Introduction to Statistics and Probability

Probability : Contingency Tables

Kevin O'Brien

Spring 2014

Correlation

- Correlation is a measure of the relation between two or more variables.
- Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negative correlation while a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents a lack of correlation.

Correlation

- The most widely-used type of correlation coefficient is Pearson r, also called linear or product- moment correlation. The pearson correlation coefficient is a metric.

Correlation

- Two variables that have no linear relationship have a correlation close to zero.
- Scatter plots are a useful way of determing the likely relationship between two variables.
- The Pearson correlation coefficient is most commonly used estimate for correlation.
- Other types of correlation are tbe ***Spearman Rho*** and the ***Kendal Tau*** correlation coefficients.

Correlation

- Correlation is a measure of strength of **Linear Relationship** between two variables.
- The Pearson correlation coefficient (denoted r) is the most comonly used statistical estimate for correlation.
- Correlation estimates are defined to be between -1 and 1. It is not possible to have a correlation value outside this range of values

$$-1 \leq r \leq 1$$

- Additionally correlation estimates are not denominated in any units. (Contrast this to standard deviation, which is denominated in the same units as the mean).

Correlation

- A strong positive linear relationship describes a relationship between two variables whereby an increase in one variable will closely coincide with an increase in the other variable.
- Conversely a strong negative linear relationship describes a relationship whereby an increase in one variable closely coincides with a decrease in the other.

Correlation

- The Pearson correlation estimate, which us based on sample data, is denoted r (although related metrics use capital R).
- This measure is used as an estimate for the Population correlation, denoted by the greek letter ρ (pronounced "Rho"). The estimate is computed using summation identities.

Outliers

Outliers can greatly influence the computed value of an estimate. Correlation is closely related to Simple linear regression models, in that both are concerned with the linear relationship between variables. However Linear Regression has a different emphasis. Simple Linear Regression describes one independent variable (IV) and the response of the dependent variable (DV).

Correlation and Causality

Implicit is simple linear regression is the notion of causality. The dependent variable changes as the independent variable changes. The converse is not true. ¡some examples : hot temperature / ice cream example¿ . Correlation is not concerned with causality at all, hence the often used expression ”causation does not imply causality ”.