# Statistical Decision Making Cheatshet

Broderick Westrope

April 26, 2022

## Contents

# 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

**Poportion**

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

$$\text{Population proportion} = p$$
$$\text{Sample proportion} = \hat{p}$$

**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.

$$\text{Population mean} = \mu = \frac{\Sigma_{i=1}^{N} x_i}{N}$$
$$\text{Sample mean} = \bar{x} = \frac{\Sigma_{i=1}^{n} x_i}{n}$$

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.

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

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.

$$\text{Population standard deviation} = \sigma = \sqrt{\frac{\Sigma_{i=1}^{N}(x_i - \mu)^2}{N}}$$
$$\text{Sample standard deviation} = s = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

## 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:

$$\text{Population: } z = \frac{x - \mu}{\sigma}$$

$$\text{Sample: } z = \frac{x - \bar{x}}{s}$$

### 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.

**Statistical Significance**

A sample is more statistically significant the less likely it is to happen under the assumption that $H_0$ is true. That is, the rarer it is the more important it is as evidence for $H_A$ and against $H_0$.

**Randomization Distribution (RD)**

Similar to a bootstrap distribution but for operations such as calculating a $p$-value. It approximates the sampling distribution, by simulating samples of the original experiment, using a random process that follows the original data collection as closely as possible, but enforces the null hypothesis.
The closer an observed value is to being an outlier amongst the simulated values, the more statistically significant it is.

### $p$-value

The $p$-value (of a sample set in a statistical test) estimates the probability of getting a result/sample at-least as extreme as the observed sample if we assume $H_0$ is true. This is often tested by creating a randomization distribution which **must** follow the assumption that $H_0$ is true. It can be thought of as the probability that the next generated sample will be as extreme as the observed sample, meaning a low enough $p$-value shows the observed sample is more rare and hence is evidence for $H_A$ and against $H_0$.
**NOTE:** When $H_A$ is two-tailed we must double the $p$-value to account for each tail.

#### Computing $p$-values

Assuming $H_0 : \pi = a$ for some parameter $\pi$ and an observed statistic $o$. $\mathbb{P} < (x)$ denotes the probability of being less-than $x$. $H_A$ is calculated differently based on its tail:

- Left-Tailed ($H_A : \pi < a$): The $p$-value for $o$ is $\mathbb{P} \leq (o)$.
- Right-Tailed ($H_A : \pi < a$): The $p$-value for $o$ is $\mathbb{P} \geq (o) = 1 - \mathbb{P} < (o)$.
- Two-Tailed ($H_A : \pi < a$): We double the smaller tail:
  - When $o < a$: $p$-value for $o$ is $2 \times \mathbb{P} \leq (o)$.
  - When $o > a$: $p$-value for $o$ is $2 \times \mathbb{P} \geq (o) = 2 \times (1 - \mathbb{P} < (o))$.

### Significance Level

The significance level is a threshold we define **before** looking at data, and if the $p$-value is below this threshold then it is enough evidence to reject $H_0$ and accept $H_A$. It is usually denoted by $\alpha$. Common levels are $\alpha = 0.05, \alpha = 0.01, \alpha = 0.1$.

### Errors

Type I error is when we incorrectly reject $H_0$ for $H_A$ (this is avoided with a lower $\alpha$).
Type II error is when we incorrectly accept $H_0$ (this is avoided with a higher $\alpha$).
Since Type II just states "the results were not strong enough to contradict our null/default hypothesis" it is (generally) less important to avoid these over Type I. In order to choose a significance level, you must consider how serious making each error would be.

### Methods for Generating a Randomization Distribution

Common ways to generate a randomization distribution:

- Testing a proportion: We sample from a population with the null proportion using the sample size in the original sample.
- Testing a mean: To keep the variability of the simulated samples the same as the original sample, we shift the original sample so that the mean of the shifted values is at the null mean. We sample with replacement from these shifted values to obtain a sample with the same sample size as the original sample.
- Testing a difference in means: To match the null assumption of no difference between the two groups, we deal all the values randomly to the two groups, matching the two sample sizes in the original sample.
- Testing a difference in proportions: To match the null assumption of no difference between the two groups, we deal all the values randomly to the two groups, matching the two sample sizes in the original sample.
- Testing a correlation: To match a null hypothesis of zero correlation, we randomly assign the values of one variable to the values of the other variable.

# 4 Hypothesis Tests

Lorem Ipsum