

Statistical Decision Making Cheatsheet

Broderick Westrope

April 13, 2022

Contents

1	Collecting Data	2
2	Describing Data	4
3	Hypothesis Tests	9
4	Hypothesis Tests	10

1 Collecting Data

Data

Data consists of a set of cases (rows) and for each case, we measure a set of variables (columns). Each variable is either categorical or quantitative.

Quantitative Variables

Quantitative variables contain numerical values, where numerical operations such as addition or computing the mean make sense (E.g. number of days before hard drive failed).

Categorical Variables

Categorical variables are limited to a set of categories (groups); the recorded value from each case must be one of the categories (E.g. gender).

Explanatory & Response Variables

When asking "Does the knowledge of variable X help to explain or predict the variable Y ?" we call X the explanatory variable and Y the response variable, since X explains Y and Y responds to the value of X .

Sampling Bias

Sampling bias occurs when the sampling method causes the value of the variable of interest in the sample to be different from its value in the population.

To avoid bias we ensure each object/case in the population has an equal chance of being selected as a case in the sample.

Association

Variables X and Y are associated if the observed values of X are related to the observed values of Y (For whatever reason).

Causation

Variables X and Y are causally associated if changing the value of X causes the value of Y to change.

To infer causation, you must control the explanatory variable, and see how the response variable changes.

Confounding Variable

A third variable that is associated to both the explanatory and response variables is called a confounding variable. Other names are confounding factor or lurking variable. A confounding variable can offer a plausible explanation for the association between two variables

You cannot assume that there is no confounding variable, just because you cannot think of one.

Observational Study

An observational study is a study where the variables of interest are observed, but not controlled, by the researcher. Causation cannot be identified in an observational study.

Randomised Experiment

In a randomized experiment, the values of the explanatory variables are determined randomly before the response variable is measured.

2 Describing Data

Proportion

Instead of absolute counts, we tend to measure the proportion of a category.

Population proportion = p

Sample proportion = \hat{p}

Parameter

A parameter is a number describing some aspect of the population. The value of a parameter for a population is constant. (E.g. population proportion p)

Statistic

A statistic is a number describing some aspect of a sample. The value of a statistic may be different for each sample. (E.g. sample proportion \hat{p})

Visualising Variables

Visualising One Categorical Variable

A frequency table for a single categorical variable is a table that lists for each category how often it occurs.

A frequency table of proportions is called a relative frequency table. Obviously, the relative frequencies always add to 1.

We can also present the values of a frequency table using a bar chart (best for absolute values) or pie chart (best for relative frequency).

Visualising Two Categorical Variables

We can summarize the relationship between two categorical variables using a two-way table, which consists of two frequency/relative frequency tables (one on each axis) intersecting.

We can also present the values of a two-way table as either a stacked bar chart or side-by-side bar chart.

Visualising One Quantitative Variable

A histogram divides the range of a quantitative variable into several bins; each bin is visualised by a bar whose height represents the number of cases in the bin. This is essentially turning a quantitative variable into a categorical variable, where the number of variables/bins is range/bin-size. A box plot can also be used to visualize a single variable's five number summary, IQR, and its outliers.

Visualising Two Quantitative Variables

A scatter plot is used to visualize two quantitative variables and see the correlation between the two, as well as the spread.

One Categorical & One Quantitative

A side-by-side box plot allows us to see the box plot values for two different categories and easily compare their IQR, five number summary, and outliers.

Histogram Distribution

A histogram can be:

- Left skewed, symmetric, or right skewed, where the skew direction is the side with a longer tail.
- Bell-shaped

Mean

The mean is the average position of the data on the x -axis. We can use the mean to measure the centre of a distribution of quantitative values.

$$\text{Population mean} = \mu = \frac{\sum_{i=1}^N x_i}{N}$$

$$\text{Sample mean} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

where N is the population size and n is the sample size.

Median

The median is the point on the x -axis that splits the data into two equal sets: half the data are smaller than the median, and half the data are greater than the median. To compute the median you must order the data; the median is the “middle number” in the ordered data set.

$$\text{Median}(X) = \begin{cases} X[\frac{n}{2}] & \text{if } n \text{ is even} \\ \frac{(X[\frac{n-1}{2}] + X[\frac{n+1}{2}])}{2} & \text{if } n \text{ is odd} \end{cases}$$

where X is the ordered list of values in the data set and n is the number of values in the list X .

Standard Deviation

The standard deviation gives us an indication of the width of a distribution; it can be thought of as the typical distance of a data point from the mean.

$$\text{Population standard deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

95% Rule

If a distribution is approximately symmetric and bell shaped, then about 95% of the data fall within two standard deviations of the mean. That is, approximately 95% of the data will be in the interval:

$$\text{Population: } (\mu - 2\sigma, \mu + 2\sigma)$$

$$\text{Sample: } (\bar{x} - 2s, \bar{x} + 2s)$$

z -Score

The z -score of a quantitative value shows us the number of standard deviations the value is from the mean and on which side of the mean it lies. A z -score measures how extreme a value is. The z -score of a value x is:

$$\begin{aligned}\text{Population: } z &= \frac{x - \mu}{\sigma} \\ \text{Sample: } z &= \frac{x - \bar{x}}{s}\end{aligned}$$

5 Number Summary

The 5 Number Summary contains five statistics which divide the sample into four equal portions, giving a good indication of the position and spread of the data. The 5 Number Summary contains:

- Minimum Value
- Q_1
- Median
- Q_3
- Maximum Value

where Q_1 and Q_3 are the first and third quartiles.

Percentiles & Quartiles

The p -th percentile is the number on the x -axis for which $p\%$ of the sample are smaller (and $(100 - p)\%$ of the sample are larger). A few percentiles have special names:

- The 25-th percentile is called the first quartile (Q_1).
- The 50-th percentile is the median.
- The 75-th percentile is called the third quartile (Q_3).

Range

The range is the distance separating the maximum and minimum values in the sample:
 $range = max - min$.

Interquartile Range

The interquartile range (IQR) is the distance separating the third and first quartiles in the sample: $IQR = Q_3 - Q_1$.

Outliers

An outlier is a point that seems unusual when compared to the rest of the sample. Determining whether a point is an outlier requires that we understand the data.

Outlier General Rule

A general rule is that a value is an outlier if it is smaller than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$.

Correlation

The correlation (short for Pearson Correlation Coefficient) is a measure of the strength and direction of linear association between two quantitative variables. The correlation is between -1 and 1 where the closer to 0 the less there is a linear correlation, and the sign represents the direction of the correlation.

The correlation is defined as the mean product of the z-scores of the two variables:

$$\text{Population correlation} = \rho = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\Sigma(x_i - \mu_x)^2 \Sigma(y_i - \mu_y)^2}}$$

$$\text{Sample correlation} = r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

where x_i are the values of the x -variable, and y_i are the values of the y -variable.

3 Hypothesis Tests

Statistical Test

A test that uses sample data to attempt to answer a question about the corresponding population.

Null & Alternative Hypotheses

The null hypothesis and alternative hypothesis are competing claims about the data within a population which needs to be tested using the corresponding sample data. The null hypothesis is our default and is of the form of an equality. Whereas, the alternative hypothesis is what we are trying to prove is true and is of the form of a type of inequality ($<$, \neq , $>$). For example when trying to prove a coin-toss is unfair:

$$\text{Null Hypothesis} = H_0 : p = 0.5$$

$$\text{Alternative Hypothesis} = H_a : p > 0.5$$

where p is the proportion of heads (tails) for example.

4 Hypothesis Tests

Lorem Ipsum