

AMA 2634

Introduction to Statistics

2024 Fall

Lecture 7

Point and Interval Estimation



A point estimate is a specific numerical value estimate of a parameter.

A good point estimate of the population mean μ is the sample mean $\bar{X}$.



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.

- $\mathbb{E}(\bar{X}) = \mu$, the sample mean is an unbiased estimator of the population mean μ
- $\mathbb{E}(s^2) = \sigma^2$, the sample variance is an unbiased estimator of the population variance σ^2
- (try to prove above two statements)



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.

- $\bar{X} \rightarrow \mu$, the sample mean is a consistent estimator of the population mean μ
- $s^2 \rightarrow \sigma^2$, the sample mean is a consistent estimator of the population mean μ
- (try to prove above two statements by Chebyshev's inequality)



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.

- Let $X_1, \dots, X_n$ be an i.i.d. sample from a **normal** distribution with mean μ and variance σ^2 .
- Let $T_n = T_n(X_1, \dots, X_n)$ be an **unbiased** estimator of the sample mean μ .
- It is known $\text{Var}(\bar{X}) \leq \text{Var}(T_n)$ for any unbiased estimator T_n .

Let $X_1, \dots, X_n$ be an i.i.d. sample from a population with mean μ and variance σ^2 .

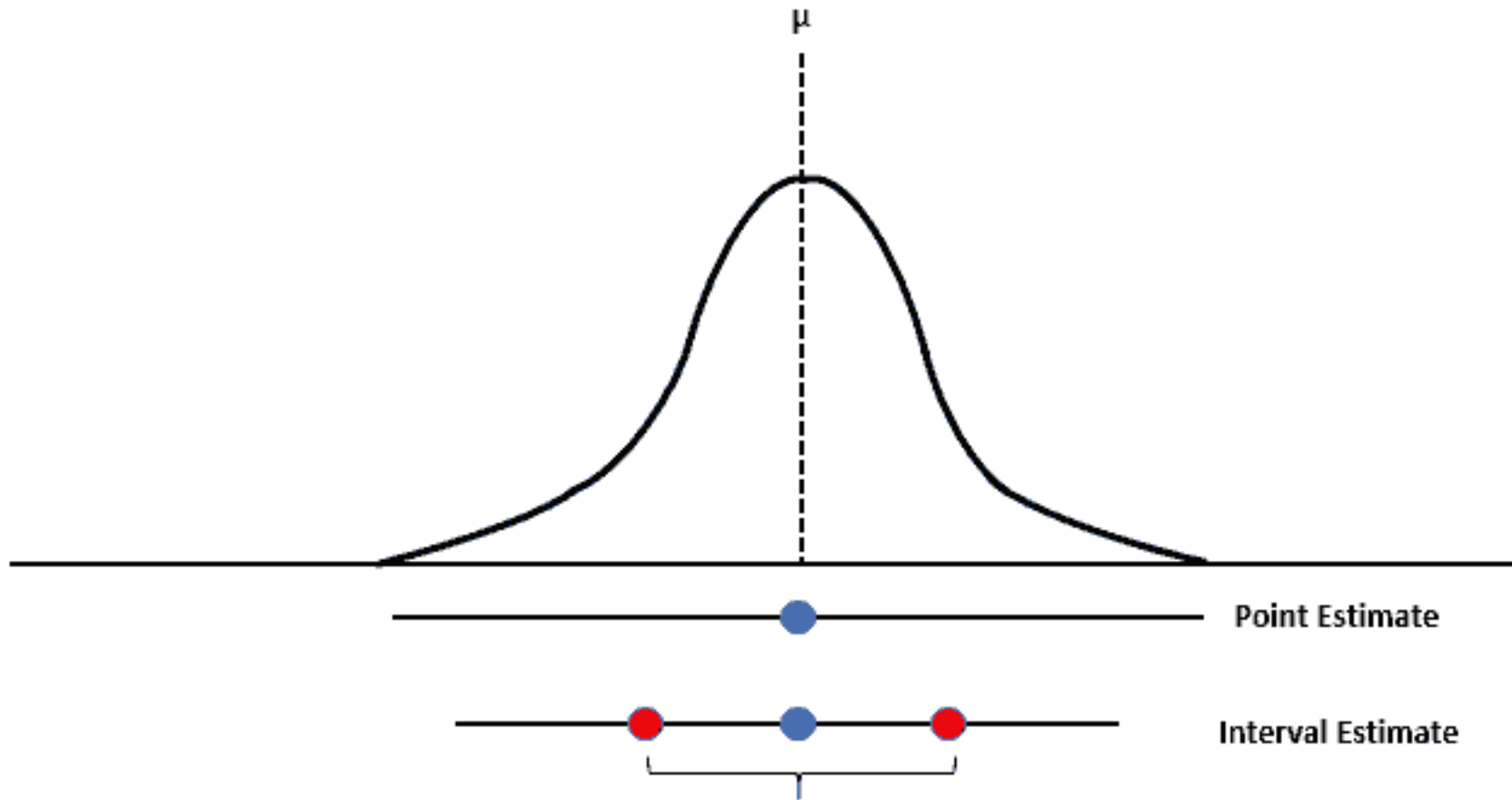
- X_1 is an **unbiased** estimator of the population mean μ , but it is **not consistent**.
- $\bar{X} + \frac{1}{n}$ is a **consistent** estimator of the population mean μ , but it is **not unbiased**.
- $\frac{X_1 + X_2 + \dots + X_{n/2}}{n/2}$ is an **unbiased and consistent** estimator of the population mean μ



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.

Point Estimate and Interval Estimate



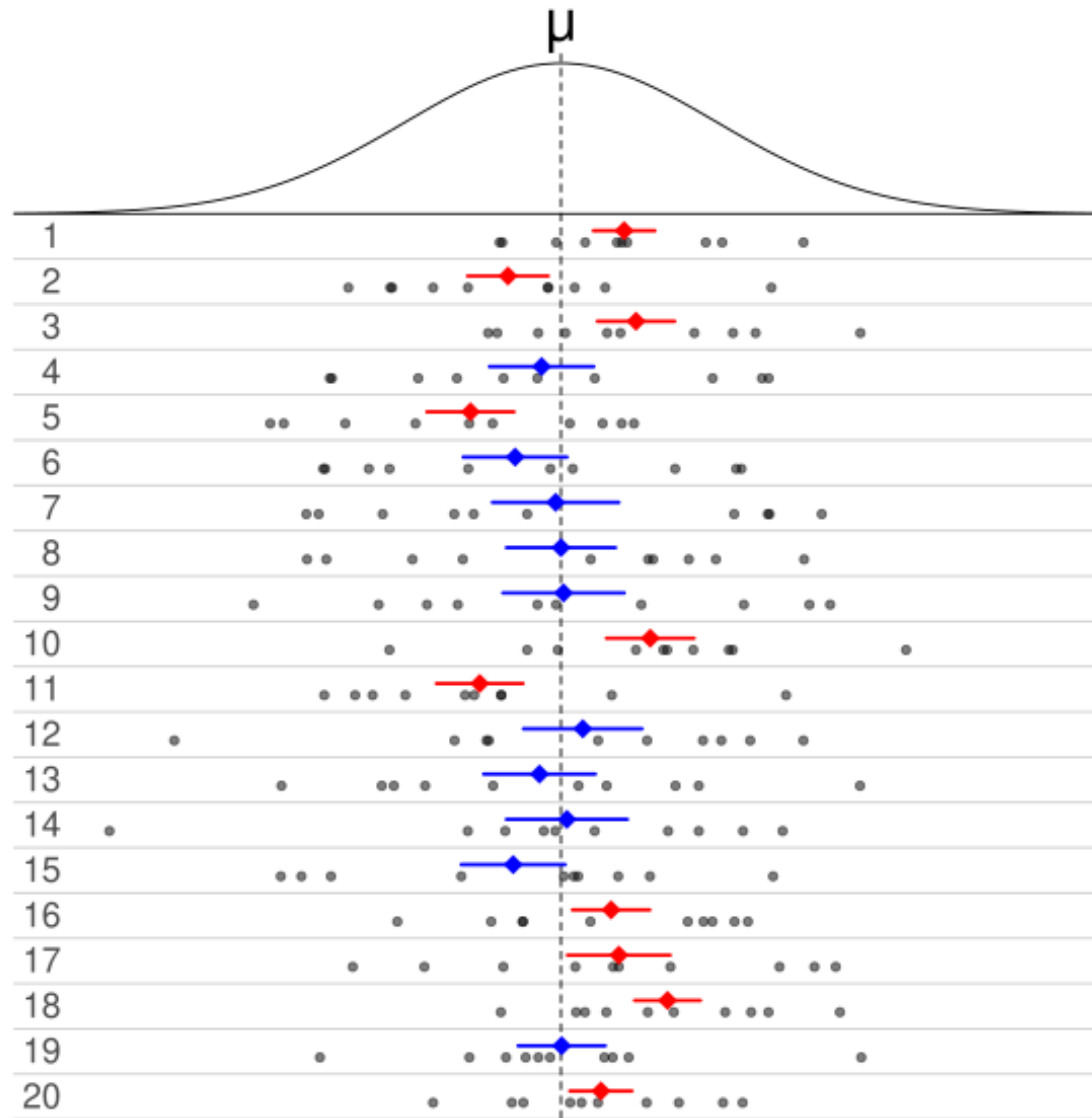
Source: <https://sixsigmastudyguide.com/point-and-interval-estimation/>



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.

Confidence Interval



Source: https://en.wikipedia.org/wiki/File:Normal_distribution_50%25_CI_illustration.svg

Confidence Intervals



Given $\alpha \in (0,1)$, the confidence interval of sample mean with confidence level $1 - \alpha$ is

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

For a 90% confidence interval: $\mathbf{z}_{\alpha/2} = 1.65$

For a 95% confidence interval: $\mathbf{z}_{\alpha/2} = 1.96$

For a 99% confidence interval: $\mathbf{z}_{\alpha/2} = 2.58$



By central limit theorem

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Then $\mathbb{P}(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1 - \alpha$

Plug in $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, we have

$$\mathbb{P}\left(-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}\right) = 1 - \alpha$$

Solve for μ



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.

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

Example: Days to Sell an Aveo



A researcher wishes to estimate the number of days it takes an automobile dealer to sell a Chevrolet Aveo. A 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.

The best point estimate of the mean is 54 days.

$$\bar{X} = 54, \sigma = 6.0, n = 50, 95\% \rightarrow z = 1.96$$

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

$$\bar{X} = 54, \sigma = 6.0, n = 50, 95\% \rightarrow z = 1.96$$

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

$$54 - 1.96 \left(\frac{6.0}{\sqrt{50}} \right) < \mu < 54 + 1.96 \left(\frac{6.0}{\sqrt{50}} \right)$$

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

$$52.3 < \mu < 55.7$$

$$52 < \mu < 56$$

One can say with 95% confidence that the interval between 52 and 56 days contains the population mean, based on a sample of 50 automobiles.

Example: Waiting Times



A survey of 30 emergency room patients found that the average waiting time for treatment was 174.3 minutes.

Assuming that the population standard deviation is 46.5 minutes, find the best point estimate of the population mean and the 99% confidence of the population mean.

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

$$174.3 - 2.58 \left(\frac{46.5}{\sqrt{30}} \right) < \mu < \bar{X} + 2.58 \left(\frac{46.5}{\sqrt{30}} \right)$$

$$174.3 - 21.9 < \mu < 174.3 + 21.9$$

$$152.4 < \mu < 196.2$$

Hence, one can be 99% confident that the mean waiting time for emergency room treatment is between 152.4 and 196.2 minutes.

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: Depth of a River



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. How large a sample is required?

$$99\% \rightarrow z = 2.58, E = 2, \sigma = 4.33$$

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2 = \left(\frac{2.58 \cdot 4.33}{2} \right)^2 = 31.2 = 32$$

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



The value of σ , when it is not known, should be estimated by using s , the standard deviation of the sample.

When s is used, especially when the sample size is small (less than 30), critical values greater than the values for $Z_{\alpha/2}$ 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**.

Student's t -distribution with degrees of freedom ν can be defined as the distribution of the random variable T with

$$T = \frac{Z}{\sqrt{V/\nu}}$$

where

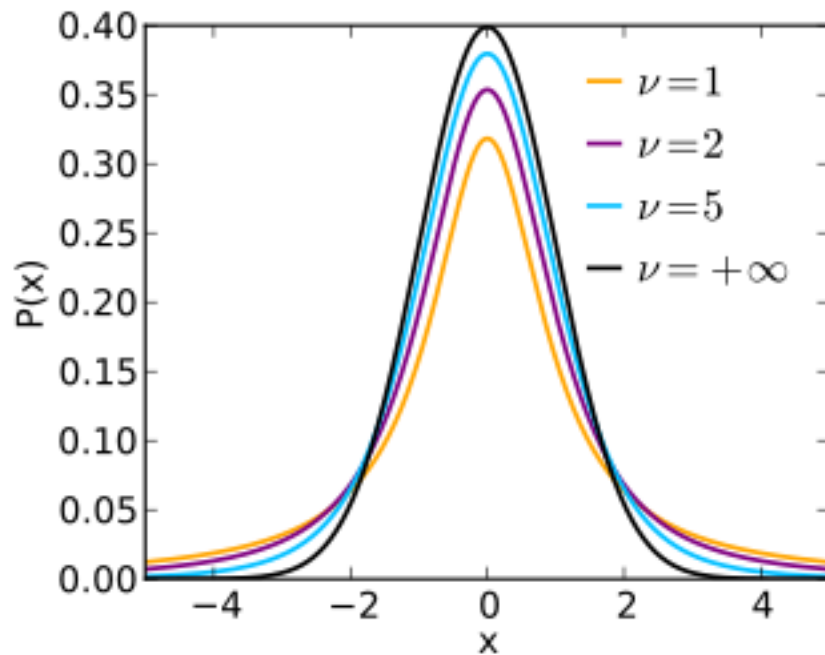
- Z is a standard normal random variable with expected value 0 and variance 1;
- V has a chi-squared distribution with degree of freedom ν
- Z and V are independent



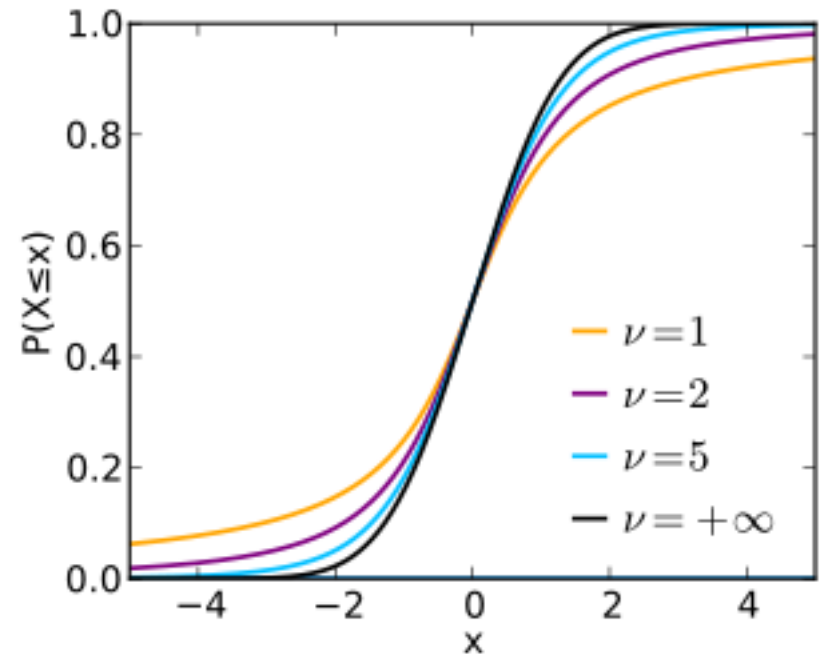
The t distribution is similar to the standard normal distribution in these ways:

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 never touches the x axis.

Characteristics of the t Distribution



Probability density function



Cumulative distribution function

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 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, $d.f. = n - 1$.

Formula for a Specific Confidence Interval for the Mean When σ is unknown and $n < 30$

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

The degrees of freedom are $n - 1$.

Example: Sleeping Time



Ten randomly selected people were asked how long they slept at night. The mean time was 7.1 hours, and the standard deviation was 0.78 hour. Find the 95% confidence interval of the mean time. Assume the variable is normally distributed.

Since σ is unknown and s must replace it, the t distribution must be used for the confidence interval. Hence, with 9 degrees of freedom, $t_{\alpha/2} = 2.262$.

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

$$7.1 - 2.262 \left(\frac{0.78}{\sqrt{10}} \right) < \mu < 7.1 + 2.262 \left(\frac{0.78}{\sqrt{10}} \right)$$

$$7.1 - 2.262 \left(\frac{0.78}{\sqrt{10}} \right) < \mu < 7.1 + 2.262 \left(\frac{0.78}{\sqrt{10}} \right)$$

$$7.1 - 0.56 < \mu < 7.1 + 0.56$$

$$6.5 < \mu < 7.7$$

One can be 95% confident that the population mean is between 6.5 and 7.7 hours.

Example: Home Fires by Candles



The data represent a sample of the number of home fires started by candles for the past several years. Find the 99% confidence interval for the mean number of home fires started by candles each year.

5460 5900 6090 6310 7160 8440 9930

Step 1: Find the mean and standard deviation. The mean is $\bar{X} = 7041.4$ and standard deviation $s = 1610.3$.

Step 2: Find $t_{\alpha/2}$ in the t -Table. The confidence level is 99%, and the degrees of freedom d.f. = 6

$$t_{.005} = 3.707.$$

Step 3: Substitute in the formula.

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

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

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

$$4785.2 < \mu < 9297.6$$

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.



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.

In a recent survey of 150 households, 54 had central air conditioning. Find $\hat{p}$ and $\hat{q}$, where $\hat{p}$ is the proportion of households that have central air conditioning.

Since $X = 54$ and $n = 150$,

$$\hat{p} = \frac{X}{n} = \frac{54}{150} = 0.36 = 36\%$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.36 = 0.64 = 64\%$$

Formula for a Specific Confidence Interval for a Proportion

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

when $np \geq 5$ and $nq \geq 5$.



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 of the true proportion of students who paid for their education by student loans.

Example: Covering College Costs



Since $\alpha = 1 - 0.90 = 0.10$, $z_{\alpha/2} = 1.65$.

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

$$\hat{p} = \frac{323}{1404} = 0.23 \qquad \hat{q} = 1 - \hat{p} = 1 - 0.23 = 0.77$$

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

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

$$0.211 < p < 0.249$$

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

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: Religious Books



A survey of 1721 people found that 15.9% of individuals purchase religious books at a Christian bookstore. Find the 95% confidence interval of the true proportion of people who purchase their religious books at a Christian bookstore.

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

$$0.159 - 1.96 \sqrt{\frac{(0.159)(0.841)}{1721}} < p < 0.159 + 1.96 \sqrt{\frac{(0.159)(0.841)}{1721}}$$

$$0.142 < p < 0.176$$

You can say with 95% confidence that the true percentage is between 14.2% and 17.6%.

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 the next whole number.

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.

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

The researcher should interview a sample of at least 2305 people.

A researcher wishes to estimate the percentage of M&M's that are brown. He wants to be 95% confident and be accurate within 3% of the true proportion.

How large a sample size would be necessary?

Since no prior knowledge of $\hat{p}$ is known, assign a value of 0.5 and then $\hat{q} = 1 - 0.5 = 0.5$. Substitute in the formula, using $E = 0.03$.

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2 = (0.5)(0.5)\left(\frac{1.96}{0.03}\right)^2 = 1067.1$$

Hence, a sample size of 1068 would be needed.



When products that fit together (such as pipes) are manufactured, it is important to keep the variations of the diameters of the products as small as possible; otherwise, they will not fit together properly and will have to be scrapped.

In the manufacture of medicines, the variance and standard deviation of the medication in the pills play an important role in making sure patients receive the proper dosage.

For these reasons, confidence intervals for variances and standard deviations are necessary.

If $Z_1, \dots, Z_K$, are independent, standard normal random variables, then the sum of their squares

$$Q = \sum_{i=1}^K Z_i^2$$

is distributed according to the chi-squared distribution with K degrees of freedom. This is usually denoted as

$$Q \sim \chi^2(K) \text{ or } Q \sim \chi_K^2$$

The chi-squared distribution has one parameter: a positive integer K that specifies the number of **degrees of freedom (d.f.)**.

The chi-square distribution is used to calculate confidence intervals for variances and standard deviations.

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.

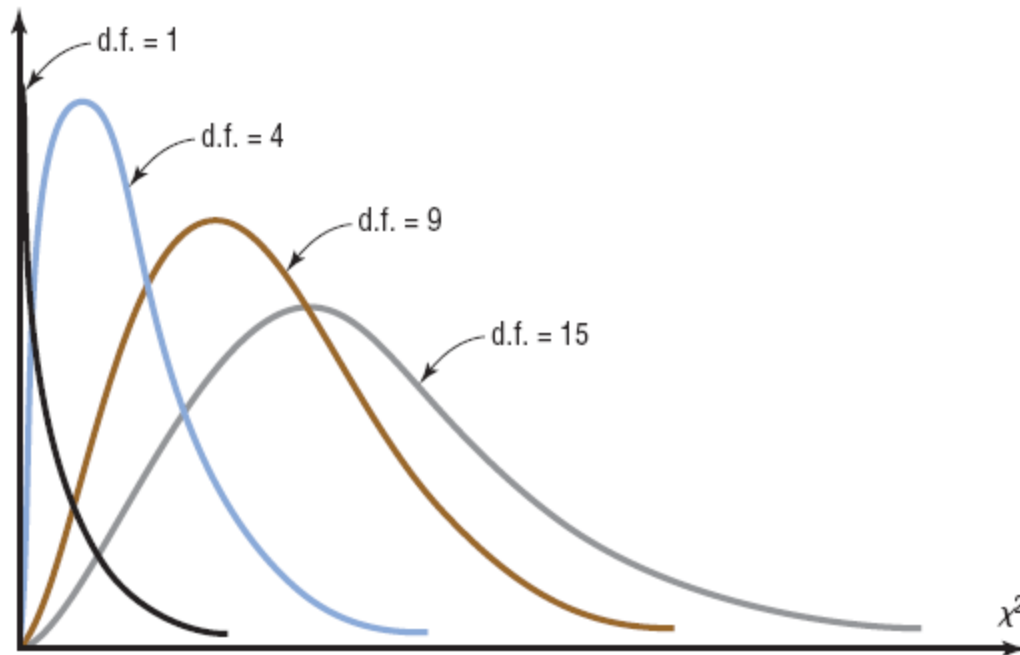
The symbol for chi-square is χ^2 (Greek letter chi).

A chi-square variable cannot be negative, and the distributions are skewed to the right.

Chi-Square Distributions



- 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%.



$$\frac{(n-1)s^2}{\chi_{\text{right}}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\text{left}}^2}, \quad \text{d.f.} = n-1$$

Formula for the Confidence Interval for a Standard Deviation

$$\sqrt{\frac{(n-1)s^2}{\chi_{\text{right}}^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_{\text{left}}^2}}, \quad \text{d.f.} = n-1$$

Example: Nicotine Content



Find the 95% confidence interval for the variance and standard deviation of the nicotine content of cigarettes manufactured if a sample of 20 cigarettes has a standard deviation of 1.6 milligrams.

To find χ^2_{right} , subtract $1 - 0.95 = 0.05$. Divide by 2 to get 0.025.

To find χ^2_{left} , subtract $1 - 0.025$ to get 0.975.

In Table G, the 0.025 and 0.975 columns with the d.f. 19 row yield values of 32.852 and 8.907, respectively.

$$\frac{(n-1)s^2}{\chi_{\text{right}}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\text{left}}^2}$$
$$\frac{(19)(1.6)^2}{32.852} < \sigma^2 < \frac{(19)(1.6)^2}{8.907}$$
$$1.5 < \sigma^2 < 5.5$$

You can be 95% confident that the true variance for the nicotine content is between 1.5 and 5.5 milligrams.

$$\sqrt{1.5} < \sigma < \sqrt{5.5}$$
$$1.2 < \sigma < 2.3$$

You can be 95% confident that the true standard deviation is between 1.2 and 2.3 milligrams.

Example: Cost of Ski Lift Tickets



Find the 90% confidence interval for the variance and standard deviation for the price in dollars of an adult single-day ski lift ticket. The data represent a selected sample of nationwide ski resorts. Assume the variable is normally distributed.

59 54 53 52 51 39 49 46 49 48

Using technology, we find the variance of the data is $s^2 = 28.2$.

From table, the 0.05 and 0.95 columns with the d.f. 9 row yield values of 16.919 and 3.325, respectively.

$$\frac{(n-1)s^2}{\chi_{\text{right}}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\text{left}}^2}$$
$$\frac{(9)(28.2)}{16.919} < \sigma^2 < \frac{(9)(28.2)}{3.325}$$
$$15.0 < \sigma^2 < 76.3$$

You can be 90% confident that the true variance for the cost of ski lift tickets is between 15.0 and 76.3.

$$\sqrt{15.0} < \sigma < \sqrt{76.3}$$
$$3.87 < \sigma < 8.73$$

You can be 95% confident that the true standard deviation is between \$3.87 and \$8.73.