

Confidence Intervals

Tom Bruning

2020-09-10

In this chapter, you learn:

- To construct and interpret confidence interval estimates for the population mean and the population proportion
- To determine the sample size necessary to develop a confidence interval for the population mean or population proportion

Point and Interval Estimates

- A point estimate is a single number,
- A confidence interval provides additional information about the variability of the estimate

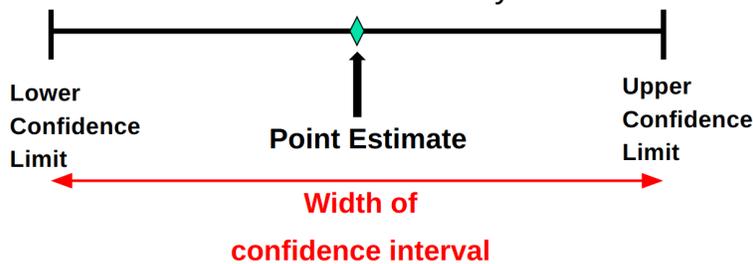


Figure 1: Confidence Interval

Point Estimate

We can estimate a population parameter: μ (population mean) or π (population proportion) with the corresponding sample statistic: \bar{x} (sample mean) or p (sample proportion).

Table 1: Point Estimates

Population Parameter	Sample Statistic	Point Estimate
μ	\bar{x}	
π	p	

Confidence Intervals

- How much uncertainty is associated with a point estimate of a population parameter?
- An *interval estimate* provides more information about a population characteristic than does a *point estimate*.
- Such interval estimates are called **confidence intervals**.

Confidence Intervals Estimates

An interval gives a range of values:

- Takes into consideration variation in sample statistics from sample to sample
- Based on observations from 1 sample
- Gives information about closeness to unknown population parameters
- Stated in terms of level of confidence
 - e.g. 95% confident, 99% confident
 - Can never be 100% confident

Confidence Interval Example

Cereal fill example

- Population has $\mu = 368$ and $\sigma = 15$.
- If you take a sample of size $n = 25$ you know (from Chapter 7)
 - $368 \pm 1.96 \times \frac{15}{\sqrt{25}} = (362.12, 373.88)$. 95% of the intervals formed in this manner will contain μ .
 - * When you don't know μ , you use \bar{X} to estimate μ .
 - If $X = 362.3$ the interval is $362.3 \pm 1.96 \times \frac{15}{\sqrt{25}} = (356.42, 368.18)$
 - Since $356.42 \leq \mu \leq 368.18$ the interval based on this sample makes a correct statement about μ .

Recall from Chapter 7: $\mu \pm Z \times \sigma_{\bar{x}}$
and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

But what about the intervals from other possible samples of size 25?

Sample #	\bar{X}	Lower Limit	Upper Limit	Contains μ ?
1	362.30	356.42	368.18	Yes
2	369.50	363.62	375.38	Yes
3	360.00	354.12	365.88	No
4	362.12	356.24	368.00	Yes
5	373.88	368.00	379.76	Yes

- In practice you only take one sample of size n .
- In practice you do not know μ so you do not know if the interval actually contains μ .
- However you do know that 95% of the intervals formed in this manner may contain μ .
- Thus, based on the one sample, you actually selected you can be 95% confident your interval may contain μ (this is a 95% confidence interval)

Note: 95% confidence is based on the fact that we used $Z = 1.96$.

General Formula

The general formula for all confidence intervals is:

Point Estimate \pm (Critical Value)(Standard Error), where:

- Point Estimate is the sample statistic estimating the population parameter of interest.
- Critical Value is a table value based on the sampling distribution of the point estimate and the desired confidence level.
- Standard Error is the standard deviation of the point estimate.

Confidence Level, $(1-\alpha)$

- Suppose confidence level = 95%.
- Also written $(1 - \alpha) = 0.95$, (so $\alpha = 0.05$).
- A relative frequency interpretation:
 - 95% of all the confidence intervals that can be constructed will contain the unknown true parameter.
- A specific interval either will contain or will not contain the true parameter.
 - No probability involved in a specific interval.

Confidence Interval for μ (σ Known)

Assumptions:

- Population standard deviation σ is known.
- Population is normally distributed.
- If population is not normal, use large sample ($n > 30$).

Confidence interval estimate:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (1)$$

where:

- \bar{X} is the point estimate

- $Z_{\alpha/2}$ is the normal distribution critical value for a probability of $\alpha/2$ in each tail
- $\frac{\sigma}{\sqrt{n}}$ is the standard error

Common Levels of Confidence

- Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	Confidence Coefficient $1 - \alpha$	$Z_{\alpha/2}$ value
80.0%	0.800	1.280
90.0%	0.900	1.645
95.0%	0.950	1.960
98.0%	0.980	2.330
99.0%	0.990	2.580
99.8%	0.998	3.080
99.9%	0.999	3.270

Example

- A sample of 11 circuits from a large normal population has a mean resistance of 2.22 ohms. We know from past testing that the population standard deviation is 0.35 ohms.
- Determine a 95% confidence interval for the true mean resistance of the population.

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 2.22 \pm (1.96) \frac{.35}{\sqrt{11}} = 2.22 \pm 0.2068 \quad (2)$$

Interpretation

- We are 95% confident that the true mean resistance is between 2.0132 and 2.4268 ohms.
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner may contain the true mean.

Do You Ever Truly Know σ ?

- Probably not!

In virtually all real world business situations, σ is not known.

If there is a situation where σ is known then μ is also known (since to calculate σ you need to know μ).

If you truly know μ there would be no need to gather a sample to estimate it.

Confidence Interval for μ (σ Unknown)

- If the population standard deviation σ is unknown, we can substitute the sample standard deviation, S
- This introduces extra uncertainty, since S is variable from sample to sample.
- So we use the *t distribution* instead of the normal distribution.
- Assumptions:
 - Population standard deviation is unknown.
 - Population is normally distributed.
 - If population is not normal, use large sample ($n > 30$).

Use Student's t Distribution

Confidence Interval Estimate:

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} \quad (3)$$

(where $t_{\alpha/2}$ is the critical value of the t distribution with $n - 1$ degrees of freedom and an area of $\alpha/2$ in each tail).

Student's t Distribution

- The t is a family of distributions.
- The $t_{\alpha/2}$ value depends on *degrees of freedom (d.f.)*.
- Number of observations that are free to vary after sample mean has been calculated.
 - d.f. = $n - 1$

Degrees of Freedom (df)

Idea: Number of observations that are free to vary after sample mean has been calculated.

Example: Suppose the mean of 3 numbers is 8.0

Let $X_1 = 7$ Let $X_2 = 8$ What is X_3 ?

If the mean of these three values is 8.0, then X_3 must be 9.

(i.e., X_3 is not free to vary)

Here, $n = 3$, so degrees of freedom = $n - 1 = 3 - 1 = 2$

(2 values can be any numbers, but the third is not free to vary for a given mean)

Table 4: Confidence Levels by degrees of freedom (df)

CL	t(df=10)	t(df=20)	t(df=30)	Z(df=inf)
.80	1.372	1.325	1.310	1.280
.90	1.812	1.725	1.697	1.645
.95	2.228	2.086	2.042	1.960
.99	3.169	2.845	2.750	2.580

Note: $t \rightarrow Z$ as n increases.

Figure 2

Figure 2 shows part of this table in graphic form. I have plotted the *Normal Curve*, which has the highest peak, the *student-t* distribution with 5 degrees of freedom, which is middle plot, and the *student-t* distribution with 3 degrees of freedom, which is the plot with the lowest peak. Although the peak is interesting, the important feature is that the tails of the two *student t* distributions are the ‘fattest’. This allows for a greater interval width using the same confidence interval (80%, 90%, 95%, etc.). It is a more conservative measurement than the *normal* distribution.

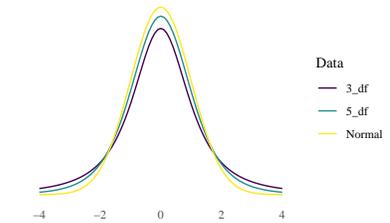


Figure 2: Shapes of Distributions

Example of *t* distribution confidence interval

A random sample of $n = 25$ has $\bar{X} = 50$ and $S = 8$. Form a 95% confidence interval for μ .

$$\text{d.f.} = n - 1 = 24, \text{ so } t_{\alpha/2} = t_{.025} = 2.064$$

$$\text{The confidence interval is: } \bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = \bar{X} \pm (2.064) \frac{8}{\sqrt{25}}$$

$$(2.064) \frac{8}{\sqrt{25}} = 3.302$$

$$\text{The confidence interval is } 46.698 \leq \mu \leq 53.302$$

Interpreting this interval requires the assumption that the population you are sampling from is approximately a normal distribution (especially since n is only 25).

This condition can be checked by creating a:

- Normal probability plot or
- Boxplot

Confidence Intervals for the Population Proportion, π

An interval estimate for the population proportion (π) can be calculated by adding an allowance for uncertainty to the sample proportion (p)

Recall that the distribution of the sample proportion is approximately normal if the sample size is large, and must have $np > 5$ and

$n(1-p) > 5$ and the standard error of the proportion is:

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \quad (4)$$

Confidence Interval Endpoints

Upper and lower confidence limits for the population proportion are calculated with the formula

$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \quad (5)$$

where: - $Z_{\alpha/2}$ is the standard normal value for the level of confidence desired

- p is the sample proportion

- n is the sample size

Note: must have $np > 5$ and $n(1-p) > 5$

Example

A random sample of 100 people shows that 25 are left-handed. Form a 95% confidence interval for the true proportion of left-handers.

$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} = \frac{25}{100} \pm 1.96 \sqrt{\frac{(.25)(.75)}{100}} = \frac{25}{100} \pm 1.96(0.0433) \quad (6)$$

So: We are 95% confident that $\bar{X} \pm 0.0433$ contains the population proportion.

$$0.1651 \leq p \leq 0.3349 \quad (7)$$

Interpretation

We are 95% confident that the true percentage of left-handers in the population is between 16.51% and 33.49%.

Although the interval from 0.1651 to 0.3349 may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.

Determining Sample Size

Sampling Error

The required sample size can be found to reach a desired *margin of error* (e) with a specified level of confidence ($1 - \alpha$).

The margin of error is also called *sampling error*, it is:

- the amount of imprecision in the estimate of the population parameter

- the amount added and subtracted to the point estimate to form the confidence interval.

For the Mean

$$e = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \rightarrow \text{Now solve for } n \rightarrow n = \frac{Z_{\alpha/2}^2 \sigma^2}{e^2}$$

To determine the required sample size for the mean, you must know:

- The desired level of confidence $(1 - \alpha)$, which determines the critical value, $Z_{\alpha/2}$
- The acceptable sampling error, e
- The standard deviation, σ

Required Sample Size Example

If $\sigma = 45$, what sample size is needed to estimate the mean within ± 5 with 90% confidence?

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{e^2} = \frac{(1.645^2)(45^2)}{5^2} = 219.19 \quad (8)$$

so the required sample size is 220 (always round up).

If σ is unknown

If unknown, σ can be estimated when using the required sample size formula:

- Use a value for σ that is expected to be at least as large as the true σ .
- Select a pilot sample and estimate σ with the sample standard deviation, S

Determining Sample Size For the Population

$$e = Z \left(\sqrt{\frac{\pi(1-\pi)}{n}} \right) \rightarrow \text{Solve for } n \rightarrow n = \frac{Z_{\alpha/2}^2 (\pi(1-\pi))}{e^2}$$

To determine the required sample size for the proportion, you must know:

- The desired level of confidence $(1 - \alpha)$, which determines the critical value, $Z_{\alpha/2}$
- The acceptable sampling error, e
- The true proportion of events of interest, π
 - π can be estimated with a pilot sample if necessary (or conservatively use 0.5 as an estimate of π)

Required Sample Size Example

How large a sample would be necessary to estimate the true proportion defective in a large population within $\pm 3\%$, with 95% confidence?

(Assume a pilot sample yields $p = 0.12$)

Solution:

For 95% confidence, use $Z_{\alpha/2} = 1.96$

$e = 0.03$

$p = 0.12$, so use this to estimate π

$$n = \frac{Z_{\alpha/2}^2(\pi(1 - \pi))}{e^2} = \frac{(1.96)^2(.12)(.88)}{(0.03)^2} = 450.74 \quad (9)$$

So: use $n = 451$

Ethical Issues

- A confidence interval estimate (reflecting sampling error) should always be included when reporting a point estimate.
- The level of confidence should always be reported.
- The sample size should be reported.
- An interpretation of the confidence interval estimate should also be provided.

Final Note

The important thing to remember is that the margin of error, confidence interval, is generally a function three things, the degree of confidence required, the sample size and the percentage being estimated. Thus, sampling error will **decrease** as:

- The sample size (or number of interviews) gets bigger;
- The percentage estimated approaches 0% or 100% or
- The need to be certain about the result (e.g. the “confidence level”) gets smaller.