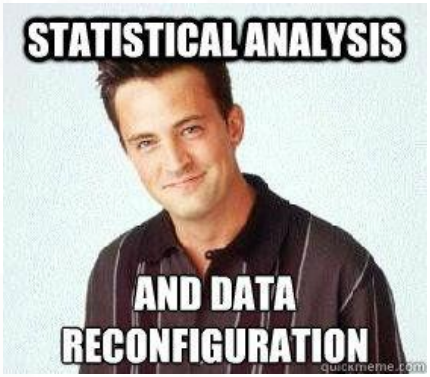
COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

Introduction to Statistical Analysis



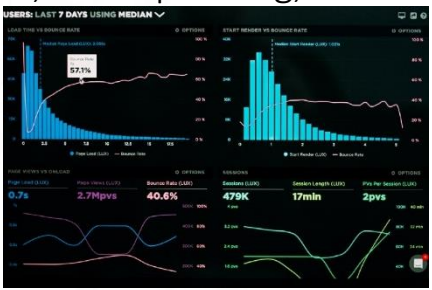
comillas.edu

Contents

Introduction

Statistics, in its broadest sense, refers to a collection of **tools and methods for evaluating, interpreting, displaying, and making decisions based on data**. Some individuals refer to statistics as the mathematical analysis of technical data.

Data scientists should learn statistics because **statistics relate data to the questions** organizations face across all disciplines, such as how to increase revenue, limit spending, etc.



Data analysts use data visualization and statistical methods to **describe dataset characteristics** such as size, quantity, and accuracy to better understand the nature of the data.

Key words

- **Data**: collections of observations (such as measurements, genders and survey responses).
- **Population**: complete collection of all individuals (scores, people, measurements, and so on) to be studied.
- **Sample**: subcollection of members selected from a population.
- **Census**: collection of data from every member of the population.
- **Parameter**: numerical measurement describing some characteristic of a population.
- **Statistic**: is a numerical measurement describing some characteristic of a sample.

Types of statistical analysis

Statistical analysis is the process of collecting and analyzing data to identify patterns and trends, remove bias and inform decision-making.

- **Descriptive statistics**: Summarize the data. Involves summary charts, graphs and tables depicting the data for easier comprehension. Some statistics are the mode, median and mean, as well as range, variance and standard deviation.
- **Inferential statistics**: Take the data from a representative sample and use it to draw larger truths. Statistical inference relies upon estimating uncertainty in predictions. It usually results in estimates, confidence interval and credible intervals.

Types of statistical analysis

Types of statistical analysis:

- **Observational**: Collect data without interfering on how the data arises. Only establish associations between variables.
  - **Retrospective**: uses past data
  - **Prospective**: data are collected throughout the study.
- **Experiment**: Randomly assign subjects to treatments. Establish causal connections.

Types of statistical analysis

Types of statistical analysis:

Study: Breakfast cereal keeps girls slim

USA TODAY
Sept 8, 2005

[...]

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute with funding from the National Institutes of Health (NIH) and cereal-maker General Mills.

[...]

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio, and Maryland who were tracked between the ages of 9 and 19.

[...]

As part of the survey, the girls were asked once a year what they had eaten during the previous three days.

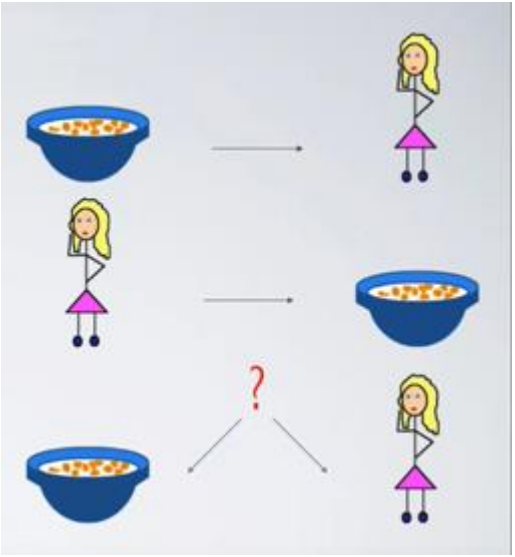
[...]

Types of statistical analysis

1. Eating breakfast causes girls to be slimmer

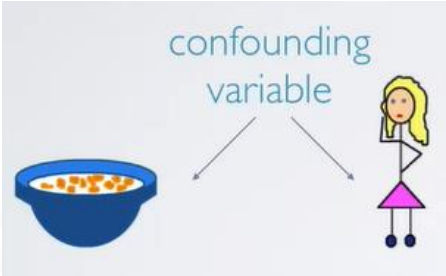
2. Being slim causes girls to eat breakfast

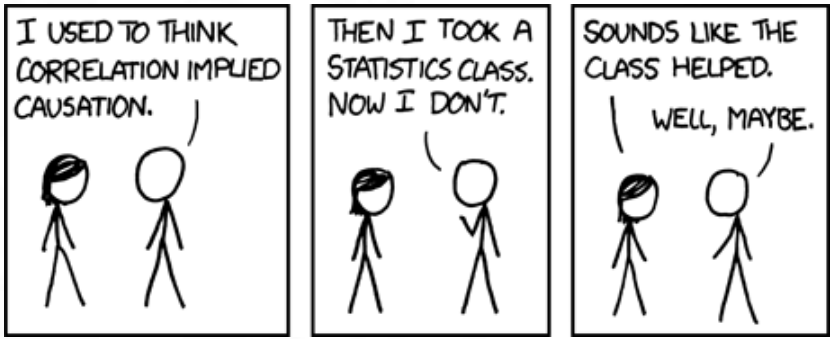
3. Third variable is responsible for both



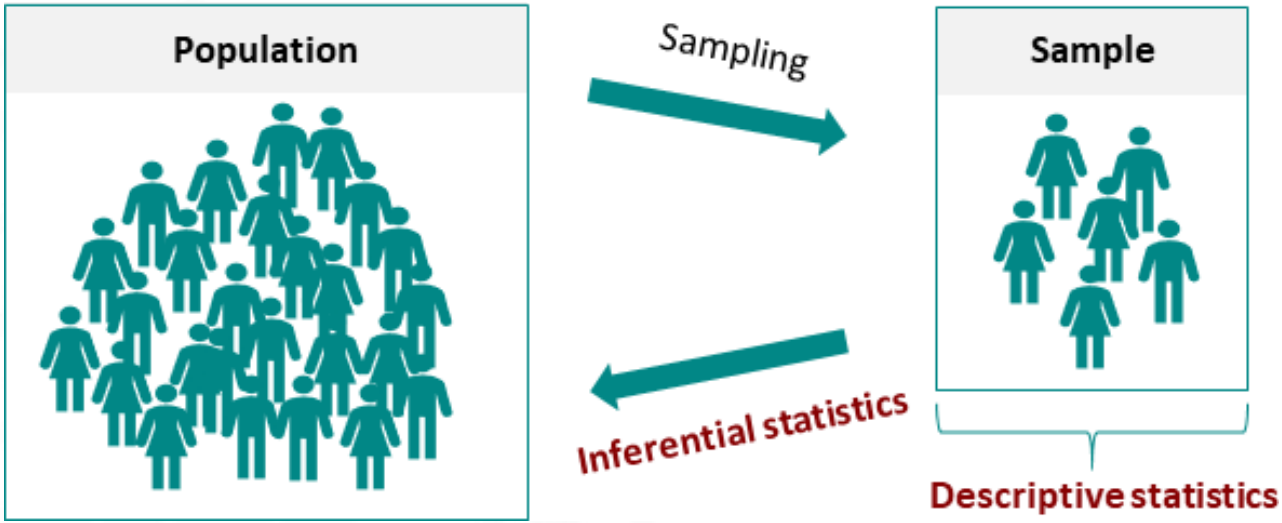
Types of statistical analysis

A confounding variable is an exogenous variable that affect both the explanatory and the response variable, and that make it seem like there is a relationship between them.





Types of statistical analysis



A few sources of sampling bias

- **Convenience sample:** Individuals who are easily accessible are more likely to be included in the sample.
- **Non-response:** If only a (non-random) fraction of the randomly sampled people respond to a survey such that the sample is no longer representative of the population.
- **Voluntary response**: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue

A few sources of sampling bias

A retail store considering updates to their credit card policies randomly samples 1000 of their credit card holders to survey on the phone. The phone calls are made during business hours, therefore there is a lower rate of responses from members who work during these hours. What type of bias is this indicative of?

- Convenience sample
- Voluntary response
- Non-response
- None of the above

A few sources of sampling bias

A retail store considering updates to their credit card policies randomly samples 1000 of their credit card holders to survey on the phone. The phone calls are made during business hours, therefore there is a lower rate of responses from members who work during these hours. What type of bias is this indicative of?

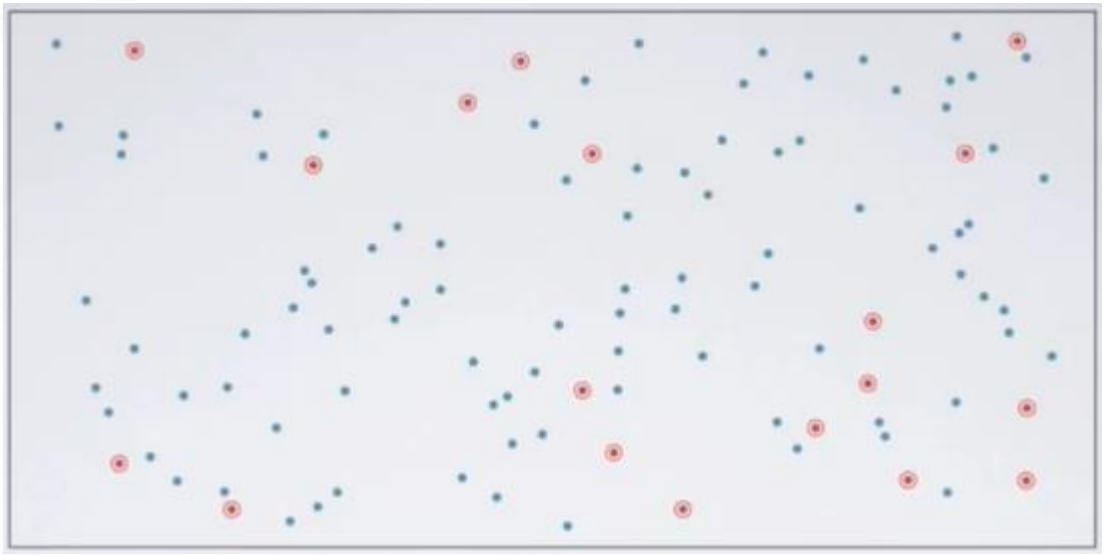
- Convenience sample
- Voluntary response
- Non-response
- None of the above

There is an initial random sample, but not everyone in this random sample is reached. Therefore, the issue is non-response of the sampled individuals.

Types of sampling

Simple Random Sample

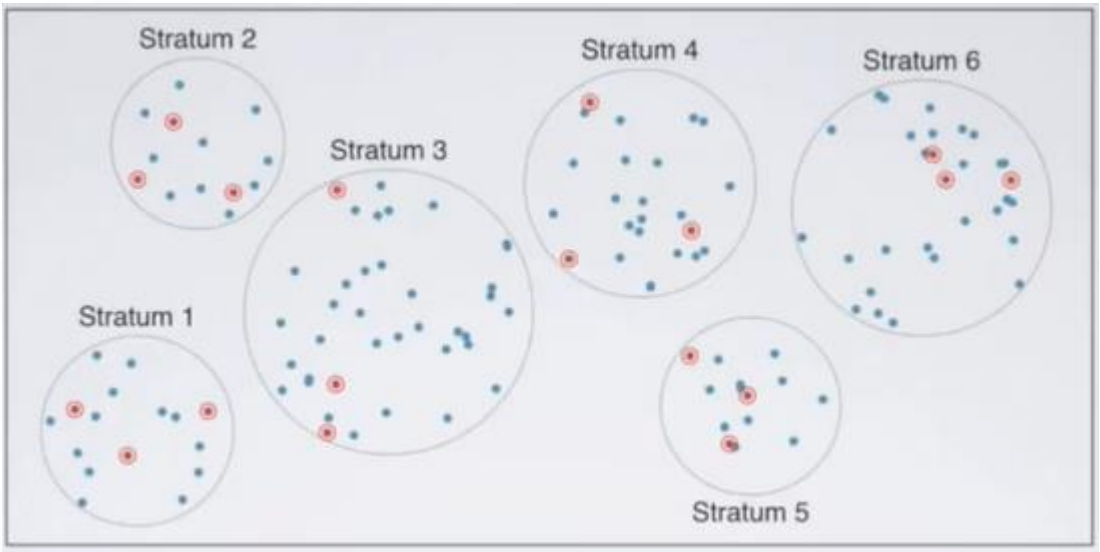
Each case is equally likely to be selected.



Types of sampling

Stratified sample

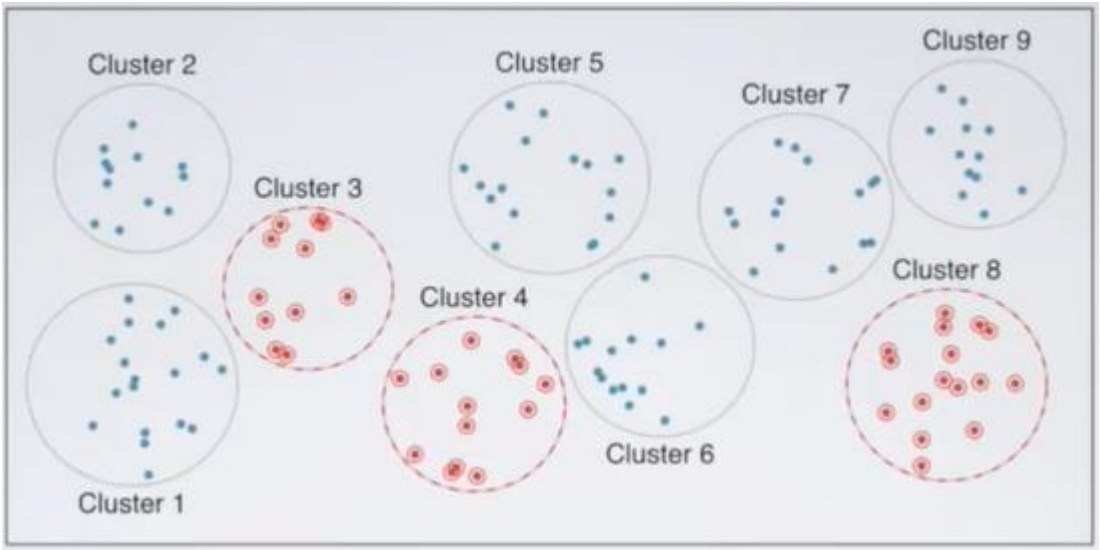
Divide the population into homogenous strata, then SRS from each stratum



Types of sampling

Cluster sample

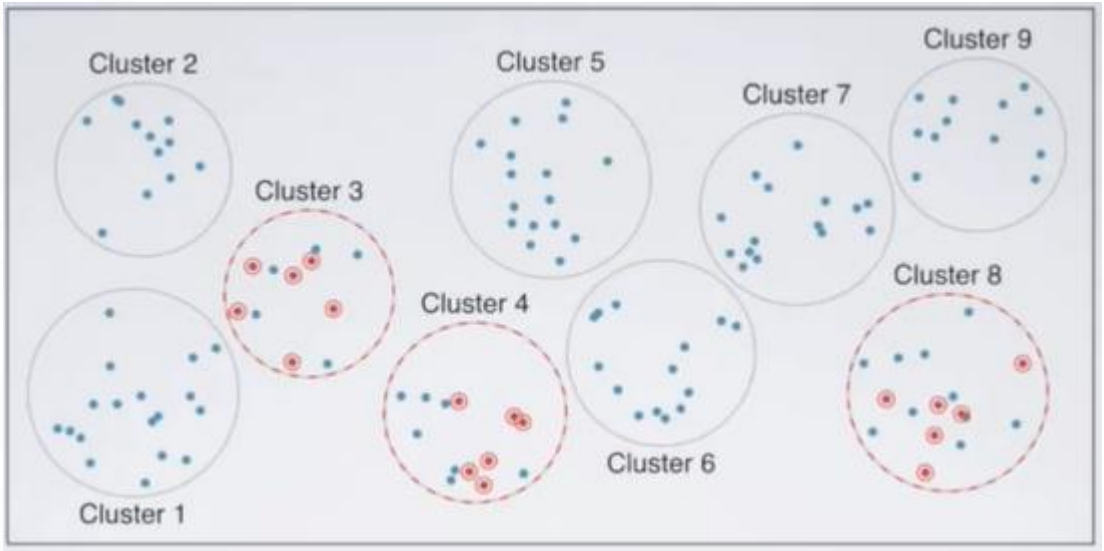
Divide the population clusters, SRS a few cluster, then sample all observations within these clusters.



Types of sampling

Multistage sample

Divide the population clusters, SRS a few cluster, then SRS within these clusters.



Types of sampling

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

- Cluster sampling, where each cluster is a neighborhood
- Stratified sampling, where each stratum is a neighborhood
- Simple random sampling

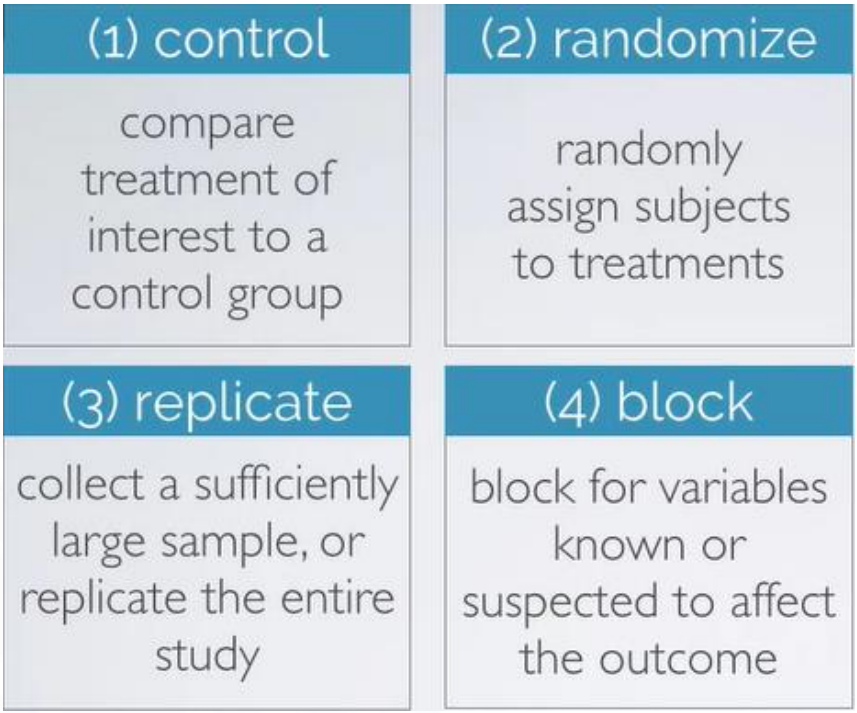
Types of sampling

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

- Cluster sampling, where each cluster is a neighborhood
- Stratified sampling, where each stratum is a neighborhood
- Simple random sampling

As neighborhoods are too different, there would be an enormous number of clusters too different to each others, so basically all samples would be selected.

Principles of experimental design



Blocking vs explanatory variables

- **Explanatory variables** (factors): conditions we can impose on experimental units.
- **Blocking variables**: characteristics that the experimental units come with, that we would like to control for.
- Blocking is like stratifying:
  - Blocking during random assignment
  - Stratifying during random sampling

Types of sampling

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are represented equally under different conditions. Which of the below is correct?

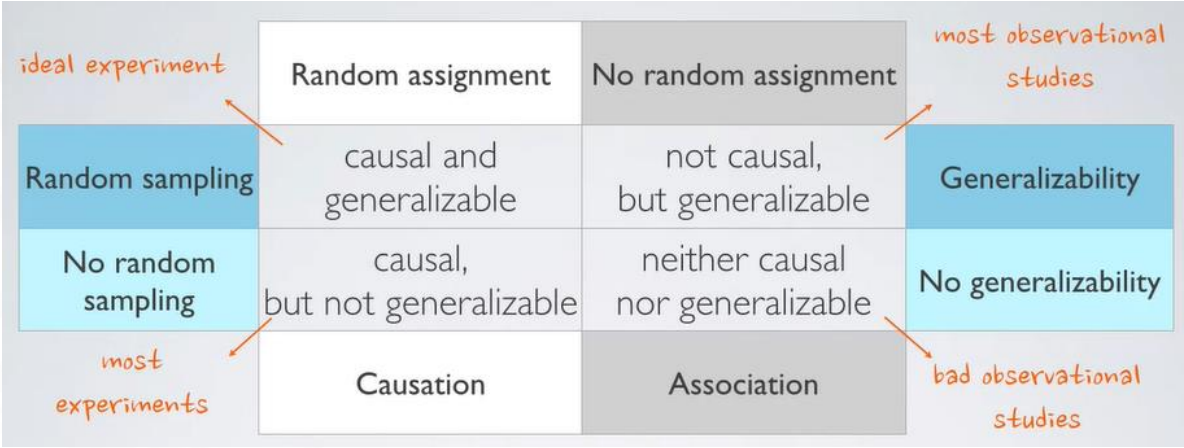
- There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
- There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)
- There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Types of sampling

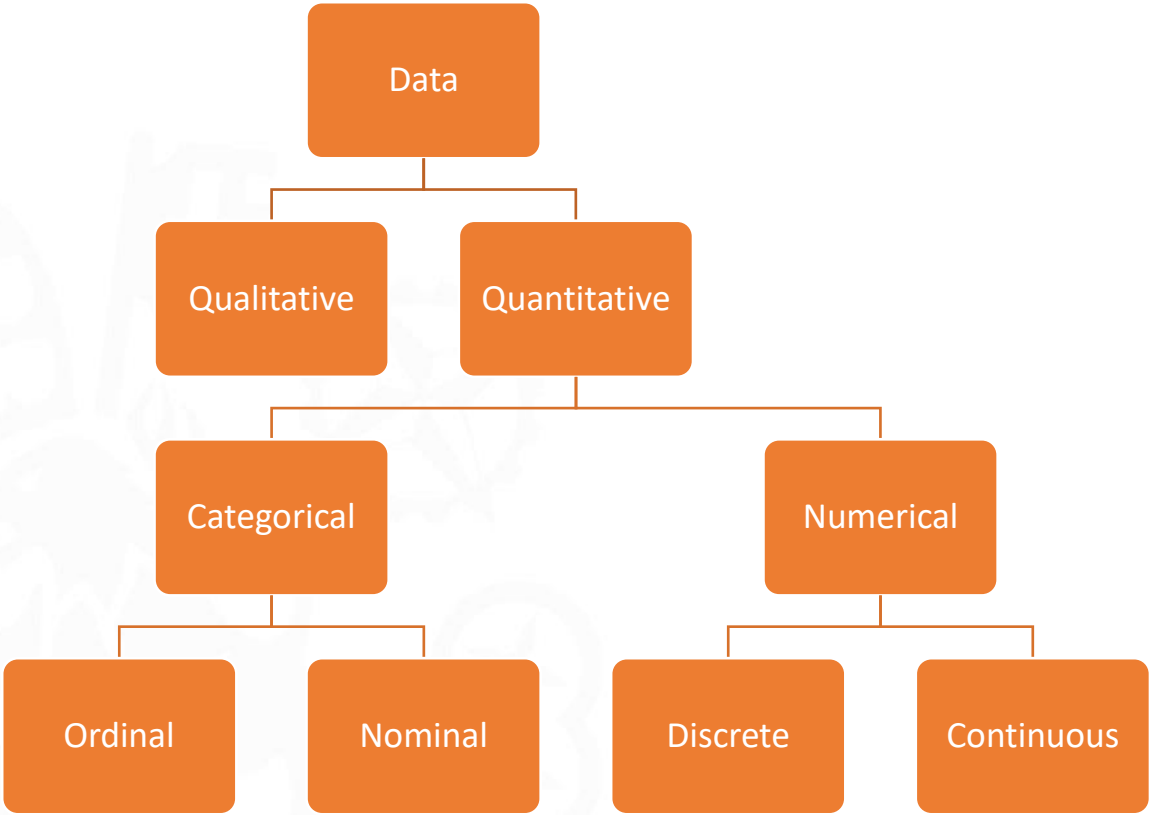
A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are represented equally under different conditions. Which of the below is correct?

- There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
- There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)
- There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Random sampling & random assignment

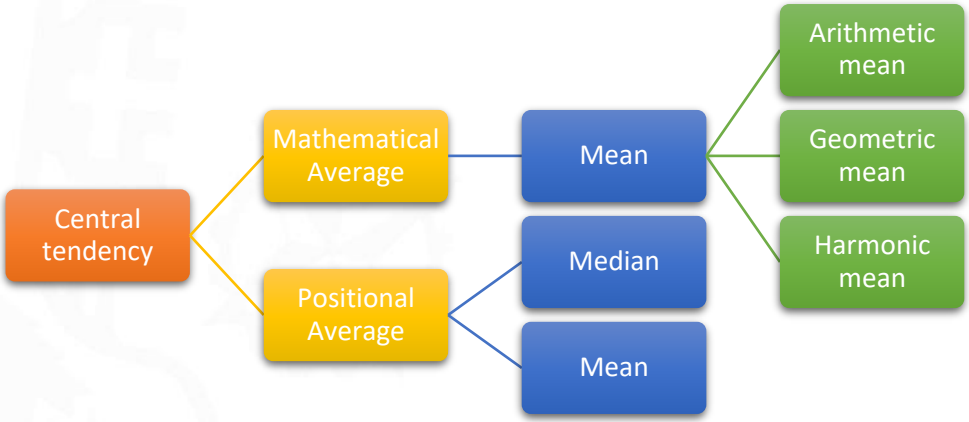


Types of data



Measures of central tendency

A measure of central tendency is a single value that **describes a set of data by identifying the central position within that set of data**. They are also called **measures of central location**.



The **mean**, **median** and **mode** are all valid measures of central tendency, but under different conditions, some measures become more appropriate to use than others.

Measures of central tendency

The **mean** (or average) is the most popular measure of central tendency. It can be used with **both discrete and continuous data**.

$$\bar{x} = \frac{\sum_i x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Where $\bar{x}$ is the **sample mean**, n is the number of values in the sample and x_i are the values of the variable in position i. If instead we look for the population mean, we use "mu":

$$\mu = \frac{\sum_i x_i}{n}$$

The mean is the **most common value in the dataset**. However, the mean is not often one of the actual values that you have observed in your data set. One of its important properties is that it minimizes error in the prediction of any one value in your data set.

Measures of central tendency

The previous definition was that of the **arithmetic mean**. If any values of the sample are more important, then we could calculate the **weighted mean (WM)**:

$$WM = \frac{\sum w \cdot x}{\sum w}$$

When values change exponentially, or sample follows a skewed distribution that can be made symmetrical by a log transformation (Box-Cox), a more appropriate value is the **geometric mean (GM)**:

$$GM = \sqrt[n]{\prod_i x_i}$$

When there are several groups, the average sample size of all the groups can be determined using the **harmonic mean (HM)**:

$$HM = \frac{n}{\sum(\frac{1}{x})}$$

Measures of central tendency

The mean has one main disadvantage: it is particularly susceptible to the **influence of outliers**. For example, consider the wages of staff at a factory below:

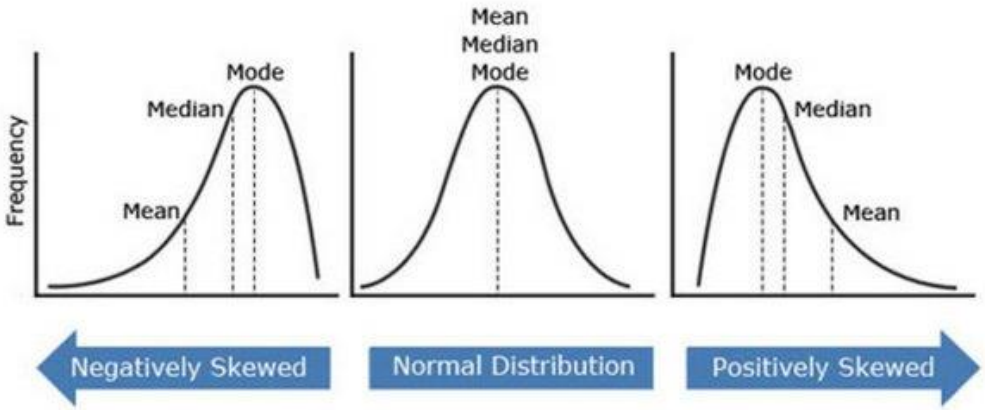
Staff	1	2	3	4	5	6	7	8	9
Salary	15k	18k	16k	14k	15k	15k	12k	17k	95k

The mean salary for these ten staff is $24.1k. However, inspecting the raw data suggests that this **mean value might not be the best way** to accurately reflect the typical salary of a worker.

In this situation, we would like to have a better measure of central tendency. As we will find out later, **taking the median** would be a better measure of central tendency in this situation.

Measures of central tendency

Another time when we usually prefer the median over the mean (or mode) is when **our data is skewed**. As the data becomes skewed, the mean loses its ability to provide the best central location. However, **the median best retains this position** and is not as strongly influenced by the skewed values.



Measures of central tendency

The **median** is the **middle score for a set of data** that has been arranged in order of magnitude. The median is **less affected by outliers and skewed data**.

Suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92

First data must be rearranged by order of magnitude:

14	35	45	55	55	**56**	56	65	87	89	92

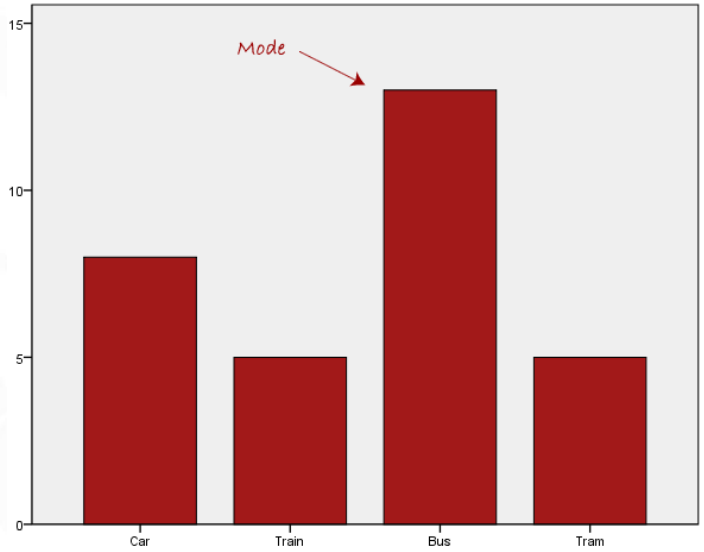
The middle position is the median of the dataset. If there is not an odd number of positions, the **average of the two middle scores** is the median.

14	35	45	55	**55**	**56**	56	65	87	89

Measures of central tendency

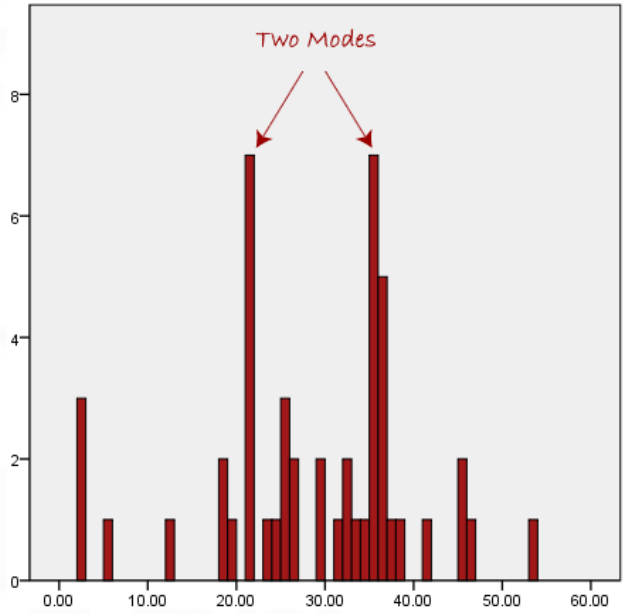
The **mode** is the **most frequent score** in our data set. On a histogram it represents the highest bar. You can sometimes consider the mode as being **the most popular option**.

Normally, the mode is used for categorical data:



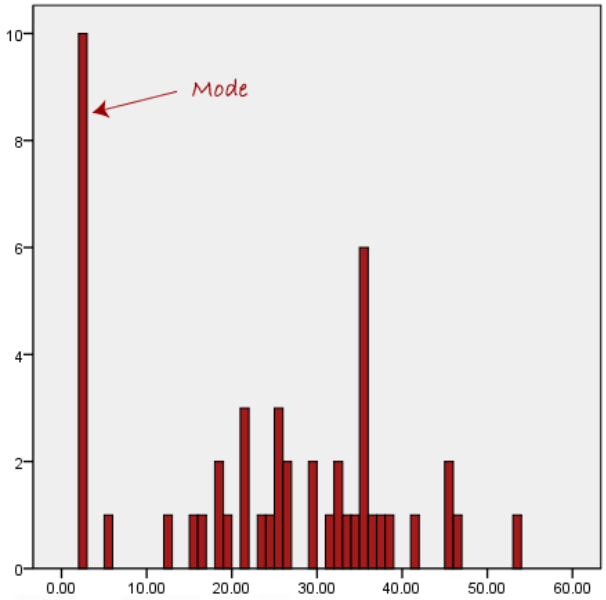
Measures of central tendency

However, one of the problems with the mode is **that it is not unique**, so it leaves us with problems when we have two or more values that share the highest frequency:



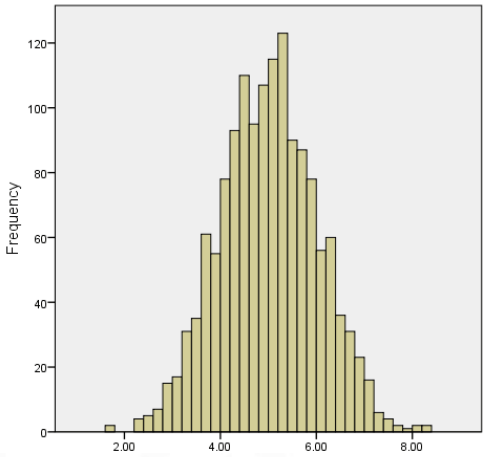
Measures of central tendency

Another problem with the mode is that it will not resemble the central location when the **most common mark is far away from the rest** of the data in the data set



Measures of central tendency

If your dataset follows a normal distribution, both the mean or the median as your measure of central tendency. In fact, in any symmetrical distribution the mean, median and mode are equal.



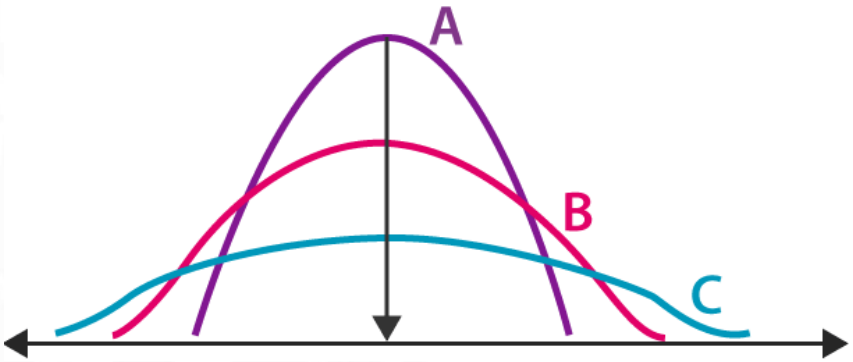
If dealing with a normal distribution, and tests of normality show that the data is non-normal, it is customary to use the median instead of the mean.

Measures of dispersion

A **measure of dispersion**, also called a **measure of spread**, is used to describe the **variability in a sample or population**.

A measure of spread **gives us an idea of how well the measures for central tendency represents the data**. If the spread of values in the data set is large, the mean is not as representative of the data as if the spread of data is small.

The most common measures are **range, quartiles, absolute deviation and standard deviation**.

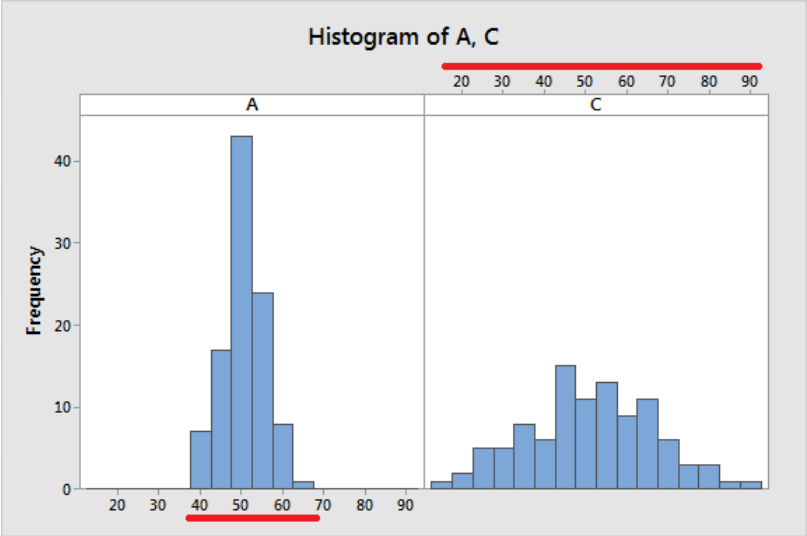


Measures of dispersion

The **range** is the difference between the highest and lowest scores in a data set and is the simplest measure of spread:

$$range(x) = \max(x) - \min(x)$$

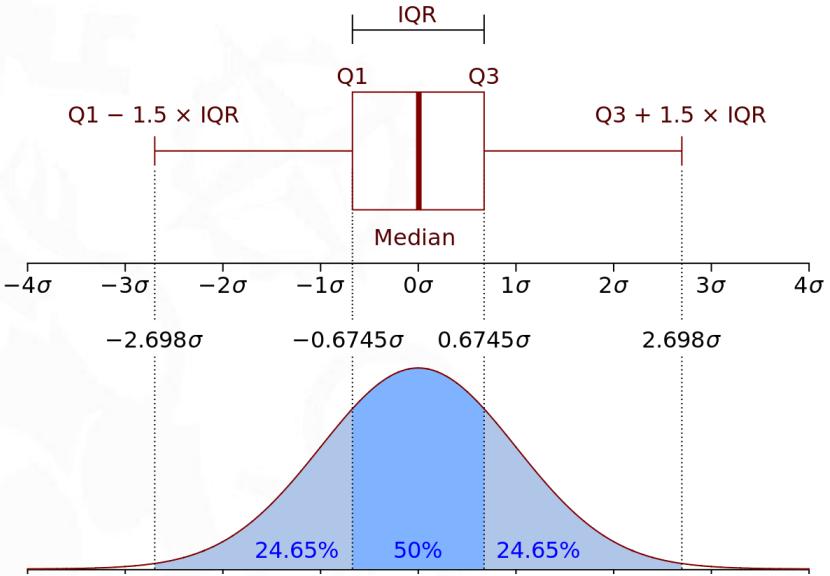
This can be useful if you are measuring a variable that has either a **critical low or high threshold (or both)** that should not be crossed.



Measures of dispersion

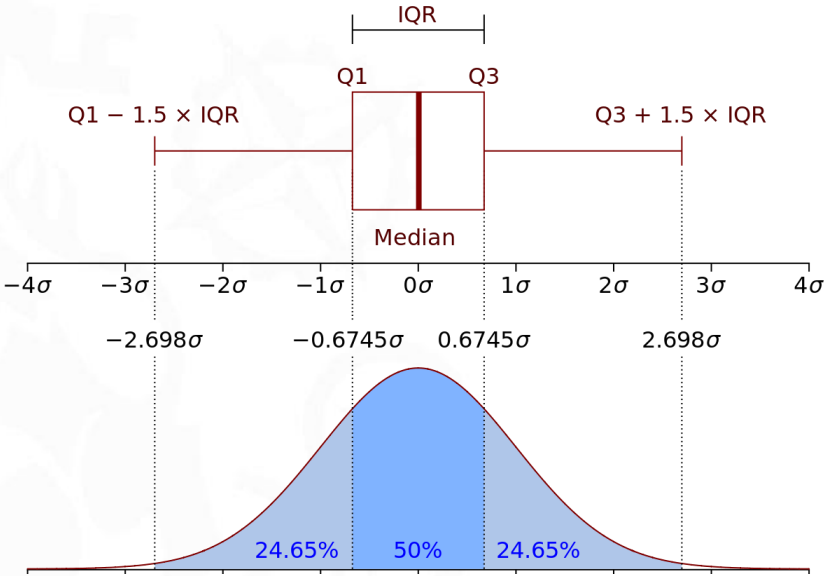
Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.

Quartiles are much **less affected by outliers or a skewed data set** than the equivalent measures of mean and standard deviation.



Measures of dispersion

A common way of expressing quartiles is as an **interquartile range**. The interquartile range describes the **difference between the third quartile (Q3) and the first quartile (Q1)**, telling us about **the range of the middle half of the scores** in the distribution.



Measures of dispersion

Quartiles are useful, but they are also somewhat limited because they do not consider every score in our group of data. To get a more representative idea of spread we need to consider the actual values of each score in a data set. The **absolute deviation**, **variance** and **standard deviation** are such measures.

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

$$|s_x| = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n-1}$$

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

Measures of dispersion

As a measure of variability, the **variance** is useful. If the scores in our group of data are spread out, the variance will be a large number. Conversely, if the scores are spread closely around the mean, the variance will be a smaller number.

However, there are two potential problems with the variance. First, because the deviations of scores from the mean are 'squared', **this gives more weight to extreme scores**. If our data contains outliers this can give undo weight to these scores. Secondly, **the variance is not in the same units as the scores in our data set**: variance is measured in the units squared. This means we cannot place it on our frequency distribution and cannot directly relate its value to the values in our data set.

Measures of dispersion

As a measure of variability, the **variance** is useful. If the scores in our group of data are spread out, the variance will be a large number. Conversely, if the scores are spread closely around the mean, the variance will be a smaller number.

However, there are two potential problems with the variance. First, because the deviations of scores from the mean are 'squared', **this gives more weight to extreme scores**. If our data contains outliers this can give undo weight to these scores. Secondly, **the variance is not in the same units as the scores in our data set**: variance is measured in the units squared. This means we cannot place it on our frequency distribution and cannot directly relate its value to the values in our data set.

Probability & probability distributions

We are normally interested in knowing the **population standard deviation**. Therefore, you would normally calculate the population standard deviation if: (1) you have the entire population or (2) you have a sample of a larger population, but you are only interested in this sample and do not wish to generalize your findings to the population. However, in statistics, we are usually presented with a sample from which we wish to estimate the population standard deviation.

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{n}}$$

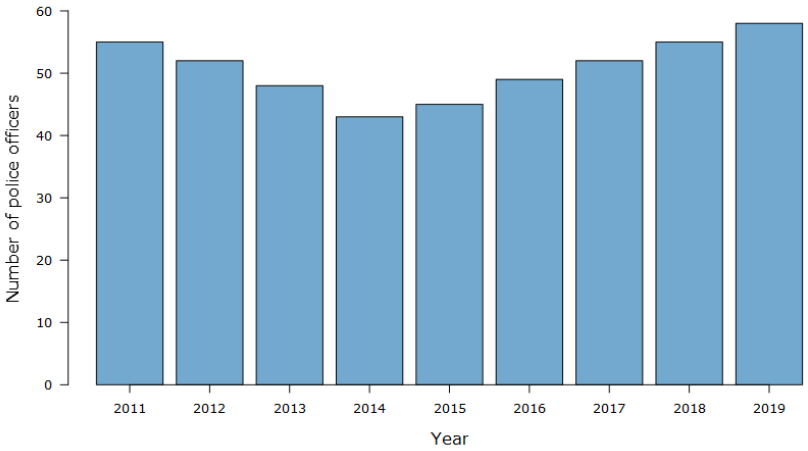
The standard deviation is used in conjunction with the mean to summarize **continuous data**, not categorical data. In addition, the standard deviation, like the mean, is normally only appropriate when the continuous data is not significantly skewed or has outliers.

Graphical representation of data

Bar chart

The important point to note about bar charts is their bar length or height—**the greater their length or height, the greater their value**. Bar charts usually present **categorical variables**, discrete variables or continuous variables grouped in class intervals.

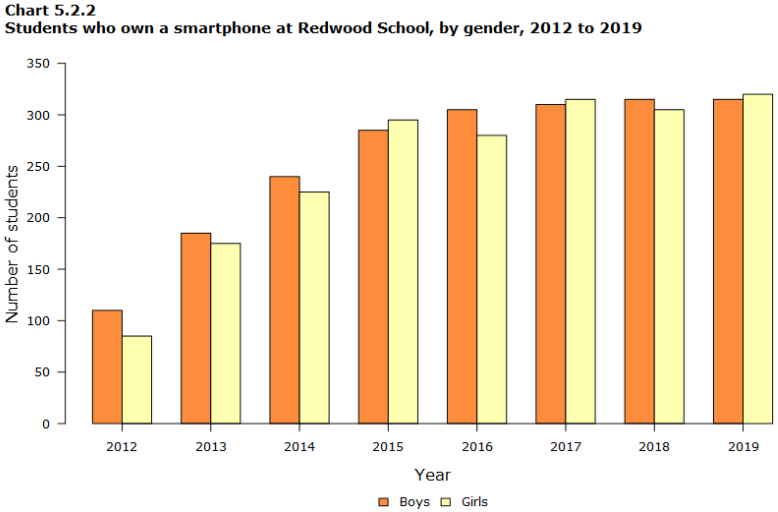
Chart 5.2.1
Number of police officers in Crimeville, 2011 to 2019



Graphical representation of data

Grouped bar chart

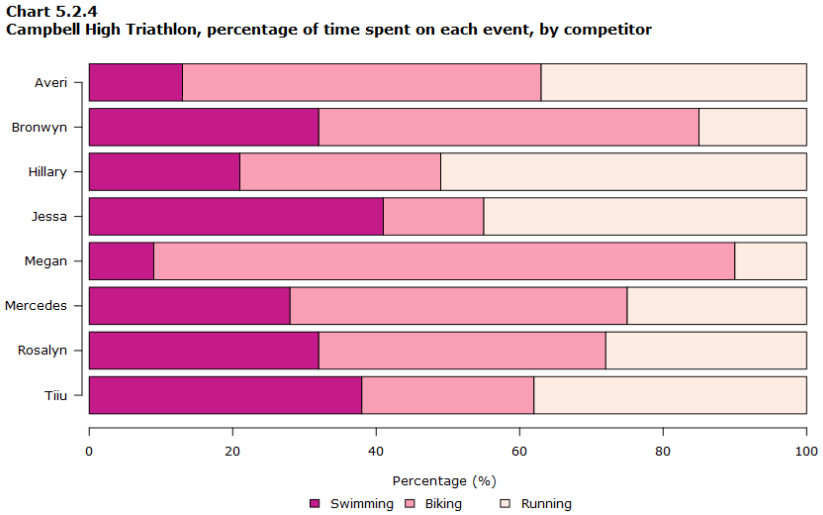
The grouped bar chart is another effective means of comparing sets of data about the same places or items. It gives two or more pieces of information for each item on the x-axis.



Graphical representation of data

Stacked bar chart

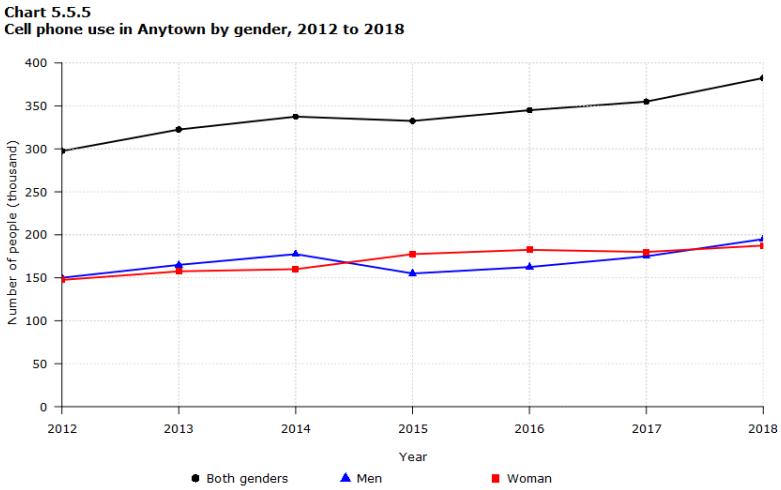
The stacked bar chart is a preliminary data analysis tool used to show segments of totals. The stacked bar chart can be very difficult to analyze if too many items are in each stack. It can contrast values, but not necessarily in the simplest manner.



Graphical representation of data

Line chart

Line charts visual characteristics reveal data trends clearly. A line chart is a visual comparison of how **two numeric variables**—shown on the x- and y-axes—are related or vary with each other.



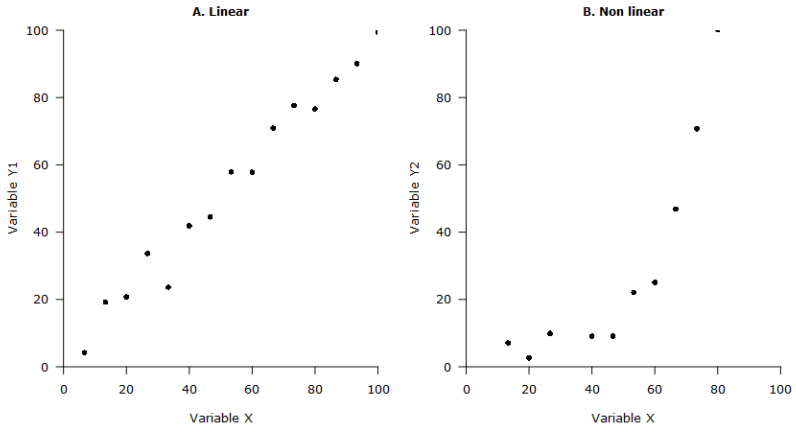
Graphical representation of data

Scatter plot

It is particularly useful when the values of the variables of the **y-axis** are **dependent** upon the values of the variable of the **x-axis**.

The pattern of the data points on the scatterplot reveals the relationship between the variables

Chart 5.6.2
Linear relation or non linear relationship



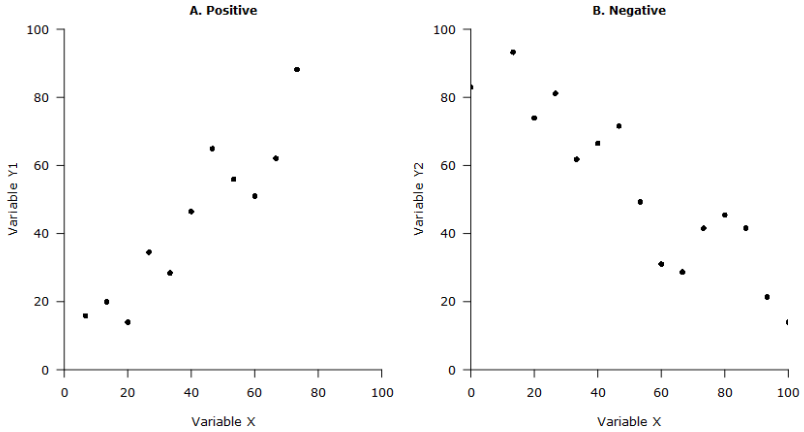
Graphical representation of data

Scatter plot

It is particularly useful when the values of the variables of the **y-axis** are **dependent** upon the values of the variable of the **x-axis**.

The pattern of the data points on the scatterplot reveals the relationship between the variables

Chart 5.6.3
Positive relation or negative relationship



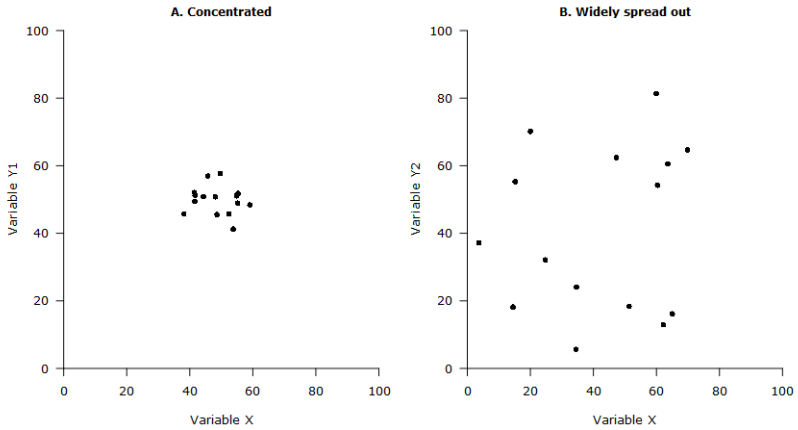
Graphical representation of data

Scatter plot

It is particularly useful when the values of the variables of the **y-axis** are **dependent** upon the values of the variable of the **x-axis**.

The pattern of the data points on the scatterplot reveals the relationship between the variables

Chart 5.6.4
Concentrated data or widely spread out data

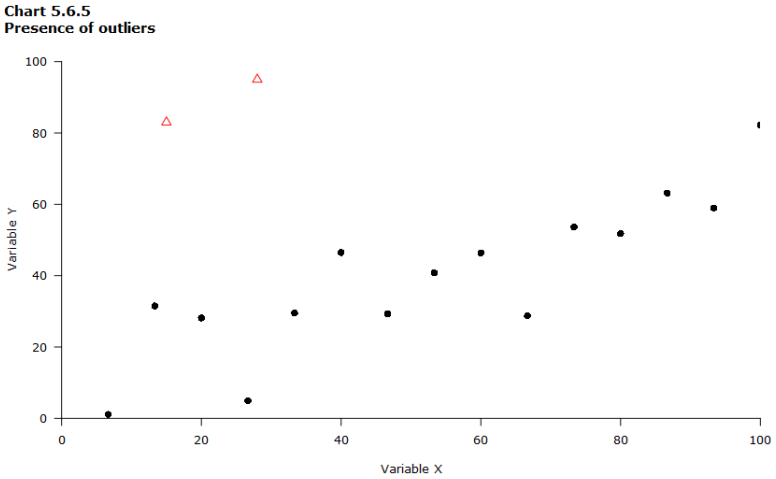


Graphical representation of data

Scatter plot

It is particularly useful when the values of the variables of the **y-axis** are **dependent** upon the values of the variable of the **x-axis**.

The pattern of the data points on the scatterplot reveals the relationship between the variables

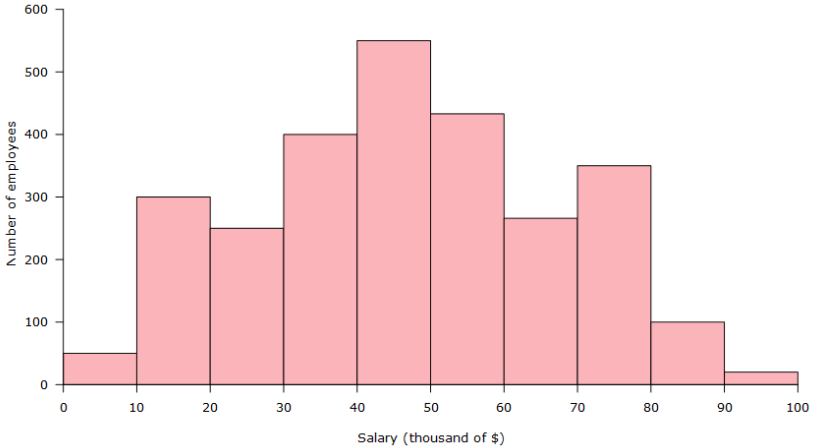


Graphical representation of data

Histogram

The histogram is a type of graph where the diagram consists of rectangles, the **area is proportional to the frequency of a variable** and the **width is equal to the class interval**. It is used to summarize **discrete or continuous data** that are measured on an interval scale.

Chart 5.7.1
Distribution of salaries of the employees of ABC Corporation



Graphical representation of data

Histogram vs Bar chart

Comparison terms	Bar chart	Histogram
Usage	To compare different categories of data.	To display the distribution of a variable.
Type of variable	Categorical variables	Numeric variables
Rendering	Each data point is rendered as a separate bar.	The data points are grouped and rendered based on the bin value. The entire range of data values is divided into a series of non-overlapping intervals.
Space between bars	Can have space.	No space.
Reordering bars	Can be reordered.	Cannot be reordered.

Bibliography

- Sivji, A. (2018). *Testing 101: Introduction to Testing*. URL: https://alysivji.github.io/testing-101-introduction-to-testing.html
- Meszaros, G. *Four-Phase Test*. URL: http://xunitpatterns.com/Four%20Phase%20Test.html
- Fowler, M. (2013). *GivenWhenThen*. URL: https://martinfowler.com/bliki/GivenWhenThen.html
- Cooke, J. (2017). *Arrange Act Assert pattern for Python developers*. URL: https://jamescooke.info/arrange-act-assert-pattern-for-python-developers.html
- Knight, A. (2018). *Behavior-Driven Python at PyCon 2018*. URL: https://www.youtube.com/watch?v=EtIAbfCrsFI
- De Caro, J. (2018). *A Beginner's Guide to Testing: Error Handling Edge Cases*. URL: https://www.freecodecamp.org/news/a-beginners-guide-to-testing-implement-these-quick-checks-to-test-your-code-d50027ad5eed
- Kanat-Alexander, M. (2013). *The Philosophy of Testing*. URL: https://www.codesimplicity.com/post/the-philosophy-of-testing/

Bibliography

- Okken, B. *Test & Code in Python: Developing Software with Automated Tests*. Podcast. URL: https://testandcode.com/