

SCST211

APPLIED STATISTICS

Instructor: Abrar Abu awf

CHAPTER 7

CONFIDENCE INTERVALS

AND SAMPLE SIZE

OUTLINE

Introduction

7-1 Confidence Intervals for the Mean When σ Is Known

7-2 Confidence Intervals for the Mean When σ Is Unknown

7-3 Confidence Intervals and Sample Size for Proportions

7-4 Confidence Intervals for Variances and Standard Deviations

INTRODUCTION

- One aspect of inferential statistics is estimation, which is the process of estimating the value of a parameter from information obtained from a sample. For example, consider the following statements:

“The average amount employees and employers pay for health insurance is \$11,664 per year.”

“Eight percent of the people surveyed in the United States said that they participate in skiing in the winter time.”

“Consumers spent an average of \$126 for Valentine’s Day this year.”

- Since the populations from which these values were obtained are large, these values are only **estimates** of the true parameters and are derived from data collected from samples.
- The statistical procedures for estimating the population mean be explained in this chapter.

7-1 CONFIDENCE INTERVALS FOR THE MEAN WHEN σ IS KNOWN

- A **point estimate** is a specific numerical value estimate of a parameter. The best point estimate of the population mean μ is the sample mean $\bar{X}$.
- Sample measures (i.e., statistics) are used to estimate population measures (i.e., parameters). These statistics are called **estimators**.
- **Three Properties of a Good Estimator:**
 1. The estimator should be an **unbiased estimator**. That is, the expected value or the mean of the estimates obtained from samples of a given size is equal to the parameter being estimated.
 2. The estimator should be consistent. For a **consistent estimator**, as sample size increases, the value of the estimator approaches the value of the parameter estimated.
 3. The estimator should be a **relatively efficient estimator**. That is, of all the statistics that can be used to estimate a parameter, the relatively efficient estimator has the smallest variance.

Confidence Intervals

- An **interval estimate** of a parameter is an interval or a range of values used to estimate the parameter. This estimate may or may not contain the value of the parameter being estimated.
- The **confidence level** of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated.
- A **confidence interval** is a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate.
- Three common confidence intervals are used: the **90%**, the **95%**, and the **99%** confidence intervals.

Formula for the Confidence Interval of the Mean for a Specific α When σ is Known

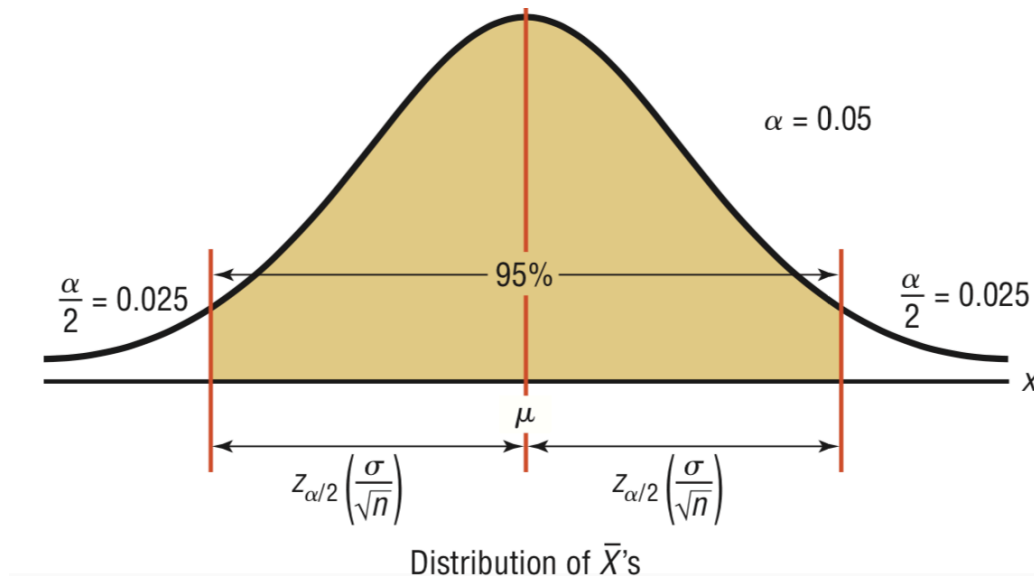
$$\bar{X} - z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{X} + z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

For a 90% confidence interval, $z_{\alpha/2} = 1.65$; for a 95% confidence interval, $z_{\alpha/2} = 1.96$; and for a 99% confidence interval, $z_{\alpha/2} = 2.58$.

Assumptions for Finding a Confidence Interval for a Mean When σ Is Known

1. The sample is a random sample.
2. Either $n \geq 30$ or the population is normally distributed when $n < 30$.

The term $\mathbf{Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}}$ is called the **marginal error** (also called the **maximum error of the estimate**). For a specific value, say, $\alpha = 0.05$, 95% of the sample means will fall within this error value on either side of the population mean. See the following figure



- The **margin of error**, also called the *maximum error of the estimate*, is the maximum likely difference between the point estimate of a parameter and the actual value of the parameter

The algebraic derivation of the formula for a confidence interval

As explained in Chapter 6, the sampling distribution of the mean is approximately normal when large samples ($n \geq 30$) are taken from a population. Also,

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Furthermore, there is a probability of $1 - \alpha$ that a z will have a value between $-z_{\frac{\alpha}{2}}$

and $+z_{\frac{\alpha}{2}}$. Hence,
$$-z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < +z_{\frac{\alpha}{2}}$$

By using algebra, the formula can be rewritten as
$$-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < +z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Subtracting $\bar{X}$ from both sides and from the middle gives

$$-\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Multiplying by -1 gives
$$\bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Reversing the inequality yields the formula for the confidence interval:

$$\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Example 7-1: A researcher wishes to estimate the number of days it takes an automobile dealer to sell a Chevrolet Aveo. A **random sample of 50** cars had a **mean** time on the dealer's lot of **54** days. Assume the **population standard deviation** to be **6.0** days. Find the best point estimate of the population mean and the 95% confidence interval of the population mean.

Solution: The best point estimate of the population mean is 54 days.

$$n = 50, \bar{x} = 54, \sigma = 6, z_{\frac{\alpha}{2}} = 1.96$$

the 95% confidence interval of μ :

$$\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$54 - (1.96) \cdot \left(\frac{6}{\sqrt{50}} \right) < \mu < 54 + (1.96) \cdot \left(\frac{6}{\sqrt{50}} \right)$$

$$54 - 1.7 < \mu < 54 + 1.7$$

$$52.3 < \mu < 55.7$$

Hence, one can say with 95% confidence that the interval between 52.3 and 55.7 days does contain the population mean, based on a sample of 50 automobiles.

Example 7-2: A large department store found that it **averages 362** customers per hour. Assume that the **standard deviation** is **29.6** and a **random sample** of **40** hours was used to determine the average. Find the 99% confidence interval of the population mean.

Solution: $n = 40, \bar{x} = 362, \sigma = 29.6, z_{\frac{\alpha}{2}} = 2.58$

the 99% confidence interval of μ :

$$\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$362 - (2.58) \cdot \left(\frac{29.6}{\sqrt{40}} \right) < \mu < 362 + (2.58) \cdot \left(\frac{29.6}{\sqrt{40}} \right)$$

$$362 - 12.1 < \mu < 362 + 12.1$$

$$349.9 < \mu < 374.1$$

$$350 < \mu < 374$$

Hence, one can be 99% confident (rounded values) that the mean number of customers that the store averages is between 350 and 374 customers per hour.

Example 7-3: The following data represent a **random sample** of the assets (in millions of dollars) of **30** credit unions in southwestern Pennsylvania. Assume the population **standard deviation** is **14.405**. Find the 90% confidence interval of the mean.

12.23	16.56	4.39
2.89	1.24	2.17
13.19	9.16	1.42
73.25	1.91	14.64
11.59	6.69	1.06
8.74	3.17	18.13
7.92	4.78	16.85
40.22	2.42	21.58
5.01	1.47	12.24
2.27	12.77	2.76

Solution:

Step 1 Find the mean for the data. Use the formula shown in Chapter 3 or your calculator. The mean $\bar{X} = 11.091$. Assume the standard deviation of the population is 14.405.

Step 2 Find $\frac{\alpha}{2}$. Since the 90% confidence interval is to be used, $\alpha = 1 - 0.90 = 0.10$, and $\frac{\alpha}{2} = \frac{0.10}{2} = \mathbf{0.05}$

Step 3 Find $z_{\frac{\alpha}{2}}$. Subtract 0.05 from 1.000 to get 0.9500. The corresponding z value obtained from Table E is 1.65. (Note: This value is found by using the z value for an area between 0.9495 and 0.9505. A more precise z value obtained mathematically is 1.645 and is sometimes used; however, 1.65 will be used in this text.)

Step 4 Substitute in the formula
the 90% confidence interval of μ :

$$\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$11.091 - (1.65) \cdot \left(\frac{14.405}{\sqrt{30}} \right) < \mu < 11.091 + (1.65) \cdot \left(\frac{14.405}{\sqrt{30}} \right)$$

$$11.091 - 4.339 < \mu < 11.091 + 4.339$$

$$6.752 < \mu < 15.430$$

Hence, one can be 90% confident that the population mean of the assets of all credit unions is between \$6.752 million and \$15.430 million, based on a sample of 30 credit unions.

Sample Size

The formula for sample size is derived from the margin of error formula

$$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

and this formula is solved for n as follows:

$$E \sqrt{n} = z_{\frac{\alpha}{2}} \cdot \sigma$$

$$\sqrt{n} = \frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E}$$

Hence,

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2$$

Formula for the Minimum Sample Size Needed for an Interval Estimate of the Population Mean

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

where E is the margin of error. If necessary, round the answer up to obtain a whole number. That is, if there is any fraction or decimal portion in the answer, use the next whole number for sample size n .

Example 7-4: A scientist wishes to estimate the average depth of a river. He wants to be **99%** confident that the estimate is accurate within **2** feet. From a previous study, the **standard deviation** of the depths measured was **4.33** feet.

Solution: Since $\alpha = 0.01$ (or $1 - 0.99$), $z_{\frac{\alpha}{2}} = 2.58$ and $E = 2$.
Substituting in the formula,

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2 = \left(\frac{(2.58)(4.33)}{2} \right)^2 = 31.2$$

Round the value 31.2 up to 32. Therefore, to be 99% confident that the estimate is within 2 feet of the true mean depth, the scientist needs a sample of at least 32 measurements.

- In most cases in statistics, we round off. However, when determining sample size, we always round up to the next whole number.

7-2 CONFIDENCE INTERVALS FOR THE MEAN WHEN σ IS UNKNOWN

When σ is known and the sample size is 30 or more, or the population is normally distributed if the sample size is less than 30, the confidence interval for the mean can be found by using the z distribution. However, most of the time, the value of σ is not known, so it must be estimated by using s , namely, the standard deviation of the sample. When s is used, especially when the sample size is small, critical values greater than the values for $z_{\frac{\alpha}{2}}$ are used in confidence intervals in order to keep the interval at a given level, such as the 95%. These values are taken from the *Student t distribution*, most often called the ***t* distribution**.

Characteristics of the t Distribution

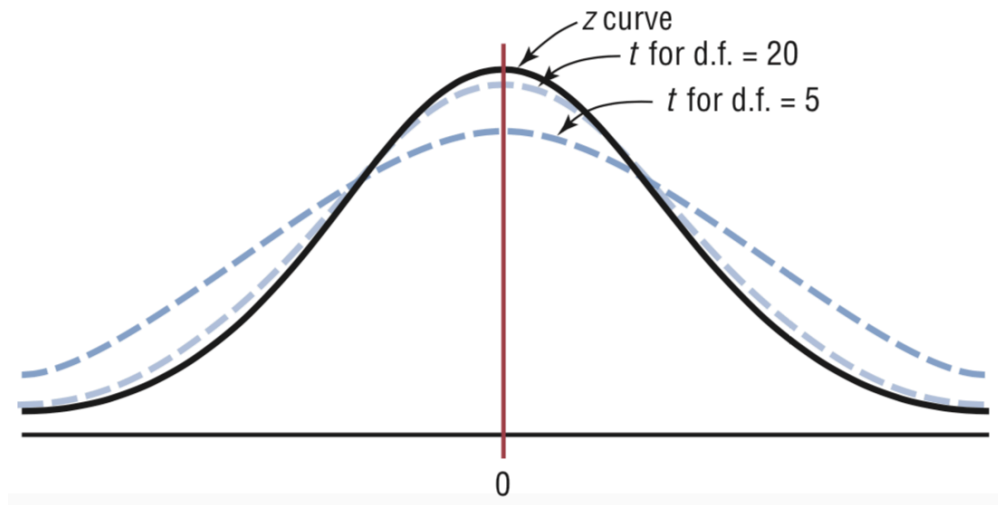
The t distribution is similar to the standard normal distribution in these ways:

1. It is bell-shaped.
2. It is symmetric about the mean.
3. The mean, median, and mode are equal to 0 and are located at the center of the distribution.
4. The curve approaches but never touches the x axis.

The t distribution differs from the standard normal distribution in the following ways:

1. The variance is greater than 1.
2. The t distribution is actually a family of curves based on the concept of *degrees of freedom*, which is related to sample size.
3. As the sample size increases, the t distribution approaches the standard normal distribution.

- The degrees of freedom are the number of values that are free to vary after a sample statistic has been computed, and they tell the researcher which specific curve to use when a distribution consists of a family of curves.
- The symbol d.f. will be used for degrees of freedom.
- The degrees of freedom for a confidence interval for the mean are found by subtracting 1 from the sample size. That is, **d.f. = $n - 1$** .



- The formula for finding the confidence interval using the t distribution has a critical value $t_{\frac{\alpha}{2}}$.

Example 7-5: Find the $t_{\frac{\alpha}{2}}$ value for a 95% confidence interval when the sample size is 22.

Solution:

$$1 - \alpha = 0.95$$

$$\text{The d.f.} = n - 1 = 22 - 1 = 21$$

$$t_{\frac{\alpha}{2}} = 2.080$$

TABLE F The t Distribution

d.f.	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947
16		1.337	1.746	2.120	2.583	2.921
17		1.333	1.740	2.110	2.567	2.898
18		1.330	1.734	2.101	2.552	2.878
19		1.328	1.729	2.093	2.539	2.861
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797
25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779

Formula for a Specific Confidence Interval for the Mean When σ Is Unknown

$$\bar{X} - t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) < \mu < \bar{X} + t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

The degrees of freedom are $n - 1$.

Assumptions for Finding a Confidence Interval for a Mean When σ Is Unknown

1. The sample is a random sample.
2. Either $n \geq 30$ or the population is normally distributed when $n < 30$.

Example 7-6: A random sample of **10** children found that their **average** growth for the first year was **9.8** inches. Assume the variable is normally distributed and the **sample standard deviation** is **0.96** inch. Find the 95% confidence interval of the population mean for growth during the first year.

Solution: $n = 10$, $\bar{x} = 9.8$, $s = 0.96$,

$$1 - \alpha = 0.95$$

$$\text{d.f.} = n - 1 = 10 - 1 = 9$$

$$t_{\frac{\alpha}{2}} = 2.262$$

TABLE F The <i>t</i> Distribution						
	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169

the 95% confidence interval of μ :

$$\bar{X} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

$$9.8 - (2.262) \cdot \left(\frac{0.96}{\sqrt{10}} \right) < \mu < 9.8 + (2.262) \cdot \left(\frac{0.96}{\sqrt{10}} \right)$$

$$9.8 - 0.69 < \mu < 9.8 + 0.69$$

$$9.11 < \mu < 10.49$$

Therefore, one can be 95% confident that the population mean of the first-year growth is between 9.11 and 10.49 inches.

Example 7-7: The data represent a random sample of the number of home fires started by candles for the past several years. (Data are from the National Fire Protection Association.) Find the 99% confidence interval for the mean number of home fires started by candles each year.

5460 5900 6090 6310 7160 8440 9930

Solution: $n = 7$, $\bar{X} = 7041.4$, $s = 1610.3$,

$$1 - \alpha = 0.99$$

$$\text{d.f.} = n - 1 = 7 - 1 = 6$$

$$t_{\frac{\alpha}{2}} = 3.707$$

TABLE F The <i>t</i> Distribution						
	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499

the 99% confidence interval of μ :

$$\bar{X} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

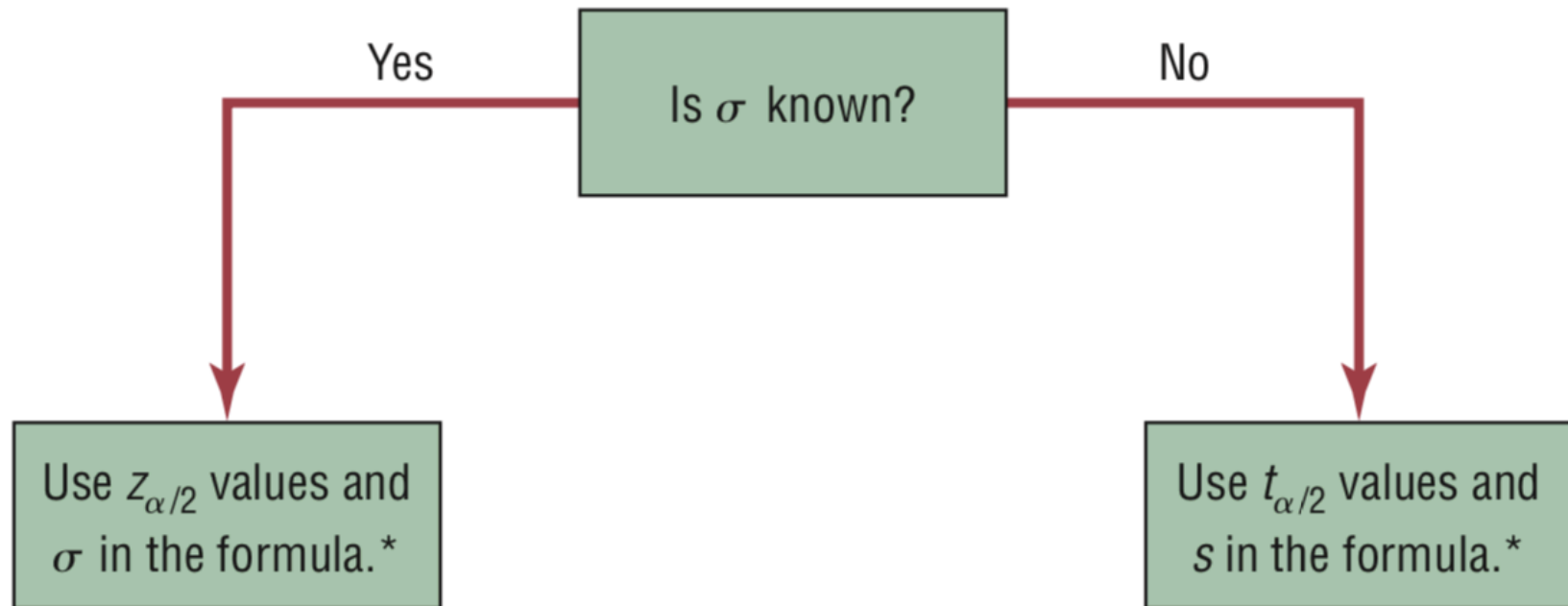
$$7041.4 - (3.707) \cdot \left(\frac{1610.3}{\sqrt{7}} \right) < \mu < 7041.4 + (3.707) \cdot \left(\frac{1610.3}{\sqrt{7}} \right)$$

$$7041.4 - 2256.2 < \mu < 7041.4 + 2256.2$$

$$4785.2 < \mu < 9297.6$$

One can be 99% confident that the population mean number of home fires started by candles each year is between 4785.2 and 9297.6, based on a sample of home fires occurring over a period of 7 years.

When to Use the z or t Distribution



*If $n < 30$, the variable must be normally distributed.

7-3 CONFIDENCE INTERVALS AND SAMPLE SIZE FOR PROPORTIONS

- A *USA TODAY* Snapshots feature stated that 12% of the pleasure boats in the United States were named *Serenity*. The parameter 12% is called a **proportion**. It means that of all the pleasure boats in the United States, 12 out of every 100 are named *Serenity*. A proportion represents a part of a whole. It can be expressed as a fraction, decimal, or percentage. In this case, $12\% = 0.12 = \frac{12}{100}$ or $\frac{3}{25}$. Proportions can also represent probabilities. In this 100 25 case, if a pleasure boat is selected at random, the probability that it is called *Serenity* is 0.12.
- Proportions can be obtained from samples or populations.

Symbols Used in Proportion Notation

p = population proportion

$\hat{p}$ (read “ p hat”) = sample proportion

For a sample proportion,

$$\hat{p} = \frac{X}{n} \quad \text{and} \quad \hat{q} = \frac{n - X}{n} \quad \text{or} \quad \hat{q} = 1 - \hat{p}$$

where X = number of sample units that possess the characteristics of interest and n = sample size.

Example 7-8: A random sample of 200 workers found that 128 drove to work alone. Find $\hat{p}$ and $\hat{q}$, where $\hat{p}$ is the proportion of workers who drove to work alone.

Solution: In this case, $X = 128$ and $n = 200$.

$$\hat{p} = \frac{X}{n} = \frac{128}{200} = 0.64 = 64\%$$

$$\hat{q} = \frac{n - X}{n} = \frac{200 - 128}{200} = \frac{72}{200} = 0.36 = 36\%$$

Also $\hat{q}$ can be found by using the formula $\hat{q} = 1 - \hat{p}$. In this case, $\hat{q} = 1 - 0.64 = 0.36$. So 64% of the people in the survey drive to work alone, and 36% drive with others.

- Point and interval estimates for a population proportion can be made by using the sample proportion. For a point estimate of p (the population proportion), $\hat{p}$ (the sample proportion) is used.
- The confidence interval for a particular p is based on the sampling distribution of $\hat{p}$. When the sample size n is no more than 5% of the population size, the sampling distribution of $\hat{p}$ is approximately normal with a mean of p and a standard deviation of $\sqrt{\frac{pq}{n}}$, where $q = 1 - p$

Confidence Intervals

- To construct a confidence interval about a proportion, you must use the margin of error, which is

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- Confidence intervals about proportions must meet the criteria that $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$.

Formula for a Specific Confidence Interval for a Proportion

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

when $n\hat{p}$ and $n\hat{q}$ are each greater than or equal to 5.

Assumptions for Finding a Confidence Interval for a Population Proportion

1. The sample is a random sample.
2. The conditions for a binomial experiment are satisfied (see Chapter 5).

Example 7-9: A survey conducted by Sallie Mae and Gallup of 1404 respondents found that 323 students paid for their education by student loans. Find the 90% confidence interval of the true proportion of students who paid for their education by student loans.

Solution:

Step 1 Determine $\hat{p}$ and $\hat{q}$.

$$\hat{p} = \frac{X}{n} = \frac{323}{1404} = 0.23$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.23 = 0.77$$

Step 2 Determine the critical value.

$$\alpha = 1 - 0.90 = 0.10$$

$$\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$$

$$z_{\frac{\alpha}{2}} = 1.65$$

Step 3 Substitute in the formula

$$\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.23 - 1.65 \sqrt{\frac{(0.23)(0.77)}{1404}} < p < 0.23 + 1.65 \sqrt{\frac{(0.23)(0.77)}{1404}}$$

$$0.23 - 0.019 < p < 0.23 + 0.019$$

$$0.211 < p < 0.249$$

Or
$$21.1\% < p < 24.9\%$$

Hence, you can be 90% confident that the percentage of students who pay for their college education by student loans is between 21.1 and 24.9%.

Example 7-10: A survey of 1898 adults with lawns conducted by Harris Interactive Poll found that 45% of the adults said that dandelions were the toughest weeds to control in their yards. Find the 95% confidence interval of the true proportion who said that dandelions were the toughest weeds to control in their yards.

Solution:

Step 1 Determine $\hat{p}$ and $\hat{q}$.

In this case, $\hat{p}$ is already given. It is 45%, or 0.45.

$$\hat{q} = 1 - \hat{p} = 1 - 0.45 = 0.55$$

Step 2 Determine the critical value.

$$\alpha = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

$$z_{\frac{\alpha}{2}} = 1.96$$

Step 3 Substitute in the formula

$$\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.45 - 1.96 \sqrt{\frac{(0.45)(0.55)}{1898}} < p < 0.45 + 1.96 \sqrt{\frac{(0.45)(0.55)}{1898}}$$

$$0.45 - 0.022 < p < 0.45 + 0.022$$

$$0.428 < p < 0.472$$

Or

$$42.8\% < p < 47.2\%$$

Hence, you can say with 95% confidence that the true percentage of adults who consider dandelions the toughest weeds to control in their lawns is between 42.8% and 47.2%.

Sample Size for Proportions

Formula for Minimum Sample Size Needed for Interval Estimate of a Population Proportion

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2$$

If necessary, round up to obtain a whole number.

This formula can be found by solving the margin of error value for n in the formula

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Example 7-11: A researcher wishes to estimate, with 95% confidence, the proportion of people who own a home computer. A previous study shows that 40% of those interviewed had a computer at home. The researcher wishes to be accurate within 2% of the true proportion. Find the minimum sample size necessary.

Solution:

Since $z_{\frac{\alpha}{2}} = 1.96$, $E = 0.02$, $\hat{p} = 0.40$, and $\hat{q} = 0.60$, then

$$n = \hat{p}\hat{q} \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 = (0.40)(0.60) \left(\frac{1.96}{0.02} \right)^2 = 2304.96$$

which, when rounded up, is 2305 people to interview. So the researcher must interview 2305 people.

Example 7-12: In Example 7–11 assume that no previous study was done. Find the minimum sample size necessary to be accurate within 2% of the true population.

Solution:

Here we do not know the values of $\hat{p}$ and $\hat{q}$. So we use $\hat{p} = 0.5$, and $\hat{q} = 0.5$.

$$E = 0.02 \quad \text{and} \quad \frac{z_{\alpha}}{2} = 1.96$$

$$n = \hat{p}\hat{q} \left(\frac{\frac{z_{\alpha}}{2} \cdot \sigma}{E} \right)^2 = (0.5)(0.5) \left(\frac{1.96}{0.02} \right)^2 = 2401$$

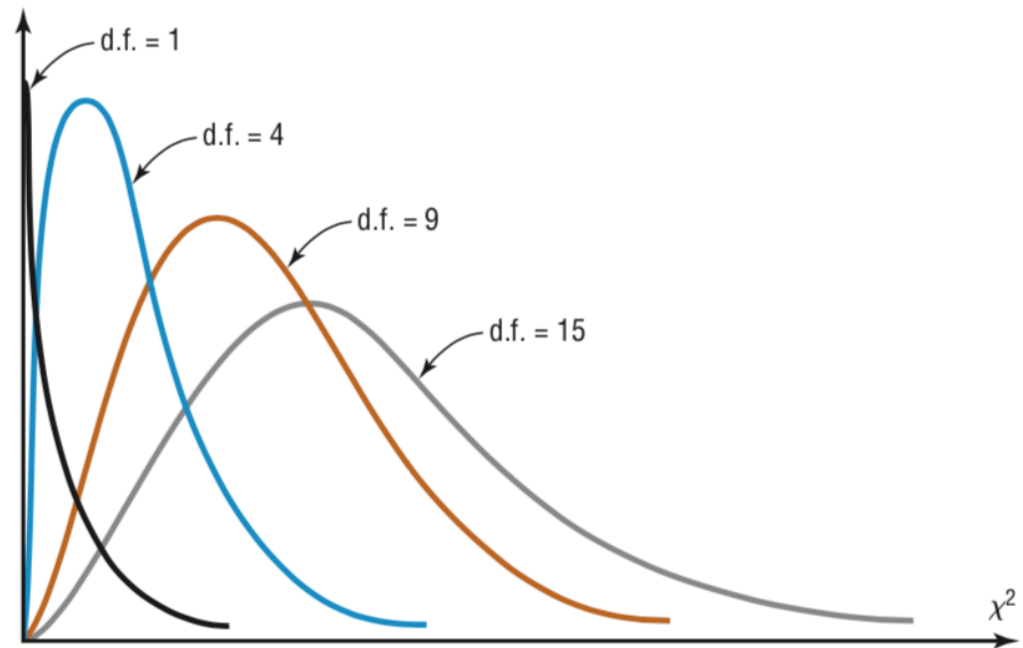
which, 2401 people must be interviewed when $\hat{p}$ is unknown. This is 96 more people than needed if $\hat{p}$ is known.

7-4 CONFIDENCE INTERVALS FOR VARIANCES AND STANDARD DEVIATIONS

- In statistics, the variance and standard deviation of a variable are as important as the mean. To calculate these confidence intervals, a new statistical distribution is needed. It is called the **chi-square distribution**.
- The chi-square variable is similar to the t variable in that its distribution is a family of curves based on the number of degrees of freedom.
- The symbol for chi-square is χ^2 (Greek letter chi, pronounced “ki”).

- The chi-square distribution is obtained from the values of $\frac{(n-1)s^2}{\sigma^2}$ when random samples are selected from a normally distributed population whose variance is σ^2 .
- A chi-square variable cannot be negative, and the distributions are skewed to the right. At about 100 degrees of freedom, the chi-square

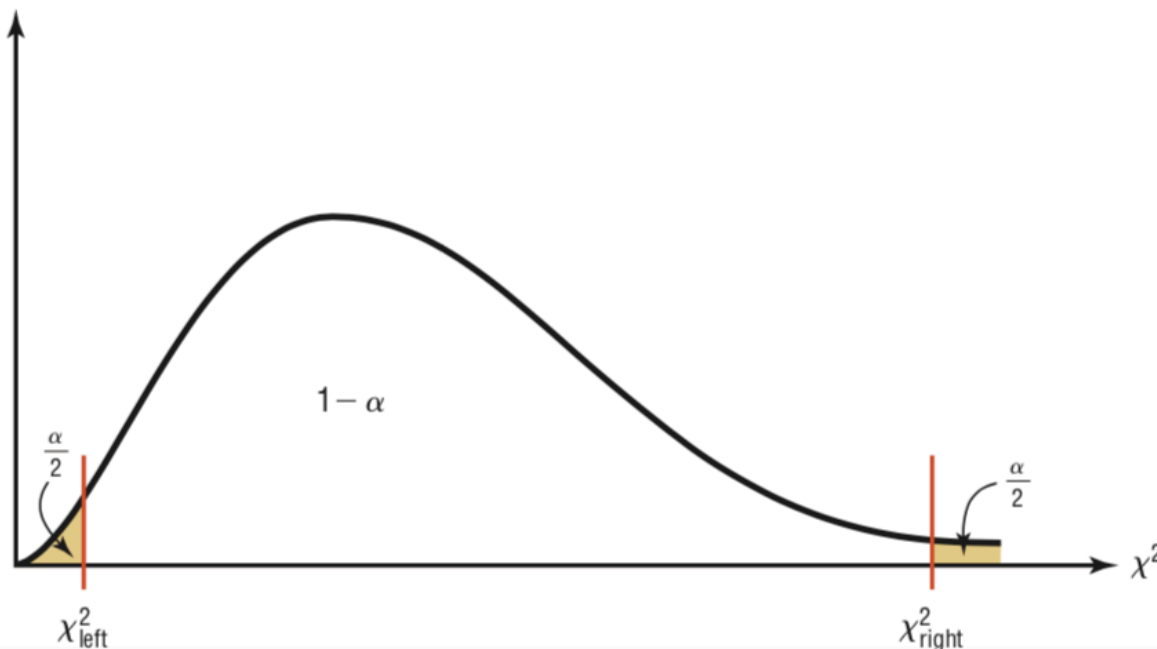
distribution becomes somewhat symmetric. The area under each chi-square distribution is equal to 1.00, or 100%.



Characteristics of the Chi-Square Distribution

1. All chi-square values are greater than or equal to 0.
2. The chi-square distribution is a family of curves based on the degrees of freedom.
3. The area under each chi-square distribution curve is equal to 1.
4. The chi-square distributions are positively skewed.

Chi-Square Distribution for d.f. = $n - 1$



Example 7-13: Find the values for χ^2_{right} and χ^2_{left} for a 90% confidence interval when $n = 25$.

Solution: To find χ^2_{right} , subtract $1 - 0.90 = 0.10$; then divide 0.10 by 2 to get 0.05.

To find χ^2_{left} , subtract $1 - 0.05 = 0.95$.

Then use the 0.95 and 0.05 columns with d.f. = $n - 1$
 $= 1 - 25 = 24$

Table G										
The Chi-square Distribution										
Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1										
2										
⋮										
24				13.848			36.415			
				χ^2_{left}			χ^2_{right}			

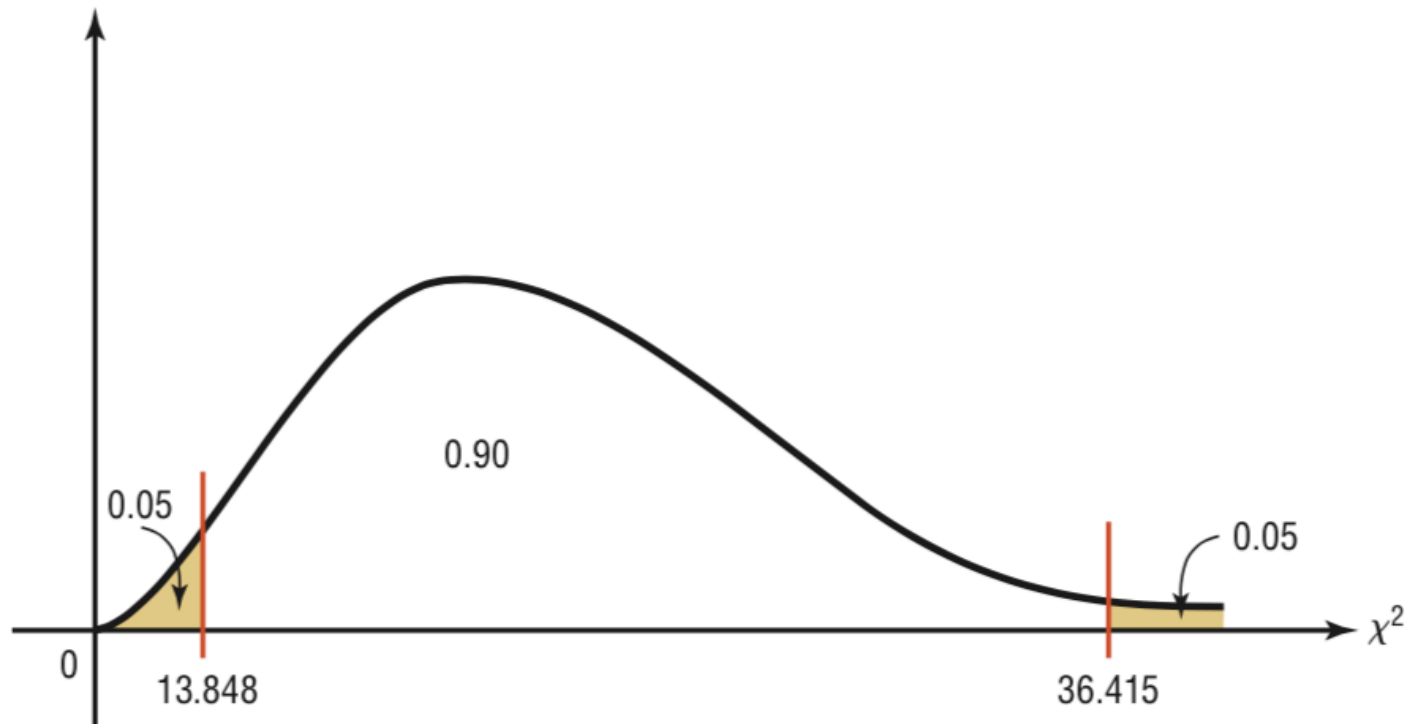
TABLE G The Chi-Square Distribution

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928

The values are

$$\chi^2_{\text{right}} = 36.415$$

$$\chi^2_{\text{left}} = 13.848$$



- If the number for the degrees of freedom is not given in the table, use the closest lower value in the table. For example, for d.f. = 53, use d.f. = 50. This is a conservative approach.
- Useful estimates for σ^2 and σ are s^2 and s , respectively.
- To find confidence intervals for variances and standard deviations, you must assume that the variable is normally distributed.

The formulas for the confidence intervals are shown here.

Formula for the Confidence Interval for a Variance

$$\frac{(n - 1)s^2}{\chi_{\text{right}}^2} < \sigma^2 < \frac{(n - 1)s^2}{\chi_{\text{left}}^2}$$

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

Formula for the Confidence Interval for a Standard Deviation

$$\sqrt{\frac{(n - 1)s^2}{\chi_{\text{right}}^2}} < \sigma < \sqrt{\frac{(n - 1)s^2}{\chi_{\text{left}}^2}}$$

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

Recall that s^2 is the symbol for the sample variance and s is the symbol for the sample standard deviation.

Assumptions for Finding a Confidence Interval for a Variance or Standard Deviation

1. The sample is a random sample.
2. The population must be normally distributed.

Example 7-14: Find the 95% confidence interval for the variance and standard deviation of the nicotine content of cigarettes manufactured if a random sample of 20 cigarettes has a standard deviation of 1.6 milligrams. Assume the variable is normally distributed.

Solution: Since $\alpha = 0.05$, the two critical values, respectively, for the 0.025 and 0.975 levels for 19 degrees of freedom are 32.852 and 8.907. The 95% confidence interval for the variance is found by substituting in the formula.

$$\begin{aligned}\frac{(n-1)s^2}{\chi_{\text{right}}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\text{left}}^2} \\ \frac{(20-1)1.6^2}{32.852} < \sigma^2 < \frac{(20-1)1.6^2}{8.907} \\ 1.5 < \sigma^2 < 5.5\end{aligned}$$

Hence, you can be 95% confident that the true variance for the nicotine content is between 1.5 and 5.5.

For the standard deviation, the confidence interval is

$$\sqrt{1.5} < \sigma < \sqrt{5.5}$$

$$1.2 < \sigma < 2.3$$

Hence, you can be 95% confident that the true standard deviation for the nicotine content of all cigarettes manufactured is between 1.2 and 2.3 milligrams based on a sample of 20 cigarettes.

TABLE G The Chi-Square Distribution

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928

Example 7-15: Find the 90% confidence interval for the variance and standard deviation for the number of named storms per year in the Atlantic basin. A random sample of 10 years has been used. Assume the distribution is approximately normal.

10	5	12	11	13
15	19	18	14	16

Solution:

Step 1 Find the variance for the data. Use the formulas in Chapter 3 or your calculator. The variance $s^2 = 16.9$.

Step 2 Find χ^2_{right} and χ^2_{left} from Table G in Appendix A, using $10 - 1 = 9$ degrees of freedom.

In this case, use $\alpha = 0.05$ and 0.95 ; $\chi^2_{\text{right}} = 3.325$; $\chi^2_{\text{left}} = 16.919$.

Step 3 Substitute in the formula.

$$\frac{(n - 1)s^2}{\chi^2_{\text{right}}} < \sigma^2 < \frac{(n - 1)s^2}{\chi^2_{\text{left}}}$$
$$\frac{(10 - 1)16.9^2}{3.325} < \sigma^2 < \frac{(10 - 1)16.9^2}{16.919}$$

$$8.99 < \sigma^2 < 45.74$$

$$\sqrt{8.99} < \sigma < \sqrt{45.74}$$

$$3.0 < \sigma < 6.9$$

Hence, you can be 90% confident that the standard deviation for the number of named storms is between 3.0 and 6.8 based on a random sample of 10 years.

IMPORTANT FORMULAS

Formula for the confidence interval of the mean when σ is known (when $n \geq 30$, s can be used if σ is unknown):

$$\bar{X} - z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{X} + z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

Formula for the sample size for means:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

where E is the margin of error.

Formula for the confidence interval of the mean when σ is unknown:

$$\bar{X} - t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) < \mu < \bar{X} + t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

Formula for the confidence interval for a proportion:

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $\hat{p} = X/n$ and $\hat{q} = 1 - \hat{p}$.

Formula for the sample size for proportions:

$$n = \hat{p}\hat{q} \left(\frac{z_{\alpha/2}}{E} \right)^2$$

Formula for the confidence interval for a variance:

$$\frac{(n-1)s^2}{\chi^2_{\text{right}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\text{left}}}$$

Formula for the confidence interval for a standard deviation:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\text{right}}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\text{left}}}}$$