Data Science Career Track

The Art of Statistics, Chapter 7: How sure can we be about what is going on?

Take-Away Notes

This chapter describes the various concepts of uncertainty that ought to be used when drawing conclusions with statistics and communicating those claims to others.

When we take samples from a population about which we'd like to discover things, there will be variability in those samples. But if we have been prudent and avoided internal biases (for example by taking random samples) the summary statistics of the samples should be close to those of the study population.

A **margin of error** is a plausible range in which a true feature of a population may lie, following a survey. A **confidence interval**, by contrast, is an estimated range within which an unknown parameter may plausibly lie. A confidence interval of 95% for a given value *v* gives a lower limit L and an upper limit U, and then entails that, before observing the data, there's a 95% probability that the random interval (L, U) contains *v*.

- The **Central Limit Theorem** is true: this is the claim that the sample mean of a set of random variables tends to have a normal sampling distribution, regardless of the shape of the underlying sampling distribution of the random variable.
- Since The Central Limit Theorem is true, and close to 95% of a normal distribution lies between the mean +- 2 standard deviations, a common approximation for a 95% confidence interval is the estimate +- 2 standard errors.

- The **standard error** of a sample mean is its standard deviation, when that sample mean is considered as a random variable.

Bootstrapping is a means of generating confidence intervals and the distribution of test statistics by resampling the observed data with replacement, rather than by assuming a particular probability model for the underlying random variable.

- 'Bootstrapping' has the name it does because we're learning about the variability of an estimate without having to make any assumptions about the shape of the population distribution (or pulling ourselves up by our own bootstraps).
- We bootstrap data when we do not want to make assumptions about the shape of the population.
- Crucially, bootstrap distributions allow us to quantify our uncertainty about estimates. For example, we can find the range of values containing 95% of the means of the bootstrap resamples, and this can be a 95% uncertainty interval for the original estimates.