Parameters and Statistics

Manpreet S Katari

Expected Value

Discrete values

$$E(X) = \sum_{x} x f(x)$$

Continuous values

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx.$$

Variance and Standard deviation

Variance quantifies the amount of dispersion of the data.

$$\operatorname{Var}(X) = E[X - E(X)]^2$$

Standard deviation is the square root of variance

$$\operatorname{SD}(X) = \sqrt{\operatorname{Var}(X)}.$$

Consideration for statistical analysis

It is often impossible to make observations of all population values. So we use estimators to estimate the population parameters.

- If the sample is the entire population than it is descriptive statistics and not inferential
- Statistical analysis is necessary to make results of an experiment easier to interpret
- Observations should be obtained independently from the population
- All statistical analysis should be accompanied with graphical representation.

Types of Estimators

Point estimators to predict a specific value of a parameter

Interval estimators estimate the bounds of a parameter.

Arithmetic Mean

Sum of all variables and divide by the number of

Values summed.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Geometric Mean

This is useful when values are part of a log normal distribution or are dealing with rates. It is the n^{th} root of the product of n values.

$$GM = \sqrt[n]{\prod_{i=1}^{n} X_i}.$$

Harmonic Mean

The reciprocal of the mean of a set of reciprocal values.

Useful for calculating values such as average velocity or in population genetics.

The harmonic mean of population sizes from non-overlapping generations can be used to estimate effective population size.

$$\mathrm{HM} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{X_i} \right)^{-1}$$

Trimmed Mean

Arithmetic mean of a portion of the data.

This is to be used with caution - you are ignoring real values that were obtained from your experiment.

Portion of the data to use for calculation mean is : $1 - 2\kappa$

$\kappa = 0$ -> arithmetic mean

$\kappa = 0.5$ -> median

Mode and Median

Mode is the most frequently obtained value

Median is the 50th percentile of the pdf.

Sample Variance and sample standard deviation

The sum of squared deviations divided by n-1.

The n-1 is for degrees of freedom. It provides the number of independent variables that are able to estimate the parameter. It's -1 because we know the mean of the sample.

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}.$$

Sample standard deviation is simply the square root of the variance

Sample coefficient of variation

It is simply scaling standard deviation by dividing it with the arithmetic mean.

$$\frac{\mathrm{SD}(X)}{E(X)}$$

Interquartile range

The quantile function in R gives you the opposite of the density function.

Given a probability, it will give you the quantile or the value where that probability is likely.

Quantiles are usually divided into 10s and Quartiles are in quarter.

First and third quartile are 25% and 75% of the values sorted.

So the IQR is the range of values from 25%-75% of the data.