ntic
master
revolucionamos la comunicación

UNIVERSIDAD
COMPLUTENSE
DE MADRID

Minería de Datos y Modelización Predictiva

Conceptos preliminares de Estadística

Guillermo Villarino

1. Descriptive statistics

1.1. Types of variables

1.2. Key measures in descriptive statistics

1.3. Statistical graphics

1.4. Bivariate analysis

2. Inferential statistics

2.1. Preliminary concepts

2.2. Hypothesis testing

2.3. Confidence intervals

What is statistics?

Statistics is the science of collecting, organizing, analysing, presenting and interpreting data to assist in making better decisions

- Statistics is a multidisciplinary science since it is used to make decisions in different fields of research such as economics, business, health, etc.

- Statistics is one of the most powerful tool for assessing behavioural questions related to a certain population.

- Statistics is not magic and we must be critical with conclusions drawn in many studies due to its lack of information about the statistical procedure used.

.

Descriptive vs Inferential Statistics

Descriptive statistics

- ✓ Collect
- ✓ Organize
- ✓ Summarize
- ✓ Display
- ✓ Analize

Inferential statistics

- ✓ Test hypothesis about value of population paramenter based on sample statistic
- ✓ Predict and forecast value of population paramenters

Inferential statistics is based on probability calculus that allows us to know how a certain measure is distributed under a set of premises

Descriptive vs Inferential Statistics

Descriptive statistics

Suppose we want to describe the test scores measured in a numerical scale from 0 to 100 points in a specific class of 30 students. We record all of the test scores and calculate the summary statistics and produce graphs.

We can compute the mean score or the proportion of students that passed the test and even show a histogram to see the shape of the distribution.

Conclusions drawn gives us a pretty good picture of this specific class. There is no uncertainty surrounding these statistics because we gathered the scores for everyone in the class. However, we can’t take these results and extrapolate to a larger population of students.

Descriptive vs Inferential Statistics

Inferential statistics

For inferential statistics, we need to **define the target population** and then draw a random sample from that population.

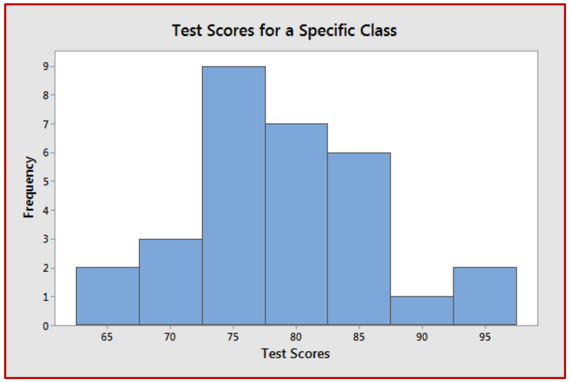
Let’s define our population as 8th-grade students in public schools in the State of Pennsylvania in the United States. We need to devise a random sampling plan to help ensure a representative sample. This process can actually be arduous.

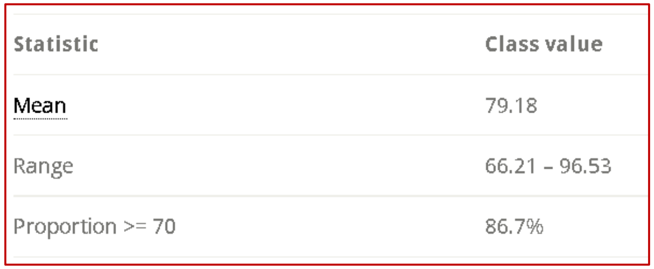
For the sake of this example, assume that we are provided a list of names for the entire population and draw a random sample of 100 students from it and obtain their test scores. Note that these students will not be in one class, but from many different classes in different schools across the state.

For inferential statistics, we can calculate the point estimate for the mean, standard deviation, and proportion for our random sample. However, it is staggeringly improbable that any of these point estimates are exactly correct, and there is no way to know for sure anyway. Because we can’t measure all subjects in this population, there is a margin of error around these statistics.

Descriptive vs Inferential Statistics. Expected output

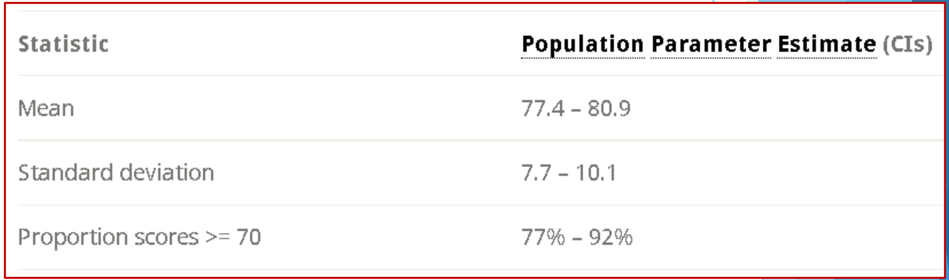
Descriptive statistics





Statistic	Class value
Mean	79.18
Range	66.21 – 96.53
Proportion >= 70	86.7%

Inferential statistics



Statistic	Population Parameter Estimate (CIs)
Mean	77.4 – 80.9
Standard deviation	7.7 – 10.1
Proportion scores >= 70	77% – 92%

In inferential statistics we expect for a single parameter to have a range of possible values. This range is usually known as Confidence Interval (CI) which will be wider the higher level confidence is.

1. Descriptive statistics

- Descriptive statistics involves describing, summarising and organizing the data in order to be understandable.

- Descriptive statistics has no any prospective propose.

- It is used to analyse the behaviour of the collected data giving us an idea of how a particular sample is distributed.

- Graphical displays are often used along with quantitative measures to enable clarity of communication.

- It is crucial being clear about the proper descriptive analysis for each type of variable in order to not make mistakes in our study.

1.1. Types of variables

Variables can be classified as qualitative (aka, categorical) or quantitative (aka, numeric).

- **Qualitative.** Qualitative variables take on values that are names or labels. The colour of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd, terrier) would be examples of qualitative or categorical variables.

- **Quantitative.** Quantitative variables are numeric. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be a quantitative variable.

1.1. Types of variables

Statistical data are often classified according to the number of variables being studied.

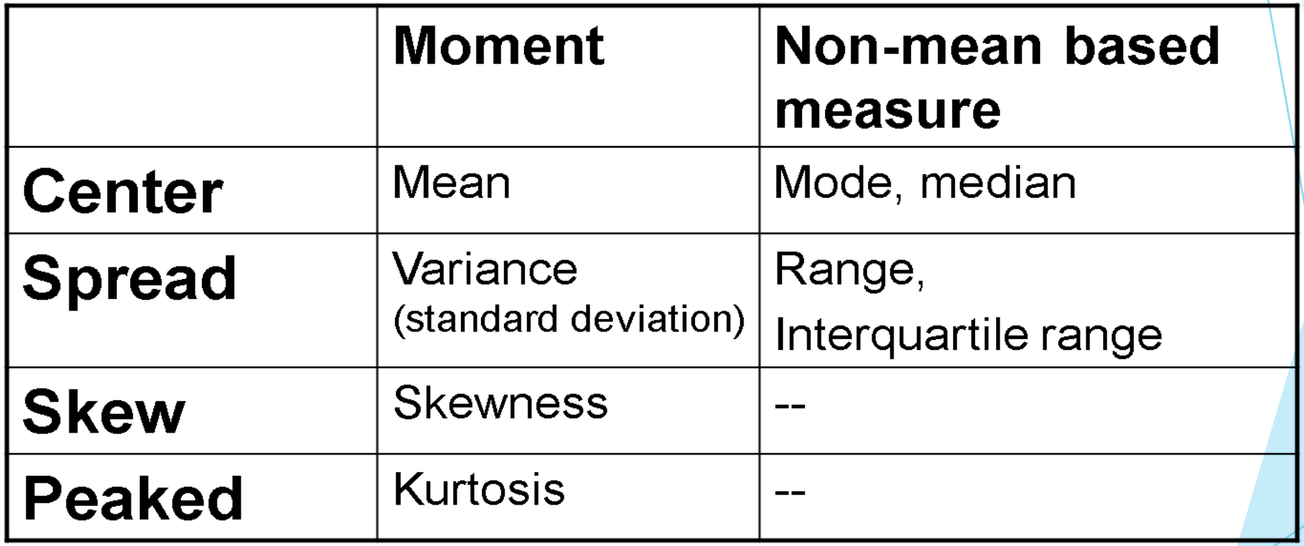
1 variable

- **Univariate data.** When we conduct a study that looks at only one variable, we say that we are working with univariate data. Suppose, for example, that we conducted a survey to estimate the average weight of high school students. Since we are only working with one variable (weight), we would be working with univariate data.

- **Bivariate data.** . When we conduct a study that examines the relationship between two variables, we are working with bivariate data. Suppose we conducted a study to see if there were a relationship between the height and weight of high school students. Since we are working with two variables (height and weight), we would be working with bivariate data

1.2. Key measures in descriptive statistics

In statistics there are three main types of measures: measures of central tendency, measures of variance/dispersion and measures of shape.



	Moment	Non-mean based measure
Center	Mean	Mode, median
Spread	Variance (standard deviation)	Range, Interquartile range
Skew	Skewness	--
Peaked	Kurtosis	--

Momentos de tendencia central: Media

1.2.1. Central Tendency

The most popular centrality measure. The mean

The mean is the most used measure of central tendency or typical value. It is very easy to handle mathematically, simple to calculate, and usually (symmetrical-like distributions) falls in the middle of the data set. To calculate the mean, we simply add up all values and divide by the sample size.

$$\boxed{\frac{\sum_{i=1}^{n} x_i}{n} \equiv \mu \equiv \overline{X}}$$

The position of the mean significantly varies with the shape of the distribution, and it is affected by the presence of *outliers* that are values lying far away from the general distribution.

Other central tendency measures.

As mentioned before, the mean is affected by the shape of the distribution it came from. Therefore, the mean is not always the best measure of centrality. Fortunately we can use other measures not affected by the shape of the distribution for assessing the central tendency.

- **Median.** The median is the middle value in a sorted numerical variable and divides the lower half from the higher half of the scores of this variable regardless the shape of the distribution.

- **Mode.** The mode is the most frequent value of a distribution. We can use the mode for describing both numerical (discrete) and categorical data but it is quite more common for describing categorical data.

1.2.2. Variability or dispersion

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. There are four frequently used measures of variability: range, interquartile range (IQR), variance, and standard deviation.

Range and IQR

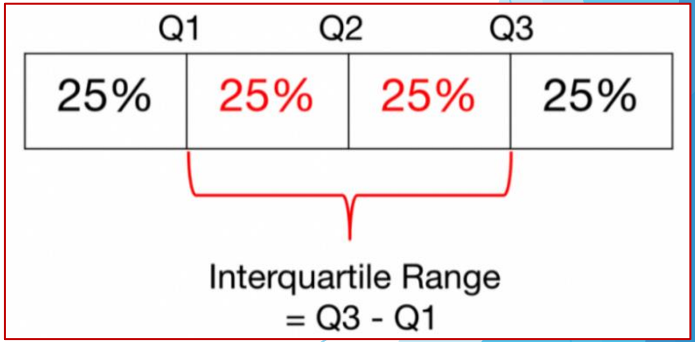
The range describes the difference between the largest and the smallest points in your data and the interquartile range (IQR) is a measure of statistical dispersion between upper (75th) and lower (25th) quartiles

Quartiles.

- 25% of the data points lie below Q1 and 75% lie above it.

- 50% of the data points lie below Q2 and 50% lie above it. Q2 is nothing but Median.

- 75% of the data points lie below Q3 and 25% lie above it.



Variance

The variance is computed by finding the difference between every data point and the mean, squaring them, summing them up and then taking the average of those numbers. The squares are used during the calculation because they weight outliers more heavily than points that are near to the mean. This prevents for differences above the mean neutralise those below the mean.

$$\boxed{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n} \equiv \sigma^2}$$

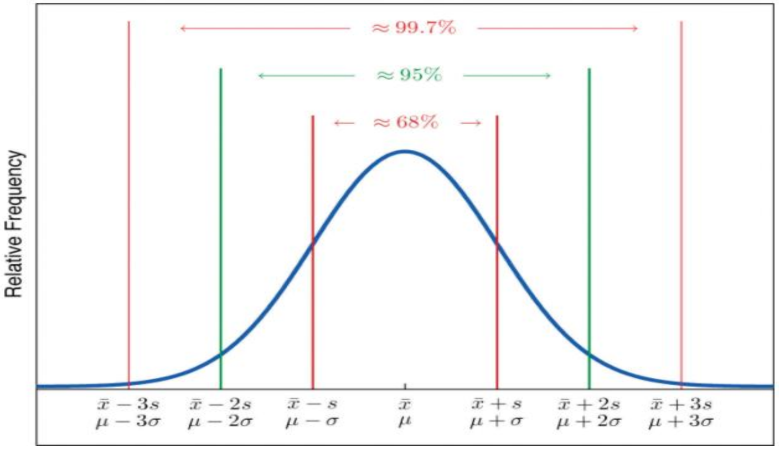
The main problem with the variance is the unit of measure that is squared from the original due to the formula and it make not easy to interpret its value.

Standard deviation

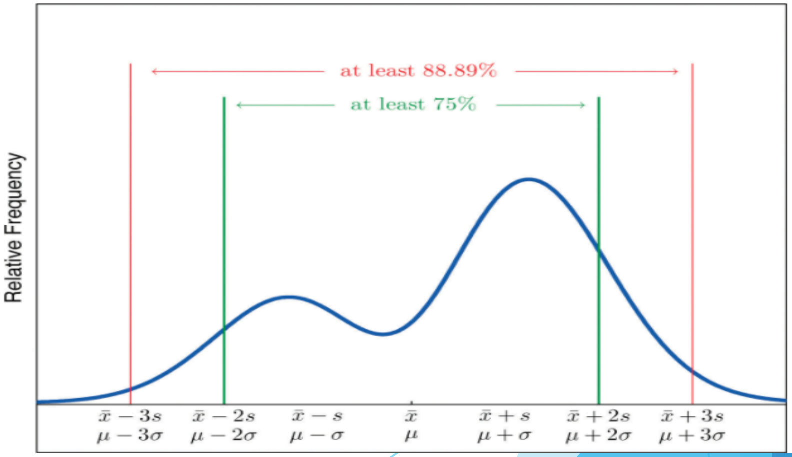
The standard deviation is used more often because it is in the original unit. It is simply the square root of the variance and because of that, it is returned to the original unit of measurement.

$$\boxed{\sqrt{\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}} \equiv \sigma}$$

Empirical Rule



Chebyshev’s inequality

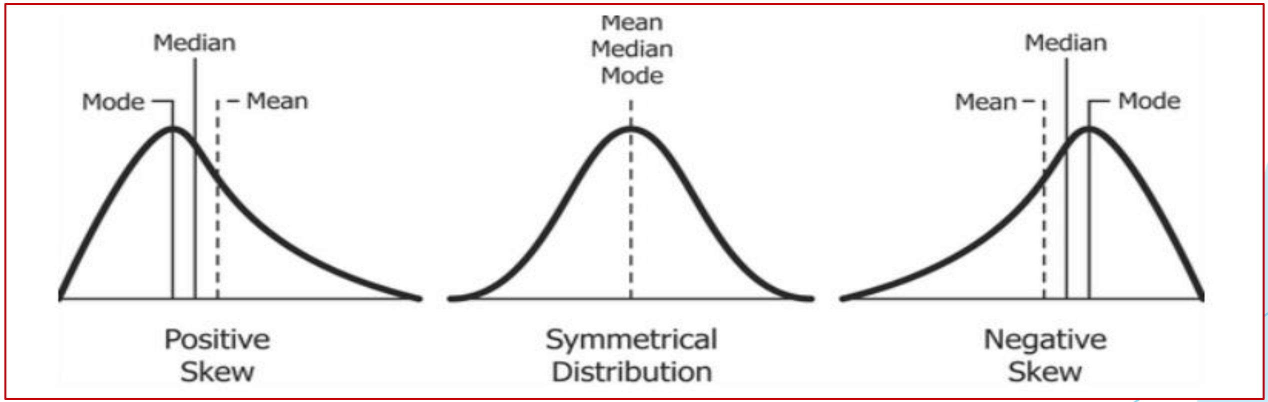


1.2.3. Shape

Skewness

Skewness measure somehow the difference in horizontal shape between the observed distribution and its corresponding bell-shaped (Normal) distribution.

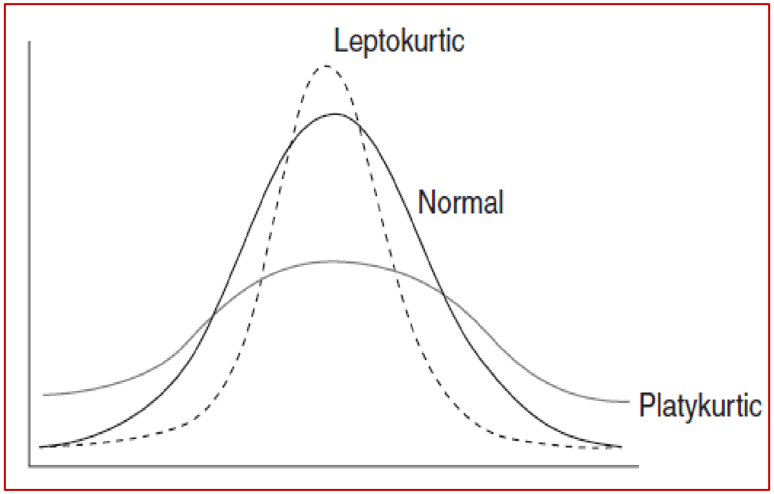
If there are a few scores creating an elongated tail at the higher end of the distribution, it is said to be positively skewed. If the tail is pulled out toward the lower end of the distribution, the shape is called negatively skewed.



Kurtosis

Kurtosis measures somehow the difference in vertical shape between the observed distribution and its corresponding bell-shaped (Normal) distribution.

When a distribution is symmetrical but has a peak that is higher than that found in a normal distribution, it is called leptokurtic. When a distribution is flatter than a normal distribution, it is called platykurtic

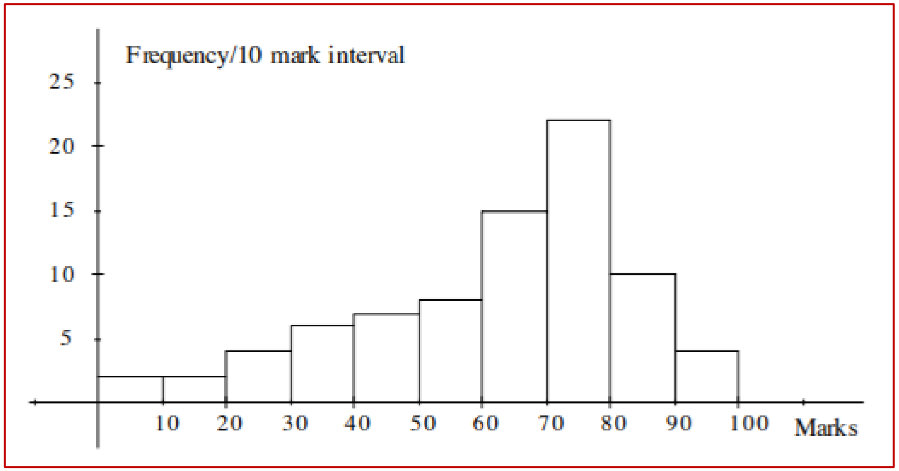


1.3. Statistical graphics

Graphical support is often useful in statistical analysis due to the global idea one can have just by looking at the appropriate plot.

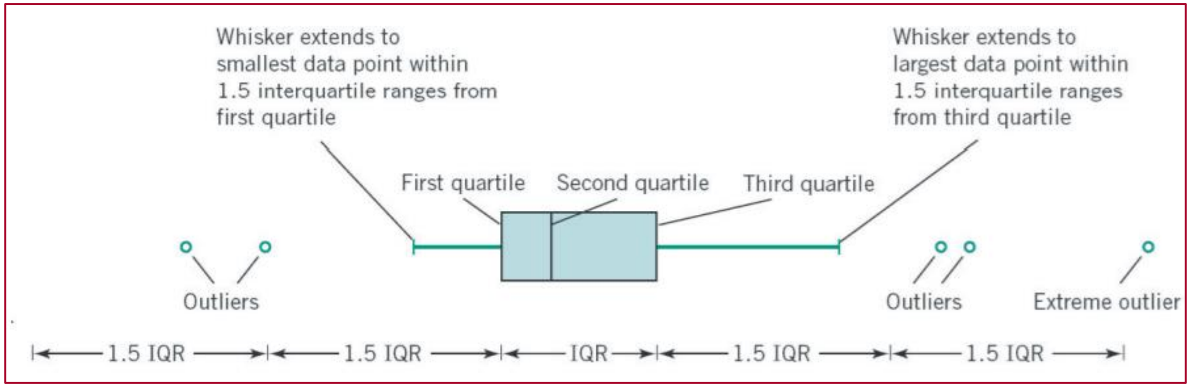
Histogram

Graphical support is often useful in statistical analysis due to the global idea one can have just by looking at the appropriate plot.



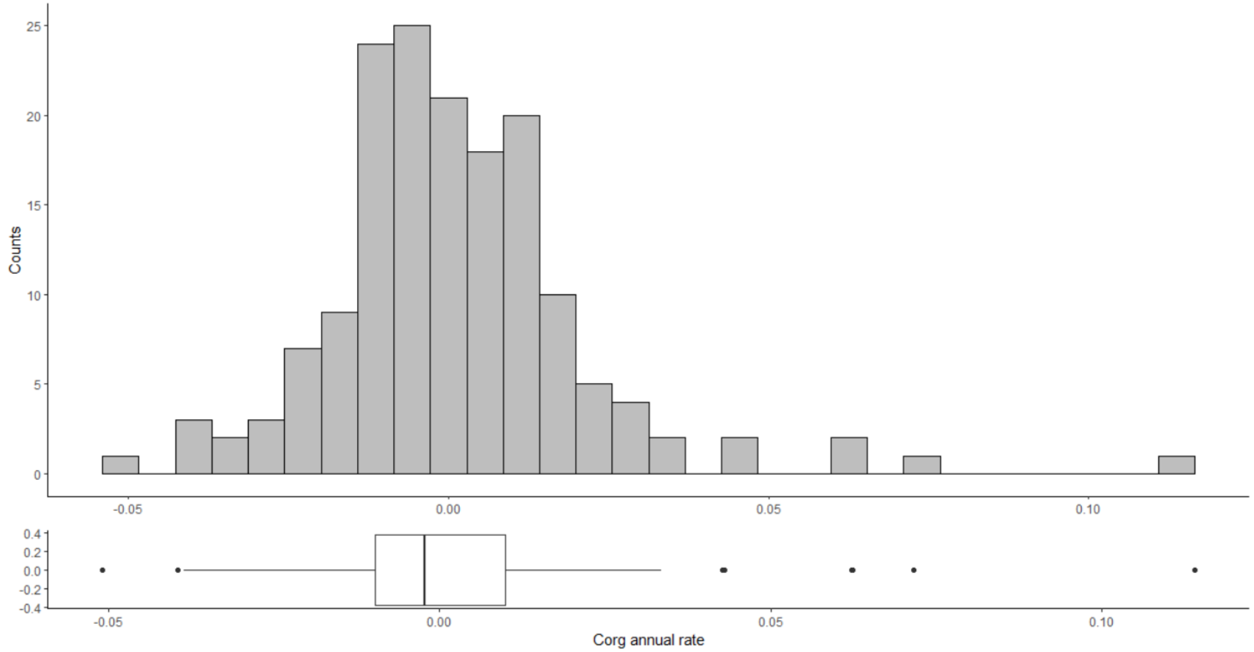
Boxplot

The boxplot is graphic representation of the distribution of scores on a numeric variable that includes the range, the median, and the interquartile range. The box in this graph contains some very useful information.


Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Relationship between histogram and boxplot

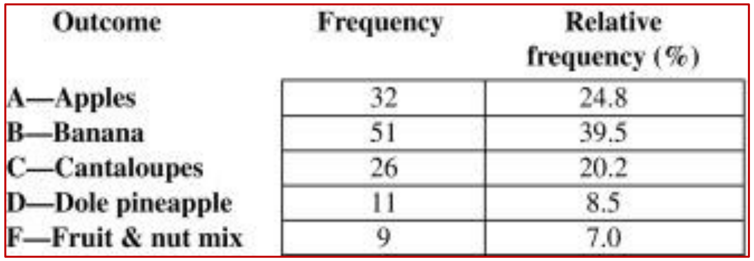
It is useful displaying both graphics together in order to establish a connexion between them. On the one hand, histogram makes more visible the shape of the distribution. On the other hand, boxplot helps identifying quartiles and median of the distribution.



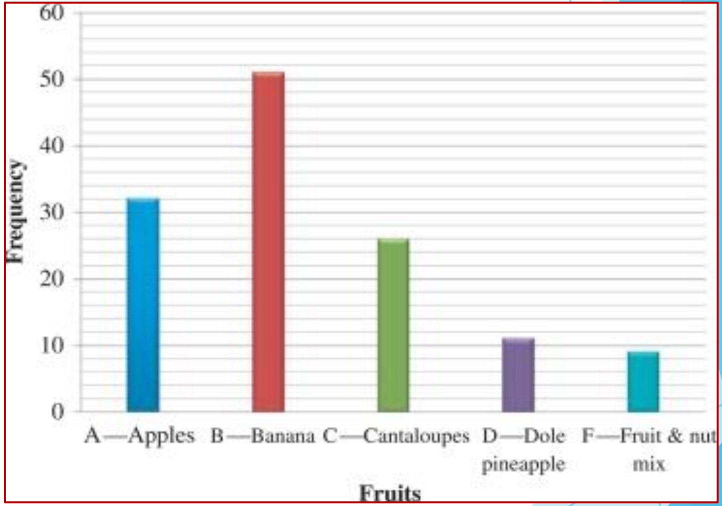
Bar plot

A bar plot is representative of a frequency distribution of a **qualitative** variable by means of rectangles whose heights are proportional to relative frequencies placed over each category. It is always related to a frequency table.

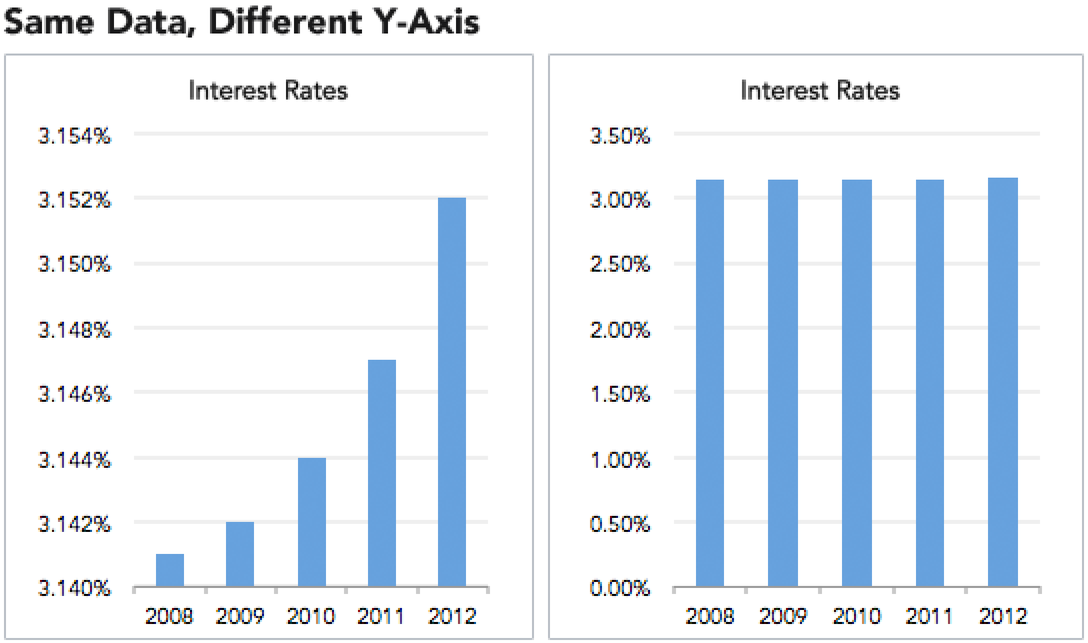
Favourite snack in a sample of 129 people summarised by the frequency table below and its corresponding bar plot on the right.



Outcome	Frequency	Relative frequency (%)
A—Apples	32	24.8
B—Banana	51	39.5
C—Cantaloupes	26	20.2
D—Dole pineapple	11	8.5
F—Fruit & nut mix	9	7.0



Be careful with the axis scale when interpreting bar plots!



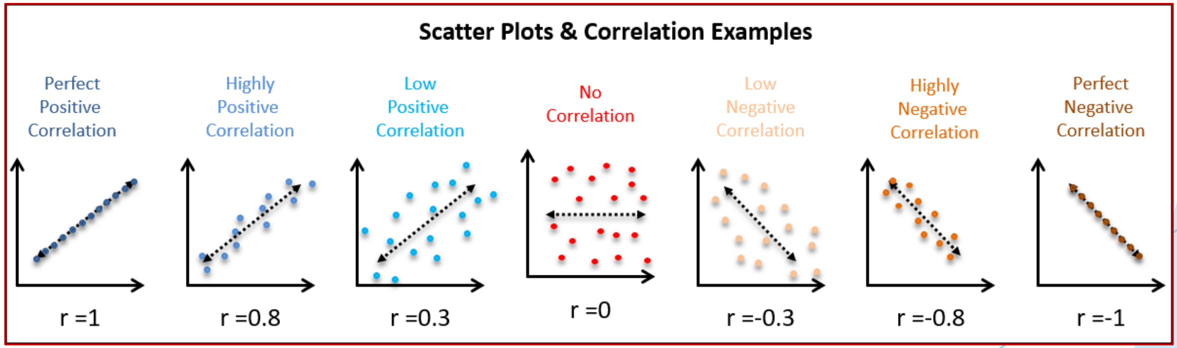
Source

1.4. Bivariate analysis

Bivariate analysis is the simultaneous analysis of two variables (attributes). It explores the concept of relationship between two variables, whether there exists an association, or whether there are differences between two variables. There are three types of bivariate analysis..

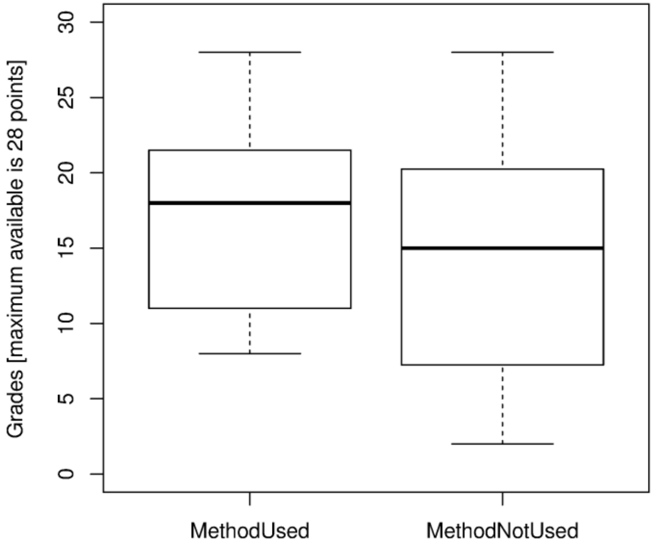
Two continuous variables

Linear correlation, denoted by r, quantifies the strength of a linear relationship between two numerical variables. When there is no correlation between two variables, there is no tendency for the values of one quantity to increase or decrease with the values of the second quantity. The scatterplot give us a visual idea about the relationship between both numerical variables.



Continuous vs Categorical

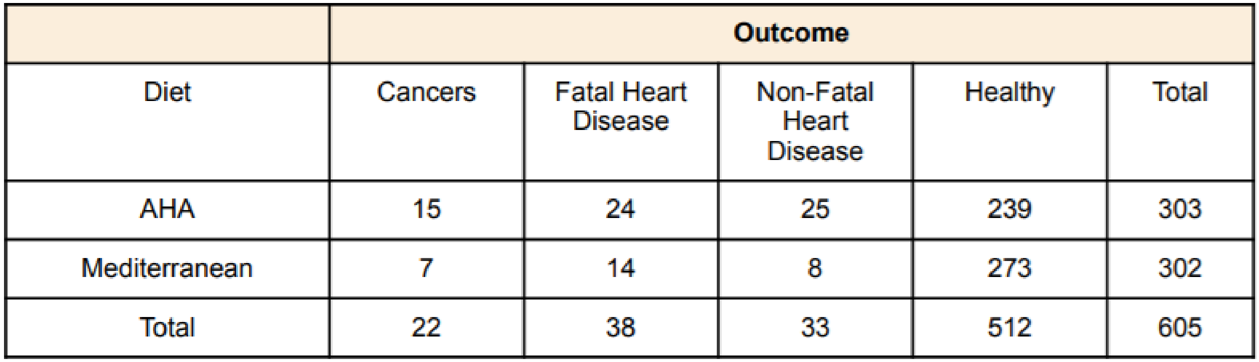
A parallel boxplot is a graphical representation of the distribution of a continuous variables into the groups given by the levels of a nominal variable or factor. The question arising in this scenario is how different a score is distributed for the individuals of each group. This is related with inferential analysis and t-test for mean difference.



Two categorical variables

A contingency table is a frequency distribution for a two-way statistical classification consisting in a matrix of information where one factor is represented by rows and the second factor is represented columns and the count of individuals that belong to exactly one cell within the matrix that is one row and one column.

Table below shows the data from a Mediterranean Diet and Health case study in which two characteristics were measured for 605 people. Diet type and outcome from a health test.

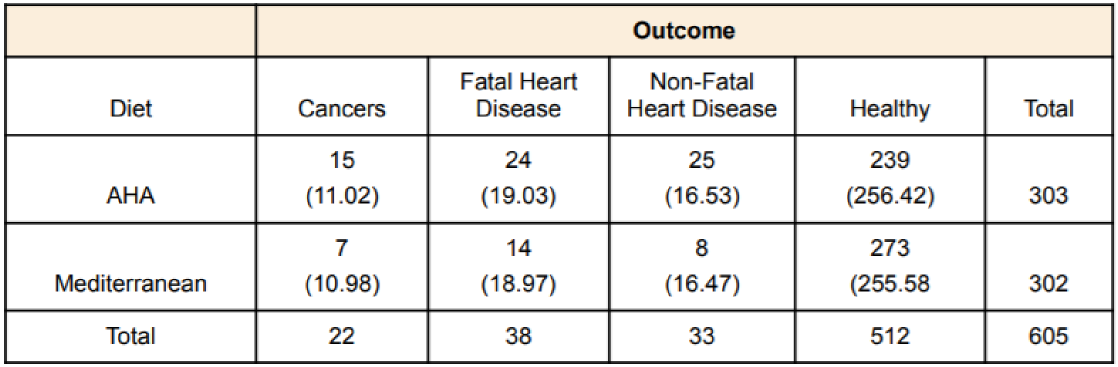


	Outcome				
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

Chi-squared test

The Chi-squared test is the most used bivariate test for categorical data. It is based in probabilities and how they are distributed in case of independency of subsets. When two subsets are independent, there is no intersection so there are no individuals belonging to each subset simultaneously. In this case the probability of the joint distribution is the product of the marginal probabilities.

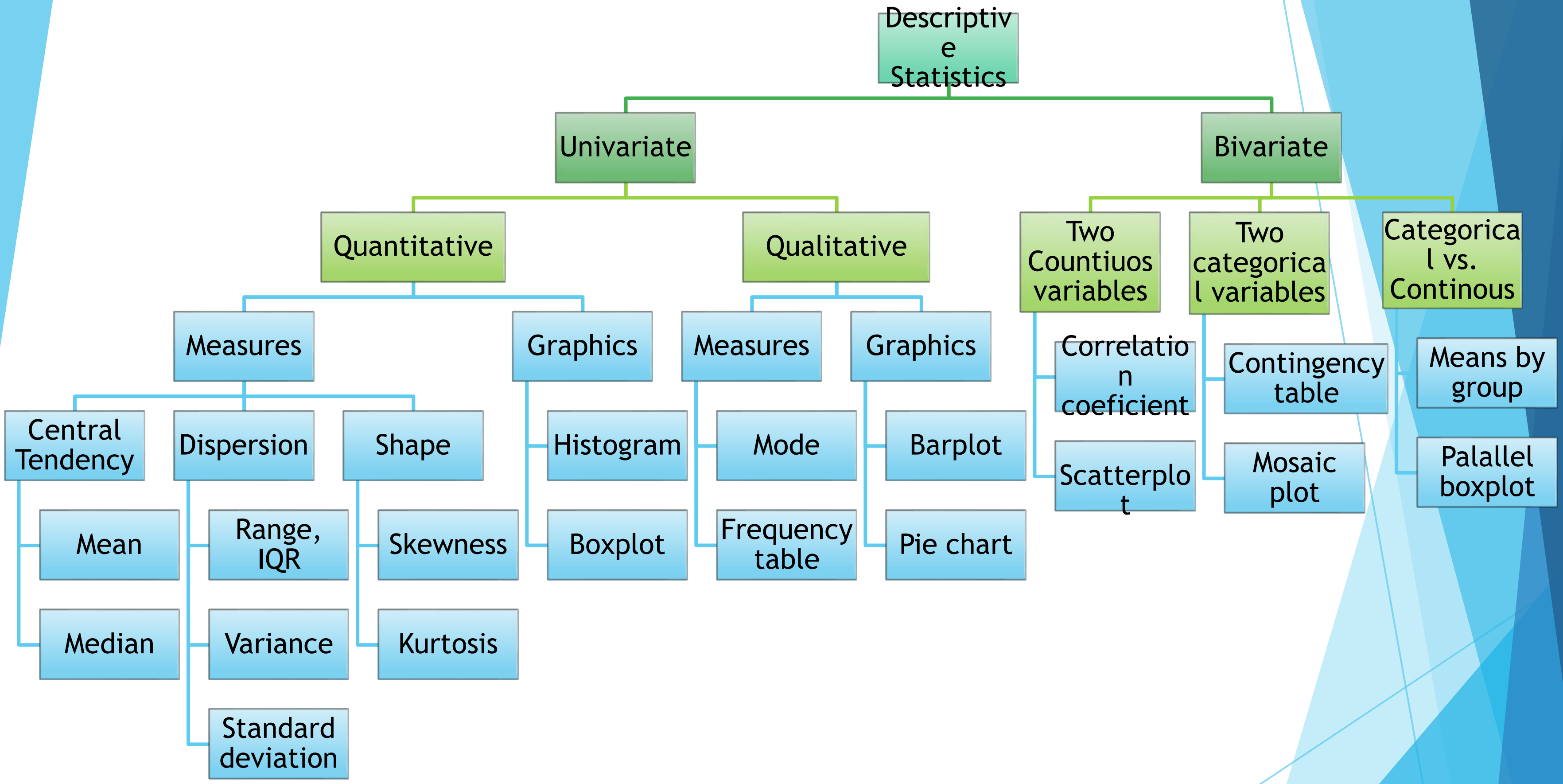
This test takes advantage of this property and assess whether two variables are independent. We cannot infer causality but only association in the relationship found through this analysis!



	Outcome				
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15 (11.02)	24 (19.03)	25 (16.53)	239 (256.42)	303
Mediterranean	7 (10.98)	14 (18.97)	8 (16.47)	273 (255.58	302
Total	22	38	33	512	605

$$x_3^2 = \sum \frac{(E-O)^2}{E} = 16.55$$

Decision tree in descriptive statistics



2. Inferential statistics

- Inferential statistics refer to the use of sample data to reach some conclusions (i.e., make some inferences) about the characteristics of the larger population, that the sample is supposed to represent.

- Although researchers are sometimes interested in simply describing the characteristics of a sample, for the most part we are much more concerned with what our sample tells us about the population from which the sample was drawn.

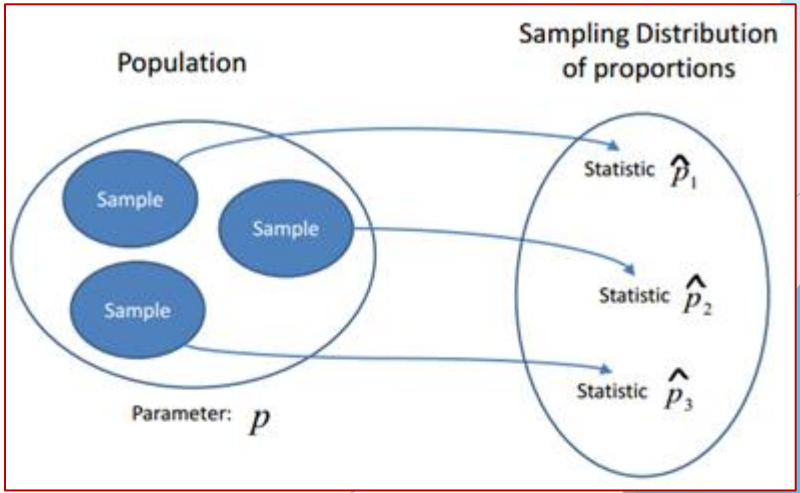
- Since we infer characteristic from the sample to the population, one of the most important aspects is controlling the sampling design to represent the population with the minimum bias.

2.1. Preliminary concepts

In statistical inference we usually handle the following concepts:

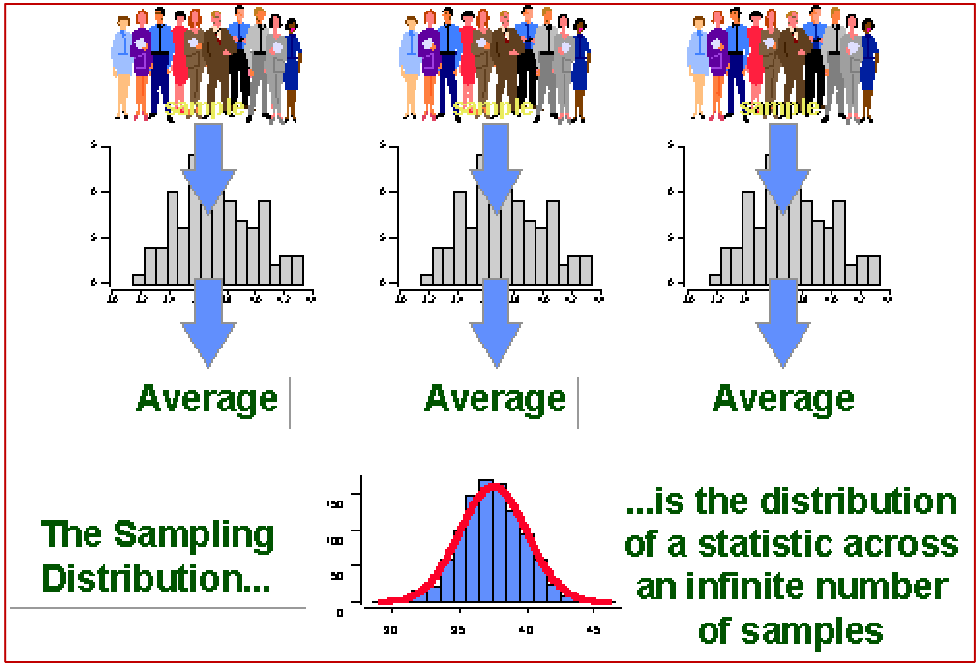
- **Population and sample.** A population is an individual or group that represents all the members of a certain group or category of interest. A sample is a subset drawn from the larger population

- **Parameter vs. Statistic:** Statistics are values derived from sample data, whereas parameters are values that are either derived from, or applied to, population data. Therefore, the statistical game consists of estimating the parameters of the population from the statistics of a representative sample.



Sampling distribution

The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population. It depends on multiple factors – the statistic, sample size, sampling process, and the overall population.



The Sampling Distribution...

...is the distribution of a statistic across an infinite number of samples

Central limit theorem

The Central Limit Theorem states that whenever a random sample is taken from any distribution, then the sample mean will be approximately normally distributed with the same mean as the original distribution and its variance divided by the sample size.

Thus, the larger the value of the sample size is, the better the approximation to the normal is due to the decrease in variance.

This is very useful when it comes to inference. For example, it allows us (if the sample size is fairly large) to use hypothesis tests which assume normality even if our data appear non-normal.

This is because the tests use the sample mean, which the Central Limit Theorem tells us will be approximately normally distributed.

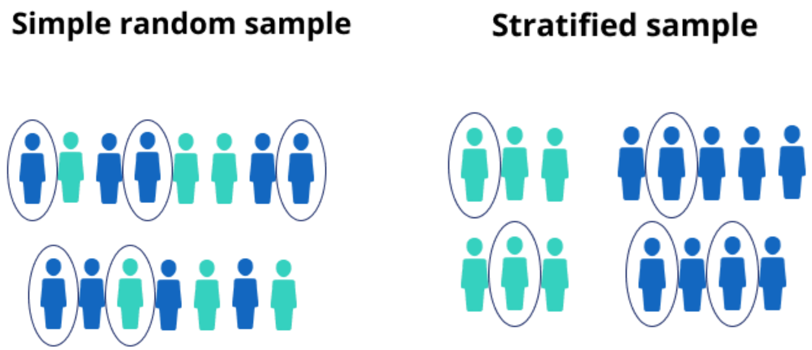
1.- Sampling distributions and the central limit theorem.
2.- The Central Limit Theorem.

2.2. Sampling methods

Among the probability-based sampling methods we can find:

- **Random sampling.** Every member of a defined population has an equal chance of being selected into a sample and any group of n individuals is equally likely to be chosen as any other group of n individuals.

- **Stratified sampling:** The population is first divided into subgroups (or strata) who all share a similar characteristic then a random sampling is conducted in each group. It is used when we might reasonably expect the measurement of interest to vary between the different subgroups, and we want to ensure representation from all the subgroups.



Non-probabilistic sampling

In several studies is not possible or affordable a probabilistic sampling

- **Convenience sampling.** Participants are selected based on availability and willingness to take part. Useful results can be obtained, but the results are prone to significant bias, because those who volunteer to take part may be different from those who choose not to (volunteer bias), and the sample may not be representative of other characteristics, such as age or sex.

- **Quota sampling:** This method of sampling is often used by market researchers. Interviewers are given a quota of subjects of a specified type to attempt to recruit.

- **Snowball sampling:** This method is commonly used in social sciences when investigating hard-to-reach groups. Existing subjects are asked to nominate further subjects known to them, so the sample increases in size like a rolling snowball.

Bias is a term which refers to how far the average statistic lies from the parameter it is estimating. Errors from chance will cancel each other out in the long run, those from bias will not.

2.2. Hypothesis testing

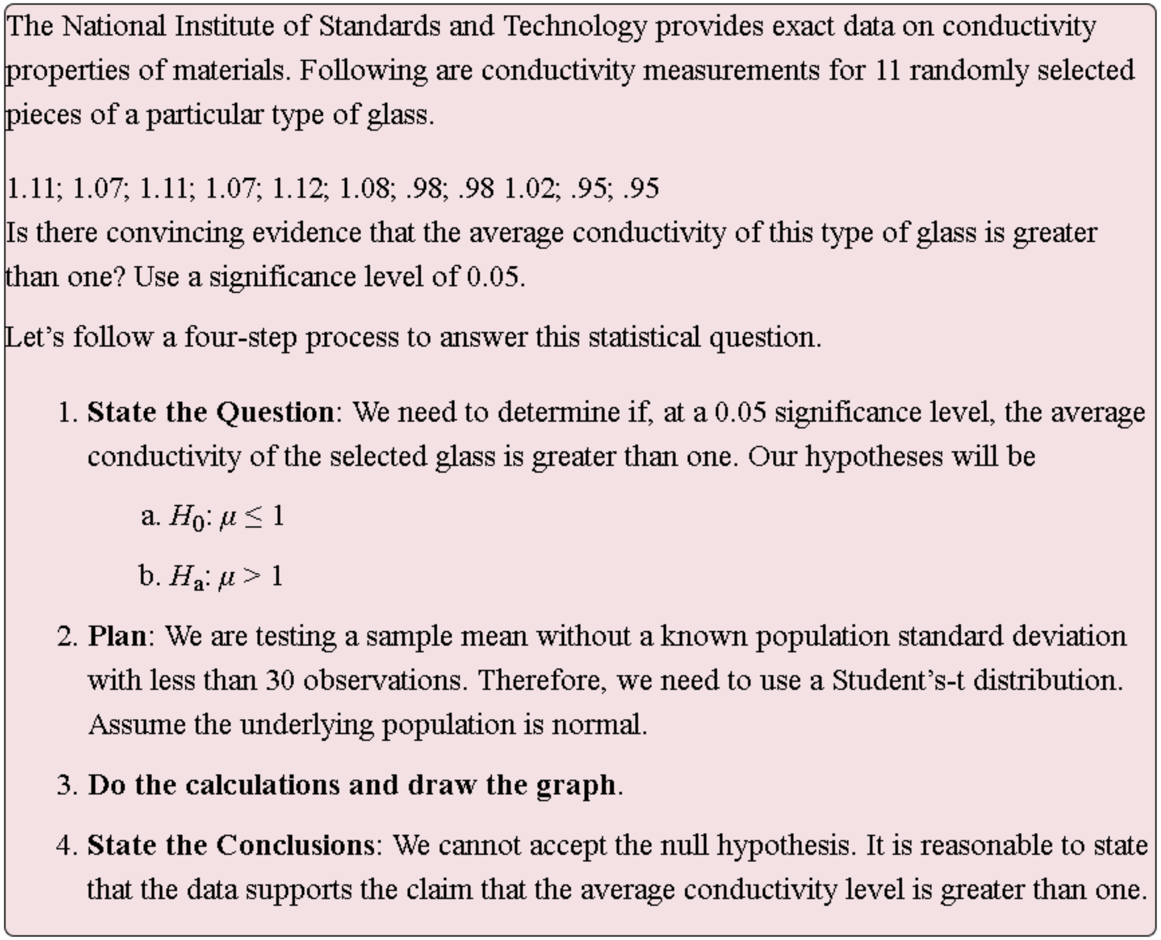
One of the most powerful tools for researching and decision making based on data is hypothesis testing due to its capability for giving a level of confidence for the estimated parameters and statistical significance to the conclusions reached within the analysis carried on.

- **Null hypothesis.** The null hypothesis is typically denoted as H0. The null hypothesis states the "status quo". This hypothesis is assumed to be true until there is evidence to suggest otherwise.

- **Alternative hypothesis:** The alternative hypothesis is typically denoted as H1. This is the statement that one wants to conclude. It is also called the research hypothesis.

If we reject the null hypothesis based on our research (i.e., we find that it is unlikely that the pattern arose by chance), then we can say our test lends support to our hypothesis. But if there is not enough evidence for refuse H0, meaning that it could have arisen by chance, then we say the test is inconsistent with our hypothesis.

Hypothesis testing example



The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95

Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05.

Let’s follow a four-step process to answer this statistical question.

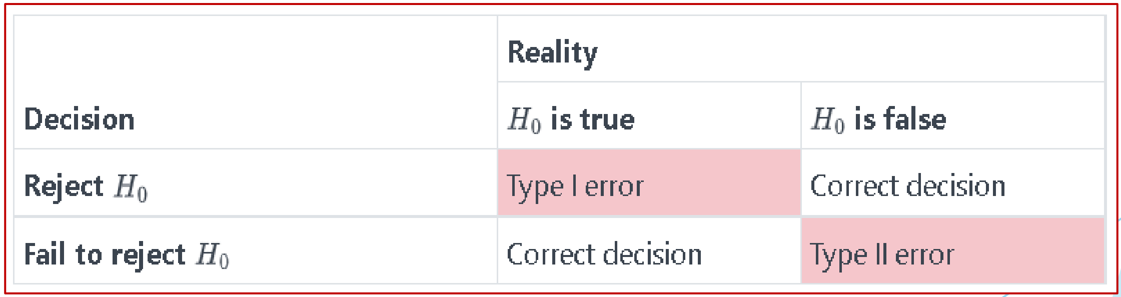
1. **State the Question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be
   a. H_0: $\mu \leq 1$
   b. H_a: $\mu > 1$
2. **Plan**: We are testing a sample mean without a known population standard deviation with less than 30 observations. Therefore, we need to use a Student’s-t distribution. Assume the underlying population is normal.
3. **Do the calculations and draw the graph**.
4. **State the Conclusions**: We cannot accept the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Null hypothesis and sampling distribution

In hypothesis testing we are handling statistics of the sample and our main aim is to infer the range of possible values the true parameter could lie within. When talking about sample statistic the concept of sampling distribution is always involved.

The sampling distribution under the null hypothesis is the one we can know while the distributional behaviour under the alternative is unknown.

- **Type I error:** Reject the null hypothesis when it is actually correct, and it is related to the significance and confidence level of the test.
- **Type II error:** Not reject the null hypothesis when it is actually false and is related to the power of the test.



	Reality	
Decision	**H_0 is true**	**H_0 is false**
Reject H_0	Type I error	Correct decision
Fail to reject H_0	Correct decision	Type II error

Concept of p-value

The probability value or p-value is by far the best known (and sometimes most feared) among the statistical concepts handled in everyday analysis.

- It has the property for standing a rule of thumb to interpret the results given by any kind of statistical test.

- The p-value of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone, if the null hypothesis H0, is true.

- The p-value is the probability of wrongly rejecting the null hypothesis if it is in fact true.

- Once the test is conducted, the p-value is compared with the actual significance level of our test and, if it is smaller, the result is significant. That is, if the null hypothesis were to be rejected at the 5% significance level, this would be reported as "$p < 0.05$".

Hypothesis testing for a single mean

In one-sample analysis we are often interested in contrasting whether the mean of the population (μ) lies above or below (or is simply different) a certain quantity ($\mu0$) usually fixed based on research criteria.

$H0: \mu > \mu0$
$H1: \mu \leq \mu0$

$H0: \mu = \mu0$
$H1: \mu \neq \mu0$

$H0: \mu < \mu0$
$H1: \mu \geq \mu0$

1. We know the true variance of the population so that we can use it to conduct the analysis without any additional source of variability. In this case, the sampling distribution will be Normal.
2. More often we will not know the true variance of the population so we must estimate it from the information given by the sample adding a new source of variability and therefore varying the sample distribution of the parameter which will be a Student's t-distribution.

Hypothesis testing for difference between two means

In two-sample analysis we are often interested in contrasting whether the mean of one population (μ1) lies above or below (or is simply different) the mean of a second population (μ2). This statement is commonly presented as:

H0: μ1- μ2 > 0
H1: μ1- μ2 ≤ 0

H0: μ1- μ2 = 0
H1: μ1- μ2 ≠ 0

H0: μ1- μ2 < 0
H1: μ1- μ2 ≥ 0

1. We know the true variance of the population so that we can use it to conduct the analysis without any additional source of variability. In this case, the sampling distribution will be Normal.

2. More often we will not know the true variance of the population so we must estimate it from the information given by the sample adding a new source of variability and therefore varying the sample distribution of the parameter which will be a Student's t-distribution.

In this case, the standard error of the sampling distribution has a more complicated formula

Hypothesis testing for a single proportion

In one-sample test for proportions we are often interested in contrasting whether the ratio of people having a certain characteristic or behaviour (p) lies above or below (or is simply different) a certain proportion (p0) usually fixed based on research criteria.

H0: p > p0
H1: p ≤ p0

H0: p = p0
H1: p ≠ p0

H0: p < p0
H1: p ≥ p0

- In this situation our population parameter is the true proportion of people having the studied characteristic.
- Our statistic will be the observed proportion in the sample.
- The sampling distribution is the Binomial distribution (see page 202 in this reference) that models the proportion of success, p, in several trials, n, of an experiment (like flip a coin several times), because it perfectly fits this phenomenon.

Hypothesis testing for difference between two proportions

In two-sample analysis we are often interested in contrasting whether the proportion of people having a certain characteristic in one population (p1) lies above or below (or is simply different) this proportion in a second population (p2).

H0: $p1 - p2 > 0$
H1: $p1 - p2 \leq 0$

H0: $p1 - p2 = 0$
H1: $p1 - p2 \neq 0$

H0: $p1 - p2 < 0$
H1: $p1 - p2 \geq 0$

- In this situation our population parameter is the difference between the true proportion of people having the studied characteristic in each population.
- Our statistic will be the observed difference in proportion between the corresponding samples.
- The sampling distribution is again the Binomial distribution that will be approximated by a Normal distribution thanks to the Central Limit Theorem.

In this case, the standard error of the sampling distribution has a more complicated formula

2.3. Confidence intervals

A confidence interval (CI) is another type of estimate but, instead of being just one number, it is an interval of numbers.

- It provides a range of reasonable values in which we expect the population parameter to fall.

- There is no guarantee that a given confidence interval does capture the parameter, but there is a predictable probability of success.

- It is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken.

CI = (estimated parameter - margin of error, estimated parameter + margin of error)

Confidence interval for a single mean with known variance

A confidence interval for a population mean usually note by µ, when the standard deviation of the population is known (noted as σ), is based on the conclusion of the Central Limit Theorem that stands the sampling distribution of the sample means follow an approximately normal distribution.

$$\boxed{\bar{x} - z_c \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_c \frac{\sigma}{\sqrt{n}}}$$

Where zc denotes the critical value given by the Normal distribution for the decided level of confidence, σ denotes the standard deviation of the population and n the sample size.

Confidence interval for a single mean with unknown variance

A confidence interval for a population mean usually note by µ, when the standard deviation of the population is unknown, it must be estimated by the sample variance and therefore the sampling distribution changes becoming a t-distribution.

$$\bar{x} - t_c \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_c \frac{s}{\sqrt{n}}$$

Where tc denotes the critical value given by the Student's t-distribution for the decided level of confidence, s denotes the standard deviation of the sample and n the sample size.

Confidence interval for a single proportion

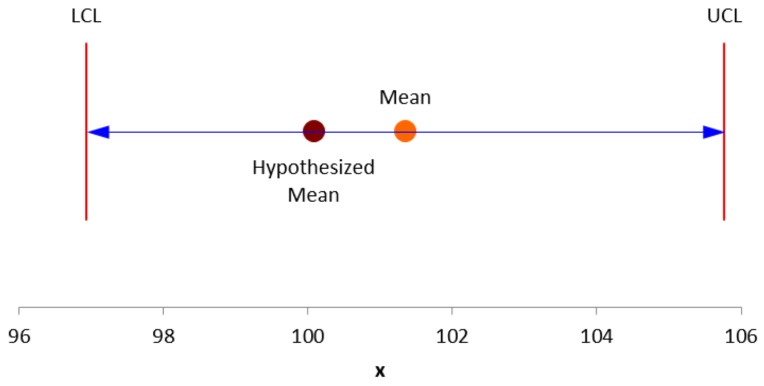
When dealing with proportions, we already know that sampling distribution is Binomial, but it will be approximate by a Normal distribution to get the critical values for the predefined level of confidence.

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \boldsymbol{p} < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where zα/2 denotes the critical value given by the Normal distribution for the decided level of confidence (1-α), p-hat denotes the proportion observed in the sample and its standard deviation is given by the expression inside the square root. As usual n denotes the sample size.

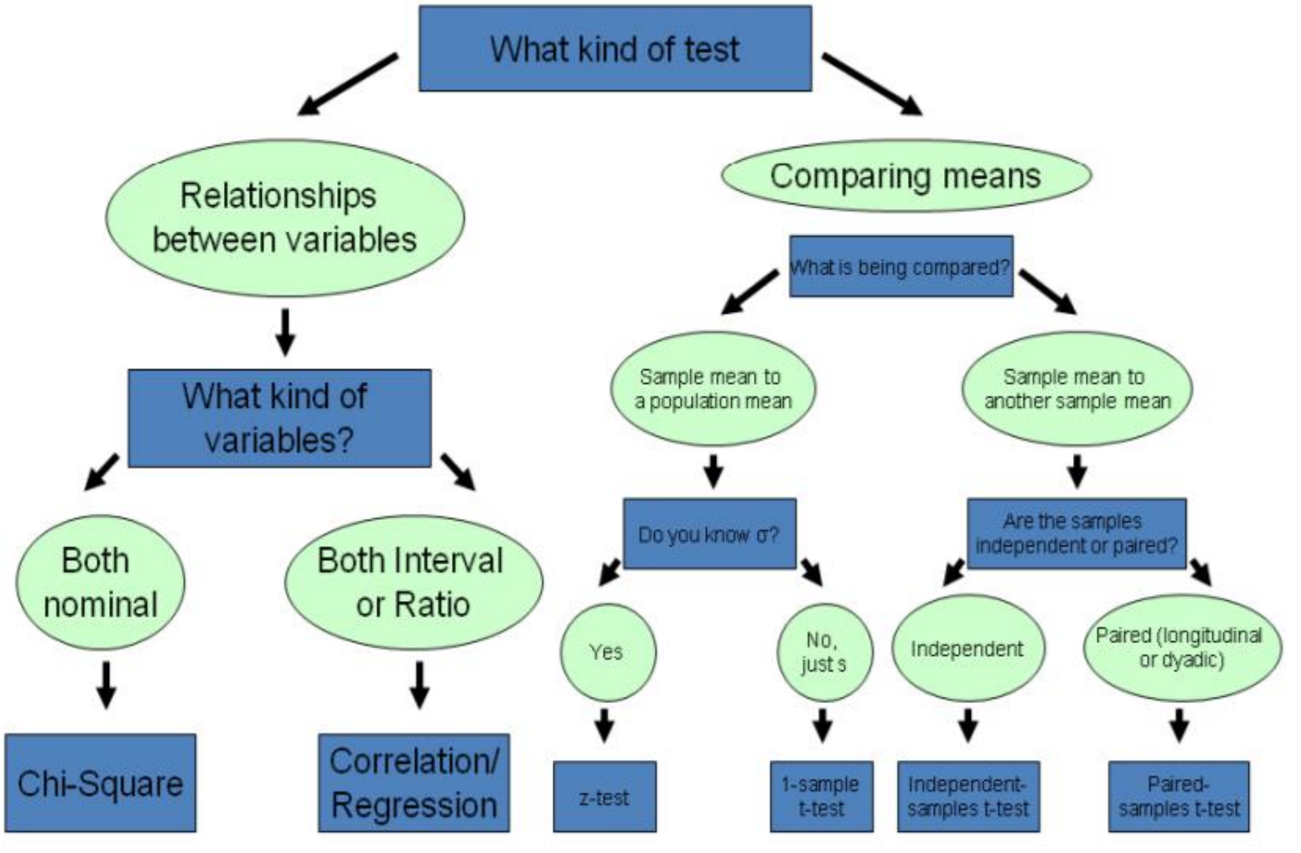
Confidence interval and hypothesis testing

There exist a clear connexion between CI’s and hypothesis testing since both of them arises from the same statistical procedure of probability calculus.



- If the interval does not contain the null hypothesis value, then we will reject the null hypothesis.
- If the interval contains the null hypothesis value, then we will fail to reject the null hypothesis.

Decision tree in inferential statistics



What kind of test

Relationships between variables

What kind of variables?

Both nominal

Chi-Square

Both Interval or Ratio

Correlation/ Regression

Comparing means

What is being compared?

Sample mean to a population mean

Do you know σ?

Yes

z-test

No, just s

1-sample t-test

Sample mean to another sample mean

Are the samples independent or paired?

Independent

Independent-samples t-test

Paired (longitudinal or dyadic)

Paired-samples t-test