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

This guide will focus on how to construct and interpret a confidence interval using the **normal distribution**. For information on confidence intervals using other distributions see [Guide: More on confidence intervals].

The normal distribution

```
#| '!! shinylive warning !!': |
#|   shinylive does not work in self-contained HTML documents.
#|   Please set `embed-resources: false` in your metadata.
#| include: false
#| standalone: true
#| viewerHeight: 340
library(shiny)
library(bslib)
library(ggplot2)

ui <- page_sidebar(
  title = "Normal Distribution Tails Explained",
  sidebar = sidebar(
    numericInput("mu",
                 "Mean ( ):",
                 value = 0,
                 step = 0.1),
    numericInput("sigma",
                 "Standard Deviation ( ):",
                 value = 1,
                 min = 0.1,
                 step = 0.1),
    hr(),
    sliderInput("tail_threshold",
                "Tails:",
                min = 1,
                max = 3,
                value = 2,
                step = 0.1),
    hr(),
    p("The distribution curve is shown in blue."),
    p("Red areas represent the tails beyond the threshold."),
    p("Use the threshold slider to see how tail areas change.")
  ),
  card(
    card_header("Normal Distribution with Highlighted Tails"),
    plotOutput("tail_plot", height = "500px")
  )
)

server <- function(input, output, session) {
  output$tail_plot <- renderPlot({
    x_range <- c(input$mu - 4 * input$sigma, input$mu + 4 * input$sigma)
    x <- seq(x_range[1], x_range[2], length.out = 1000)

    y <- dnorm(x, mean = input$mu, sd = input$sigma)

    left_tail <- input$mu - input$tail_threshold * input$sigma
    right_tail <- input$mu + input$tail_threshold * input$sigma

    df <- data.frame(x = x, y = y)

    p <- ggplot(df, aes(x = x, y = y)) +
      geom_line(color = "blue", size = 1.2) +
      geom_area(data = df[df$x >= left_tail & df$x <= right_tail, ],
                alpha = 0.2, fill = "lightblue") +
      geom_area(data = df[df$x <= left_tail, ],
                alpha = 0.6, fill = "red") +
      geom_area(data = df[df$x >= right_tail, ],
                alpha = 0.6, fill = "red") +
      geom_vline(xintercept = left_tail, color = "red", size = 1, linetype = "dashed"
      geom_vline(xintercept = right_tail, color = "red", size = 1, linetype = "dashed
      geom_vline(xintercept = input$mu, color = "blue", size = 1, linetype = "dotted"
      labs(
        title = "The normal distribution",
        subtitle = paste("  =", input$mu, ",   =", input$sigma)
      ) +
      theme_minimal() +
      theme(
        plot.title = element_text(size = 16, hjust = 0.5),
        plot.subtitle = element_text(size = 12, hjust = 0.5),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10)
      )

    max_y <- max(y)
    p + annotate("text",
                 x = input$mu,
                 y = max_y * 0.9,
                 label = " ",
                 color = "blue",
                 size = 5) +
        annotate("text",
                 x = left_tail,
                 y = max_y * 0.5,
                 label = paste("-", input$tail_threshold, " "),
                 color = "red",
                 size = 4,
                 hjust = 1.1) +
        annotate("text",
                 x = right_tail,
                 y = max_y * 0.5,
                     label = paste("+", input$tail_threshold, " "),
                     color = "red",
                     size = 4,
                     hjust = -0.1)
      })
    }

    shinyApp(ui = ui, server = server)
```

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.

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.

Definition of the normal distribution

A random variable X which is normally distributed appears as

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

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

- mu $= \mu =$ the mean.
- sigma $= \sigma =$ the standard deviation, which tells you how spread out your data is from your mean.

The extremes of the distribution are the values furthest away from the mean and are called the **tails**. The total of both tails is alpha (α) so one tail $= \frac{\alpha}{2}$. Under the normal distribution, the alpha value generates a 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.

Definition of the Z value using the normal distribution

A Z value is written $Z_{\frac{\alpha}{2}}$, where

- $\alpha =$ alpha $=$ the sum of the two tails

So for:

$$Z_{\frac{\alpha}{2}} = Z_{0.025}$$

$$\alpha = 0.05$$

Example 1

Use the Z value calculator below to find the Z values for different values of alpha.Try: - $\alpha = 0.5$ - $\alpha = 0.1$ - $\alpha = 0.01$

```
#| '!! shinylive warning !!': |
#|   shinylive does not work in self-contained HTML documents.
#|   Please set `embed-resources: false` in your metadata.
#| standalone: true
#| viewerHeight: 340
#| include: false
library(shiny)
```

Using this, you can define your confidence level which you will need when constructing a confidence interval.

What is a confidence level?

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.

Using the normal distribution the confidence level, alpha value, and Z value are used to construct a confidence interval.

So, 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 in economics, medicine, and they can measure uncertainties in everyday life. For example, weather forecasting or outcomes in sports.

Definition of confidence interval using the normal distribution

A **confidence interval** (CI) is written as

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

Which you read as $x% bar plus or minus the **standard error**. Where:

- $\bar{x}$ = the sample mean
- SE = the standard error

$$SE = Z_{\frac{\alpha}{2}} \frac{\sigma}{n}$$

Where:

- $Z_{\frac{\alpha}{2}}$ = Z value

- $\sigma =$ the sample standard deviation
- $n =$ the sample size

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 use a normal distribution to model the mean weight of their best selling bag of assorted sweets.
Use the slider to adjust the values of α and see the tails change.

```
#| '!! shinylive warning !!': |
#|   shinylive does not work in self-contained HTML documents.
#|   Please set `embed-resources: false` in your metadata.
#| standalone: true
#| viewerHeight: 340
#| include: false
library(shiny)
library(bslib)
library(ggplot2)

ui <- page_sidebar(
  title = "Cantor's Confectionery: The normal distribution for assorted sweets",
  sidebar = sidebar(
    radioButtons("alpha_level",
                 "Significance level ( ):",
                 choices = list("  = 0.10" = 0.10,
                               "  = 0.05" = 0.05,
                               "  = 0.01" = 0.01),
                 selected = 0.05)
  ),
  card(
    card_header("Weight Distribution of Assorted Sweets"),
    plotOutput("tail_plot", height = "500px")
  )
)

server <- function(input, output, session) {
  output$tail_plot <- renderPlot({
    mu <- 100
    sigma <- 12
    alpha <- as.numeric(input$alpha_level)

    x_range <- c(mu - 4 * sigma, mu + 4 * sigma)
    x <- seq(x_range[1], x_range[2], length.out = 1000)

    y <- dnorm(x, mean = mu, sd = sigma)

    z_critical <- qnorm(1 - alpha/2)
    left_tail <- mu - z_critical * sigma
    right_tail <- mu + z_critical * sigma

    df <- data.frame(x = x, y = y)
```

From the distribution,

$$\mu = 100g$$

means that from the sample, the average weight of the assorted bags of sweets is $100g$.

$$\sigma = 12g$$

means that from the sample, the standard deviation of the assorted bags of sweets is $12g$.

Example 3

Cantor's Confectionery want to use the normal distribution 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 $75g$, the calculated standard deviation is $10g$.
Step 1: What do you need?

- the sample size is 100 bags of sweets so $n = 100$
- the average weight of the sample is $75g$ so $\bar{x} = 75$
- the standard deviation of the sample is $10g$ 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

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{n}$$

$$= 75 \pm 1.960(\frac{10}{100})$$

$$= [74.804, 75.196]$$

Step 4: Check your work

$$74.804 + 75.196 = 150$$

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

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

Example 4

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 $84g$ with a standard deviation of $9.5g$. 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

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{n}$$

$$= 84 \pm 1.960(\frac{9.5}{77})$$

$$= [83.758, 84.242]$$

Example 3 and 4 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

  $$[74.804, 75.196]g$$

  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

$$[83.758, 84.242]g$$

ℹ Example 5

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 $84.242g$".

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 $[83.758, 84.242]g$".

ℹ Example 6

Cantor's Confectionery release a new "family sized" bag of sweets to compete with Lovelace's Lollies. They test 150 bags, in grams (g), from the new batch and generate the following 99% confidence interval

$$[126.789, 132.443]$$

with a sample mean of $129.616g$. What is the sample standard deviation?

The sample mean is given as

$$\bar{x} = 129.616$$

The confidence level is 99% so under the normal distribution

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

The sample size is

$$n = 150$$

So,

$$[83.758, 84.242]g$$

To work out the sample standard deviation (σ), you must work back from the confidence interval

$$[126.789, 132.443] = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{n}$$

$$[126.789, 132.443] = 129.616 \pm 2.576(\frac{\sigma}{150})$$

You can then use either bound. So, using the lower bound

$$126.789 = 129.616 - 2.576(\frac{\sigma}{150})$$

$$-2.827 = \frac{-2.576(\sigma)}{150}$$

$$\frac{2.827(150)}{2.576} = \sigma$$

So,

$$\sigma = 164.616g$$

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.