The Foundations of Data Science

Japjot Singh

October 24, 2018

1 Estimation

How to make justifianble conclusions about an unknown parameter, based on the data from a random sample? How can inferential thinking be leveraged to estimate a numerical parameter and quantity the error in the estimate?

1.1 Percentiles

Numerical data can be sorted from increasing to decreasing order, thus providing each of the data with a *rank order*, a pecentile is the value at a particular rank (i.e the median is the 50th percentile because 50% of the values in a data set are above the median). However it is important to remember that ties, equal data values (i.e. all students getting 75 on an exam), be taken into account when defining percentiles.

Definition 1. ***Percentile****: let p be a number between 0 and 100. the pth percentile of a collection is the smallest value in the collection that is at least as large as p% of the values.*

Method for finding the pth percentile:

1. Sort the collection in increasing order
2. Set k to p% of n, $k = \frac{p}{100} * n$
3. If k is an integer, take the kth element of the sorted collection
4. If k is not an integer, round it up to the next integer and find that element in the sorted collection

Let `heights` be a an array containing a random samply of heights of 7th grade boys in Orange County. To calculate the 85th percentile of heights, simply do `percentile(95, heights)`.

Quartiles

The *first quartile* of a mnumerical collection is the 25th percentile. The second quartile is the median, and the third quatile is the 75th percentile.

1.2 Bootstrap

The motivation behind bootsrap is to make statistically sound conclusions from a limited random sample. Instead of generating new samples from the population, the bootstrap leverages *resampling*, drawing samples from the original sample, to generate new samples.

By the law of averages, as the sample size gets larger and larger the distribution of the sample begins to resemble that of the population, and as a result a sample statistic (mean, median, etc) will approach the population statistic.

Since a large random sample is likely to resemble the population from which it is drawn a data scientist can leverage the bootstrap method: **treat the original sample** as if it were the **population** itself, and **draw from the sample**, at random **with replacement**, **the same number of times as the original sample size**.

Sampling the same number of times as the original sample size ensures that the variability of the sample size is accounted for. Drawing samples **with replacement** allows for the possibility of samples different from the original, where some parts that where in the original sample are left out.

By the law of averages, the distribution of the original sample resembles the distribution of the popiulation, so then the distributions of all the resamples are likely to resemble the distribution of the original sample. Thus the distributions of the resamples will also resemble distribution of the population.

1.3 Confidence Intervals

Usually, data scientists don't know the value of a parameter and will seek to estimate it with a certain level of certainty.

To calculate a statistic with some degree of confidence first bootstrap the random sample of the population with replacement with the number of replications as the sample size. Then, for 95% confidence, calculate the 2.5th and 97.5th percentiles and the two numbers will be the lower and upper bound respectively of the confidence interval.

Some key notes when using Boostrap:

- Start with a large random sample, Bootstrapping hinges on the Law of Averages, which in turn relies upon large random samples as large random samples (and hence resamples of the sample) resemble the population
- It is good to replicate the resampling procedure as many times as possible, targetting to resample 10,000 times is a good benchmark
- There are some situations where bootstrapping has limitations and is **not** effective:
  - When estimating the minimum or maximum value of a population or a very low or very high percentile, or if parameteres are greatly influenced by rare elements of the population
  - The probability distribution of the statistic is not roughly bell shaped
  - the original sample is very small, rougly less than 10 or 15