Statistics

## Fundamental Statistics

One of the pillars of the medical sciences are *clinical tests* – to know whether some treatment works, one tests it against something whose effect is known (either a placebo, or known treatment). Depending on the results of this test, we may know whether the treatment actually works or not.

These clinical tests have two parts: their execution and their analysis. In this chapter, you will learn about data, and how to analyze it.

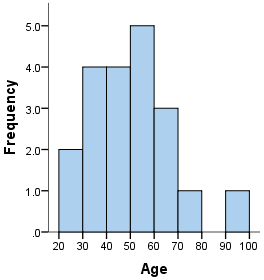
### Data sets and histograms

Clinical trial data is composed by data sets: for every patient, one has one or more attributes that are measured. For example, one may have the age, height and weight for all the patients.

Data sets can be represented as histograms. For example, suppose we have the following data set of patient ages:

36 25 38 46 55 68 72 55 36 38 67 45 22 48 91 46 52 61 58 55 36 25 38 46 55 68 72 55 36 38 67 45 22 48 91 46 52 61 58 55

can be represented as:



Histogram

The height of the bar between 20 and 30 counts the amount of patients aged between 20 and 30 – in this case, 2 people. The same for 30 and 40, 40 and 50, and so on. The choice of the values 20, 30, and so on is arbitrary – one can chose different points and obtain a different histogram from the same data.

Histograms give us a visual representation of the data set, showing a distinctive shape, which can be used to analyze the data conceptually.

### Statistical Measurements

Every data set has distinctive characteristics, but there are some measurements that are useful very often. In the following, instead of talking about measurements as numbers, we will talk about them using and so on. Each one is a number, corresponding to patient .

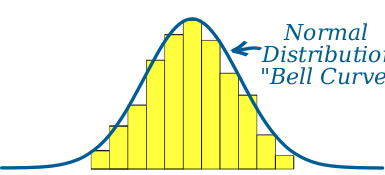
The most important statistical measurements are:

* Minimum: the smallest value in the data set.
* Maximum: the largest value in the data set
* Median: the value which splits the data set in two equal parts
* Average: an intermediate value defined by:
* Standard deviation: a measurement of the width of the data set:
* Standard error: a measurement of the error when measuring the mean:

## Statistical Tests

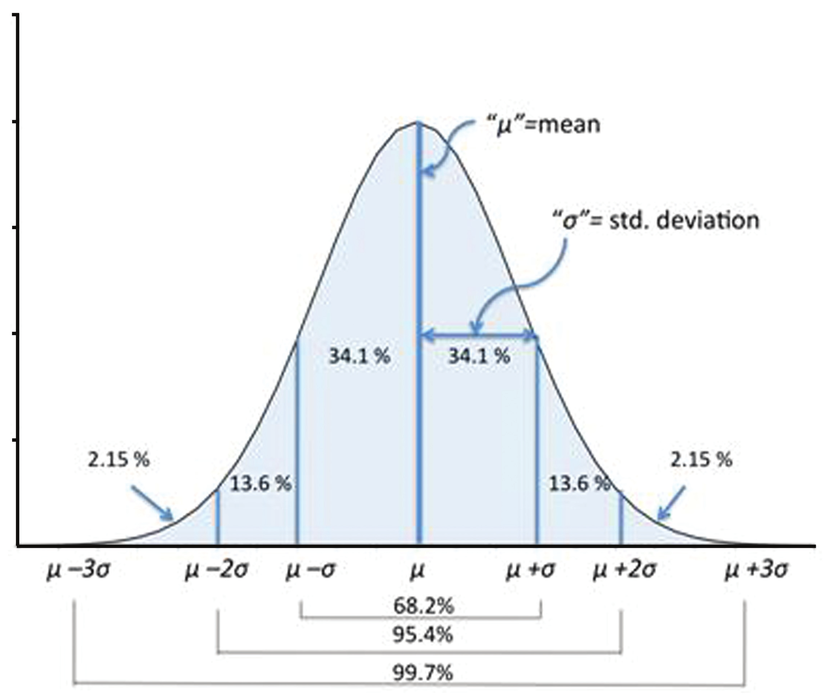
### Normal Distribution

The normal distribution happens very often in nature. It has the following shape:



Normal Distribution

It is determined by just two parameters: it’s average and it’s standard deviation:



Normal Distribution: Average and Standard Deviation

A frequent question in the medical sciences is: are two data sets “the same”? In the example Excel file, we may ask the following question:

Do mothers who smoke have children with the same height or weight? Does smoking has an effect on baby size?

For a normal distribution we can test whether the average and the standard deviation are the same.

### Student Test

[Wikipedia Reference](https://en.wikipedia.org/wiki/Student's_t-testhttps://en.wikipedia.org/wiki/Student's_t-test)

The goal of Student’s test is to check whether the distributions’ averages are different. The Student test formula is:

In Excel, this test can be done using: T.TEST(A1:A100,B1:B100,1,3)

### Fisher test

[Wikipedia Reference](https://en.wikipedia.org/wiki/F-test_of_equality_of_variances)

The goal of Fisher’s test is to check whether the distributions’ standard deviations are different. The Fisher test formula is:

In Excel, this test can be done using: F.TEST(A1:A100,B1:B100)

### In Excel

#### Filtering Values

You will need to filter values in Excel. Supposing that the selection values are on column B, and then the desired values are on column A:

=AVERAGE(IF(B1:B100=0,A1:A100))  
=STDEV.P(IF(B1:B100=0,A1:A100))

Select the cell, enter the formula above, and type CTRL-SHIFT-ENTER.

#### Executing Tests

We are going to calculate the average and the standard deviation of the chosen attribute (for example, weight), both for smoking and non-smoking mothers. These values are going to be slightly different, which brings the question: is this difference statistically relevant?

To know whether the difference between averages is relevant, we apply a Student t-test (T.TEST). For the difference between standard deviations, we apply a Fisher F-test (F.TEST). The result of both is a **p-value**.

How to interpret p-values.? The number answers the question:

If both distributions were the same, what would be the chance of getting the result we observe?

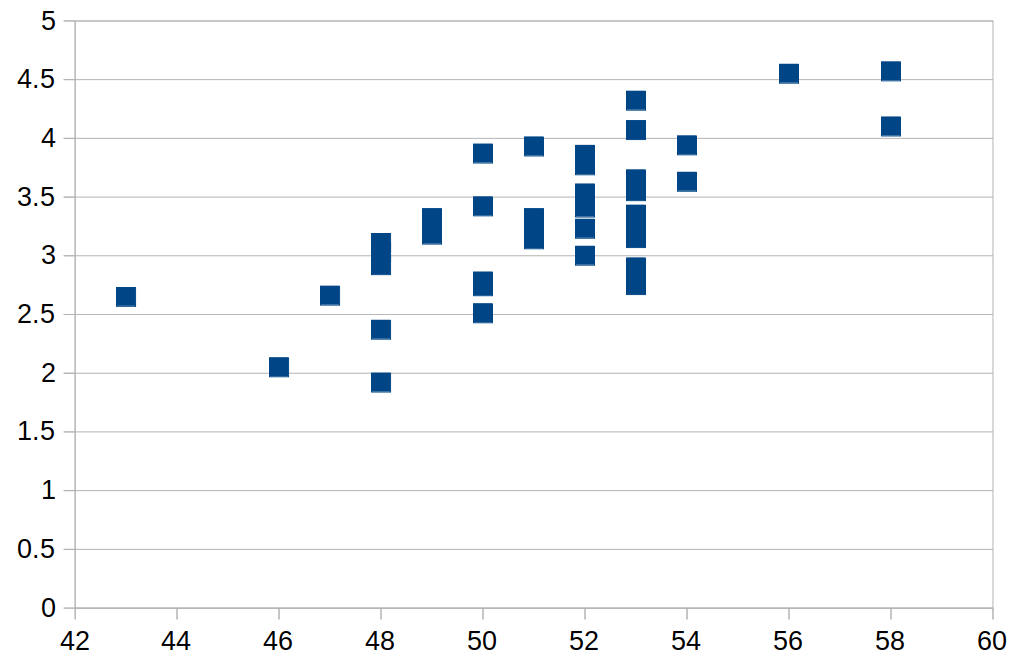
That is, the lower you number you get, the more the result is meaningful. Very often, values less than 0.05 are considered significant. It means that there is less than 5% of chance that this difference would be observed randomly.

## Correlation

Correlation is a measure of how two things behave similarly. For example:

* People’s height is **positively correlated** with their weight – that’s because a bigger height in general means a bigger weight.
* In a house, the amount of people is **negatively correlated** with the amount of space per person – that’s because the more people you have, the less space each one can have for themselves.

These behaviors can be seen in a graph. For example, look at this graph:



Correlation

As we can see, the values tend to grow from the left to the right, together. This is the sign of a positive correlation.

Mathematically, there is a formula for correlation, which takes value from -1 up to 1. It goes from negatively correlated (-1), to uncorrelated (0) to positively correlated (1).

### Formula

In order to calculate correlation, one needs sequences pairs of numbers. These sequences will be written and . Writing for the average of and for the correlation of , then the formula for **correlation** is:

## In Excel

In Excel, the correlation between two data sets can be calculated using the CORREL(A1:A100,B1:B100) function, replacing the data ranges by the desired values.

### Exercises

In the babies data set, there are many measurements. For any pair of measurements, we can:

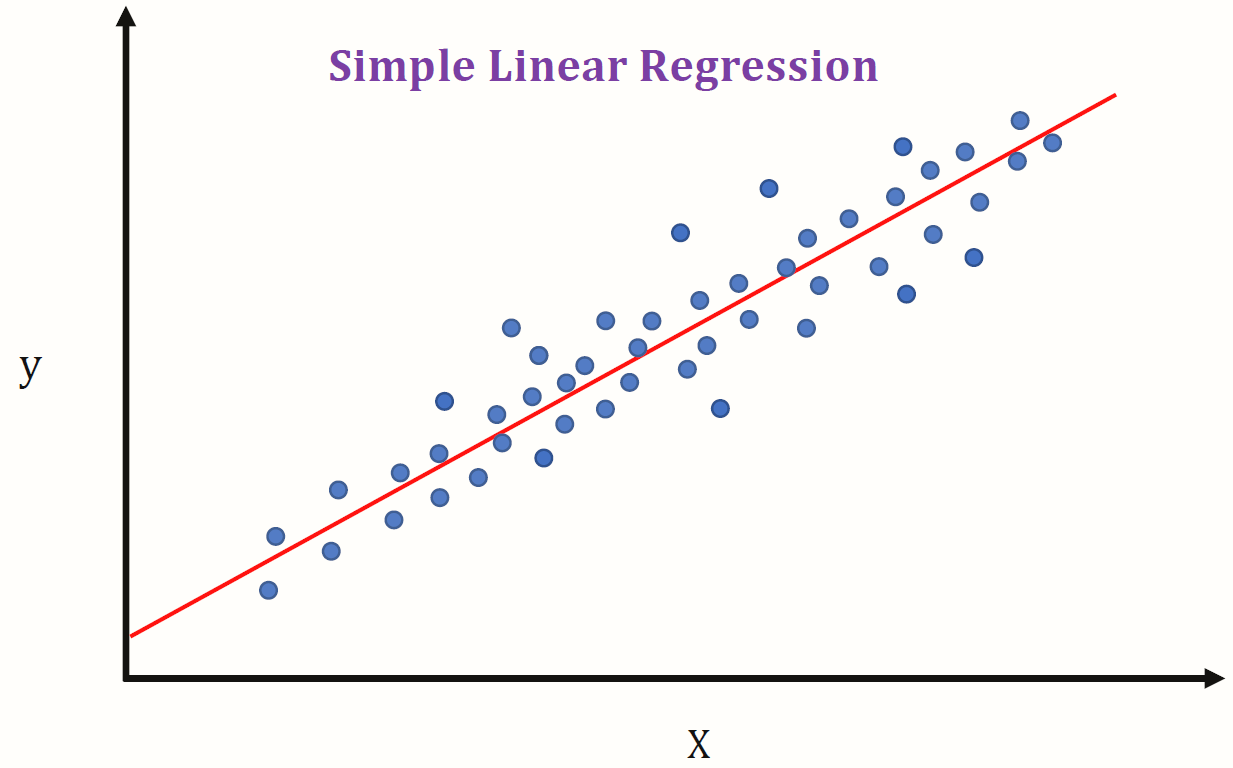
* Create a X-Y scatter plot to visualize the data
* Analyze it, and try to infer if there is a positive, negative or no correlation.
* In a separate cell, measure the correlation using CORREL.

Some interesting pairs for doing it are:

* Size and weight
* Size and number of cigarettes smoked by the mother
* Weight and number of cigarettes smoked by the mother
* Size and father height
* Size and father years of education

## Regression

In the section above we studied correlation. In some cases, such as weight and size, there is a very clear relation between the two variables. This deserves a question: can we model this relation with a straight line?



Regression

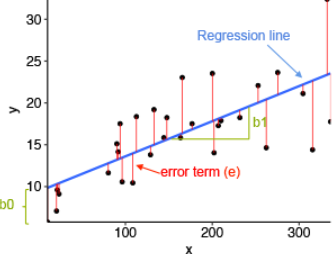
Look at the graph above: each blue dot has a coordinate . The straight line, is also composed by pairs of coordinates, such as the points, but its points follow a rule:

In this formula, there are two coefficients:

* is the intersection coefficient: when
* is the inclination coefficient: it measures the speed of increase or decrease of the variable

How should we draw this line? After all, we can draw any line in the graph, and it will be a model – but its quality can be better or worse, depending on the coefficients.

In order to choose the coefficients, we evaluate the error term:



Regression error

The error term measures the difference between the measured value (one of the values) and the value predicted by the model (calculated by ). The best model is the one which makes the error, measured by:

as small as possible. The square () is there to make calculations easier.

If we use this criterion, we can calculate the regression coefficients as:

where represents the mean, the regression coefficient, and the standard deviation.

### Practice

Choose a pair of measurements. For this pair we will create a table with:

* Average of ()
* Average of ()
* Standard deviation of ()
* Standard deviation of ()
* Correlation of and ()
* The coefficient
* The coefficient
* The optimal
* The optimal

Next to it, two columns:

* One with the prediction
* Another with the error