

10.3: R and Bootstrapping

All generalities like standard deviation and mean are normally taken from sample but meant to represent the whole statistical population. Therefore, it is possible that these estimations could be seriously wrong. Statistical techniques like *bootstrapping* were designed to minimize the risk of these errors. Bootstrap is based only on the given sample but try to estimate the whole population. The idea of bootstrap was inspired by from Buerger and Raspe “Baron Munchausen’s miraculous adventures”, where the main character pulls himself (along with his horse) out of a swamp by his hair (Figure 10.3.1). Statistical bootstrap was actively promoted by Bradley Efron since 1970s but was not used frequently until 2000s because it is computationally intensive. In essence, *bootstrap* is the re-sampling strategy which replaces part of sample with the subsample of its own. In R, we can simply `sample()` our data *with the replacement*.



Figure 10.3.1 Baron Munchausen pulls himself out of swamp. (Illustration of Gustave Doré.)

First, we will bootstrap the mean (Figure 10.3.2) using the advanced `boot` package:

(Note that here and in many other places in this book number of replicates is 100. For the working purposes, however, we recommend it to be at least 1,000.)

Package `boot` allows to calculate the 95% confidence interval:

More basic `bootstrap` package bootstraps in a simpler way. To demonstrate, we will use the `spur.txt` data file. This data is a result of measurements of spur length on 1511 *Dactylorhiza* orchid flowers. The length of spur is important because only pollinators with mouth parts comparable to spur length can successfully pollinate these flowers.

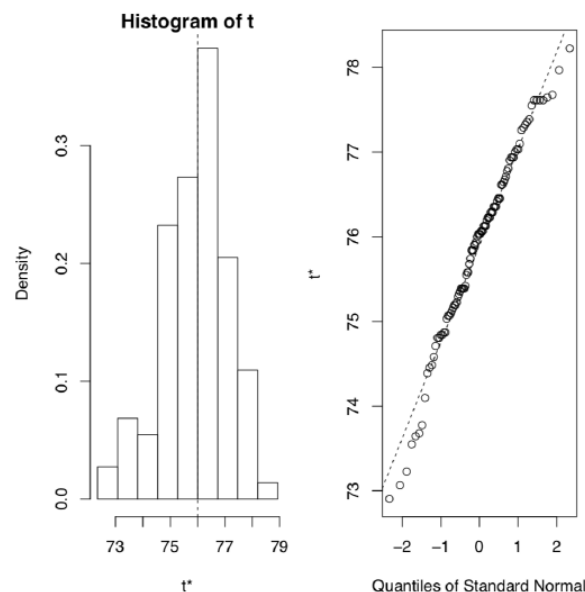


Figure 10.3.2 Graphical representation of bootstrapping sample median.

Jackknife is similar to the bootstrap but in that case observations will be taking out of the sample one by one without replacement:

This is possible to bootstrap standard deviation and mean of this data even without any extra package, with `for` cycle and `sample()`: (Alternatively, `tt` could be an empty data frame, but this way takes more computer time which is important for bootstrap. What we did above, is the *pre-allocation*, useful way to save time and memory.)

Actually, spur length distribution does not follow the normal law (check it yourself). It is better then to estimate median and median absolute deviation (instead of mean and standard deviation), or median and 95% range:

(Note the use of `replicate()` function, this is another member of `apply()` family.)

This approach allows also to bootstrap almost any measures. Let us, for example, bootstrap 95% confidence interval for Lyubishchev's K:

Bootstrap and jackknife are related with numerous *resampling techniques*. There are multiple R packages (like `coin`) providing resampling tests and related procedures:

Bootstrap is also widely used in the machine learning. Above there was an example of `Jclust()` function from the `asmisc.r` set. There also are `BootA()`, `BootRF()` and `BootKNN()` to bootstrap non-supervised and supervised results.

) plants. Use bootstrap and resampling methods.

This page titled 10.3: R and Bootstrapping 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.