Statistics &Data Analysis Concepts for Data Science and ML 8

8

Estimating Population Values: Confidence Intervals

Learning Objectives

- The main objective of this chapter is to help you understand why and how the unknown population parameters are estimated using sample statistics. After completing this chapter, you should be able to:
- Understand the ideas behind the point and interval estimates or confidence intervals
- Estimate the population mean with a known population standard deviation using the normal distribution
- Estimate the population mean when the population standard deviation is not known
- Learn when it is appropriate to use the normal distribution and when to use the t-distribution to develop the confidence intervals for the mean

Learning Objectives…cont.

- Develop the confidence interval for the population proportion
- Estimate the population variance using the chi-square distribution and known sample variance
- Determine the appropriate sample size required to estimate the population mean and population proportion

Statistical Inference

Statistical inference is an extremely important area of statistics and is used to estimate the unknown ***population parameters*** such as,

- μ = the population mean
- σ^2 = the population variance
- σ = the population standard deviation, etc.
- p = population proportion
- N= the population size

The above population parameters are estimated using the corresponding ***sample statistic***

- $\overline{x}$ = the sample mean
- s^2 = the sample variance
- s = the sample standard deviation, etc.
- $\overline{p}$ = sample proportion
- n= the sample size

Basic Concepts

Estimation is the simplest form of inferential statistics in which a sample statistic is used to draw conclusion regarding the unknown ***population parameter.***

An ***estimate*** is a numerical value assigned to the unknown population parameter. In statistical analysis, the calculated value of a ***sample statistic*** serves as the estimate. This statistic is known as the ***estimator*** of the unknown parameter.

Estimation or parameter estimation comes under the broad topic of ***statistical inference.***

Statistical inference is the process of using information from sample data to draw a conclusion about the population from which the sample was drawn.

Basic Concepts…cont.

There are two major areas of statistical inference :

(1) Parameter Estimation, and

(2) Hypothesis Testing.

The objective of parameter estimation is to estimate the unknown population parameter using the sample statistic. Two types of estimates are used in parameter estimation: ***point estimate*** *and* ***interval estimate****.*

Point Estimates

A point estimate is a single sample statistic that is used to estimate a population parameter. It is a single numerical quantity.

Example:

1. ***The sample mean*** $\bar{x}$ ***is the point estimate of the population mean*** μ, ***calculated using***

$$\bar{x} = \frac{\sum x}{n}$$

2. ***The sample variance*** s^2 ***is the point estimate of*** *population variance* σ^2 ***calculated using***

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{or,} \quad s^2 = \frac{\sum x^2 - \left[\frac{\sum x^2}{n}\right]}{n-1}$$

3. ***The sample standard deviation, s is the point estimate of the population standard deviation,*** σ ***calculated using,*** $s = \sqrt{s^2}$

3. ***The sample proportion,*** $\bar{p}$ ***is the point estimate of the population proportion, p calculated as*** $\bar{p} = \frac{x}{n}$

Interval Estimates

- An ***interval estimate*** provides an interval or range of values that is used to estimate a population parameter.
- To construct an interval estimate, we find an ***interval*** about the point estimate so that we can be highly confident that it contains the parameter to be estimated.
- An interval with high confidence means that it has a high probability of containing the unknown population parameter to be estimated.

Interval Estimates…cont.

An interval estimate acknowledges that the sampling procedure is subject to error, and therefore, any computed statistic may fall above or below its population parameter target.

The interval estimate is represented by an interval or range of possible values so it implies the presence of uncertainty. An interval estimate is represented in one of the following ways:

$16.8 \leq \mu \leq 18.6$ or, (16.8 to 18.6) or, (16.8 - 18.6)

A formal way of writing an interval estimate is "$L \leq \mu \leq U$" where, L is the lower limit and U is the upper limit of the interval. The symbol μ indicates the population mean is estimated. The interval estimate involves certain probability known as the confidence level.

Estimator

The sample statistic that is used to estimate a population parameter is known as an estimator.

Example: ***the sample variance*** s^2 ***is an estimator of the population variance*** σ^2 ***and the sample mean*** $\bar{x}$ ***may be an estimator of the population mean,*** μ.

The value of the estimator is called an estimate which is used to estimate or predict the value of a population parameter.

Several statistics can be used as estimators such as the mean, median, mode, variance, etc. The most desirable feature of an estimator is that it has a value close to the unknown value of the population parameter.

Choosing an Estimator

The following are the basic questions we should ask while choosing an estimator:

(i) Which statistic will be the most reliable estimator?

(ii) Which will require the least expenditure of resources in terms of sample size?

Properties of Estimator

(1)Unbiasedness:

An estimator is unbiased if the expected value of the estimator is equal to the actual value of the corresponding population parameter that is,

$$E(\bar{x}) = \mu$$

The above expression means that the sample mean $\bar{x}$ is an unbiased estimator of the population mean (μ). Similarly, the sample proportion ($\bar{p}$) is an unbiased estimator of the population proportion (p).

Unbiasedness explains why we define the sample variance, s^2 as :

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

In the above equation, the divisor is (n-1) instead of n. It can be shown that for s^2 to be an unbiased estimator of σ, we must have (n-1) in the denominator.

Properties of Estimator…cont.

(2) Consistency:

A statistic is said to be a consistent estimator of a population parameter if the value of the statistic comes very close to the value of the population parameter as the sample size increases.

(3) Efficiency:

Efficiency refers to the standard error of the statistic. An efficient estimator is one that has a smaller standard error when compared to the standard error of the other statistic.

Commonly used Estimator

The commonly used estimators are the sample mean, $\bar{x}$ and the sample variance, s^2 .

- *$\bar{x}$ is the most desirable estimator of μ because it is unbiased, consistent, and more efficient than other estimators, including the sample median.*

- *$\bar{x}$ has a readily obtainable normal sampling distribution when the sample size (n) is large.*

The sample variance s^2 is also an unbiased and consistent estimator of the population variance σ^2 .

Confidence Interval Estimate

In many situations, a point estimate does not provide enough information about the parameter of interest. In such cases, an interval estimate is desirable. The general form of an interval estimate is:

$$L \leq \mu \leq U$$

To construct an interval estimate of unknown parameter β, we must find two statistics L and U such that

$$P\{L \leq \beta \leq U\} = 1-\alpha$$

The resulting interval $L \leq \beta \leq U$ is called a 100 $(1-\alpha)$ percent confidence interval for the unknown parameter β. L and U are known as the lower and upper confidence limits respectively, and $(1-\alpha)$ is known as the confidence level.

Confidence Level and Confidence Interval

A confidence level is the probability attached to a confidence interval. A 95% confidence interval means that the interval is estimated with a 95% probability or confidence level. This means that there is a 95% chance that the estimated interval would include the unknown population parameter being estimated.

Interpretation of Confidence Interval

The confidence interval $\mathbf{L \leq \beta \leq U}$ *means that if many random samples are collected and a 100 (1-*α*) percent confidence interval computed from each sample for* β*, then 100 (1-*α*) % of these intervals will contain the true value* β*.*

The interval $\mathbf{L \leq \beta \leq U}$ is known as a two-sided or two-tailed interval.

Wider the confidence interval, the more confident we are that the interval actually contains the unknown population parameter being estimated. On the other hand, the wider the interval, the less information we have about the true value of β*. In an ideal situation, we would like to obtain a relatively short interval with high confidence.*

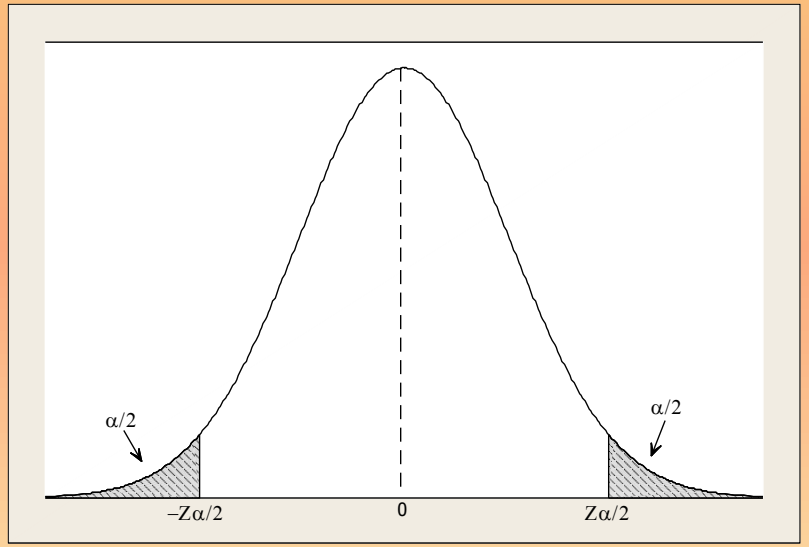
Interpretation of Confidence Interval…cont.

In general:

- ***Higher the confidence level, wider is the interval (We have high confidence that the interval contains the unknown population parameter but a wide interval may be less accurate). In other words, we gain confidence but loose accuracy.***
- ***In a similar way, smaller the confidence level, the shorter is the interval (We lose confidence but gain accuracy).***

CONFIDENCE INTERVAL FOR THE MEAN, KNOWN VARIANCE σ^2 (OR σ KNOWN)

Let X be a random variable with an unknown mean μ and known variance σ^2. A random sample of size *n* ($x_1, x_2 \ldots, x_n$) is taken. A 100 (1-α) % confidence interval on μ can be obtained by considering the sampling distribution of the sampling mean $\bar{x}$. We know that the sample mean $\bar{x}$ follows a normal distribution as the sample size n increases. For a large sample *n* the sampling distribution of the sample mean is almost always normal. The sampling distribution is given by:



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

The distribution of the sample mean is normal and is shown in the figure.

From the figure we see that:

$$P\left\{-z_{\alpha/2} \le z \le z_{\alpha/2}\right\} = 1 - \alpha$$

or,

$$P\left\{-z_{\alpha/2} \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right\} = 1 - \alpha$$

The previous equation can be rearranged to give:

$$P\left\{\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\right\} = 1-\alpha$$

This leads to:

$$\left\{\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\right\}$$

This is a 100(1- a) percent confidence interval for the population mean μ .

CONFIDENCE INTERVAL FOR THE MEAN, UNKNOWN VARIANCE σ^2 (OR, σ UNKNOWN)

When the population standard deviation σ is unknown, the appropriate distribution to use is the t-distribution with the assumption that the population from which the sample is drawn is normal or approximately normal. ***The confidence interval for this case is as follows:***

$$\bar{x} \pm t_{n-1,\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

or,

$$\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

$t_{n-1,\alpha/2}$ = the t-value for (n-1) degrees of freedom and appropriate α that depends on the confidence level of the problem

s is the sample standard deviation, and n is the sample size.

SUMMARY OF THE FORMULAS FOR ESTIMATING THE POPULATION MEAN (μ)

Case (1): The population standard deviation σ is known (or known variance)

In this case the appropriate distribution to use is the ***Normal distribution*** and the confidence interval is given by:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \text{(A)}$$

or,

$$\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \text{(B)}$$

Note that in this equation, the true mean or the population mean μ is estimated using the sample mean. The Z value is based on the confidence level.

SUMMARY OF THE FORMULAS FOR ESTIMATING THE POPULATION MEAN (μ)

Case (2): Population standard deviation σ is unknown (or unknown variance)

In this case, the appropriate distribution to use is the ***t-distribution*** with the assumption that the population from which the sample is drawn is normally distributed. The confidence interval for this case is given by the following formula:

$$\bar{x} \pm t_{n-1,\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) \qquad \text{(C)}$$

$$\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \qquad \text{(D)}$$

or,

where, $t_{n-1,a/2}$ = t-value from the t-table for (n-1) degrees of freedom and $\alpha/2$.

Important Note:

- ***The confidence interval estimate in equations (A or B in the previous slide) are used when the population from which the sample is drawn has a normal distribution.***
- ***If the population is normally distributed, equation (C or D) is exact and can be used for any sample size (small or large). If the population is not normal, equation (C) provides an approximate confidence interval.***
- ***In case the distribution of the population is highly skewed, a larger sample (n= 50 or more) will provide a better estimate.***
- ***For small sample sizes (n <30) the confidence interval using t-distribution in equation (C or D) should only be used if the population approximately follows a normal distribution. In most cases, if the population standard deviation σ is not known equation (C) is used.***

SUMMARY OF THE FORMULAS FOR ESTIMATING THE POPULATION MEAN (μ)

Case (3) : Population standard deviation σ is unknown and the sample size is very large.

If the population standard deviation σ is unknown but the sample size is very large then the population standard deviation can be approximated by the sample standard deviation and the confidence interval can be calculated using the following formula using a normal distribution:

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \qquad \text{or,} \qquad \bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

The general practice is to use a t-distribution for the confidence interval whenever the population standard deviation is unknown and the sample size is less than 30 (or, n < 30).

In cases where the sample size is very large (n ≥ 30) and the population standard deviation in not known both the normal and t-distributions will provide similar results because the t-value gets close to the Z-value as the sample size increases. For very large samples the t-value and the z-value are close.

ESTIMATING THE POPULATION PROPORTION (P) USING THE SAMPLE PROPORTION ($\bar{p}$)

Confidence Interval for proportion (p) based on the Normal Approximation (when the sample size n ≥ 30) is given by

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

or,

$$\bar{p} - z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \le p \le \bar{p} + z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where, p is the population proportion and $\bar{p}$ is the sample proportion.

Margin of Error

The general form of a confidence interval is of the form:

Point estimate ± Margin of error

Thus, the general form of confidence interval for the mean is

$\overline{x}$ ***± Margin of error***

where, $\overline{x}$ is the point estimate of the population mean.

The general form of the confidence interval for the proportion is

$\overline{p}$ ***± Margin of error***

where, $\overline{p}$ is the point estimate of the population proportion

The confidence interval estimates for the mean and proportion discussed earlier follow the general form explained above. Using these equations, the margin of errors for the confidence intervals for the mean and proportion are summarized in a table (next slide).

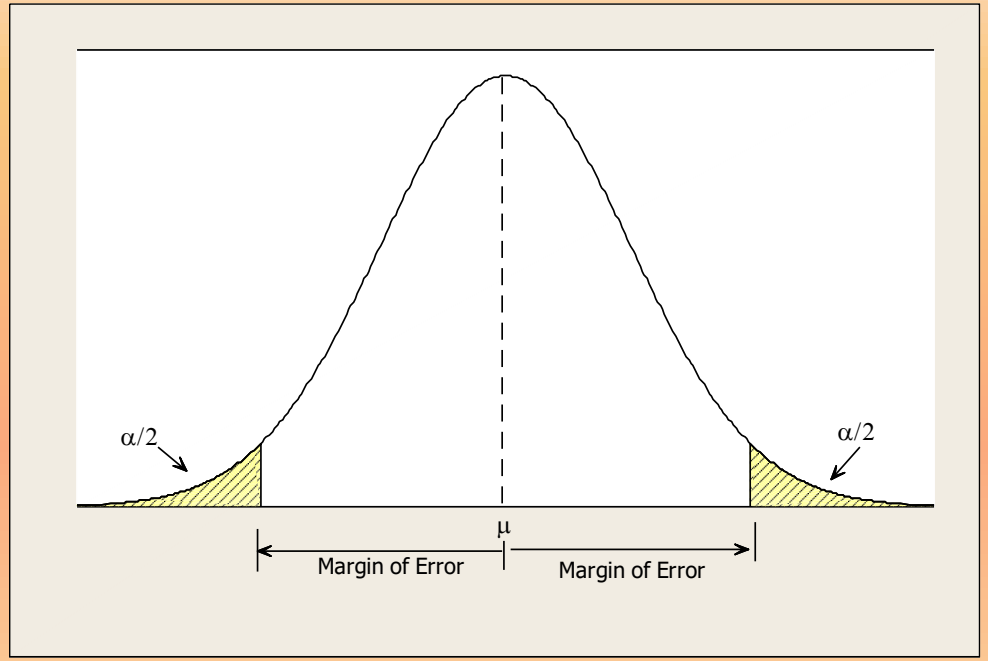
Margin of Error…cont.

Confidence Intervals and Margins of Error

Confidence Interval Estimate for the Mean		
Requirements	***Confidence Interval Estimate***	***Margin of Error***
(1) The population standard deviation σ known	$\overline{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
(2) The population standard deviation σ unknown; population is normal --- If the population is normally distributed, then the t-distribution can be used for both small or large samples.	$\overline{x} \pm t_{n-1,\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$	$t_{n-1,\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$
Confidence Interval Estimate for the Proportion		
Large sample size, Normal approximation	$\overline{p} \pm z_{\alpha/2} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$	$z_{\alpha/2} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$

Margin of Error…cont.

Note: The margin of error is one-half the width of the confidence interval. See Figure below.



Margin of Error as the half-width of a Confidence Interval

Example 1

The assembly time of a particular electronic appliance is of interest. A random sample of 25 assembly times (in minutes) is given below.

22.8	29.3	27.2	30.2	24.0	23.2	22.9	30.3	27.1	31.2	27.0
32.0	28.6	24.1	28.9	26.8	23.4	25.1	26.6	25.7	28.1	31.5
24.8	25.2	26.8								

(a) Find the mean and standard deviation of the assembly time data.

The sample mean and standard deviation of the above sample are:

$$\bar{x} = 26.912$$

$$s = 2.809$$

Use your calculator to find these values

(b) Determine an 80% confidence interval for the average assembly time of all workers using the 25 observations.

Since the population standard deviation σ is unknown, use t-distribution to compute the confidence interval. The confidence interval formula in this case is:

$$\bar{x} \pm t_{n-1,\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$

$$26.91 \pm t_{24,0.10}\left(\frac{2.809}{\sqrt{25}}\right)$$

$$26.91 \pm (1.318)\frac{2.809}{\sqrt{25}}$$

$$26.17 \le \mu \le 27.65$$

Example 1…cont.

Note that for an 80% confidence interval, α=0.20 and we need the value of $t_{24,0.10}$. From the *t-table* for 24 degrees of freedom and $\alpha = 0.10$, this value is 1.318. (part of the t-table is shown in below). To obtain this value, read down the column of 0.10 and degrees of freedom of 24 in the table.

	Area in Upper Tail				
Degrees of Freedom(Df)	0.100	0.050	0.025	0.010	0.005
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
:					

Example 1…cont.

(c) Determine a 90% confidence interval for the average assembly time of all workers using the sample size of 25.

Using the confidence interval formula in part (b), the confidence interval is

$$26.91 \pm t_{24,0.05}\left(\frac{2.809}{\sqrt{25}}\right)$$

$$26.91 \pm (1.711)\frac{2.809}{\sqrt{25}}$$

$$25.95 \le \mu \le 27.87$$

(d) Determine a 95% confidence interval for the average assembly time.

$$26.91 \pm t_{24,0.025}\left(\frac{2.809}{\sqrt{25}}\right)$$

$$26.91 \pm (2.064)\frac{2.809}{\sqrt{25}}$$

$$25.75 \le \mu \le 28.07$$

(e) Compare your answers obtained in parts [b], [c], and [d]. What happens when the confidence interval is increased, and when the confidence level is decreased?

As the confidence level increases, the interval gets wider. You gain confidence but lose accuracy.

Example 2.

The life in hours of a 20 watt halogen bulb is known to be approximately normally distributed with a standard deviation $\sigma = 50$ hours. A random sample of 25 bulbs has a mean life of $\bar{x} = 1500$ hours.

(a) Construct a 95% confidence interval on the mean life.

The sample size is small (n < 30) but you can still use the normal distribution to calculate the confidence interval. This is because the population standard deviation is known. Note that

$n = 25; \bar{x} = 1500$

$\sigma = 50$

The confidence interval can be calculated as shown

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

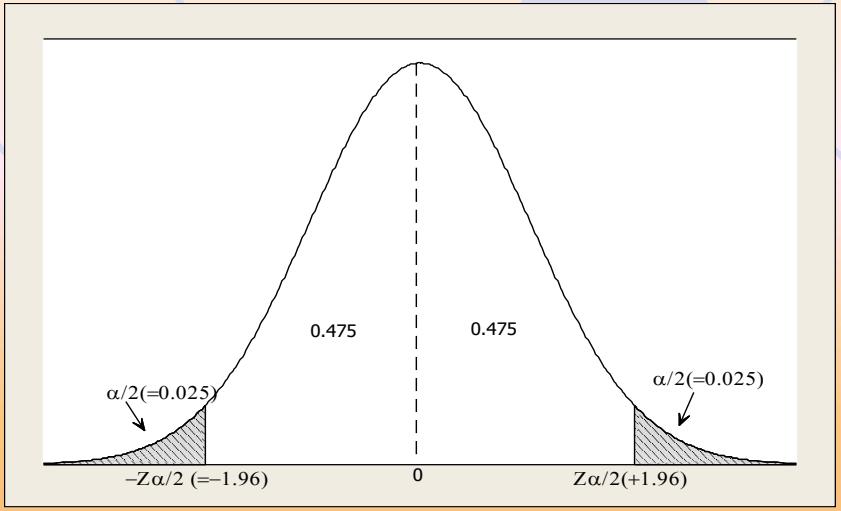
$$1500 \pm 1.96\left(\frac{50}{\sqrt{25}}\right)$$

$$1480.4 \le \mu \le 1519.6$$

In the above interval, $z_{\alpha/2} = 1.96$ is the z value for an area of 0.475 obtained from the normal table (see the figure on the next slide). To obtain this z value, locate 0.475 in the normal table and then read the corresponding z value in the extreme left row and the topmost column above of 0.475 in the table. This value is 1.96. If you don't find the exact area value of 0.475 in the table, find the closest value of the area to determine the z value. Note that $z_{\alpha/2} = 1.96$ value is for a confidence level of 95%. If you change the confidence level, then the z value will change accordingly.

Example 2…cont.

Z-value for a 95% C.I. using the Normal Distribution



The table shows the Z values for commonly used confidence levels. You should verify these values using the normal table on the next page.

Confidence Level	α	$\alpha/2$	Corresponding Z values ($z_{\alpha/2}$)
90%	0.10	0.05	$z_{\alpha/2}=1.645$
95%	0.05	0.025	$z_{\alpha/2}=1.96$
98%	0.02	0.01	$z_{\alpha/2}=2.33$
99%	0.01	0.005	$z_{\alpha/2}=2.58$

Standard Normal Distribution Table

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499

Example 2…cont.

(b) Construct a 95% confidence interval on the mean life based on a sample size of 36

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$1500 \pm 1.96\left(\frac{50}{\sqrt{36}}\right)$$

$$1483.7 \le \mu \le 1516.3$$

(c) Construct a 95% confidence interval on the mean life based on a sample size.

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$1500 \pm 1.96\left(\frac{50}{\sqrt{100}}\right)$$

$$1490.2 \le \mu \le 1509.8$$

(d) Based on the above answers, discuss the impact of increasing sample size on the confidence interval.

As the sample size increases, the width of the confidence interval becomes smaller. This means that the interval becomes more precise.

Example 3

A random sample of size n=18 was taken from a manufacturing process that produces metal rods for an automobile suspension system. Of primary concern is the diameter of the manufactured rods. The diameter is believed to follow a normal distribution. The measurements below show the eighteen diameters in millimeter (mm).

10.85, 11.40, 10.81, 10.24, 10.23, 9.49, 9.89, 10.11, 10.57, 11.21, 10.10, 11.22, 10.31, 11.24, 9.51, 10.52, 9.92, 8.33.

What is a 95% confidence interval on the mean diameter?

Solution:

First, find the mean and the standard deviation of the 18 values. These are

$$\bar{x} = 10.33$$

$$s = 0.77$$

The confidence interval can be obtained using the t-distribution and is shown below.

$$\bar{x} \pm t_{n-1,\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

$$10.33 \pm (2.110) \left(\frac{0.77}{\sqrt{18}} \right)$$

$$9.95 \le \mu \le 10.71$$

With a 95% confidence (or probability), we can say that the average diameter is between 9.95 and 10.71 mm.

Example 4

The average life of a sample of 36 tires of particular brand is 38,400 miles. If it is known that the average lifetime of the tires is approximately normally distributed with a standard deviation of 3,600 miles, construct 80%, 90%, 95%, and 99% confidence intervals for the average tire life. Compare and comment on the confidence interval estimates.

Solution: Given:

$$n = 36$$
$$\bar{x} = 38,400$$
$$\sigma = 3,600$$

Since the sample size is large (n ≥ 30), and the population standard deviation is known, the appropriate confidence interval formula is

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

(a) 80% confidence interval

$$38,000 \pm 1.28\left(\frac{3,600}{\sqrt{36}}\right)$$
$$37,232 \le \mu \le 38,768$$

(b) 90% confidence interval

$$38,000 \pm 1.645\left(\frac{3,600}{\sqrt{36}}\right)$$
$$37,013 \le \mu \le 38,987$$

Example 4...cont.

(c) 95% confidence interval

$$38,000 \pm 1.96\left(\frac{3,600}{\sqrt{36}}\right)$$

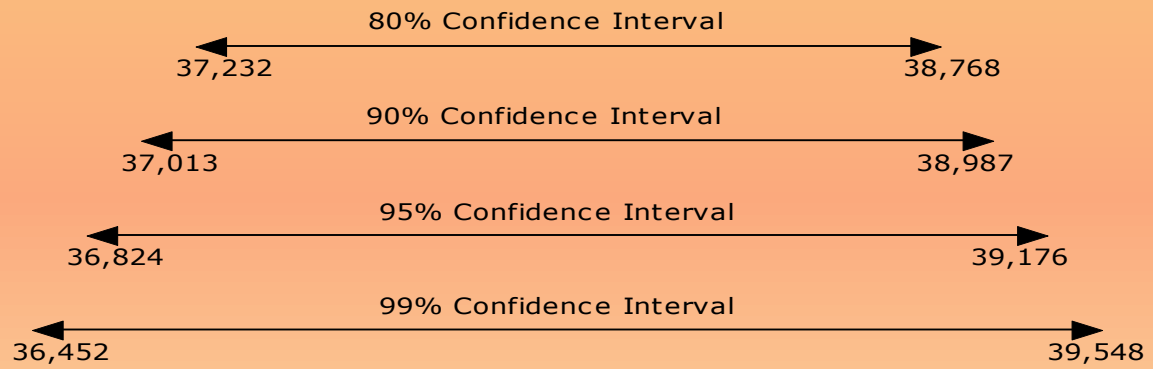
$$36,824 \le \mu \le 39,176$$

(d) 99% confidence interval

$$38,000 \pm 2.58\left(\frac{3,600}{\sqrt{36}}\right)$$

$$36,452 \le \mu \le 39,548$$

Note: the values in the above confidence interval calculations are obtained from the normal table. Refer to the standard normal table on page 29 and verify the values of z. Figure below shows the confidence intervals graphically.



Note: higher the confidence level, wider the length of the interval. This indicates that for a larger confidence interval, we gain confidence. There is higher chance that the true value of the parameter being estimated is contained in the interval but at the same time, we lose accuracy.

Example 5

The average college graduate has nearly $20,000 in debt. The average credit card debt has increased 47 percent between 1989 and 2004 for 25-to 34-year-olds and 11 percent for 18- to 24-*year*-olds. Nearly one in five 18- to 24-year-olds is in "debt hardship," up from 12 percent in 1989. (Source: Demos.org, "The Economic State of Young America," May 2008). Suppose that a telephone survey of 100 college graduates indicated a sample average debt of $22,500 with a standard deviation of $4,250.

(a) Find a 95% confidence interval using both the t-distribution and the normal distribution for the average amount of debt.

$n = 100$

$\bar{x} = \$22,500$

$s = \$4,250$

A 95% confidence interval using a t-distribution

$$\bar{x} \pm t_{n-1,\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm t_{100-1,0.025}\left(\frac{s}{\sqrt{n}}\right)$$

$$22500 \pm (1.984)\left(\frac{4250}{\sqrt{100}}\right)$$

$$21656.80 \le \mu \le 23343.20$$

A 95% confidence interval using a normal distribution

$$\bar{x} \pm Z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$22500 \pm 1.96\left(\frac{4250}{\sqrt{100}}\right)$$

$$22500 \pm 833$$

$$21667; 23333$$

$$or, 21667 \le \mu \le 23333$$

Example 5…cont.

Note: In this case, since the population standard deviation is unknown, a t-distribution should be used to compute the confidence interval. However, for such a large sample (n=100), the population standard deviation can be approximated by the sample standard deviation, s and the confidence interval using both the normal and t-distribution produced very close results.

(b) Find the margin of error?

The margin of error can be calculated as shown.

$$Z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$= 1.96\left(\frac{4250}{\sqrt{100}}\right)$$

$$= 833$$

(c) Show that the margin of error is one-half of the width of confidence interval. The confidence interval for this problem calculated in part (a) was

The confidence interval for this problem calculated in part (a):

$$21667 \le \mu \le 23333$$

From this interval the width of the interval is (23333-21667) = 1666. Since the margin of error is one-half of the width of the interval, we see that 1666/2 = 833 which is the margin of error.

Example 6 – confidence interval for the proportion

Out of a random sample of 200 students at a university, 180 stated that they carried a cell phone. Based on this survey result, construct a 90% and a 95% confidence interval estimate of p for the proportion of all the students at the university who carry a cell phone.

Solution: The confidence interval formula for estimating the population proportion is $\bar{p} \pm Z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$

Where $\bar{p}$ is the sample proportion that can be obtained as $\bar{p} = \frac{x}{n} = \frac{180}{200} = 0.90$. For a 90% confidence interval, α=0.10 and we need the value of $Z_{\alpha/2} = Z_{0.05}$. This is equivalent to finding a Z value for an area of 0.45. From the normal table, this value is 1.645. Thus, the 90% confidence interval is

$$0.90 \pm (1.645)\sqrt{0.90(1-0.90/200}$$
$$0.90 \pm 0.035$$
$$0.865 \le p \le 0.935$$

A 95% confidence interval can be calculated in the same way. The interval is as follows:

$$0.90 \pm (1.96)\sqrt{0.90(1-0.90/200}$$
$$0.90 \pm 0.042$$
$$0.858 \le p \le 0.942$$

or, with a 95% confidence, we can assert that the true percentage of the students carrying a cell phone is between 85.8 and 94.2 percent.

Sample Size Determination

Determine the sample size (n) for estimating population mean (μ)

The confidence interval formula to estimate μ when the population standard deviation is known is given by:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where, $Z * \sigma / n$ is the accuracy (or, the margin of error) of the confidence interval. This means that when using $\bar{x}$ to estimate μ, the error $E = |\bar{x} - \mu|$ is less than $Z * \sigma / n$ with confidence $100(1-\alpha)$. Thus, in determining the sample size n, we can be $100(1-\alpha)$ % confident that the error in estimating μ is less than a specified error, E. The expression for the sample size is given by

Formula to determine the sample size ***(n)*** to estimate μ

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

where, the value of Z depends upon the confidence level desired, E is the specified error, and σ is the population standard deviation.

Determine the sample size for estimating the population proportion (p)

The sample size required to estimate a population proportion is derived using the normal approximation

$$z = \frac{(\bar{p} - p)}{\sqrt{p(1-p)/n}}$$

Solving the above expression for n, the required sample size is:

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2}$$

where, n is the required sample size, the value of z is based on the confidence level desired, or z is the value corresponding to an area of $(1-\alpha)/2$ from the mean of a standard normal curve. The population proportion of success is *p*, and E is the error of estimation.

Example 7

A human resources (HR) consulting firm would like to estimate with 95% confidence the mean starting salary for entry level graduates in management. Past data indicate that the entry level salary can be approximated by a normal distribution with a standard deviation σ=$5000.

(a) How large a sample size should the firm choose if the error in estimating the salary is less than $2000 with 95% confidence?

Solution : The sample size n is given by

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

Since, $Z_{\alpha/2} = Z_{0.025} = 1.96$, σ=5000, and the error E=2000, the required sample size will be

$$n = \left(\frac{(1.96)5000}{2000} \right)^2 = 24.01 \approx 25$$

(b) ***How large a sample is needed if the error in estimating the salary is (i) less than $1000 (ii) less than $500 (with 95% confidence)?***

The required sample size (i),

$$n = \left(\frac{(1.96)5000}{1000} \right)^2 = 96.04 \approx 97$$

(ii)

$$n = \left(\frac{(1.96)5000}{500} \right)^2 = 385$$

Example 8

A survey of 180 randomly selected economists 144 revealed that the economy of the United States will not slip into a recession for at least a year.

(a) Construct a 90% confidence intervals for the percentage of economists who believe that a recession will not occur for at least a year. Interpret your result.

Solution: Note that $\bar{p} = \frac{x}{n} = \frac{144}{180} = 0.80$

The 90% confidence interval is as follows:

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$0.80 \pm (1.645)\sqrt{\frac{0.80(1-0.80)}{180}}$$

$$0.80 \pm 0.05$$

$$0.75 \le p \le 0.85$$

(b) Construct a 95% confidence interval for the percentage of economists who believe that a recession will not occur in at least a year.

$$\bar{p} \pm Z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$0.80 \pm (1.95)\sqrt{\frac{0.80(1-0.80)}{180}}$$

$$0.80 \pm 0.058$$

$$0.742 \le p \le 0.858$$

(c) Compare and comment on the results obtained in parts [a] and [b].

With higher confidence level, the confidence interval became wider.

(d) Determine the sample size for this study if we want to be 95% confident that the margin of error in estimating this percentage (the percentage of economists who believe that a recession will not occur for at least a year) is less than 0.05.

The required sample size can be determined using

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2}$$

Since, $z_{\alpha/2} = z_{0.025} = 1.96$, the population proportion, p =0.5 (note that if the value of p is not known, we can specify this value to be 0.5), the sample size n is

$$n = \frac{(1.96)^2 0.5(0.5)}{(0.05)^2} \approx 385$$

Answer part [d] if the error is less than 0.10 and all other values remaining the same.

The required sample size is

$$n = \frac{(1.96)^2 0.5(0.5)}{(0.10)^2} \approx 97$$

Note the effect of changing the error requirement from 0.5 to 0.1 in parts [c] and [d].

Example 9

A pressure seal used in an assembly must be able to withstand a maximum load of 6,000 pounds per square inch (psi) before bursting. If the average maximum load of a sample of seals taken from a shipment is less than 6,000 psi, then the quality control must reject the entire shipment. How large a sample is required if the quality engineer wishes to be 95% confident that the error in estimating this quantity is no more than 15 psi, or the probability that the sample mean differs from the population mean by no more than 15 psi is 0.95. From the past experience, it is known that the standard deviation for bursting pressures of this seal is 150 psi.

Solution:

The sample size n can be calculate using

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

Since, $Z_{\alpha/2} = Z_{0.025} = 1.96$, σ=150, and the error E=15, the required sample size will be as follows:

$$n = \left(\frac{(1.96)150}{15} \right)^2 = 385$$

Confidence Interval for the Variance- (Optional)

Suppose $x_1, x_2, \ldots, x_n$ be the random sample of size n drawn from a normally distributed population with unknown mean μ and variance σ^2. If s^2 is the sample variance given by the following expression:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

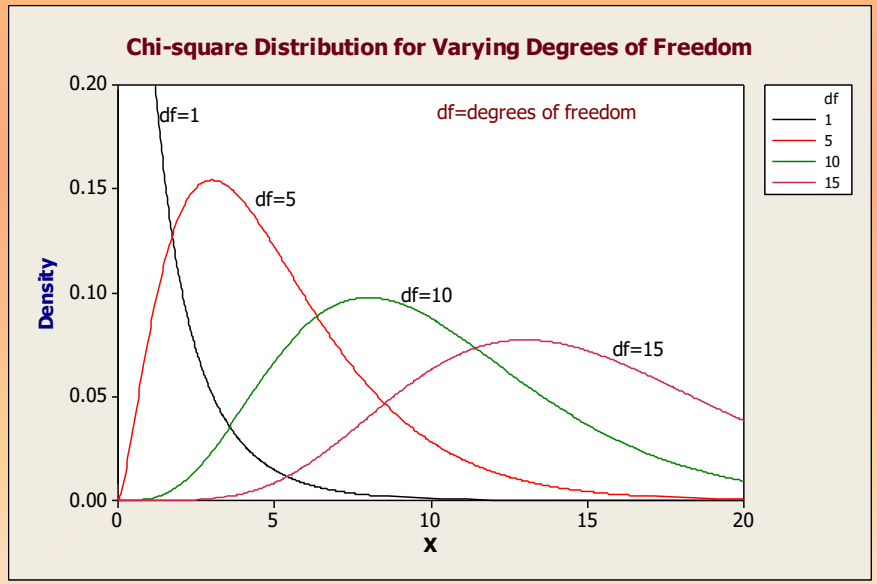
then it can be proved that

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum (x_i - \bar{x})^2}{\sigma^2}$$

has a ***chi-square distribution*** with (n-1) degrees of freedom.

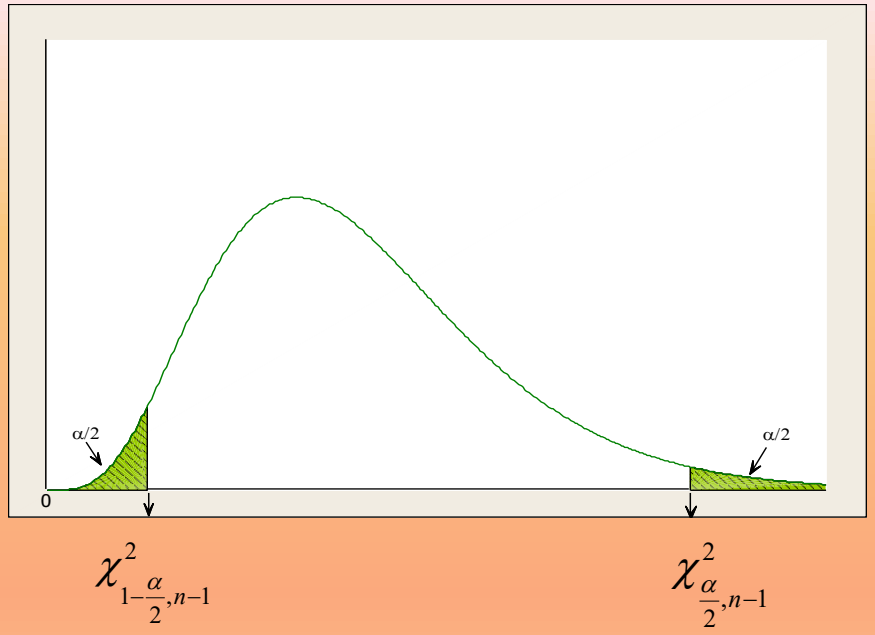
The confidence interval formula for the variance can be developed using a ***chi-square distribution***.

The shape of the chi-square distribution depends upon the degrees of freedom. Figure on the right shows the chi-square distribution for different degrees of freedom.



Developing the Confidence Interval for the Variance:

To develop the confidence interval, refer to the figure below



From the Figure

$$p[\chi^2_{1-\alpha/2,n-1} \le \chi^2 \le \chi^2_{\alpha/2,n-1}] = 1-\alpha$$

or,

$$p[\chi^2_{1-\alpha/2,n-1} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2,n-1}] = 1-\alpha$$

Multiplying each element of the above equation by $(n-1)s^2$ results into

$$p[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}] = 1-\alpha$$

It follows that a 100(1- α) percent two-sided confidence interval for is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

Example 10

Two identical machines are used to fill boxes with dishwashing detergent. A random sample of size n=20 boxes was taken from this process. Since the machines are identical and are used to fill the same size of detergent boxes, the boxes produced using these two machines are mixed. Of primary concern is the variance in the weight of the detergent boxes. The data below show the weight of twenty boxes in ounces.

Wt. in Ounce

71.24	71.39	70.84	70.11	71.09	70.88	71.03	70.42	71.30
71.49	70.66	71.04	71.31	71.40	71.25	70.80	70.63	71.43
71.37		70.89						

(a) Find a 99% two-sided confidence interval of the population variance, σ^2.

Solution:

First, find the variance of the sample data using a calculator. The variance is $s^2 = 0.139$

A 99% confidence interval for the variance can be calculated as shown below.

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} = \frac{(19)(0.139)}{\chi^2_{0.005,19}} \le \sigma^2 \le \frac{(19)(0.139)}{\chi^2_{0.995,19}} = \frac{(19)(0.139)}{38.582} \le \sigma^2 \le \frac{(19)(0.139)}{6.844}$$

Therefore a 99% confidence interval for the variance is: $0.0685 \le \sigma^2 \le 0.3859$

The values $\chi^2_{0.005,19}$ and $\chi^2_{0.995,19}$ are obtained from the chi-square table.

Estimating Population Values: Confidence Intervals

Confidence interval formulas to estimate the population mean, μ and Population Proportion (p)

Confidence interval formulas to estimate the population mean, μ

Case (1): Large sample (n), σ known: use normal distribution

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or,

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Margin of Error for estimating μ when σ is known

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

Case (2): Large sample (n), σ unknown: use normal distribution

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

or,

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Case (3): Small sample (n), σ unknown: use t- distribution

$$\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \quad \text{or,}$$

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

Confidence interval formulas to estimate the population proportion, p

Assumption: sample size n is large so that normal approximation can be used

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

or,

$$\bar{p} - z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \le p \le \bar{p} + z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

In the above formula,

p = population proportion

$\bar{p} = \frac{x}{n}$ *(Sample proportion)*

Determine the sample size (n)

Determine the sample size (n) to estimate μ

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

E=error or accuracy

n= sample size

Determine the sample size (n) to estimate p

$$n = \frac{(z_{\alpha/2})^2 p(1-p)}{E^2}$$

p= population proportion (if p is not known or given, use p=0.5)

Some commonly used confidence intervals and corresponding Z-values

90% Confidence Interval	*Z=1.645*
95% Confidence Interval	*Z=1.96*
99% Confidence Interval	*Z=2.58*

Some less commonly used confidence intervals and their Z-values

80% Confidence Interval	*Z=1.28*
94% Confidence Interval	*Z= 1.88*
96% Confidence Interval	*Z=2.05*