FASS512: Confidence intervals. Hypothesis testing.

Professor Patrick Rebuschat, [p.rebuschat@lancaster.ac.uk](mailto:p.rebuschat@lancaster.ac.uk)

Please work through the following handout at your own pace.

As in the previous handouts, please type the commands in your computer. That is, don’t just read the commands on the paper, please type every single one of them.

Before running the commands, think about what you expect to happen. If you are able to do this, that’s a good sign that you are starting to understand the R language. ☺

This handout assumes that you have completed all previous handouts. If you haven’t, please do this before working on the following handout. Handouts are available on [Moodle](https://modules.lancaster.ac.uk/course/view.php?id=37879).

**References for this handout**

Many of the examples and data files from our class come from these excellent textbooks:

* Andrews, M. (2021). *Doing data science in R*. Sage.
* Brown, D. S. (2021). *Statistics and data visualization using R. The art and practice of data analysis*. Sage.
* Cumming, G. & Calin-Jaegeman, R. J. (2017). *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. Routledge.
* Crawley, M. J. (2013). *The R book*. Wiley.
* Fogarty, B. J. (2019). *Quantitative social science data with R*. Sage.
* Winter, B. (2019). *Statistics for linguists. An introduction using R*. Routledge.

Are you ready? Then let’s start on the next page! ☟

Task 1: *Confidence intervals*

Practice calculating 95% confidence intervals in R.

library(tidyverse)

We assume μ and σ are unknown. We use t-distribution rather than z-distribution.

View(MASS::nlschools)

summary(MASS::nlschools)

We place the IQ scores in the data set in a new object, raw\_IQ.

all\_IQ\_scores <- MASS::nlschools$IQ

TWO WAYS OF DOING THIS

*Example 1*

To calculate the 95% CIs for the IQ variable, we need to know the following:

* sample mean M
* critical t-value for 95% confidence and for our sample size (corresponds to degrees of freedom, N-1)
* standard error SE
* standard deviation SD
* sample size N

Calculating the sample mean M.

mean <- mean(all\_IQ\_scores) # M = 11.83406

Calculating the sample standard deviation SD.

sd <- sd(all\_IQ\_scores) # SD = 2.06889

We can find out the sample size by using the length() function.

sample\_size <- length(all\_IQ\_scores) # N = 2,287

To calculate the standard error SE we divided the standard deviation by the square root of the sample size.

standard\_error <- (sd / sqrt(sample\_size)) # SE = 0.04326177

We can find out the critical t-value for 95% confidence, bearing in mind the sample size 2,287.

standard\_error

[1] 0.04326177

Finally, we also need to determine the critical value for 95% confidence from the t-distribution. The qt() function does this for us. The first argument is the probability .05, which we here divide by two as we are interested in scores in two tails, the left and the right one. The second argument is the degrees of freedom (2,286). Remember: The degrees of freedom is N – 1, so 2,287 – 1 = 2,286. Remember that the t-value can be positive or negative, depending on the tail in question. Below, we also add the abs() function, which makes sure that we get a positive value.

t\_value <- abs(qt(p = .05 / 2, df = 2286))

t\_value <- abs(qt(.05 / 2, 2286))

t\_value

[1] 1.961002

We’re finally ready to calculate the 95% CIs for the IQ scores.

These commands tell us the 95% lower and upper limits.

lower\_limit <- mean - (t\_value \* standard\_error)

upper\_limit <- mean + (t\_value \* standard\_error)

intervals\_95 <- c(lower\_limit, upper\_limit)

intervals\_95

[1] 11.74923 11.91890

*Example 2: Using the ESCI for R tools*

Alternative:

Bob Calin-Jageman, 2020 workshop materials, retrieved here:

<https://osf.io/d89xg/wiki/tools:%20esci%20for%20R/>

Installing esci for R

install.packages("devtools")

devtools::install\_github("rcalinjageman/esci")

Estimate a mean

To calculate:

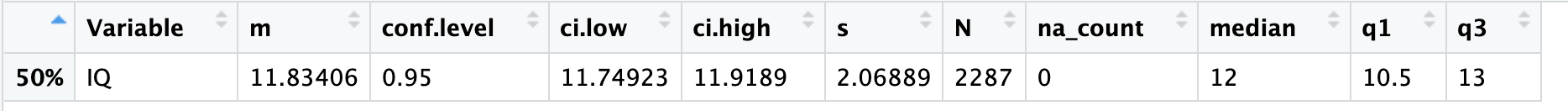
estimate <- estimateMean(MASS::nlschools,

IQ,

conf.level = .95)

To make it particularly easy to read off the values, I like to use the View() function with estimate$summary\_data as the argument.

View(estimate$summary\_data)



Another example, the Nettle (1999) data set about languages. We have used this before.

library(esci)

Create the object with the data. This dataset is from Winter (2019), who in turn took the data from Nettle’s (1999) book Linguistic Diversity.

languages <- read.csv('nettle\_1999\_climate.csv')

Let’s have a look.

View(languages)

The data set lists 74 countries (Country), and for each country tells us population size (Population), the area (Area), the mean growing season (MGS, measured in months), and the number of languages spoken (Langs).

Let’s calculate the 95% CIs for number of languages spoken (Langs). What is the point estimate? What is the 95% confidence interval?

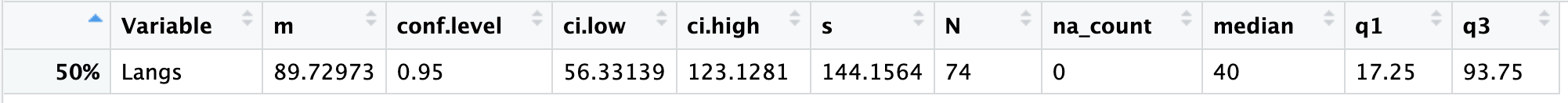
estimate <- estimateMean(languages,

Langs,

conf.level = .95

)

View(estimate$summary\_data)



We can also use ggplot to visualize the distribution and the 95% CIs.

plotEstimatedMean(estimate)

Just like in other plots, you can customize the plot with the usual ggplot options (labels, titles, colors, themes, line width and type, etc. etc. etc. etc.).

Below, we first create an object called langs\_CI.

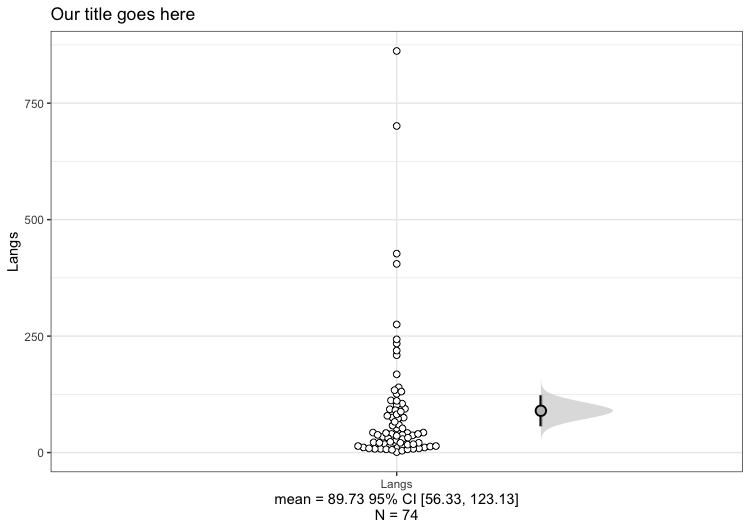
langs\_CI <- plotEstimatedMean(estimate)

Then, we customize the plot by adding the title “Our titles goes here”.

langs\_CI <- langs\_CI + ggtitle("Our title goes here")

If you now run the command langs\_CI, you will see your plot. (Note: The comments and arrows won’t be visible, of course, I added them for explanation.)

langs\_CI



The 74 observations, one bubble per country.

The 95% CI. The point estimate (M) is in the center.

The plot below displays the distribution of scores (74 observations, each one transparent bubble) and the 95% CI to the right.

What does the figure tell us about the distribution of Langs, the number of languages spoken across the 74 countries? How precise is the CI?