Excel **MASTER** Series

Confidence Intervals in Excel

The Complete Guide

Excel Statistical Master

Clear and Simple
yet THOROUGH
Statistical Instruction for the
Graduate Student and
Business Manager
with
LOTS of Worked-Out Problems
and Screen Shots

Mark Harmon MBA

Confidence Intervals

in Excel

The Excel Statistical Master

By Mark Harmon

mark@ExcelMasterSeries.com

www.ExcelMasterSeries.com

ISBN: 978-0-9833070-1-3

Table of Contents

(Click On Chapters To Go To Them)

Confidence Intervals in Excel

Confidence Intervals are estimates of a population's average or proportion based upon sample data drawn from the population. A Confidence Interval is a range of values in which the mean is likely to fall with a specified level of confidence or certainty.

Contents

Basic Explanation of Confidence Intervals

The Confidence Interval is an interval in which the true population mean or proportion probably lies based upon a much smaller random sample taken from that population.

Confidence Intervals for means are calculated differently than Confidence Intervals for proportions. The first half of this course module will discuss calculating a Confidence Interval for a population mean. The second half will cover calculating a Confidence Interval for a population proportion.

First we will briefly discuss the difference between sampling for mean and sampling for proportion:

Mean Sampling vs. Proportion Sampling

What determines whether a mean is being estimated or a proportion is being estimated is the number of possible outcomes of each sample taken.

Proportion samples have only two possible outcomes.
For example, if you are comparing the proportion of Republicans in two different cities, each sample has only two possible values; the person sampled either is a Republican or is not.

Mean samples have multiple possible outcomes.
For example, if you are comparing the mean age of people in two different cities, each sample can have numerous values; the person sampled could be anywhere from 1 to 110 years old.

Below is a description of how to calculate a Confidence Interval for a population's mean. Note that everything is almost the same as the calculation of the Confidence Interval for a population proportion, except sample standard error.

Confidence Interval of a Population Mean

The Confidence Interval of a Mean is an interval in which the true population mean probably lies based upon a much smaller random sample taken from that population.

A 95% Confidence Interval of a Mean is the interval that has a 95% chance of containing the true population mean.

The width of a Confidence Interval is affected by the sample size. The larger the sample size, the more accurate and tighter is the estimate of the true population mean. The larger the sample size, the smaller will be the Confidence Interval. Samples taken must be random and also be representative of the population.

Calculate Confidence Intervals Using Large Samples (n>30)

Confidence Intervals are usually calculated and plotted on a Normal curve. If the sample size is less than 30, the population must be known to be Normally distributed. If small-sample data (n<30) is used to plot the Confidence Interval of the Mean for a population that is not Normally distributed, the result can be totally wrong.

Probably the most common major mistake in statistics is to apply Normal or t-distribution tests to small-sample data taken from a population of unknown distribution. Typically the actual distribution of a population is not known.

If the population's underlying distribution is not known (usually it is not), then only large samples (n>30) are valid for creating a Confidence Interval of the Mean. The most important theorem of statistics, the Central Limit Theorem, explains the reason for this.

The Central Limit Theorem

The Central Limit Theorem is statistics' most fundamental theorem. In a nutshell, it states the following: Random sample data can be plotted on a Normal curve to estimate a population's mean no matter how the population is distributed, as long as sample size is large (n>30).

The above definition of the Central Limit Theorem is the most practical and easy-to-understand. The following definition of this theorem is a bit more technical and will satisfy statisticians (but basically says the same thing as the above): No matter how the population is distributed, the sampling distribution of the mean approaches the Normal curve as sample size becomes large.

Levels of Confidence and Significance

Level of Significance, α ("alpha"), equals the maximum allowed percent of error. If the maximum allowed error is 5%, then α = 0.05.

Level of Confidence is the desired degree of certainty. A 95% Confidence Level is the most common. A 95% Confidence Level would correspond to a 95% Confidence Interval of the Mean. This would state that the actual population mean has a 95% probability of lying within the calculated interval. A 95% Confidence Level corresponds to a 5% Level of Significance, or α = 0.05. The Confidence Level therefore equals 1 - α.

Population Mean vs. Sample Mean

Population Mean = μ ("mu") (This is what we are trying to estimate)

Sample Mean = x_{avg}

Standard Deviation and Standard Error

Standard Deviation is a measure of statistical dispersion. It's formula is the following:

SQRT ([SUM (X - x_{avg})^2] / N).

There is no need to memorize the formula because you can plug in Excel's STDEV function discussed later. Standard Deviation equals the square root of the Variance.

Population Standard Deviation = σ ("sigma")

Sample Standard Deviation = S

Standard Error is an estimate of population Standard Deviation from data taken from a sample. If the population Standard Deviation, σ, is known, then the Sample Standard Error, s_{xavg}, can be calculated. If only the Sample Standard Deviation, S, is known, then Sample Standard Error, S_{xavg}, can be estimated by substituting Sample Standard Deviation, S, for Population Standard Deviation, σ, as follows:

Sample Standard Error = S_{xavg} = σ / SQRT(n) ≈ S / SQRT(n)

σ = Population standard deviation

S = Sample standard deviation

n = sample size

Region of Certainty vs. Region of Uncertainty

Region of Certainty is the area under the Normal curve that corresponds to the required Level of Confidence. If a 95% percent Level of Confidence is required, then the Region of Certainty will contain 95% of the area under the Normal curve. **The outer boundaries of the Region of Certainty will be the outer boundaries of the Confidence Interval**.

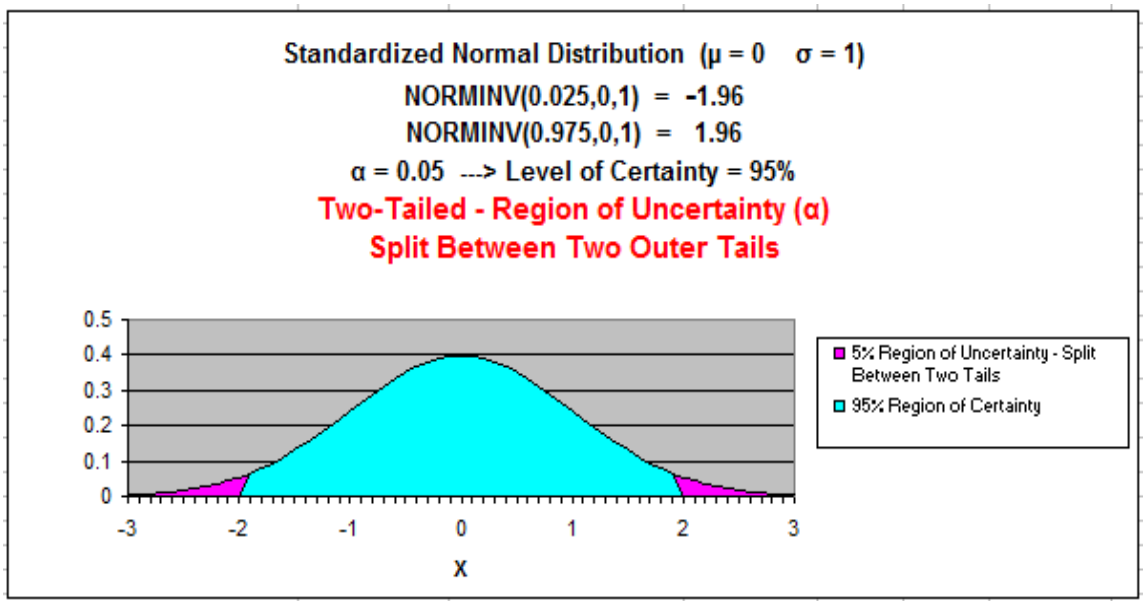
The Region of Certainty, and therefore the Confidence Interval, will be centered about the mean. Half of the Confidence Interval is on one side of the mean and half on the other side.

Region of Uncertainty is the area under the Normal curve that is outside of the Region of Certainty. Half of the Region of Uncertainty will exist in the right outer tail of the Normal curve and the other half in the left outer tail. This is similar to the concept of the "two-tailed test" that is used in Hypothesis testing in further sections of this course. The concepts of one- and two-tailed testing are not used when calculating Confidence Intervals. Just remember that the Region of Certainty, and therefore the Confidence Interval, are always centered about the mean on the Normal curve.

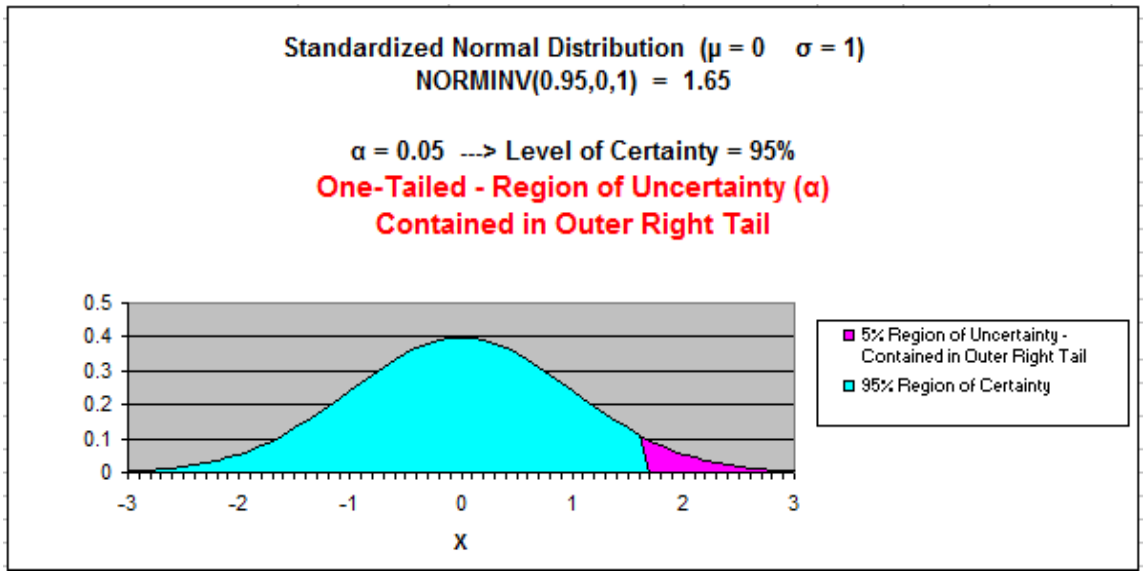
Relationship Between Region of Certainty, Uncertainty, and Alpha

The Region of Uncertainty corresponds to α ("alpha"). If α = 0.05, then that Region of Uncertainty contains 5% of the area under the Normal curve. Half of that area (2.5%) is in each outer tail. The 95% area centered about the mean will be the Region of Certainty. The outer boundaries of this Region of Certainty will be the outer boundaries of the 95% Confidence Interval. The Level of Confidence is 95% and the Level of Significance, or maximum error allowed, is 5%.

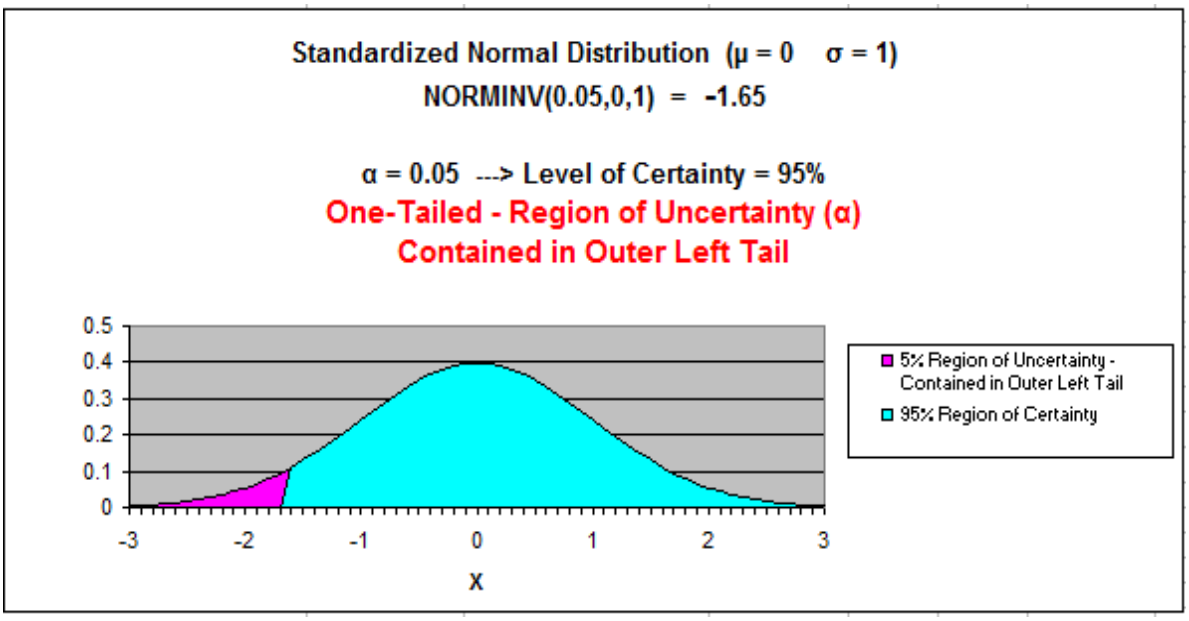
Illustrating a two-tail test – Similar to what is used for calculating Confidence Interval



Illustrating a one-tailed test – Right Tail



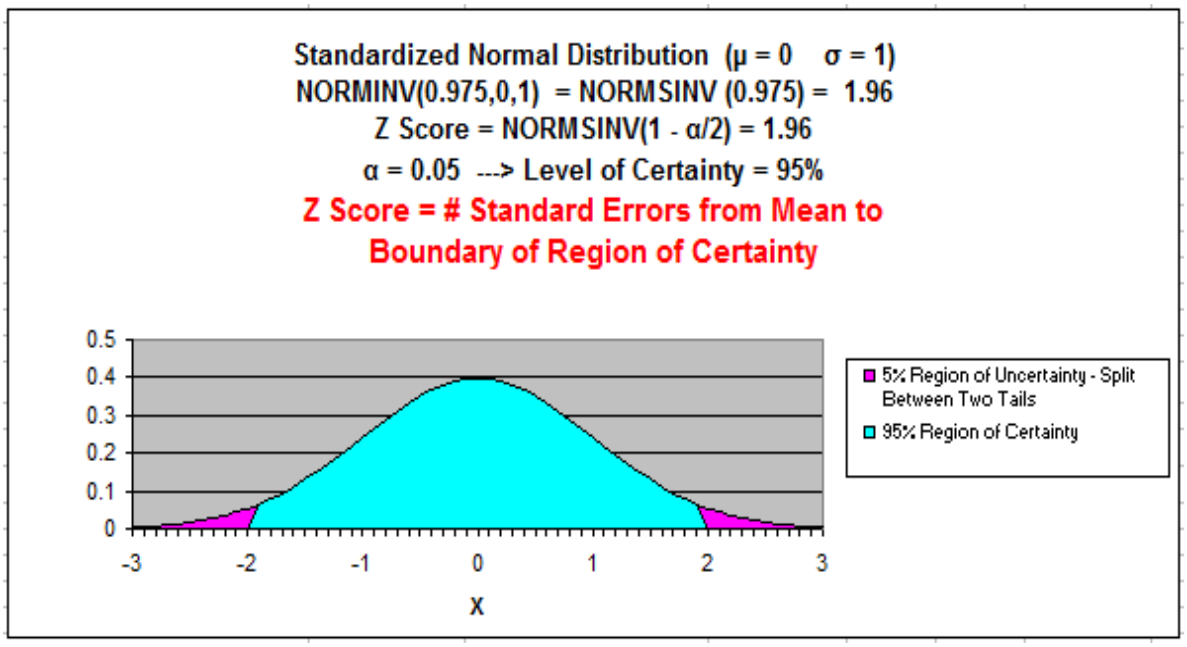
Illustrating a one-tailed test – Left Tail



Z Score

Z Score is the number of Standard Errors from the mean
to outer right boundary of the Region of Certainty (and
therefore to the outer right boundary of the Confidence Interval).
Standard Errors are used and not Standard Deviations
because sample data is being used to calculate the
Confidence Interval.

Z Score is calculated by the following Excel function:
$Z\ Score_{(1-\alpha)}$ = NORMSINV (1 - α/2) - This will be discussed below.



It is very important to note that on a Standardized Normal Curve, the Distance from the mean to boundary of the Region of Certainty equals the number of standard errors from the mean to boundary, which is the Z Score.

The above is only true for a Standardized Normal Curve. It is NOT true for a Non-Standardized Normal curve.

Excel Functions Used When Calculating Confidence Interval of Mean

COUNT (Highlighted block of cells) = Sample size = n
----> Counts number of cells in highlighted block

STDEV (Highlighted block of cells) = Standard deviation
----> Calculates Standard Deviation of all cells in highlighted block

AVERAGE (Highlighted block of cells) = Mean
----> Calculates the mean of all cells in highlighted block

NORMSINV (1 - α/2) = **Z Score$_{(1 - \alpha)}$**
= Number of Standard errors from mean to boundary of Confidence Interval. Note that (1 - α/2) = the entire area in the Normal curve to the left of outer right boundary of the Region of Certainty, or Confidence Interval. This includes the entire Region of Certainty and the half of the Region of Uncertainty that exists in the left tail.

For example:

Level of Confidence = 95% for a 95% Confidence Interval
Level of Significance = 5% (α = 0.05)
1 - α = 0.95 = 95%

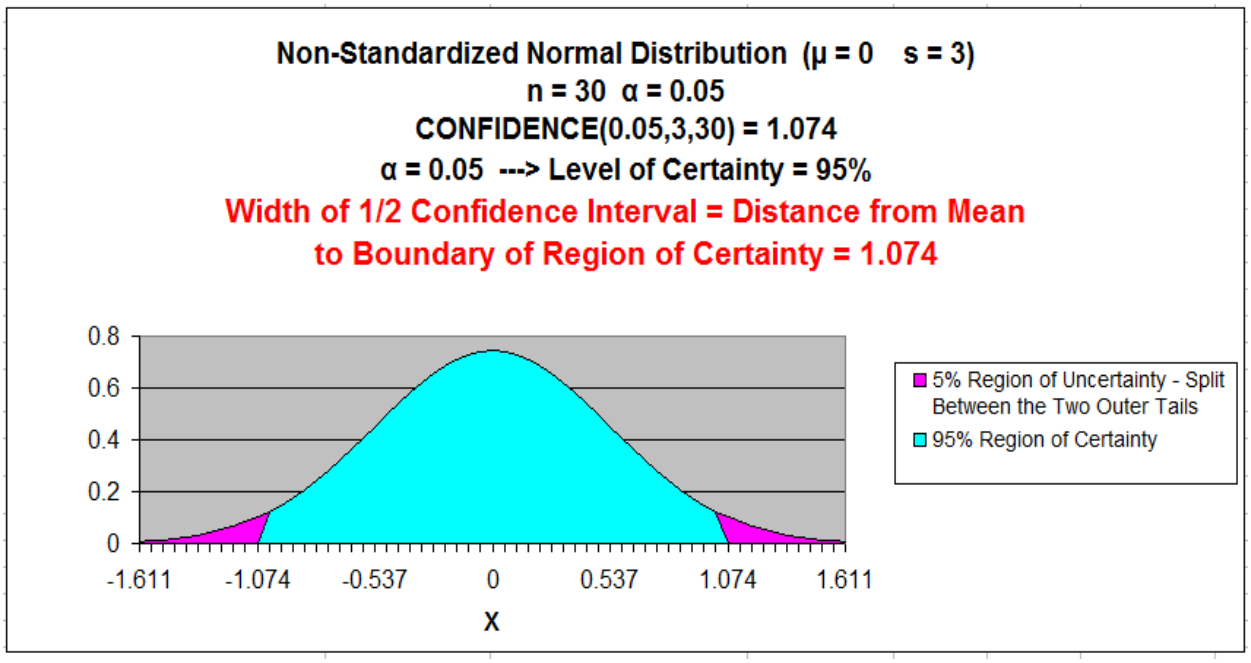
Z Score$_{95\%}$ = NORMSINV(1 – α/2) = NORMSINV (1 - .05/2) = NORMSINV(1 - 0.025)

Z Score$_{95\%}$ = NORMSINV (0.975) = 1.96

The outer right boundary of the 95% Confidence Interval, and the Region of Certainty, is 1.96 Standard Errors from the mean. The left boundary is the same distance from the mean because the Confidence Interval is centered about the mean.

CONFIDENCE (α, s, n) = **Width of half of the Confidence Interval**

α = Level of Significance
s = Sample Standard Deviation - Note that this is not Standard Error.
s is calculated by applying STDEV to the sample values.
n = Sample size - Apply COUNT to sample values.



Formulas for Calculating Confidence Interval Boundaries from Sample Data

Confidence Interval Boundaries = Sample mean +/- Z Score$_{(1-\alpha)}$ * Sample Standard Error

Confidence Interval Boundaries = x_{avg} +/- Z Score$_{(1-\alpha)}$ * S_{xavg}

Sample Mean = x_{avg} = AVERAGE (Highlighted block of cell containing samples)

Z Score$_{(1 - \alpha)}$ = NORMSINV (1 - α/2)

Sample Standard Error = S_{xavg} = σ / SQRT(n) ≈ s / SQRT(n)

Sample size = n = COUNT (Highlighted block of cells containing samples)

Sample Standard Deviation = s = STDEV (Highlighted block of cells containing samples)

CONFIDENCE (α, s, n) = **Width of half of the Confidence Interval**

CONFIDENCE (α, s, n) = Z Score$_{(1-\alpha)}$ * S_{xavg}

So:

Confidence Interval Boundaries = x_{avg} +/- Z Score$_{(1-\alpha)}$ * S_{xavg}

Confidence Interval Boundaries = x_{avg} +/- CONFIDENCE (α, s, n)

Problem 1: Calculate a Confidence Interval from a Random Sample of Test Scores

Problem: Given the following set of 32 random test scores taken from a much larger population, calculate with 95% certainty an interval in which the population mean test score must fall. In other words, calculate the 95% Confidence Interval for the population test score mean. The random sample of 32 tests scores is shown on the next page.

32 Random Test Scores Samples from a Much Larger Population

220	300
370	410
500	540
640	660
220	300
370	410
500	540
640	660
220	300
370	410
500	540
640	660
220	300
370	410
500	540
640	660

Level of Confidence = 95% = 1 - α

Level of Significance = α = 0.05

Sample Size = n = COUNT (Yellow Highlighted block of cells) = 32

Sample Mean = X_{avg} = AVERAGE (Yellow Highlighted block of cells) = 455

Sample Standard Deviation = S = STDEV (Yellow Highlighted block of cells) = 149.8

Sample Standard Error = $S_{xavg} = \sigma / SQRT(n) \approx S / SQRT(n)$

$S_{xavg} = \sigma / SQRT(n) \approx 149.8 / SQRT(32) = 26.5$

Z Score$_{(1-\alpha)}$ = Z Score$_{95\%}$ = NORMSINV (1 - α/2)

= NORMSINV (1 - 0.025) = NORMSINV (0.975) = 1.96

Width of Half the Confidence Interval = CONFIDENCE (α, S, n)

= CONFIDENCE (0.05, 149.8, 32) = 51.9

Also, equivalently:

Width of Half the Confidence Interval = Z Score$_{(1-\alpha)}$ * S_{xavg}

= 1.96 * 26.5 = 51.9

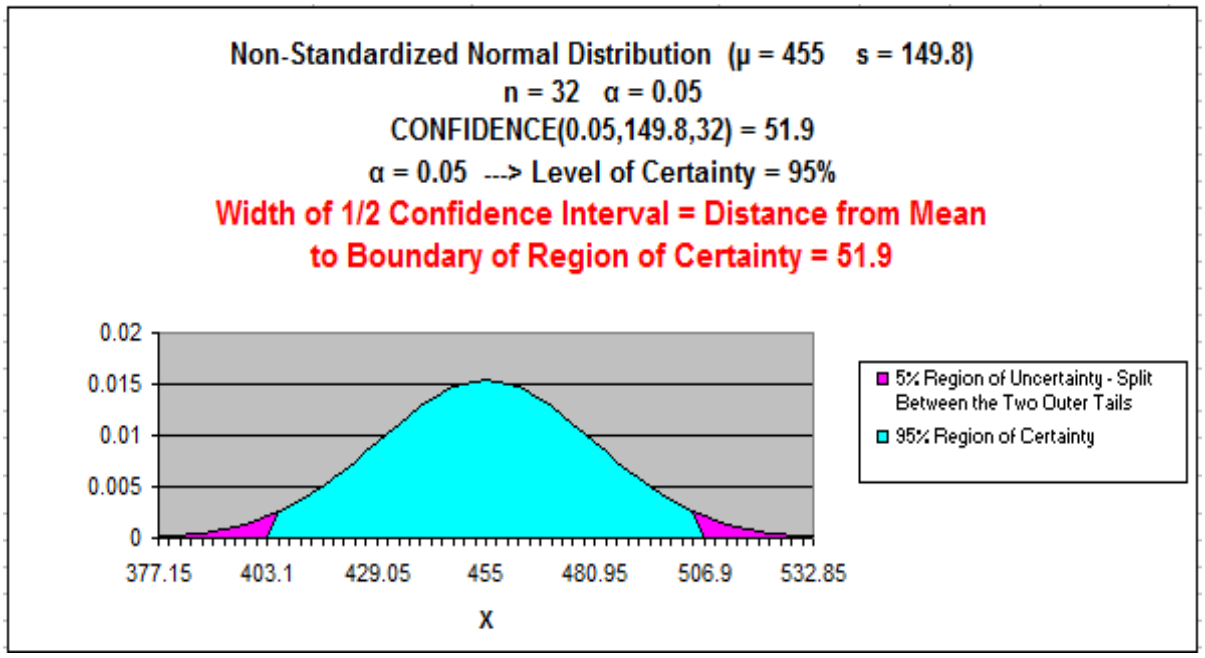
Confidence Interval Boundaries = X_{avg} +/- Z Score$_{(1-\alpha)}$ * S_{xavg}

= 455 +/- (1.96)*(26.5) = 455 +/- 51.9 = 403.1 to 506.9

Also, equivalently:

Confidence Interval Boundaries = X_{avg} +/- CONFIDENCE (α, S, n)

= 455 +/- 51.9 = 403.1 to 506.9


Non-Standardized Normal Distribution (μ = 455 s = 149.8)
n = 32 α = 0.05
CONFIDENCE(0.05,149.8,32) = 51.9
α = 0.05 ---> Level of Certainty = 95%
Width of 1/2 Confidence Interval = Distance from Mean
to Boundary of Region of Certainty = 51.9
0.02
0.015
0.01
0.005
0
377.15
403.1
429.05
455
480.95
506.9
532.85
X
5% Region of Uncertainty - Split Between the Two Outer Tails
95% Region of Certainty

Problem 2: Calculate a Confidence Interval of Daily Sales Based Upon Sample Mean and Standard Deviation

Problem: Average daily demand for books sold in a small Barnes and Noble store is 455 books with a standard deviation of 200. This average and standard deviation are taken from sale data collected every day for a period of 60 days. What is the range that the true average daily book sales lies in with 95% certainty?

Level of Confidence = 95% = 1 - α

Level of Significance = α = 0.05

Sample Size = n = 60

Sample Mean = x_{avg} = 455

Sample Standard Deviation = s = 200

Sample Standard Error = $s_{x_{avg}}$ = σ / SQRT(n) ≈ s / SQRT(n)

$s_{x_{avg}}$ = σ / SQRT(n) ≈ 200 / SQRT(60) = 25.8

Z Score$_{(1 - \alpha)}$ = Z Score$_{95\%}$ = NORMSINV (1 - α/2)

= NORMSINV (1 - 0.025) = NORMSINV (0.975) = 1.96

Width of Half the Confidence Interval = CONFIDENCE (α, s, n)

= CONFIDENCE (0.05, 200, 60) = 50.6

Also, equivalently:

Width of Half the Confidence Interval = Z Score$_{(1-\alpha)}$ * $s_{x_{avg}}$

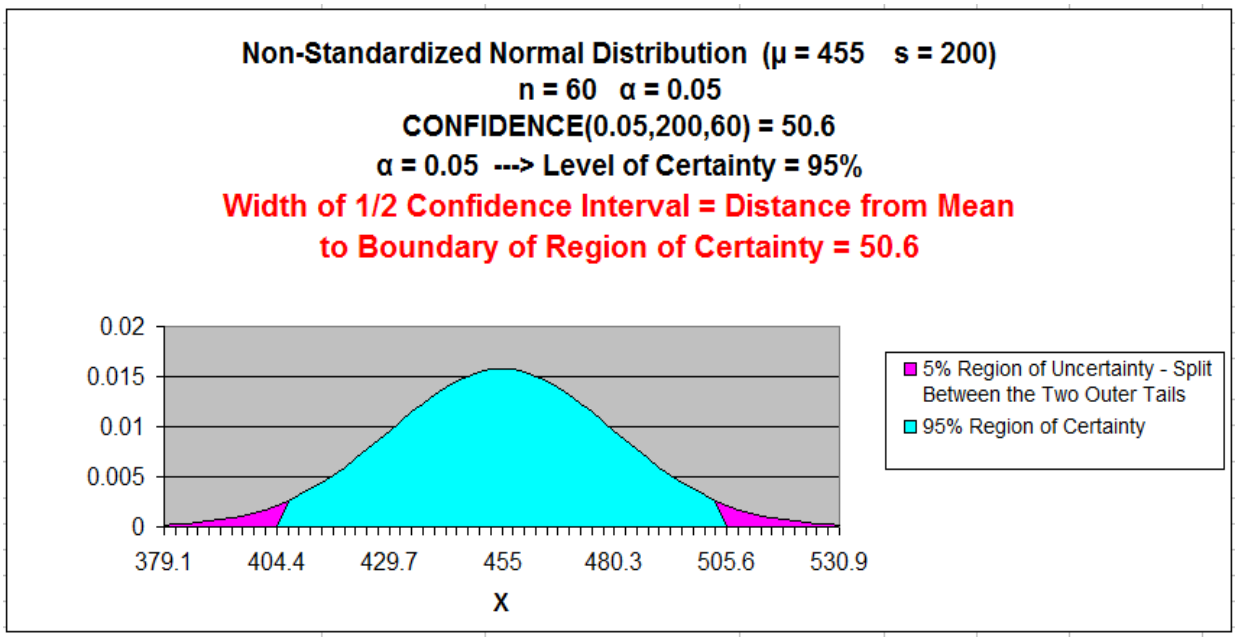
= 1.96 * 25.8 = 50.6

Confidence Interval Boundaries = x_{avg} +/- Z Score$_{(1-\alpha)}$ * $s_{x_{avg}}$

= 455 +/- (1.96)*(25.8) = 455 +/- 50.6 = 404.4 to 505.6

Confidence Interval Boundaries = x_{avg} +/- CONFIDENCE (α, s, n)

= 455 +/- 50.6 = 404.4 to 505.6



Problem 3: Calculate an Exact Range of 95% of Sales Based Upon the Upon the Population Mean and Standard Deviation

Problem: Average daily demand for books sold in a large Barnes and Noble store is 5,000 books with a standard deviation of 200. This average and standard deviation are taken from sale data collected every day for a period of 5 years. What is the range that 95% of the daily unit book sales fall in? The daily sales data is Normally distributed.

This problem is not a Confidence Interval problem. We do not need to create an estimate of the population mean (a Confidence Interval) because we know exactly what it is. We are given the population mean and population standard deviation.

We do need to know how the population is distributed in order to calculate the interval that contains 95% of all population data. Given that the population is Normally distributed, we simply need to map the region of this Normal curve that contains 95% of the total area and is centered about the mean as follows:

Population mean = $\mu = 5{,}000$

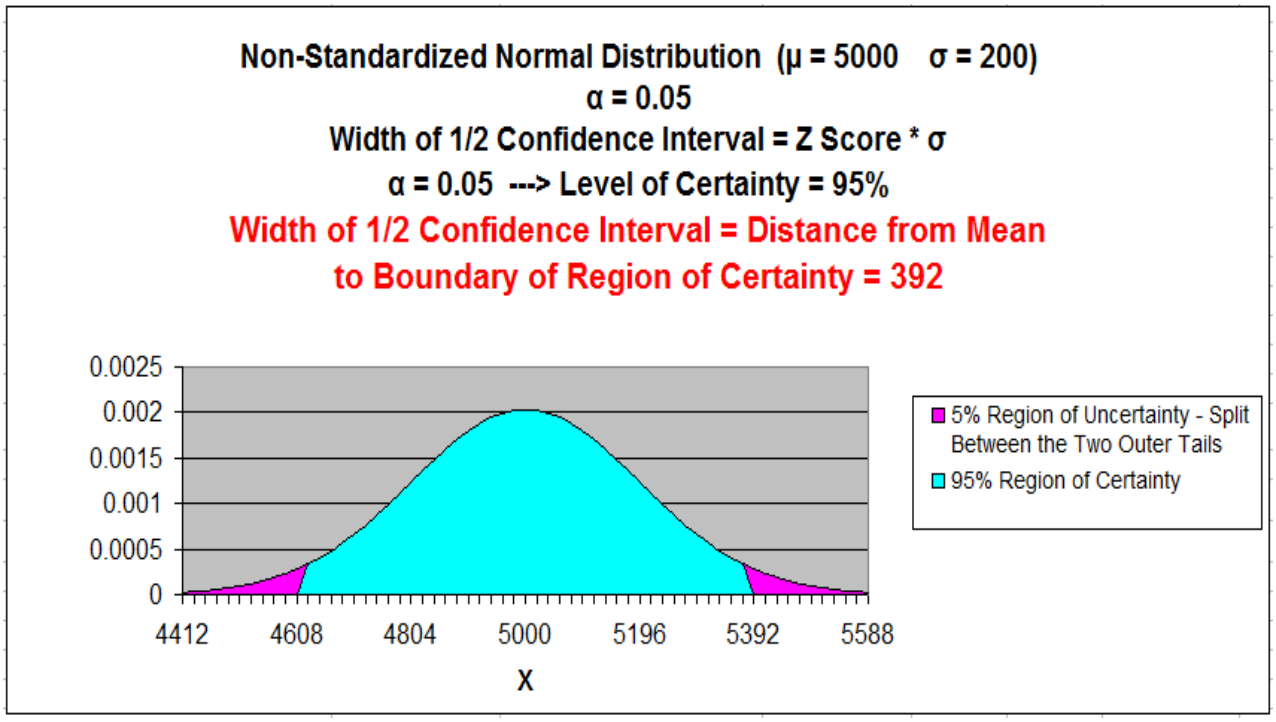
Population Standard Deviation = $\sigma = 200$

Range Containing 95% of Sales Data = μ +/- [$Z\ Score_{95\%}$ * σ]

= 5,000 +/- [NORMSINV(0.975) * 200]

= 5,000 +/- 392

= 4,608 to 5,392


Non-Standardized Normal Distribution (μ = 5000 σ = 200)
α = 0.05
Width of 1/2 Confidence Interval = Z Score * σ
α = 0.05 ---> Level of Certainty = 95%
Width of 1/2 Confidence Interval = Distance from Mean
to Boundary of Region of Certainty = 392
0.0025
0.002
0.0015
0.001
0.0005
0
4412
4608
4804
5000
5196
5392
5588
X
5% Region of Uncertainty - Split Between the Two Outer Tails
95% Region of Certainty

Determining Minimum Sample Size (n) to Keep Confidence Interval of the Mean within a Certain Tolerance

The larger the sample size, the more accurate and tighter will be the prediction of a population's mean. Stated another way, the larger the sample size, the smaller will be the Confidence Interval of the population's mean. Width of the Confidence Interval is reduced when sample size is increased.

Quite often a population's mean needs to be estimated with some level of certainty to within plus or minus a specified tolerance. This specified tolerance is half the width of the Confidence Interval. Sample size directly affects the width of the Confidence Interval. The relationship between sample size and width of the Confidence Interval is shown as follows:

CONFIDENCE (α, s, n) = **Width of half of the Confidence Interval**

CONFIDENCE (α, s, n) = $Z\ Score_{(1-\alpha)} * S_{xavg}$

Width of Half of the Confidence Interval = $Z\ Score_{(1-\alpha)} * S_{xavg}$

Width of Half of the Confidence Interval = $Z\ Score_{(1-\alpha)} * \sigma\ /\ SQRT(n)$

$\approx Z\ Score_{(1-\alpha)} * s\ /\ SQRT(n)$

With algebraic manipulation of the above we have:

$SQRT(n) = Z\ Score_{(1-\alpha)} * \sigma\ /$ Width of Half of the Confidence Interval

$n = [\ Z\ Score_{(1-\alpha)}\ ]^2 * [\sigma]^2\ /\ [$ Width of Half of the Confidence Interval $]^2$

Also, if we only have sample standard deviation, s, and not population standard deviation, σ :

$n \approx [\ Z\ Score_{(1-\alpha)}\ ]^2 * [s]^2\ /\ [$ Width of Half of the Confidence Interval $]^2$

Problem 4: Determine the Minimum Number of Sales Territories to Sample In Order To Limit the 95% Confidence Interval to a Certain Width

Problem: A national sales manager in charge of 5,000 similar territories ran a nationwide promotion. He then collected sales data from a random sample of the territories to evaluate sales increase. From the sample, the average sales increase per territory was $10,000 with a standard deviation of $500. How many territories would he have had to have sampled to be 95% sure that the actual nationwide average territory sales increase was no more than $50 different than average territory sales increase from the sample he took?

Level of Confidence = 95% = 1 - α

Level of Significance = α = 0.05

Sample Size = n = ?

Sample Mean = x_{avg} ------> Note this does not need to be known to solve this problem

Sample Standard Deviation = s = 500

Z Score$_{(1 - \alpha)}$ = Z Score$_{95\%}$ = NORMSINV (1 - α/2)

= NORMSINV (1 - 0.025) = NORMSINV (0.975) = 1.96

Width of Half the Confidence Interval = 50

$$n \approx [\text{Z Score}_{(1-\alpha)}]^2 * [s]^2 / [\text{Width of Half of the Confidence Interval}]^2$$

$$n \approx [1.96]^2 * [500]^2 / [50]^2 = 384$$

The sales manager would have to sample at least 384 territories to be 95% certain that nationwide territory average was within +/- $50 of the sample territory average. Note that the 95% confidence interval is $10,000 +/- $50 and this interval has a width = $100 if sample size is 384.

****

Confidence Interval of a Population Proportion

Creating a Confidence Interval for a population's proportion is very similar to creating a Confidence Interval for a population's mean. The only real difference is how the standard error is calculated. Everything else is the same. The method of calculating a Confidence Interval for a population mean was covered in detail earlier in this module. First, the difference between using sampling to estimate a population mean and using sampling to estimate a population proportion will be explained below:

Mean Sampling vs. Proportion Sampling

What determines whether a mean is being estimated or a proportion is being estimated is the number of possible outcomes of each sample taken.

Proportion samples have only two possible outcomes. For example, if you are comparing the proportion of Republicans in two different cities, each sample has only two possible values; the person sampled either is a Republican or is not.

Mean samples have multiple possible outcomes. For example, if you are comparing the mean age of people in two different cities, each sample can have numerous values; the person sampled could be anywhere from 1 to 110 years old.

Below is a description of how to calculate a Confidence Interval for a population's proportion. Note that everything is almost the same as the calculation of the Confidence Interval for a mean, except sample standard error.

Levels of Confidence and Significance

Level of Significance, α ("alpha"), equals the maximum allowed percent of error. If the maximum allowed error is 5%, then α = 0.05.

Level of Confidence is selected by the user. A 95% Level is the most common. A 95% Confidence Level would correspond to a 95% Confidence Interval of the Proportion. This would state that the actual population Proportion has a 95% probability of lying within the calculated interval. A 95% Confidence Level corresponds to a 5% Level of Significance, or α = 0.05. The Confidence Level therefore equals 1 - α.

Population Proportion vs. Sample Proportion

Population Proportion = $\mu_p = p$ (This is what we are trying to estimate)

Sample Proportion = p_{avg}

Standard Deviation and Standard Error

Standard Deviation is not calculated during the creation of Confidence Interval for a population proportion.

Standard Error is an estimate of population Standard Deviation from data taken from a sample. Sample Standard Error will be an estimate taken from the sample proportion, p_{avg}, and sample size, n. This is the major difference between calculating a Confidence Interval for a proportion and for a mean. Binomial distribution rules apply to proportions because a proportion sample has only two possible outcomes, just like a binomial variable.

Sample Standard Error of a Proportion = $\sigma_{pavg} = \text{SQRT}(p * q / n) \approx s_{pavg}$

Estimated Sample Standard Error of a Proportion = $s_{pavg} = \text{SQRT}(p_{avg} * q_{avg} / n)$

p = Population proportion - This is the unknown that will be estimated with a Confidence Interval
q = 1 - p
n = sample size
p_{avg} = Sample proportion
$q_{avg} = 1 - p_{avg}$

Region of Certainty vs. Region of Uncertainty

Region of Certainty is the area under the Normal curve that corresponds to the required Level of Confidence. If a 95% percent Level of Confidence is required, then the Region of Certainty will contain 95% of the area under the Normal curve. **The outer boundaries of the Region of Certainty will be the outer boundaries of the Confidence Interval**.

The Region of Certainty, and therefore the Confidence Interval, will be centered about the mean. Half of the Confidence Interval is on one side of the mean and half on the other side.

Region of Uncertainty is the area under the Normal curve that is outside of the Region of Certainty. Half of the Region of Uncertainty will exist in the right outer tail of the Normal curve and the other half in the left outer tail. This is similar to the concept of the "two-tailed test" that is used in Hypothesis testing in further sections of this course. The concepts of one and two-tailed testing are not used when calculating Confidence Intervals. Just remember that the Region of Certainty, and therefore the Confidence Interval, are always centered about the mean on the Normal curve.

Relationship Between Region of Certainty, Uncertainty, and Alpha

The Region of Uncertainty corresponds to α ("alpha"). If α = 0.05, then that Region of Uncertainty contains 5% of the area under the Normal curve. Half of that area (2.5%) is in each outer tail. The 95% area centered about the mean will be the Region of Certainty. The outer boundaries of this Region of Certainty will be the outer boundaries of the 95% Confidence Interval. The Level of Confidence is 95% and the Level of Significance, or maximum error allowed, is 5%.

Z Score

Z Score is the number of Standard Errors from the mean to outer right boundary of the Region of Certainty (and therefore to the outer right boundary of the Confidence Interval). Standard Errors are used and not Standard Deviations because sample data is being used to calculate the Confidence Interval.

Z Score is calculated by the following Excel function:

$Z\ Score_{(1-\alpha)}$ = NORMSINV (1 - α/2) - This will be discussed shortly.

Excel Functions Used When Calculating Confidence Interval for a Population Proportion

Note that Excel functions STDEV and AVERAGE are not used when working with proportions. The CONFIDENCE function is not used either.

COUNT (Highlighted block of cells) = Sample size = n
----> Counts number of cells in highlighted block

NORMSINV (1 - α/2) = **Z Score$_{(1 - \alpha)}$**
= Number of Standard Errors from mean to boundary of Confidence Interval. Note that (1 - α/2) = the entire area in the Normal curve to the left of outer right boundary of the Region of Certainty, or Confidence Interval. This includes the entire Region of Certainty and the half of the Region of Uncertainty that exists in the left tail.

For example:

Level of Confidence = 95% for a 95% Confidence Interval
Level of Significance = 5% (α = 0.05)
1 - α = 0.95 = 95%

Z Score$_{95\%}$ = NORMSINV (1 - α/2) = NORMSINV (1 - .05/2) = NORMSINV(1 - 0.025)
Z Score$_{95\%}$ = NORMSINV (0.975) = 1.96

The outer right boundary of the 95% Confidence Interval, and the Region of Certainty, is 1.96 Standard Errors from the mean. The left boundary is the same distance from the mean because the Confidence Interval is centered about the mean.

Formula for Calculating Confidence Interval Boundaries from Sample Data for a Population Proportion

Confidence Interval Boundaries = Sample proportion +/- Z Score$_{(1-\alpha)}$ * Sample Standard Error

Confidence Interval Boundaries = p_{avg} +/- Z Score$_{(1-\alpha)}$ * s_{pavg}

Sample Proportion = p_{avg}

Z Score$_{(1 - \alpha)}$ = NORMSINV (1 - α/2)

Sample Standard Error of a Proportion = $\sigma_{pavg} \approx s_{pavg}$ = SQRT ($p_{avg} * q_{avg} / n$)

Sample size = n = COUNT (Highlighted block of cells containing samples)

Confidence Interval Boundaries = p_{avg} +/- Z Score$_{(1-\alpha)}$ * s_{pavg}

Problem 5: Determine Confidence Interval of Shoppers Who Prefer to Pay By Credit Card Based Upon Sample Data

Problem: A random sample of 1,000 shoppers was taken. 70% preferred to pay with a credit card. 30% preferred to pay with cash. Determine the 95% Confidence Interval for the proportion of the general population that prefers to pay with a credit card.

Level of Confidence = 95% = 1 - α

Level of Significance = α = 0.05

Sample Size = n = 1,000

Sample Proportion = p_{avg} = 0.70

$q_{avg} = 1 - p_{avg} = 0.30$

Sample Standard Error of a Proportion = $\sigma_{pavg} \approx s_{pavg}$ = SQRT ($p_{avg} * q_{avg} / n$)

s_{pavg} = SQRT (0.70 * 0.30 / 1,000) = 0.014

Z Score$_{(1 - \alpha)}$ = Z Score$_{95\%}$ = NORMSINV (1 - α/2)

= NORMSINV (1 - 0.025) = NORMSINV (0.975) = 1.96

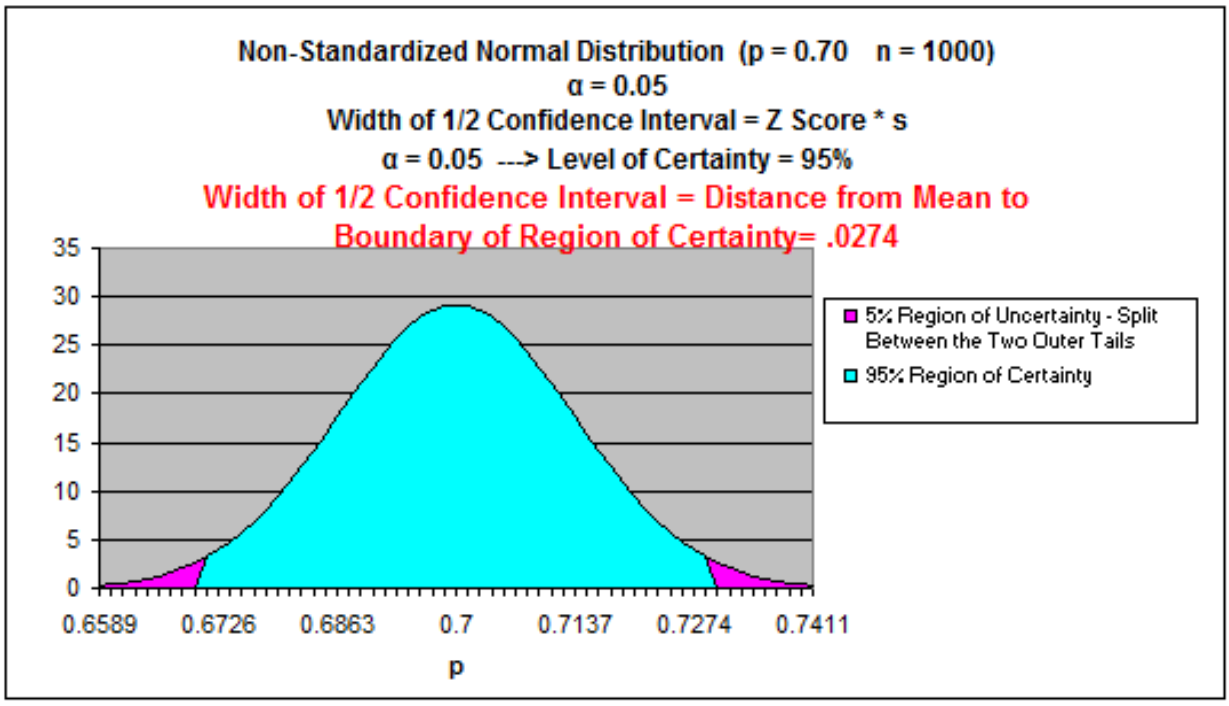
Width of Half the Confidence Interval = Z Score$_{(1-\alpha)}$ * s_{pavg}

= 1.96 * 0.014 = 0.0274

Confidence Interval Boundaries = p_{avg} +/- Z Score$_{(1-\alpha)}$ * s_{pavg}

= 0.70 +/- (1.96) * (0.014)

= 0.70 +/- 0.0274 = 0.6726 to 0.7274 = 67.26% to 72.74%


Non-Standardized Normal Distribution (p = 0.70 n = 1000)
α = 0.05
Width of 1/2 Confidence Interval = Z Score * s
α = 0.05 ---> Level of Certainty = 95%
Width of 1/2 Confidence Interval = Distance from Mean to Boundary of Region of Certainty= .0274
35
30
25
20
15
10
5
0
0.6589
0.6726
0.6863
0.7
0.7137
0.7274
0.7411
p
5% Region of Uncertainty - Split Between the Two Outer Tails
95% Region of Certainty

Determining Minimum Sample Size (n) to Keep Confidence Interval of the Proportion within a Certain Tolerance

The larger the sample size, the more accurate and tighter will be the prediction of a population's mean. Stated another way, the larger the sample size, the smaller will be the Confidence Interval of the population's mean. Width of the Confidence Interval is reduced when sample size is increased.

Quite often a population's mean needs to be estimated with some level of certainty to within plus or minus a specified tolerance. This specified tolerance is half the width of the confidence interval. Sample size directly affects the width of the Confidence Interval. The relationship between sample size and width of the Confidence Interval is shown as follows:

Width of Half the Confidence Interval = $Z\ Score_{(1-\alpha)} * S_{pavg}$

$S_{pavg} = SQRT\ (\ p_{avg} * q_{avg} / n\ )$

Width of Half the Confidence Interval = $Z\ Score_{(1-\alpha)} * SQRT\ (\ p_{avg} * q_{avg} / n\ )$

[**Width of Half the Confidence Interval**]^2 = $[\ Z\ Score_{(1-\alpha)}\ ]^2 * (\ p_{avg} * q_{avg} / n\ )$

$n = [\ Z\ Score_{(1-\alpha)}\ ]^2 * (\ p_{avg} * q_{avg}\ ) / [$ **Width of Half the Confidence Interval** $]^2$

Problem 6: Determine the Minimum Sample Size of Voters to be 95% Certain that the Population Proportion is no more than 1% Different from Sample Proportion

Problem: A random survey was conducted in one city to learn voting preferences. 40% of voters surveyed said they would vote Republican. 60% of the voters surveyed said they would vote Democrat. Determine the minimum number of voters that had to be surveyed to be 95% certain that the results were accurate within +/- 1%.

Level of Confidence = 95% = 1 - α

Level of Significance = α = 0.05

p_{avg} = 0.40

$q_{avg} = 1 - p_{avg}$ = 0.60

Width of Half the Confidence Interval = 0.01 ---> (1%)

Z Score$_{(1 - \alpha)}$ = Z Score$_{95\%}$ = NORMSINV (1 - α/2)

= NORMSINV (1 - 0.025) = NORMSINV (0.975) = 1.96

$$n = [\text{Z Score}_{(1-\alpha)}]^2 * (p_{avg} * q_{avg}) / [\text{Width of Half the Confidence Interval}]^2$$

$$n = [1.96]^2 * (0.40 * 0.60) / [0.01]^2 = 9{,}220$$

At least 9,220 random voters had to be surveyed to be 95% certain that the population proportion is no more than 1% different from the sample.

Meet the Author

Mark Harmon is a master number cruncher. Creating overloaded Excel spreadsheets loaded with complicated statistical analysis is his idea of a good time. His profession as an Internet marketing manager provides him with the opportunity and the need to perform plenty of meaningful statistical analysis at his job.

Mark Harmon is also a natural teacher. As an adjunct professor, he spent five years teaching more than thirty semester-long courses in marketing and finance at the Anglo-American College in Prague and the International University in Vienna, Austria. During that five-year time period, he also worked as an independent marketing consultant in the Czech Republic and performed long-term assignments for more than one hundred clients. His years of teaching and consulting have honed his ability to present difficult subject matter in an easy-to-understand way.

Harmon received a degree in electrical engineering from Villanova University and MBA in marketing from the Wharton School.