

4.3: Confidence intervals

We are ready now to make the first step in the world of inferential statistics and use *statistical tests*. They were invented to solve the main question of statistical analysis (Figure 4.3.1): how to estimate anything about *population* using only its *sample*? This sounds like a magic. How to estimate the whole population if we know nothing about it? However, it is possible if we know some data law, feature which our population should follow. For example, the population could exhibit one of *standard data distributions*.

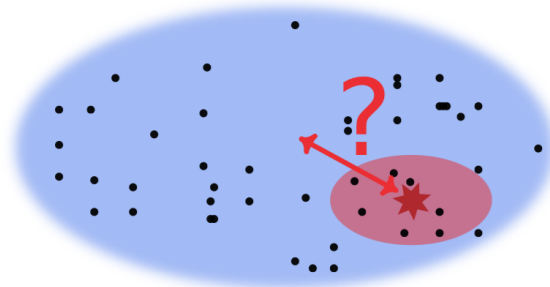


Figure 4.3.1 Graphic representation of the main statistical question: how to estimate population (blue) from sample (red)? Red arrow relates with the confidence interval. To answer “big red” question, one needs the *p*-value.

Let us first to calculate *confidence interval*. This interval *predict* with a given probability (usually 95%) where the particular central tendency (mean or median) is located within population. Do not mix it with the 95% quantiles, these measures have a different nature.

We start from checking the *hypothesis* that the *population mean is equal to 0*. This is our *null hypothesis*, H_0 , that we wish to accept or reject based on the test results.

Here we used a variant of *t-test* for univariate data which in turn uses the standard *Student’s t-distribution*. First, this test obtains a specific *statistic* from the original data set, so-called *t-statistic*. The test statistic is a single measure of some attribute of a sample; it reduces all the data to one value and with a help of standard distribution, allows to re-create the “virtual population”.

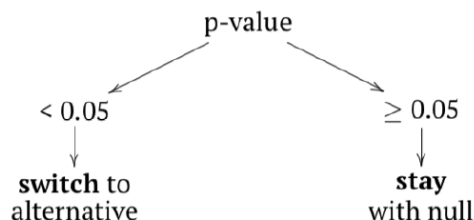
Student test comes with some price: you should assume that your population is “parametric”, “normal”, i.e. interpretable with a normal distribution (dart game distribution, see the glossary).

Second, this test estimates if the statistic derived from our data can reasonably come from the distribution defined by our original assumption. This principle lies at the heart of calculating *p-value*. The latter is the probability of obtaining our test statistic if the initial assumption, *null hypothesis* was true (in the above case, mean tree height equals 0).

What do we see in the output of the test? *t-statistic* equals 66.41 at 30 degrees of freedom ($df = 30$). *P*-value is really low (2.2×10^{-16}), almost zero, and definitely much lower then the “sacred” confidence level of 0.05.

Therefore, we *reject the null hypothesis*, or our initial assumption that mean tree height equals to 0 and consequently, go with the *alternative hypothesis* which is a logical opposite of our initial assumption (i.e., “height is *not* equal to 0”):

However, what is really important at the moment, is the *confidence interval*—a range into which the true, population mean should fall with given probability (95%). Here it is narrow, spanning from 73.7 to 78.3 and *does not include zero*. The last means again that null hypothesis is not supported.



If your data does not go well with normal distribution, you need more universal (but less powerful) *Wilcoxon rank-sum test*. It uses *median* instead of mean to calculate the test statistic *V*. Our null hypothesis will be that *population median is equal to zero*:

(Please ignore warning messages, they simply say that our data has ties: two salaries are identical.)

Here we will also reject our null hypothesis with a high degree of certainty. Passing an argument `conf.int=TRUE` will return the confidence interval for population median—it is broad (because sample size is small) but does not include zero.

This page titled 4.3: Confidence intervals is shared under a [Public Domain](#) license and was authored, remixed, and/or curated by Alexey Shipunov via source content that was edited to the style and standards of the LibreTexts platform.