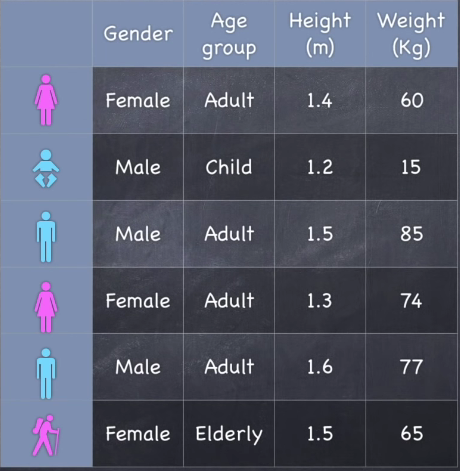
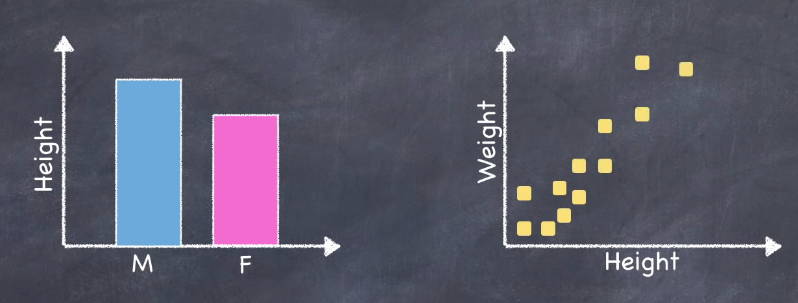
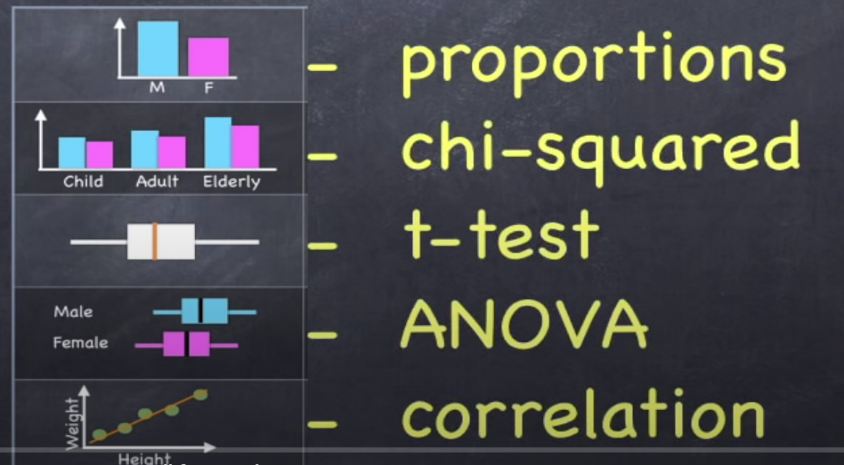
**! ! !...Statistics made easy…! ! !   
Learn about the t-test, the chi square test, the p value and more**

<https://youtu.be/I10q6fjPxJ0>

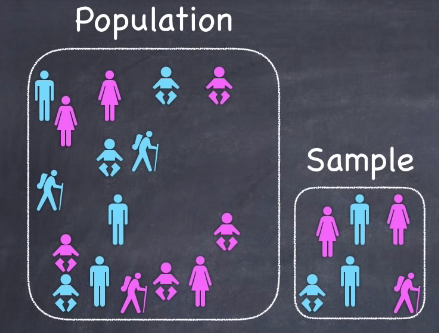
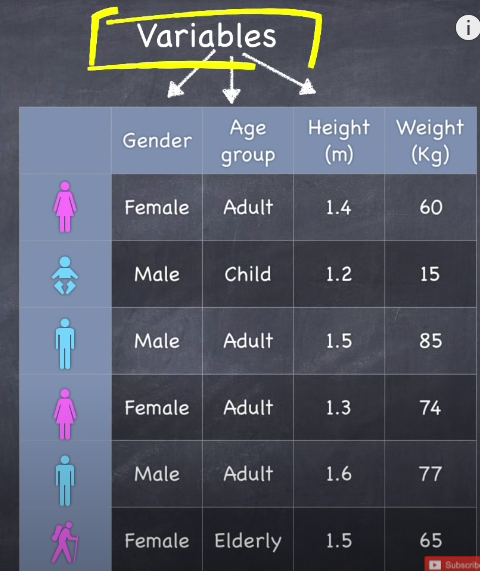
When we look at sample data for the most part, we see two things:

* We see differences between groups, so men are taller than women.
* We see relationships between variables like taller people weigh more than shorter people.

The big question is that are those differences / associations / relationships real?

We're going to take a look at a very simple data set and we're going to see how by looking at various combinations of variables and variable types we can identify very specific differences between groups and very specific relationships between variables and also, we will see when and how to use statistical tests and how to interpret the results.

Let's imagine that we have a research question and it's about the height and the weight of people living in Ireland. Of course, we can't measure the height in the weight of the entire population so instead we take a random sample of the population and we measure the weight and the height of that sample and we took some additional information like gender and age group from each of the people in our sample and we arranged these data in a spreadsheet or data set with the various attributes in columns and these are called **variables**.

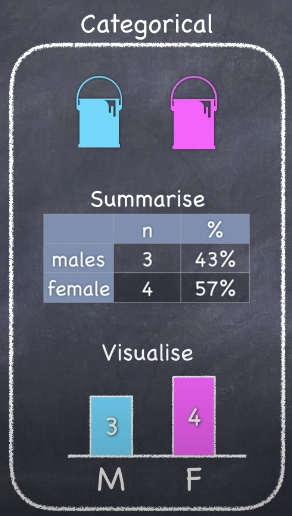
 

Most data sets will contain two types of variables categorical and numerical.

* **Categorical variables** like gender contains categories. As the name suggests think of them as groups or buckets that the data can be arranged into. In this case males and females.
  + **Nominal:** When there is not any order in the data, e.g., Classes-Red/White/Blue.
  + **Ordinal:** With order, e.g., Rank-1/2/3/4/5
* **Numeric variables** like height are numbers. As the name suggests and can be arranged on a number line.
  + **Discrete:** E.g., Count-1,2,3,4,5…
  + **Continuous:** E.g., Lelgth-1.5m, 20m, 55.5m

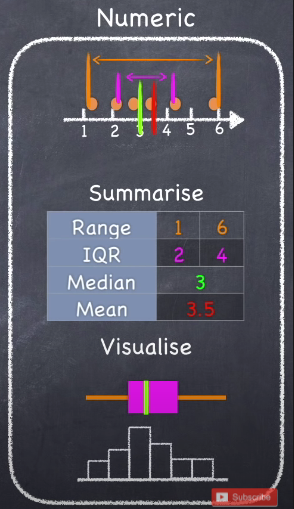


**To better understand our data and to make sense of it we summarize and visualize it.**

To summarize and visualize **categorical data**, we can:

* Count up the number of observations in any given category.
* Can represent them in a table or on a bar chart.

To summarize and visualize **numerical data**, we can:

* Check spread or distribution of the data:
  + Range
  + Interquartile Range
  + Standard Deviation
* To get a sense of the middle of the data:
  + Mean: It is the **average** and is the most commonly used summary value to represent numerical data.
  + Median: Divides the data into two equal halves.
* Box Plot: Visual representation of the interquartile range and the median of the data.
* Histogram: Gives us the shape of the data

We can see that this process of summarizing and visualizing the data takes it from being just numbers and words on a spreadsheet and turns it into something that is:

* meaningful to us.
* something that we can get our heads around.
* something that we can think about.

Now in this very simple data set we've got two categorical and two numeric variables and things start to get interesting when we start looking at combinations of variables.

So, for example when we take a look at a categorical and a numeric variable like gender and height:

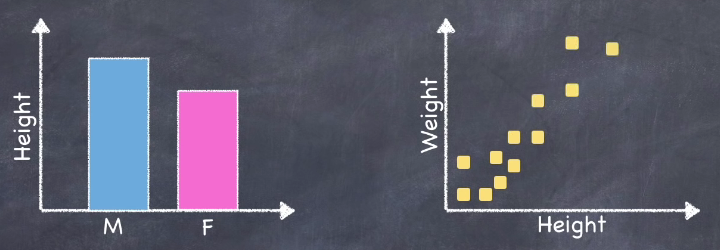
* We can group the data by gender which is the categorical variable.
* And create a summary of the numeric variable in this case.



Looking at the summary we can see that in our sample data men are on average taller than women.

There are other possible combinations of variables that we could have looked at we could have looked at like:

* Height and Weight which are both numeric.
* Gender and Age group both categorical, etc.

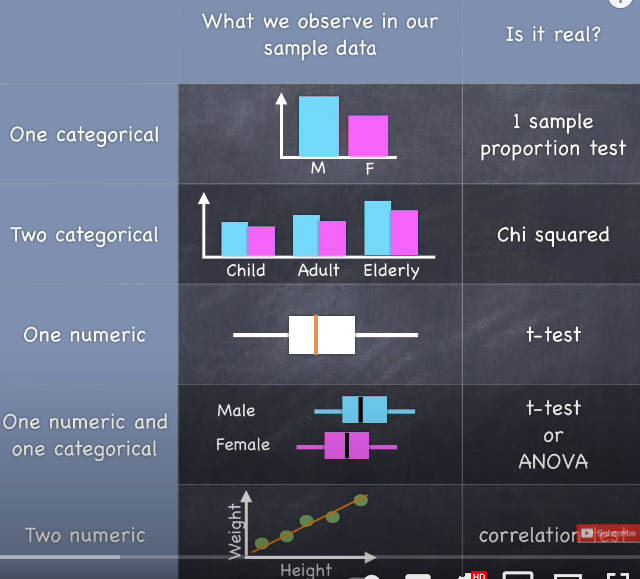


In each case we might see either differences between groups or relationships between variables.

And in each of these cases there are specific statistical tests that we can apply to see if what we are seeing in the sample data has implications for what we think about the wider population, can we infer anything, is what we are seeing statistically significant?

Let's take a quick look at the five most important combinations of data that we have and we'll look at:

* Firstly, what might we observe in our sample data given that sort of combination of data types.
* Secondly what statistical test we might apply to determine whether or not we can infer anything about the wider population.



Various data type combinations and tests:

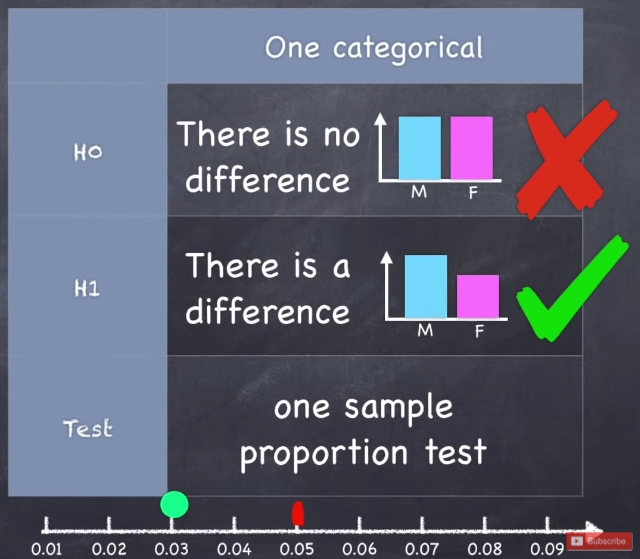
* Single categorical variable - **one sample proportion test**.
* Two categorical variables - **chi-square test.**
* Single numerical variable - **t-test.**
* A categorical and a numeric variable - **t-test -** or **analysis of variance** or **ANOVA** if we have more than 2 categories in a categorical variable**.**
* Two numeric variables - **correlation test.**

It's not good science to take a data set and just randomly stab around blindly hoping to find something that's statistically significant.

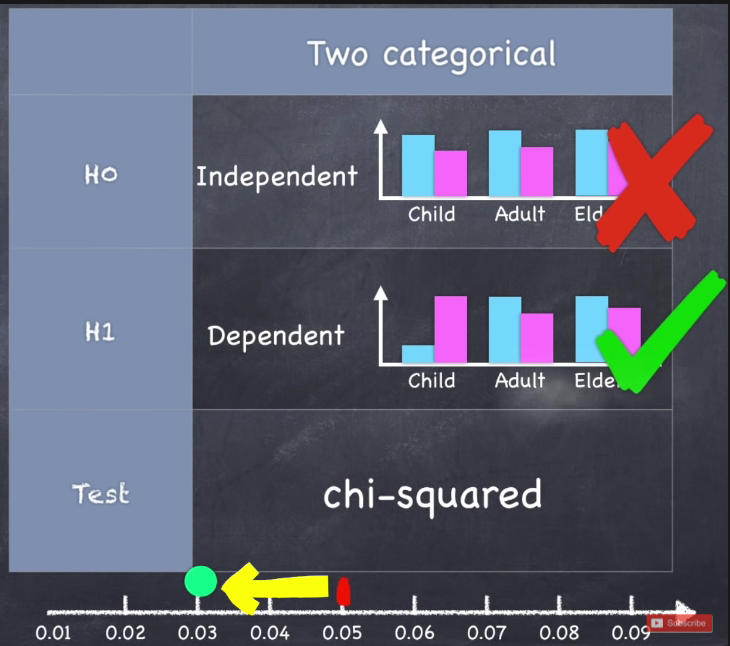
Before you interrogate the data, you start off by:

* Defining question/hypothesis.
* Define null hypothesis.
* Identify the alpha value that we’re going to use
* And then we analyse the data.

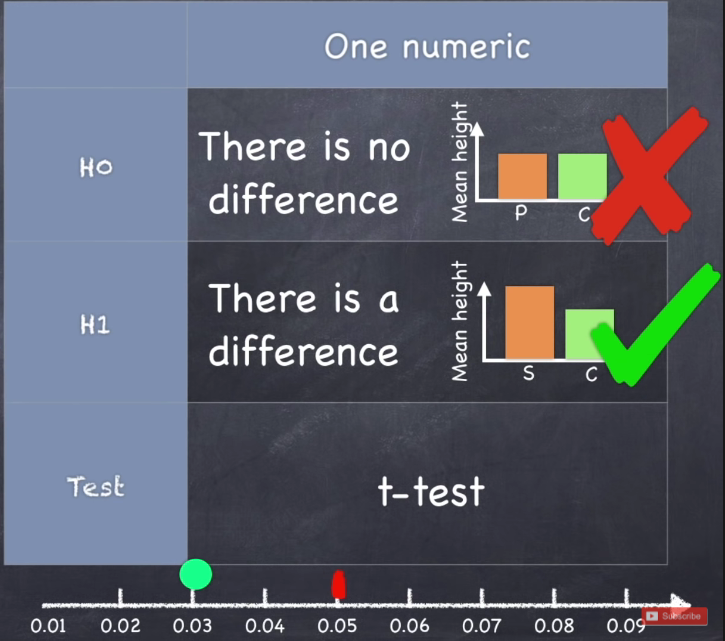
Let's look at what we can do with just **one categorical variable** like Gender.

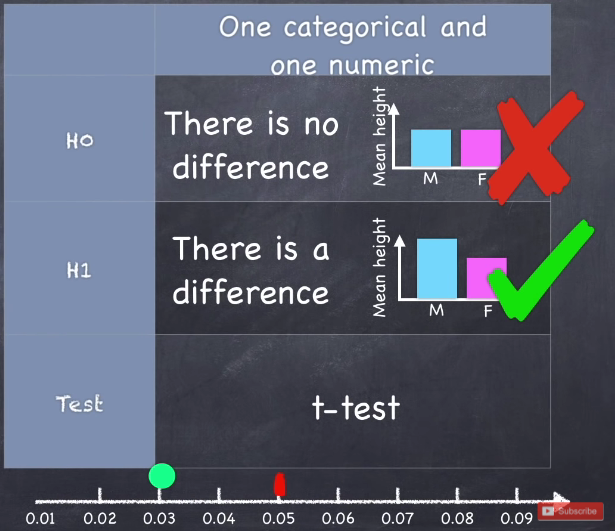
* We might ask the question, is there a difference in the number of men and women in the population?
* Now we could state that as a hypothesis - which is that, there is a difference between the number of men and women in the population.
* And we could check to see whether or not we think that that is the case.
* And when we look at our sample data well, we do in fact see that there's a difference in the proportion of men and women. So should we get excited? Well, no not yet!!!
* Remember this is just sample data, we could have by chance selected a sample that just happened to show a difference.
* So, let's consider the possibility that in actual fact there is no difference in the number of men and women in the population and we call that our **null hypothesis.**
* And if that were true how likely would it or what is the probability that we would see the difference that we have observed.
* If we can show that the probability is low then we can have a degree of confidence that the null hypothesis is wrong and we can reject it.
* But before we calculate this probability i.e., **p-value** we must be clear about how small is small enough below what value of p would we **reject the null hypothesis** and we must decide on that cut-off before we calculate the p-value.
* And we call that cut-off the **alpha value** and for the rest of the examples we're going to use an alpha value of 0.05 or five percent.
* Now we've really got two scenarios we've got the null hypothesis which is:
  + Null Hypothesis: There's no difference in population (male vs female).
  + Alternative Hypothesis: There is a difference in population (male vs female).
* Next step is to apply a statistical test and, in this case, we're doing **one sample proportion test** and we generate a p-value.
* If the p is less than the alpha then we can reject the null hypothesis and state that the difference that we observe is statistically significant.

Now let's look at what we can do with **two categorical variables** like Gender and Age Group.

* We may have a research question like does the proportion of males and females differ across these groups?
* So, our hypothesis is that the number of men and women that we observe is dependent on the age category that we look at, in other words the proportions are dependent on the age category.
* Now we can collect our sample data we look at it and we can see that yes in fact the proportions do change across the age groups.
* In other words, in our sample data the proportions are dependent on age category.
* Now is that due to chance? Well let's test the idea that the proportions are all the same, that they are independent of age category that's our **Null Hypothesis.**
* Here we can conduct a **chi-square test** and that gives us a **p-value** and if the p-value is less than the **Alpha** we can reject the null hypothesis and state that our observation is statistically significant.

Now let's look at what we can do with **one numeric variable** like Height.

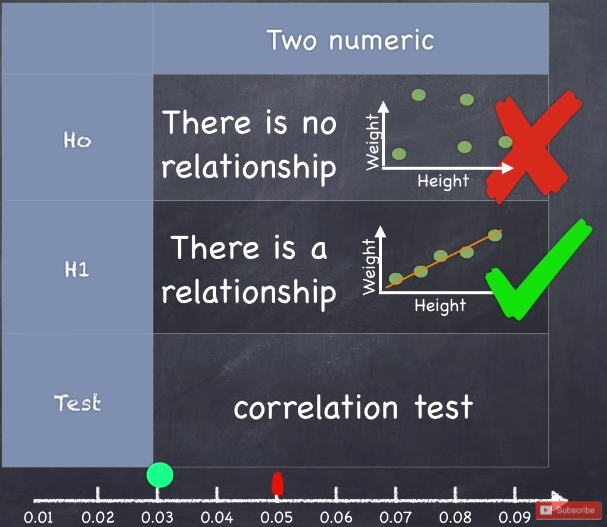
* We don't have any groups to look for differences between and we don't have another numeric variable to look for some sort of associational relationship.
* So, what questions can we ask?
* We might have some theoretical value that we want to compare our data to.
* For example, in the case of average height, we might have some historic data we might wonder if the current population is significantly different from that of historic data.
* So, our question might be, is the average height different from a previously established height?
* Let's imagine that the previously established height was 1.4 meters and we want to know if the average height in our current population is different to that.
* Our hypothesis is that there is a difference.
* Again, we collect some sample data we find that the average height is indeed different from the historic height.
* Is that statistically significant?
* Well, if there were no difference what would the chances be that we observed the difference that we do or a greater difference.
* We conduct a **t-test** comparing the averages.
* If the p-value is less than the alpha then we can reject the null hypothesis and state that the observed difference is statistically significant.

Now let's look at a **categorical variable** and **a numerical variable** like Gender and Height.

* We may ask the question is there a difference between the average height of men and women?
* In this case our hypothesis is that there is a difference in our sample, we do observe a difference.
* Let’s, assume that there's no difference.
* We conduct a t-test which gives us a p-value.
* If the p is less than the Alpha will reject the null and we state that the observation is statistically significant.

NOTE: If we had a categorical variable with more than two categories like age group that's got three categories then instead of doing a **t-test** we would do an **analysis of variance or ANOVA**.

Now let's look at what we can do with **two numeric variable** like Height and Weight.

* Here we might start with the question, is there a relationship between height and weight?
* Our hypothesis is that there is a relationship.
* We collect sample data; we look at it and we do see some sort of relationship.
* Is it real? Let's assume that it's not. Let's assume that there's no correlation between the two variables.
* And if it weren't real then what are the chances that we'd see the relationship that we do here?
* Here we do **correlation test**.
* Now a correlation test is going to give us two things:
  + Firstly, it's going to give you a **Correlation Coefficient** which tells us something about the nature of the association between the two variables.
    - Correlation Coefficient is a number between negative 1 and 1.
    - It looks at the relationship between two numeric variables.
    - If as the X variable gets larger and the Y variable gets smaller, we say that they are negatively correlated.
    - If they are perfectly negatively correlated then the correlation coefficient will be -1.
    - If there's no relationship between the two variables then the correlation coefficient will be 0.
    - If X goes up Y goes up then they are said to be perfectly corelated and the correlation coefficient will be 1.
    - And of course, you can have any value in between.
    - And by the way it doesn't matter which of your variables is on the x and the y axis.
  + Secondly, it gives us a p-value.
* And again, if the p-value is less than the Alpha we can reject the null hypothesis and state that the correlation that we see is statistically significant and the correlation that we see can be represented by a number that we call the correlation coefficient.