

Summary/Review

Estimation and Inference

Inferential Statistics consist in learning characteristics of the population from a sample. The population characteristics are parameters, while the sample characteristics are statistics. A parametric model, uses a certain number of parameters like mean and standard deviation.

The most common way of estimating parameters in a parametric model is through maximum likelihood estimation.

Through a hypothesis test, you test for a specific value of the parameter.

Estimation represents a process of determining a population parameter based on a model fitted to the data.

The most common distribution functions are: uniform, normal, log normal, exponential, and poisson.

A frequentist approach focuses in observing man repeats of an experiment. A bayesian approach describes parameters through probability distributions.

Hypothesis Testing

A hypothesis is a statement about a population parameter. You commonly have two hypothesis: the null hypothesis and the alternative hypothesis.

A hypothesis test gives you a rule to decide for which values of the test statistic you accept the null hypothesis and for which values you reject the null hypothesis and accept he alternative hypothesis.

A type 1 error occurs when an effect is due to chance, but we find it to be significant in the model.

A type 2 error occurs when we ascribe the effect to chance, but the effect is non-coincidental.

Significance level and p-values

A significance level is a probability threshold below which the null hypothesis can be rejected. You must choose the significance level before computing the test statistic. It is usually .01 or .05.

A p-value is the smallest significance level at which the null hypothesis would be rejected. The confidence interval contains the values of the statistic for which we accept the null hypothesis.

Correlations are useful as effects can help predict an outcome, but correlation does not imply causation.

When making recommendations, one should take into consideration confounding variables and the fact that correlation across two variables do not imply that an increase or decrease in one of them will drive an increase or decrease of the other.

Spurious correlations happen in data. They are just coincidences given a particular data sample.