On replicability, verifiability and falsifiability

Table of content

1. Replicability and falsifiability in the physical sciences

In the sciences we develop hypotheses, namely statements we believe are true about a physical phenomenon but that we have not yet shown to be true. The general statement we make about the phenomenon has to be worded in a very specific way. The statement has to be worded in a way that it can be tested for confirmation or non-confirmation. For example, the following statements are worded in such a way that their concept can be tested:

- "Water above 0^oC is liquid". This is not true, water contracts between 0^oC and 4^oC and can therefore become solid. This is why ice (a solid) floats on water (a liquid).
- "The solubility of salt increase with increasing temperature";
- "Fatigue testing gives significantly better results on the failure rates of steel compared to hardness and tensile testing".
- "Impurities in NaCl crystals results in a change in the lattice spacing in the growth direction."
- "The magnetic field of an infinite rectilinear current causes particles to follow a spiralling motion."

Science then relies on statements or hypotheses that can be falsified. Now, the paradox is that we then conduct experiments in order to confirm our hypothesis (not in order to falsify it). If our results confirm our hypothesis then we have evidence that supports our hypothesis. Note that we don't yet have evidence that our hypothesis is true. Our hypothesis becomes true only after more and more evidence is found which supports it. So one positive experiment may support our hypothesis and two positive experiments will more positively support our hypothesis. Then 20 positive experiments will so strongly support your hypothesis that we may consider our hypothesis to be verified, i.e. true.. But without a hypothesis which can be falsified there is no way we can confirm the hypothesis to be true. So if our hypothesis cannot be falsified then is no way we can test it experimentally to see if it is true.

2. Some examples on falsifiability and replicability

It is useful to study falsifiability and replicability together, so we start with some examples which illustrate how hypotheses can be replicated in order to confirm or falsify them.

2.1. Example 1: Measuring the temperature of boiling water – version 1

Statement: Water always boils at 100^0C at all heights off the Earth's surface.

Question: Is this statement falsifiable? Can we perform experiments that would, in principle (say, in our imagination, even if not in practice), show that water boils at lower or higher temperatures above the Earth's surface? Yes. So this statement is in fact a hypothesis.

Experiment: Boil some water at ground level and measure its temperature. This should happen (near enough) at 100^0C.

Replicate the experiment: Now boil some water at different heights. You will find that water boils at different temperatures for different heights.

Height/elevation (metres)	**Boiling point (^{o}C)**
0	100
1143	96
2896	90.1
8848 (Everest)	72

So we see that by replicating the experiment we have found our hypothesis to not be true. We have falsified our hypothesis.

2.2. Example 2: Measuring the temperature of boiling water – version 2

Statement: Water always boils at 100^0C at any pressure at ground level.

Question: Is this statement falsifiable? Can we perform experiments that would, in principle (say, in our imagination, even if not in practice), show that water boils at lower or higher temperatures above the Earth's surface? Yes. So this statement is in fact a hypothesis.

Experiment: Boil some water at ground level and measure its temperature. This should happen (near enough) at 100^0C.

Replicate the experiment: Now boil some water at ground level using different pressures. The pressure at the surface of the Earth is said to be 1 atmosphere. We will measure the temperature of boiling water measured in multiples of atmospheres. You will find that when the pressure increase water requires higher temperatures in order to boil.

Pressure at ground level (Atmospheres)	**Boiling point (^oC)**
1	100
2	121
3	134
4	144

So we see that by replicating the experiment we have found our hypothesis to not be true. We have falsified our hypothesis.

3. Repeating an experiment versus replicating an experiment

Replicating an experiment does not mean repeating the experiment. In terms of conducting experiments, replication and repetition are two different things. As an example consider the experiment of determining the boiling point of water. Then let us have the following:

i) *Equipment*: Kettle, mercury thermometer, tap water,

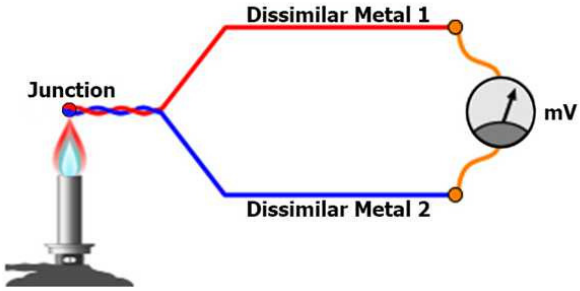
ii) *Environment*: The experiment performed at home,

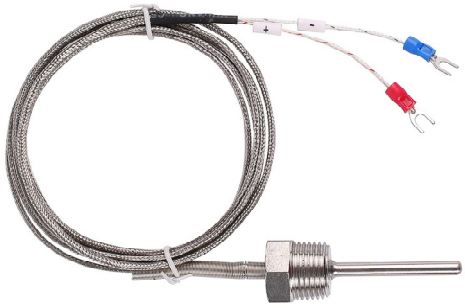
iii) *Procedure/protocol*: "Procedure 1"

Given the above we can say that

- *repeating* an experiment consists of using the same equipment in the same environment with the same instructions/protocols;
- *replicating* an experiment consists of conducting the experiment using different equipment (say, using a beaker because maybe the glass affects the boiling temperature, or using a different type of thermometer such as a thermocouple or a pyrometer), and/or different environment (say in a clean room), and/or different procedures (say, "Procedures 2").

(*Optional information*: A thermocouple is an instrument consisting of two different metal wires. When these two wires are brought into contact with heat the heat generates an electric current. Since the wires are made of different materials they heat up at different rates. This causes heat to travel faster along one wire than the other which then affects the speed at which electric current flows along each wire. This difference in speed causes a voltage which can be measured using a voltmeter. The higher the voltage the hotter the heat source. A schematic of a thermocouple is shown in diagram (a) below, with a photo of a real thermocouple shown in photo (b) below.





(a) (b)

A pyrometer is a type of remote-sensing thermometer used to measure the temperature of distant objects. It is a device that from a distance determines the temperature of a surface from the amount of the thermal radiation it emits. An example of a pyrometer is shown in the photo below.

)

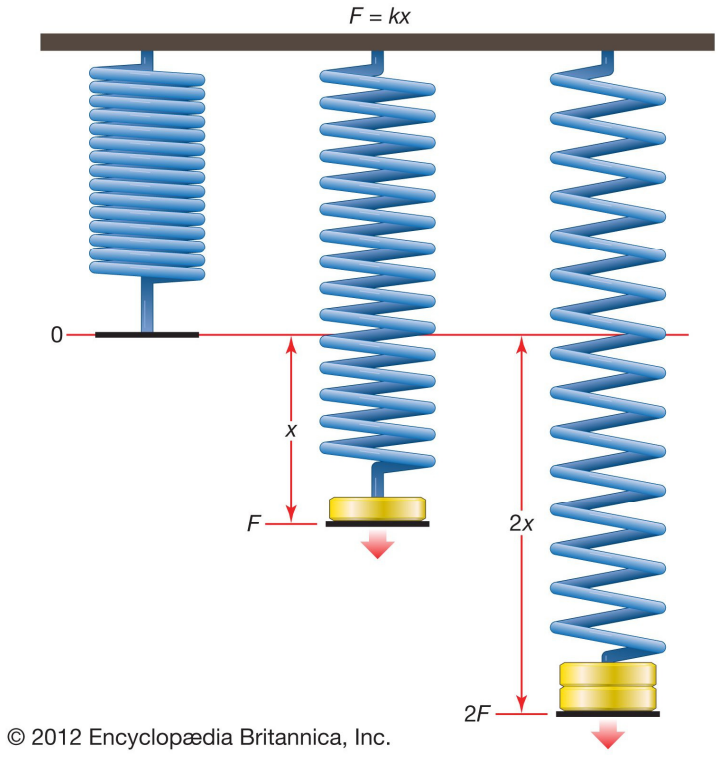
3.1. Example: Hooke's law of elasticity

In 1676 Robert Hooke mathematically described how the stretch in a spring is related to the weight at the end of the spring. Obviously the heavier the weight the greater the stretch, but if we know how much a spring has stretched with a 1kg weight, what is the effect on stretching for a 2kg weight? Hooke found that the amount of stretch is proportional to the weight (the stretch also depends on the type of material the spring is made of. Different materials have different springiness and therefore different stretchability).

The formula Hooke devised is $F = -kx$, where F is the force from the weight (in kg) attached to the end of the spring, x is the amount of stretch (in cm or m) that has occurred and k is a number which describes the inherent stretchability of the material which the spring is made of (i.e. k is different if the spring is rubber compared to metal compared to plastic, etc.). What this formula says is that

- if we double the weight we double the amount by which the spring has stretched from its unstretched length;
- if we triple the weight we triple the amount by which the spring has stretched from its unstretched length,

etc. This is what "proportional" means.



Statement: Assuming we don't stretch the spring too far (otherwise it will not spring back), the stretch of a spring for given weights is given by the formula $F = -kx$.

Question: Is this statement falsifiable? Can we perform experiments that would, in principle (say, in our imagination, even if not in practice), show that for a given weight and stretched amount, if we double/triple/quadruple/… the weight we will double/triple/quadruple/… the stretched amount of the spring. Yes, this statement is falsifiable. So it is in fact a hypothesis.

Experiment: Measure the unstretched length of a spring. Assuming you know the value of k, put a 1kg weight on the end of the spring and measure how much it has stretched. Then put a 2kg weight on the end of the spring and measure how much it has stretched. Confirm the results are proportional, implying that $F = -kx$ is a correct formula.

Replicate the experiment: We want to confirm the formula which means we want to confirm the idea that stretchiness is proportional to weight. We can do this by i) using a spring of a different length (longer or shorter), or ii) using springs of different materials (rubber, different metals, plastic, etc.)

Questions

1) Suppose person A conducts an experiment using certain equipment, a certain protocol, and obtains data which we can call data set X. Now suppose person A performs the experiment again with the same equipment and protocol and gets different data. Has the experiment been replicated?

2) Suppose person A conducts an experiment using certain equipment and a specific protocol, and obtains data which we can call data set X. Now suppose person A performs the experiment again with the same equipment and protocol and gets the same data set X. Has the experiment been replicated? Suppose now that person A gets most of data set X but not all of it. Has the experiment been replicated?

3) Suppose person A conducts an experiment using certain equipment and a specific protocol, and obtains data which we can call data set X. Now suppose person B performs the same experiment with the same equipment and protocol to get the same data set X. Has the experiment been replicated? If person B gets a different set of data has the experiment been falsified?

4) Suppose person A conducts an experiment using certain equipment and a specific protocol, and obtains data which we can call data set X. Now suppose person A performs a different experiment, say using different equipment and/or protocol and gets the same data set X. Has the experiment been replicated?

Suppose person A conducts an experiment using certain equipment and a specific protocol, and obtains data which we can call data set X. Now suppose that 10 other people, independent of A attempt to replicate A's experiment to obtain similar results. Three of those people are not able to replicate A's results, one person's attempt is inconclusive, and the remaining 6 people are able to replicate A's results with varying degrees of accuracy. Has A's results been replicated?

4. Commentary on replicability and falsifiability in the physical sciences

So what is replicability and falsifiability in the sciences?

4.1. On falsifiability

As was stated earlier a statement or theory about natural phenomena can only be called scientific if it can be falsified. A statement/theory which cannot be falsified is not considered scientific. So, for any statement/theory to be called scientific it has to be stated in such a way that it can be verified or falsified. For example,

- "The extension of a spring is proportional to the weight attached to the spring". This is true. It is known as Hooke's law;
- "All substances expand when heated". This is not true: water and silver iodide contract when heated within a range of temperatures (silver iodide contracts when heated between 80^{o}C and 141^{o}C), and rubber expands when it gets colder. Also, a weird substance called zirconium tungstate keeps shrinking between −273^{o}C and 770^{o}C.
- "Ultra-sonic fatigue testing of cold-formed ferrite steel at lower-critical temperatures is a more accurate predictor of failure than sub-sonic fatigue testing."
- "NaCl crystals have lattice shapes."
- "Material X shows 10% reduction in noise compared to …"

Then, showing the validity or falsifiability of a theory refers to being able confirm, or not, the predictions that theory makes about the physical phenomenon it is describing. This confirmation or falsification comes as a result of conducting experiments. If experiments produce data which confirms the theory, and if the experiments can be replicated, then there is good evidence that the theory is correct.

For example, Einstein's theory of general relativity predicted that

- light would bend around massive objects such as the sun. This was finally confirmed experimentally in 1919 by Arthur Eddington (1882 – 1944), an English astronomer and mathematician, when he went to the West African island of Principe to study a solar eclipse that was occurring at the time. Relativity predicted that light would bend when passing near a massive object (the sun). The bending of starlight would then appear on photographic plates as stars appearing to have changed position from where they were known to be (the bending of starlight effectively creates a visual illusion that stars are in a slightly different position to where they actually are);
- the universe was constantly expanding, and was not static. In 1922 a Russian physicist called Alexander Friedmann used Einstein's theory of relativity to show that the universe had to be expanding. In 1927 a Belgian physicist/mathematicina called George LeMaitre also (independently) calculated that the universe should be expanding. The actual verification of these results had to wait until 1929 when Edwin Hubble, an American astronomer, collected data on the speed of galaxies which confirmed that the universe had to be expanding.

In both cases above Einstein's theory made predictions which were testable and therefore verifiable or falsifiable.

4.2. On replicability

We have seen examples of replicability earlier in these notes. It should be noted that that replication is not the same as repetition. To repeat an experiment is to use exactly the same equipment and set-up of equipment; to use exactly the same protocol or experimental procedure; to use exactly the same environmental conditions (if this is relevant), etc. In this case we should obtain exactly the same results.

However, in replicating an experiment we can change some or all of these aspects. For example, in statistics we can change the sample used to calculate the mean or standard deviation. We can then use all the same tests previously used in order to confirm a hypothesis about the mean or standard deviation. Then, if the hypothesis is confirmed for different samples it is probable that the hypothesis is valid for the population as a whole.

As an example, in the physical sciences consider a company which produces thousands of a certain item. Batches of this item are then processed or treated. Finally, tests or measurements are conducted. Several options might be available to obtain ten test values. Some possibilities are:

- One finished and treated item might be measured repeatedly to obtain ten test results. Only one item was measured so there is no replication of the measurement. The repeated measurements help only in identifying observational error;
- The last batch from the production line is chosen. Then, ten finished and treated items are taken from this batch. All these ten items are measured once. This is not a full replication of the measurements because the ten items all come from the same single last batch (this is not a random choice of batches since the batch was chosen on the basis of being the last in the production line). In other words, the ten items are not taken as randomly as they could be;
- Ten items are taken at random from the production line. This can be said to be a full replication of the measurement, since each item is not identical (each item is likely to have small, subtle differences due to the inherent nature of the manufacturing process).

So, it can be said that results from an experiment which replicates another experiment will be more confirming of a theory than repetition of the same original experiment (if you want to know the time look at two different watches or ask two different people, instead of looking at your watch twice).

It should be clear that the aim of repetition or replicability is to obtain the same results when the experiment is conducted again, or when the experiment is conducted by different people. But such a case is only true for an undergraduate performing an experiment in the laboratory when testing a known theory (why is this true only in this situation?). In the real world obtaining the same results when the experiment is repeated/replicated by different people in different conditions is not totally true. Instead scientist look for results which agrees to within an acceptable level of accuracy. In other words, scientists accept a certain degree of deviation between results of experiments repeated or replicated by other people.

Some experiments can never be replicated by other people independent of the original researchers, nor can those independent people ever use alternative experimental equipment and protocols. An example of this is any practical physics experiment conducted at CERN, (Conseil Européen pour la recherche nucléaire). Here experiments are carried out by bombarding particles into each other at speeds close to the speed of light. The equipment used is so expensive and complex that it is impossible for one or two or three people to reproduce it and replicate the results. Furthermore, in order to conduct a CERN experiment you need hundreds of people involved in all aspects of the process. Again, two r three people would not be enough to conduct such a huge scale experiment.

So at a basic level we have the following:

- a theory has to be able to make predictions,
- these predictions should be testable,
- the tests will confirm or deny the predictions made by the theory.
- repeating or replicating the tests should produce the same results, implying stronger confirmation of the theory. If the results are not the same they should be within an accepted range of values.

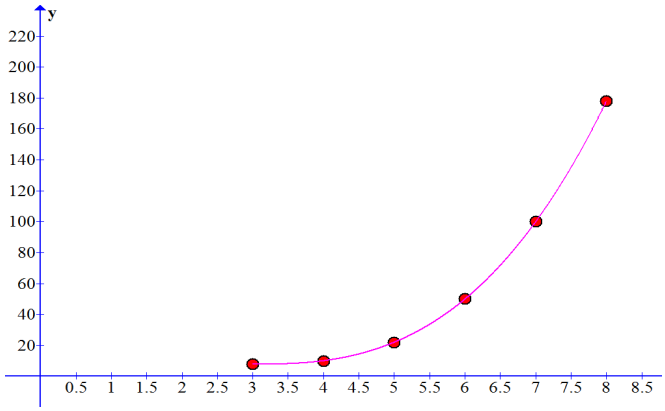
By this approach a theory can be verified or falsified. This is what separate science from non-science.

5. Commentary on replicability and falsifiability in statistics

Replicability, verifiability and falsifiability is much more problematic if you use statistics as your method of analysis. This is because statistics deals with the probabilities and trends in data not with definite exact relationships. For example, data and graph (i) below illustrates an exact relationship between the *x* data and the *y* data. But data and graph (ii) does not. It only suggests a trend. So no exact answer will be obtained when using methods for analysing trend data.

x	*y*
3	8
4	10
5	22
6	50
7	100
8	178

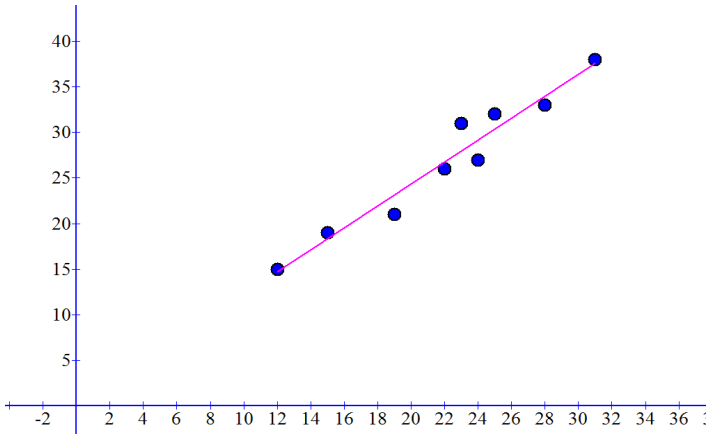
Data (i)



Graph (i)

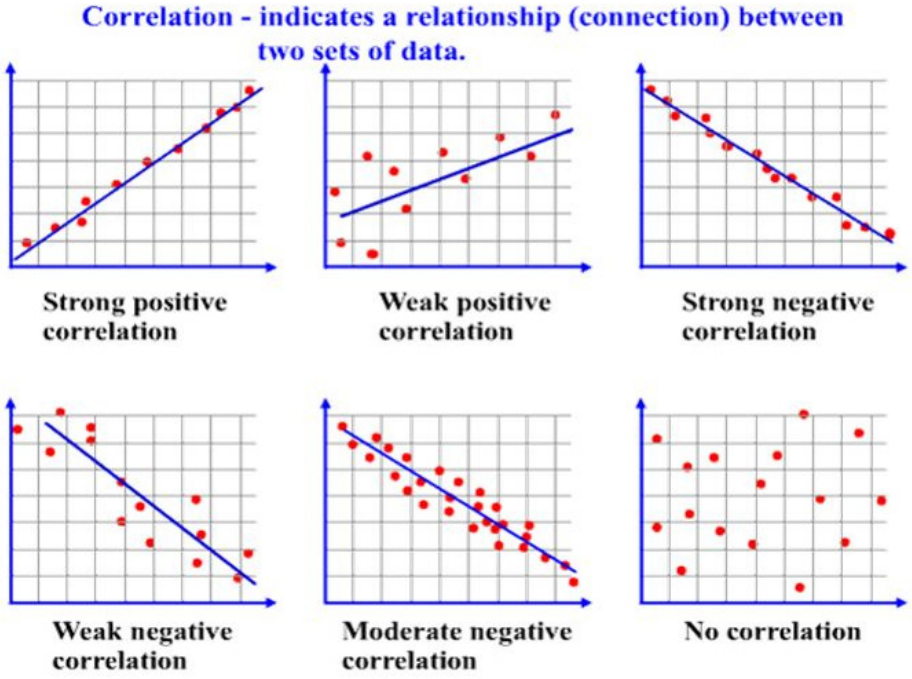
x	*y*
12	15
22	26
19	21
15	19
31	38
25	32
28	33
24	27
23	31

Data (ii)



Graph (ii)

Some trends are strong, some are weak as illustrated below (the strength of a trend is called correlation).



Factors which determine the answers you obtain in such situations include

- how you collected the data, i.e. the experimental procedure you used;
- where and when you collected the data, i.e. is location important and time of day important when collecting data, or was there anything influencing the ability to collect good quality data, etc.

and how much data you collect, i.e. too little data will not give you a probability or a trend which is viable for the purposes of making future predictions.

For example, consider one of the most basic analysis done in statistics, which is to compare the means, say the mean heights, between two groups, say the population of two countries. The question which can be asked is,

Is there a statistical difference between the mean height of two countries?

The first thing to notice is the language used in this question. It doesn't say "Is there a difference between …?". It says "Is there a statistical difference between …?". Without going into any details this means that we are prepared to accept a certain amount of difference or deviation between the two results. Then,

- if the difference lies within a certain range we accept that the mean heights are *statistically* the same (or similar) between the two countries,
- if the difference lies outside a certain range we accept that the mean heights are *statistically* different (or not similar) between the two countries.

This then is how repeatability, replicability, and falsifiability works in statistics. We are dealing with the probability that the mean (or the spread of the data or the linearity of data or some other thing we are measuring) is *statistically significantly different* between two groups, or before an experiment compared to after an experiment. Any results obtained are valid only for large scale data not individual cases.

Even within a single group there are difficulties replicating an experiment in order to confirm or falsify results. Here is a concrete example using numbers. Suppose we have a population of 100 people, and we have collected their heights (in metres), as shown in the table below. The mean height of the population (all 100 people) is 1.79m and the spread (standard deviation) is 19.5cm.

2.15	1.68	1.67	1.78	2.04	2.13	1.53	1.65	1.49	1.90
1.97	1.90	1.82	2.06	1.71	1.39	1.68	1.91	1.82	2.16
1.83	1.57	1.60	1.38	1.88	1.89	1.33	1.60	1.58	1.19
1.76	1.65	1.99	1.64	1.67	1.69	1.96	2.00	1.59	1.99
1.61	1.96	2.06	1.87	1.80	1.94	1.53	1.95	1.73	1.84
1.44	1.62	1.84	1.78	1.82	1.81	1.81	1.74	2.10	1.85
1.73	1.80	2.06	1.80	1.79	2.00	1.76	1.59	1.81	1.60
1.92	1.68	1.98	1.73	1.88	1.80	1.90	1.78	1.79	1.87
1.77	1.98	1.87	1.63	2.09	1.49	1.63	2.04	1.74	1.78
1.85	1.58	2.14	1.91	1.80	1.55	2.22	1.82	1.74	1.74

We know that it is impossible at a practical level to measure the heights of the whole population so instead we take samples. When you collect a sample of data from a population the mean and spread you calculate will not be identical to that of the population.

Now consider five different experimenters conducting the experiment of measuring heights of 10 people. Experimenter 1 takes his sample to be row 1 of the table above, experimenter 2 takes row 2, etc. the table below show their results.

Sample of 10 people	**Mean (m)**	**Spread (m) (called Standard Deviation)**
Experimenter 1 (row 1)	1.80	0.242
Experimenter 2 (row 2)	1.84	0.217
Experimenter 3 (row 3)	1.59	0.236
Experimenter 4 (row 4)	1.79	0.169
Experimenter 5 (row 5)	1.83	0.167

In fact if we take *any* sample of 10 people from the population of 100 people the means and spread will be different between all samples. Note that the mean of Experimenter 4 is the same as the population mean. This was just luck (and even then we wouldn't know that this mean was the same as that of the population).

Exercise

Refer to the slide for four questions to consider based on the example above.

In general, statistics is about making inferences about the population based on samples we have collected. We therefore have to takes samples in order to estimate the mean and spread of the population. But as we have seen, different samples will give us different means and spread. So we will never obtain the exact same results from experiments which use statistical analyses. In the strict sense these experiments are not replicable. They are not even repeatable unless you use the exact same data. This is something we have to live with. Does this mean that the experiment has been falsified? No. Statistics is a useful method of analysis since it provides information about large/big data. We mitigate errors and difference in results by taking large samples, or taking multiple samples. In doing this we reduce the difference/error in results between two experiments conducted independently of each other.

Statistics is then about obtaining consistent results within an error margin deemed acceptable, and these results are valid only for large scale data not individual cases. There is no such thing as truth in statistics, only probabilities, trends, degrees of significance or confidence (but you are free to argue the case against this).

Exercises

1) How does your discipline think of replicability and falsifiability? How does the idea of replicability and falsifiability apply to your disciplines?
To what extent, if at all, can you apply these ideas in your discipline?

2) Find statements or theories in your discipline which are falsifiable. What makes them falsifiable? How have they been verified to be true?

3) Karl Popper (1902 – 1994) was an Austrian-British philosopher of science who rejected the idea that science progressed by using inductive reasoning. In other words he did not agree that scientific theories could be confirmed as valid by collecting data from experiments which proved the theory true.

Instead he said that science progressed using empirical falsification. In other words, he said that scientific theories could never be proven, only falsified. More and more experimental data could be collected which supported a theory, but this did not prove the theory true. However, if the data collected did not agree with the theory then the theory would be shown to be false.

Karl Popper's definition of the confirmation or not of a scientific theory is a very strict one. It means that we only need one piece of data to not confirm the physical theory for that theory to be false, i.e.

- "I have a theory which predicts some observations";
- "I conduct an experiment whose aim is to confirm this observation";
- "I don't get the observations I expect";
- "Does this mean my theory is false?";

Questions:

- Is this a practical way to conduct science?
- How many times in your own experiments (lab based or computer based) during your undergraduate degree were you able to confirm, one-hundred percent, the truth of a physical phenomenon?

- How many times did you obtain data that did not confirm the theory? Did you therefore reject the theory?
- How does this perspective apply to the research you will conduct in your future careers?

4) In my MSc mathematics project I studied the motion of an electron in the Earth's magnetic field. Mine was a maths/theoretical physics project (not an experimental physics project). The aim was to set up relevant equations of motion for the electron as ODEs (Ordinary Differential Equation) and then write a computer program that would solve the ODEs so as to trace the motion of the electron. I then plotted the results using 3D graphing software. The intention of my project was only to replicate already known results, and the graph illustrated the already known path of the electron. I was essentially verifying the already known physics of the motion of the electron.

So, in your case recall one experiment you did in the labs during your undergraduate degree.

a) Who designed the experiment?

b) Who set up the equipment for the experiment?

c) Who made sure the equipment was properly tested and calibrated?

d) Was the experiment designed for you to find new results or was it designed for you to replicated results already known?

e) At the end of your experiment did you get the results expected? If not, why not? What were the reasons for you not getting the expected results?

f) What about your final year project? If you did experimental or computational or design work in your final year project, were you replicating previous work or were you producing original/new results?

5) Do some research into the theory of the phlogiston (a theory about combustion) or Blondot's N-rays. Can the experiments supposedly demonstrating these products be repeated/verified/falsified?