MTH410

Quantitative Business Analysis

Module 5: Confidence Intervals

Surveys and polls provide a large amount of data for those who initiate them, but what do the responses really mean, and how do statisticians use them to make the definitive statements we hear all around us? In Module 5, we will work to understand how sampling and inference are used in everyday life and how probabilities and interval estimation can help us use data from surveys and polls to make decisions.

Learning Outcomes

1. Evaluate the concepts of sampling and inference and their value in decision making.
2. Calculate probabilities of sample proportions and of sample means.
3. Demonstrate understanding of interval estimation methods.

For Your Success & Readings

As you read the lecture and textbook, try to mentally reference how sampling and point estimation might work in your real life to better understand groups of people and clusters of data. This module has some excellent, basic information you will need should you ever need to run a survey on customers, employees, or even product quality.

To navigate through this module successfully, keep the following in mind:

- This week, you will complete the third Critical Thinking Assignment. Review the assignment early in the week and contact your instructor if you have any questions or concerns.

Required

- Chapters 7 & 8 in *Introductory Business Statistics*

Recommended

- Brussolo, M.E. (2018). **Understanding the central limit theorem the easy way: A simulation experiment** (https://www.mdpi.com/2504-3900/2/21/1322). *Proceedings, 2*(21), 1322.
- Platikanova, M., Hristova, P., & Milcheva, H. (2018). Mathematical model for forecasting the influence of atmospheric pollution on population morbidity in Stara Zagora municipality (Bulgaria). **Open Access Macedonian Journal of Medical Sciences** (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5985859/), *6*(5), 934–939. doi:10.3889/oamjms.2018.205

1. Sampling Distributions

Sampling allows us to collect data to answer a research question about a population without obtaining information on each entity in the population. For example, if we wanted to know how many online adult students did their homework while at their workplace, we could determine how many students would comprise a statistically significant sample, and survey them. If 50% of the sample responded that they did, in fact, do their homework while at work, we could say that, based on the sample data, 50% of all adult learners do their homework while at work.

It may be, however, that a sample does not fully represent the population being studied. This can occur because the population is so highly diverse that a selected sample alone is not representative of the population, or because the sample is being taken from an ongoing process in which the sampled population is conceptually infinite. For example, if we wanted to examine the qualities of online instructors, there are so many variations in personalities, classroom contributions, grading factors, and communication styles that it would not be possible to take a select group of instructors as a sample of the entire population of online instructors. In such a case, a random sample (one that is comprised of data that have an equal chance of being included or excluded) can be drawn to represent the probability distribution of the population.

Another way to predict population attributes for a study is to use **point estimation**. This method uses a single value determined from the sample from which the mean is derived to become the point estimate. For example, if you wanted to determine the increase in pay that the average graduate earns post-graduation, you could examine the range of post-graduation pay increases for a sample group, determine the mean amount, and use it as representative of the population.

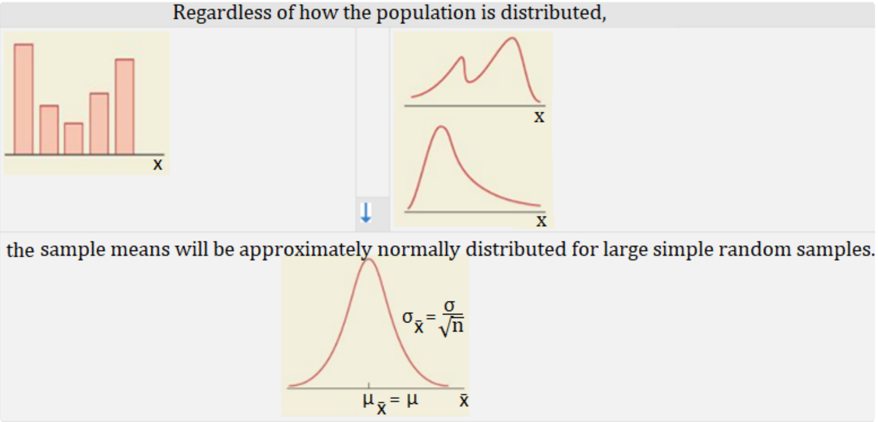
If, however, you ran the study again using a different sample group, you would likely end up with a different point estimate; and if you ran it again with five other sample groups, you would probably have five more point estimates. All of these point estimates can then be used to determine the mean of all of the point estimates, the standard deviation of the point estimates, and the probability distribution of the point estimates, otherwise referred to as the **sampling distribution**.

The last step in identifying the characteristics of the sampling distribution of the mean is to determine the **shape of the sampling distribution**.

If many simple random samples of size *n* are drawn from a population that is normally distributed, the sample means will also be normally distributed. The standard deviation of the distribution (the standard error of the distribution) will be smaller when the number of samples *n* is larger. Formulas that will be used with the **sampling distribution of the mean**,   , follow:

Sample Mean:	**Sample Size:** n	**Population Mean of** X: μ	**Population Standard Deviation of** X: σ
Mean			
Standard Deviation			
z**-Score**			

Even if we do not know whether the population is normally distributed, if the sample size is large (30 or more), the sampling distribution of the mean can be assumed to be approximately normal. This result is due to the **Central Limit Theorem for Means.** The larger the sample size, the better the approximation of the normal distribution.


Regardless of how the population is distributed,

the sample means will be approximately normally distributed for large simple random samples.

Source: Weiers, 2011

Here is a further explanation of the Central Limit Theorem for Means. As the sample size, n, becomes larger, the distribution of approaches a bell-shaped curve. This is true regardless of the shape or distribution of X: **Central Limit Theorem: Summary** (https://statistical-engineering.com/clt-summary/)

See **Figures 7.3, 7.4, and 7.5 of our textbook** (https://cnx.org/contents/tWu56V64@35.2:Mjy3YF-Z@20/7-2-Using-the-Central-Limit-Theorem) (pages 311–315) explaining the same concept

Regardless of how the random variable X is distributed, the distribution of approaches a bell-shaped curve as the sample size increases.

Recall this example from Module 4 on the normal distribution:

Suppose the height of an adult male is normally distributed with a mean μ = 69 in. and standard deviation of σ = 2.5 in.

Find the probability that a randomly chosen person has height greater than 73 in. (6' 1").

We first found the z-score: $z = (x - \mu)/\sigma = (73\text{-}69)/2.5 = 1.6$

Then, we were able to find the answer: $P(X > 73) = P(Z > 1.6) = 0.0548$.

Suppose we modify the above problem as follows:

Suppose the height of an adult male has an *unknown distribution* and has a mean μ = 69 in. and standard deviation of σ = 2.5 in.

1. What is the mean height for the sample mean distribution?
2. What is the standard deviation of the sample mean distribution of 36 persons?
3. Find the probability that a randomly chosen sample of 36 persons has a **mean** height greater than 70 in. (5' 10").

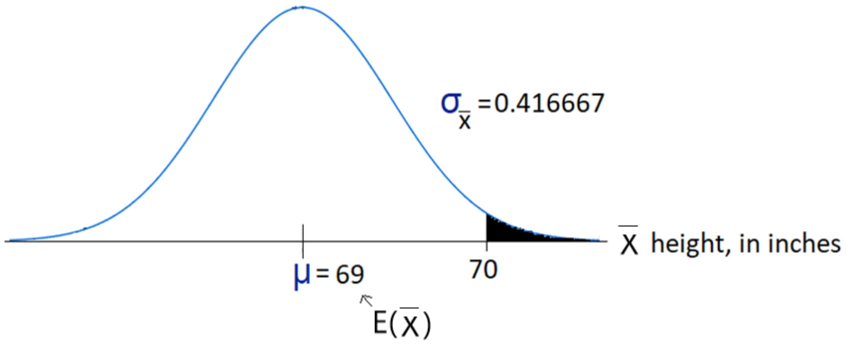
Click "Solution" to check your thinking.

Solution

- The mean height of the sample mean distribution is the mean of X:
- The standard deviation of the sample mean distribution is:
- We want to find this probability concerning the sample mean:

Notice that the main difference between this problem and the one right above it from Module 4 is we are now interested in a probability of the sample mean. Also, we are not assuming that the height is normally distributed. **We do not need to know the distribution of height.**

According to the Central Limit Theorem for Means, the distribution of is approximately normal. As a rule of thumb, one uses the Central Limit Theorem for Means if the sample size for is greater than 30. Since that is the case here (n = 36), we may assume that is approximately normally distributed.



Apply the z-score formula of to the value of 70:

Thus, by the Central Limit Theorem for Means, we have:

We used Table A11 on page 607 of Appendix A of the textbook (Standard Normal Probability Distribution: Z Table). The table only has positive z-values, to the right of z=0 under the standard normal distribution (see Figure A2 on page 606).

We thus subtracted the area in the table, $P(0 < Z < 2.40)$, from 0.5: 0.5 - $P(0 < Z < 2.40) = .5 - 0.4918 = 0.0082$

A quicker way to compute normal probabilities is to use a **normal distribution calculator** (https://www.mathportal.org/calculators/statistics-calculator/normal-distribution-calculator.php).

In the above example, to find $P(Z > 2.40)$, we can input the mean and standard deviation of the standard normal distribution into the calculator, μ = 0 and σ = 1. Then, we would select the radio button corresponding to the "greater than" probability and fill-in the blank for the right endpoint, 2.40. Finally we would press "Compute" to obtain the answer:

$P(Z > 2.40) = 0.0082$

We can also use the **Central Limit Theorem for Sums**. This follows by a modification of the Central Limit Theorem for Means and by using properties of the variance:

Sample Mean:	**Sample Size**: n	**Population Mean of** X: μ	**Population Standard Deviation of** X: σ
Mean			
Standard Deviation			
z-**Score**			

In the example above, we can find the probability that the sums are less than 2490, . The z-score is

Thus,

We used Table A11 on page 607 of Appendix A of the textbook to find $P(0 < Z < 0.40)$.

The **mean for the sample sum distribution** is

In addition, the **standard deviation for the sample sum distribution** is

Central Limit Theorem for Proportions

The sample **proportion**, $p' = x/n$, is the point estimator of the population proportion, p. To determine how close the sample proportion, p', is to the population proportion, p, we need to understand the properties of the sampling distribution of p'. In other words, we need to understand the expected value (mean) of p', the standard deviation of p', and the shape or form of the distribution of p'.

The practical use of the sampling distribution of p' is that it can be used to provide probability information about the difference between the sample proportion, p', and the population proportion, p. For example, if you wanted to know the probability of obtaining a value of p' that is within a given range, you can use the formula for the standard deviation of p' on p. 320 to determine the standard deviation of p'. You can then use Table A11 on page 607 of Appendix A to determine the z value (the number of standard deviations a variable is from the mean) to determine the sample proportion within the desired range.

If both np and $n(1-p)$ are greater than 5, then the **sampling distribution of the proportion** will be approximately normally distributed.
The formulas used for the sampling distribution of the sample proportion, p', are as follows:

Sample Proportion: $p' = x/n$	**Sample Size**: n	**Population Proportion**: p	**Sample Proportion of Failures:** $q' = 1-p'$
Mean	$E(p') = p$		
Standard Deviation			
z-**Score**			

When sampling from a finite population without replacement, we need a correction factor for our standard error calculation. It needs to be applied whenever the sample is greater than 5% of the population size. The second term in each of the below cases is the **finite population correction** factor for p':

Population Size: N	
Standard Deviation	
z-**Score**	

Note: There is a **finite population correction factor for the mean**, too. This is the first formula in Section 7.4. However, we will not use that formula, unless otherwise noted.

EXAMPLE

Suppose 38% of the population of factory workers belonging to a certain union prefers working the third shift. Find the probability that in a sample of 36 workers, fewer than 40% of them will prefer working the night shift.

Click "Solution" to check your thinking.

Solution

We need to find *P*. The population proportion is *p* = 0.38, and the sample size is *n* = 36. First, we need to determine if one may use the normal distribution. We need to determine if both *np* and *n*(1 - *p*) are greater than 5:

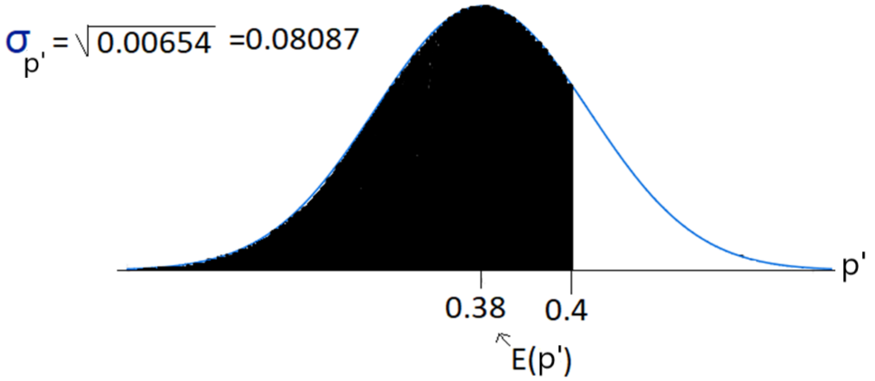
np = 36(0.38) = 13.7 > 5

and

n(1 - *p*) = 36(1 - 0.38) = 22.3 > 5

Thus, we may proceed in using the normal distribution. Had at least one of the above inequalities been false, then we would have needed to use the binomial distribution. The *z*-score corresponding to the sampling distribution of *p*’ is

The shaded area represents the probability that we want to find:



The above is a graph of the normal distribution with mean 0.38 and standard deviation 0.08087. The shaded area to the left represents the probability that fewer than 40% prefer the night shift.

We used Table A11 on page 607 of Appendix A of the textbook to find the probability *P*(0 < *Z* < 0.25).

1.1. More Examples of the Central Limit Theorem for Means

Let’s look at a few more examples of the Central Limit Theorem for Means.

EXAMPLE

Suppose that the mean time taken to perform a certain task at a factory is 15 minutes with a standard deviation of 7 minutes. A sample, with size n = 33, was randomly drawn from the population.

a. What is the mean of the sample mean distribution?
b. What is the standard deviation for the sample mean distribution?
c. Find the probability that the sample mean will be at most 18.

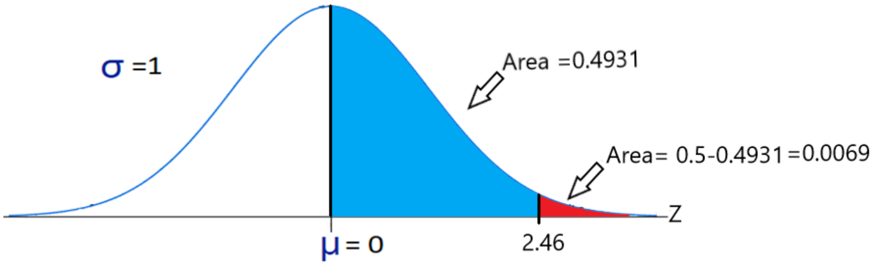
Click “Solution” to check your thinking.

Solution

In this problem, we are not given any information about how the random variable, X, is distributed. However, by the Central Limit Theorem for Means, we can assume that the random variable is approximately normally distributed. This is because the sample size, n, is large enough; larger than 30.

- The mean of the sample mean distribution is
- The standard deviation for the sample mean distribution is .
- We need to find $P($ <18). The z-score for sample means is:

The result $P(z < -2.46) = P(Z > 2.46)$ follows by symmetry of the normal distribution. See the following diagram. The blue area is $P(0 < Z < 2.46)$. The probability $P(0 < Z < 2.46)$ is from Table A11 on page 607 of Appendix A of the textbook.



1.2. More Examples of the Central Limit Theorem for Sums

Let’s continue looking at examples—this time of the Central Limit Theorem for Sums.

EXAMPLE

a. What is the mean of the sample sum distribution?
b. Find the standard deviation for the sample sum distribution.
c. Find the probability that the total number of hours to perform the task by 33 employees is more than 500 hours.

Click “Solution” when ready, to check your thinking.

Solution

In this problem, we are not given any information about how the random variable, *X*, is distributed. However, by the Central Limit Theorem for Sums, we can assume that the random variable is approximately normally distributed. This is because the sample size, *n*, is large enough; larger than 30.

- The mean of the sample sum distribution is
- The standard deviation for the sample sum distribution is
- We need to find the probability of the sum of employee hours is greater than 500 hours, . The *z*-score for the sums is:

2. Confidence Intervals

One can use sample data to determine point and interval estimates regarding the population mean or proportion. *Click on the tabs below to learn more about each.*

Point EstimateInterval Estimate

A **point estimate** is a single number that estimates the exact value of the population parameter of interest.

The **interval estimate** is a range of values that should include the actual population parameter.

Some examples of unbiased point estimators for the population parameters are shown in the table below.

Population Parameter	**Unbiased Estimator**	**Formula**
Mean, μ		
Variance, σ^2	s^2	–
Proportion, p	p’	

The purpose of **interval estimation** is to provide information about how close the point estimate, provided by the sample, is to the value of the population parameter. In order to develop an interval estimate of population mean, either the population standard deviation or the sample standard deviation must be used to compute the margin of error. In general, the larger the standard deviation, the larger the margin of error; and if the mean is smaller than the population mean, the correlation between the mean and standard deviation causes the margin of error to be small.

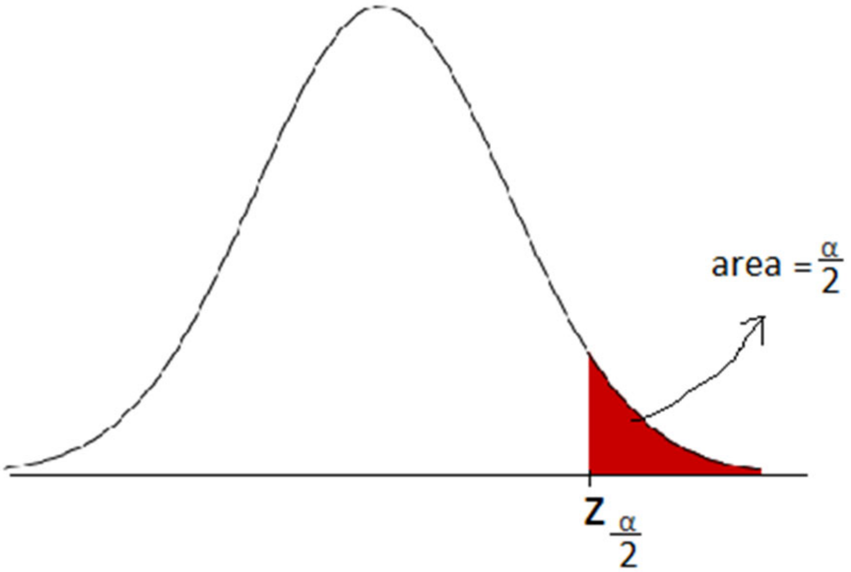
When the margin of error is given or calculated, the general form of an interval estimate of a population mean when a standard deviation is known is found through the formula:

Confidence Interval Estimate for the Population Mean, σ Known

where

is the Confidence Level (CL)

is the *z-score* providing an area of in upper tail of the standard normal probability distribution:



The above is the standard normal curve (mean is 0 and standard deviation is 1) whose shaded area to the right is . The value on the *x*-axis represents the *z*-value whose area to the right is

The term

is called the Error Bound, EBM. EBM stands for “Error Bound for a Population Mean.” The EBM is an example of a **Margin of Error**.

All confidence intervals are of the form

(Point Estimate) ± (Margin of Error)

In the previous formula, the Point Estimate of the population mean, is the sample mean, μ, .

Estimating Confidence Interval for when the Population Standard Deviation Is Known

Suppose the sample mean is 10, population standard deviation is $\sigma = 5$, the sample size is $n = 100$. We want to construct the 95% confidence interval. Then 1 - = .95. So = .05 and = .025. Then .

Applying this result to the formula listed above, we get the following:

The Error Bound is EBM=0.98

The Lower Limit of the confidence interval is 10 – 0.98 = 9.02.

The Upper Limit of the confidence interval is 10 + 0.98 = 10.98.

The 95% confidence interval is (9.02,10.98).

The interpretation is that we estimate with 95% confidence that the population mean is between 9.02 and 10.98.

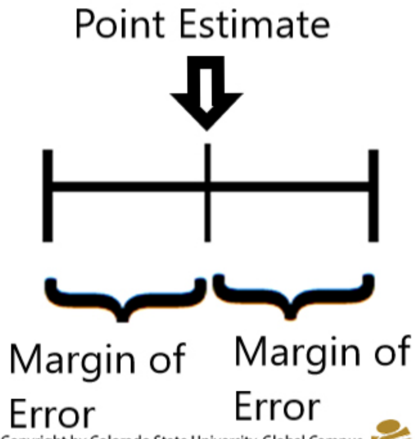
To find , we used Table A11 on page 607 of Appendix A of the textbook (Standard Normal Probability Distribution: Z Table). The table only has positive *z*-values, to the right of $z = 0$ under the standard normal distribution (see Figure A2 on page 606):

We looked up 0.475 on Table A11 on page 607 and obtained a *z*-score of

The meaning of a 95% confidence interval in the above example is if many such confidence intervals are similarly obtained, then approximately 95% of them will contain the population mean, μ. It does not guarantee that the confidence interval contains the population mean.

For a picture of what the above statement means, see the first figure in the webpage **"Understanding Hypothesis Tests: Confidence Intervals and Confidence Levels"** (http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests%3A-confidence-intervals-and-confidence-levels) on the Minitab Blog website. In the figure, the red confidence interval does not contain the population mean, μ. The others do contain the population mean, μ.

All confidence intervals are centered at the Point Estimate (midpoint) and half the length of the confidence interval is the Margin of Error:



© Copyright by Colorado State University-Global Campus

EXAMPLE

Suppose the confidence interval for the population mean is (35, 55). Find the Point estimate, , and the Error Bound (EBM).

Click "Solution" to check your thinking.

Solution

As shown above, the midpoint of the interval is the point estimate:

Also, half the length of the interval is the Margin of Error or EBM is as follows:

What happens if we do not know the standard deviation? In such cases, the sample data are used to estimate both the population mean and the population standard deviation, and we use the formula listed below:

If the sample size, n, is large ($n \geq 30$), we may use the sample standard deviation, s, and the normal distribution:

If the sample size, n, is small ($n < 30$) and σ is unknown we use the sample standard deviation, s, and the t-distribution if the underlying population is approximately normally distributed:

where

s is the sample standard deviation

is the Confidence Level (CL)

is the *t-score* providing an area of in the upper tail of the *t* distribution with = *n* - 1 degrees of freedom (*df*):

Each t-distribution is associated with different degrees of freedom. See Figure 8.8 on page 344 for different t-distributions along with their degrees of freedom. In general, more than 30 degrees of freedom will produce a t-distribution close to the normal distribution.

Estimating the Confidence Interval for μ when the Population Standard Deviation Is Unknown (n < 30)

Suppose the sample mean time taken to help customers at a certain store is 10 minutes with sample standard deviation is *s* = 5 minutes. Suppose the sample size is 22, and the time taken to help customers is approximately distributed. We want to construct the 95% confidence interval for the mean time taken to help customers. The Confidence Level is 1 – = .95. So = .05 and = .025. Using degrees of freedom = *df* = *n* – 1 = 22 – 1 = 21, we get the t-score:

Applying this result to the formula listed above, we get the following:

The Error Bound is EBM = 2.217.

The Lower Limit of the confidence interval is 10 – 2.217 = 7.783.

The Upper Limit of the confidence interval is 10 + 2.217 = 12.217.

The 95% confidence interval is (7.783, 12.217).

The interpretation is that we estimate with 95% confidence that the population mean time to help customers is between 9.02 minutes and 10.98 minutes.

To find , we used Table A12 on pages 608–609 of Appendix A of the textbook (Student's t-Distribution). The table contains t-values with a right-tail area of . We looked up = 21 degrees of freedom and = 0.025 to obtain = 2.080.

In the above example, one may also use **an online calculator** (http://www.danielsoper.com/statcalc/calculator.aspx?id=10) to find the critical value .

Degree of Freedom: 21

Probability Level: 0.025

Press **Calculate** and you will get:

t-value (right-tail): 2.07961385

With **Excel** you may use this command, where the probability is cumulative:

=T.INV(probability, degrees of freedom)

The Excel command will obtain :

=T.INV(0.025, 21)

= −2.07961

Note: In other textbooks, the t-distribution is **always** used if the value σ **is unknown,** even if the sample size is large**.**

Confidence Intervals for a Population Proportion:

If we want to determine the confidence interval estimate for the population proportion, we will use the following formula. We must use the sample proportion *p*’ as a point estimate of the unknown population proportion *p* and in estimating the standard deviation of the sampling distribution of the sample proportion.

Confidence interval limits for the population proportion:

= sample proportion =

= number of trials,

= *z*-score corresponing to the level of confidence

= Error Bound for a Population Proportion, EBP (Margin of Error)

NOTE: This assumes *np*’ and *nq*’ are both greater than 5.

EXAMPLE

A researcher wants to estimate the proportion of its customers who prefer Phone Plan A over the others. A survey of 200 customers determined that 120 preferred Phone Plan A. Construct the 99% confidence interval for the population proportion of customers who prefer Phone Plan A.

Click “Solution” to check your thinking.

Solution

The sample proportion of successes and failures are *p*’ and *q*’, respectively:

First check to see if and : 200(0.6) = 120 > 5 and 200(0.4) = 80 > 5

Thus, we may use the confidence interval formula.

The *z*-score is

$p' \pm \quad 0.6 \pm 2.58 \times \quad 0.6 \pm 2.58 \times 0.034641 = 0.6 \pm 0.089$

The confidence interval is:

The confidence interval is (0.511, 0.689) and the Error Bound (EBP) is 0.089. The interpretation is that we estimate with 99% confidence that the population proportion of customers who prefer Phone Plan A is between 51.1% and 68.9%.

2.1. More on Finding a Confidence Interval for the Mean for a Small Sample Size

If the sample size is small and the variable, *X*, is approximately normally distributed, we use the t-distribution. A sample size might be small due to expense of a large size or other constraints. For example, a small sample might be all that is available now. The following is an example of a small sample size.

EXAMPLE

A study was done on the average startup cost of small franchises that sell flowers. A business magazine provided data on the startup costs of 10 flower store franchises. The mean startup cost is computed to be $80 thousand with a sample standard deviation of $12 thousand. Find the 99% confidence interval for the population mean startup cost. Assume that the startup cost is approximately normally distributed.

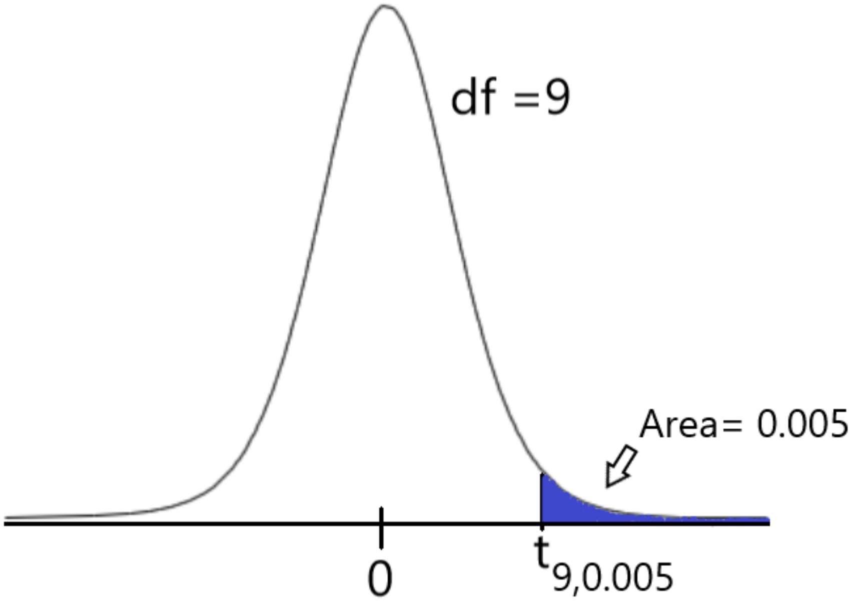
Click “Solution” to check your thinking.

Solution

This is a small sample (n < 30). Hence, we will use the t-distribution. The sample standard deviation is s = $12 and sample mean is . The Confidence Level (CL), is:

The value is .

The degrees of freedom are = $df = n - 1 = 10 - 1 = 9$. The t value satisfies that its right tail area is 0.005, as shown:



To find , we used Table A12 on pages 608–609 of Appendix A of the textbook (Student's t-Distribution). The table contains t-values with right-tail area as shown. We looked up = 9 degrees of freedom and right-tail area = 0.005 to obtain .

Applying these results to the confidence interval formula, we get the following:

The Error Bound is EBM = 12.333.

The Lower Limit of the confidence interval is 80 – 12.333 = 67.667.

The Upper Limit of the confidence interval is 80 + 12.333 = 92.333.

The 95% confidence interval is (67.667, 92.333).

The interpretation is that we estimate with 99% confidence that the population mean startup cost is between $67.667 thousand and $92.333 thousand.

2.2. More on Finding a Confidence Interval for a Proportion

It is often the case that we are interested in proportions. However, we are unable to sample the entire population. Confidence intervals for proportions provide an interval estimate of the sample proportion based on probability.

EXAMPLE

A bank manager conducted a survey of 50 local subprime mortgages and determined that 33% of them recently defaulted. Construct the 90% confidence interval for the population proportion of subprime mortgages that recently defaulted.

Click "Solution" to check your thinking.

Solution

The sample proportion is $p' = 0.33$ and $q' = 1-0.33 = 0.67$. The sample size is $n = 50$.

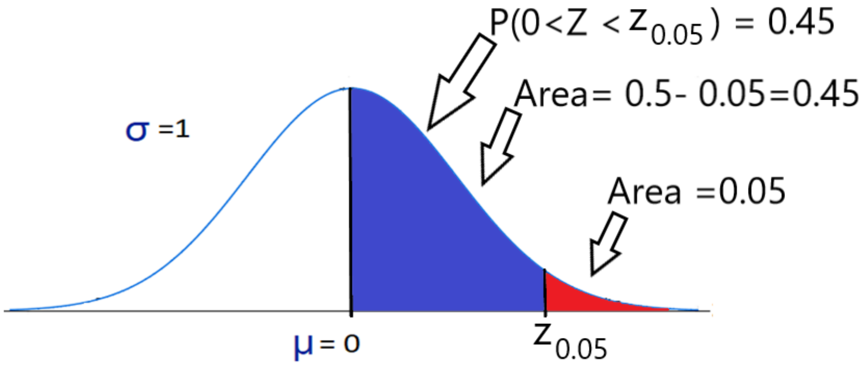
First check to see if $np' > 5$ and $nq' > 5$: $50(0.33) = 16.5 > 5$ and $50(0.67) = 33.5 > 5$

We may therefore use the confidence interval formula.

The Confidence Level (CL), is:

The value is .

To find the z value, , we used Table A11 on page 607 of Appendix A of the textbook (Standard Normal Probability Distribution: Z Table). The table only has positive z-values, to the right of $z = 0$ under the standard normal distribution (see Figure A2 on page 606): $0.5 – 0.05 = 0.45$



We look up the probability of 0.45 on Table A11 on page 607, and the value is not listed on the table. The probability value of 0.45 is between 0.4495 and 0.4505. Therefore, is between 1.64 and 1.65. To find the approximate value of , we will use the average of 1.64 and 1.65:

The confidence interval can now be found: $= 0.33 \pm 1.645 \times 0.066498 = 0.33 \pm 0.109$

The Error Bound is EBP= 0.109.

The Lower Limit of the confidence interval is $0.33 – 0.109 = 0.221$.

The Upper Limit of the confidence interval is $0.33 + 0.109 = 0.439$.

The 90% confidence interval is (0.221, 0.439).

The interpretation is that we estimate with 90% confidence that the population proportion of local subprime mortgages that defaulted is between 22.1% and 43.9%.

3. Finding the Sample Size

Sample Size Estimate for the Population Mean

An interesting problem in statistics is in determining the sample size to assure a certain margin of error. From a practical research perspective, larger sample sizes provide better approximations. However, the more highly skewed the population, the larger the sample size needs to be to obtain a good approximation. In most cases, a sample size of thirty or more will provide good approximate confidence intervals.

From the Error Bound term (or Margin of Error) in the confidence interval formula (for a known population standard deviation),

n =

one can algebraically solve for *n* (the sample size) and get the **Sample Size Formula for the Population Mean**:

= required sample size, = the *z*-value corresponding to the desired level of confidence

= known (or estimated) value of the population standard deviation

= acceptable error = acceptable difference between the population mean and the sample mean (margin of error)

The above formula provides the required sample size for estimating a **population mean** at a desired confidence level.

If the population standard deviation is unknown, one may replace by *s*, the sample standard deviation, and use this alternative version of the sample size formula:

EXAMPLE

Suppose one is interested in the mean weight of certain containers. Suppose, from a prior calculation, the sample standard deviation was 7.2 lbs. How large a sample is required to estimate the mean weight of containers to within 3.5 lbs. with 95% confidence?

Click "Solution" to check your thinking.

Each t-distribution is associated with different degrees of freedom. See Figure 8.8 on page 344 for different t-distributions along with their degrees of freedom. In general, more than 30 degrees of freedom will produce a t-distribution close to the normal distribution.

Solution

e = 3.5 lbs., and − = 1 − 0.9 = 0.05. Thus − − 10.20

<u>Round up the above value to the next largest whole number</u> to get an answer of *n* = 17 samples or more.

Sample Size for the Population Proportion

If we need to find the sample size for estimating a **population proportion** at a desired level of confidence and margin of error, we would use the following formula:

= required sample size, = the *z*-value corresponding to the desired level of confidence

= the estimated value of the population proportion (Use *p*’ = 0.5 if you have no idea of the actual value of *p*’.)

= acceptable error = acceptable difference between the population proportion and the sample proportion

3.1. More on Finding Sample Size for the Mean

In statistics, it is often a good idea to determine ahead of time the number of samples that one should collect. The determination of the sample size when constructing a confidence interval for a mean will depend on the maximum value of the Margin of Error (*e*), standard deviation (*s*), and Confidence Level ( ).

EXAMPLE

A researcher is interested in constructing a 90% confidence interval for the average cost of a certain type of backpack. From a preliminary study, the sample standard deviation is $13. If the Error Bound is at most $5, how many samples should be collected?

Click "Solution" to check your thinking.

Solution

$n =$

Use the sample size formula for a population mean:

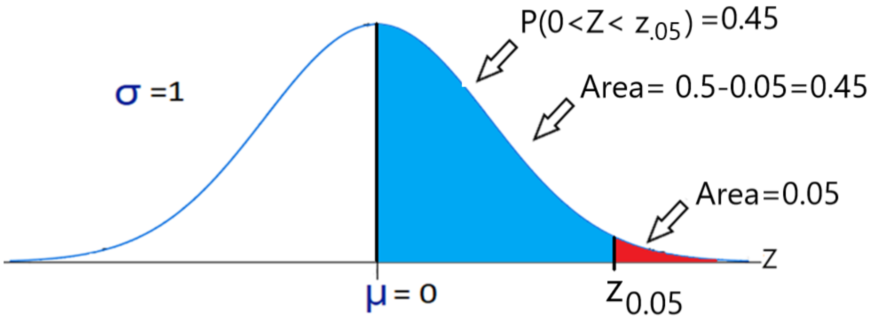
The sample standard deviation is *s* = $13 and the maximum value of the Margin of Error (acceptable error) is *e* = $5. The Confidence Level (CL), is:

The α value is = 0.1. The *z*-score is:

The value was found by going to the table of probabilities on Appendix A on page 607.

Since the table only has *z*-scores that are positive, we subtracted 0.04 from 0.5: 0.5 – 0.05 = 0.45

Thus, We need to find the *z*-score so that :



The probability value of 0.45 is not in the table. The value of 0.49 is between table values 0.4495 and 0.4505.

The value of is between the corresponding *z* values 1.64 and 1.65. Compute the average of 1.64 and 1.65:

$n = \quad = \quad = \quad = 18.3$

The sample size can now be computed:

Round up the answer of 18.3 to the next largest whole number. We get *n* = 19 samples.

3.2. Finding the Sample Size for a Proportion

The determination of the sample size when constructing a confidence interval for a proportion will depend on the maximum value of the Margin of Error (*e*), sample proportion estimate (*p*'), and Confidence Level (1 –).

EXAMPLE

Suppose the marketing department wants to determine the sample size for a 98% confidence interval of the proportion of people who buy their products after watching a commercial. If the Error Bound is at most 3%, how many people should be sampled?

Click "Solution" to check your thinking.

Solution

Recall that we need to use the sample size formula for a population proportion: $n =$

However, we do not have any preliminary estimate of the sample proportion *p*'. As mentioned previously, if there is no preliminary estimate of the sample proportion *p*', then use *p*'=0.5 and *q*'=0.5. The Confidence Level (CL), is: 1 – = 0.98

The value is = 0.02. The *z*-score is:

The acceptable error is *e* = 0.03.

$n =$ $=$ $= 1501.6$

Thus, we get the following:

If we round up the answer of 1501.6 to the next largest whole number, we get *n* = 1502 people.

This means that if we sample 1502 people and construct a 98% confidence interval, the margin of error will be at most 0.03.

The value = 2.325 was found by going to the table of probabilities on Appendix A on page 607.

Since the table only has *z*-scores that are positive, we subtracted 0.01 from 0.5: 05 – 0.01 = 0.49

Thus, we need to find the *z*-score so that *P* (0<*Z*<)= 0.49.

However, the probability value 0.49 is not in the table. The value of 0.49 is between table values of 0.4898 and 0.4901.

Thus, is between the corresponding z values 2.32 and 2.33. We computed the average of 2.32 and 2.33:

Also, always round your final answer for sample size to the next largest whole number if the answer is not an integer. For example, if the value of *n* is 125.1, then round the value to the next largest whole number, 126.

4. Summary

We first covered the Central Limit Theorem and its applications. The Central Limit Theorem is among the most important in statistics. A reason why is that the Central Limit Theorem applies to any continuous random variable, provided that the sample size is large enough.

We then covered confidence intervals. A confidence interval uses probability to give us a range (interval) of possible values of the location of the parameter. We also covered the sample size formulas. Sample size formulas can give us an estimate of how many samples we need to collect.

Confidence intervals form one of the two major branches of inferential statistics. In the next module, we will cover the second major branch of inferential statistics, hypothesis testing.

Here is the list of the objectives that we have covered and are part of the Mastery Exercises in Knewton Alta:

- Use the Central Limit Theorem for Means to find the sample mean and the sample standard deviation in business examples
- Use the Central Limit Theorem for Sums to find the sample mean and sample standard deviation
- Use both forms of the Central Limit Theorem to compute probability
- Determine the z-score for a stated confidence level and compute the error bound in business applications
- Calculate and interpret the confidence interval for a population mean with a known standard deviation in business examples
- Find the sample size required to estimate a population mean with a given confidence level in business applications
- Determine the degrees of freedom to find and interpret the t-score of a normally distributed random variable in business applications
- Use the Student's t-distribution to calculate the confidence interval for a population mean with an unknown standard deviation in business applications
- Find the confidence interval, given a population proportion in business examples
- Calculate the sample size required to estimate a population proportion with a given confidence level in business applications
- Work backwards to calculate the error bound and sample mean, given the confidence interval in business contexts

Check Your Understanding

Embedded Media Content! Please use a browser to view this content.

References

None