Understanding Clinical Research: Behind the Statistics

Keynotes

Last updated: 13 May 2016

Clinical Research Notes

Download the notes by clicking File -> Download as in the toolbar.

File edit access is not needed.

Table of Contents:

Week 1 : Getting things started by defining different study types

Getting to know study types

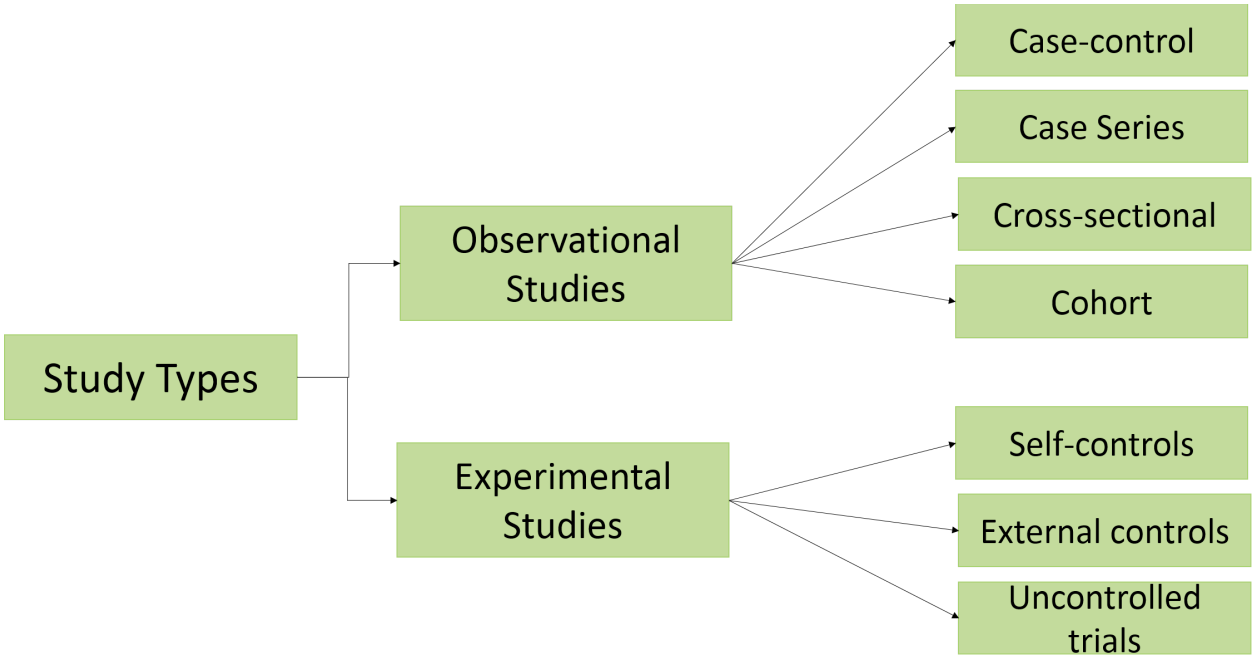
Do you know your cross-sectional from your cohort study? Your retrospective case-control series from your dependent control trials? Study types can be very confusing. Yet it is essential to know what type of study you are reading or planning to conduct. Defining a study into a specific type tells us a lot about what we can learn from the outcomes, what strengths and weaknesses underpin its design and what statistical analysis we can expect from the data.

There are various classification systems and it is even possible to combine aspects of different study types to create new research designs. We'll start this course off with an intuitive classification system that views studies as either observational or experimental (interventional).

Observational and experimental studies

In this first lecture I covered clinical study types, using the classification system that divides all studies into either observational or experimental. I also took a look at meta-analyses and systematic reviews.

Let's summarise the key characteristics of the different study types that I mentioned. The diagram below gives an overview.



In observational studies:

- subjects and variables pertaining to them are observed and described
- no treatment or intervention takes place other than the continuation of normal work-flow, i.e. healthcare workers are allowed to carry on with their normal patient management or treatment plans
- there are four main observational study types: case series, case-control studies, cross-sectional studies and cohort studies

In experimental studies:

- subjects are subjected to treatments or interventions based on a predesignated plan
- healthcare workers are not allowed to continue their routine care, but must alter their actions based on the design of the study
- this usually results in at least two groups of patients or subjects that follow a different plan and these can then be compared to each other
- the main idea behind an experimental study is to remove bias
- if a study involves humans, this is known as a clinical trial
- the main experimental study types are: trials with independent concurrent controls, trials with self-controls, trials with external controls, and uncontrolled trials

I also introduced the topic of ***meta-analyse****s* and ***systematic reviews****. Meta-analysis* uses pre-existing research and combines their results to obtain an overall conclusion. They aim to overcome one of the most common problems that beset clinical research and that is small sample sizes, resulting in underpowered results. A *systematic review* is a literature review that sums up the best available information on a specific research question and includes results from research into the specific field, published guidelines, expert (group) opinion and meta-analyses.

Now that you know how to distinguish between the various clinical studies, I'll cover each of the study types in the upcoming lecture more in-depth.

Getting to Know Study Types: Case Series

A case series is perhaps the simplest of all study types and reports a simple descriptive account of a characteristic observed in a group of subjects. It is also known by the terms clinical series or clinical audit.

A case series:

- observes and describes subjects
- can take place over a defined period or at an instant in time
- is purely analytical and requires no research hypotheses
- is commonly used to identify interesting observations for future research or planning

Also simple by nature and design, cases-series are nevertheless important first steps in many research areas. They identify numbers involved, i.e. how many patients are seen, diagnosed, under-threat, etc. and describe various characteristics regarding these subjects.

We are by nature biased and as humans have a tendency to remember only selected cases or events. We are usually poor at seeing patterns in large numbers or over extended periods and by examining case-series (audits), we find interesting and sometimes surprising results and these can lead to further research and even a change in management.

Paper referenced in video:

1. *Donald, K. a, Walker, K. G., Kilborn, T., Carrara, H., Langerak, N. G., Eley, B., & Wilmshurst, J. M. (2015). HIV Encephalopathy: pediatric case series description and insights from the clinic coalface. AIDS Research and Therapy, 12, 1–10. doi:10.1186/s12981-014-0042-7*

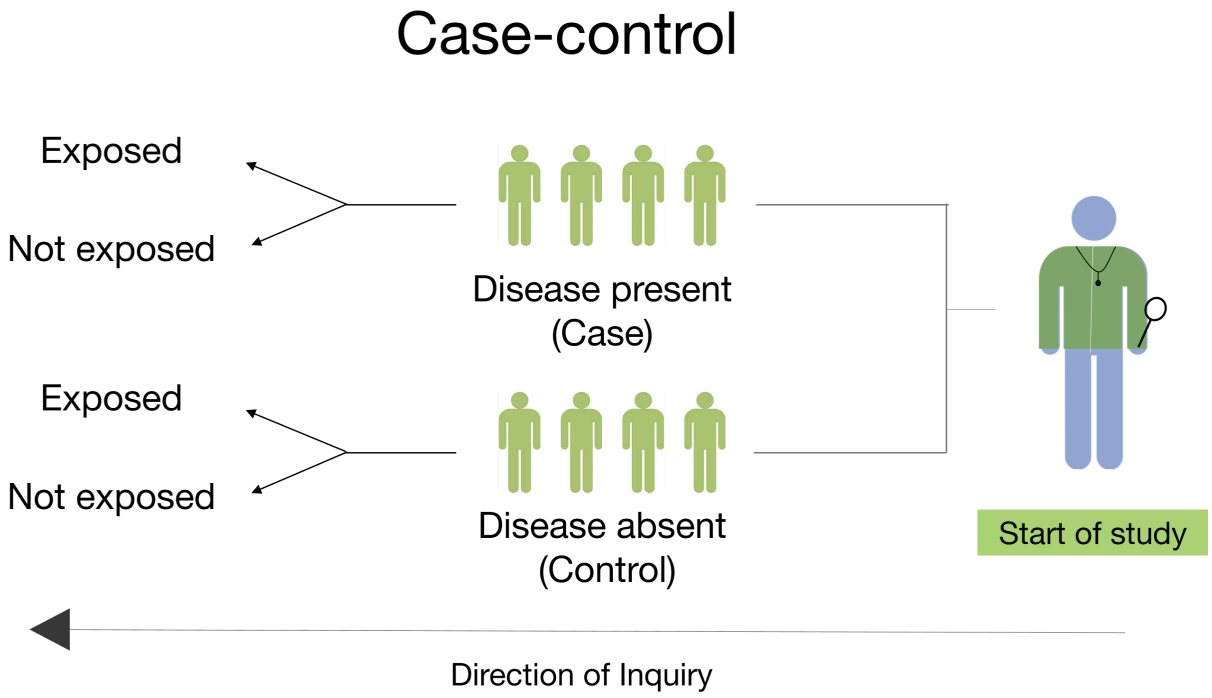
Case-control Studies

Now that I have covered the topic of case-control studies in the video lecture, let's summarise and expand on what we've learned.

A case-control study:

- selects subjects on the basis of a presence (cases) and absence (controls) of an outcome or disease
- looks back in time to find variables and risk factors that differ between groups
- can attempt to determine the relationship between the exposure to risk factors (or any measured variable) and the disease
- case-control studies can include more than two groups

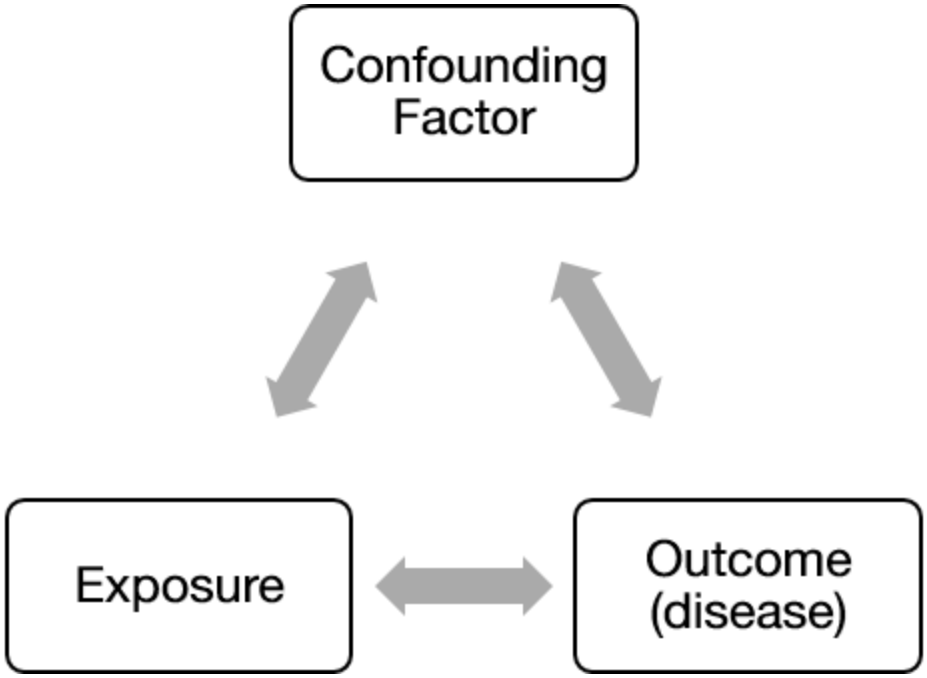
To illustrate these points, consider an example where patients undergo some form of invasive surgical intervention. You might note that after the same surgical procedure, some develop infection at the wound site and some do not. Those with the wound infection (termed surgical site infection) make up the cases and those without, the controls. We can now gather data on various variables such as gender, age, admission temperature, etc. and compare these between the two groups. Note how such data on these variables all existed prior to the occurrence of the wound infection. The study looks back in time and the data is collected retrospectively.



What is a drawback of case-control studies?

The main drawback is ***confounding***, which refers to a false association between the exposure and outcome. This occurs when there is a third variable, we call this the *confounding factor* which is associated with both the risk factor and the disease.

Let's consider another example. Several studies find an association between alcohol intake (exposure) and heart disease (outcome). Here, a group of patients with heart disease will form the cases and a group without heart disease, the controls and we look back in time at their alcohol consumption. If, by statistical analysis, we find a higher alcohol consumption in the heart disease group than in the control group, we may think that drinking alcohol causes heart disease. But another confounding factor, i.e. smoking, may be related to both alcohol intake and heart disease. If the study does not consider this confounding factor, this relationship between the exposure and outcome may be misinterpreted. The confounding factor, in this case smoking, needs to be controlled in order to find the true association.



You can review the example of the case-control study in the paper I discussed in the lecture

References:

1. Yung J, Yuen JWM, Ou Y, Loke AY. Factors Associated with Atopy in Toddlers: A Case-Control Study. Tchounwou PB, ed. International Journal of Environmental Research and Public Health. 2015;12(3):2501-2520. doi:10.3390/ijerph120302501.

Cross-sectional studies

Let's review the characteristics of cross-sectional studies.

A cross-sectional study:

- identifies a population or sub-population rather than individuals
- takes place at a point in time or over a (relatively) short period
- can measure a range of variables across groups at the same time
- is often conducted in the form of a survey
- can be a quick, easy and a cost effective way of collecting information
- can be included in other study designs such as case-control and cohort studies

- is commonly used to measure prevalence of an outcome or disease, i.e. epidemiological studies

What are the potential drawbacks of cross-sectional studies?

Bias

Response bias is when an individual is more likely to respond if they possess a particular characteristic or set of characteristics. For example, HIV negative individuals may be more comfortable responding to a survey discussing their status compared to HIV positive individuals. A variety of technical difficulties or even age may also influence responders. Once bias exists in the group of responders, it can lead to seeing of the data and inappropriate conclusions can be drawn from the results. This can have devastating consequences as these studies are sometimes used to plan large scale interventions.

Separating Cause and Effect

Cross-sectional studies may not provide accurate information on *cause and effect*. This is because the study takes place at a moment in time, and does not consider the sequence of events. Exposure and outcome are assessed at the same time. In most cases we are unable to determine whether the disease outcome followed the exposure, or the exposure resulted from the outcome. Therefore it is almost impossible to infer causality.

You may find it useful to review the papers I discussed in the video, which are good examples of cross-sectional studies.

References:

1. Lawrenson JG, Evans JR. Advice about diet and smoking for people with or at risk of age-related macular degeneration: a cross-sectional survey of eye care professionals in the UK. BMC Public Health. 2013;13:564. doi:10.1186/1471-2458-13-564.
2. 2. Sartorius B, Veerman LJ, Manyema M, Chola L, Hofman K (2015) Determinants of Obesity and Associated Population Attributability, South Africa: Empirical Evidence from a National Panel Survey, 2008-2012. PLoS ONE 10(6): e0130218. doi:10.1371/journal.pone.0130218

Cohort studies

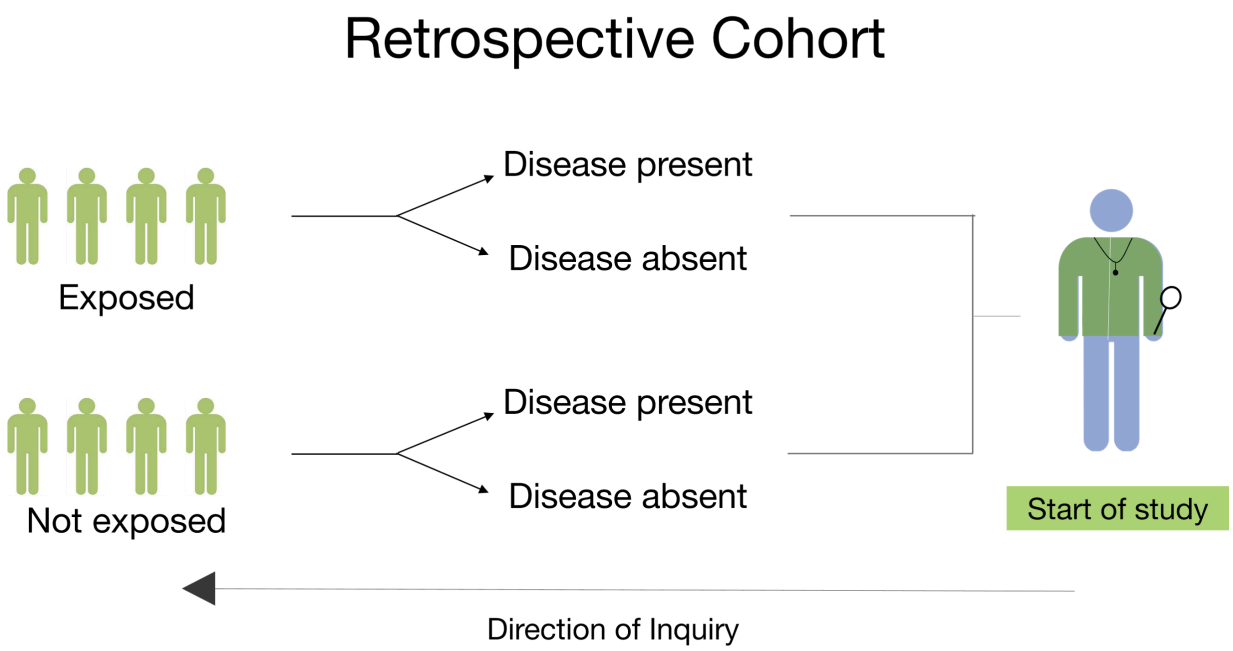
As I outlined in the lecture, a cohort study:

- begins by identifying subjects (the cohort) with a common trait such as a disease or risk factor
- observes a cohort over time

- can be conducted retrospectively or prospectively

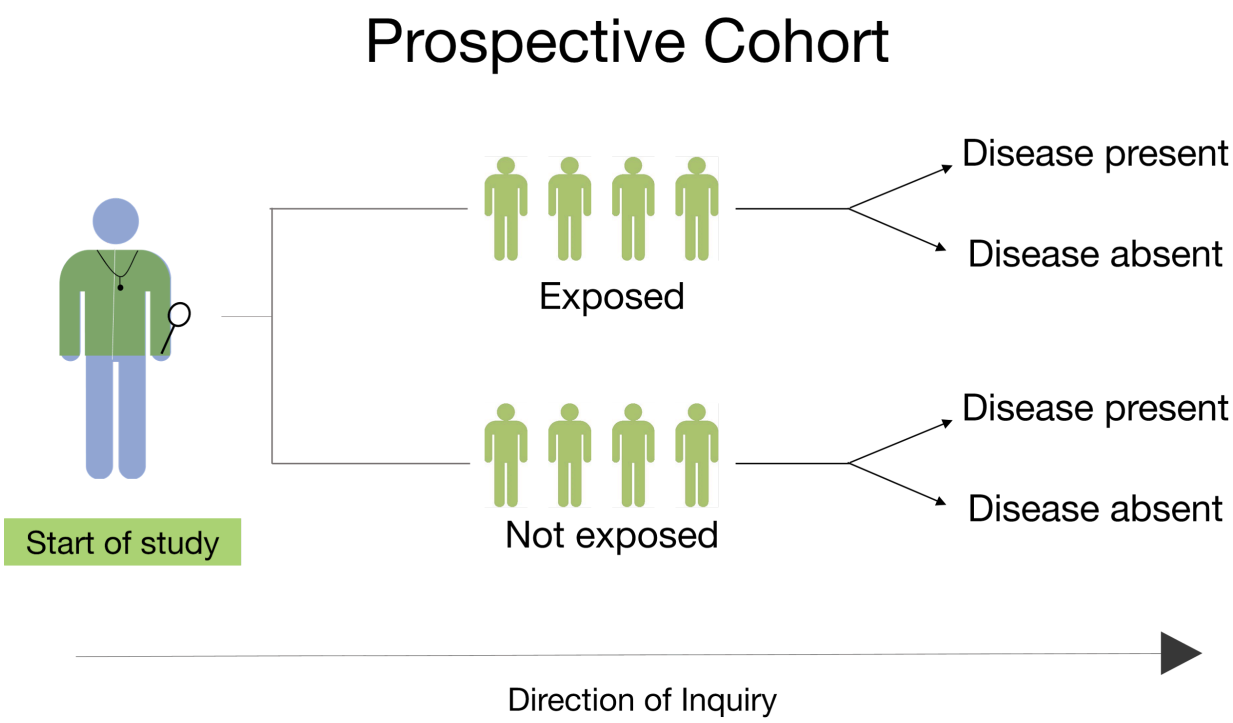
Retrospective Cohort Studies

A *retrospective* study uses existing data to identify a population and exposure status. Since we are looking back in time both the exposure and outcome have already occurred before the start of the investigation. It may be difficult to go back in time and find the required data on exposure, as any data collected was not designed to be used as part of a study. However, in cases where reliable records are on-hand, retrospective cohort studies can be useful.



Prospective Cohort Studies

In a *prospective cohort* study, the researcher identifies subjects comprising a cohort and their exposure status at the beginning of the study. They are followed over time to see whether the outcome (disease) develops or not. This usually allows for better data collection, as the actual data collection tools are in place, with required data clearly defined.



The term *cohort* is often confused. It simply refers to a group of subjects and for the purposes of research, they usually have some common trait. We often use this term when referring to this type of study, but you will also note it in the other forms of observational studies. When used there, it is simply a generic term. When used in the case of cohort studies it refers to the fact that the data gathered for the research points to events that occurred after the groups (cohorts) were identified.

To get back to our earlier example of wound infection patients (that we used in the case-control section), the patient with and without wound infection could be considered cohorts and we consider what happened to them after the diagnosis of their wound infection. We might then consider length of hospital stay, total cost, or the occurrence of any events after the development of the wound infection (or at least after the surgery for those without wound infection). The defining fact, though, is that we are looking forward in time from the wound infection in contrast to case-control series, where we look back at events before the development of the wound infection.

You may find it useful to review the paper I discussed in the video, which is a good example of a cohort study.

Paper discussed in the video:

Le Roux, D. M., Myer, L., Nicol, M. P., & Zar, H. J. (2015). Incidence and severity of childhood pneumonia in the first year of life in a South African birth cohort: the Drakenstein Child Health a. The Lancet Global Health, 3(2), e95–e103. doi:10.1016/S2214-109X(14)70360-2

Experimental studies

In *experimental* studies (as opposed to observational studies, which we discussed earlier), an active intervention takes place. These interventions can take many forms such as medication, surgery, psychological support and many others.

Experimental studies:

- aim to reduce bias inherent in observational studies
- usually involve two groups or more, of which at least one is the control group
- have a control group that receives no intervention or a sham intervention (placebo)

Randomization

To reduce bias, true *randomization* is required. That means that every member of a population must have an equal opportunity (random chance) to be included in a sample group. That necessitates the availability of a full list of the population and some method of randomly selecting from that list.

In practical terms this means that every subject that forms part of the trial, must have an equal opportunity of ending up in any of the groups. Usually it also means that all of these subjects are also taken from a non-selected group, i.e. in a non-biased way. For example, if we want to investigate the effectiveness of a new drug on hypertension, we must not only be certain that all patients have an equal opportunity to receive either the drug or a placebo, but that all the participants are randomly selected as a whole. If all the participants come from a selected group, say, from a clinic for the aged, there is bias in the selection process. In this case, the researchers must report that their results are only applicable to this set of the population.

Blinding

If the participants do not know whether they are in the control group or not, this is termed *blinding*. When the researchers are similarly unaware of the grouping, it is termed *double-blinding*. Therefore, in a double-blinded trial, both participants and researchers are unaware of grouping until after data collection. This method is preferable but not always possible, i.e. in a surgical procedure. In these cases the observer taking measurements after the intervention may be blinded to the intervention and the surgeon is excluded from the data collection or measurements.

The pinnacle of clinical research is usually seen to be the *randomized, double-blind controlled trial*. It often provides the strongest evidence to prove causation.

The control group can be set up in a variety of ways:

Trials with independent concurrent controls

In trials with *independent concurrent controls*, the controls are included in the trial at the same time as the study participants, and data points are collected at the same time. In practical terms this means that a participant cannot be in both groups, nor are homozygotic twins allowed.

Trials with self-controls

In trials with *self-controls*, subjects are treated as the control and treatment groups. Data is collected on subjects before and after the intervention.

The most elegant subtype of this form of trial is the *cross-over study*. Two groups are formed each with their own intervention. Most commonly one group will receive a placebo. They form their own controls, therefore data is collected on both groups before and after the intervention**.** After the intervention and data collection, a period of no intervention takes place. Before intervention resumes, individuals in the placebo group are swapped with individuals in the treatment group. The placebo group then becomes the treatment group and the treatment group becomes the placebo group.

Trials with external controls

Trials with *external controls* compares a current intervention group to a group outside of the research sample. The most common external control is a *historical control*, which compares the intervention group with a group tested at an earlier time. For example, a published paper can serve as a historical control.

Uncontrolled trials

In these studies an intervention takes place, but there are *no controls*. All patients receive the same intervention and the outcomes are observed. The hypothesis is that there will be varying outcomes and reasons for these can be elucidated from the data. No attempt is made to evaluate the intervention itself, as it is not being compared to either a placebo or an alternative form of intervention.

You may find it useful to review the paper I discussed in the video, which is a good example of an experimental study.

Paper mentioned in video:

1. Thurtell MJ, Joshi AC, Leone AC, et al. Cross-Over Trial of Gabapentin and Memantine as Treatment for Acquired Nystagmus. Annals of neurology. 2010;67(5):676-680. doi:10.1002/ana.21991.

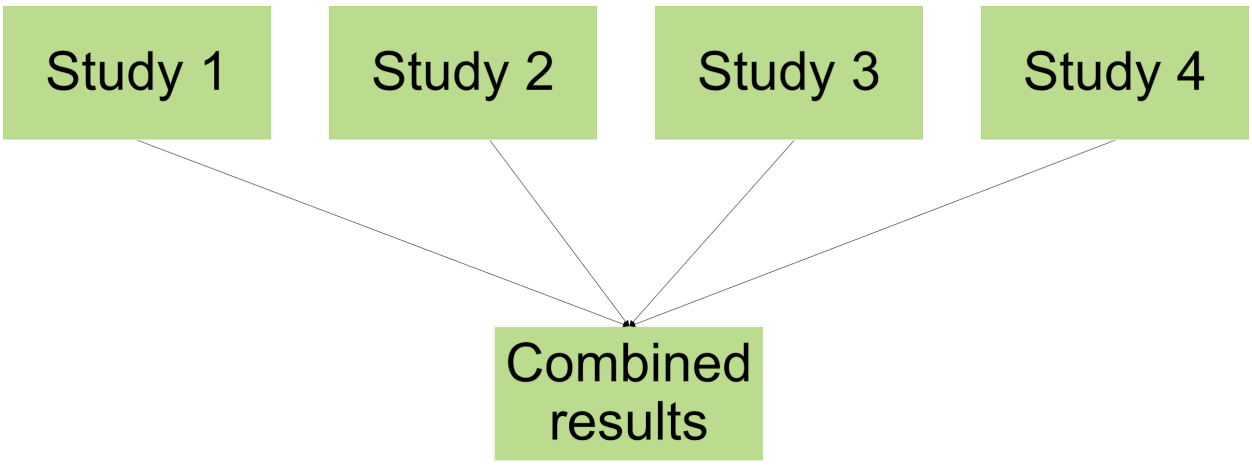
Meta-analysis and systematic review

We're almost at the end of week 1! I've covered observational and experimental studies in previous videos, and I'm concluding this lesson with a discussion on *meta-analyses* and *systematic reviews*.

Meta-analysis

A meta-analysis:

- uses pre-existing research studies and combines their statistical results to draw an overall conclusion
- centers around a common measurement such as finding an average or mean
- is useful for combining individual studies of inadequate size or power to strengthen results
- uses inclusion and exclusion criteria to select papers to be analysed



What is a drawback of meta-analysis?

One possible drawback, which makes meta-analyses less useful, is the ever-present danger of *publication bias*. Publication bias is well recognized and refers to the fact that through various positive and negative incentives, it is much more likely to find positive results in the published literature, i.e. statistically significant results. It is much less common to find negative results. Even

though meta-analysis is used to increase the power of a result and make it more generalizable, its results may still be poor if the studies on which it is based are biased towards positive, statistically significant results.

Systematic Review

How often do you come across research studies that contradict one another's findings? One study reports that carbohydrates are bad for you, another study says carbohydrates are required as part of a balanced diet. When looking for research evidence, we need to look beyond a single study. This is where *systematic reviews* fit in.

A systematic review:

- summarises a comprehensive amount of published research
- helps you find a definite word on a research question
- can include a meta-analysis
- can use frameworks such as the PRISMA to structure the review

What is the difference between a systematic review and a meta-analysis?

There is some overlap and not everyone stick clearly to the strict definitions of these two types of research (although, as I mentioned in the lesson, clear guidelines for both have been accepted by most researchers and publishers). The aim of both is to collect and use previously published data. Most systematic reviews include a meta-analysis, but they are rather like to circles in a Venn-diagram, with some overlap (intersection).

A meta-analysis is indeed the combination of previously published research. This combination can then be used for re-analysis. Combining these results into bigger numbers may result in improved results. It is often seen as a quantitative look at previously published data. There may or may not be some narrative to the meta-analysis giving a little background information and knowledge about the subject.

A systematic review also collected previously published work, but takes a more qualitative look and is usually much more involved than a meta-analysis. It really aims to be the most definitive word on a topic or research question. Certain procedures are follows and are clearly set out in the design of the study, so as to minimise bias in the selection and analysis of the data. Objective techniques are then used to analyse the data. The focus is on the magnitude of the effect , rather than on statistical significance (meta-analysis). It adds a lot of detail and explanation of the topic at hand and includes a lot of narrative.

There is then also the narrative review. These are found in publications such as the various North American Clinics Journals. There is reference to previous published data, but the trend is more towards the style of a textbook.

Next up, we've developed a practice quiz for you to check your understanding. Good luck!

Papers mentioned in the video:

1. Geng L, Sun C, Bai J. Single Incision versus Conventional Laparoscopic Cholecystectomy Outcomes: A Meta-Analysis of Randomized Controlled Trials. Hills RK, ed. PLoS ONE. 2013;8(10):e76530. doi:10.1371/journal.pone.0076530.

Week 2: Describing your data

The spectrum of data types

Definitions

There are eight key definitions which I introduced in the video lecture that I will be using throughout the rest of this course.

Descriptive statistics

- The use of statistical tools to summarize and describe a set of data values
- Human beings usually find it difficult to create meaning from long lists of numbers or words
- Summarizing the numbers or counting the occurrences of words and expressing that summary with single values makes much more sense to us
- In descriptive statistics, no attempt is made to compare any data sets or groups

Inferential statistics

- The investigation of specified elements which allow us to make inferences about a larger population (i.e., beyond the sample size)
- Here we compare groups of subjects or individuals
- It is normally not possible to include each subject or individual in a population in a study, therefore we use statistics and infer that the results we get, apply to the larger population

Population

- A group of individuals that share at least one characteristic in common
- On a macro level, this might refer to all of humanity
- At the level of a clinical research, this might refer to every individual with a certain disease, or risk factor, which might still be an enormous number of individuals
- It is quite possible to have quite small population, i.e. in the case of very rare condition
- The findings of a study infer its results to a larger population; we make use of the findings to manage the population to which those study findings infer

Sample

- A sample is a selection of members within the population (I'll discuss different ways of selecting a sample a bit later in this course)
- Research is conducted using that sample set of members and any results can be inferred to the population from which the sample was taken
- This use of statistical analysis makes clinical research possible as it is usually near impossible to include the complete population

Parameter

- A statistical value that is calculated from all the values in a whole population, is termed a parameter
- If we knew the age of every individual on earth and calculated the mean or average age, that age would be a parameter

Statistic

- A statistical value that is calculated from all the values in a sample, is termed a statistic
- The mean or average age of all the participants in a study would be a statistic

Variable

- There are many ways to define a variable, but for use in this course I will refer to a variable as a group name for any data values that are collected for a study
- Examples would include age, presence of risk factor, admission temperature, infective organism, systolic blood pressure
- This invariably becomes the column names in a data spreadsheet, with each row representing the findings for an individual in a study

Data point

- I refer to a data point as a single example value for a variable, i.e. a patient might have a systolic blood pressure (the variable) or 120 mm Hg (the data point)

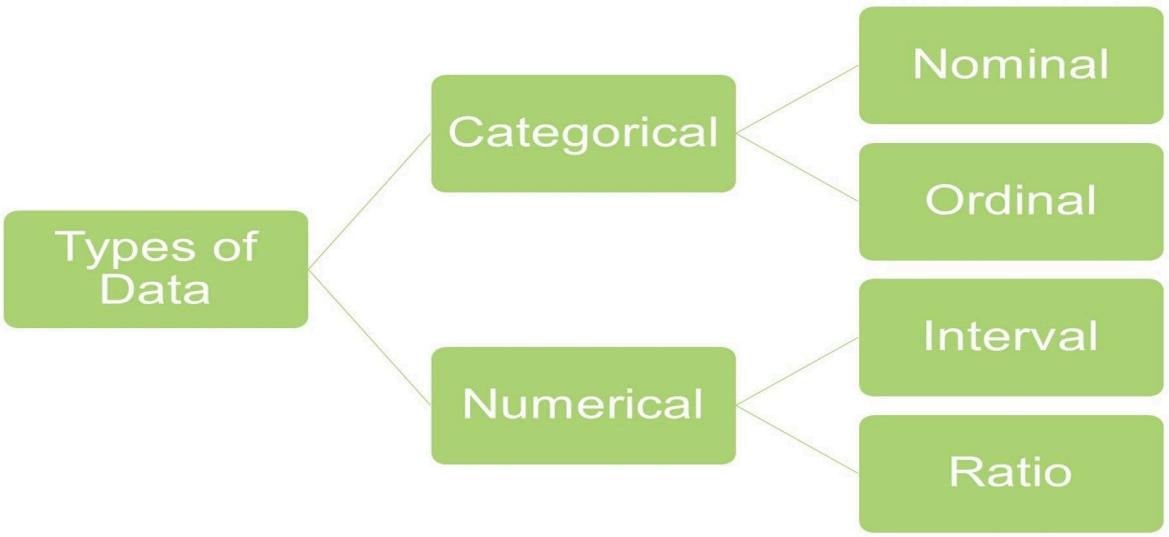
Let's use this knowledge in a quick example: Say we want to test the effectiveness of levothyroxine as treatment for hypothyrodism in South African subjects between the ages of 18-24. It will be physically impossible to find every individual in the country with hypothyrodism and collect their data. However, we can collect a representative sample of the population. For example, we can calculate the average (sample statistic) of the thyroid stimulating hormone level before and treatment. We can use this average to infer results about the population parameter.

In this example thyroid stimulating hormone level would be a variable and the actual numerical value for each patient would represent individual data points.

Data types

I will be using two classification systems for data - categorical and numerical, each of which has two sub-divisions.

- Categorical (including nominal and ordinal data), refers to categories or things, not mathematical values
- Numerical (further defined as being either interval or ratio data) refers to data which is about measurement and counting



In short categorical data types refer to words. Although words can be counted, the words themselves only represent categories. This can be said for diseases. Acute cholecystitis (infection of the gallbladder) and acute cholangitis (infection of the bile ducts) are both diseases of the biliary (bile) tract. As words, these diseases represent categorical entities. Although I can count how many patients have one of these conditions, the diseases themselves are not numerical entities. The same would go for gender, medications, and many other examples.

Just to make things a bit more difficult, actual numbers are sometimes categorical and not numerical. A good, illustrative example would be choosing from a rating system for indicating the

severity of pain. I could ask a patient to rate the severity of the pain they experience after surgery on a scale from 0 (zero) to 10 (ten). These are numbers, but they do NOT represent numerical values. I can never say that a patient who chooses 6 (six) has twice as much pain as someone who chooses 3 (three). There is no fixed difference between each of these numbers. They are not quantifiable. As such, they represent categorical values.

As the name implies, numerical data refers to actual numbers. We distinguish numerical number values from categorical number values in that there is a fixed difference between them. The difference between 3 (three) and 4 (four) is the exact same difference as that between 101 (one-hundred-and-one) and 102 (one-hundred-and-two).

I will also cover another numerical classification type: discrete and continuous variables. Discrete values as the name implies exist as little islands which are not connected (no land between them). Think of the roll of a die. With a normal six-sided die you cannot roll a three-and-a-half. Continuous numerical values on the other hand have (for practical purposes) many values in-between other values. They are infinitely divisible (within reasonable limits).

Nominal categorical data

Nominal categorical data:

- are data points that either represents words (‘yes’ or ‘no’) or concepts (like gender or heart disease) which have no mathematical value
- have no natural order to the values or words - i.e. ‘nominal’ - for example: gender, or your profession

Be careful of categorical concepts that may be perceived as having some order. Usually these are open to interpretation. Someone might suggest that heart disease is worse than kidney disease or *vice versa*. This, though, depends on so many points of view. Don't make things too complicated. In general, it is easy to spot the nominal categorical data type.

Readings referred to in this video as examples of nominal categorical data:

1. De Moraes, A. G., Racedo Africano, C. J., Hoskote, S. S., Reddy, D. R. S., Tedja, R., Thakur, L., … Smischney, N. J. (2015). Ketamine and Propofol Combination (“Ketofol”) for Endotracheal Intubations in Critically Ill Patients: A Case Series. The American Journal of Case Reports, 16, 81–86. doi:10.12659/AJCR.892424 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4332295/

Ordinal categorical data

If categorical data have some natural order or a logical ranking to the data points, it is termed ordinal categorical data, i.e. they can be placed in some increasing or decreasing order.

I gave the example of a pain score from 1 (one) to 10 (ten). Even though these are numbers, no mathematical operation can be performed on these digits. They are ordered in magnitude from 1 to 10. But there is no standardized measurement of these rankings and therefore no indication that the interval between the specific scores is of the same value.

Other common examples include survey questions: where a participant can rate their agreement with a statement on a scale, say 1 (one), indicating that they don't agree at all, to 5 (five), indicating that they fully agree. Likert style answers such as totally disagree, disagree, neither agree nor disagree, agree and totally agree can also be converted to numbers, i.e. 1 (one) to 5 (five). Although they can be ranked, they still have no inherent numerical value and as such remain ordinal categorical data values.

References:

1. Lawrenson JG, Evans JR. Advice about diet and smoking for people with or at risk of age-related macular degeneration: a cross-sectional survey of eye care professionals in the UK. BMC Public Health. 2013;13:564. doi:10.1186/1471-2458-13-564.

Numerical data types

As opposed to categorical data types (words, things, concepts, rating numbers), numerical data types involve actual numbers. Numerical data is quantitative data - for example, the weights of the babies attending a clinic, the doses of medicine, or the blood pressure of different patients. They can be compared and you can do calculations on the values. From a mathematical point of view, there are fixed differences between values. The difference between a systolic blood pressure value of 110 and 120 mm Hg is the same as between 150 and 160 mm Hg (being 10 mm Hg).

There are two types of numerical data - interval and ratio.

Interval

With interval data, the difference between each value is the same, which means the definition as 'I' used above holds. The difference between 1 and 2 degrees Celsius is the same as the difference between 3 and 4 degrees Celsius (there is a 1 degree difference). However, temperatures expressed in degrees Celsius (or Fahrenheit) do not have a ‘true zero’ because 0 (zero) degrees Celsius is not a

true zero. This means that with numerical interval data (like temperature) we can order the data and we can add and subtract, but we cannot divide and multiply the data (we can't do ratios without a 'true zero'). 10 degrees plus 10 degrees is 20 degrees, but 20 degrees is not twice as hot as 10 degrees Celsius. Ratio type numerical data requires a true zero.

Ratio

This type applies to data that have a true 0 (zero), which means you can establish a meaningful relationship between the data points as related to the 0 (zero) value eg. age from birth (0) or white blood cell count or number of clinic visits (from 0). A systolic blood pressure of 200 mm Hg is indeed twice as high as a pressure of 100 mm Hg.

Summary

Nominal categorical = naming and describing (eg. gender)

Ordinal categorical = some ordering or natural ranking (eg. pain scales)

Interval numerical = meaningful increments of difference (eg. temperature)

Ratio numerical = can establish a base-line relationship between the data with the absolute 0 (eg. age)

Why do we need to spend time distinguishing data?

You have to use very different statistical tests for different types of data, and without understanding what data type values (data points) reflect, it is easy to make false claims or use incorrect statistical tests.

Discrete and continuous variables

Another important way to classify the data you are looking at, is to distinguish between discrete or continuous types of data.

Discrete data:

- has a finite set of values
- cannot be subdivided (rolling of the dice is an example, you can only roll a 6, not a 6.5!)

- a good example are binomial values, where only two values are present, for example, a patient develops a complications, or they do not

Continuous data:

- has infinite possibilities of subdivisions (for example, 1.1, 1.11. 1.111 etc.)
- an example I used was the measure of blood pressure, and the possibility of taking ever more detailed readings depending on the sensitivity of the equipment that is being used
- is mostly seen in a practical manner, i.e. although we can keep on halving the number of red blood cells per litre of blood and eventually end up with a single (discrete) cell, the absolutely large numbers we are dealing with make red blood cell count a continuous data value

In the next lesson, I will look at why knowledge about the data type is so important. Spoiler: The statistical tests used is very different depending on which types of data you are working with.

Summarising data through simple descriptive statistics

Describing the data: measures of central tendency and dispersion

Research papers show summarized data values, (usually) without showing the actual data set. Instead, key methods are used to convey the essence of the data to the reader. This summary is also the first step towards understanding the research data. As humans we cannot make sense of large sets of numbers or values. Instead, we rely on summaries of these values to aid this understanding.

There are three common methods of representing a set of values by a single number - the mean, median and mode. Collectively, these are all measures of central tendency, or sometimes, point estimates.

Most papers will also describe the actual size of the spread of the data points, also known as the dispersion. This is where you will come across terms such as range, quartiles, percentiles, variance and the more common, standard deviation, often abbreviated as SD.

Measures of central tendency

Let's summarize what we've just learned about mean, median and mode. They are measures of central tendency. As the name literally implies, they represent some value that tends to the middle of all the data point values in a set. What they achieve in reality, is to summarize a set of data point values for us, replacing a whole set of values with a single value, that is somehow representative of all the values in the set.

As humans we are poor at interpreting meaning from a large set of numbers. It is easier to take meaning from a dataset if we could consider a single value that represents that set of numbers. In order for all of this to be meaningful, the measure of central tendency must be an accurate reflection of all the actual values. No one method would suffice for this purpose and therefore we have at least these three.

Mean

- or average refers to the simple mathematical concept of adding up all the data point values for a variable in a dataset and dividing that sum by the number of values in the set
- is a meaningful way to represent a set of numbers that do not have outliers (values that are way different from the large majority of numbers)
- Example: the average or mean for this data set is 15 ((3+4+5+8+10+12+63)/7 = 15)).

3	4	5	8	10	12	63

Median

- is a calculated value that falls right in the middle of all the other values. That means that half of the values are higher than and half are lower than this value, irrespective of how high or low they are (what their actual values are)
- are used when there are values that might skew your data, i.e. a few of the values are much different from the majority of the values
- in the example above (under mean) the value 15 is intuitively a bit of an overestimation and only one of the values (63) is larger than it, making it somehow unrepresentative of the other values (3, 4, 5, 8, 10, and 12)
- Example (below): The calculation is based on whether there are an odd or even number of values. If odd, this is easy. There are seven numbers (odd). The number 8 appears in the data set and three of the values (3, 4, 5) are lower than 8 and three (10, 12, 63) are higher than 8, therefore the median is 8. In case of an even number of values, the average of the middle two is taken to reach the median.

3	4	5	8	10	12	63

Mode

- is the **data value that appears most frequently**
- is used to describe categorical values
- returns the value that occurs most commonly in a dataset, which means that some datasets might have more than one mode, leading to the terms bimodal (for two modes) and multimodal (for more than two modes)
- Example: Think of a questionnaire in which participants could choose between values of 0 to 10 to indicate their amount of pain after procedure:

4	5	6	7	8	10	4	2	1	3	4	4	8	9	4	10

Then arrange the numbers in order:

1	2	3	**4**	**4**	**4**	**4**	**4**	5	6	7	8	8	9	10	10

It is evident now that most chose the value of 4, so 4 would be the mode.

Note: It would be incorrect to use mean and median values when it comes to categorical data. **Mode is most appropriate for categorical data types**, and the only measure of central tendency that can be applied to nominal categorical data types. In the case of ordinal categorical data such as in our pain score example above, or with a Likert scale, a mean score of 5.5625 would be meaningless. Even if we rounded this off to 5.6, it would be difficult to explain what .6 of a pain unit is. If you consider it carefully, even median suffers from the same shortcoming.

Readings in this video

1. Mgelea E et al. Detecting virological failure in HIV-infected Tanzanian children. S Afr Med J. 2014;104(10):696-9. doi:10.7196/samj.7807] http://www.samj.org.za/index.php/samj/article/view/7807/6241
2. Naidoo, S., Wand, H., Abbai, N., & Ramjee, G. (2014). High prevalence and incidence of sexually transmitted infections among women living in Kwazulu-Natal, South Africa. AIDS Research and Therapy, 11(1), 31. doi:10.1186/1742-6405-11-31 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4168991/pdf/1742-6405-11-31.pdf you will note that they represented their age and duration analysis by the median. In their case they described a median age of 28, and a median duration of 12 months.

Measures of dispersion

There are several measures of dispersion such as range, quartiles, percentiles, variance and the more common standard deviation (SD). Whereas measures of central tendency give us a single-value representation of a dataset, measures of dispersion summarizes for us how spread out the dataset is.

Range

- refers to the difference between the minimum and maximum values
- is usually expressed by noting both the values
- it is used when simply describing data, i.e. when no inference is called for

Quartiles

- divide the group of values into four equal quarters, in the same way that median divides the dataset into two equal parts
- has a zeroth value, which represents the minimum value is a dataset, and a fourth quartile which represents the maximum value
- has a first quartile, which represents the value in the dataset, which will divide that set into a quarter of the values being smaller than the first quartile value and three-quarters being larger than that value
- has a third quartile, which represents the value in the dataset, which will divide that set into three-quarters of the values being smaller than the third quartile value and one quarter being larger than that value
- has a second quartile value, which divides the dataset into two equal sets and is nothing other than the median
- The zeroth value is the same as the minimum value and the fourth quartile value is the same as the maximum value.

Percentile

- looks at your data in finer detail and instead of simply cutting your values into quarters, you can calculate a value for any percentage of your data points
- turns the first quartile into the 25th percentile, the median (or second quartile) into a 50th percentile and a third quartile into a 75th percentile (and all of these are just different expression of the same thing)
- also includes a percentile rank that gives a percentage of values that fall below any value in your set that you decide on, i.e. a value of 99 might have a percentile rank of 13 meaning that 13% of the values in the set are less than 99 and 87% are larger than 99

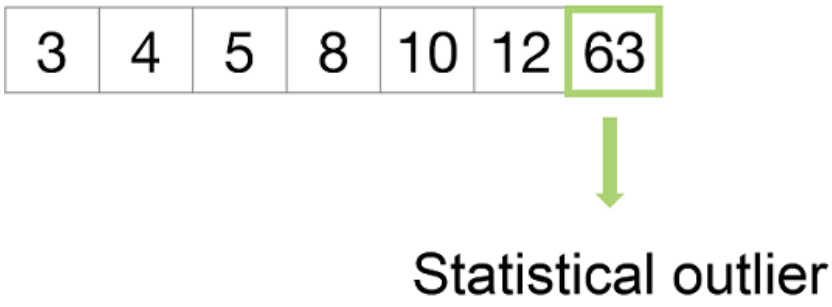
The Interquartile Range and Outliers

The interquartile range (IQR) is the difference between the values of the first and third quartiles. A simple subtraction. It is used to determine statistical outliers.

Extreme or atypical values which fall far out of the range of data points are termed ‘outliers’ and can be excluded.

For example:

Remember our initial sample values from the lecture?



With this small data set we can intuitively see that 63 is an outlier. When datasets are much larger this might not be so easy and outliers can be detected by multiplying the interquartile range (IQR) by 1.5. This value is subtracted from the first-quartile and added to the third-quartile. Any value in the data set that is lower or higher than these values can be considered statistical outliers.

Outlier values will have the biggest impact on the calculation of mean (rather than on the mode or median). Such values can be omitted from analysis if it is reasonable to do so (i.e. incorrect data input or machine error) and the researcher states that this was done and why. If the value(s) is / are rechecked and confirmed as valid, special statistical techniques can help reduce the skewing effect.

Variance and standard deviation

The method of describing the extent of **dispersion or spread of data values in relation to the mean** is referred to as the variance. We use the square root of the variance, which is called the standard deviation (SD).

- Imagine all the data values in a data set are represented by dots on a straight line, i.e. the familiar x-axis from graphs at school. A dot can also be placed on this line representing the mean value. Now the distance between each point and the mean is taken and then averaged, so as to get an average distance of how far all the points are from the mean. Note that we want distance away from the mean, i.e. not negative values (some values will be smaller than the mean). For this mathematical reason all the differences are squared, resulting in all positive values.
- The average of all these values is the variance. The square root of this is then the SD, the average distance that all the data points are away from the mean.
- As an illustration, the data values of 1, 2, 3, 20, 38, 39, and 40 have a much wider spread (standard deviation), than 17, 18, 19, 20, 21, 22, and 23. Both sets have an average of 20, but the first has a much wider spread or SD. When comparing the results of two group we should always be circumspect when large standard deviations are reported and especially so when the values of the standard deviations overlap for the two groups.

Readings referred to in this video:

1. Mgelea E et al. Detecting virological failure in HIV-infected Tanzanian children. S Afr Med J. 2014;104(10):696-9. doi:10.7196/samj.7807] http://www.samj.org.za/index.php/samj/article/view/7807/6241
2. Naidoo, S., Wand, H., Abbai, N., & Ramjee, G. (2014). High prevalence and incidence of sexually transmitted infections among women living in Kwazulu-Natal, South Africa. AIDS Research and Therapy, 11(1), 31. doi:10.1186/1742-6405-11-31 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4168991/pdf/1742-6405-11-31.pdf

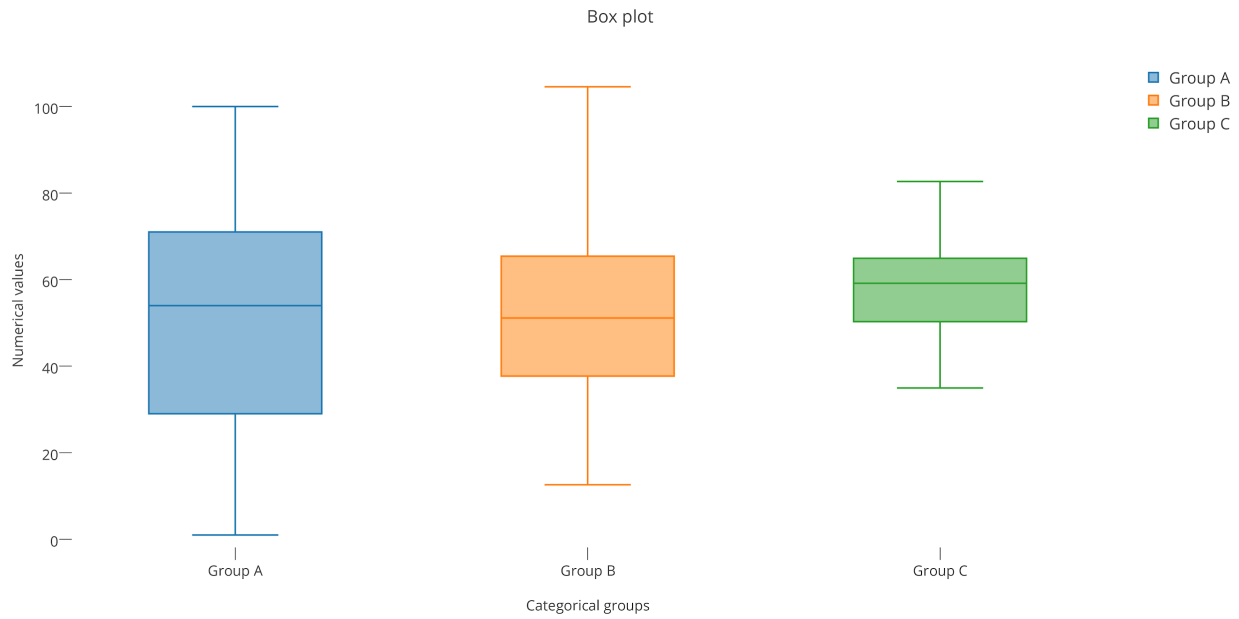
Plots, graphs and figures

Most published articles, poster presentations and indeed almost all research presentations make use of graphs, plots and figures. The graphical representation of data allows for compact, visual and information rich consumptions of data.

It is invariable much easier to understand complex categorical and numerical data when it is represented in pictures. Now that you have a good understanding of the different data types, we will take a look at the various ways graph them.

Box and whisker plots

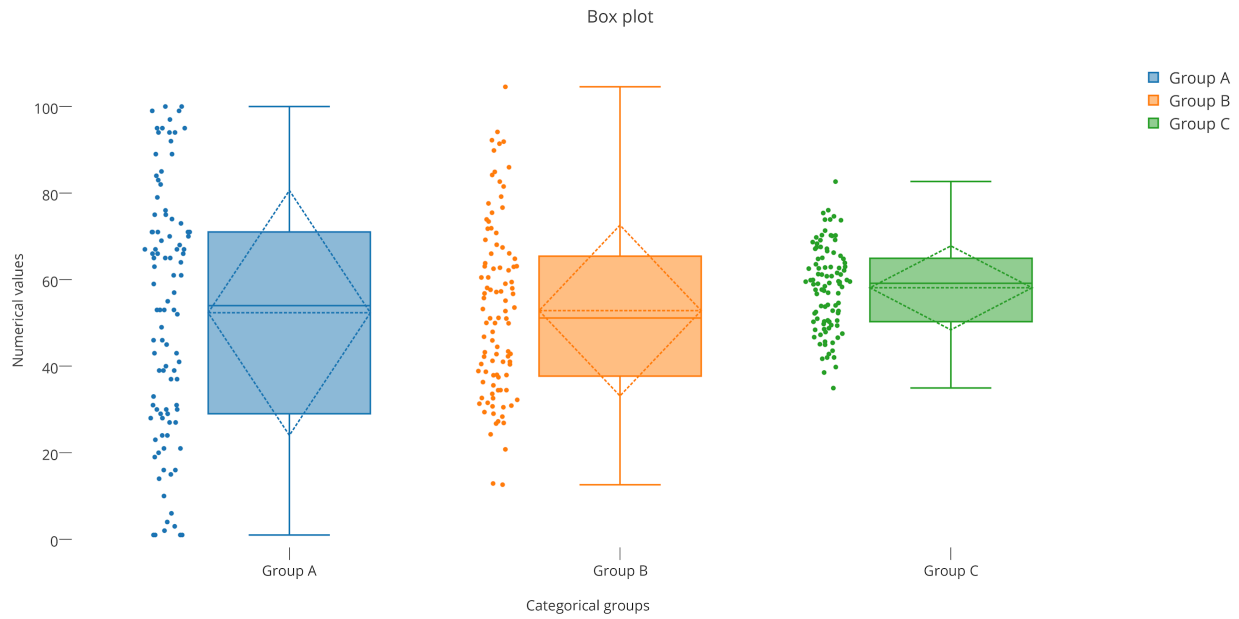
A box and whisker plot provides us with information on both the measures of central tendency as well as measures of spread. It takes the list of numerical data point values for categorical variables and makes use of quartiles to construct a rectangular block.



In the example above we see three categorical groups (Group A, B and C) on the *x*-axis and some numerical data type on the *y*-axis. The rectangular block has three lines. The bottom line (bottom of the rectangle if drawn vertically) represents the first quartile value for the list of numerical values and the top line (top of the rectangle) indicates the third quartile value. The middle line represents the median.

The whiskers can represent several values and the authors of a paper should make it clear what their whisker values represent. Possible values include: minimum and maximum values, values beyond which statistical outliers are found (one-and-a-half times the interquartile range below and above the first and third quartiles), one standard deviation below and above the mean or a variety of percentiles (2nd and 98th or 9th and 91st).

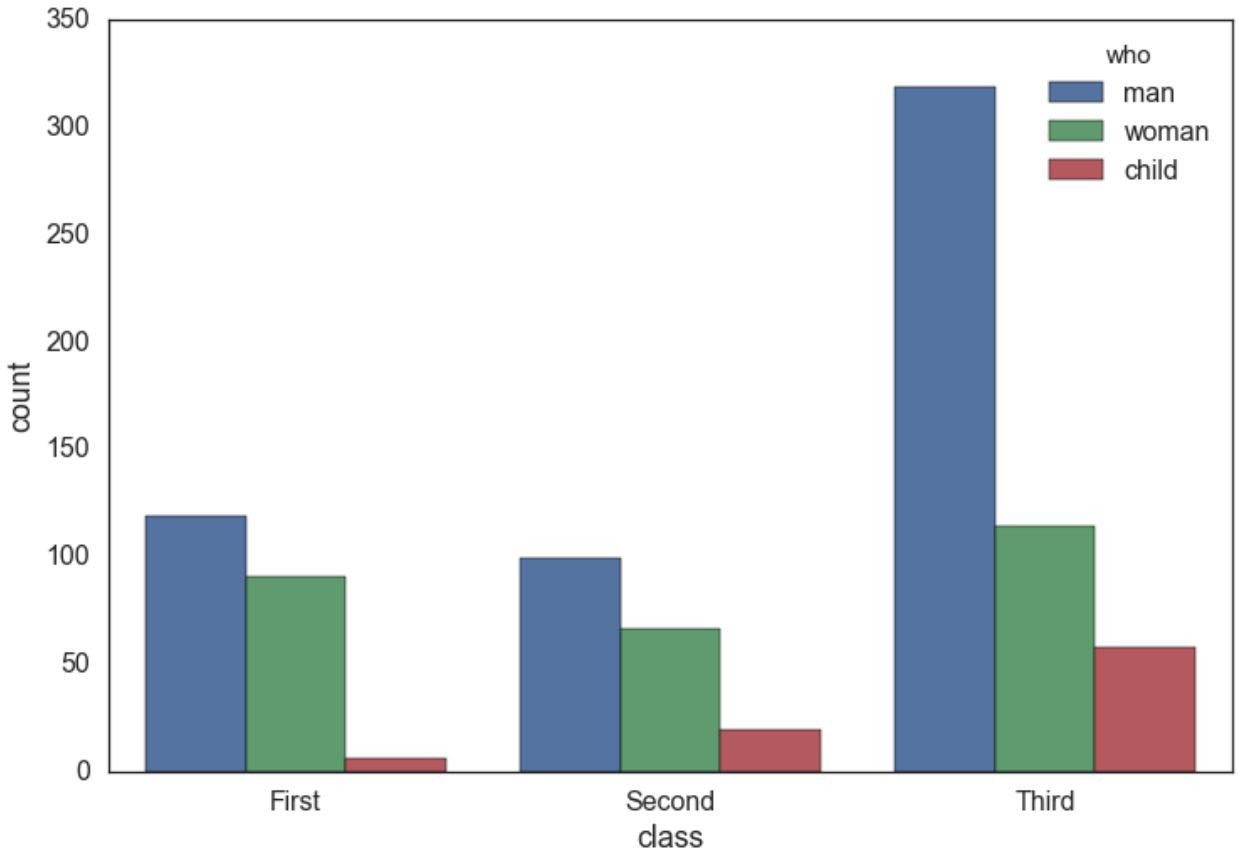
Some authors also add the actual data points to these plots. The mean and standard deviation can also be added as we can see in the graph below (indicated by dotted lines).



For both graphs above the whiskers indicate the minimum and maximum values. Note clearly how the box and whisker plot gives us an idea of the spread and central tendency of numerical data points for various categories.

Count plots

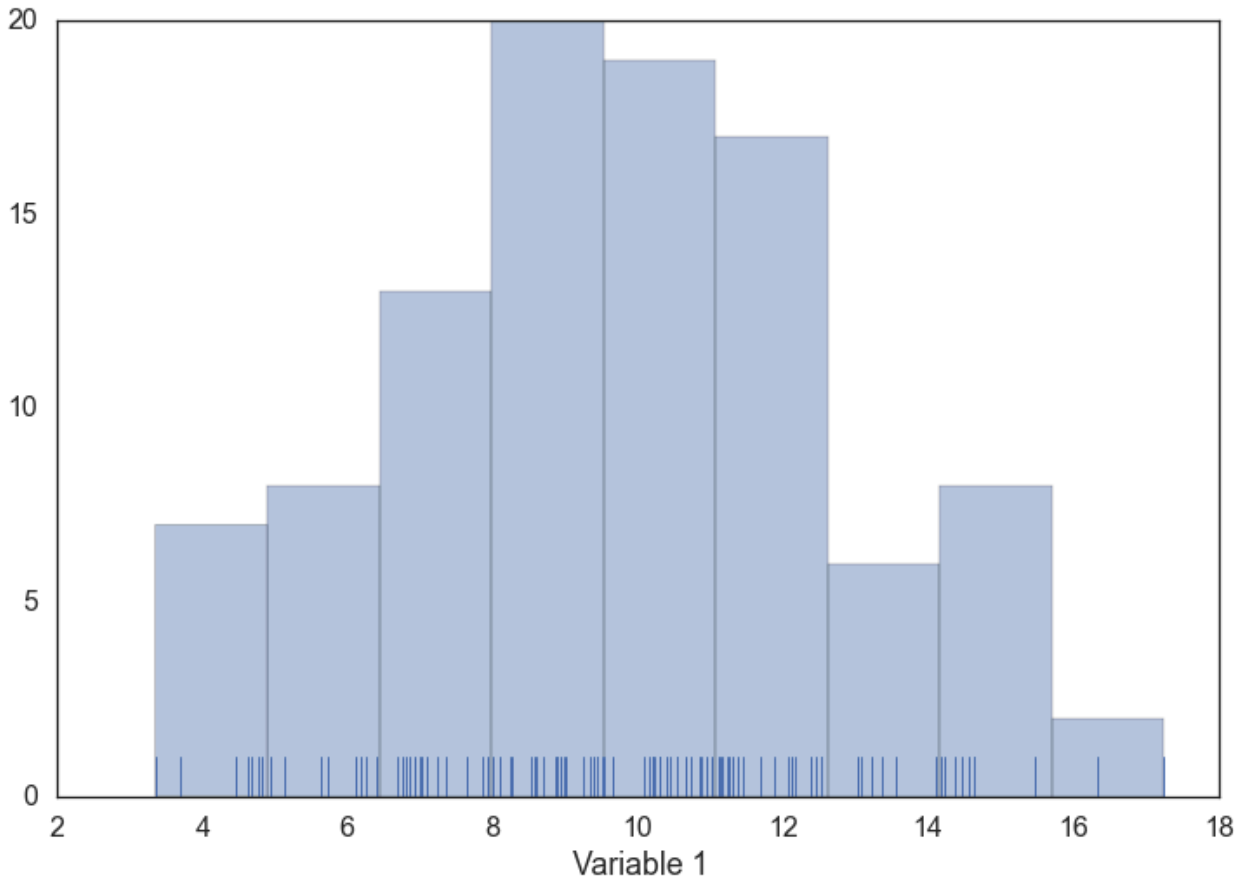
Count plots tell us how many times a categorical data point value or discrete numerical value occurred. Count plots are often referred to as bar plots, but the term *bar plot* is often misused.



This count plot is taken from a famous data set and represents the number of passengers on the titanic divided by age and gender.

Histogram

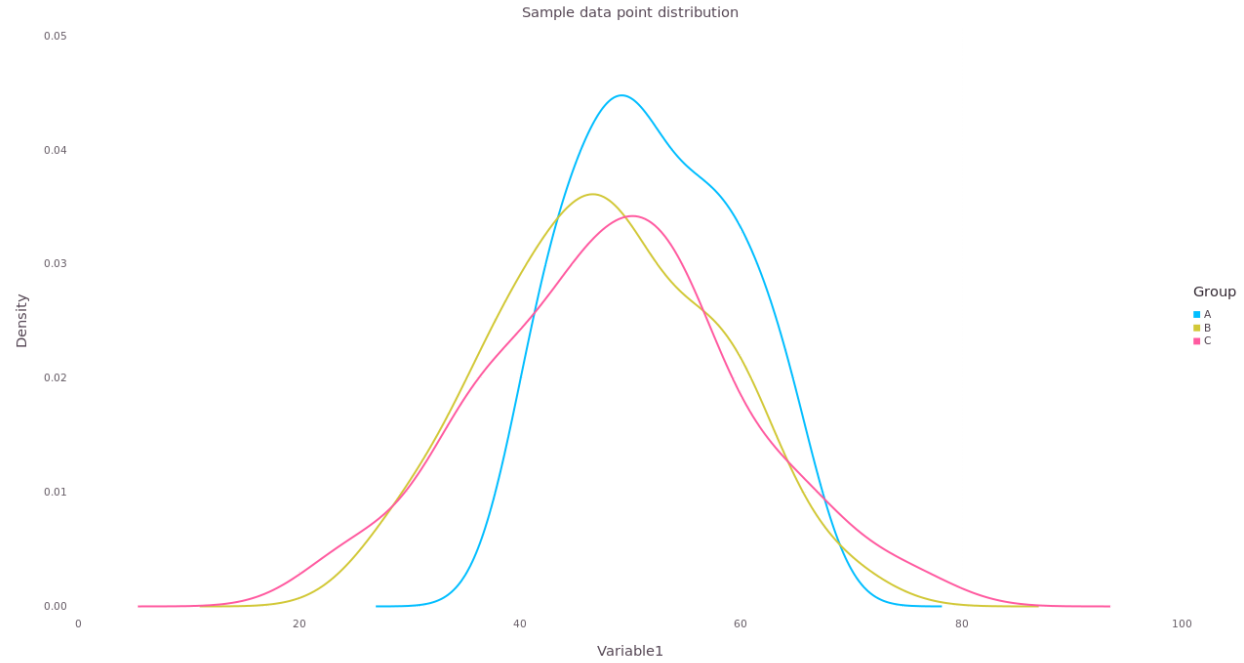
A histogram is different from a count plot in that it takes numerical values on the *x*-axis. From these, so-called bins, are constructed. A bin represents a minimum and a maximum cut-off value. Think of the counting numbers 0 to 20. Imagine now have a list of 200 values, all taken from 0 to 20, something like 1, 15, 15, 13, 12, 20, 13, 13, 13, 14, ... and so on. I can construct bins with a size of five, i.e. 0 to 5, 6 to 10, 11 to 15, and 16 to 20. I can now count how many of my list of 200 fall into each bin. From the deprecated list here, there is already 8 values in the 11 to 15 bin. The size of the bins are completely arbitrary and the choice is up to the researcher(s) involved in a project.



The graph above actually adds what is termed a *rug plot* at the bottom, showing the actual data points, which then represents how many there are in each bin.

Distribution plots

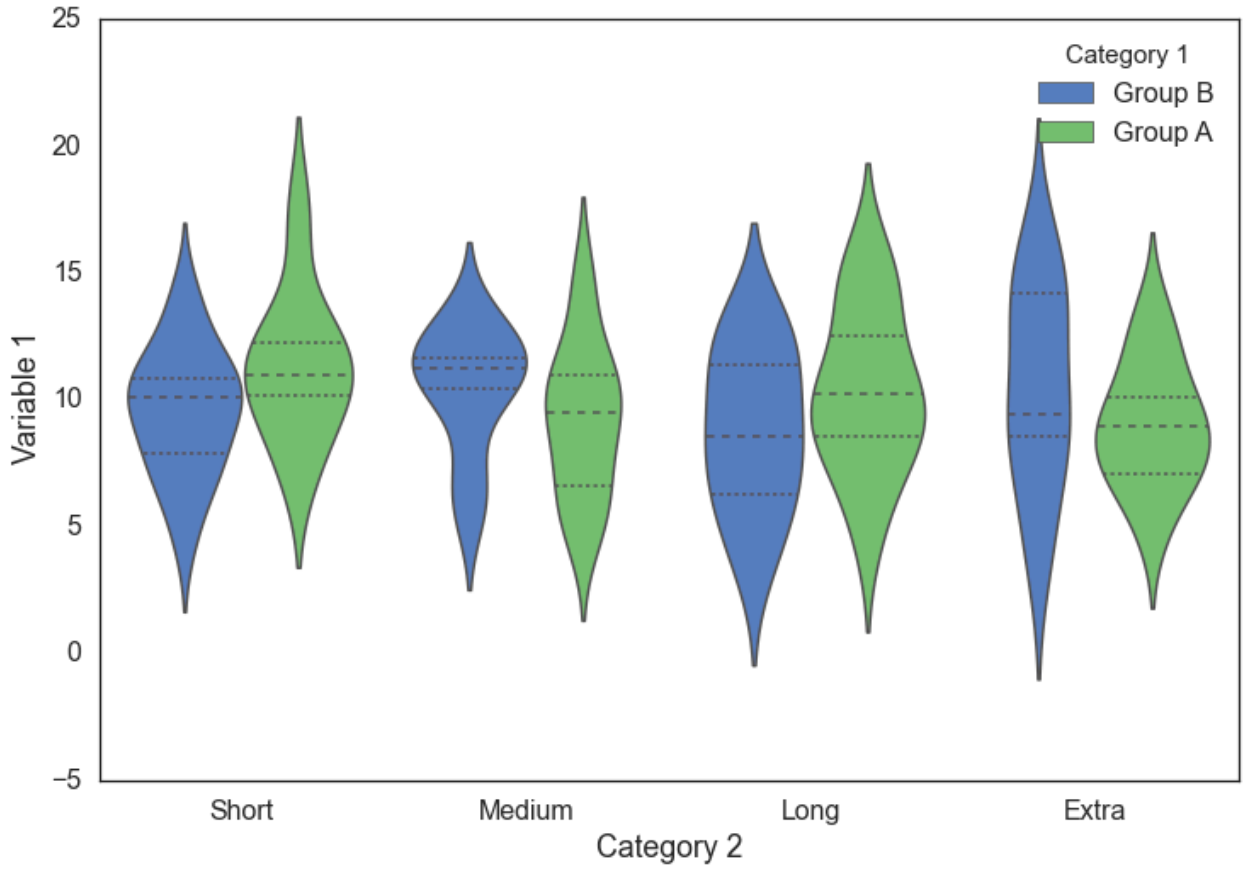
These plots take histograms to the next level and, through mathematical equations, gives us a visual representation of the distribution of the data.



Note that, as with the histogram, the *x*-axis is a numerical variable. If you look closely you'll see the strange values on the*y*-axis. Later in the course we will learn that it is from these graphs that we calculate p-values. For now, it is clear that they give us a good indication of what the distribution shape is like. For the graph above, showing densities for three groups, we note that some values (at around 50) occur much more commonly than values at 20 or 80 as indicated by the height of the lines at each of these *x*-axis values. The distributions here are all nearly bell-shaped, called the normal distribution.

Violin plots

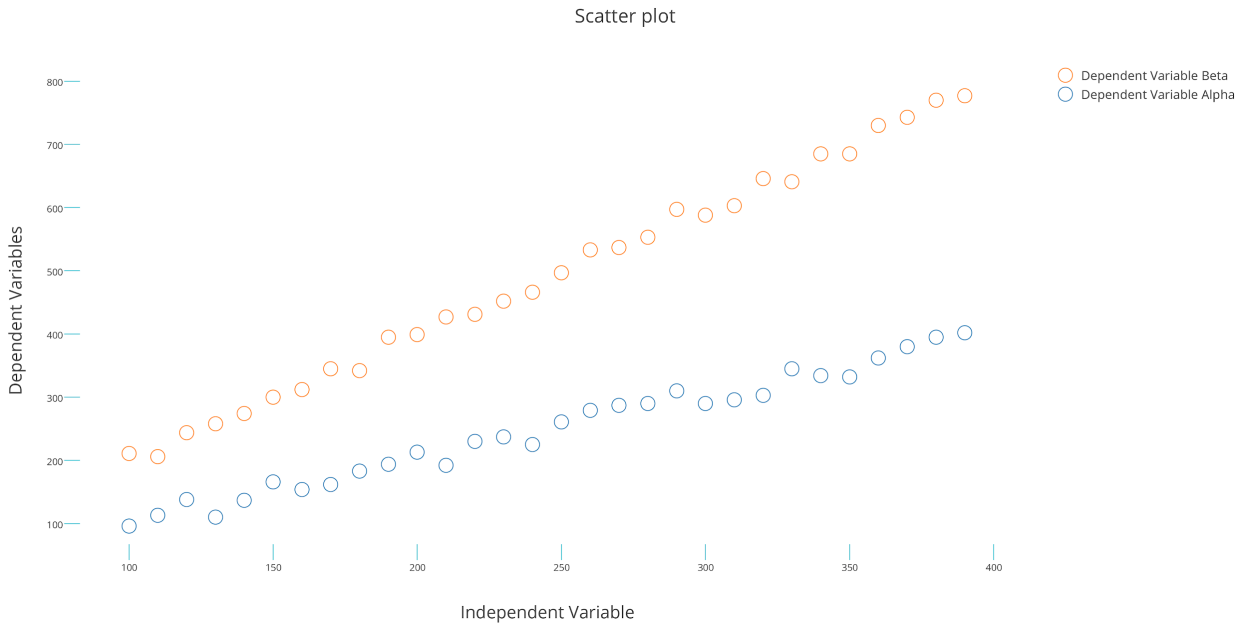
Violin plots combine box and whisker plots and density plots. they actually take the density plots, turn them on their sides and mirror them.



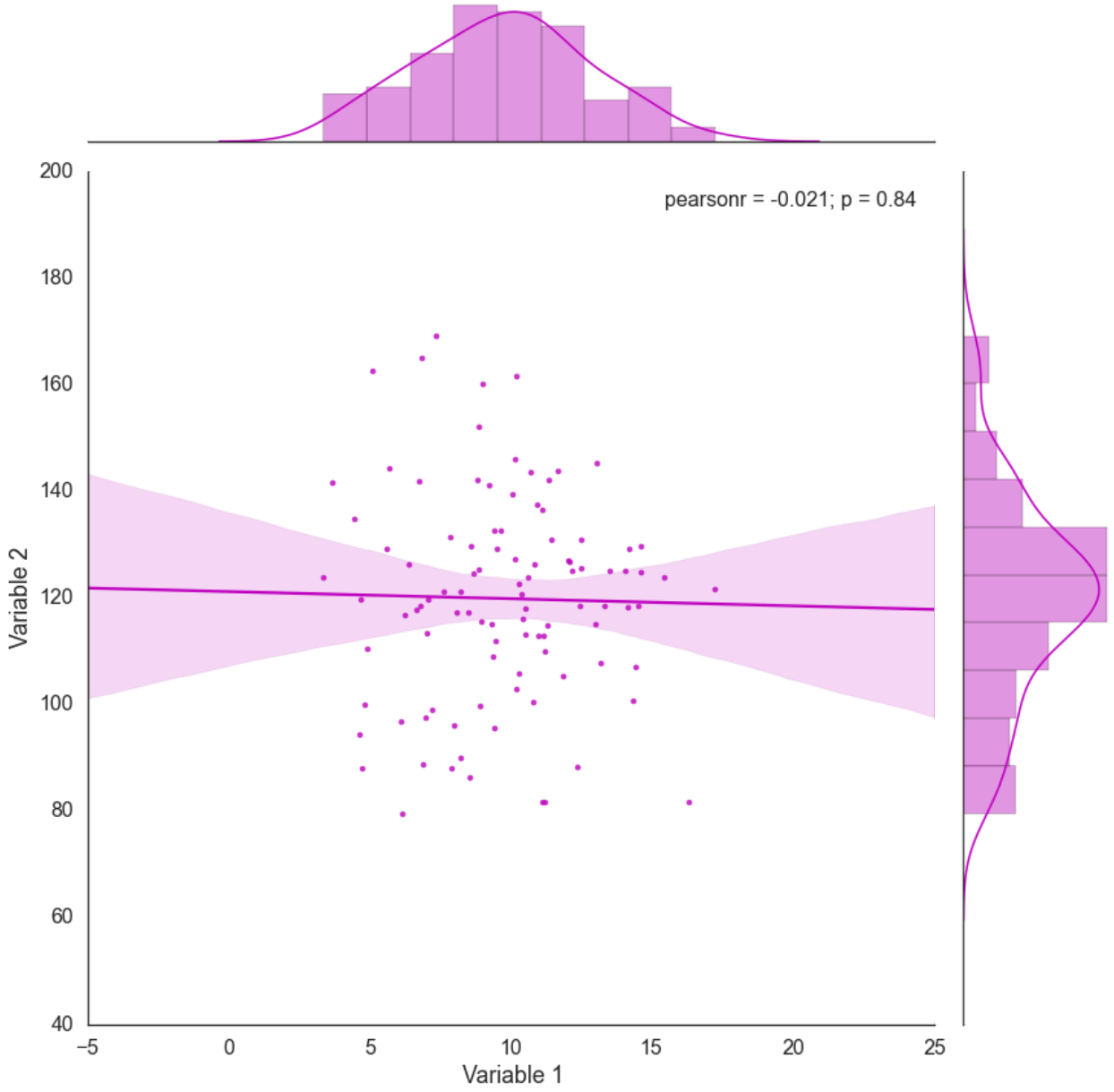
In the graph above we have dotted lines indicating the median and first and third quartiles as with box and whisker plots, but we get a much better idea of the distribution of the numerical values.

Scatter plots

Scatter plots combine sets of numerical values. each dot has a value from two numerical data point sets, so as to make a *x*- and a *y*-coordinate. Think of a single patient with a white cell count and a red cell count value.

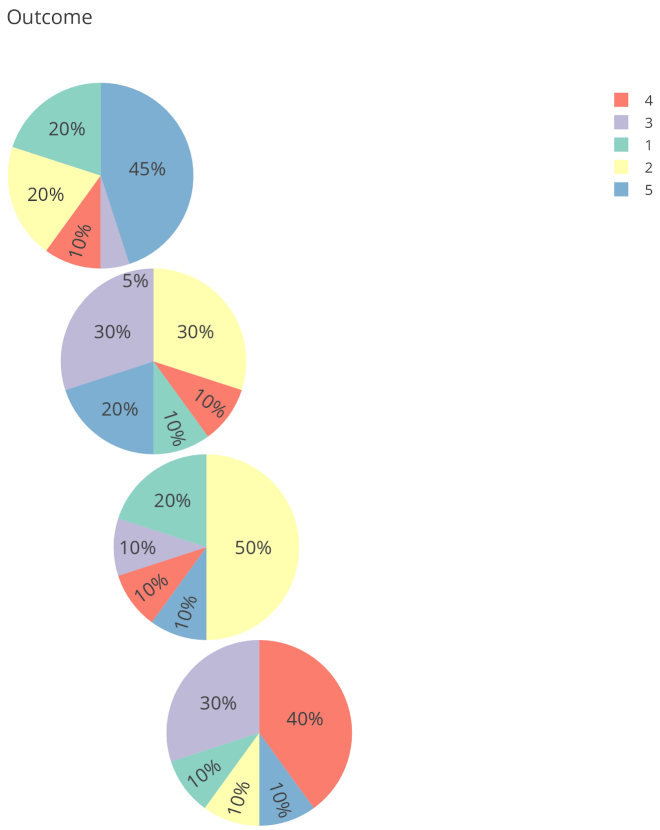


From the graph above it is clear that both axes have numerical values. Through mathematical equations, we can even create lines that represent all these points. These are useful for creating predictions. Based on any value on the *x*-axis, we can calculate a predicted value on the *y*-axis. Later we will see that this is a form of linear regression and we can use it to calculate how well two sets of values are correlated. The graph below show such a line and even adds a histogram and density plot for each of the two sets of numerical variables.



Pie chart

Lastly, we have to mention the poor old pie chart. Often frowned upon in scientific circles it nonetheless has its place.



The plot above divides up each circle by how many of each of the values 1-through-5 occurred in each data set.

Sampling

Introduction

This course tackles the problem of *how* healthcare research is conducted, but have you ever wondered, on a very fundamental level, *why* healthcare research is conducted? Well, the *why* is actually quite easy. We would like to find satisfactory answers to questions we have about human diseases and their prevention and management. There are endless numbers of questions. The *how* is more tricky.

To answer questions related to healthcare, we need to investigate, well, humans. Problem is, there are so many of us. It is almost always impossible to examine all humans, even when it comes to some pretty rare diseases and conditions. Can you imagine the effort and the cost?

To solve this problem, we take only a (relatively) small group of individuals from a larger population that contains people with the disease or trait that we would like to investigate. When we analyse the data pertaining to the sample selection and get answers to our questions, we need to be sure that these answers are useful when used in managing everyone who was not in the sample.

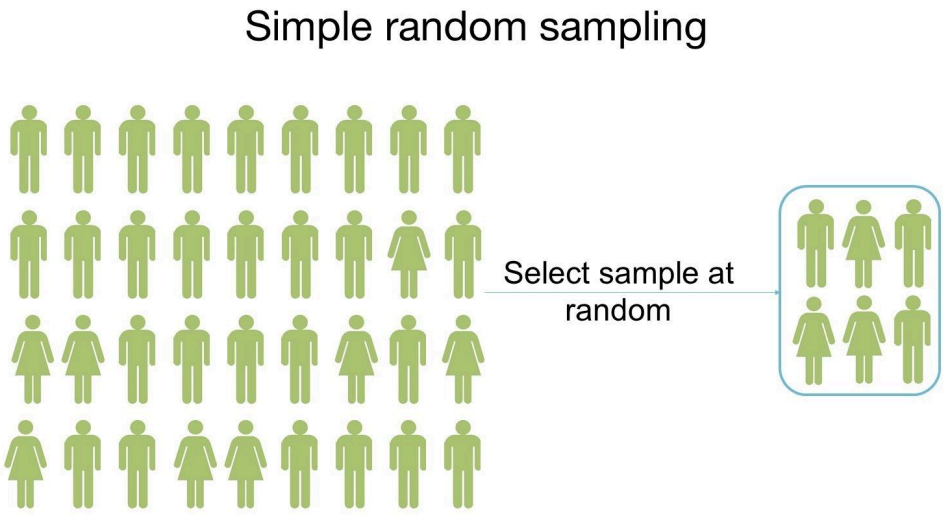
In order to do this, we must be sure that the sample properly reflects the larger population. The larger the sample size, the higher the likelihood that the results of our analyses correctly infers to the population. When a sample is not properly representative of the population to which the results will infer, some sort of bias was introduced in the selection process of the sample. All research must strive to minimize bias, or if it occurred, to properly account for it.

This section explains a few of the methods that are used to sample participants for studies from larger populations.

Types of sampling

Simple random sampling

In *simple random sampling* a master list of the whole population is available and each individual on that list has an equal likelihood of being chosen to be part of the sample.

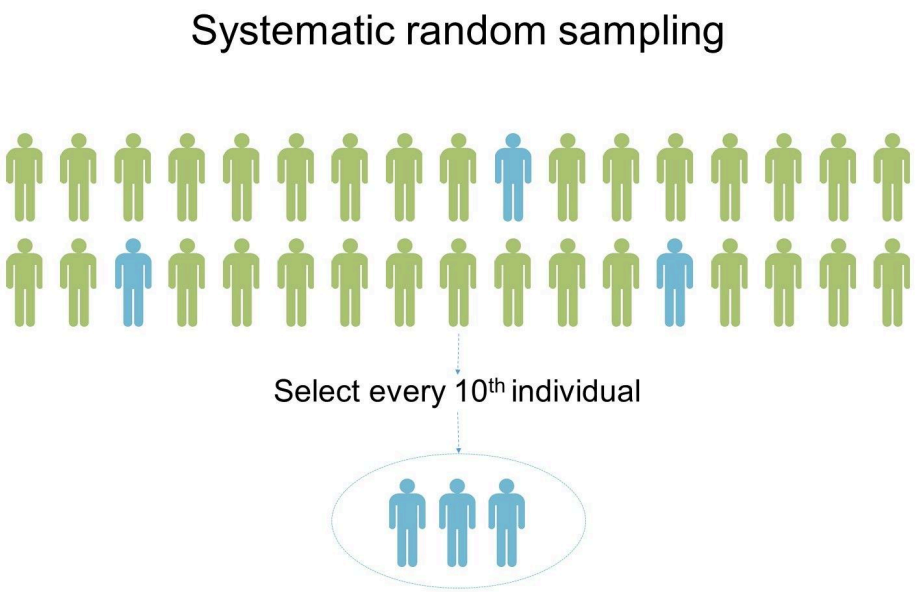


This form of sampling can be used in two settings. On a larger scale we might have a master list of patients (or people with a common trait) who we would like to investigate. We can draw from that

list using simple random sampling. On a smaller scale, we might already have a list of participants for a clinical trial. We now need to divide them into different groups. A good example would be dividing our participants into two groups, group one receiving an active drug and group two, a placebo. Each individual participant must have an equal likelihood of getting either drug. This can be achieved by simple random sampling.

Systematic random sampling

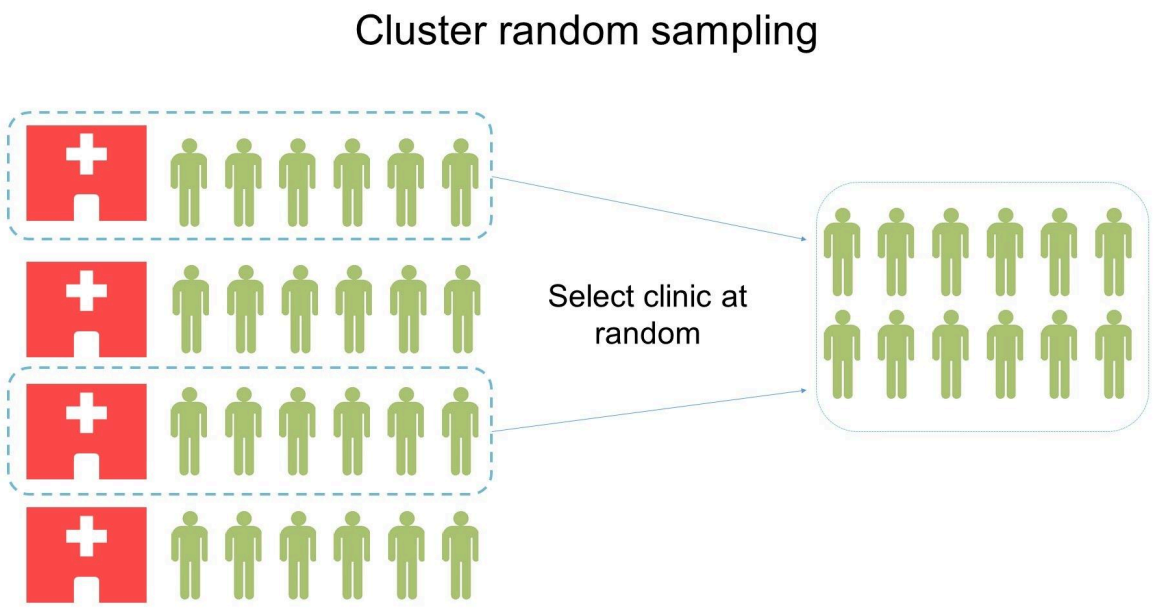
In *systematic random sampling*, the selection process iterates over every decided number of individuals, i.e. every 10th or 100th individual on a master list.



We can again consider two scenarios. In the first, we are dealing again with finding participants for a study and in the second, we already have our participants selected and now need to divide them into groups.

Cluster random sampling

In *cluster sampling*, the groups of individuals that are included are somehow clustered, i.e. all in the same space, location, or allowed time-frame. There are many forms of cluster random sampling. We often have to deal with the fact that a master list simply does not exist or is too costly or difficult to obtain. Clustering groups of individuals greatly simplifies the selection process.

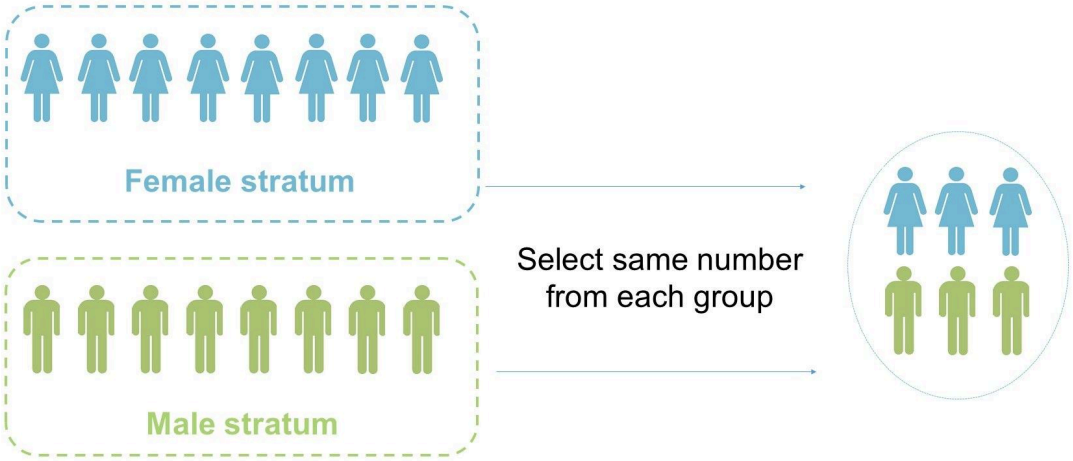


We could even see the trivial case of a case series or case-control series as having made use of clustering. A study might compare those with and without a trait, say a postoperative wound complication. The sample is taken from a population who attended a certain hospital over a certain time period.

Stratified random sampling

In *stratified sampling* individuals are chosen because of some common, mutually exclusive trait. The most common such trait is gender, but may also be socio-economic class, age, and many others.

Stratified random sampling



Week 3: Building an intuitive understanding of statistical analysis

From area to probability

P-values

When you read medically related research papers you are likely to come across the concept of probability and the term p-value, along with the gold standard of statistical significance - 0.05.

What is the p-value?

- The p-value explains a probability of an event occurring
- It is based on the calculation of a geometrical area
- The mathematics behind a p-value draws a curve and simply calculates the area under a certain part of that curve

Rolling dice

I explained the notion of probability by using the common example of rolling dice:

- For each die, there is an equal likelihood of rolling a one, two, three, four, five, or a six
- The probability of rolling a one is one out of six or 16.67% (it is customary to write probability as a fraction (between zero and one) as opposed to a percentage, so let's make that 0.1667)
- It is **impossible** to have a negative probability (a less than 0% chance of something happening) or a probability of more than one (more than a 100% chance of something happening)
- If we consider all probabilities in the rolling of our die, it adds to one (0.1667 times six equals one)
- This is our probability space, nothing exists outside of it

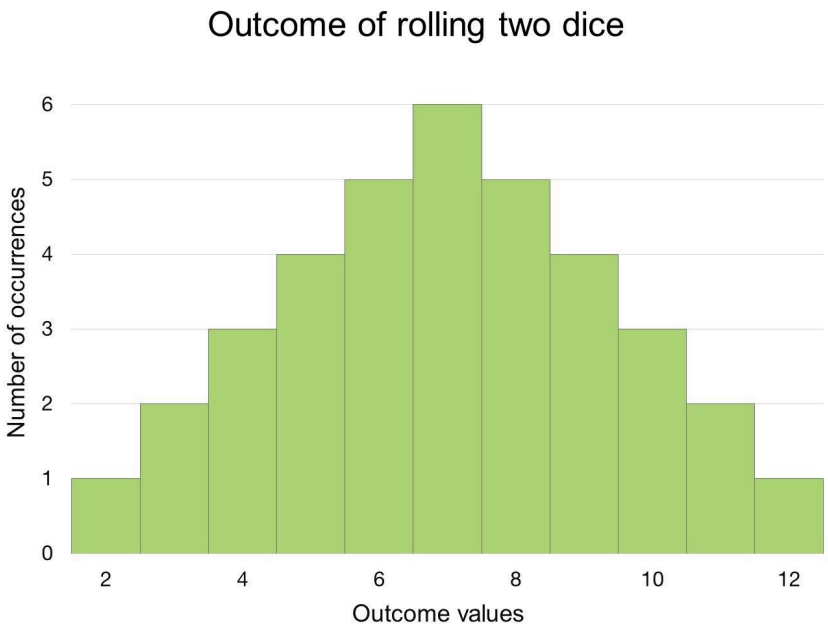
- By looking at it from a different (more of a medical statistics) point of view, I could roll a five and ask the question: "What was the probability of finding a five?", to which we would answer, $p = 0.1667$
- I could also ask: "What is the likelihood of rolling a five or more?", to which the answer is, $p = 0.333$

Example of rolling a pair of dice:

- We hardly ever deal with single participants in a study, so let's ramp things up to a pair of dice
- If you roll a pair of dice, adding the values that lands face-up, will leave you with possible values between two and 12
- Note that there are 36 possible outcomes (using the fact that rolling, for example, a one and a six is *not* the same as rolling a six and a one)
- Since there are six way of rolling a total of seven, the chances are six in 36 or 0.1667
- Rolling two sixes or two one are less likely at one out of 36 each, or 0.0278 (a 2.78% chance)

Outcome	Combinations						Total
2	1+1						1
3	1+2	2+1					2
4	1+3	3+1	2+2				3
5	1+4	4+1	2+3	3+2			4
6	1+5	5+1	2+4	4+2	3+3		5
7	1+6	6+1	2+5	5+2	4+3	3+4	6
8	2+6	6+2	3+5	5+3	4+4		5
9	3+6	6+3	4+5	5+4			4
10	4+6	6+4	5+5				3
11	5+6	6+5					2
12	6+6						1

We can make a chart of these called a *histogram*. It shows how many times each outcome can occur. You'll note the actual outcomes on the horizontal axis and the number of ways of achieving each outcomes on the vertical axis.



Equating geometrical area to probability

We could also chart a *probability plot*. So, instead of the actual number of times, we chart the probability on the y-axis. It is just the number of occurrences divided by the total number of outcomes.

Continuous data types

We started this lesson off by looking at discrete outcomes, i.e. the rolling of dice. The outcomes were discrete by way of the fact that no values exist between the whole numbers two to 12 (rolling two dice). That made the shape of the probability graphs quite easy, with a base width of one and all in the shape of little rectangles. Continuous variables on the other hand are considered to be infinitely divisible which makes it very difficult to measure the geometric width of those areas, not to mention the fact that the top ends of the little rectangles are curved and not straight.

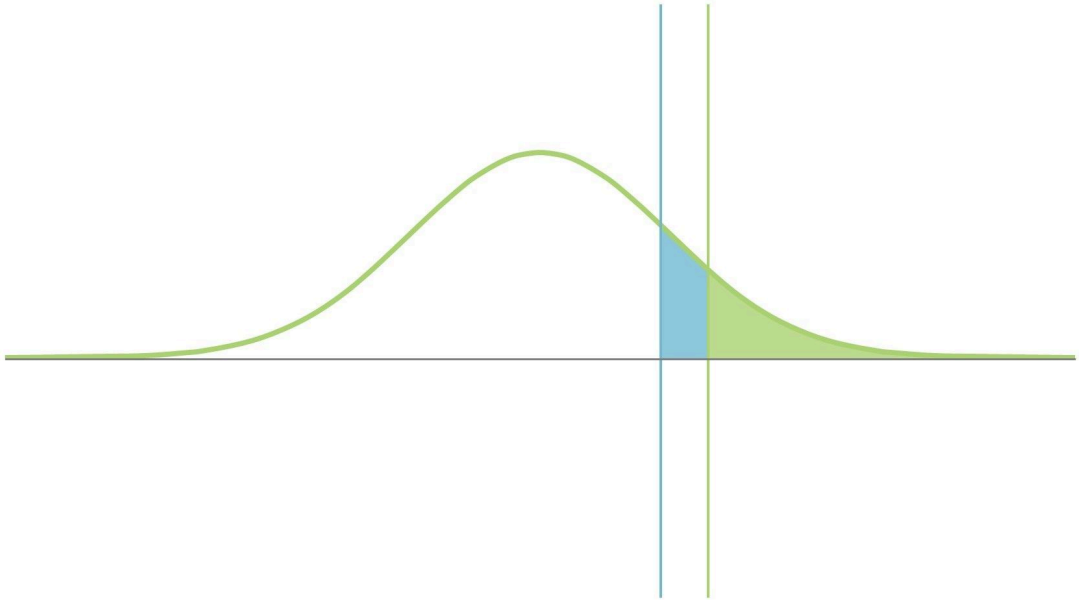
Integral calculus solves the problem of determining the area of an irregular shape (as long as we have a nice mathematical function for that shape). Our bigger problem is the fact that (for continuous variables) it is no longer possible to ask what the probability of finding a single value (outcomes) is. A single value does not exist as in the case of discrete data types. Remember that with continuous data types we can (in theoretical terms at least) infinitely divide the width of the base. Now, we can only ask what the probability (area under the curve) is between two values, or

more commonly what the area is for a value larger than or smaller than a given value (stretching out to positive and negative infinity on both sides).

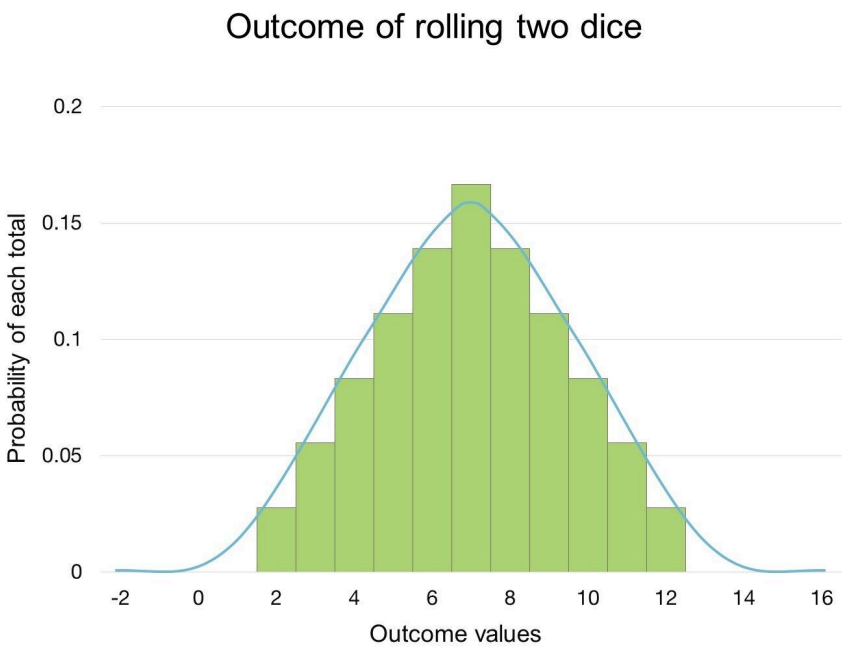
So, how does this work?

The graph below (not to scale), illustrates the *p*-value. Let's go back to the example of researching the white cell count of two groups of patients. Imagine that group one has a certain average white cell count and so does group two. There is a difference between these averages. The question is whether this difference is statistically significant? The mathematics behind the calculation is going to use some values calculated from the data point values and represent in a (bell-shaped) curve (as in the graph below).

If you chose a *p*-value of less than 0.05 to indicate a significant difference and you decided in your hypothesis that one group will have an average higher than the other, the maths will work out a cut-off on the x-axis which will indicate an area under the curve (the green) of 0.05 (5% of the total area). It will then mark the difference in averages of your data and see what the area under the curve was for this (in blue). You can see it was larger than 0.05, so the difference was not statistically significant.



And there you have it. An intuitive understanding of the *p*-value. It only gets better from here!



- As if by magic, the height of the rectangular bars are now equal to the the likelihood (*p*-value or sorts) of rolling a particular total
- Note how the height of the centre rectangle (and outcome of seven) is 0.1667 (there are six ways of rolling a seven and therefore we calculate a *p*-value 6 / 36 = 0.1667.
- If you look at each individual rectangular bar and if you consider the width of each to be one, the area of each rectangle (height times width) gives you the probability of rolling that number (the *p*-value) (for the sake of completeness, we should actually be using the probability density function, but this is an example of discrete data types and the results are the same)
- If we want to know what the probability is of rolling a 10 or more is, we simply calculate the area of the rectangles from 10 and up

The *p* then refers to probability, as in *the chance or likelihood of an event occurring (i*n healthcare research, an event is an outcome of an experiment).

An example of an experiment is comparing the difference in the average of some blood test value between two groups of patients, with the *p*-value representing the probability that the particular difference was found. If the probability was sufficiently low, we infer that the two sets of participants represent two separate sets of populations, which are then significantly different from each other.

The heart of inferential statistics: Central limit theorem

Central limit theorem

Now that you are aware of the fact that the *p*-value represents the area under a very beautiful and symmetric curve (for continuous data type variables at least) something may start to concern you. If it hasn’t, let’s spell it out. Is the probability curve always so symmetric? Surely, when you look at the occurrence of data point values for variables in a research project (experiment), they are not symmetrically arranged.

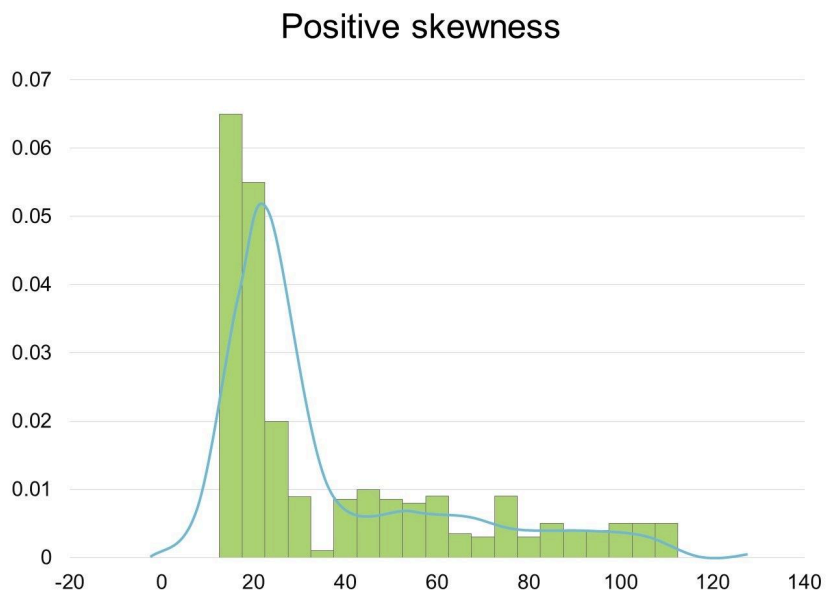
In this lesson, we get the answer to this question. We will learn that this specific difference between the means of the two groups is but one of many, many, many (really many) differences that are possible. We will also see that some differences occur much more commonly than others. The answer lies in a mathematical theorem called the Central Limit Theorem (CLT). As usual, don't be alarmed, we won’t go near the math. A few simple visual graphs will explain it quite painlessly.

Skewness and kurtosis

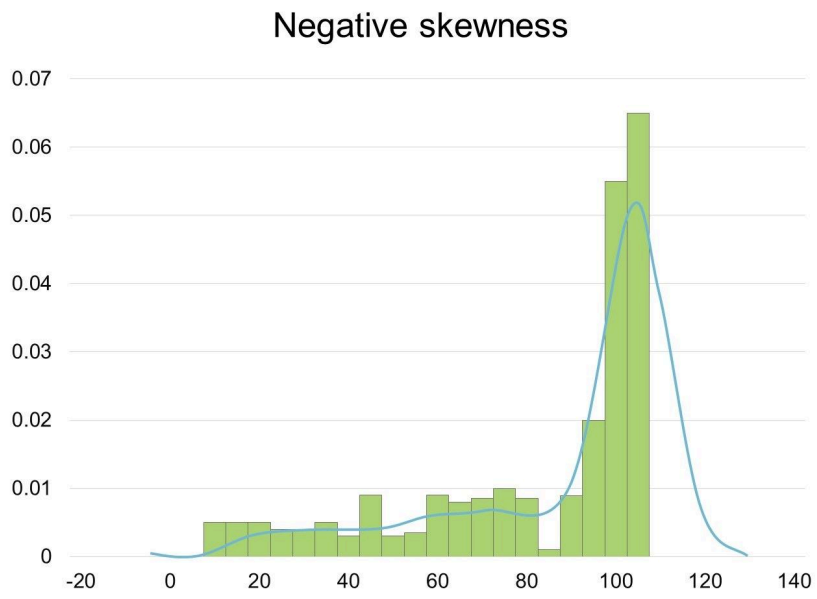
As I mentioned, data can be very non-symmetrical in its distribution. To be clear, by distribution I mean physically counting how many times each individual value comes up in a set of data point values. The two terms that describe non-symmetric distribution of data point values are *skewness* and *kurtosis*.

Skewness

Skewness is rather self-explanatory and is commonly present in clinical research. It is a marker that shows that there are more occurrences of certain data point values at one end of a spectrum than another. Below is a graph showing the age distribution of participants in a hypothetical research project. Note how most individuals were on the younger side. *Younger* data point values occur more commonly (although there seems to be some very old people in this study). The skewness in this instance is right-tailed. It tails off to the right, which means it is *positively* skewed.



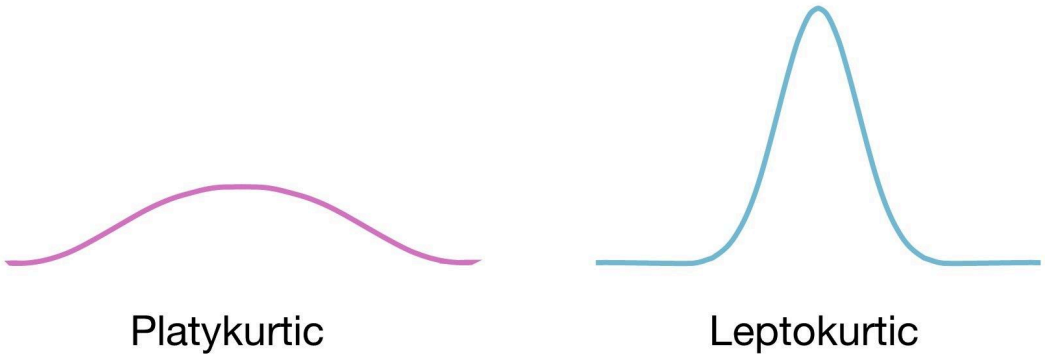
On the other hand, *negative skewness* would indicate the data is left-tailed.



Kurtosis

Kurtosis refers to the spread of your data values.

- A *platykurtic* curve is flatter and broader than normal as a result of having few scores around the mean. Large sections under the curve are forced into the tail, thereby (falsely) increasing the probability of finding a value quite far from the mean.
- A *mesokurtic* curve takes the middle ground with a medium curve from average distributions.
- In a *leptokurtic* curve is more peaked, where many values are centred around the mean.
- Remember, in this section we are discussing the distribution of the actual data point values in a study, but the terms used here can also refer to the curve that is eventually constructed when we calculate a p-value. As we will see later, these are quite different things (the curve of actual data point values and the p-value curve calculated from the data point values). This is a very important distinction.



Combinations

Combinations lie at the heart of the Central Limit Theorem and also inferential statistical analysis. It is the key for understanding the curve that we get when attempting to calculate the *p*-value.

Combinations refer to the number of ways a selection of objects taken from a group of objects can be arranged.

- In a simple example we might consider how many combination of two colors we can make from a total choice of four colors, say red, green, blue, and black. We could choose: red + green, red + blue, red + black, green + blue, green + black, and finally, blue + black (noting that choosing blue + black is the same as choosing black + blue). That is six possible combination choosing a two color combination from four choices.
- Many countries in the world have lotteries in which you pick a few numbers and hand over some money for a chance to win a large cash prize should those numbers pop up in a draw. The order doesn't matter, so we are dealing with combinations. So, if you had to choose six numbers between say one and 47, how many combinations could come up? It's a staggering 10,737,573. Over 10 million. Your choice of six numbers is but one of all of those. That means that you chances of picking the right combination is less than one in 10 million! Most lotteries have even more numbers to choose from! To put things into perspective (just for those who play the lottery), a choice of the numbers 1, 2, 3, 4, 5, and 6 (a very, very unlikely choice) is just as likely to come up as your favorite choice of 13, 17, 28, 29, 30, and 47! Good luck!

Combination has some serious implications for clinical research, though.

For example, a research project decides to focus on 30 patients for a study. The researcher chose the first 30 patients to walk through the door at the local hypertension (high blood pressure) clinic and notes down their ages. If a different group of patients was selected on a different day, there would be completely different data. The sample group that you end up with (the chosen 30) is but one of many, many, many that you could have had! If 1000 people attended the clinic and you had to choose 30, the number of possible combinations would be larger than 2.4 times 10 to the power 57. Billions upon billions upon billions!

This is how the distribution curve for the outcomes of studies (from which the p-value is calculated) are constructed. Be it the difference in means between two or more groups, or proportions of choices for a cross-sectional studies Likert-style questions. There are (mathematically) an almost uncountable different number of outcomes (given the same variables to be studied) and the one found in an actual study, is but one of those.

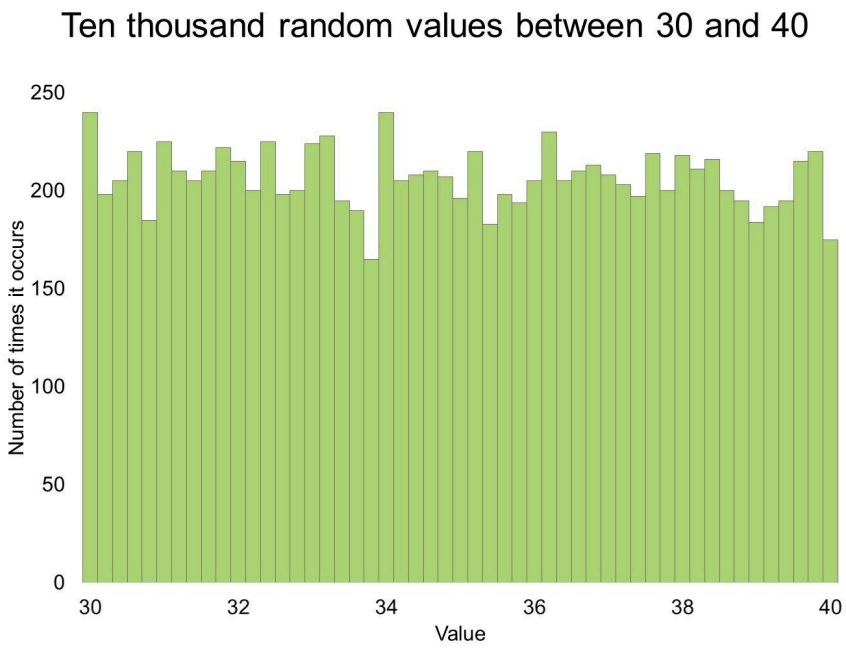
Central limit theorem

We saw in the previous section on combinations that when you compare the difference in averages between two groups, your answer is but one of many that exist. The Central Limit Theorem states that if we were to plot all the possible differences, the resulting graph would form a smooth,

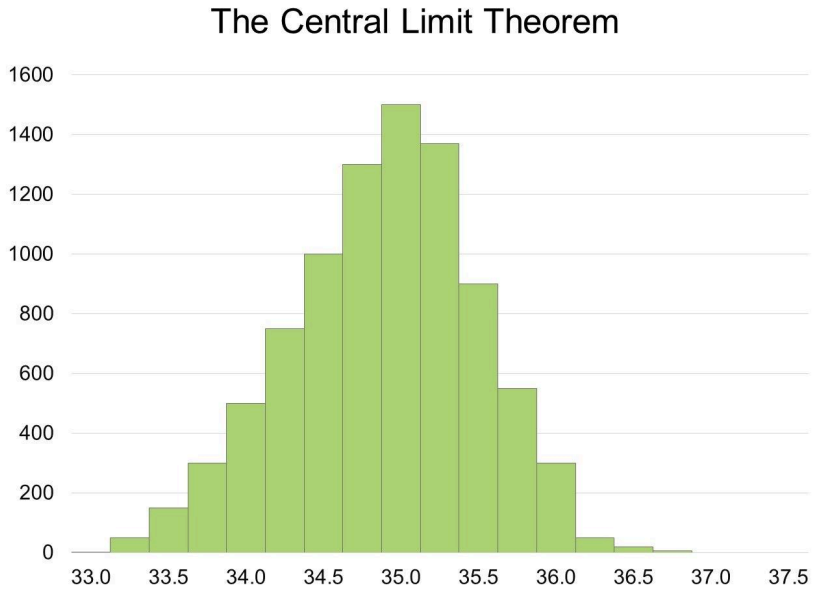
symmetrical curve. Therefore, we can do statistical analysis and look for the area under the curve to calculate our *p*-values.

The mathematics behind the calculation of the *p*-value constructs an estimation of all the possible outcomes (or differences as in our example).

Let's look at a visual representation of the data. In the first graph below we asked a computer program to give us 10,000 random values between 30 and 40. As you can see, there is no pattern.



Let's suggest that these 10,000 values represent a population and we need to randomly select 30 individuals from the population to represent our study sample. So, let's instruct the computer to take 30 random samples from these 10,000 values and calculate the average for those 30. Now, let's repeat this process 1000 times. We are in essence repeating our medical study 1000 times! The result of the occurrence of all the averages is shown in the graph below. The Central Limit Theorem predicts, a lovely smooth, symmetric distribution. Just ready and waiting for some statistical analysis.



Every time a medical study is conducted, the data point values (and their measures of central tendency and dispersion) are just one example of countless others. Some will occur more commonly than others and it is the Central Limit Theorem that allows us to calculate how likely it was to find a result as extreme as the one found in any particular study.

Distributions: the shape of data

Distributions

We all know that certain things occur more commonly than others. We all accept that there are more days with lower temperatures in winter than days with higher temperatures. In the northern hemisphere there will be more days in January that are less than 10 degrees Celsius (50 degrees Fahrenheit) than there are days that are more than 20 degrees Celsius (60 degrees Fahrenheit).

Actual data point values for any imaginable variable comes in a variety of, shall we say, shapes or patterns of spread. The proper term for this is a *distribution*. The most familiar shape is the *normal distribution*. Data from this type of distribution is symmetric and forms what many refer to as a bell-shaped curve. Most values center around the average and taper off to both ends.

If we turn to healthcare, we can imagine that certain hemoglobin level occur more commonly than other in a normal population. There is a distribution to the data point values. In deciding which type of statistical test to use, we are concerned with the distribution that the parameter takes in the population. As we will see later, we do not always know what the shape of distribution is and we

can only calculate if our sample data point values might come from a population in which that variable is normally (or otherwise) distributed.

It turns out that there are many forms of data distributions for both discrete and continuous data type variables. Even more so, averages, standard deviations, and other statistics also have distributions. This follows naturally from the Central Limit Theorem we looked at before. We saw that if we could repeat an experiment thousands of times, each time selecting a new combination of subjects, some average values or differences in averages between two groups would form a symmetrical distribution.

It is important to understand the various types of distributions because, as mentioned, distribution types have an influence on the choice of statistical analysis that should be performed on them. It would be quite incorrect to do the famous *t*-test on data values for a sample that does not come from a variable with a normal distribution in the population from which the sample was taken.

Unfortunately, most data is not shared openly and we have to trust the integrity of the authors and that they chose an appropriate test for their data. The onus then also rests on you to be aware of the various distributions and what tests to perform when conducting your own research, as well as to scrutinize these choices when reading the literature.

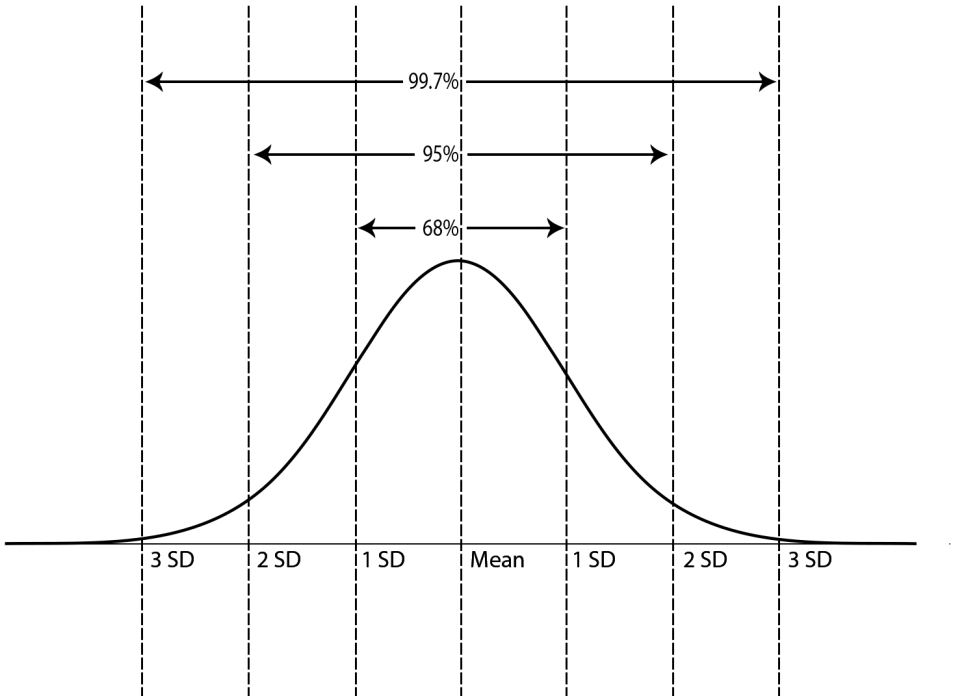
In this lesson you will note that I refer to two main types of distributions. First, there is the distribution pattern taken by the actual data point values in a study sample (or the distribution of that variable in the underlying population from which the sample was taken). Then there is the distribution that can be created from the data point values by way of the Central Limit Theorem. There are two of these distributions and they are the *Z*- and the *t-distributions* (both sharing a beautifully symmetric, bell-shaped pattern, allowing us to calculate a p-value from them).

Normal distribution

The normal distribution is perhaps the most important distribution. We need to know that data point values for a sample are taken from a population in which that variable is normally distributed before we decide on what type of statistical test to use. Furthermore, the distribution of all possible outcomes (through the Central Limit Theorem) is normally distributed.

The normal distribution has the following properties:

- most values are centered around the mean
- as you move away from the mean, there are fewer data points
- symmetrical in nature
- bell-shaped curve
- almost all data points (99.7%) occur within 3 standard deviations of the mean



Most variables that we use in clinical research have data point values that are normally distributed, i.e. the sample data points we have, come from a population for whom the values are normally distributed. As mentioned in the introduction, it is important to know this, because we have to know what distribution pattern the variable has in the population in order to decide on the correct statistical analysis tool to use.

It is worthwhile to repeat the fact that actual data point values for a variable (i.e. age, height, white cell count, etc.) have a distribution (both in a sample and in the population), but that through the Central Limit Theorem, where we calculate how often certain values or differences in values occur, we have a guaranteed normal distribution.

Sampling distribution

As was mentioned in the previous section, actual data point values for variables, be it from a sample or from a population, has a distribution. Then we have the distribution that is guaranteed through the Central Limit Theorem. Whether we are talking about means, the difference in means between two or more groups, or even the difference in medians, a plot of how many times each will occur given an almost unlimited repeat of a study will be normally distributed, allows us do inferential statistical analysis.

So, imagine, once again, being in a clinic and entering the ages of consecutive patients in a spreadsheet. Some ages will occur more or less commonly, allowing us to plot a histogram of the values. It would show the distribution of our actual values. These patients come from a much, much larger population. In this population, ages will also come in a certain distribution pattern, which may be different from our sample distribution.

In the video on the Central Limit Theorem we learn that our sample or the difference between two sample groups is but one of an enormous number of differences that could occur. This larger distribution will always be normally distributed according to the Central Limit Theorem and refers to another meaning of the term distribution, termed a *sampling distribution* (in other words, the type of distribution we get through the mathematics of the Central Limit Theorem is called a sampling distribution).

Z-distribution

The Z-distribution is one of the sampling distributions. In effect, it is a graph of a mathematical function. This function takes some values from parameters in the population (as opposed to only from our sample statistics) and constructs a symmetrical, bell-shaped curve from which we can do our statistical analysis.

It is not commonly used in medical statistics as it requires knowledge of some population parameters and in most cases, these are unknown. It is much more common to use the t-distribution, which only requires knowledge that is available from the data point values of a sample.

t-distribution

This is the most commonly used sampling distribution. It requires only calculations that can be made from the data point values for a variable that is known from the sample set data for a study.

The *t*-distribution is also symmetrical and bell-shaped and is a mathematical equation that follows from the Central Limit Theorem allowing us to do inferential statistical analysis.

One of the values that has to be calculated when using the t-distribution, is called *degrees of freedom*. It is quite an interesting concept with many interpretations and uses. In the context we use it here, it depends on the number of participants in a study and is easily calculated as the difference between the total number of participants and the total number of groups. So, if we have a total of 60 subjects in a study and we have them divided into two groups, we will have a degree of freedom equal to 58, being 60 minus two. The larger the degrees of freedom, the more the shape of the *t*-distribution resembles that of the normal distribution.

The effect of this is that all studies should aim to have as many participants as possible and the a larger sample set would allow for the sample to be more representative of the population, increasing the power of the study.

Week 4: The important first steps: Hypothesis testing and confidence levels

Hypothesis testing

This course represents an inferential view on statistics. In doing so we make comparisons and calculate how significant those differences are, or rather how likely it is to have found the specific difference when comparing results.

This form of statistical analysis makes use of what is termed hypothesis testing. Any research question, pertaining to comparisons, has two hypotheses. We call them the null and alternate (maintained, research, test) hypothesis and in light of comparisons, they are actually quite easy to understand.

The null hypothesis

It is of utmost importance that researchers remain unbiased towards their research and particularly the outcome of their research. It is natural to become excited about a new academic theory and most researchers would like to find significant results.

The proper scientific method, though, is to maintain a point of no departure. This means, that until data is collected and analyzed, we state a null hypothesis. This means that there will be no difference when a comparison is done through data collection and analyses.

Being all scientific about it, though, we do set a threshold for our testing and if the analyses finds that this threshold has been breached, we no longer have to accept the null hypothesis. In actual fact, we can now reject it.

This method, called the scientific method, forms the bedrock of evidence-based medicine.

As scientists we believe in the need for proof.

The alternative hypothesis

The alternative hypothesis states that there is a difference when a comparison is done. Almost all comparisons will show differences, but we are interested in the size of that difference, and more importantly in how likely it is to have found the difference that a specific study finds.

We have seen through the Central Limit Theorem that some differences will occur more commonly than others. We also know through the use of combinations, that there are many, many differences indeed.

The alternative hypothesis is accepted when an arbitrary threshold is passed. This threshold is a line in the sand and states that any difference found beyond this point is very unlikely to have occurred. This allows us to reject the null hypothesis and accept the alternate hypothesis, that is, that we have found a statistically significant difference or results.

In the next section we will see that there is more than one way to state the alternative hypothesis and it is of enormous importance

The alternative hypothesis

Two ways of stating the alternative hypothesis

As I have alluded to in the previous section, there are two ways to state the alternative hypothesis. It is exactly for the reason of different ways of stating the alternative hypothesis, that it is so important to state the null and alternative hypotheses prior to any data collection or analyses.

Let's consider comparing the cholesterol level of two groups of patients. Group one is on a new test drug and group two is taking a placebo. After many weeks of dutifully taking their medication, the cholesterol levels of each participant is taken.

Since the data point values for the variable total cholesterol level is continuous and ratio-type numerical, we can compare the means (or as we shall see later, the medians) between the two groups.

As good scientists and researchers, though, we stated our hypotheses way before this point! The null hypothesis would be easy. There will be no difference between the means (or medians) of these two groups.

What about the alternate hypothesis? It's clear that we can state this in two ways. It might be natural, especially if we have put a lot of money and effort into this new drug to state that the treatment group (group one) will have lower cholesterol than the placebo or control group (group two). We could also simply state that there will be a difference. Either of the two groups might have a lower mean (or median) cholesterol level.

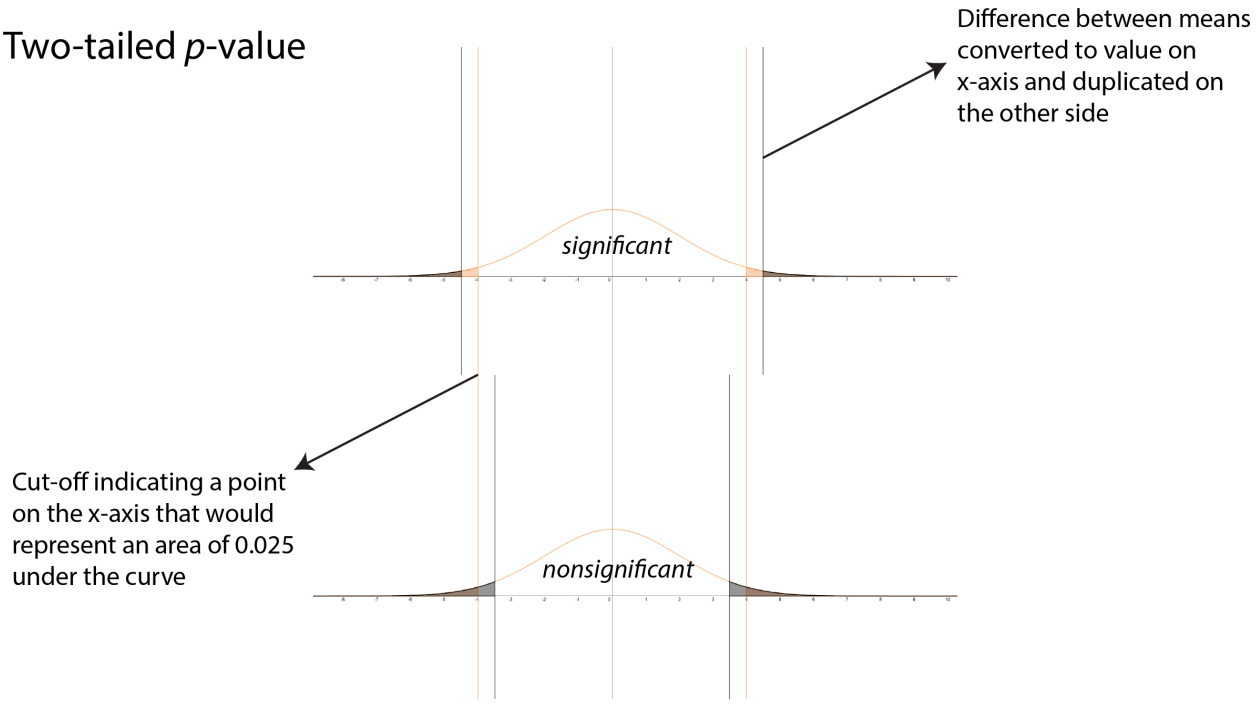
This choice has a fundamental effect on the calculated *p*-value. For the exact same data collection and analyses, one of these two ways of stating the alternative hypothesis has a *p*-value of half of the other. That is why it so important to state these hypotheses before commencing a study. A non-significant p-value of 0.08 can very easily be changed after the fact to a significant 0.04 by simply restating the alternate hypothesis.

The two-tailed test

When an alternative hypothesis states simply that there will be a difference, we conduct what is termed a two-tailed test.

Let's consider again our cholesterol example for before. Through the use of combinations and the Central Limit Theorem we will be able to construct a symmetrical bell-shaped *t*-distribution curve from our data point values. A computer program will construct this graph and will also know what cut-off values on the x-axis would represent an area under the curve to represent 0.05 (5%) of the total area.

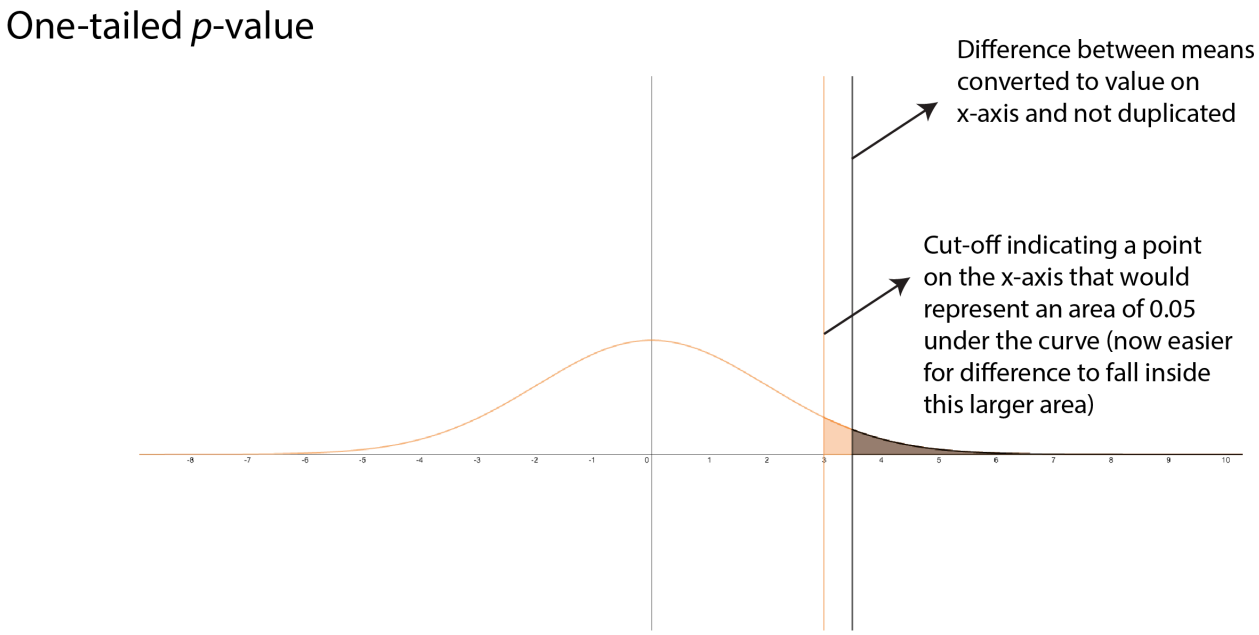
This being a two-tailed test, though, it will split this to both sides, with 0.025 (2.5%) on either side.



When comparing the means (or medians) of the two groups, one will be less than the other and depending which one you subtract from which, you will get either a positive or a negative answer. This is converted to a value (units of standard error) that can also be plotted on the x-axis. Since this is a two-tailed test, though, it is reflected on the other side. The computer can calculate the area under the curve for both sides, called a two-tailed *p*-value. From the graphs it is clear that this form of alternate hypothesis statement is less likely to yield a low *p*-value. The area is doubled and the cut-off marks representing the 0.025 (2.5%) level are further away from the mean.

The one-tailed test

A researcher must be very clear when using a one-tailed test. For the exact same study and data point values, the *p*-value for a one-tailed test would be half of that of a two-tailed test. The graph is only represented on one side (not duplicated on the other) and all of the area representing a value of 0.05 (5%) is on one side and therefore the cut-off is closer to the mean, making it easier for a difference to fall beyond it.



The choice between a one-tailed and two-tailed approach

There is no magic behind this choice. A researcher should make this choice by convincing her or his peers through logical arguments or prior investigations. Anyone who reads a research paper should be convinced that the choice was logical to make.

Hypothesis testing errors

All research starts with a question that needs an answer. Everyone might have their own opinion, but an investigator needs to look for the answer by designing an experiment and investigating the outcome.

A concern is that the investigator may introduce bias, even unintentionally. To avoid bias, most healthcare research follows a process involving hypothesis testing. The hypothesis is a clear statement of what is to be investigated and should be determined before research begins.

To begin, let's go over the important definitions I discussed in this lesson:

- The null hypothesis predicts that there will be no difference between the variables of investigation.
- The alternative hypothesis, also known as the test or research hypothesis predicts that there will be a difference or significant relationship.

There are two types of alternative hypotheses. One that predicts the direction of the hypothesis (e.g. A will be more than B or A will be less than B), known as a one-tailed test. The other states there will be a significant relationship but does not state in which way (e.g. A can be more or less than B). The latter is known as a two-tailed test.

For example, if we want to investigate the difference between the white blood cell count on admission between HIV-positive and HIV-negative patients, we could have the following hypotheses:

- Null hypothesis: there is no difference between the admission white blood cell count between HIV-positive and HIV-negative patients.
- Alternative hypothesis (one-tailed) : the admission white cell count of HIV-positive patients will be higher than the admission white blood cell count of HIV-negative patients.
- Alternative hypothesis (two-tailed) : the admission white blood cell count of HIV-positive patients will differ from that of HIV-negative patients.

Type I and II errors

Wrong conclusions can sometimes be made about our research findings. For example, a study finds that drug A has no effect on patients, when in fact it has major consequences. How might such mistakes be made, and how do we identify them? Since such experiments are about studying a sample and making inferences about a population (which we have not studied), these mistakes can conceivably occur.

Table I shows two possible types of errors that exist within hypothesis testing. A Type I error is one where we falsely reject the null hypothesis. For instance, a Type I error could occur when we conclude that there is a difference in the white blood cell count between HIV-positive and HIV-negative patients when in fact no such difference exists.

A Type II statistical error relates to failing to reject the null hypothesis, when in fact a difference was present. This is where we would conclude that there is no difference between the white blood cell count between HIV-positive and HIV-negative patients exists when in reality a difference exists.

	The null hypothesis is valid	The null hypothesis is invalid
Reject the null hypothesis	Type I error	✓
Fail to reject the null hypothesis	✓	Type II error

Why do we say “we fail to reject the null hypothesis” instead of “we accept the null hypothesis”? This is tricky, but just because we fail to the reject the null hypothesis, this does not mean we have enough evidence to prove it is true.

In the next lesson, I’ll cover the topic of confidence intervals.

Confidence in your results

Introduction to confidence intervals

Confidence intervals are very often quoted in the medical literature, yet it is often mentioned that it is a poorly understood statistical concept amongst healthcare personnel.

In this course we are concentrating on inferential statistics. We realize that most published literature is based on the concept of taking a sample of individuals from a population and analyzing some data point values we gather from them. These results are then used to infer some understanding about the population. We do this because we simply cannot investigate the whole population.

This method, though, is fraught with the danger of introducing bias. The sample that is selected for analysis might not properly represent the population. If a sample statistic such as a mean is calculated, it will differ from the population mean (if we were able to test the whole population). If a sample was taken with proper care, though, this sample mean should be fairly close to the population mean.

This is what confidence intervals are all about. We construct a range of values (lower and upper limit, or lower and upper maximum), which is symmetrically formed around the sample mean (in this example) and we infer that the population mean should fall between those values.

We can never be entirely sure about this, though, so we decide on how confident we want to be in the bounds that we set, allowing us to calculate these bounds.

Confidence intervals can be constructed around more than just means and in reality, they have a slightly more complex meaning than what I have laid out here. All will be revealed in this lesson, allowing you to fully understand what is meant by all those confidence intervals that you come across in the literature.

Reference:

1. Dinh T et al. Impact of Maternal HIV Seroconversion during Pregnancy on Early Mother to Child Transmission of HIV (MTCT) Measured at 4-8 Weeks Postpartum in South Africa 2011- 2012: A National Population-Based Evaluation PLoS ONE 10(5) DOI: 10.1371/journal.pone.0125525. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0125525

Confidence levels

What are they?

As I've mentioned in the previous section, we only examine a small sample of subject from a much larger population. The aim is though, to use the results of our studies when managing that population. Since we only investigate a small sample, any statistic that we calculate based on the data point gathered from them, will not necessarily reflect the population parameter, which is what we are really interested in.

Somehow, we should be able to take a stab at what that population parameter might be. Thinking on a very large scale, the true population parameter for any variable could be anything from negative infinity to positive infinity! That sounds odd, but makes mathematical sense. Let's take age for instance. Imagine I have the mean age of a sample of patients and am wondering about the true mean age in the population from which that sample was taken. Now, no one can be -1000 years old, neither can they be +1000 years old. Remember the Central Limit Theorem, though? It posited that the analysis of a variable from a sample was just one of many (by way of combinations). That distribution graph is a mathematical construct and stretches from negative to positive infinity. To be sure, the occurrences of these values are basically nil and in practice they are. So let's say (mathematically) that there are values of -1000 and +1000 (and for that matter negative and positive infinity) and I use these bounds as my guess as to what the mean value in the population as a whole is. With that wide a guess I can be 100% confident that this population mean would fall between those bounds.

This 100% represent my confidence level and I can set this arbitrarily for any sample statistic.

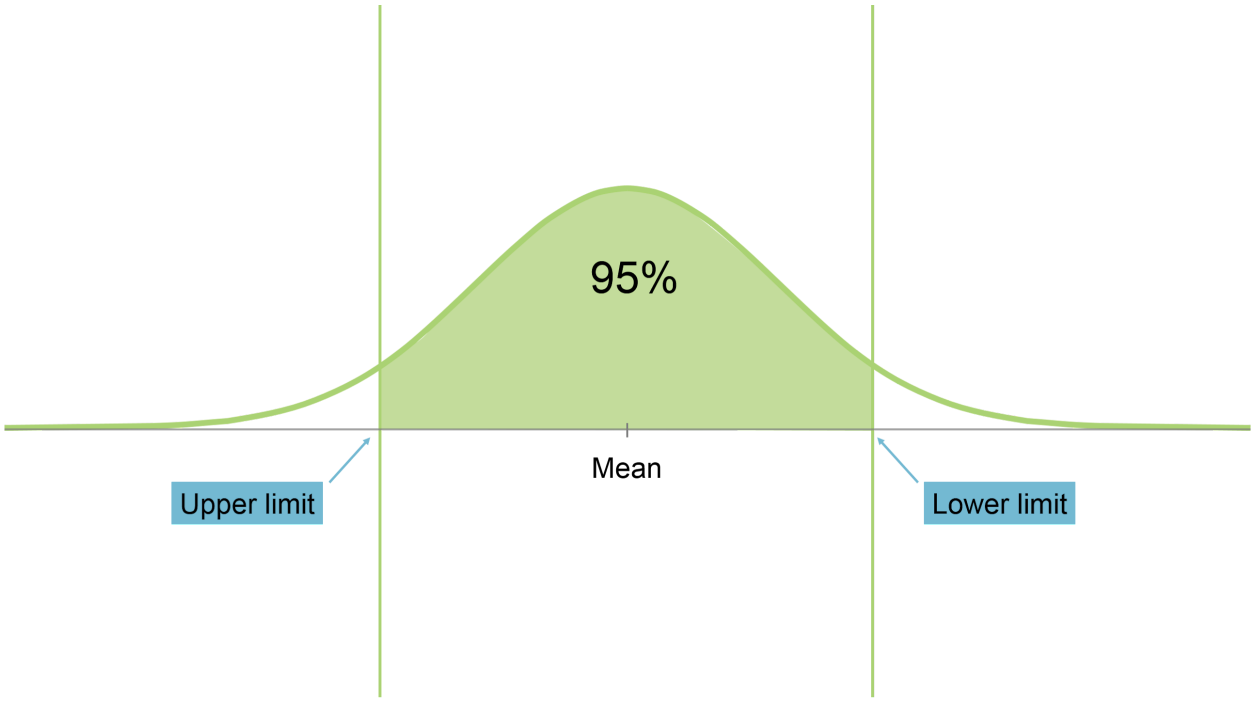
What happens if I shrink the bounds?

Enough with the non-sensical values. What happens if for argument's sake the sample mean age was 55 and I suggest that the population mean is 45 to 65. If have shrunk the bounds, but now, logically, I should lose some confidence in my guess now. Indeed, that is what happens. The confidence level goes down. If I shrink it to 54 to 56, there is a much greater chance that the population mean escapes these bounds and the confidence level would be much smaller.

A 95% confidence level

It is customary to use a confidence level of 95% and that is the value that you will notice most often in the literature when confidence intervals are quoted. The 95% refers to the confidence levels and the values represent the actual interval.

The mathematics behind confidence intervals constructs a distribution around the sample statistic based on the data point values and calculates what area would be covered by 95% (for a 95% confidence level) of the curve. The x-axis values are reconstituted to actual values which are then the lower and upper values of the interval.



Confidence intervals

Now that we understand what confidence levels are, we are ready to define the proper interpretation of confidence intervals.

It might be natural to suggest that given a 95% confidence level, that we are 95% confident that the true population parameter lies between the intervals given by that confidence level. Revisiting our last example we might have a sample statistic of 55 years for the mean of our sample and with a 95% confidence level construct intervals of 51 to 59 years. This would commonly be written as a mean age of 55 years (96% CI, 51-59). It would be incorrect though to suggest that there is a 95% chance of the population mean age being between 51 and 59 years.

The true interpretation of confidence intervals

Consider that both the sample statistics (mean age of the participants in a study) and the population parameter (mean age of the population from which the sample was taken) exist in reality. They are both absolutes. Given this, the population parameter either does or does not fall inside of the confidence interval. It is all or nothing.

The true meaning of the confidence level of say 95% is that if the study is repeated 100 times (each with its own random sample set of patients drawn from the population, each with its own mean and 95% confidence intervals), 95 of these studies will correctly have the population parameter correctly within the intervals and 5 would not. There is no way to know which one you have for any given study.

Week 5: Which test should you use?

Introduction to parametric tests

Finally in this course we get to grips with some real inferential statistics and we start things off with parametric tests. Inferential statistics is all about comparing different sample subject to each other. Most commonly we deal with numerical data point values, for which we can calculate measures of central tendency and dispersion.

When using parametric tests, we use the mean or average as our measure of central tendency. Commonly we will have two (or more) groups and for any given variable, say for instance white cell count, we could calculate a mean value for each group. A difference will exist between means of the groups and through the use of statistical tests we could calculate how common certain differences should occur given many repetitions of a study and also how likely it was then to have found a difference at least as wide as the one for the particular at hand.

Parametric tests are are the most commonly used tests, but with this common use come some very strict rules or assumptions. If these are not met and parametric tests are used, the subsequent p-values might not be a true reflection what we can expect in the population.

Types of parametric tests

I will discuss three main types of parametric tests in this lesson.

t-tests

These are truly the most commonly used and most people are familiar with Student's t-test. There is more than one *t*-test depending on various factors.

As a group, though, they are used to compare the point estimate for some numerical variable between two groups.

ANOVA

ANOVA is the acronym for analysis of variance. As opposed to t-test, ANOVA can compare a point estimate for a numerical variable between more than two groups.

Linear regression

When comparing two or more groups, we have the fact that although the data point values for a variable are numerical in type, the two groups themselves are not. We might call the groups A and B, or one and two, or test and control. As such they refer to categorical data types.

In linear regression we directly compare a numerical value to a numerical value. For this, we need pairs of values and in essence we look for a correlation between these. Can we find that a change in the set that is represented by the first value in all the pairs causes a predictable change in the set made up by all the second values in the pair.

As an example we might correlate the number of cigarettes smoked per day to blood pressure level. To do this we would need a sample of participants and for each have a value for cigarettes smoked per day and blood pressure value. As we will see later, correlation does not prove causation!

Student’s t-test :Introduction

We have learned that this is at least one of the most common statistical tests. It takes numerical data point values for some variables in two groups of subjects in a study and compares them to each other.

William Gosset (developer of Student's t-test) used the *t*-distribution, which we covered earlier. It is a bell-shaped, symmetrical curve and represents a distribution of all the differences (in the mean) between two groups should the same study be repeated multiple times. Some will occur very often and some will not. The *t*-distribution uses the concept of degrees of freedom. This value refers to the total number of participants in a study (both groups) and minus the number of groups (which is two). The higher the degrees of freedom, the more accurately the *t*-distribution follows the normal distribution and the mathematics behind this posits in some way, a more accurate *p*-value. This is another reason to include as large a sample size as possible.

Once the graph for this distribution is constructed, it becomes possible to calculate where on the x-axis the cut-off representing a desired area would be. The actual difference is also converted to the same units (called standard errors) and the area for this calculated as we have seen before.

Since we are trying to mimic the normal distribution, we have one of the most crucial assumptions for the use of the *t*-test and other parametric tests. We need assurances that the sample of subjects in a study was taken from a population in which the variable that is being tested is normally distributed. If not, we cannot use parametric tests.

Types of *t*-tests

There are a variety of *t*-tests. Commonly we will have two independent groups. If we were to compare the average cholesterol levels between two groups of patients, participants in these two groups must be independent of each other, i.e. we cannot have the same individual appear in both groups. A special type of *t*-test exists if the two groups do not contain independent individuals as would happen if the groups are made up of homozygotic (identical) twins and we test the same

variable in the same group of participants before and after an intervention (with the two sets of data constituting the two groups).

There is also two variations of the *t*-test based on equal and unequal variances. It is important to consider the difference in the variances (square of the standard deviation) for the data point values for the two groups. If there is a big difference a *t*-test assuming unequal variances should be used.

ANOVA

ANOVA is the acronym for analysis of variance. As opposed to the *t*-test, ANOVA can compare the means of more than two groups. There are a number of different ANOVA tests, which can deal with more than one factor.

The most common type of ANOVA test is one-way ANOVA. Here we simply use a single factor (variable) and compare more than two groups to each other.

ANOVA looks at both variations of values inside of groups and between groups and constructs a distribution based on the same criteria that we have based on the Central Limit Theorem and combinations.

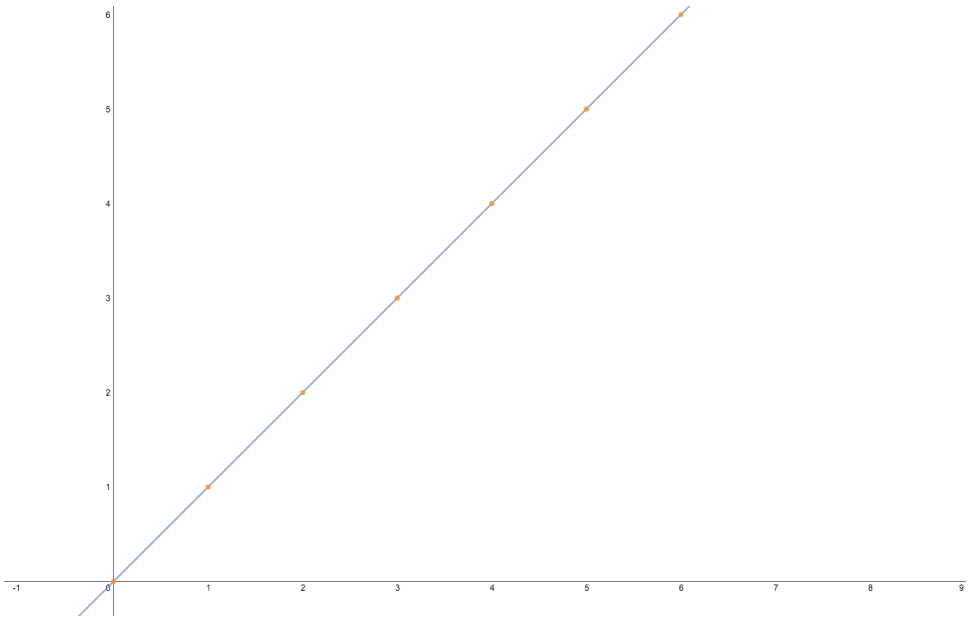
When comparing more than two groups, it is essential to start with analysis of variance. Only when a significant value is calculated should a researcher continue with comparing two groups directly using a *t*-test, so as to look for significant differences. If the analysis of variance does not return a significant result, it is pointless to run *t*-tests between groups and if done any significant finding should be ignored.

Linear Regression

Up until now we have been comparing numerical values between two categorical groups. We have had examples of comparing white cell count values between groups A and B, cholesterol values between groups taking a new test drug and a placebo. The variable cholesterol and white cell count contain data point that are ratio-type numerical and continuous, but the groups themselves are categorical (group A and B or test and control).

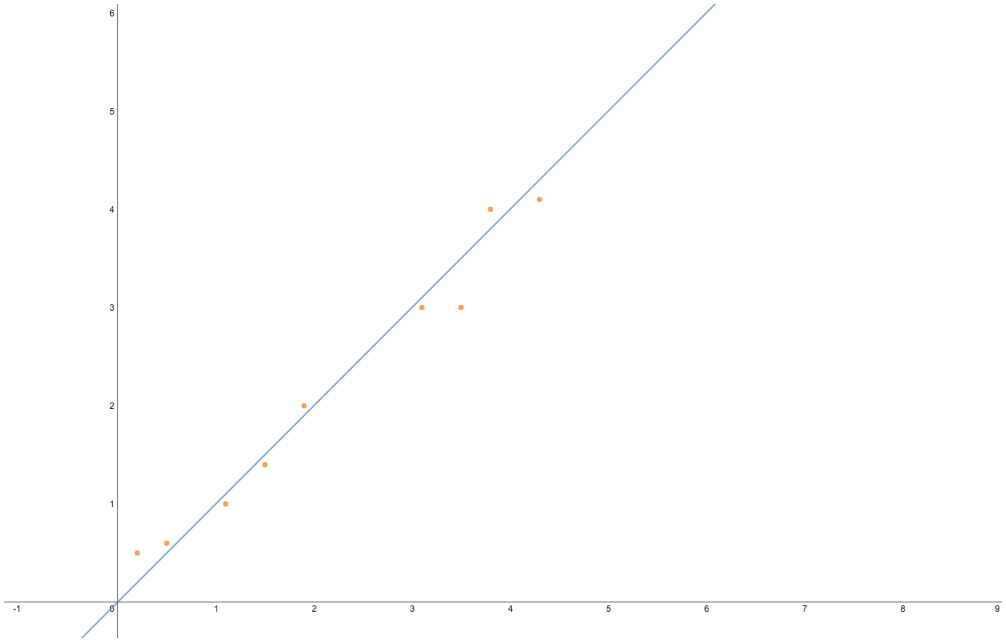
We can compare numerical values directly to each other through the use of linear regression.

In linear regression we are looking for a correlation between two sets of values. These sets must come in pairs. The graph below show a familiar plot of the mathematical equation $y = x$. We can plot sets of values on this line, i.e. (0,0), (1,1), (2,2), etc. This is how linear regression is done. The first value is all the pair come from one set of data point values and the second from a second set of data point values. I've mentioned an example before, looking at the number of cigarettes smoked per day versus blood pressure value.

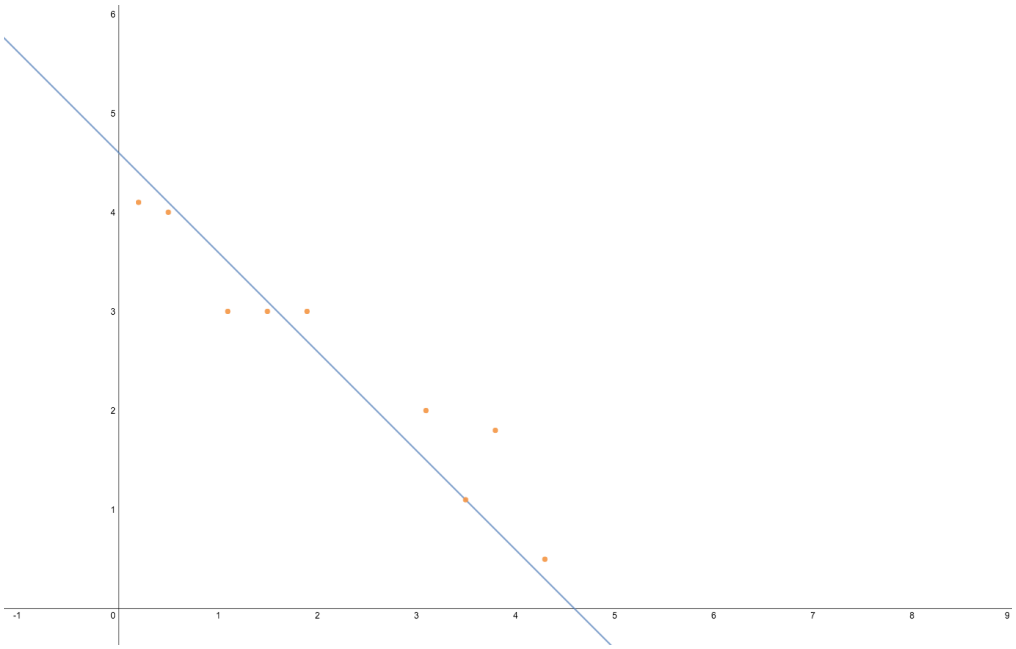


Linear regression looks at the correlation between these two sets of values. Does one depend on the other. Is there a change in the one, as the other changes.

Sets of data points will almost never fall on a straight line, but the mathematics underlying linear regression can try and make a straight line out of all the data point sets. When we do this we note that we usually get a *direction*. Most sets are either positively or negatively correlated. With positive correlation, one variable (called the dependent variable, which is on the y-axis), increases as the other (called the independent variable, which is on the x-axis) also increases.



There is also a negative correlation and as you might imagine, the dependent variables decreases as the independent variable increases.



Strength of the correlation

As you would have noticed, some of the data point pairs are quite a distance away from the linear regression line. With statistical analysis we can calculate how strongly the pairs of values are correlated and express that strength as a correlation coefficient, *r*. This correlation coefficient ranges from -1 to +1, with negative one being absolute negative correlation. This means that all the dots would fall on the line and there is perfect movement in the one variable as the other moves. With a positive one correlation we have the opposite. In most real-life situations, the correlation coefficient will fall somewhere in between.

There is also the zero value, which means, no correlation at all.

Correlating variables are always fascinating, but comes with a big warning. Any correlation between variables does not necessarily mean causation. Just because two variables are correlated does not mean the change in one is caused by a change in the other. There might be a third factor influencing both. Proof of a causal relationship requires much more than linear regression.

Nonparametric testing for your non-normal data

Nonparametric tests

When comparing two or more groups of numerical data values, we have a host of statistical tools on hand. We have seen all the t-tests for use when comparing two groups and well as ANOVA for comparing more than two groups. As mentioned in the previous lecture, though, the use of these tests requires that some assumptions are met.

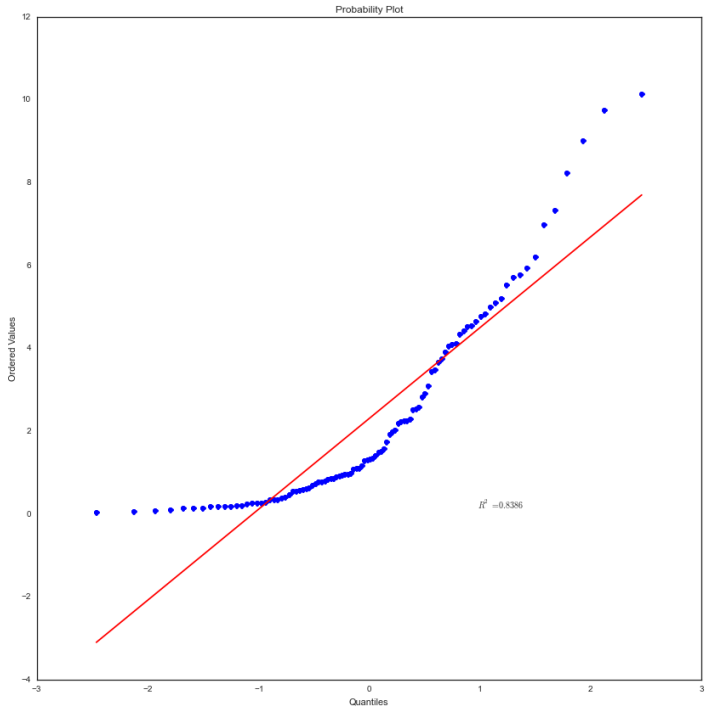
Chief among these was the fact that the data point values for a specific variable must come from a population in which that same variable is normally distributed. We are talking about a population and therefore a parameter, hence the term *parametric* tests.

Unfortunately we do not have access to the data point values for the whole population. Fortunately there are statistical tests to measure the likelihood that the data point values in the underlying population are normally distributed and they are done based on the data point values that are available from the sample of participants.

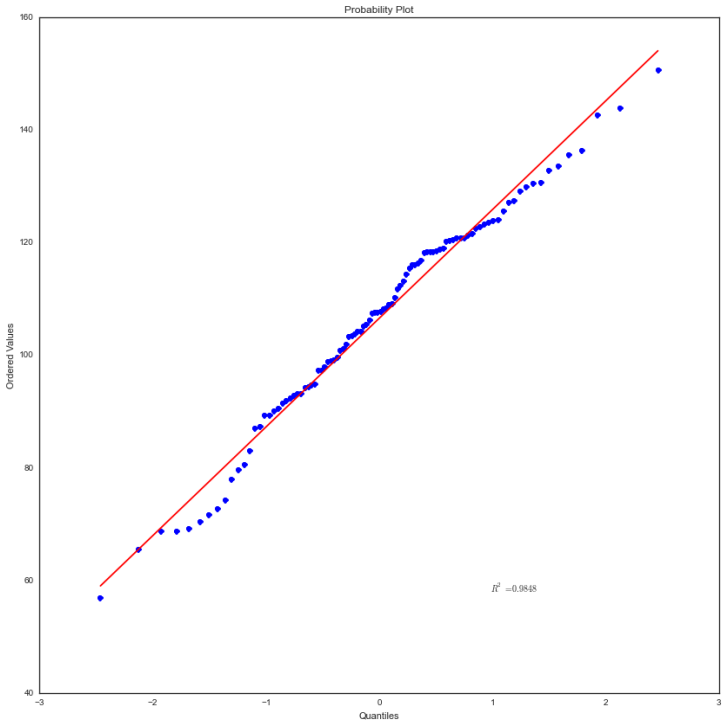
Checking for normality

There are a variety of ways to check whether parametric tests should be used. I'll mention two here. The first is quite visual and makes use of what is called a QQ plot. In a QQ plot, all the data point values from a sample group for a given variable are sorted in ascending order and each is assigned its percentile rank value, called its quantile. This refers to the percentage of values in the set that falls below that specific value. This is plotted against the quantiles of values from a distribution that we need to test against. In deciding if a parametric test should be used, this would be the normal distribution.

In the first image below we have a computer generated plot showing data point values that do not follow the hypothetical straight (red) line if the data point values for the sample were taken from a population in which that variable was normally distributed.



In the image below, we note the opposite and would all agree that these points are a much closer match for the red line normal distribution.



If you look closely at these graphs, you will also note an *R*-squared value. That is the square of the correlation coefficient,*r*, which we met in the lesson on linear regression. Note the value of 0.99 (very close to 1) for the second set of values, versus only 0.84 for the first.

The second method that I will mention here is the *Kolmogorov-Smirnov test*. It is in itself a form of a non-parametric test and can compare a set of data point values against a reference probability distribution, most notably, the normal distribution. As with all statistical tests of comparison it calculates a *p*-value and under the null hypothesis, the sample data point values are drawn from the same distribution against which it is tested. The alternative hypothesis would state that it is not and would demand the use of a non-parametric test for comparing sets of data point values.

Non-parametric tests to the rescue

Since we are going to compare numerical values to each other and in most cases make use of the t-distribution of sample means (trying to mimic the normal distribution) it is clear to see that when the data point values seem not to come from a population in which those data point values are normally distributed would lead to creating incorrect areas under the curve (p-value). In these cases, it is much better to use the set of non-parametric tests.

Nonparametric tests

The most common tests used in the literature to compare numerical data point values are *t*-tests, analysis of variance, and linear regression.

These tests can be very accurate when used appropriately, but do suffer from the fact that fairly stringent assumptions must be made for their use. Not all researchers comment on whether these assumptions were met before choosing to use these tests.

Fortunately, there are alternative tests for when the assumptions fail and these are called nonparametric tests. They go by names such as the Mann-Whitney-U test and Wilcoxon sign-rank test.

In this lesson we will take a look at when the assumptions for parametric tests are not met so that you can spot them in the literature and we will build an intuitive understanding of the basis for these tests. They deserve a lot more attention and use in healthcare research.

Key concepts

- Common parametric tests for the analysis of numerical data type values include the various *t-tests, analysis of variance,* and *linear regression*
- The most important assumption that must be met for their appropriate use is that the sample data point must be shown to come from a population in which the parameter is normally distributed

- The term parametric stems from the word parameter, which should give you a clue as to the underlying population parameter
- Testing whether the sample data point are from a population in which the parameter is normally distributed can be done by checking for skewness in the sample data or by the use of quantile (distribution) plots, amongst others
- When this assumption is not met, it is not appropriate to use parametric tests
- The inappropriate use of parametric analyses may lead to false conclusions
- Nonparametric tests are slightly less sensitive at picking up differences between groups
- Nonparametric tests can be used for numerical data types as well as ordinal categorical data types
- When data points are not from a normal distribution the mean (on which parametric tests are based) are not good point estimates. In these cases it is better to consider the median
- Comparing medians makes use of *signs*, *ranks*, *sign ranks* and *rank sums*
- When using *signs* all of the sample data points are grouped together and each value is assigned a label of either zero or (plus) one based on whether they are at or lower than a suspected median (zero) or higher than that median (one)
- The closer the suspected value is to the true median, the closer the sum of the signs should be to one half of the size of the sample
- A distribution can be created from a *rank* where all the sample values are placed in ascending order and ranked from one, with ties resolved by giving them an average rank value
- The sign value is multiplied by the (absolute value) rank number to give the *sign rank* value
- In the *rank sums* method specific tables can be used to compare values in groups and making lists of which values 'beat' which values with the specific outcome one of many possible outcomes
- The *Mann-Whitney-U* (or Mann-Whitney-Wilcoxon or Wilcoxon-Mann-Whitney or Wilcoxon rank-sum) test uses the rank-sum distribution
- The Kruskal-Wallis test is the nonparametric equivalent to the one-way analysis of variance test
- If a Kruskal-Wallis finds a significant p-value then the individual groups can be compared using the Mann-Whitney-U test
- The Wilcoxon sign-rank test is analogous to the parametric paired-sample t-test
- Spearman's rank correlation is analogous to linear regression
- The alternative Kendall's rank test can be used for more accuracy when the Spearman's rank correlation rejects the null hypothesis

References:

1. Parazzi P, et al. Ventilatory abnormalities in patients with cystic fibrosis undergoing the submaximal treadmill exercise test, Biomed Central Pulmonary Medicine, 2015, 15:63, http://www.biomedcentral.com/1471-2466/15/63
2. Bello, A, et al. Knowledge of pregnant women about birth defects, Biomed Central Pregnancy Childbirth , 2013, 13:45,http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3598521/

3. Craig E, et al. Risk factors for overweight and overfatness in rural South African children and adolescents, Journal of Public Health, 2015 http://jpubhealth.oxfordjournals.org/content/early/2015/03/04/pubmed.fdv016.full
4. Paul R, et al. Study of platelet aggregation in acute coronary syndrome with special reference to metabolic syndrome, International Journal of Applied Medical Research 2013 Jul-Dec; 3(2): 117–121, http://www.ijabmr.org/article.asp?issn=2229-516X;year=2013;volume=3;issue=2;spage=117;epage=121;aulast=Paul

Week 6: Categorical data and analyzing accuracy of results

Comparing categorical data

In the previous sections we looked at comparing numerical data types. What about methods to analyze categorical data, though? The most often used statistical test for categorical data types, both nominal and ordinal, is the chi-square test.

In this section we will use the *chi-squared* (χ^2) *distribution* to perform a *goodness-of-fit test* and have a look at how to test sample data points for independence.

The chi-squared goodness-of-fit test

Before we get to the more commonly used chi-squared test for independence, let's start off with the goodness-of-fit test. This test allows us to see whether the distribution pattern of our data fits a predicted distribution. This is called a *goodness-of-fit test*. In essence we will predict a distribution of values, go out and measure some actual data and see how well our prediction faired.

Since we are dealing with a frequency distribution, we have to count how many times a data value occurs and divide it by the sum-total of values. Let's consider predicting the five most common emergency surgical procedures over the next five months. The prediction estimates that appendectomies will be most common, making up 40% of the total, cholecystectomies making up 30% of the total, incision and drainage (I&D) of abscesses making up 20% of the total and with 5% each we predict repairing perforated peptic ulcers and major lower limb amputations. Over the next five months we actually note the following: 290 appendectomies, 256 cholecystectomies, 146 I&D procedures, 64 perforated peptic ulcer repairs and 44 amputations.

A chi-square goodness-of-fit test allows us to if the observed frequency distribution fits our prediction. Our null hypothesis would be that the actual frequency distribution can be described by the expected distribution and the test hypothesis states that these would differ.

The actual values are already available. We do need to calculate the expected (predicted) values, though. Fortunately we know the total, which in our example above was 800 (adding all the actual values) and we can construct values based on the predicted percentages above. This will leave it with 40% of 800, which is 320. Compare this to the actual value of 290. For the 30%, 20% and two 5% values we get 240, 160, 40 and 40, which compares to the observed values of 256, 146, 64 and 44.

The chi-square value is calculated from the differences between the observed and expected values and from this can calculate a probability of having found this difference (a *p*-value). If it is less than

our chosen value of significance we can reject the null hypothesis and accept the test hypothesis. If not, we cannot reject the null hypothesis.

The chi-squared test for independence

This is also called the χ^2-test for association (when considering treatment and condition) and is the common form that we see in clinical literature. It is performed by constructing so called contingency tables.

The null hypothesis states that there is no association between treatment and condition, with the alternative hypothesis stating that there is an association. Below is a table contingency table, clearly showing the categorical (ordinal) nature of the data.

Assessment (categorical)	Treatment	Placebo	Totals
Considerable improvement	27	5	32
Moderate improvement	11	12	23
No change	3	2	5
Moderate deterioration	4	13	17
Considerable deterioration	5	7	12
Death	4	14	18
Totals	54	53	107

This table represents the observed totals from a hypothetical study and simply counts the number of occurrences of each outcome, i.e. 27 patients in the treatment group were assessed as having improved considerably, whereas only five did so in the placebo group. Note how totals occur for both the rows and columns of the data. Thirty-two patients in total (both treatment and placebo groups) showed considerable improvement. There were 54 patients in the treatment group and 53 in the placebo group.

From this table an expected table can be calculated. Mathematical analysis of these two tables results in a χ^2-value, which is converted to a p-value. For a p-value less than a chosen value of significance we can reject the null hypothesis, thereby accepting the alternate hypothesis, that there

is an association between treatment and outcome. When viewing the observed table above, that would mean that there is a difference in proportions between the treatment and placebo columns. Stated differently, which (treatment) group a patient is in, does affect the outcome (there is independence).

The calculation for the p-value using the chi-squared test makes use of the concept of degrees of freedom. This is a simple calculation and multiplies two values. The first one subtract 1 from the number of columns and the second subtracts 1 from the number of rows. In our example above, we have two columns and six rows. Subtracting 1 from each yields 1 and 5. Multiplying them yields a value of five for the degrees of freedom.

Fisher's exact test

There are cases in which the χ^2-test does become inaccurate. This happens when the numbers are quite small, with totals in the order of five or less. There is actually a rule called Cochran's rule which states that more than 80% of the values in the expected table (above) must be larger than five. If not, Fisher's exact test should be used. Fisher's test, though, only considers two columns and two rows. So in order to use it, the categorical numbers above must be reduced by combining some of the categories. In the example above we might combine considerable improvement, moderate improvement and no change into a single row and all the deteriorations and death in a second row, leaving us with a two column and two row consistency table (observed table).

The calculation for Fisher's test uses factorials. Five factorial (written as 5!) means 5 x 4 x 3 x 2 x 1 = 120 and 3! is 3 x 2 x 1 = 6. For interest's sake 1! = 1 and 0! is also equal to 1. As you might realise, factorial values increase in size quite considerably. In the example above we had a value of 27 and 27! is a value with 29 numbers. That is billions and billions. When such large values are used, a researcher must make sure that his or her computer can accurately manage such large numbers and not make rounding mistakes. Fisher's exact test should not be used when not required due to small sample sizes.

Sensitivity, specificity, and predictive values

Considering medical investigations

The terms *sensitivity* and *specificity*, as well as *positive* and *negative predictive values* are used quite often in the medical literature and it is very important in the day-to-day management of patients that healthcare workers are familiar with these terms.

These four terms are used when we consider medical investigations and tests. They look at the problem from two points of view. In the case of sensitivity and specificity we consider how many

patients will be correctly indicated as suffering from a disease or not suffering from a disease. From this vantage point, no test has been ordered and we do not have the results.

This is in contrast to the use of positive and negative predictive value. From this point of view we already have the test results and need to know how to interpret a positive and a negative finding.

Sensitivity and specificity help us to decide which test to use and the predictive values help us to decide how to interpret the results once we have them.

Sensitivity and specificity

Let's consider the choice of a medical test or investigation. We need to be aware of the fact that tests are not completely accurate. False positive and negative results do occur. With a false positive result the patient really does not have the disease, yet, the results return positive. In contrast to this we have the false negative result. Although the test returns negative, the patient really does have the disease. As a good example we often note headlines scrutinising the use of screening tests such as mammography, where false positive tests lead to both unnecessary psychological stresses and further investigation, even surgery.

Sensitivity refers to how often a test returns positive when patients really have the disease. It requires some gold-standard by which we absolutely know that the patient has the disease. So imagine we have a set of 100 patients with a known disease. If we subject all of them to a new (or different) test, the sensitivity will be the percentage of times that this test returns a positive result. If 86 of these tests come back as positive, we have a 86% true positive rate, stated as a sensitivity of 86%. That means that in 14% of cases of using this test, we will get a false negative and might miss the fact that a patient might have the disease that we were investigating.

Specificity refers to how often a test returns a negative results in the absence of a disease. Once again, it requires the presence of some gold-standard, whereby we absolutely know that a disease is absent. Let's use a hundred patients again, all known not to have a certain disease. If we subject them to a test and 86 of those tests comes back as negative, we have a specificity of 86%. This would also mean that 14% of patients will have a false positive result and might be subjected to unnecessary further investigations and even interventions.

The great majority of medical tests and investigations cannot absolutely discriminate between those with and without the presence of a disease and we have to be circumspect when choosing any test.

Here we have an example:

	With disease	Without disease	
Test positive	90 (True positive)	90 (False positive)	180
Test negative	10 (False negative)	810 (True negative)	820
	100	900	1000

We note 1000 patients, 100 with a disease and 900 without. They are all subjected to a new test and 180 return a positive result and 820 a negative result. You will note the indication of false positives and negatives.

The equations for sensitivity and specificity are shown below.

$$\text{Sensitivity} = \frac{\text{True positive (TP)}}{\text{With disease}} = \frac{90}{100} = 90\%$$

$$\text{Specificity} = \frac{\text{True negative (TN)}}{\text{Without disease}} = \frac{810}{900} = 90\%$$

This gives us a sensitivity of (90/100) 90% and a specificity of (810/900) 90%.

Predictive values

Let's turn the tables and now consider how to interpret test results once they return. Again we could imagine that some tests will return both false positive and false negative results.

We express positive predictive values as the percentage of patients with a positive test result that turns out to have the disease and we express negative predictive value as the percentage of

patients with a negative test result that turns out not to have the disease. The figure below gives the simple formulae for predictive values.

$$\text{Positive predictive value (PPV)}$$

$$= \frac{\text{True positive}}{(\text{True positive}) + (\text{False positive})} = \frac{\text{True positive}}{\text{Test positive}} = \frac{90}{180}$$

$$= 50\%$$

$$\text{Negative predictive value (NPV)}$$

$$= \frac{\text{True negative}}{(\text{True negative}) + (\text{False negative})} = \frac{\text{True negative}}{\text{Test negative}} = \frac{810}{820}$$

$$= 99\%$$

The gives us a positive predictive value of (90/180) of 50% (only half of patients with a positive result will actually have the disease) and a negative predictive value of (810/820) of 99%, which means that only almost all patient with a negative result will actually not have the disease.

Predictive values are very dependent on the prevalence of a disease. Here we chose a sample set of 1000 patients in which the disease existed in only 10%. It is this low prevalence which gives us the poor positive predictive value. When interpreting positive and negative predictive values, you must always compare the prevalence of the disease in the study sample versus the prevalence of the disease in the patient population that you see. (There are mathematical methods of converting results to different levels of prevalence.)

Reference:

1. Sutton PA, Humes DJ, Purcell G, et al. The Role of Routine Assays of Serum Amylase and Lipase for the Diagnosis of Acute Abdominal Pain. Annals of The Royal College of Surgeons of England. 2009;91(5):381-384. doi:10.1308/003588409X392135. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2758431/