Introduction to confidence intervals

Millie Harris

Summary

In statistics, to estimate an unknown parameter you can construct a confidence interval. This is the range of values you expect the true estimate to fall between if you were to repeat the study several times, with a certain level of confidence. This study guide introduces confidence intervals, confidence levels, and Z values using the normal distribution.

Before reading this guide, it is recommended that you read Guide: Introduction to hypothesis testing, Guide: Introduction to probability, and Guide: Expected value, variance, standard deviation

What is a confidence interval?

If you were conducting a study and took several different samples of data, the mean for that data could be slightly different each time. So, when estimating population means, instead of providing one value, you can specify a **range of values** which is likely to contain the true mean. This is called a **confidence interval** (CI).

CIs are a vital tool used to measure uncertainties in everyday life. For example:

- In politics, confidence intervals can be used to show the uncertainty in polling estimates.
- In economics, confidence intervals can be used to show uncertainties in market trends and inflation.
- In medicine and biology, confidence intervals can be used to show uncertainties around effects like mean weight loss, drug effectiveness, or survival rates.
- In sports, confidence intervals can be used by coaches to measure the true performance levels of athletes.

This guide will focus on how to construct and interpret a confidence intervals using the **normal distribution**, Z Values, two-tailed alpha values, and confidence levels. For information on confidence intervals using other distributions see [Guide: More on confidence intervals].

The normal distribution

The **normal distribution,** or sometimes called the "bell-curve", is a symmetrical graph which represents data that clusters around the mean. It depends on two parameters: the mean (μ) and the standard deviation (σ). In situations where $\mu = 0$ and $\sigma = 1$, you have a **standard** normal distribution. For more on μ and σ see Guide: Expected value, variance, standard deviation

It is commonly used in statistics for data sets such as average height, average IQ, or average age of people. You would expect these samples to have a lot of values near the mean and less values on each extreme.

For more information on the normal distribution see Factsheet: Normal Distribution

i Definition of the normal distribution

For samples where μ is the mean and σ is the standard deviation, a random variable X which is normally distributed is represented as

$$X \sim N(\mu, \sigma)$$

which you read as X is normally distributed with parameters mu and sigma.

The tails of the normal distribution

Because of the **bell shape** of the normal distribution, there are less values at each end. The extremes of the normal distribution are called the **tails**.

- The tails themselves are the **areas at each end of the curve.** The total area of both tails is called alpha (α). So, the area of each tail is $\frac{\alpha}{2}$.

To construct a confidence interval you need both tails, because you are looking at values on both extremes, and so you will construct what is called a **two-tailed test.**

! α is 1 minus the CL

- $1-\alpha$ is the confidence level. So you only need **one** in order to generate a confidence interval.

What is a confidence level?

i Definition of a confidence level

A **confidence level** (CL) suggests that if you were to repeat the study many times, you would expect the true estimate to fall within CL$\%$ of the results.
A CL is typically represented using a percentage or decimal. For example, a 95% CL can be represented as 0.95.

The value of α (and/ or the CL) is decided before constructing the confidence interval. Every value of α gives scores on the x axis which leaves that much $\frac{\alpha}{2}$ in each tail. Because of the **symmetry** of the normal distribution, these scores are plus and minus each other. This is called the Z value.

What is a Z value (Z score)?

Using the normal distribution, a Z **value (sometimes called** Z **score or standard score)** is a known test statistic. It shows how many standard deviations above or below the mean an observed data point is.

i Definition of the Z value using the normal distribution

A Z value is written

$$Z_{\frac{\alpha}{2}}$$

To know $Z_{\frac{\alpha}{2}}$, you need to specify the α (and/ or CL) value. You can then use the calculator to find out $Z_{\frac{\alpha}{2}}$.

i **Example 1**

Use the Z value calculator below to find the Z values for these alphas. They have been chosen as these are the most commonly used alpha values in statistics.
Try:

- $\alpha = 0.05$
- $\alpha = 0.1$
- $\alpha = 0.01$

Using the normal distribution, alpha value (and/ or the confidence level), and the corresponding

Z value, you can construct a confidence interval.

So, how do you construct a confidence interval?

i Definition of confidence interval using the normal distribution

A **confidence interval** (CI) is made up of the **sample mean** $(\bar{x})$ and the **standard error** SE. It is constructed by

$$\bar{x} \pm SE$$

Which you read as x bar plus or minus the standard error.
The **standard error** is made up of the Z value $(Z_{\frac{\alpha}{2}})$, the sample **standard deviation** (σ), and the **sample size** (n). It is written as

$$Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

So your confidence interval is written as

$$[\bar{x} - SE, \bar{x} + SE]$$

where $\bar{x} - SE$ is called your **lower bound** and $\bar{x} + SE$ is called your **upper bound**.

General steps for constructing a confidence interval using the normal distribution

Step 1: What do you need?

- the sample mean $(\bar{x})$
- the alpha value (α) or the confidence level (CL) and corresponding Z value
- the sample standard deviation (σ)
- the sample size (n)

Step 2: Use your α (or CL) and the Z value calculator to find

$$Z_{\frac{\alpha}{2}}$$

Step 3: Construct the confidence interval

$$[\bar{x} - SE, \bar{x} + SE]$$

$$= [\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

Step 4: Check your work!
The average of your confidence interval should equal your sample mean.
For example,

$$\bar{x} = 50, [49.2, 50.8]$$

sum the upper and lower bound of the confidence interval, so

$$49.2 + 50.8 = 100$$

and when you divide by 2

$$\frac{100}{2} = 50 = \bar{x}$$

Example 2

Cantor's Confectionery have purchased a new computer to help monitor the quality of their products. They test the average weight of their best selling products. The computer outputs a graph. What can be understood from this?

- This is a normal distribution for the mean weight of one of Cantor's Confectioneries best selling products.
- This is a sample and so the mean is the sample mean $(\bar{x})$ with known standard deviation (σ).
- When you select different values for alpha (α), the tails of the normal distribution change. This means that when constructing a confidence interval the Z values will be different.

Example 3

Cantor's Confectionery use the normal distribution from example 1 to construct a 95% confidence interval for the mean weight of their bags of sweets. They take a sample of 100 bags which have an average weight of 75 grams, the calculated standard deviation is 10 grams.

Step 1: What do you need?

- the sample size is 100 bags of sweets so $n = 100$
- the average weight of the sample is 75 grams so $\bar{x} = 75$
- the standard deviation of the sample is 10 grams so $\sigma = 10$
- the confidence level is 95% so $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$

Step 2: Use the Z value calculator to identify

$$Z_{\frac{0.05}{2}} = Z_{0.025} = 1.960$$

Step 3: Construct the confidence interval

$$[\bar{x} - Z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{0.025}\frac{\sigma}{\sqrt{n}}]$$

add in all values from Step 1

$$= [75 - 1.960(\frac{10}{\sqrt{100}}), 75 + 1.960(\frac{10}{\sqrt{100}})]$$

$$= [73.04, 76.96]$$

Step 4: Check your work

$$73.04 + 76.96 = 150$$

$$\frac{150}{2} = 75$$

where $75 =$ the sample mean, so your confidence interval is correct.

Example 4

Using the test from example 3, Cantor's Confectionery ask the computer to work out a 90% confidence interval and 99% confidence interval. It outputs the following results:

$$\text{A } 90\% \text{ CI } = [73.36, 76.64] \text{A } 99\% \text{ CI } = [72.424, 77.576]$$

What does this suggest about the confidence levels?
For all three examples, the sample mean falls within the confidence interval. But, as the confidence level **increases** so does the **width** of the confidence interval.
If you were to think about the tails of these normal graphs, as the CL increases the amount of area underneath each extreme decreases. This means you have more values in the middle of the graph, which is why you have more values in the confidence interval.

Example 5

There is a new shop in town! Lovelace's Lollies are claiming their products are better than Cantor's Confectionery. They take a sample of 77 of their best selling boxes of lollies. The average weight of these boxes is 84 grams with a standard deviation of 9.5 grams. They construct a 95% confidence interval.

- $\bar{x} = 84$
- $Z_{\frac{0.05}{2}} = Z_{0.025} = 1.960$
- $\sigma = 9.5$
- $n = 77$

So to construct the confidence interval

$$[\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

$$= [84 - 1.960(\frac{9.5}{\sqrt{77}}), 84 + 1.960(\frac{9.5}{\sqrt{77}})]$$

$$= [81.88, 86.12]$$

Example 3 and 5 in context

Cantor's Confectionery and Lovelace's Lollies have both constructed a 95% confidence interval from a sample of their best selling products. What does this tell you?

- If Cantor's Confectionery were to repeat the study several times, they would expect the average weight of a bag of sweets to **lie between** 73.04 grams and 76.96 grams, **with** 95% **confidence.**
- If Lovelace's Lollies were to repeat the study several times, they would expect the average weight of a box of lollies to **lie between** 81.88 and 86.12 grams, **with** 95% **confidence.**

Example 6

The rivalry between Cantor's Confectionery and Lovelace's Lollies has reached the News with the following headline
"Here to compete with Cantor's Confectionery: Lovelace's Lollies **guarantee** that 95% of all boxes of lollies will weigh 86.12 grams".
There are issues with this statement.
From the definition of a confidence level, you know that a confidence level suggests that if you were to repeat the study many times, you would expect the true estimate to fall within CL$\%$ of the results.
This is not the same as saying CL$\%$ of the products weigh a certain amount.
Instead, the News headline should read
"Here to compete with Cantor's Confectionery: A study on a sample of Lovelace's Lollies suggest that if more boxes were to be sampled, they expect 95% of boxes would weigh between 81.88 and 86.12 grams".

Example 7

Cantor's Confectionery release a new "family-sized" bag of sweets to compete with Lovelace's Lollies. They weigh 81 bags, in grams, from the new batch and generate the computer generates the following 99% confidence interval

$$[100.00, 104.00]$$

with a sample mean of 102 grams.
Cantor's Confectionery want to know the standard deviation of this sample to compare with their standard deviation of 10 grams from example 3 and Lovelace's Lollies standard deviation of 9.5 grams from example 5
The sample mean is given as

$$\bar{x} = 102$$

The confidence level is 99% so under the normal distribution

$$Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.576$$

The sample size is

$$n = 81$$

So, to work out the sample standard deviation (σ), you must work back from the confidence interval
You can then use **either bound**. So, using the lower bound you know that

$$100 = 102 - 2.576(\frac{\sigma}{\sqrt{81}})$$

by rearranging the equation

$$-2 = -\frac{2.576(\sigma)}{\sqrt{81}}$$

and so,

$$\sigma = 7$$

Which is a lower standard deviation than both their previous test and Lovelace's Lollies test.

Quick check problems

1. What would your CL be for $\alpha = 0.05$?

(a) 85%

(b) 90%

(c) 95%

2. What is the Z value for $\alpha = 0.05$, to 2 decimal places?

3. X is normally distributed with which parameters?

4. What are the extremes of the normal distribution called?

Further reading

For more questions on the subject, please go to Questions: Confidence Intervals.

Version history

v1.0: initial version created 12/06 by Millie Harris as part of a University of St Andrews VIP project.