Introduction to Statistics

Alessia Mondolo

July 18, 2019

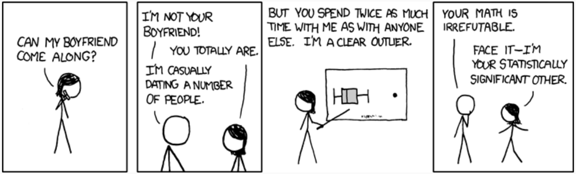
Academy AI

Formalities

Statistics Definition

Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

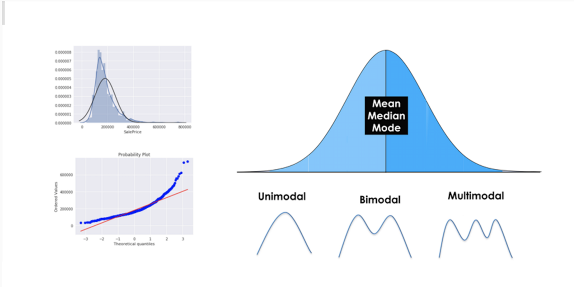
Statistical analysis – used to manipulate summarize, and investigate data, so that useful decision-making information results.


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

In other words: **There are two ways of lying, the first one by not telling the truth or the second one by making up statistics -or hiding them-.**

Descriptive statistics

Methods of organizing, summarizing, and presenting data in an informative way. In order words, **how is your data**.

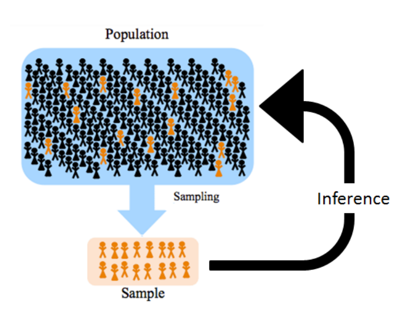


Unintended pun: Always treat your data very well or it will be mean to you.

Inferential Statistics

The methods used to determine something about a population on the basis of a sample.

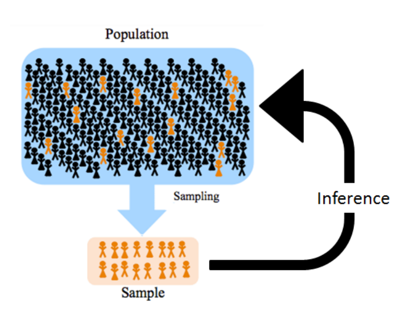
1. **Population –** The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
2. **Sample –** A portion, or part, of the population of interest



Inferential Statistics

Inference is the process of drawing conclusions or making decisions about a population based on sample results

1. **Estimation** – Estimate the population mean weight using the sample mean weight
2. **Hypothesis testing** – e.g., Test the claim that the population mean weight is 70 kg



Sampling

A sample should have the same characteristics as the population it is representing. Because of this, **randomness in the sampling is often very important**. Sampling can be:

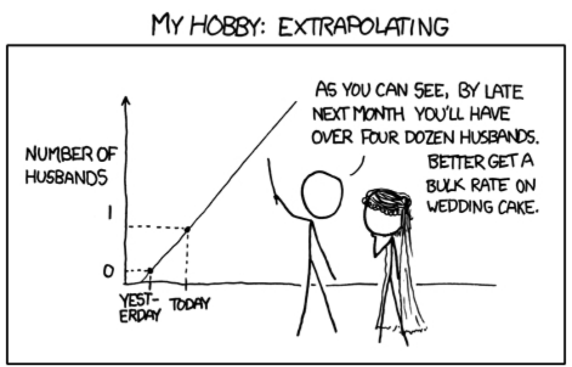
1. **With replacement –** a member of the population may be chosen more than once (e.g. picking the candy from the bowl)
2. **Without replacement –** a member of the population may be chosen only once (e.g. lottery ticket)

Sampling methods can be:

1. **Random –** each member of the population has an equal chance of being selected
2. **Non-random –** a member of the population may be chosen only once (e.g. lottery ticket)

The actual process of sampling causes errors, some common one's include a sample size that is **not large enough** or representative of the population.

Caution: Sampling and Extrapolation


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

Statistical Features

Central Value

A fundamental concept in summary statistics is that of a **central value** for a **set of observations** and the extent to which the central value characterizes the whole set of data.

Center measurement is a summary measure of the overall level of a dataset.

Commonly used methods are **mean, median, mode, geometric mean** etc.

Methods of central measurement

Measures of central value such as the mean or median must be coupled with measures of **data dispersion** (e.g., average distance from the mean) to indicate **how well the central value characterizes the data** as a whole.

To understand how well a central value characterizes a set of observations, let us consider the following two sets of data:

A:	70	80	90
B:	60	80	100

Table 1: Spain average life expectation sampling

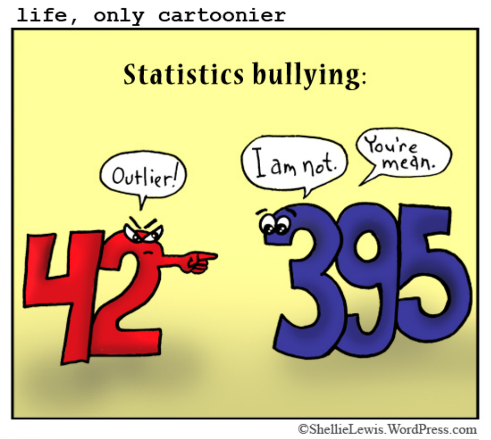
While both A and B have a mean of 80 years old, the distance between observations in A is shorter. In this case, A is a better representation of the data. **If those were samples from two different countries, where would you rather live?**

Methods of Central Measurement

Sample: *9, 3, 7, 7, 5*

- **Mean:** Summing up all the observation and dividing by number of observations. Mean of the sample is $(9+1+7+7+5)/5 = 5.8$
- **Median**: The middle value in an ordered sequence of observations. Ordered sample: 1, 5, **7**, 7, 9 = 7
- **Mode**: The value that is observed most frequently. Mode(Sample) = 7

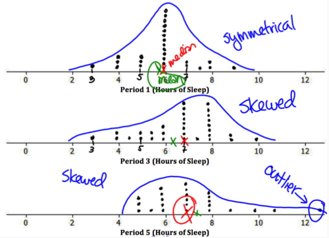
Statistical Bullying



Mean or Median?

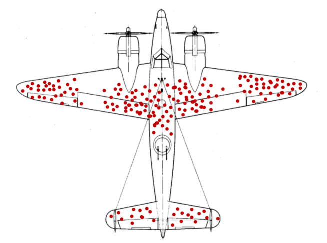
Generally, we pick the Mean, but the **Median is less sensitive to outliers** (extreme scores) than the mean and thus a better measure than the mean for **highly skewed distributions**, e.g. family income.

For example mean of 20, 30, 40, and 990 is (20+30+40+990)/4 =270. The **median** of these four observations is (30+40)/2 = **35**. Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.



Bias

Statistical bias is a feature of a statistical technique or of its results where the **expected value of the results** differs from the **true underlying quantitative parameter** being estimated.

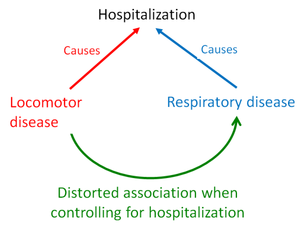


For example, a statistical analysis of American Planes coming back from fights with the Japanese pilots in WW2 was performed. What are the most critical parts of the plane? In this case, we're doing a selection bias since we're sistematically excluding the planes that never made it back

Types of Bias I

1. **Selection bias** happens when a particular group is more likely to be selected for a study than others. **Survivorship bias** is a subset of this.
2. **Spectrum bias** comes from evaluating diagnostic tests on biased patient samples, leading to an overestimate of the sensitivity and specificity of the test.
3. **Funding bias** may lead to the selection of outcomes, test samples, or test procedures that favor a study's financial sponsor. Ex: *Coca-Cola and the surprising amount of studies indicating that sugar is not a risk for our health.*
4. **Reporting bias** involves a skew in the availability of data, such that observations of a certain kind are more likely to be reported. Ex: *Black people acting suspicious vs white people acting suspicious.*

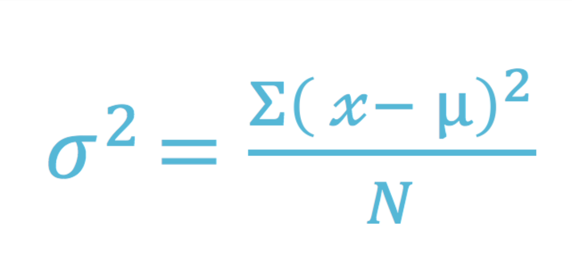
Types of Bias II



1. Analytical bias arises due to the way that the results are evaluated.
2. Exclusion bias arise due to the systematic exclusion of certain individuals from the study. Ex: *Lack of proper drug test studies with women due to interference with pregnancy.*
3. **Recall bias** arises due to differences in the accuracy or completeness of participant recollections of past events. e.g. *How many times do you lie each week?*
4. **Observer bias** arises when the researcher subconsciously influences the experiment due to cognitive bias. E.g. R*esearchers wanting their experiments to be successful.*

Variance

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean.

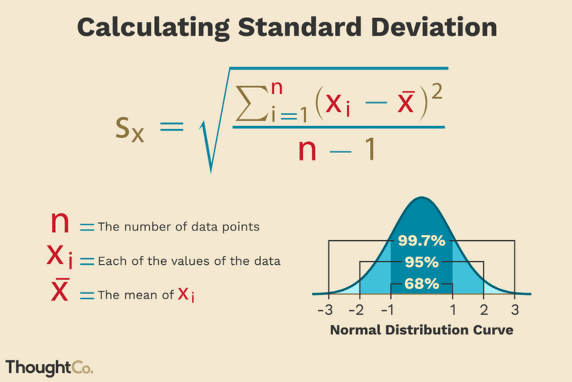


$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

Figure 1: Scary math notation that means the same :)

Standard deviation

The **standard deviation** is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A **low standard deviation** indicates that the **data points** tend to be **close to the mean** of the set, while a **high standard deviation** indicates that the **data points are spread out** over a wider range of values.



Calculating Standard Deviation

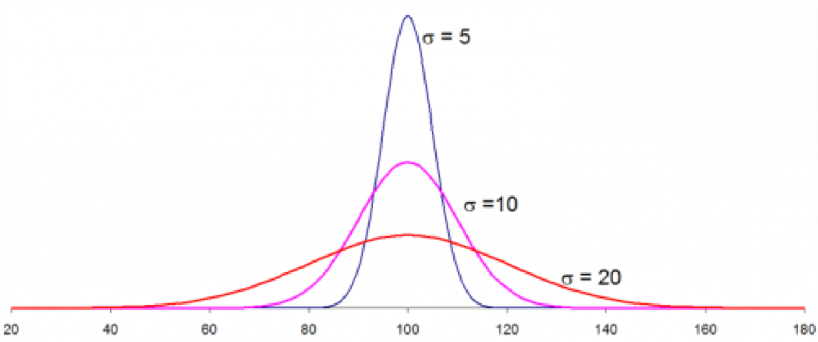
$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

n = The number of data points

x_i = Each of the values of the data

$\bar{x}$ = The mean of x_i

Standard deviation



Quartiles

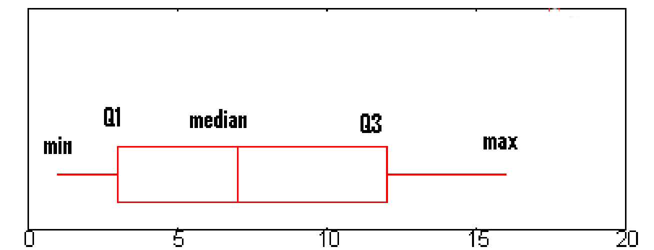
Quantiles: cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.

Quartiles: the three cut points that will divide a dataset into four equal-sized groups:

- First quartile (Q1): the middle number between the smallest number and the median of the data set; it splits off the lowest 25% of data from the highest 75%;
- Second quartile (Q2): the median of the data; it cuts the data set in half;
- Third quartile (Q3): the middle value between the median and the highest value of the data set; it splits off the highest 25% of data from the lowest 75%.

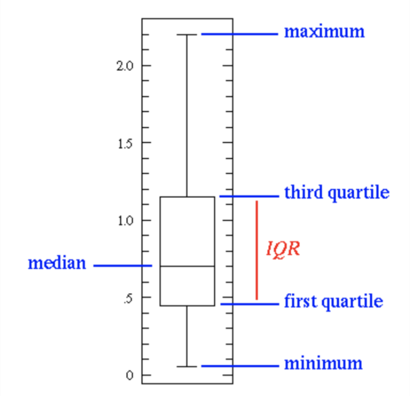
Five Number Summary

Five Number Summary: The five number summary of a distribution consists of the smallest (Minimum) observation, the first quartile (Q1), The median (Q2), the third quartile (Q3), and the largest (Maximum) observation written in order from smallest to largest.



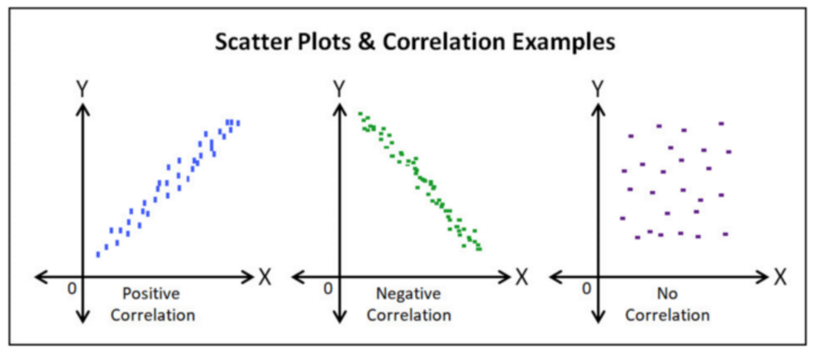
Box plot

Box Plot: A box plot is a **graph** of the **five number summary**. The central box spans the quartiles. A line within the box marks the median. Lines extending above and below the box mark the smallest and the largest observations



Correlation

A correlation is a statistic intended to quantify the strength of the relationship between two variables. The correlation coefficient r quantifies the strength and direction of the linear relationship between two quantitative variables.



Distributions

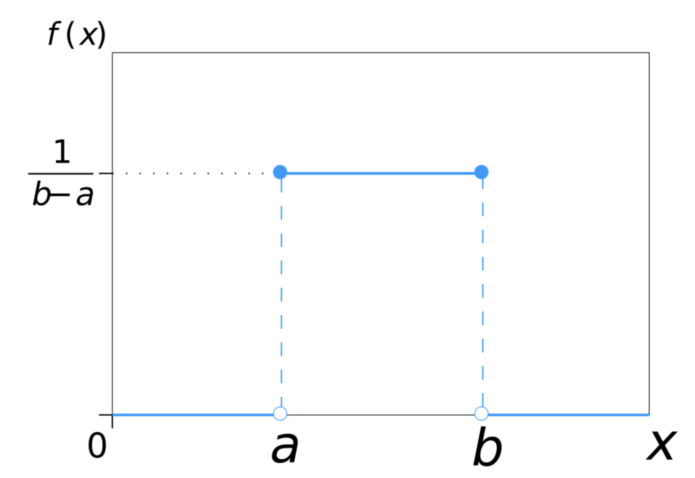
Uniform distribution

The probability distribution function of the continuous uniform distribution is:

$$f(x) = \begin{cases} \frac{1}{b-a} & for\ a \leq x \leq b \\ 0 & for x < a\ or\ x > b \end{cases}$$

Since any interval of numbers of equal width has an equal probability of being observed, the curve describing the distribution is a rectangle, with constant height across the interval and 0 height elsewhere. Since the area under the curve must be equal to 1, the length of the interval determines the height of the curve. The following figure shows a uniform distribution in interval (a,b). Notice since the area needs to be 1. The height is set to 1/(b-a).

Uniform distribution



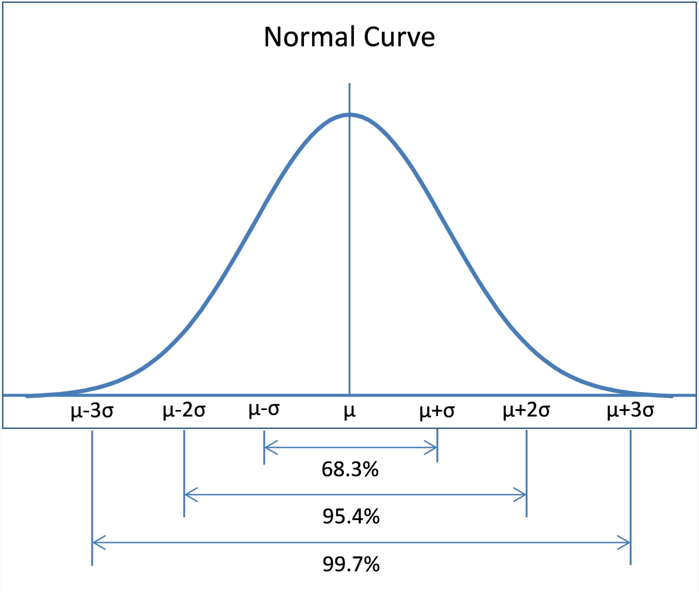
Normal distribution

A normal distribution has a bell-shaped density curve described by its mean μ and standard deviation σ. The density curve is symmetrical, centered about its mean, with its spread determined by its standard deviation showing that data near the mean are more frequent in occurrence than data far from the mean. The probability distribution function of a normal density curve with mean μ and standard deviation σ at a given point x is given by:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A distribution with mean 0 and standard deviation 1 is called a standard normal distribution.

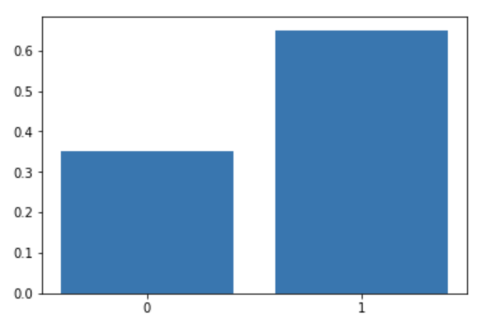
Normal distribution



Bernoulli distribution

Bernoulli distribution is a discrete probability distribution of a random variable which has only two outcomes (”success” or a ”failure”).

For example, probability (p) of scoring a goal in last 10 minutes is 0.35 (success), probability of not scoring a goal in last 10 minutes (failure) is 1 - p = 0.65.


0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1