Sampling and hypothesis tests

Part 2: Central limit theorem, bootstrap and confidence interval

By: Noureddin Sadawi, PhD

University of London

Central limit theorem

- If we have a numeric dataset (its distribution does not matter).
- If we take a sufficient number of samples from this data and calculate the mean of each sample, these means will approximate a normal distribution (see next topic).
- The more samples we take, and the bigger they are, the closer to a normal distribution the sample means will be.
- Moreover, the mean of these means will be approximately the same as that of the original dataset.
- It has many applications in statistical studies.

Standard error

- ‘The standard error is a single metric that sums up the variability in the sampling distribution for a statistic.
- The standard error can be estimated using a statistic based on the standard deviation s of the sample values, and the sample size n.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

$$Standard\ error = SE = \frac{s}{\sqrt{n}}$$

Key ideas

- 'The frequency distribution of a sample statistic tells us how that metric would turn out differently from sample to sample.
- This sampling distribution can be estimated via the bootstrap, or via formulas that rely on the central limit theorem.
- A key metric that sums up the variability of a sample statistic is its standard error.'

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

The bootstrap

- One easy and effective way to estimate the sampling distribution of a statistic, or of model parameters, is to draw additional samples, with replacement, from the sample itself and recalculate the statistic or model for each resample.
- This procedure is called **the bootstrap**, and it does not necessarily involve any assumptions about the data or the sample statistic being normally distributed.'

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key ideas

- ‘The bootstrap (sampling with replacement from a data set) is a powerful tool for assessing the variability of a sample statistic.
- The bootstrap can be applied in similar fashion in a wide variety of circumstances, without extensive study of mathematical approximations to sampling distributions.
- It also allows us to estimate sampling distributions for statistics where no mathematical approximation has been developed.
- When applied to predictive models, aggregating multiple bootstrap sample predictions (bagging) outperforms the use of a single model.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Confidence intervals (CIs)

The confidence interval (CI) is another way to understand the potential error in a sample estimate.

'**Confidence level**

- The percentage of confidence intervals, constructed in the same way from the same population, that are expected to contain the statistic of interest.

Interval endpoints

- The top and bottom of the confidence interval.'

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Confidence intervals are used to represent an estimate as a range instead of a single number.

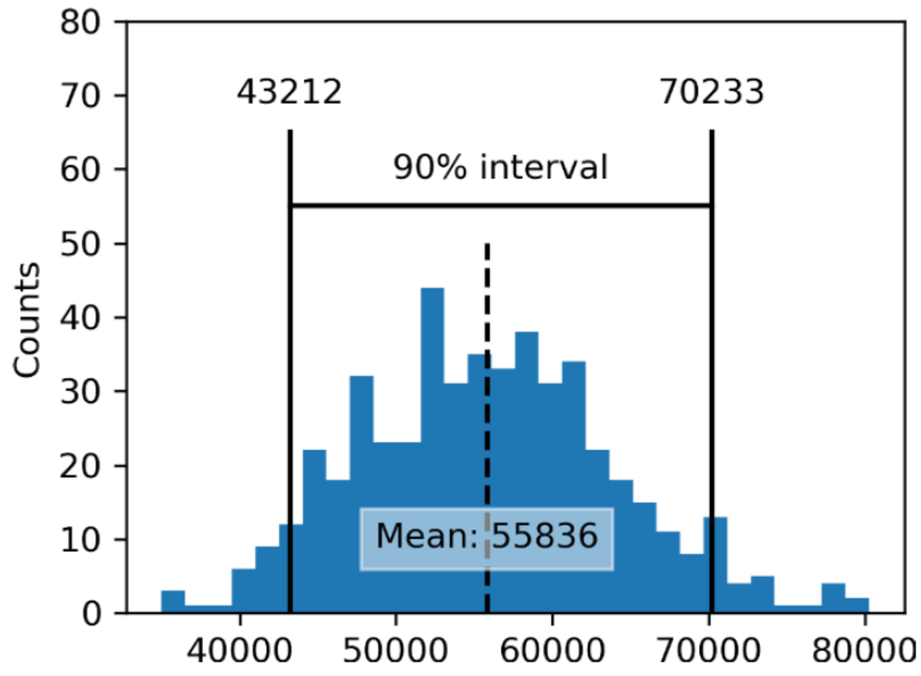
Algorithm

‘Given a sample of size n, and a sample statistic of interest, the algorithm for a bootstrap confidence interval is as follows:

1. Draw a random sample of size n with replacement from the data (a resample).
2. Record the statistic of interest for the resample.
3. Repeat steps 1–2 many (R) times.
4. For an x% confidence interval, trim [(100-x) / 2]% of the R resample results from either end of the distribution.
5. The trim points are the endpoints of an x% bootstrap confidence interval.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example



Bootstrap confidence interval for the annual income of loan applicants, based on a sample of 20.

Source: (Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key ideas

- ‘Confidence intervals are the typical way to present estimates as an interval range.
- The more data you have, the less variable a sample estimate will be.
- The lower the level of confidence you can tolerate, the narrower the confidence interval will be.
- The bootstrap is an effective way to construct confidence intervals.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).