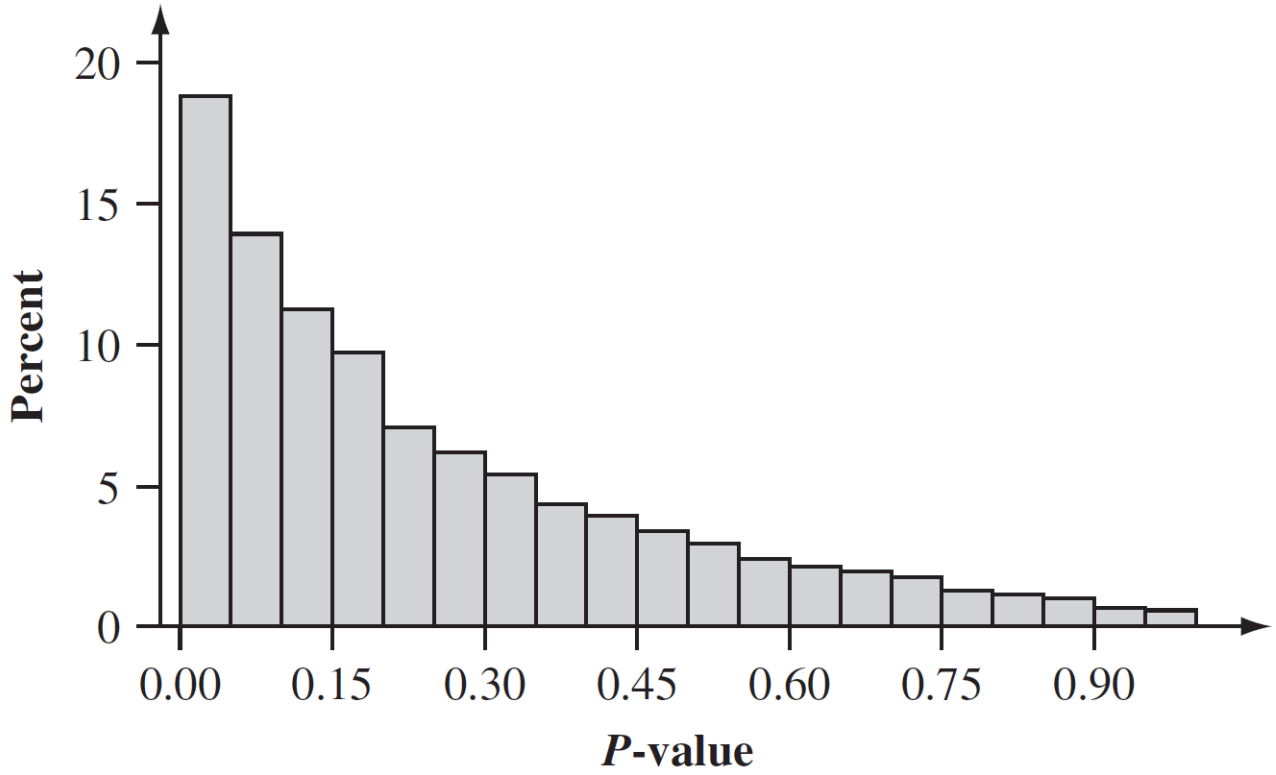
About 4.5% of these *P*-values are in the first class interval from 0 to .05.

Thus when using a significance level of .05, the null hypothesis is rejected in roughly 4.5% of these 10,000 tests.

If we continued to generate samples and carry out the test for each sample at significance level .05, in the long run 5% of the *P*-values would be in the first class interval.

Distribution of p-values

A histogram of the *P*-values when we simulate under an alternative hypothesis. There is a much greater tendency for the *P*-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$.



(b) $\mu = 21$

Distribution of p-values

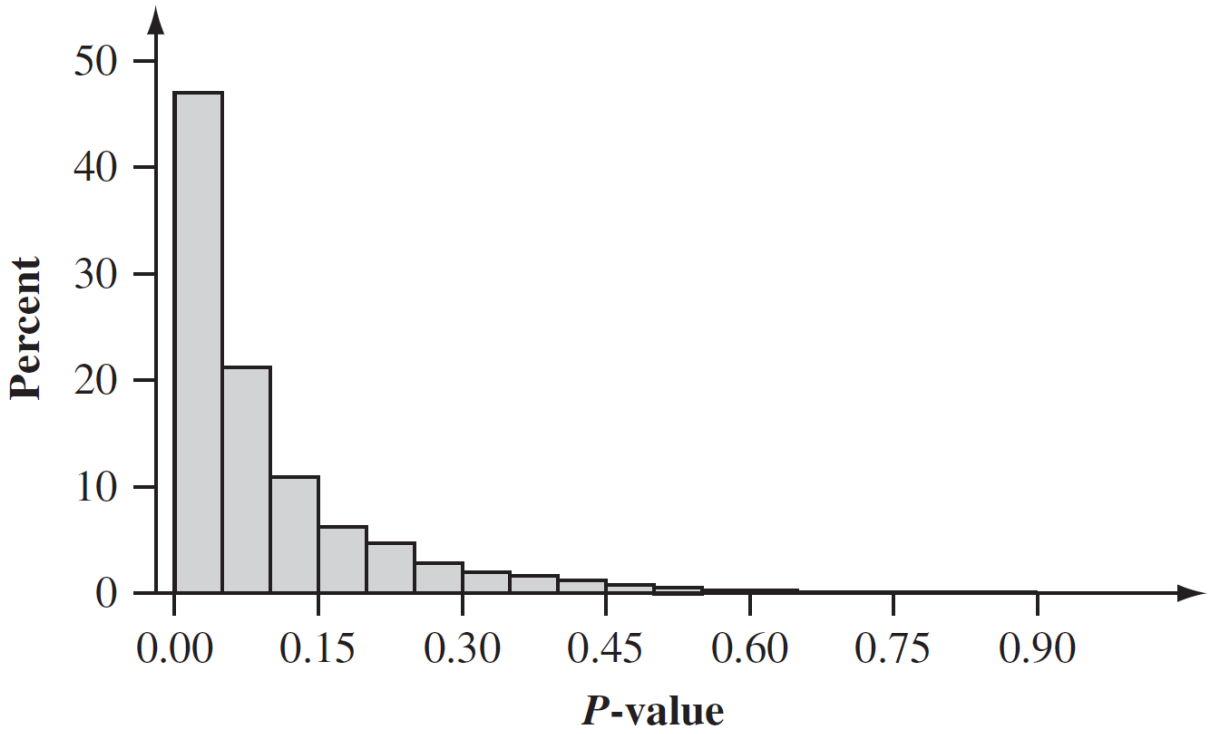
Again H_0 is rejected at significance level .05 whenever the *P*-value is at most .05 (in the first bin).

Unfortunately, this is the case for only about 19% of the *P*-values. So only about 19% of the 10,000 tests correctly reject the null hypothesis; for the other 81%, a type II error is committed.

The difficulty is that the sample size is quite small and 21 is not very different from the value asserted by the null hypothesis.

Distribution of p-values

Figure below illustrates what happens to the *P*-value when H_0 is false because $\mu = 22$.



(c) $\mu = 22$

Distribution of p-values

The histogram is even more concentrated toward values close to 0 than was the case when $\mu = 21$.

In general, as μ moves further to the right of the null value 20, the distribution of the *P*-value will become more and more concentrated on values close to 0.

Even here a bit fewer than 50% of the *P*-values are smaller than .05. So it is still slightly more likely than not that the null hypothesis is incorrectly not rejected. Only for values of μ much larger than 20 (e.g., at least 24 or 25) is it highly likely that the *P*-value will be smaller than .05 and thus give the correct conclusion.

Proportions: Large-Sample Tests

The estimator $\hat{p} = X/n$ is unbiased $(E(\hat{p}) = p)$, has approximately a normal distribution, and its standard deviation is $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$.

When H_0 is true, $E(\hat{p}) = p_0$ and $\sigma_{\hat{p}} = \sqrt{p_0(1-p_0)/n}$, so $\sigma_{\hat{p}}$ does not involve any unknown parameters. It then follows that when n is large and H_0 is true, the test statistic

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

has approximately a standard normal distribution.

Proportions: Large-Sample Tests

Alternative Hypothesis	**Rejection Region**
$H_a: p > p_0$	$z \geq z_\alpha$ (upper-tailed)
$H_a: p < p_0$	$z \leq -z_\alpha$ (lower-tailed)
$H_a: p \neq p_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed)

These test procedures are valid provided that $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

Example

Natural cork in wine bottles is subject to deterioration, and as a result wine in such bottles may experience contamination.

The article "Effects of Bottle Closure Type on Consumer Perceptions of Wine Quality" (*Amer. J. of Enology and Viticulture*, 2007: 182–191) reported that, in a tasting of commercial chardonnays, 16 of 91 bottles were considered spoiled to some extent by cork-associated characteristics.

Does this data provide strong evidence for concluding that more than 15% of all such bottles are contaminated in this way? Use a significance level equal to 0.10.

TWO SAMPLE TESTING

Normal Population, Known Variances

In general:

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$

Test statistic value: $z = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}}$

Test Procedures for Normal Populations with Known Variances

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$

Alternative Hypothesis	Rejection Region for Level α Test
$H_a: \mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$ (upper-tailed)
$H_a: \mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$ (lower-tailed)
$H_a: \mu_1 - \mu_2 \neq \Delta_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed)

Example 1

Analysis of a random sample consisting of 20 specimens of cold-rolled steel to determine yield strengths resulted in a sample average strength of $\bar{x} = 29.8$ ksi.

A second random sample of 25 two-sided galvanized steel specimens gave a sample average strength of $\bar{y} = 34.7$ ksi.

Assuming that the two yield-strength distributions are normal with $\sigma_1 = 4.0$ and $\sigma_2 = 5.0$, does the data indicate that the corresponding true average yield strengths μ_1 and μ_2 are different?

Let's carry out a test at significance level $\alpha = 0.01$

Large-Sample Tests

The assumptions of normal population distributions and known values of σ_1 and σ_2 are fortunately unnecessary when both sample sizes are sufficiently large. WHY?

Furthermore, using S_1^2 and S_2^2 in place of σ_1^2 and σ_2^2 gives a variable whose distribution is approximately standard normal:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

These tests are usually appropriate if both $m > 30$ and $n > 30$.

Example

Data on daily calorie intake both for a sample of teens who said they did not typically eat fast food and another sample of teens who said they did usually eat fast food.

Eat Fast Food	Sample Size	Sample Mean	Sample SD
No	663	2258	1519
Yes	413	2637	1138

Does this data provide strong evidence for concluding that true average calorie intake for teens who typically eat fast food exceeds more than 200 calories per day the true average intake for those who don't typically eat fast food?

Let's investigate by carrying out a test of hypotheses at a significance level of approximately .05.

The Two-Sample *t* Test

When the population distribution are both normal, the standardized variable

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \qquad \textbf{(9.2)}$$

has approximately a *t* distribution with df *v* estimated from the data by

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

The Two-Sample *t* Test

The **two-sample *t* test** for testing $H_0: \mu_1 - \mu_2 = \Delta_0$ is as follows:

Test statistic value: $t =$

$$\frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

The Two-Sample t Test

Alternative Hypothesis	**Rejection Region for Approximate Level α Test**
H_a: $\mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha,\nu}$ (upper-tailed)
H_a: $\mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha,\nu}$ (lower-tailed)
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	either $t \geq t_{\alpha/2,\nu}$ or $t \leq -t_{\alpha/2,\nu}$ (two-tailed)

A Test for Proportion Differences

Theoretically, we know that:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{pq\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

has approximately a standard normal distribution when H_0 is true.

However, this Z cannot serve as a test statistic because the value of p is unknown—H_0 asserts only that there is a common value of p, but does not say what that value is.

A Large-Sample Test Procedure

Under the null hypothesis, we assume that $p_1 = p_2 = p$, instead of separate samples of size m and n from two different populations (two different binomial distributions). So, we really have a single sample of size $m + n$ from one population with proportion p.

The total number of individuals in this combined sample having the characteristic of interest is $X + Y$.

The estimator of p is then

(9.5)

$$\hat{p} = \frac{X + Y}{m + n} = \frac{m}{m + n} \cdot \hat{p}_1 + \frac{n}{m + n} \cdot \hat{p}_2$$

A Large-Sample Test Procedure

Using $\hat{p}$ and $\hat{q} = 1 - \hat{p}$ in place of p and q in our old equation gives a test statistic having approximately a standard normal distribution when H_0 is true.

Null hypothesis: H_0: $p_1 - p_2 = 0$

Test statistic value (large samples):

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

A Large-Sample Test Procedure

Alternative Hypothesis	Rejection Region for Approximate Level α Test
H_a: $p_1 - p_2 > 0$	$z \geq z_a$
H_a: $p_1 - p_2 < 0$	$z \leq -z_a$
H_a: $p_1 - p_2 \neq 0$	either $z \geq z_{a/2}$ or $z \leq -z_{a/2}$

A *P*-value is calculated in the same way as for previous *z* tests.

The test can safely be used as long as $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2,$ and $n\hat{q}_2$ are all at least 10.

The *F* Test for Equality of Variances

The *F* Distribution

The *F* probability distribution has two parameters, denoted by v_1 and v_2. The parameter v_1 is called the *numerator degrees of freedom*, and v_2 is the *denominator degrees of freedom*.

A random variable that has an *F* distribution cannot assume a negative value. The density function is complicated and will not be used explicitly, so it's not shown.

There is an important connection between an *F* variable and chi-squared variables.

The *F* Distribution

If X_1 and X_2 are independent chi-squared rv's with v_1 and v_2 df, respectively, then the rv

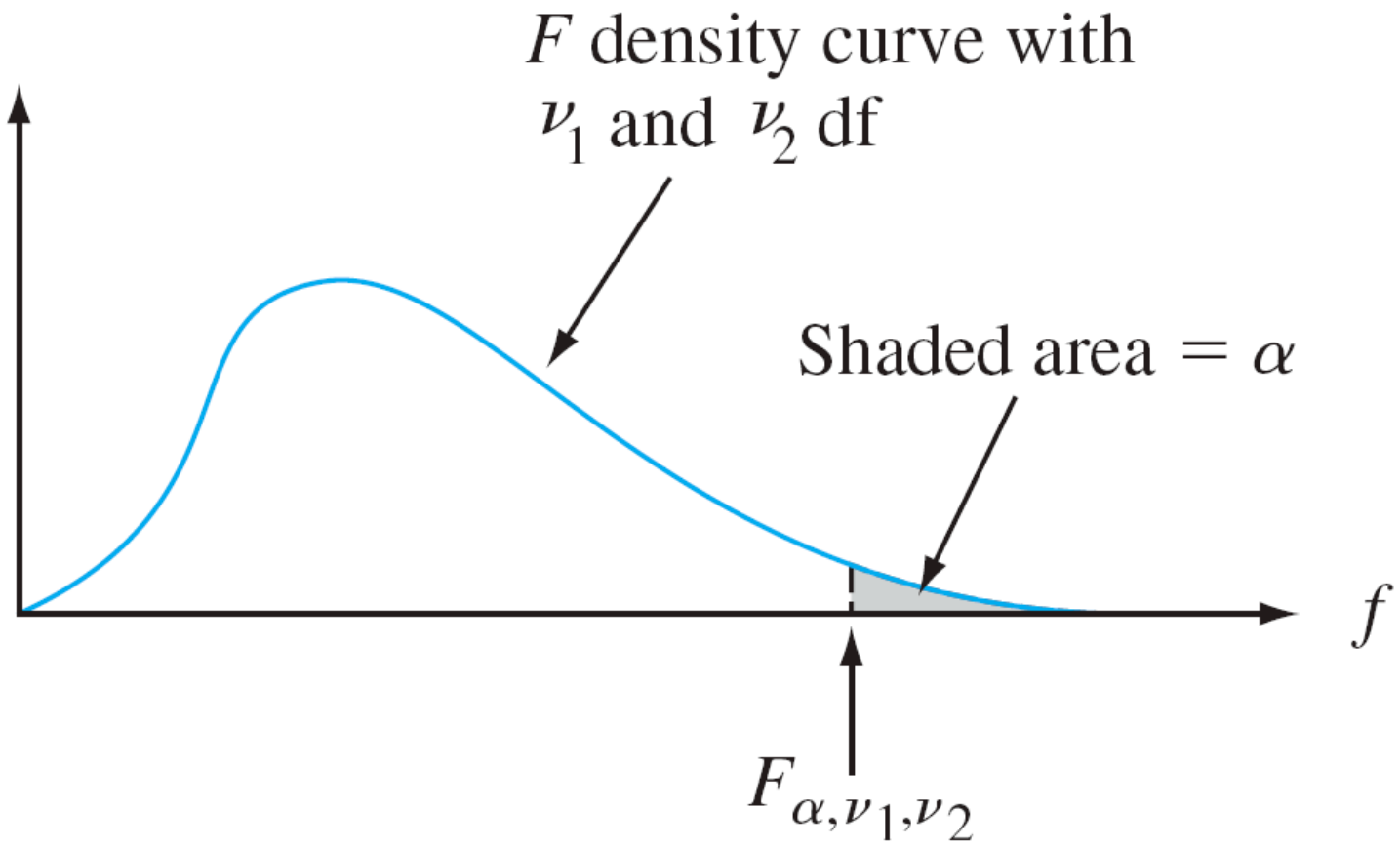
$$F = \frac{X_1/v_1}{X_2/v_2}$$

can be shown to have an *F* distribution.

Recall that a chi-squared distribution was obtain by summing squared standard Normal variables (such as squared deviations for example). So a scaled ratio of two variances is a ratio of two scaled chi-squared variables.

The *F* Distribution

Figure below illustrates a typical *F* density function.



The *F* Distribution

We use F_{α,v_1,v_2} for the value on the horizontal axis that captures α of the area under the *F* density curve with v_1 and v_2 df in the upper tail.

The density curve is not symmetric, so it would seem that both upper- and lower-tail critical values must be tabulated. This is not necessary, though, because of the fact that

$$F_{1-\alpha,v_1,v_2} = 1/F_{\alpha,v_2,v_1}.$$

The *F* Distribution

We use F_{α,v_1,v_2} for the value on the horizontal axis that captures α of the area under the *F* density curve with v_1 and v_2 df in the upper tail.

The density curve is not symmetric, so it would seem that both upper- and lower-tail critical values must be tabulated. This is not necessary, though, because of the fact that

$$F_{1-\alpha,v_1,v_2} = 1/F_{\alpha,v_2,v_1}.$$

For example, $F_{.05,6,10} = 3.22$ and $F_{.95,10,6} = 0.31 = 1/3.22$.

The *F* Test for Equality of Variances

A test procedure for hypotheses concerning the ratio σ_1^2/σ_2^2 is based on the following result.

Theorem

Let $X_1,\ldots, X_m$ be a random sample from a normal distribution with variance σ_1^2, let $Y_1,\ldots, Y_n$ be another random sample (independent of the X_i's) from a normal distribution with variance σ_2^2, and let S_1^2 and S_2^2 denote the two sample variances. Then the rv

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an *F* distribution with $v_1 = m - 1$ and $v_2 = n - 1$.

The *F* Test for Equality of Variances

This theorem results from combining the fact that the variables $(m-1)S_1^2/\sigma_1^2$ and $(n-1)S_2^2/\sigma_2^2$ each have a chi-squared distribution with $m-1$ and $n-1$ df, respectively.

Because *F* involves a ratio rather than a difference, the test statistic is the ratio of sample variances.

The claim that $\sigma_1^2 = \sigma_2^2$ is then rejected if the ratio differs by too much from 1.

The *F* Test for Equality of Variances

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

Test statistic value: $f = s_1^2/s_2^2$

Alternative Hypothesis	**Rejection Region for a Level α Test**
$H_a: \sigma_1^2 > \sigma_2^2$	$f \geq F_{\alpha,m-1,n-1}$
$H_a: \sigma_1^2 < \sigma_2^2$	$f \leq F_{1-\alpha,m-1,n-1}$
$H_a: \sigma_1^2 \neq \sigma_2^2$	either $f \geq F_{\alpha/2,m-1,n-1}$ or $f \leq F_{1-\alpha/2,m-1,n-1}$

Example

On the basis of data reported in the article "Serum Ferritin in an Elderly Population" (*J. of Gerontology*, 1979: 521–524), the authors concluded that the ferritin distribution in the elderly had a smaller variance than in the younger adults. (Serum ferritin is used in diagnosing iron deficiency.)

For a sample of 28 elderly men, the sample standard deviation of serum ferritin (mg/L) was $s_1 = 52.6$; for 26 young men, the sample standard deviation was $s_2 = 84.2$.

Does this data support the conclusion as applied to men? Use alpha = .01.