

AMA 2634

Introduction to Statistics

2024 Fall

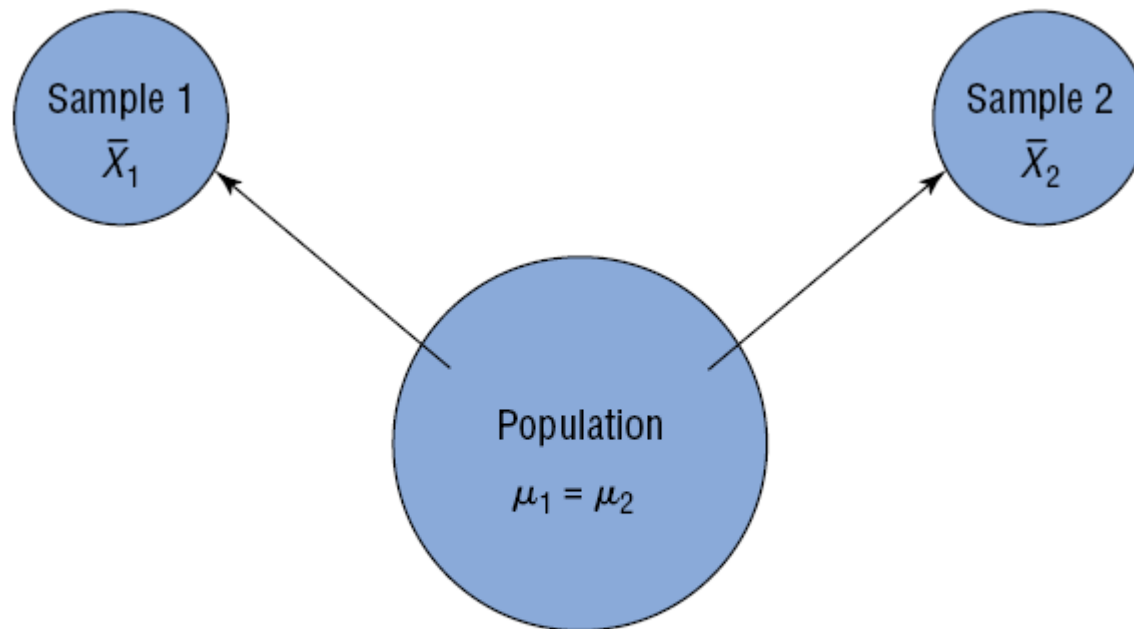
Lecture 9

Testing the Difference Between Two Means, Two Proportions, and Two Variances



Assumptions:

1. The samples must be independent of each other. That is, there can be no relationship between the subjects in each sample.
2. The standard deviations of both populations must be known, and if the sample sizes are less than 30, the populations must be normally or approximately normally distributed.



(a) Difference is not significant

Do not reject $H_0: \mu_1 = \mu_2$ since $\bar{X}_1 - \bar{X}_2$ is not significant.



(b) Difference is significant

Reject $H_0: \mu_1 = \mu_2$ since $\bar{X}_1 - \bar{X}_2$ is significant.



Formula for the z test for comparing two means from independent populations

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$



A survey found that the average hotel room rate in New Orleans is \$88.42 and the average room rate in Phoenix is \$80.61.

Assume that the data were obtained from two samples of 50 hotels each and that the standard deviations of the populations are \$5.62 and \$4.83, respectively. At $\alpha = 0.05$, can it be concluded that there is a significant difference in the rates?

Step 1: State the hypotheses and identify the claim.

$$H_0: \mu_1 = \mu_2 \text{ and } H_1: \mu_1 \neq \mu_2 \text{ (claim)}$$

Step 2: Find the critical value.

The critical value is $z = \pm 1.96$.

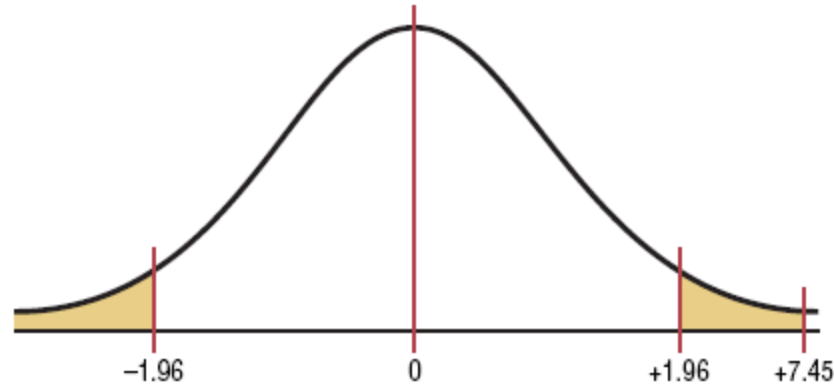


Step 3: Compute the test value.

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$
$$z = \frac{(88.42 - 80.61) - (0)}{\sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}}} = 7.45$$

Step 4: Make the decision.

Reject the null hypothesis at $\alpha = 0.05$, since $7.45 > 1.96$.



Step 5: Summarize the results.

There is enough evidence to support the claim that the means are not equal. Hence, there is a significant difference in the rates.

Example: College Sports Offerings



A researcher hypothesizes that the average number of sports that colleges offer for males is greater than the average number of sports that colleges offer for females. A sample of the number of sports offered by colleges is shown. At $\alpha = 0.10$, is there enough evidence to support the claim? Assume σ_1 and $\sigma_2 = 3.3$.

Males					Females				
6	11	11	8	15	6	8	11	13	8
6	14	8	12	18	7	5	13	14	6
6	9	5	6	9	6	5	5	7	6
6	9	18	7	6	10	7	6	5	5
15	6	11	5	5	16	10	7	8	5
9	9	5	5	8	7	5	5	6	5
8	9	6	11	6	9	18	13	7	10
9	5	11	5	8	7	8	5	7	6
7	7	5	10	7	11	4	6	8	7
10	7	10	8	11	14	12	5	8	5



Step 1: State the hypotheses and identify the claim.

$$H_0: \mu_1 = \mu_2 \text{ and } H_1: \mu_1 > \mu_2 \text{ (claim)}$$

Step 2: Compute the test value.

For the males: $\bar{X}_1 = 8.6$ and $\sigma_1 = 3.3$

For the females: $\bar{X}_2 = 7.9$ and $\sigma_2 = 3.3$

Substitute in the formula.

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(8.6 - 7.9) - (0)}{\sqrt{\frac{3.3^2}{50} + \frac{3.3^2}{50}}} = 1.06$$

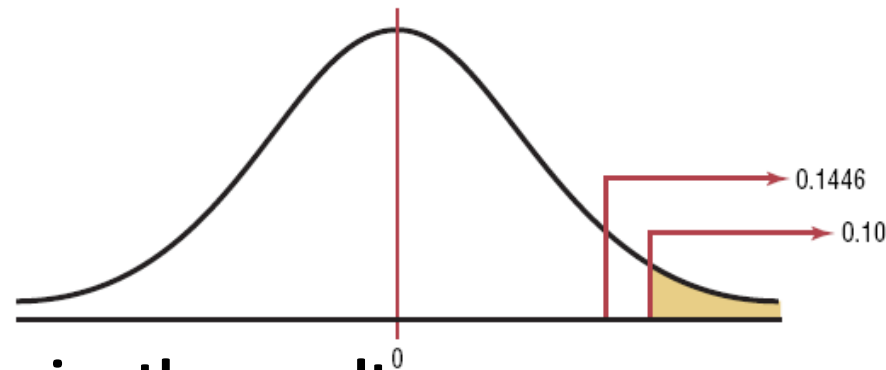
Step 3: Find the *P*-value.

For $z = 1.06$, the area is 0.8554.

The *P*-value is $1.0000 - 0.8554 = 0.1446$.

Step 4: Make the decision.

Do not reject the null hypothesis.



Step 5: Summarize the results.

There is not enough evidence to support the claim that colleges offer more sports for males than they do for females.



Formula for the z confidence interval for the difference between two means from independent populations

$$\begin{aligned} \left(\bar{X}_1 - \bar{X}_2 \right) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &< \left(\mu_1 - \mu_2 \right) \\ &< \left(\bar{X}_1 - \bar{X}_2 \right) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \end{aligned}$$

Find the 95% confidence interval for the difference between the means for the data in Example 9–1.

$$\begin{aligned}(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 \\ < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\end{aligned}$$

$$\begin{aligned}(88.42 - 80.61) - 1.96 \sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}} < \mu_1 - \mu_2 \\ < (88.42 - 80.61) + 1.96 \sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}}\end{aligned}$$

$$7.81 - 2.05 < \mu_1 - \mu_2 < 7.81 + 2.05$$

$$5.76 < \mu_1 - \mu_2 < 9.86$$



Formula for the t test for comparing two means from independent populations **with unequal variances**

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

where the degrees of freedom are equal to the smaller of $n_1 - 1$ or $n_2 - 1$.



There is a different formula for the t test for comparing two means from independent populations **with equal variances**. To determine whether two sample variances are equal, the researcher can use an F test.

Note, however, that not all statisticians are in agreement about using the F test before using the t test. Some believe that conducting the F and t tests at the same level of significance will change the overall level of significance of the t test. Their reasons are beyond the scope of this course. Because of this, we will assume that $\sigma_1 \neq \sigma_2$ in this course.

Example: Farm Sizes



The average size of a farm in Indiana County, Pennsylvania, is 199 acres. The average size of a farm in Greene County, Pennsylvania, is 199 acres. Assume the data were obtained from two samples with standard deviations of 38 and 12 acres, respectively, and sample sizes of 8 and 10, respectively. Can it be concluded at $\alpha = 0.05$ that the average size of the farms in the two counties is different? Assume the populations are normally distributed.

Step 1: State the hypotheses and identify the claim.

$$H_0: \mu_1 = \mu_2 \text{ and } H_1: \mu_1 \neq \mu_2 \text{ (claim)}$$

Step 2: Find the critical values.

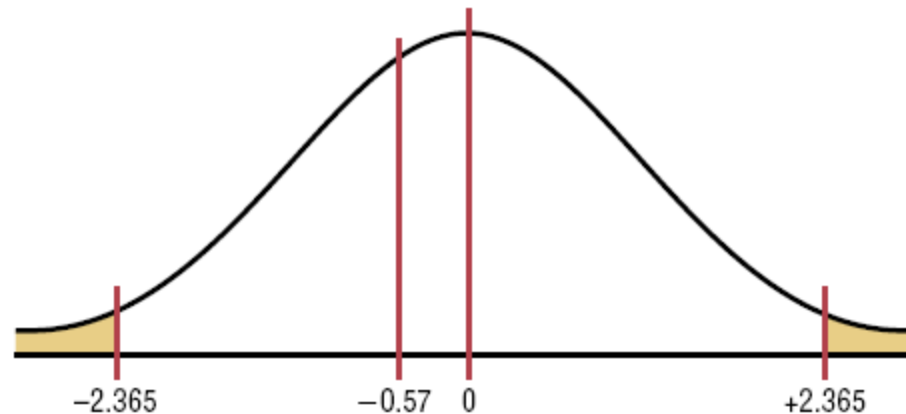
Since the test is two-tailed, $\alpha = 0.05$, and the variances are unequal, the degrees of freedom are the smaller of $n_1 - 1$ or $n_2 - 1$. In this case, the degrees of freedom are $8 - 1 = 7$. Hence, from Table F, the critical values are -2.365 and 2.365.

Step 3: Find the test value.

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(191 - 199) - (0)}{\sqrt{\frac{38^2}{8} + \frac{12^2}{10}}} \\ = -0.57$$

Step 4: Make the decision.

Do not reject the null hypothesis.



Step 5: Summarize the results.

There is not enough evidence to support the claim that the average size of the farms is different.



Formula for the t confidence interval for the difference between two means from independent populations **with unequal variances**

$$\begin{aligned} \left(\bar{X}_1 - \bar{X}_2 \right) - t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} &< \left(\mu_1 - \mu_2 \right) \\ &< \left(\bar{X}_1 - \bar{X}_2 \right) + t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \end{aligned}$$

d.f. smaller value of $n_1 - 1$ or $n_2 - 1$.

Example: Confidence Intervals



Find the 95% confidence interval for the difference between the means for the data in Example 9–4 in the textbook

$$\begin{aligned}(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} &< \mu_1 - \mu_2 \\ &< (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\end{aligned}$$

$$\begin{aligned}(191 - 199) - 2.365 \sqrt{\frac{38^2}{8} + \frac{12^2}{10}} &< \mu_1 - \mu_2 \\ &< (191 - 199) + 2.365 \sqrt{\frac{38^2}{8} + \frac{12^2}{10}} \\ -41.02 &< \mu_1 - \mu_2 < 25.02\end{aligned}$$



When the values are dependent, do a t test on the differences. Denote the differences with the symbol D , the mean of the population differences with μ_D , and the sample standard deviation of the differences with s_D .

$$t = \frac{\bar{D} - \mu_D}{s_D / \sqrt{n}}$$

with d.f. = $n - 1$ and where

$$\bar{D} = \frac{\sum D}{n} \text{ and } s_D = \sqrt{\frac{n \sum D^2 - (\sum D)^2}{n(n-1)}}$$

Example: Cholesterol Levels



A dietitian wishes to see if a person's cholesterol level will change if the diet is supplemented by a certain mineral. Six subjects were pretested, and then they took the mineral supplement for a 6-week period. The results are shown in the table. (Cholesterol level is measured in milligrams per deciliter.) Can it be concluded that the cholesterol level has been changed at $\alpha = 0.10$? Assume the variable is approximately normally distributed.

Subject	1	2	3	4	5	6
Before (X_1)	210	235	208	190	172	244
After (X_2)	190	170	210	188	173	228

Example: Cholesterol Levels



Subject	1	2	3	4	5	6
Before (X_1)	210	235	208	190	172	244
After (X_2)	190	170	210	188	173	228

Step 1: State the hypotheses and identify the claim.

$$H_0: \mu_D = 0 \text{ and } H_1: \mu_D \neq 0 \text{ (claim)}$$

Step 2: Find the critical value.

The degrees of freedom are 5. At $\alpha = 0.10$, the critical values are ± 2.015 .

Step 3: Compute the test value.

Before (X_1)	After (X_2)	$D = X_1 - X_2$	D^2
210	190	20	400
235	170	65	4225
208	210	-2	4
190	188	2	4
172	173	-1	1
244	228	16	256
		$\Sigma D = 100$	$\Sigma D^2 = 4890$

$$\bar{D} = \frac{\sum D}{n} = \frac{100}{6} = 16.7$$

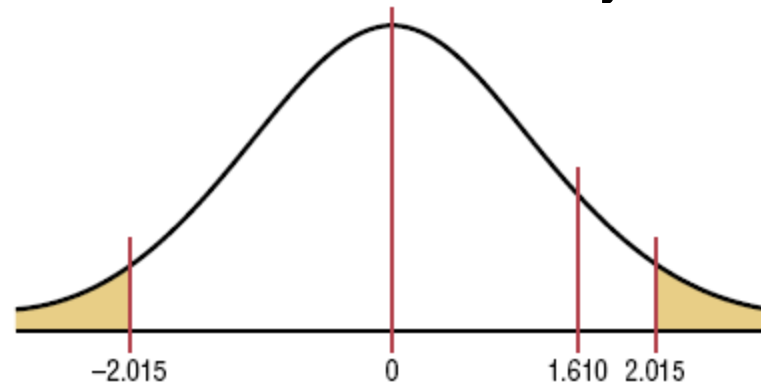
$$s_D = \sqrt{\frac{n \sum D^2 - (\sum D)^2}{n(n-1)}} = \sqrt{\frac{6 \cdot 4890 - (100)^2}{6 \cdot 5}} = 25.4$$

Step 3: Compute the test value.

$$\bar{D} = 16.7, s_D = 25.4$$

$$t = \frac{\bar{D} - \mu_D}{s_D / \sqrt{n}} = \frac{16.7 - 0}{25.4 / \sqrt{6}} = 1.610$$

Step 4: Make the decision. Do not reject the null.



Step 5: Summarize the results.

There is not enough evidence to support the claim that the mineral changes a person's cholesterol level.



Formula for the t confidence interval for the mean difference

$$\bar{D} - t_{\alpha/2} \frac{s_D}{\sqrt{n}} < \mu_D < \bar{D} + t_{\alpha/2} \frac{s_D}{\sqrt{n}}$$

$$\text{d.f.} = n - 1$$

Find the 90% confidence interval for the difference between the means for the data in Example on p.22.

$$\bar{D} - t_{\alpha/2} \frac{s_D}{\sqrt{n}} < \mu_D < \bar{D} + t_{\alpha/2} \frac{s_D}{\sqrt{n}}$$

$$16.7 - 2.015 \cdot \frac{25.4}{\sqrt{6}} < \mu_D < 16.7 + 2.015 \cdot \frac{25.4}{\sqrt{6}}$$

$$16.7 - 20.89 < \mu_D < 16.7 + 20.89$$

$$-4.19 < \mu_D < 37.59$$

Since 0 is contained in the interval, the decision is to not reject the null hypothesis $H_0: \mu_D = 0$.



$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where

$$\begin{aligned} \bar{p} &= \frac{X_1 + X_2}{n_1 + n_2} & \hat{p}_1 &= \frac{X_1}{n_1} \\ \bar{q} &= 1 - \bar{p} & \hat{p}_2 &= \frac{X_2}{n_2} \end{aligned}$$

Example: Vaccination Rates



In the nursing home study mentioned in the chapter-opening Statistics Today, the researchers found that 12 out of 34 small nursing homes had a resident vaccination rate of less than 80%, while 17 out of 24 large nursing homes had a vaccination rate of less than 80%. At $\alpha = 0.05$, test the claim that there is no difference in the proportions of the small and large nursing homes with a resident vaccination rate of less than 80%.

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{12}{34} = 0.35 \quad \text{and} \quad \hat{p}_2 = \frac{X_2}{n_2} = \frac{17}{24} = 0.71$$

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{12 + 17}{34 + 24} = \frac{29}{58} = 0.5, \bar{q} = 0.5$$

Step 1: State the hypotheses and identify the claim.

$$H_0: p_1 - p_2 = 0 \text{ (claim) and } H_1: p_1 - p_2 \neq 0$$

Step 2: Find the critical value.

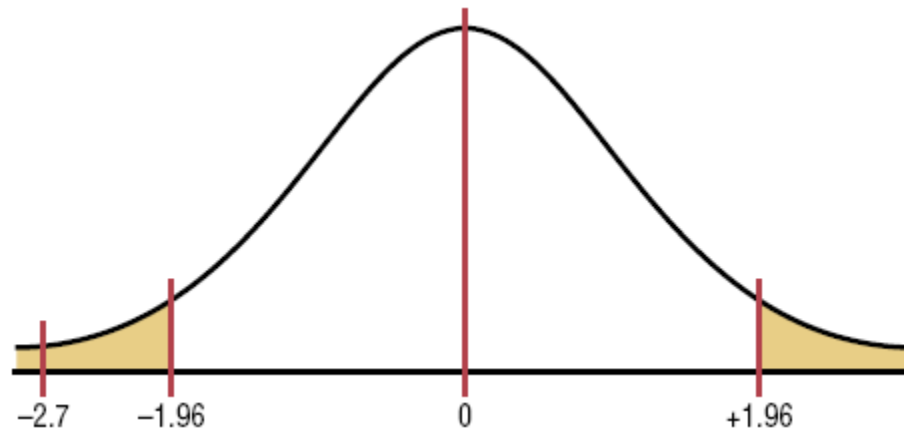
Since $\alpha = 0.05$, the critical values are -1.96 and 1.96 .

Step 3: Compute the test value.

$$\begin{aligned} z &= \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.35 - 0.71) - (0)}{\sqrt{(0.5)(0.5)\left(\frac{1}{34} + \frac{1}{24}\right)}} \\ &= -2.7 \end{aligned}$$

Step 4: Make the decision.

Reject the null hypothesis.



Step 5: Summarize the results.

There is enough evidence to reject the claim that there is no difference in the proportions of small and large nursing homes with a resident vaccination rate of less than 80%.



A survey of 1000 drivers this year showed that 29% of the people send text messages while driving. Last year a survey of 1000 drivers showed that 17% of those send text messages while driving.

At $\alpha = 0.01$, can it be concluded that there has been an increase in the number of drivers who text while driving?

Step 1: State the hypotheses and identify the claim.

$$H_0: p_1 = p_2 \text{ and } H_1: p_1 > p_2 \text{ (claim)}$$

Step 2: Find the critical value.

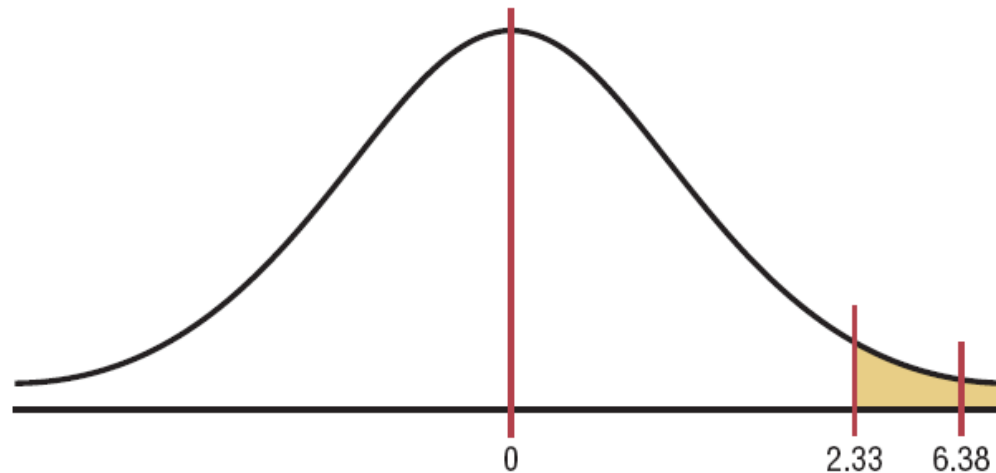
Since $\alpha = 0.01$, the critical value is 2.33.

Step 3: Compute the test value.

$$\begin{aligned} z &= \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{(0.29 - 0.17) - 0}{\sqrt{(0.23)(0.77)\left(\frac{1}{1000} + \frac{1}{1000}\right)}} = 6.38 \end{aligned}$$

Step 4: Make the decision.

Reject the null hypothesis since $6.38 > 2.33$.



Step 5: Summarize the results.

There is enough evidence to say that the proportion of drivers who send text messages is larger today than it was last year.



Formula for the confidence interval for the difference between proportions

$$\begin{aligned} (\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 \\ < (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \end{aligned}$$

Find the 95% confidence interval for the difference of the proportions for the data in Example on p.29.

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{12}{34} = 0.35 \quad \text{and} \quad \hat{q}_1 = 1 - 0.35 = 0.65$$

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{17}{24} = 0.71 \quad \text{and} \quad \hat{q}_2 = 1 - 0.71 = 0.29$$

$$\begin{aligned} (\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} &< p_1 - p_2 \\ &< (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \end{aligned}$$

$$(0.35 - 0.71) - 1.96 \sqrt{\frac{(0.35)(0.65)}{34} + \frac{(0.71)(0.29)}{24}} < p_1 - p_2$$

$$< (0.35 - 0.71) + 1.96 \sqrt{\frac{(0.35)(0.65)}{34} + \frac{(0.71)(0.29)}{24}}$$

$$-0.36 - 0.242 < p_1 - p_2 < -0.36 + 0.242$$

$$-0.602 < p_1 - p_2 < -0.118$$

Since 0 is not contained in the interval, the decision is to reject the null hypothesis $H_0: p_1 = p_2$.



In addition to comparing two means, statisticians are interested in comparing **two variances** or **standard deviations**.

For the comparison of two variances or standard deviations, an F test is used.

The F test should not be confused with the chi-square test, which compares a single sample variance to a specific population variance, as shown in Lecture Note 8.

The F -distribution with d_1 and d_2 degrees of freedom is the distribution of

$$F = \frac{U_1/d_1}{U_2/d_2}$$

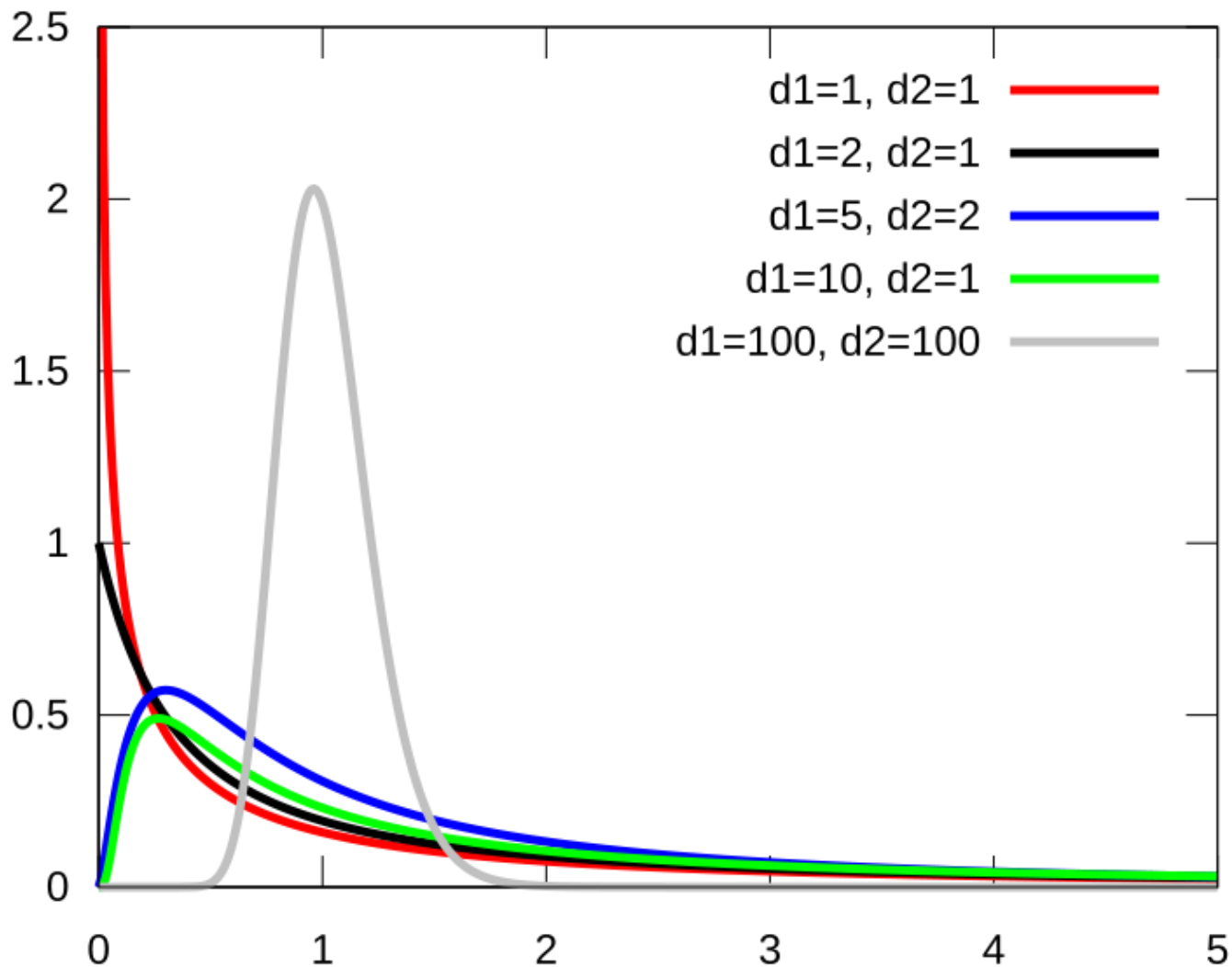
where U_1 and U_2 are independent random variable with chi-square distributions with degree of freedom d_1 and d_2 respectively. $U_1 \sim \chi^2(d_1)$ and $U_2 \sim \chi^2(d_2)$.

$$E[F] = \frac{d_2}{d_2-2} \text{ for } d_2 > 2$$

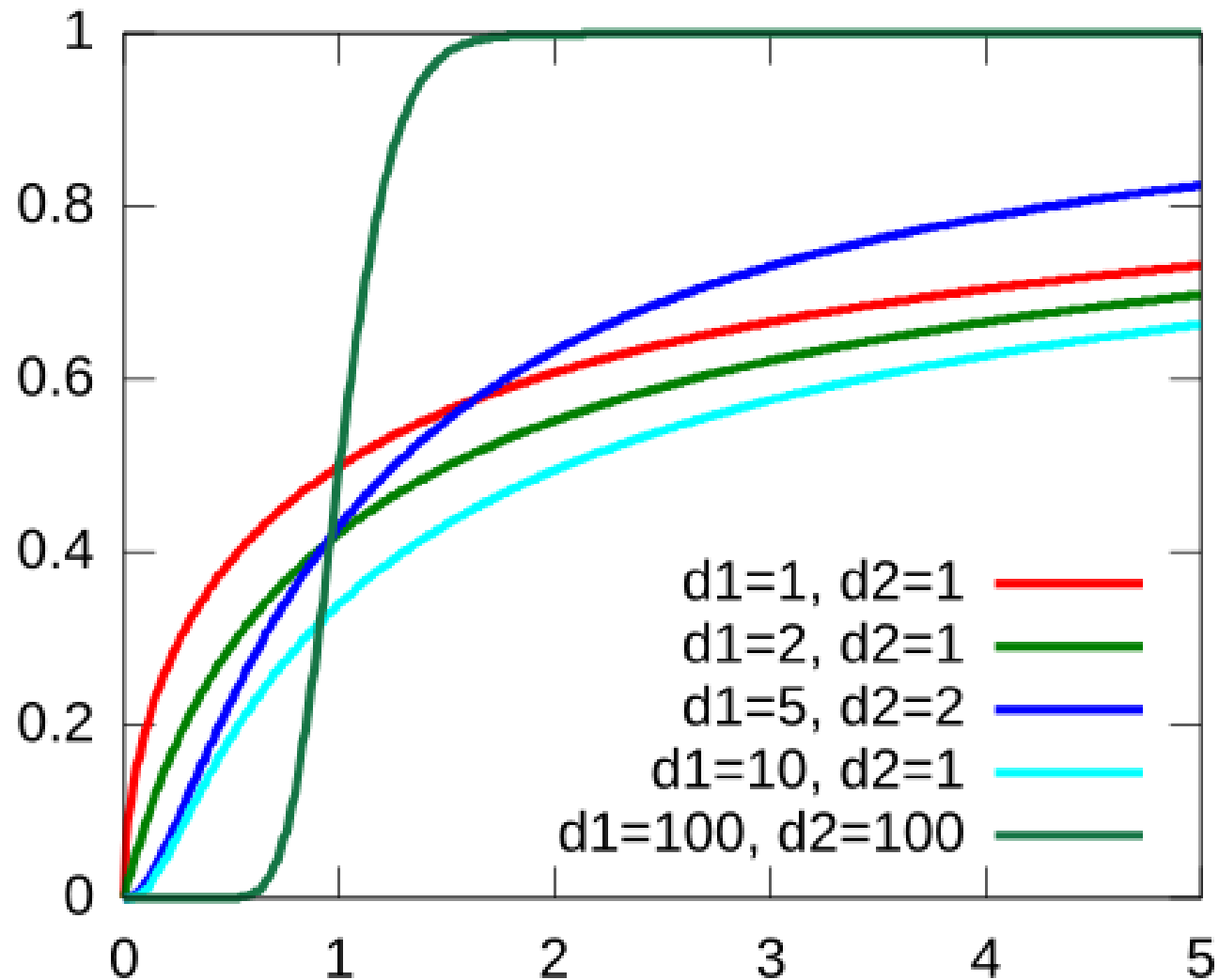


1. The values of F cannot be negative, because variances are always positive or zero.
2. The distribution is positively skewed.
3. The mean value of F is approximately equal to 1.
4. The F distribution is a family of curves based on the degrees of freedom of the variance of the numerator and the degrees of freedom of the variance of the denominator.

PDF of F Distribution



CDF of F Distribution





$$F = \frac{s_1^2}{s_2^2}$$

where the larger of the two variances is placed in the numerator regardless of the subscripts. (to force the test into the right tail. Right-tailed tests are easier to calculate.)

The F test has two terms for the degrees of freedom: that of the numerator, $n_1 - 1$, and that of the denominator, $n_2 - 1$, where n_1 is the sample size from which the larger variance was obtained.



1. The populations from which the samples were obtained must be normally distributed. (Note: The test should not be used when the distributions depart from normality.)
2. The samples must be independent of each other.



A medical researcher wishes to see whether the variance of the heart rates (in beats per minute) of smokers is different from the variance of heart rates of people who do not smoke. Two samples are selected, and the data are as shown. Using $\alpha = 0.05$, is there enough evidence to support the claim?

Smokers	Nonsmokers
$n_1 = 26$	$n_2 = 18$
$s_1^2 = 36$	$s_2^2 = 10$

Step 1: State the hypotheses and identify the claim.

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{and} \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \text{ (claim)}$$

Smokers	Nonsmokers
$n_1 = 26$	$n_2 = 18$
$s_1^2 = 36$	$s_2^2 = 10$

Step 2: Find the critical value.

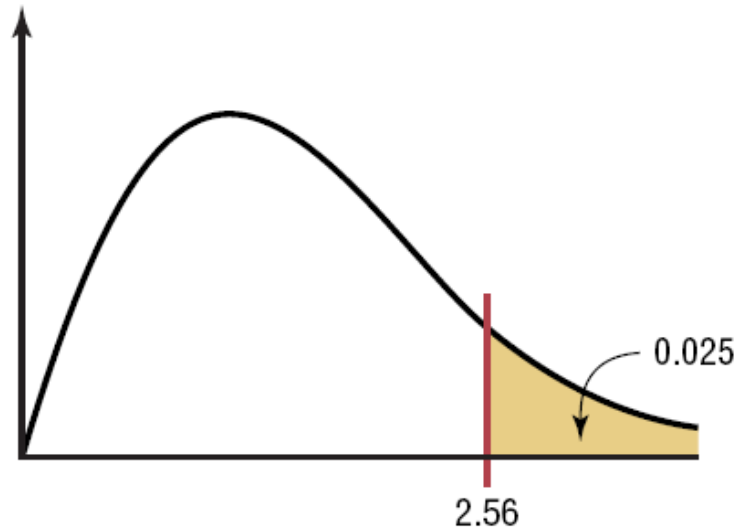
since $\alpha = 0.05$ and this is a two-tailed test. Here, d.f.N. = 25, and d.f.D. = 17. The critical value is 2.56 (d.f.N. 24 was used).

Step 3: Compute the test value.

$$F = \frac{s_1^2}{s_2^2} = \frac{36}{10} = 3.6$$

Step 4: Make the decision.

Reject the null hypothesis, since $3.6 > 2.56$.



Step 5: Summarize the results.

There is enough evidence to support the claim that the variance of the heart rates of smokers and nonsmokers is different.



The standard deviation of the average waiting time to see a doctor for non-lifethreatening problems in the emergency room at an urban hospital is 32 minutes. At a second hospital, the standard deviation is 28 minutes. If a sample of 16 patients was used in the first case and 18 in the second case, is there enough evidence to conclude that the standard deviation of the waiting times in the first hospital is greater than the standard deviation of the waiting times in the second hospital?

Step 1: State the hypotheses and identify the claim.

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{and} \quad H_1 : \sigma_1^2 > \sigma_2^2 \text{ (claim)}$$



Step 2: Find the critical value.

Here, d.f.N. = 15, d.f.D. = 17, and $\alpha = 0.01$.

The critical value is $F = 3.31$.

Step 3: Compute the test value.

$$F = \frac{s_1^2}{s_2^2} = \frac{32^2}{28^2} = 1.31$$

Step 4: Make the decision.

Do not reject the null hypothesis since $1.31 < 3.31$.

Step 5: Summarize the results.

There is not enough evidence to support the claim that the standard deviation of the waiting times of the first hospital is greater than the standard deviation of the waiting times of the second hospital.