TYPE OF DATA

Qualitative: text, images, sounds. Values vary by class and usually represented by label. Characteristics of the occurrence are not numerical. Data not given numerically such as birthplace, favorite color, favorite meal.
Nominal: categories have no natural order or structure. E.g: religion, sex, race.
Ordinal: if the data can be organized in certain order or structure. E.g: high, moderate, low.
Binary: yes or no answer
Dichotomous: existence or absence of something;

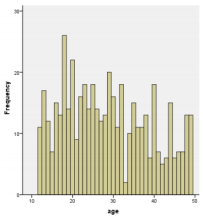
Quantitative: numbers, age, temperature. (Data that is numeric or non numeric but assigned numeric codes). Must be numerically measured and collected by counting or measuring. Not all numbers are to be suctracted, e.g: personal number.
Discrete: data with whole specific number or categories that are not distributed along a continuum. E.g: number of children, number of deaths.
Continuous: data that can take any value from a range depending on the precision of the measurement. Measured along a continuum at any place beyond the decimal point. E.g: age, weight, temperature.

Equidistant scales: values whose intervals are distributed in equal units;
Interval scales: even though there is equidistance, the difference between measures does not have the same meaning. There is no true zero. E.g: temperature (15° + 15° are not 30°).
Ratio scales: Zero is a meaningful value and there is equidistance between measures. E.g: money, age, height, weight,

MEASURES OF CENTRAL TENDENCY AND DISPERSION

Frequency distribution: simple distribution, easy graphic representation, cannot be summarized in one estimate

Age	Freq	%	Cum%
12	11	2.3	2.3
13	17	3.5	5.8
14	12	2.5	8.3
15	7	1.5	9.8
16	15	3.1	12.9
17	13	2.7	15.6
18	26	5.4	21.0
19	14	2.9	23.9
20	22	4.6	28.5



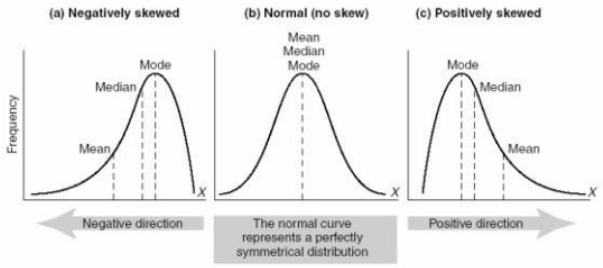
Mode: For nominal variables. Category of a variable that occurs most frequently in the data set. Missing don't count. Mode is the attribute behind the estimation of percentage/frequency. When observations tend to cluster in two or more different attributes, we have multimodal distributions.
11223344444445556. 4 is the mode. 7/13

- Minimum and maximum values are the highest and lowest values a variable can receive and not only the ones available. They are not necessarily the ones with lowest frequency.
- The range is a number, not the minimum and maximum values of the distribution. It is never negative. Eg: 3~9; 9 - 3 = range = 6

Median: is the middle value of the rank. There's 50% observations below and above it. Outliers cannot be detected in the calculation. Not affected by the influence of outliers. Formula to find the position of the median: (n+1)/2
How to find it: 1) rank the values, 2) count the number of total observations (n), 3) calculate the position with the formula and 4) find the observation.

Mean: most common measure, useful to make predictions, requires interval variable (continuous), distribution needs to be relatively normal.
How to find it: 1) add up all the observations together, divide the sum by the total number of observations.

- Skewness: measure of the asymmetry of the probability distribution of a real valued random variable. It can be negative skewed (tail is longer on the left side, mass of the distribution concentrated on the right side) or positive skewed (tail is longer on the right side, mass of the distribution concentrated on the left side).



- Skewness coefficient: indicates how skewed is the distribution of a ratio or interval variable. The distribution is more balanced as the coefficient is closer to 0.

➔ When could the median be a better central tendency measure than the mean? When the sample is small (the smaller the sample, less likely to be normally distributed)/ positively skewed (use median); the sample is not randomly taken (might be skewed); several outliers.

Dispersion: how spread out are the values around the measures of central tendency? How much do the values deviate from the measure of central tendency?
-With values close to the mean, the mean is a good predictor; peaked distribution.
-With spread values and symmetrical distribution, the mean is a good measure, but the predictive capability is reduced.
-Variability: outliers and standard deviation increase variability.

Percentiles: percentiles report the relative standing of a particular value within a statistical data set. The actual mean and standard deviation are not important neither is the actual data value. What is important is where you stand in relation to everyone else. In the case of exam scores, suppose your score is better than 90% of the class. That means that your exam score is at the 90^{th} percentile. Conversely, if your score is at the 10^{th} percentile, that means only 10% of the scores are below yours and 90% are above. Advantages are that percentiles have universal interpretation, the 95^{th} percentile always means 95% of the other values lies below yours and 5% lies above it. This also allows you to fairly compare 2 data sets that have different means and standard deviations. Remember: a percentile is not a percent. A percentile is a value (or average of 2 values) in the data set that marks a certain percentage of the way through the data. If you

are in the 80th percentile it doesn't mean that you scored 80% of the questions correctly, means that 80% of the scores were lowers than yours and 20% of the other scores were higher than yours.
How to calculate the percentile: 1) order all values in the data set; 2) multiply the percent you want (k) by the total number of values (n), obtain index number; 3) if the index is a not a whole number, round it up to the nearest number and count the values in your data set until you reach the number indicated; 4) if the index obtained is a whole number, count the values in the data set until you reach the number indicated, the percentile will be the average of the corresponding value and the following value. [x + (x+1)]/2.

For example, suppose you have 25 test scores, and in order from lowest to highest they look like this: 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99. To find the 90th percentile for these (ordered) scores, start by multiplying 90% times the total number of scores, which gives 90% ∗ 25 = 0.90 ∗ 25 = 22.5 (the index). Rounding up to the nearest whole number, you get 23.

Counting from left to right (from the smallest to the largest value in the data set), you go until you find the 23rd value in the data set. That value is 98, and it's the 90th percentile for this data set.

Now say you want to find the 20th percentile. Start by taking 0.20 x 25 = 5 (the index); this is a whole number, so the 20th percentile is the average of the 5th and 6th values in the ordered data set (62 and 66). The 20th percentile then comes to (62 + 66) ÷ 2 = 64.

The 25th percentile is also known as the 1st quartile, the 50th percentile is the median or 2nd quartile and the 75th percentile is the 3rd quartile.

Measures of dispersion: variability

Mean deviation: the average distance between each value and the mean.
How to calculate it: find the mean of all values; subtract the mean to each value and keep the absolute difference; then add up all the values together and divide the total by the number of values; the result will be the average that the values are away from the mean. Formula: Mean Deviation = Σ |X - μ | ÷ N
Σ = sum; X = each value; μ = mean; N = total number of values.

Variance: the average of the squared differences from the mean.
How to calculate it: 1) calculate the mean; 2) for each number, subtract the mean and square the result; 3) divide by the number of samples.
If we are calculating the population, divide by the total number of samples (N). If it is a sample of a bigger population, then divide by N -1 when calculating the variance.

Standard deviation: is a measure of how spread out numbers are. The symbol is σ (greek letter sigma). Formula: square root of the variance ($\sigma = \sqrt{variance}$) . Therefore the variance is σ^2.

Outlier: value that is far from the mean distribution. It is probably due to sample error/confounder. Can be found in continuous or discrete (numerical) variables.

MEASURES OF DISEASE ASSOCIATION:

- Proportion: typical calculation, frequency measure; amount out of a whole
- Rate: proportion x time; denominator can be person time (each participant contributes with a portion of person time; density) Measure of how rapidly an event happen.
- Ratio: A:B; comparison between different groups
    - Ideal study design: equally distributed, large sample size, adjusted for possible confounders. Ideally the population exposed and unexposed had to be the same, but it is not possible, so the study population is similar but not equal.

-Confounders: third variable, related to the independent variable (exposure); associated with the outcome; not causal but correlate one to the other. A confounder cannot be the mediator or causal factor for the outcome.

-Incidence is a measure of risk of developing the disease. We must specify period of time and study population. Number of new cases (specific population and time)/number of people at risk (during the same period of time). Any individual who is included in the denominator must have the potential to become part of the group that is counted in the numerator.

+ Special types of incidence:
    - Mortality Rate (incidence) – fatal cases of a disease/total population (individuals at risk and individuals who already have the disease).
    - Case fatality: fatal cases of a disease/people who already have the disease

- Prevalence: includes old and new cases; it is a proportion and not a measure of risk. Important to evaluate disease burden at a point in time and health planning. Number of cases of a disease present in the population at a specific time/number of persons in the population at that specified time.

-Factors that affect prevalence: Incidence of new cases, duration of disease, immigration and emigration, diagnosis differences, treatment (availability), reported number of cases, recovery and death.

- In a steady-state population in which the rates are not changing and in migration equals out migration, the following applies: Prevalence = Incidence x Duration of Disease.

→ Relative measures of disease association (Relative risks)

- Cumulative Incidence: measure proportion of people who develop new cases during specified time period. Exposed cases/total population at risk (assume everybody is followed for the entire time period)
- Incidence Rate: each individual contributes a measured time period to the denominator; time period vary for each person. Exposed cases/ total population at risk x years contributed by each participant at risk
- Cumulative Incidence Ratio: CumInc exposed/ CumInc unexposed

E.g: How to describe a CI of 1.3 – The incidence of disease X in exposed group was 1.3 times higher than in unexposed/ Exposure Y caused 30% increase in the risk of disease X. A CI < 1.0 describes a protective effect of an exposure. “CI = 0.33 – Unexposed have 3.3 higher risk of getting the disease (1/0,33)/ 70% lower risk of the exposed to get the disease.”

- Incidence Rate Ratio: IncRate exposed/ IncRate unexposed (increased risk per person years)
- Excess risk: CIR – 1 (x100%) (Percentage increase in incidence among exposed compared to unexposed).

→Absolute measure of association

- Cumulative Incidence Difference: CI exposed – CI unexposed

How to describe a CID of 0.0042 – There were 42 more cases per 10000 persons of disease X in exposed group compared to unexposed.
- Incidence Rate Difference: IR exposed – IR unexposed
How to describe a IRD of 0.00076/py – Exposure Y was associated with approximately 8 more cases of disease X per 10000 person-years compared to unexposed.

+Cross sectional study: snapshot of a population. You can only calculate PREVALENCE and ODDS RATIO.
+ Case control is a retrospective study. We can also use OR and PR calculation for it.
-In cross sectional studies we have random population at study and in case control we select people with disease (case) and non-diseased (control).
-With rare exposure, may be difficult to make CS and CC studies. So it'd be better to make longitudinal studies (cohort). But it has drawbacks such as longer follow up time, long latent period, larger sample size and probably more expensive.
- 2x2 table: outcome Y or N

Exposure				
	Y	A	B	Total exposed A+B
	N	C	D	Total unexposed C + D

Total diseased: (A + C) (B + D): Total not diseased

- Prevalence: prevalence in exposed over prevalence in unexposed A/(A + B) or C/(C + D) (looks like CI but there's no CI in cross sectional)
- Prevalence ratio = A/(A + B) / C/(C + D)
- Odds ratio: probability of having a disease over the probability of not having the disease in exposed and unexposed groups. OR –(A/B)/(C/D)
- Why use odds ratio: more used in logistic regression models. Also used in case control models. We don't know what comes first (exposure or disease). There is no temporality so, in this case, odds ratio give us association and not causation. Never forget to compare with the confidence interval.
How to describe an OR of 1.28 – Exposed group have 28% higher odds/probability of having the disease compared to the unexposed/ The odds of disease X in exposed group was 28% higher than in unexposed group.

→Rates: used to make comparison among groups more meaningful, rates may be used instead of raw numbers. A rate is defined as the number of cases of a particular outcome of interest that occurs over a given period of time divided by the size of the population in that time period.
- A crude rate is a single number computed as summary and disregards differences caused by age, gender, race and other characteristics. These aspects often have significant effect in the description of vital statistics.
- Rates well defined for specific subgroups are called specific rates. E.g: age specific death rate.

→Standardization: although subgroup specific rates provide a more accurate comparison among populations than the crude rates do, it would be more convenient to summarize the entire situation with a single number that adjusts for differences in composition.
-Direct method of standardization: compute the overall rates that would result if all populations had the same characteristics (same standard composition).

How to calculate it: 1)Select standard population for all of the subgroups; 2)Multiply the specific rate by the standard population of each stratum; 3) Sum up the number of deaths expected and divide by the total standard population.
- Advantages of DSMR: removes the effect of age when comparing disease occurrence between population or same population in different time points; weights applied to the stratum specific rates are the same for both populations – able to make comparisons with any rate calculated using the same standard population; comparisons across multiple sets of data because the same denominator is used.
- Disadvantages of DSMR: absolute value has no meaning, it does not relate to the real population; should ONLY be used to compare disease occurrence; depends on the choice of the standard population (should not be radically different).

- Indirect method of standardization: apply the specific rates of a known population to a population of interest under comparison, previously stratified by the variable to be controlled.
How to calculate it: 1) Multiply the specific rate of a known stratum to the same stratum of a population of interest. 2) Sum up the numbers of the expected number of deaths. 3) Calculate the Standardized Mortality Ratio (SMR): Divide the total observed number of deaths (real number of the pop of interest) by the total expected number of deaths (deaths with rates of a different population applied).
- How to interpret a SMR: =1 (risk is the same in both populations); <1 (risk is lower in the observed population compared to the reference population); >1 (risk is higher in the observed population compared to the reference population).
- Use it when we don't know what are the stratum specific death rates in one of the populations (e.g: to assess an occupational exposure to a risk factor, regions where the information is not recorded), if the numbers in some age groups are too small (we can choose rates from a large population to minimize the effects of sampling error). Used often to compare rates in a sub population with the general population.
-Advantages of ISMR: we do not need to know the stratum specific deaths of one of the populations, minimize sampling error if the numbers are too small.
-Disadvantages of ISMR: Weights applied are not the same for both populations; SMR cannot be compared between different populations (unless the age distribution is similar); does not use as much information from the study population as the direct method.

HYPOTHESIS TESTING

- Statistical inference is hypothesis testing. It is a statistical test to calculate the p value. Nearly all tests leads to a p value.
- Statistical test done in attempt to reject the null hypothesis. Also called test of statistical significance.
- The choice of the test depends on the study design (cross sectional, cohort, case control), the data type (continuous or nominal) and the distribution (normal, intervallic..). However, the interpretation of the p value is always the same. Commonly used when a comparison is made between two or more groups (e.g: height and weight of 2 different groups).
- **Efficacy**: if under ideal conditions, a drug has its effects/works.
- **Effectiveness**: if under real life circumstances, the drug has its effects/works.
*If the medication is not efficacious, it will never be effective.
- **Noncompliance**: epidemiologic term for those who don't keep up with the treatment.

- Iron supplementation for iron deficiency anemia: daily intake of 25mg cause collateral effects such as vomiting and diarrhea and is as effective as a weekly intake of 50 mg that doesn't cause collateral effects.
- To study this association: divide the population into two randomized groups (evenly distributed) in order to control for confounders.
- The idea of a study in this case is to assess whether the difference is likely to be due to the treatment rather than to chance (null hypothesis).
- A statistical test cannot demonstrate or prove the truth, it only provides evidence to reject or support the hypothesis.
- The null hypothesis (Ho) is expressed as there's no difference or relationship (careful with the word association) between the compared groups for the variable under study. G1 = G2
-The alternative hypothesis (Ha or Hi) is expressed as there is difference or relationship (careful with the term association) between the groups of interest for the variable under study.
- To run a statistical test, a null and alternative hypothesis need to be defined. The hypothesis testing will then explore whether the difference between the groups is likely to be due to chance (Null hypothesis).

→ Some concepts:

- Population: all possible values
- Samples: part of the population
- Statistical inference: generalization of the results from a sample to a population level with an estimated degree of certainty.
    - The truth lies within the entire population but as we can't study the whole population, we take samples but there's always sampling errors (never a true representative of the whole population). The smaller the sample, the higher the sampling error.
- Forms of statistical inference: hypothesis testing and estimation (confidence intervals). All good journals have been switching to CI. But there's a remaining usage of p values.
- Parameter: Numeric characteristics of a population (e.g: population mean or proportion). Expressed in greek letters (μ)
- Statistic: computed value from the data of a sample (e.g: sample mean or proportion). Expressed in roman letters (x).Try to have the same value as the parameter.
- Relationship between parameters and statistics: whole population → sample population → produce data → create statistics → statistical inference → parameter of a population.

-P value means probability and it can range from 0 to 1.

-Measure the strength of evidence against Ho. Measure the probability that an observed difference between comparison groups is due to chance.

- 0 = unlikely due to chance, real difference, due to random variation.
- 1 = likely due to chance, no real difference, not due to random variation.

- E.g: a p value of a statistical test is equal to 0.032. It means that the probability of making a type I error is 0.032. So you reject the null hypothesis, once it is very unlikely to happen (less than 3.2%)

- Hypothesis testing a parameter with evidence from sample data:

→operationalize the Ho and Ha (E.g: Ho – men and women have similar Hb levels. Ha – men and women have different Hb levels. Operationalize: Ho – men and women have anemia to the same extent. Ha – men and women have anemia to different extents.)

→Carry out statistical test

→Determine p value

→Interpretation

- Sample distribution of a mean: the sample distributions of a mean describe the behavior of a sample mean.
- The larger the standard deviation, higher the sample error.
- The probability that a p value has the null hypothesis true is correspondent to the area under the curve of the side of the normal distribution. The p value is normally given by computer software or can be obtained from a table or internet calculator.
- There's a table where you can see the corresponding value of the p value.
- Interpretation: you reject the null hypothesis if the conventional significance level of 95% (0,05) is set.
- The smaller P value, the higher evidence against the Ho.
- The higher P value, the lower evidence against the Ho.
- Convention:

		Truth	
		H_0 true	H_0 false
Decision	Retain H_0	Correct retention Confidence level 1-α	Type II error β
	Reject H_0	Type I error α	Correct rejection Power 1-β

$P > 0.10$ →No significant evidence against the Ho (Ho True)
$0.05 < P \leq 0.10$ →Marginal evidence against the Ho (Some association, Ho may be true)
$0.01 < P \leq 0.05$ →Significant evidence against the Ho (Ho false)
$P \leq 0.01$ →Highly significant evidence against the Ho (Ho highly false)

-Type I error (α): probability of rejecting the Ho when it is true. (reject when it is true)
-Confidence level: refers to the probability of correctly retaining the Ho when the Ho is true (1-α) (1-0,05 = 95% CI)
- Type II error (β) = probability of failing to reject the Ho when it is false (accept when it is false).
- Statistical power: refers to the probability of correctly rejecting the Ho when the Ho is false (1-β).
-The use of a threshold to determine statistical significance is becoming obsolete. Reporting the exact p value is now preferred. A p value of 0,05 means that the researcher would be wrong in 5% of the times (he rejects the null hypothesis when it is actually true – type I error)
- A non significant p value might also be a result of type II error (accept the null hypothesis when it is false). It can relate to small sample size (power).
- Increasing sample size, you decrease type II error and sample error.
- The main disadvantage is that the P value does not indicate the magnitude of the effect or difference.

→Two sided test: also referred as a two-tailed significance test. When the values to reject the Ho are located in both tails of the probability distribution. Ho: σ = or ≠ X
→One sided test: also referred as a one-tailed significance test. When the values to reject the Ho are located entirely in one tail of the probability distribution. Ho: p > X or p < X
-One tailed tests make it easier to reject the Ho.
-If the critical probability value is set at 0,05 in a two sided test, the probability is divided by 2, so that the critical value becomes 0.025 in each tail.

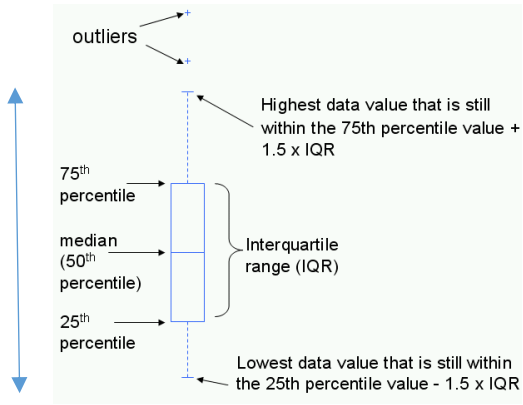
→Confidence interval versus point estimate: CI is a range of values which gives information on the strength of effect, duration, direction of the effect and sample size while the point estimate only describes statistical significance, a single number.

↙ ↑ ALPHA, ↓ BETA ↘

CONFIDENCE LEVEL ↓ ↑ POWER

→Increasing alpha with sample size fixed, increases power.
→Increasing the sample size while alpha is fixed, increases power.
→Small effect > large sample size. Large effect > small sample size.
→High variability > large sample size. Low variabilility > small sample size.

→ Box plot: Never forget to put scales near the boxplot.



→Interquartile range: value resulted from the difference of the 3rd minus the 1 st quartile
→Standard deviation: measure of how spread the numbers are.
→Range: distance between the highest and lowest value. Always an absolute measure.
→The median is a better central tendency measure when:

- ➔ The sample is small
- ➔ Sample positively skewed
- ➔ Sample not taken randomly
- ➔ Existence of several outliers (median does not reflect the presence of outliers, mean does.)

→Inflection point: where the curve turns (normal distribution)

PROBABILITY DISTRIBUTIONS: Table, equation or function that links each outcome of a statistical experiment with its probability of occurrence. Assigns probabilities to each value of a random variable.
→Frequency distributions: can be seen as probability distributions. We can calculate all the appropriate summary statistics. E.g: mean, median and variance. Graphic representation could be with bars or histograms and the area under the bar is the frequency and probability of the sample. Can be used with categorical or continuous variables.

→Binomial distribution: used with binary variables; represents the number of outcomes in *n* trials. Characterized by 2 parameters: *n* (number of trials) and *p* (probability of success in each trial – mean). Possible values range from 0 to n. The shape of the distribution with big *n* and *p* close to 0,5 is similar to the shape of a normal distribution curve. We assume that there's a fixed number of *n* trials which results in one of two mutually exclusive events (e.g: male or female), outcomes of *n* trials are independent and the probability of success *p* is constant for each trial.

"In general, if we have a sequence of *n* dichotomous trials, with constant probability of success *p* then the total number of success X is a random variable with binomial distribution."
Mean: μ = np;
Variance: σ^2 = np (1-p);
Standard deviation: σ = $\sqrt{[np\ (1-p)]}$

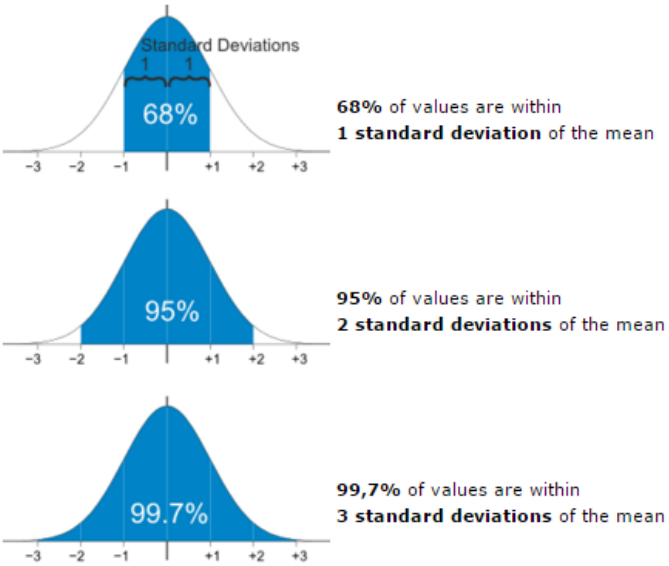
Understanding the binomial distribution formula:
To figure out the probability of EACH outcome → $p^x(1-p)^{(n-x)}$
To find out the total number of outcomes (how many probabilities of success)→ $\frac{n!}{k!(n-k)!}$

The general binomial probability formula:

$$P(X=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

In the graphic representation, we have on the X axis the number of possible values of the variables and on the Y axis we have the probability or frequencies of each.

→Normal distribution: with a large number of observations, the classes can be very narrow and the histogram can be well approximated by a smooth curve. The normal distribution is meant for continuous variables, defined from - ∞ to + ∞. Characterized by the parameters mean and standard deviation. Shape of the distribution graph is a bell shape and symmetrical, centered about its mean. We can transform a normal distribution to a standard normal distribution where μ = 0 and σ = 1. The letter Z denotes variables that follows that standard distribution. The probabilities associated are tabulated in statistical textbooks. Probability and proportion are the same in normal distribution.
-Things that closely follow a normal distribution: heights of people, size of things produced by machines, errors in measurements, marks on a test.
- The normal distribution has mode = median = mean.



- One standard deviation from the mean is also called sigma σ.
- Standardizing: convert a value to a z-score, first subtract the mean then divide by the SD.

$$Z = X - \mu/\sigma$$

Sampling distribution: two common statistics are the sample proportion and sample mean. These statistics are random variables and vary from sample to sample. As a result, sample statistics also have a distribution called sampling distribution. They also have a mean and standard deviation (now called standard error).

Central Limit Theorem: states that the sampling distribution of the mean of any independent random variable will be normal or nearly normal, IF the sample size is large enough (population distribution tends to normal, large enough is over 30; if population distribution is more skewed, has outlier or multiple peaks, there must be a higher sample size). The benefit of CLT is that allows the use of Z distribution to make estimations.

-When a sample of size *n* is selected from a population with mean μ and standard deviation σ, the sampling distribution of a mean has the following properties:

+ sample mean is equal to the population mean μ
+ standard deviation is called standard error $\sqrt{\sigma^2/n} = \sigma/n^{½}$
+ with large *n,* sampling distribution is near normal.
+ as the sample size increases, standard error decreases.
+ if sample size is small and finite, then the standard error is calculated as $\sqrt{\sigma^2(N-n)/n(N-1)}$
+ for a small sample and variable normally distributed, use the t distribution table instead of the z table.
+ small sample size, higher variability.

- A T distribution is a bit more spread than normal distribution but also symmetrical. If the variable is not normally distributed then we can't use it (use non parametrical tests such as Wilcoxon, Mann Whitney). For large samples, the t distribution becomes the same as the normal distribution (Z).

-We can also construct an interval about the population mean within which lie 95% (or 90 or 99.7%) of the sample means.

- Point estimation: population mean is unknown and use the sample mean as an estimator. This estimation is hardly exactly correct. That is why it is preferable to use the confidence interval.

- Interval estimation: reasonable values for a variable.

- Confidence interval: range of values within which the parameter of the population will lie with some level of confidence. It is built using the point estimate, adding and subtracting the margin of error.

-95% CI means that the true value of the population is included in that range. To decrease the CI, it is necessary to increase sample size (graph will be more narrow, concentrated near the mean therefore less need for a large range of CI).

-To construct a CI we would need the standard deviation of the population, which is usually unknown and that's why we use the standard error. P[X – 1,96 (σ/n ^{½}), X + 1,96 (σ/n ^{½})] ≈95%

-For 95% use 1,96; for 99% use 2,58; for 90% use 1,65.

- To find out the width of the CI: $w = 2(1.96 \times \sigma/\sqrt{n})$

If a sample size is small, to construct the CI we need to use the t distribution table and it depends on one parameter (degrees of freedom) which is obtained as the sample size minus one. (n-1). As the sample size increases, the t distribution tends to the standard normal distribution. With 50 observations, the t distribution is quite close to the normal distribution.

- We can find the confidence intervals for proportions and means. They follow the same rules.

- To define a normal standard distribution for a mean of the means or proportion:

$$Z \sim (X - \mu)/\sigma/\sqrt{n}$$

Where X is the sampling distribution mean, μ mean of the population, σ standard deviation and *n* is the sample size.

-Confidence intervals for the difference between means: two independent random samples n_1 and n_2 with means μ_1 and μ_2 and standard deviation σ_1 and σ_2, the sampling distribution of the difference $(X_1 - X_2)$ has the following properties:

+ The mean is $\mu_1 - \mu_2$;

+ Standard deviation is $(\sqrt{\sigma^2_1 / n_1 + \sigma_2^2 / n_2})$;

+ Provided that n_1 and n_2 are large, sampling distribution of the mean is approximately normal.

+ The 95% CI for the difference is the same but instead of X, use the mean difference:

$$[(\bar{x}_1 - \bar{x}_2) - 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2}, (\bar{x}_1 - \bar{x}_2) + 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2}]$$

- CI to estimate a population μ when σ is known: $\mu_{\bar{x}} = \bar{X} \pm Z\frac{\sigma}{\sqrt{n}}$ (use Z table);

- CI to estimate a population μ when σ is UNknown: $\mu = \bar{X} \pm t_{n-1}\frac{S}{\sqrt{n}}$ (use T table);

-CI for a proportion: $p \pm Z_{\alpha/2}(\sqrt{p\,(1-p)/n})$

-Hypothesis testing a large <u>sample test about one population mean:</u>

1) Determine the null and alternative hypothesis (Ho and Ha)
2) Decide the level of the test (usually 5%)
3) Draw a sample
4) Calculate the test statistic and p value
5) Draw conclusions in the context of the problem

-With a small sample, you don't have enough evidence to reject the null hypothesis.

-There's few power with small sample. To reduce type II error (accept Ho when it is false), increase sample size.

- To increase power: increase sample size and decrease standard deviation.

- Formula to use: $Z \sim (X - \mu)/\sigma/\sqrt{n}$

-<u>Sample testing of one population proportion</u>: find p value. Same formulas $Z \sim (X - p)/\sigma/\sqrt{n}$ (p is the proportion in the population instead of the mean in the population, which will be the same thing when it comes to proportions)*Don't confound it when trying to find the CI!

-Determine the null and alternative hypothesis, set a two sided test at 5% and draw the sample. Remember the both sides of the curve so add (Z > standardized found value + Z < standardized found value). Reject the null hypothesis if result is smaller than 5%.

- For proportions, knowing the population proportion implies knowing the population standard deviation. So we could have also calculated like this: $\sigma = \sqrt{[p(1-p)]}$

- <u>Sample testing two population means (paired data)</u>: calculate the difference of the means between one same group at different times or 2 related groups. The data obtained must be quantitative and randomly selected. Ho state that the mean difference is equal to 0 or mean of group 1 is equal to mean of group 2. The calculation is similar to the calculation for a single sample mean: $z = (Xd - \mu 0)/(\frac{Sd}{\sqrt{n}})$. Where Xd is the sample mean difference; $\mu 0$ is the mean difference specified in null hypothesis (=0); Sd is standard deviation of the differences; n is sample size (number of pairs).

-Sample testing two independent population means: calculated the same way as before but the formula now is $z = (Xd - \mu 0)/(\frac{\sqrt{s1^2}}{\sqrt{n1}} + \frac{\sqrt{s2^2}}{\sqrt{n2}})$. With small samples, we compare the test statistic with t distribution.

-**Chi square test**: comparisons between two categorical variables. We use it to measure the differences between what is observed and what is expected according to an assumed hypothesis. "Are the differences that we see in the table enough to reject the null hypothesis?" Main characteristics of the chi-square test: *Data must be in the form of frequencies; *Expected frequency in any cell of the table must be greater than 5; *Total number of observations must be greater than 20.
-Cross tabulation: expected frequency = row total x column total / grand total. Means the expected number of observations in case there's no relationship between these 2 values.

-Formula: $$X^2 = \sum \frac{(obs - \exp)^2}{\exp}$$

Where X^2 is the value of the chi square
Obs is the observed value (real value in each cell)
Exp is the expect value found in the previous formula
Sum it all up and square then added together.
-Minimum value is 0 and maximum is infinite. As high as the expected is, bigger the amount and higher is the cut off level.
-Degrees of freedom = (number of rolls – 1)x(number of columns - 1). Increased number of categories, increased DFs.
- Graphic representation: total are under the curve is 1. If the difference between expected and observed is similar, difference is low, the curves would be similar and you don't reject null hypothesis. If it is very different, the curves are separated, you reject null hypothesis.
- With the result of X^2, you check the significance table (degrees of freedom by significance level). Use the formula for DFs and compare with the critical value. If our X^2 value is greater than the critical value, we reject null hypothesis.
- An alternative for the chi square test for association between categorical variable with small samples: **Fisher's exact test.**
- Result is valid for any sample size and is exact in the probabilistic manner.

T test: used to compare means of populations.
Procedure for significance testing:

1) State hypothesis (null and alternative): Ho: $\mu_1 = \mu_2$ / Ha - $\mu_1 \neq \mu_2$
   (Ho means no effect, no association, no difference. Sometimes called test hypothesis because that's the only hypothesis we test).
2) Set alpha level (p value): α = 0,05 (or 0.01, 0.10).
3) Calculate the test statistic (sample value t) – compare means
4) Find the critical value of the statistic - t_v (value that we'll compare with the t value found and decide if it is significant or not). The v (niu) is calculated with the degrees of freedom (N – 1) where n is sample size. For a two sided test, each side corresponds to 0,025 (both sides are 0,05). Check table to find the t_v.
5) State decision rule: $|t| > t_{v, 0,025}$, reject Ho, accept Ha. Otherwise, accept Ho.
6) State conclusion.

-T distribution table: in one side we find the DFs (n – 1) and on the other we find the alpha. If it is a two sided test, divide alpha by two. With these two parameters in hand, check the corresponding $t_{v.}$.

- Distribution of a sample: we always suppose a population is normally distributed: μ = mean and σ = standard deviation. A sample will also be normally distributed and have X as the sample mean and s as the standard error (sample SD).

-Graphic representation: the shape of a t distribution curve is similar to a normal distribution. It only has one parameter, which is the degree of freedom. The mean is always 0. The larger the numbers of DFs, the flatter the distribution will be. The smaller DFs, steeper it'll be.

+ A normal distribution is a special kind of t distribution where DF equals to 1.

-One sample t test: you have one population and draw samples out of it. To confirm that the sample means is the same as the population mean, perform this calculation. The hypothesis and null hypothesis follow the same pattern Ho: μ = 0 (or stated value) and Ha: μ ≠ 0 (or other stated value), set alpha (usually 0.05), calculate test statistic. With large sample size, the sampling error may be due to chance.

Formulas: SE = SD/√n (Standard error = Standard deviation/sqrt of the sample size). SE measures the standard sample error.

*SD = $\sqrt{\Sigma(X^1 - X^2)^2}/\sqrt{n}$

t_x = X - μ/SE_x where the numerator is the sampling error and the denominator is the standard error. X is sample mean, μ is population mean, SE is the standard error. The meaning of t value is "how many times of sampling error compared with standard error".

-You determine the critical value by looking up in the table with the DFs and the value for alpha. If the absolute value of t is higher than the critical value, reject null hypothesis. The conclusion is: the difference between the sample mean and population mean is significantly different that it is unlikely that the sample was drawn from that population.

→IN SPSS: analyze – compare means – one sample t test. Decide the test and grouping variable. It does not give us the critical t value only the t value, DF and p value. If the p value is lower than 0,05. Reject null hypothesis. Therefore you know that the critical value was smaller than the t value.

T test with two independent samples: used when we have two independent samples (treatment and control, for instance).

Formula: $t_{x1\text{-}x2} = X_1 + X_2 – (\mu_1 - \mu_2) / SE_{diff}$ (X1 and X2 are the sample means)

- If we are testing the null hypothesis we can say that the mean difference ($\mu_1 - \mu_2$) is 0.
- In the numerator we have the sample means and in the denominator we have the standard error of the difference between means.

$SE_{diff} = \frac{\sqrt{SD1^2}}{\sqrt{N1}} + \frac{\sqrt{SD2^2}}{\sqrt{N2}}$

-State null and alternative hypothesis (Ho: $\mu_1 = \mu_2$ and Ha: $\mu_1 \neq \mu_2$), set alpha, determine critical values (now we have 2 groups so N1+N2 – 2 to find the DFs) and look up the table for the critical t value.

- Same way as before, if the absolute value of t is higher than the critical t value, reject null hypothesis/population means are different.

→IN SPSS: analyze – compare means – independent samples t test – add grouping and test variable – define grouping variables. For independent samples, SPSS will also not give us the critical t value. Instead it will show us the Levene's test for equality of variances.

-Levene's test for equality of variances: if the p (Sig = significance level) displayed is higher than 0.05 we can assume the variances are equal (normally distributed) so we check the p value of

the first row (Equal variances assumed). Otherwise, if the p displayed is lower than 0.05 equal variances are NOT assumed and we check the second row for the p value.
*The null hypothesis for levene's test is that the variances are equal and that is why when the p value is lower than 0.05, we assume the variances are not equal (not enough evidence to support Ho).
T test with two dependent samples: use when we have dependent samples – matched, paired or tied. Used to control individual differences. Can result in more powerful test than independent samples t test.
Formulas: t_x = D/SEdiff where t is the difference in means over a standard error.
SEdiff = $SD_D/\sqrt{npairs}$ where SD_D is the difference among the standard deviations of the 2 groups ($\sigma_1 - \sigma_2$). Divide this value by the sqrt of the number of pairs to get SEdiff.
-State null and alternative hypothesis (Ho: $\mu_1 = \mu_2$ or $\mu_d = 0$ and Ha: $\mu_1 \neq \mu_2$ or $\mu_d \neq 0$), set alpha, calculate the test statistic, determine the critical value of t (Number of pairs -1). If the absolute t value is higher than critical t value, reject null hypothesis.
→ IN SPSS: analyze – compare means – paired samples t test - choose the paired variables you want to compare. There will be the t value, df and p value (critical t value doesn't show up, as see before).
-The smaller the sample size, wider the confidence interval.

Anova (Analysis of variance): extends independent samples t test (one way) for normally distributed numerical variables. *Two way will not be studied in this course but it is regarding extension of dependent samples t test.
- Compares the means of groups of independent observations (Not the variance; don't be fooled by the name). Can compare three or more groups. Anova assumes that, in principle, the variances are equal (null hypothesis). It is a relatively robust procedure with respect to violations of the normality assumption (relatively uninfluenced by violations in their assumptions).
- If the sample contains K independent groups. The formulation of null and alternative hypothesis for anova would be: Ho - $\mu_1 = \mu_{2..} \ldots = \mu_k$ and Ha: $\mu_i \neq \mu_j$ (or the group means are not all equal).
- Combination of mean of the means. Suppose you have 3 groups with 20 samples each, you will have a mean for each of the 3 groups and for the 3 groups combined. If there's a big difference between the distance of one group mean to the mean of the means, we can't say that the means are similar.
-Sum of Squares Between Groups: SSB = $n(X_1 - X) + n(X_2 - X) + n(X_3 - X)$.. (where n is number of observations in each group; $X_{1,2,3..}$ is mean of each group and X is mean of the means). Combine the differences from the grand mean.
-Mean Square Error: estimates the variability of individual observations. MSE = $\Sigma \frac{(Xi - Xj)^2}{N-K}$; were N is number of observations, K is number of groups; $X_i - X_j$ is the difference within groups.
-Anova assumes that all the group variances are equal. Consider other options if group variances differ by a factor of 2 or more.
-The anova F test is based on the F statistic: F = $\frac{\frac{SSB}{(K-1)}}{MSE}$; where the numerator is the variance between groups and the denominator is the variance within groups.
-If the numerator is large (F >1) we can say there is difference between groups; reject null hypothesis. But if the variance between groups is smaller or equal within, we won't have enough evidence to say there is difference, therefore we accept null hypothesis.

-Parameters of the F statistic: K-1 and N-K (2 degrees of freedom). To get a p value we compare our F to the F statistic of these two degrees of freedom.

-Results are often displayed using an ANOVA table:

	Sum of squares	**DF**	**Mean square**	**F**	**Sig.**
	(Individual levels – individual means)²		Sum of squares/ DF	F value (use this)	P value
Between Groups	SSB (use this value)	N-1			
Within groups		N-K	MSE (use this value)		

→IN SPSS: analyze – compare means – one way anova – choose the dependent list and factor. They will present us with a similar table as above.
-If the test is significant, we can assume that there's difference between the groups. We must assess next which groups are different. In order to do so we use the post hoc tests for multiple comparisons.
-Each time a hypothesis test is performed at significance level α, there is probability α of rejecting in error. Performing multiple tests increases the chances of rejecting in error at least once.
-**The Bonferroni Correction** performs each test at significance level α. You have to multiply each p value by the number of tests performed [n(n-1)/2)]. The overall significance level (chance of any tests rejecting in error) will be less than α.
→IN SPSS: analyze – compare means – one way anova – choose dependent variable and factor variable – click on post hoc – select bonferroni.
*SPSS takes the sig. value of the LSD test and multiplies by the number of tests performed [n(n-1)/2)] and presents this value as the Bonferroni significant value. It will present us a table comparing every group. If the p value is higher than 0,05 in any of the comparisons, then there is NO difference between the groups, they are equal.

NON PARAMETRIC STATISTICS:
-In parametric statistics we assume that the data collected come from a type of probability distribution and we can make inferences about the parameters of this distribution (normal, binomial, chi square). We assume that the population studied follows a normal distribution, has same variance and so on. When we calculate the p value for an inference test, we find the probability that the sample was different due to sampling variability (by chance), we are assuming that all samples of the given sample size are normally distributed around the mean. This is why the test statistic, which is the number of standard deviations that sample mean is away from the population mean is able to be used therefore, without normality, no p value can be found.
- There is a problem with parametric statistics: when there is lack of normality none of the tests are reliable. The sampling distribution won't follow t or z distribution. The way statisticians deal with this problem is the non-parametric statistics.
- For non-parametric statistics we don't need assumptions of normality (symmetry, mean, standard deviation).
- The parameter to deal with is the MEDIAN. A mean can be easily influenced by outliers or skewness, and as we are not assuming normality, mean no longer makes sense. The median is another judge of location, considered the center of the distribution. The sample data receives a

rank and these ranks create a test statistics. Do not involve any population parameters, the data can be measure on any scale (ratio or interval, ordinal or nominal).

- Similarity between parametric and non-parametric tests:

PARAMETRIC TEST	**Goals for PT**	**NON PARAMETRIC TEST**	**Goals for NPT**
One sample T test	Test hypothesis about the *mean of a population* where the sample was taken from.	**Wilcoxon Signed Ranks Test**	Test hypothesis about the *median of the population* where the sample was taken from.
Two sample T test	See if two samples have identical population *means*	**Mann Whitney Test**	See if two samples have identical population *medians*
Chi Square Test	See if a sample fits a theorical distribution, such as the normal curve.	**Kolmogorov – Smirnov Test**	See if a sample could have come from a certain distribution.
ANOVA	See if *two or more sample means* are significantly different	**Kruskal-Wallis Test**	Test if *two or more sample medians* are significantly different.

- Advantages of NPT: robust procedure (used with all scales), easier to compute, make fewer assumptions, doesn't involve population parameters and the results may be as exact as parametric procedures.
- Disadvantages of NPT: may waste information, lose power (there is greater risk of accepting false null hypothesis; increased risk of committing type II error), null hypothesis is somewhat loosely formulated.
- **Measurement of normality**: skewness = 0 and kurtosis = 3. First test to perform before going to the test statistics. We set a null and alternative hypothesis, where the Ho states that the data is normally distributed and Ha states that the data is not normally distributed. There's 3 possible graphical methods: histogram (not suitable for small samples), Q-Q plot and P-P plot.
-For histogram in SPSS: Graphs – Legacy Dialogs – Histogram – add normality curve.
- Q-Q plot: quantile-quantile plot. It is a plot of quantiles of the first data set against the quantiles of the second data set. If the data is normally distributed then the data points will be close to the diagonal line. If they stray away from the line in an obvious non-linear trend then the data is not normally distributed. A quantile is a point below which a given fraction or percent of points lies. For instance, the 0.3 or 30% quantile is the point where 30% of the data fall below and 70% above that value.
- A Q-Q plot compares the quantiles of a data distribution with the quantiles of a standardized theorical distribution from a specified family of distributions. Use this plot if your objective is to compare the data distribution with a family of distributions that vary only in location and scale, particularly if you want to estimate the location and scale parameters from the plot. Q-Q plots tend to magnify deviations from the normal distribution on the tails, spot non normality better on the tails.
- In SPSS: analyze – descriptive statistics – Q-Q plot – select variables of interest into the variables box – test distribution set to Normal. Analyze relation between the diagonal line and the little

circles. SPSS also gives you the detrended normal Q-Q plot which is a turn of the first one in 45°. It shows the differences between the observed and expected values of a normal distribution. If the distribution is normal, the little circles must the spread around the horizontal line with no pattern.

-P-P plot: probability – probability plot or percent-percent plot. Plots a variable's cumulative proportions against the cumulative proportions of any number of test distribution. Generally used to determine whether the distribution of a variable matches a given distribution. If the selected variable matches the test distribution, the points cluster around a straight line.

- A P-P plot compares the empirical cumulative distribution function of a data set with a specified theoretical cumulative distribution function. An advantage of P-P plots is that they are discriminating in regions of high probability density, since in these regions the empirical and theoretical cumulative distributions change more rapidly than in regions of low probability density. P-P plots tend to magnify deviations from the normal distribution in the middle, spot non normality better around the mean.

-In SPSS, procedure is the same as the Q-Q plot.

-There are numerical tests to check for normality of distributions such as Kolmogorov-Smirnov test and Shapiro-Wilk.

-**Kolmogorov-Smirnov test (K-S test)**: non parametric test for the equality of continuous, one dimensional probability distributions that can be used to compare a sample with a reference probability distribution or to compare two samples. General test that detect differences in both the locations and shapes of the distributions.

- IN SPSS: analyze – descriptive statistics – explore – select dependent list (variable)– click plots and select normality plots with tests and uncheck everything for descriptive. It gives you 3 boxes, one with cases, second with descriptive analysis such as CI, mean, median, variance, standard deviation and the third with two tests of normality K-S and Shapiro Wilk.

*If the dataset is smaller than 50 or larger than 2000 elements, use the Shapiro Wilk results. If not, use the results of K-S test. If the p value is higher than 0.05, we reject alternative hypothesis and conclude that the data comes from normal distribution.

- For testing 2 or more samples, the normality test needs to be made sample by sample. Therefore, add your independent variable to the factor list. In the output you will have the results for the different groups.

- You could also split your independent variables: data – split file – click organize output by groups – select the variables on "group based on" – check sort the file by grouping variable – ok – do the explore analysis again. This way you will have the different groups sorted out by the independent grouping variable you split the file.

- An advantage of the numerical test over the graphical is that the judgment is very objective. And an advantage of the graphical over the numerical is that sometimes the numerical is not sensitive enough at low sample sizes or overly sensitive to large sample sizes, therefore, the graphical provides good judgment when numerical doesn't. Some statisticians prefer to use their experience to make a subjective judgment about the data from graphs. But if you do not have great deal of experience, then it may be better to rely on numerical methods.

- Homogeneity of variance test: only when the assumption of homogeneity of variances is valid, we can use variances to be pooled across groups to yield an estimate of variance that is used in the standard error calculations. If this assumption is ignored, the results of the statistical test are greatly distorted leading to incorrect inferences and resulting in type I error (reject when null hypothesis is true).

- How to test the homogeneity in spss? With Levene's Test of equality of variances which is produced in spss when running the independent t test. This test provides an F statistic and a

significance value (p value). We are primarily concerned about the sig level. If it is greater than 0.05 than our variances are equal. However, if p is lower than 0.05, then we have unequal variances and we have violated the assumption of homogeneity of variances.
- You could also test it for one way Anova in spss: in the one way anova – click option – select Homogeneity of variance test – select Brown-Forsythe or Welch in the statistics area.
- It will give you a box with Test of Homogeneity of Variances with the Levene score and p value. (lower than 0,05 –variances not assumed). If there was a violation of the assumption, we can still determine whether there were significant differences between the groups. Not using ANOVA, but with the Brown-Forsythe or Welch test. If the p value resulted is less than 0,05 then there are statistically significant differences between groups. If the similar variances are assumed, there will be obviously no need to consult this table.
→Caution: non parametric tests such as Kruskal-Wallis or Mann-Whitney U tests, even though they do not assume normal distributions, they assume that the shape of the data distribution is the same in each studied group. So if you have very different standard deviations, not appropriate for ANOVA, they should not be analyzed by these two non-parametrical tests. Often the best approach is to transform the data to logarithms or reciprocals, restoring the equal variance.

- Robust procedures are tests and estimates that are relatively uninfluenced by violations in their assumptions.

-**Wilcoxon signed rank test**: for 2 related medians. It requires that the differences are approximately symmetric and that the data is measured on an ordinal, interval or ratio scale. When the assumptions for the Wilcoxon signed ranks test are met but the assumptions of the t test are violated, the Wilcoxon is usually more powerful in detecting differences between the two populations. Even under appropriate conditions to the paired t test, the Wilcoxon signed ranks test is almost as powerful. It also considers information about both the sign of the differences and the magnitude of the difference between pairs, meaning it incorporates more information about the data.
-Steps:
1) set null and alternative hypothesis: Ho – Median difference is 0. Ha – Median difference is different than 0.
2) Calculate the difference between the values (B-A) and median difference.
3) Rank the absolute values of differences affixing a sign to each rank. If there's absolute differences with the same value then they will have the same rank: add up the actual rank number they would get and divide by two, then continue rank count.
E.g: **1**-0,1; **2**-0.2; **3.5**-0,5; **3.5**- 0,5; **5**-0.7; **6**-0.9..
4) Calculate the sum of the ranks for the positive and negative values (W^- and W^+). The lowest value will be the one used to consult the critical values of W table.
5) Consult a table of critical values of W for the required alpha level (usually 0.05) and the number of difference (not sample size). If the obtained value for W is greater than the value shown in the table, the null hypothesis should be retained; if less, it may be rejected and the alternate hypothesis accepted at that level of significance. (It is the opposite of the anova and the critical t value).
- IN SPSS: your data must have 3 variables (the cases and different values –drug A and drug B- so we can compare the effect of different exposures in the same cases). You can initially create a variable called difference between the two exposures assessed (B-A). Then go to analyze – nonparametric tests – legacy dialogs – 2 related samples – select the variables you want to compare in the test pairs – select Wilcoxon as the test type and Ok. SPSS will give both ranks

(negatives and positives) with the mean and sum for each. Also will give a p value and if the value is lower than 0.05 you can reject null hypothesis (W found is lower than critical W value, but spss does not give us this value).
- If your data is binary, use the Mc Nemar Test (typically used in repeated measures situation, before and after a specified event occurs, and determines whether the initial response rate is equal to the final response rate).
- If your data is categorical, use the marginal homogeneity test (tests for changes in response and is useful for detecting response changes due to experimental intervention in before and after designs).

- Mann-Whitney U test: for 2 independent medians. Most popular of the two independent samples test. The null hypothesis is that two independent samples of scores could have been drawn from the same population. This test tells us whether the difference between the samples are so great to make it unlikely that the null hypothesis is correct. Used when there's a requirement to test the difference between two samples of data; samples are independent (each participant contributes only one value to only one of the two groups); the values represent measures either ordinal or interval scales; population distribution unknown or non-normal. It is a non-directional hypothesis, you can only tell if there is difference or not and not about effectiveness. Two tailed test.
-Steps:
- Set null and alternative hypothesis.
- Merge the scores and rank them. Using the same assumption for equal scores of the Wilcoxon test (for tied scores give them the average of the ranks).
- Sum up the rankings in each group independently. Whichever group has the greater sum of ranks will also necessarily contain most of the higher scores and the two medians will be different. The sum of all ranks for 2 samples combined must be equal to [n (n+1)/2].
- Now calculate the value U. It is determined as the number of times that a score from one set of data has been given a lower ranking than a score from the other set. As there is two independent samples we will have two U values, calculated by the following formula:
Ux = Rx – (nx(nx+1)/2; where Ux is the U value for each of the groups, R is the sum of ranks of each group; nx is the sample size of each group. You will have two U values, the lowest one is the one that'll be used to consult a significance table.
-Consult a table of critical values of U for the required alpha level and sample size. The table is composed by the sample size of the largest sample and size of the smallest sample. Then check column and row for the critical U value. If the obtained value for U is greater than the value shown in the table, the null hypothesis should be retained. If it is less, then the null hypothesis will be rejected (same logic as for Wilcoxon).
- IN SPSS: analyze – non parametric tests – legacy dialogs – 2 independent samples – move dependent variable to the Test variable List and the independent variable to the grouping variable – make sure that the mann whitney u is ticked in the test area – select which grouping variable you want to compare in Define groups. SPSS will give you the descriptive table (not very useful for this matter), a rank table with the sum and mean of ranks (indicates which group had highest mean rank) and lastly a test statistics table with the U value and significance level (p value).

-**Kruskal-Wallis test:** for more than 2 independent medians. It is an extension of the Mann-Whitney U test, is the non-parametric analog of one way analysis of variance and detects differences in distribution location.

-Steps:
1)Rank all of the scores merged. Lowest scores gets the lowest rank. Tied scores get the average of the ranks that they would have obtained, had they not been tied.
2) Find the total of the ranks for each group. Just add together all of the ranks of each group in turn.
3) Find H by using the following formula: H = = $\left[\frac{12}{N(N+1)} * \Sigma \frac{Tc^2}{nc}\right] - 3 * (N+1)$; where N is total number of participants (groups combined), Tc is the rank total for each group and nc is the number of participants in each group.
4) The degrees of freedom is the number of groups – 1.
5) The significance of H will be decided based on the number of participants and the number of groups. If you have 3 groups with five or fewer participants in each group, then you need to use the special table for small sample sizes. If you have more than 5 participants per group, then use the chi square table. H will be statistically significant if it is equal to or larger than the critical value of chi square for your particular DF. Thus, p value will be lower than 0,05. (Meaning of the p value: our value of H will occur by chance with a probability of less than 0.05)
- In the table for small sample size you have K (number of groups) per sample sizes and alpha level. They give you multiple combinations of sample sizes. For instance, if K= 3, they will have 3/2/1 or 3/3/3 or 4/3/3 and so on.
- In the Chi Square table you choose the DF (k-1) and the p value (0.05, 0.01..).
6) Conclude (or not) that there is a difference between the groups. But which groups are different?
-Conduct post hoc test: compare 2 groups at a time and check for the significance (pair wise test). With independent samples, use 3 times the Mann-Whitney u test.
-IN SPSS: analyze – nonparametric tests – legacy dialogs – K independent samples – transfer the dependent variable to Test variable List – independent variable to grouping variable – check for kruskal wallis H test – in Define groups choose the 3 groups you want to compare by range 1-3 (if you have more than 3 groups, you may have to rearrange them) – check for descriptive or quartiles, if you want to. SPSS gives you the ranks for each groups with mean ranking and a test statistics box with the chi square value (which will be the H value), the DFs and the p value (never the critical value in which we compare our H value). We may also have a box plot showing the differences between the groups. If the p value is lower than 0.05 we know we can't assume the null hypothesis, where the groups medians are equal, but we still must find which groups are different from the other (conduct post hoc test/ pairwise multiple comparisons).
- In order to do so, you need to run separate MWU tests on the different combinations of the related groups. After you get the results from the MWU, you need to use a Bonferroni adjustment on the results. When you are making multiple comparisons it is more likely that you will declare a result significant when you should not.
-The Bonferroni adjustment is easy to calculate: take the significance level you were initially using (0.05) and divide by the number of tests you are running [n(n+1)/2]. The p value larger than the result you get from that will not be significant (meaning the groups are equal; if the p value is lower then we reject null hypothesis – the groups are different). We are doing our own PostHoc test on this more rigorous level.
- **Friedman's ANOVA:** differences between several related groups. It is a non-parametric analogue to a repeated measure ANOVA where the same subjects have been subjected to various conditions.
-Steps:

1) Take the variable data of the different but related groups and rank per group (you would have 1~n ranks, where n is number of groups per independent variable). Sum up the ranks per group (Ri).

2)Calculate the test statistic Fr with the formula: $Fr = \left[\frac{12}{Nk} * (k+1) * \Sigma Ri^2\right] - 3N(k+1)$;

Where N is the sample size in each group, k is number of groups and ΣRi is sum of the ranks for each group.

3) Compare the Fr statistic to a chi square distribution. If our F value is higher than the chi square given value there is statistically significant difference. If it is smaller, no evidence to reject null hypothesis will be observed (the medians are equal between the groups).

-The probability distribution of Fr can be approximated to a chi square distribution. But if the n or k is small, the approximation with chi square becomes poor and the p value should be obtained from specific Fr tables for the Friedman test.

-If the p value is significant (there is difference between the groups), post hoc multiple comparison tests can be performed.

- IN SPSS: analyze – nonparametric tests – legacy dialogs – K related samples – move the dependent variables to the Test variable box – make sure that Friedman is selected. SPSS will give you a descriptive statistics table (if you have selected it on the options), a ranking table showing the mean rank for all the related groups and a test statistics table providing the chi square value (Fr value), DFs and the significance level (if higher than 0.05, no statistically significant difference were found).

-If there were differences, run Wilcoxon test to compare the medians. But corrected to be more rigorous. If there are 3 comparison groups then α/3 = 0.0167 significance level. To be significant, the p values Wilcoxon gives us must the lower than this.

-In SPSS: analyze – nonparametric tests – legacy dialogs – 2 related tests – choose Wilcoxon – specify the 3 different pairs of groups. SPSS will give us the mean ranks and sum of ranks for all 3 comparisons and the other table will give us the test statistics with the p value. Check if it is lower than the rigorous level (0.0167 if it is 3 comparison groups) – that is the Bonferroni correction, again.

COVARIANCE and CORRELATION: they both describe how two variables are related.

Variables are positively related if they move in the same direction ↗

Variables are inversely related if they move in opposite directions ↘

*If the linear trend is positive, the covariance will be positive. If the linear trend is negative, the covariance will be negative.

Both indicate whether variables are positively or inversely related but correlation also tells you the degree to which the variables tend to move together.

-Formula for covariance? $COV(x,y)$: $\frac{\Sigma(xi-\mu x)(yi-\mu y)}{n}$

Where x is the independent variable (X axis) and Y is the dependent variable (Y axis), μx or μy are the means for each variable and n is number of population/sample. The covariance will simple tell you the direction of the association. Important: it does not have a standard unit of measurement; scale dependent.

- The correlation analysis is used to quantify two continuous variables. The association could be between an independent and dependent variable or two independent.

-If you want to determine how two variables are related, you can use correlation that will also tell you the degree (quantifies direction and strength of the linear association) to which the variables tend to move together but it does not imply causation. Correlation standardizes the

measure of interdependence between two variables and how closely the 2 variables move. The correlation coefficient ranges from -1 to +1, where:
+1 – The variables have a perfect positive correlation. If one variable moves a given amount, the second moves proportionally in the same direction. The strength of the correlation grows as the value approaches one.
0 – no relationship exists between the variables, uncorrelated. Graphic is a straight line.
-1 – The variables are perfectly negatively correlated and move in opposition to each other. If one variable increases, the other decreases proportionally. The strength grows as the value approaches -1.
*The sign indicates direction and the magnitude indicates strength of association.
-Graphical displays are useful to explore associations between variables. Not always two continuous variables will have a linear association.
- For parametrically distributed variables: Pearson correlation. $\rho = Cov\frac{x,y}{\sigma x \sigma y}$
-Test of ρ: T = ρs – 0/ s.e(ps), where the Test statistic value is the ρ of the sample divided by the standard error of the sample. If the CI level doesn't include 0, stronger the correlation will be. Even if the correlation is negative.
-For non-parametrically distributed variables: Spearman Correlation. However, we lose power when using a non-parametric test because it requires less information and gives you less, it is more conservative (you need more evidence against null hypothesis).
-In order to calculate it, you have to rank both variables; create a variable of the differences of the two ranks and square the difference. Then use the following formula: $\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$; where d^2 is the squared difference and n is the number of observations.

REGRESSION ANALYSIS:
R.A is a related technique to assess the relationship between an outcome variable and a risk factor or confounding variable. The outcome variable can be called response or dependent variable (denoted by y) and the risk factor or confounder can be called predictor, explanatory or independent variable (denoted by x).
-When there is a single continuous dependent variable and a single independent variable, the analysis is called a simple linear regression analysis. It assumes that there is a linear association between the two variables and you can make predictions out of it.
-A regression line is simply a single line that best fits the data "least squares regression". MEANING: smallest overall distance from the line to the points (least distance between the observed and corresponding values). There are some formulas that help you draw a line that defines the follows the logic "line that minimizes the squared vertical distances between the data points and the line".
-Never do a regression analysis unless you have already found at least a moderately strong correlation between the two variables. Before moving on to find the equation for the regression line, we have to identify which variable is x and which is y. Generally, Y is the one you want to predict and X is the one you are using to make the prediction.
-The formula for the best fitting line is: y = α + βx, where:
* α is the y-value when x=0 (point where the line meets the y axis), also called Y-INTERCEPT.
* β is the change in Y over the change in x, also called the SLOPE. Eg: a slope of 10/3 means as the x value increases (move right) by 3 units, the y-value moves up by 10 units. The slope is negative when the line decreases.

Formula for β = $\frac{\Sigma(xi-\mu x)(yi-\mu y)}{\Sigma(xi-\mu x)^2}$ (numerator of covariance/variance of x) OR β = ρ(Sy/Sx) (that formula is better for two numerical variables), where ρ is the correlation between X and Y and Sy and Sx are the standard deviation of the x and y values.

-Some assumptions for linear regression models:

+ y (dependent variable) is normally distributed (check it with histogram or QQ-plot) and homoscedastic (have the same standard deviation in different groups). If there is a problem in this matter, you can try to log transform the Y variable.

+ assume that the data fit a straight line (linearity)

+ mean change in y per x-unit does not depend on x (that is the definition of linearity: the change in y is constant whenever there's a increase in one unit in x = β; you can check it with a scatter plot)

+ the variance for y (=σ^2) does not depend on x (the variability of y does not change when x increases; also check it with scatter plot).

-Ho would be β = 0 (as the X variable gets larger, the associated Y variable get neither higher nor lower) and Ha β ≠ 0.

-Estimation of μ: expected value of y when we have an observed value of x. Model: μ = α + βx. Explanation: if we have already set our α and β values, we just replace the x value to find the correspondent in y (=μ).

-CI and tests for α and β: with test variables t-distributed under Ho with linear regression we can have $\frac{\hat{\alpha}-\alpha_0}{s.e.(\hat{\alpha})}, \frac{\hat{\beta}-\beta_0}{s.e.(\hat{\beta})}$

-S.e of α and β are obtained from statistical softwares.

-Variability of Y can be explained by the error (distance between what we observed and what is estimated and we wan this to be as small as possible) and the regression model (want it to be as high as possible – distance between the estimated value and the mean). "How much of the variability we can explain by the exposure?"

Formula:

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$$

Where we have the Total Sum of the Squares (observed values of Y – mean value of y) = Sum of Squares due to Error (observed values of Y – value of Y predicted by the model) + Sum of Squares due to Regression (value of Y predicted by the model – mean value of Y).

This formula could be read as: the total variability of the dependent variable, corrected by its mean, splits into two sources: regression model and the error.

-The larger SSE compared to SSR, the poorer the fit is going to be.

-Coefficient of determination (r^2): expresses the strength of the relationship between the X and Y variables. Proportion of the variation in the Y variable that is explained by the variation in the X variable. The value ranges from 0 to 1 where values near 1 mean the Y values fall almost right on the regression line, while values near 0 mean there is very little relationship between X and Yand a nice result depends on what you are trying to assess. The higher the variability, the lower the r^2 value will be. The value of R^2 is the goodness of fit.

- "How much of the probability of the outcome is explained by the exposure?" r^2 = SSR/TSS.

- r = βs*Sx/Sy; where r is the correlation coefficient of the sample, βs is estimation of the slope and S is the standard deviation. This is the link between pearson's correlation, least squares estimator of β and the r^2.

DIAGNOSTIC TEST: Sensitivity and specificity.

-Diagnostic tests are done to predict condition; the gold standard is needed to evaluate the performance of a test however the gold standard measure is not used routinely because it is usually costly, not feasible or time consuming; dichotomized results (diseased or not diseased).
- It is not hypothesis testing, it's about assessing how good the test is to identify sick and healthy patients.
- Relation between Se, Sp, NPV and PPV.

	GOLD STANDARD		TOTAL
	+	-	
TEST +	True Positive A	False positive B	PPV: TP/TP+FP
TEST -	False negative C	True negative D	NPV: TN/TN+FN
TOTAL	SE: TP/TP+FN	SP: TN/TN+FP	N

-Sensitivity: probability that a diseased individual will be identified as diseased by the test.
-Specificity: probability that a non-diseased individual will be identified as non-diseased/healthy by the test.
-Perfect situation: B + C = 0
-Predictive value: proportion of those tested who are correctly classified.
- In relation to the predictive values, it is important that the prevalence of the outcome in the population is similar to that of the study. It is more important for patients and care takers. Not really useful for epidemiology.
- Se and Sp: which is more important? Answer: depends on the purpose of the test. Cheap mass screening must be sensitive (identifies correctly the really diseased; not worried about the non-diseased) A test to confirm the presence of a disease must be specific (avoid false positives).
-Pretest probability: prevalence of the disease in the population. If a disease has low prevalence, the NPV will be high. Otherwise, high prevalence, high PPV.
-How to calculate? Take the percentage of the pre test and multiply by the Se and Sp of the test. It does not change the characteristics of a test. The Se and Sp are not changed. A good value would range in the middle (50%).
-Steps for diagnostic testing: determine whether there is a need for a Dx test; describe the selection pattern of subjects; reasonable gold standard; standardized gold standard and diagnostic test; estimate sample size for a 95% CI for Se and Sp; find sufficient number of subjects; report the results in terms of Se, Sp, PPV and NPV.
-Population screening: supposed to detect occult disease or a precursor state; immediate objective is to classify people as being likely or unlikely of having the disease and the ultimate objective is to reduce mortality and morbidity.
- Requirements for a screening method: suitable disease (serious consequences if untreated, detectable in pre clinical stage); suitable test (detects pre symptomatic phase, safe, accurate, acceptable and cost-effective); suitable program (reaches appropriate target population, efficient, good follow up of positives); good use or resources (cost of screening and follow up dx tests, costs of tx, benefits vs alternatives).
-How to avoid misclassification? Consistency check up, standardization, validated method.

-Reliability: get the same results each time (could be wrong or not but gets the same results). It does not ensure validity.
-Validity: gets the correct result each time (can't be wrong, always reliable because the result is repeatedly right). Associated with sensitivity and specificity. Lack of reliability constrains validaity.
-Accuracy: degree to which a variable actually measure what it is supposed to measure. Best way to assess is to compare with a gold standard reference. It can be threatened by systematic error.
-Precision: degree to which a variable has nearly the same value when measured several times. Best way to assess is using repeated measures. It can be threatened by random error.
-Intereliability: two different researches giving similar results.
-Intrareliability: same results given by the same reasearcher repeated times.

KAPPA COEFFICIENT: measures the agreement between 2 raters for the same evaluation (Kappa coefficient of COHEN). 1 meaning perfect agreement of the two raters and 0 agreement not better than what could be obtained by chance.
-It is used for tables in which both variables use the same values for the categories and both variables have the same number of categories.

Kappa = (Observed Ag – Ag by chance)/(1 – Ag by chance)

Observed agreement: A + D/N

Agreement by chance: (n1*m1) + (n2*m2)

		RATER A		TOTAL
		NO	YES	
RATER B	NO	A	B	n1
	YES	C	D	n2
TOTAL		m1	m2	N

*A+D = agreements; B+C = disagreements

- Kappa could be interpreted as: <0 no agreement; 0-0.2 poor; 0.21-0.4 mild; 0.41-0.6 moderate; 0.61-0.8 good; 0.8-1 very good to excellent. However, it can be harmful because the number of categories and subjects affects the magnitude of the value. (Higher K with less categories).
-In SPSS: analyze- descriptive statistics- crosstabs – click Kappa. SPSS will give you a descriptive table and another one with the kappa and p values. Interpret p value as usual.
- When you are dealing with several raters, use Fleiss Kappa.

Agreement between two methods: used when you want to assess a new measurement technique with an established one in order to see whether they agree sufficiently for the new to replace the old. This is often analyzed incorrectly with correlation coefficients. The correct way is with the Bland-Altman method.
-This is not about calibrating a new technique because often we do not know the gold standard parameters (too expensive, unethical, impossible, and too difficult). Meaning this is not about measuring a new method with previously known quantities.
- We are trying to compare two different methods on their degree of agreement, which may or may not be the gold standard.
- Many studies give the correlation coefficient (r) as an indicator of the agreement but this is wrong. Reasons:
a) r measures strength of relation between two variables, not agreement. We have perfect agreement only if the points lie along the line of equality but we will have perfect correlation if the points lie along any straight line.

b) a change in scale of measurement does not affect correlation but it affects the agreement. If we have one method that is half the other, the slope of the correlation would be 2.0 but the agreement would not agree, since one is half the other.
c) correlation depends on the range of the true quantity. (Wide quantity, wide correlation and vice versa).
d) the test of significance is irrelevant to the question of agreement
e) data with high correlation may have poor agreement.
-It is very unlikely that different methods will agree exactly for all individuals. Therefore we want to know by how much the new method is likely to differ from the old. We want to know by how much the new method is likely to differ from the old. How far apart measurements can be without causing difficulties? Ideally, it should be defined in advance to help the interpretation of the method comparison and to choose the sample size.
- First examine the data. A simple plot of the results of one method against the other without a regression line could be useful but will result in data points clustered bear the line and it will be difficult to assess between-method differences. And that is why we plot the difference between the methods against their means. Assuming that we do not know the true value, the mean of the two methods is the best estimate we have.
-If the differences are normally distributed, 95% of them will lie between (+2SD and -2SD or more precisely: d – 1,96*SD; d +1,96*SD).
The measurements themselves may not be normally distributed (and often will not) but we can check the differences by drawing a histogram. If they are skewed or have very long tails, the assumption of normality may not be valid.
- If this range of differences is not clinically important, we can use both methods interchangeably. We can refer to it as “Limits of agreement”.
- Precision of estimated limits of agreement: we might sometimes which to use standard errors and confidence intervals to see how precise our estimates are, provided that the differences follow a normal distribution. The standard error of the d is s/√n, where n is sample size and the standard error of d-2ds and d+2sds is 3s/√n. If the intervals are wide, may be due to small sample size and the great variation of differences. There can be considerable discrepancies between the two methods and the degree of agreement is not acceptable.
- For some analysis you can log transform your difference to assess the relation between two methods.
-Steps of the Blank-Altman analyzes: plot the scatter plot of the variables compared with an equality line (all points would lie if the two parameters assessed gave exactly the same reading every time); assess the correlation coefficient (but remember: having correlation does not mean that the 2 methods agree); compute new variable Difference between the method and mean; plot scatter plot with the Difference in Y axis and Mean in the X axis; determine the mean of the difference and the standard deviation; add these two parameters to the scatter plot and check the range of the limits of agreement.
- Repeatability: the repeatabilities of the two methods can limit the amount of agreement. If one method has poor repeatability (high variation in repeated measurements on the same subject), the agreement is bound to be poor. When the old method is the most variable one, even a perfect new method will not agree with it. If both have poor repeatability, the problem is even worse.
-For this matter we can take repeated measurements on a series of subjects. Plot the scatter plot of the differences against the mean for each repetition. We then calculate the mean and SD of the difference, just as before. The mean difference should be 0 and expect that 95% of the different be less than 2SDs away.

- Agreement can also be computed when distributions are not normal.

-ROC Curve: Receiver Operating Characteristics. It is a graphic representation of the Se (True Positives) and Specificity (1-Sp: false positives) for continuous variables. It is a tool to select an optimal model. It relates directly with the cost-benefit analysis in the diagnostic decision process. The Se and Sp determine the performance of the diagnostic test to classify positively or negatively across all the possible samples in the study.
-The area under the curve discriminates how good the test is. As more "skewed" to the left, better the result, better instrument being evaluated. In the y axis goes the sensitive with the true positives and on the x axis does 1-specificity (false positives).
-If the line where dividing the graph in two equal halves, there would be no discrimination. The test is no better than chance.
-IN SPSS: analyze – ROC curve – choose the test variable (instrument being evaluated) and state variable (gold standard) – make sure it's clicked on Display all the 4 boxes. SPSS will give you the graph and a table with CI, standard error and p value.

-Confounding: is a variable other than the independent that may affect the dependent variable leading to erroneous conclusions about the relationship of the variables under study. You deal with confounding variables by controlling, matching, randomizing or statistically controlling them.
- In this multicausal complex we live in, there's a variety in genetics, developmental and environment factors. This means that when we design an experiment, the samples will differ not only in relation to the independent variable but also to others variables you may or may not be aware. It may trick the results creating associations where there really isn't or cause so much variation that it's hard to detect the real relationship between the studied variables (under or overestimation). However some variables are not confounders and they only need to be adjusted.
-A condition for a factor to be called confounder is to be correlated to the exposure and associated to the outcome. E.g.: smoking as a causation for CHD. It is known that people who smoke tend to drink more alcohol and vice versa (two handed) and alcohol is one of the risk factors for CHD but CHD does not "cause" alcohol (one hand). Therefore alcohol is a real confounder.
-Case where the factor assessed is not a real confounder: diet as a causation to CHD. Cholesterol levels are related to the diet and is in on the causal pathway to CHD. Therefore it is not a confounder, it is an intermediary of the relation between diet and CHD. Another example: once again smoking and CHD. Yellow fingers are associated with the exposure (smoking) however it has nothing to do with CHD so there would be no need to adjust people with yellow fingers in this study.
-Controlling for confounders must ideally be done at the design stage of the study or during the data analysis.
-They must be equally distributed between the exposed and unexposed groups in order to have the effects neutralized.

-Multiple regression: use multiple regression when you have three or more measurement variables. One of them is the dependent variable (y). The purpose of a multiple regression is to find an equation that best predicts the Y variable as a linear function of the x variables.
- MR for prediction: estimation of an unknown Y value corresponding to a set of X values.

- MR for causation: understand the functional relationships between the dependent and independent variables.
-The main null hypothesis is that there is no relationship between the X variables and the Y variables. In other words, the Y values you predict from your multiple regression equation are no closer to the actual Y values than you would expect by chance. The alternative hypothesis is that at least one of the independent variables is associated to the dependent variables ($\beta x \neq 0$).
- Formula: $Y\hat{} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_x X_x$; where $Y\hat{}$ is the expected value of Y for a given set of X values, β_1 is the estimated slope of a regression of Y on X1 if all of the other X variables could be kept constant and so on, α is the intercept (value of Y when x=0). α could be called βo (mean response when x is 0).
-How to read the slope: "for every unit ___ of increase in X, the change in Y would be ___"
-How well the equation fits the data is expressed by r^2 "coefficient of multiple determination". Ranges from 0 (no relationship between Y and X variables) to 1 (for a perfect fit, no difference between the observed and expected Y values). The P value will be a function of the r^2, number of observations and the number of X variables.
- Using nominal variables in a multiple regression: if the independent variable is categorical (ordinal, binary, dichotomous), you can create a dummy variable. You have to code your variable based on a reference and then code the others with 0 and 1. E.g: SES (low, medium and high). Low would be the reference, medium 1 when the rest are zero or high 1 when the rest are zero. It's mandatory to recode the variables in order to make it possible to interpret. The basic idea is that for k values, you create k-1 dummy variables (this -1 is the reference group).
-Adding variables to a linear regression model will increase the unadjusted r^2 value. One way to choose the variables, called forward selection, is to do a linear regression for each of the x variables, one at a time, then pick the X variable that had the highest R^2. Next you take the variable chosen and another one and run a multiple regression. You add the X variable that increases the r^2 by the greatest amount until adding another x variable does not increase it significantly. You can also set up a desired cut off value depending on the amount of variables you want in your regression.
-Another way to do it is called backwards elimination. You start with a multiple regression using all of the X variables, then perform multiple regression with each X variable removed in turns. You eliminate the variable whose removal causes the smallest decrease in R^2. You continue removing X variable until there's a significant decrease in r^2.
-Regardless of the method chosen, it is better to have a small number of independent variables (use the ones who are really significant), the interpretation of results becomes difficult if there's too many variables. The best model is the one who explains as most as possible with the smallest number of independent variable (Model parsimony).
-Assumptions of multiple regression: variables normally distributed and homoscedastic (constant variance for all levels of X). Note that regression models (linear and multiple are robust procedures, not that sensitive to the violations of these assumptions). The expected residuals of the regression variables are supposed to be also normally distributed. Check for linearity of the dependent in relation to the independent variables. Look for the correlation (just because an individual correlation looks linear, it doesn't necessarily means it is). Check for multicollinearity (two independent variables that are highly correlated with each other. If two are highly correlated there may be inconsistency of the results and you should make an option of keeping one independent variable and discarding the other).
-Collinearity: SPSS has an output that identify covariates with high degree of collinearity, the variance inflaction factor (VIF). The VIF should not exceed 10, if it does there is a sign of collinearity.

-If there were only one variable under study, the p value would be the same as the test statistics p value but as that is not the case for MR the p value is different that the test statistics.
-IN SPSS: analyze- regression – linear – choose the dependent and independent variables – you can ask for CI and collinearity diagnostics in the statistics option. SPSS will give a table with the values of the R, r^2 (coefficient of determination – proportion of variance in the dependent variable that can be explained by the independent variables) and adjusted R squared; an anova table with the F value; coefficient table – the slope (β) values are provided on the unstandardized coefficients and the α value is the constant. The unstandardized values indicate how much the dependent variable (Y) varies with an independent variable when all the other independent variables are held constant. If p value is lower than 0, assume significance.
- A significant CI means that at least one of the correlation coefficients are valid. To be valid it should not be 0, contain 0 or be equal to the others (null hypothesis). If we already know that the range of variables does not contain 0, we do not need to know the p value. If there is an overlap of the CI, the p value will be higher than 0.05 (the subtraction of the mean differences of the variables will be 0).