Statistics or "How not to fool yourself"

Michael McEllin

When you are doing scientific research fooling yourself is the easiest thing in the World. You have some wonderful idea about how the World works, and go and look for the evidence. You plot a graph and it is doing what you predicted, so you think you have made a discovery. Wow! You would like to believe it, and it is very easy to convince yourself that what you have just measured is a real insight into nature. After all, it is what you hoped and expected would happen.

Then you show the stuff to someone else and the questions start: could there be a different explanation other than your pet theory? Could the data you see have occurred just by chance? Do you actually known how improbable that would be? Did you do the experiment again and get the same result? Can someone else do the experiment and get the same result? Before you get something published you will have to answer all these questions and more.

Let us revisit the analysis of day/night differences in cosmic ray arrival rates. Do we believe that our positive result actually means there is a real day/night difference? What other explanations might be possible? What do we do next to eliminate all but one explanation?

One of the most obvious characteristics of cosmic ray data at any one detector is that the events see to happen pretty randomly. It is very hard to distinguish any patterns. There may possibly be systematic differences between day and night, morning and evening, high and low atmospherics pressure, phases of the Moon and so on, but none of that is obvious because if there are any such effects they are well hidden by the variability in the data. In some hours the computer logs as few as 1000 events while in others it finds nearly 1500. In fact, if the arrival times of cosmic rays are completely random, with an average rate of say 1250 per hour, it would be very surprising indeed if in each hour we actually recorded just 1250 events – or even close to 1250 events. We *expect* to get more counts in some hours and in others to get fewer counts.

If we compare night-time counts with day-time counts over a week we ought, in fact, to be very surprised if we found *exactly* the same number of counts. Even if in reality the arrival rates are completely uniform between day and night, we would expect one to be higher than the other over any particular period of observation (though in some weeks day would win, and in other weeks night would have more counts).

Given that no difference would be surprising, how much difference between day and night do we need to see before we can be sure that any difference we do see is not just due to random variations? The truth is that we can never be *absolutely sure*. What we can do is to tell other scientists just how unlikely it is that the data they are being shown could have occurred by chance. Then they make up their own minds whether your results are worth thinking about.

In physics we rarely even think about showing our experimental data to other people unless the chances are pretty small – conventionally less than 3 chances in a 1000 – that we are being fooled by an unfortunate selection of data. There is nothing magic about this 3/1000 "significance level". Physicists have just agreed not to bother each other with "Look! I have something interesting!" unless we reach this threshold. It saves time and money trying to reproduce experiments supporting

claims that turn out to be premature. Different scientific fields, however, tend settle on different conventions for their own very good and valid reasons.

Biologists, social scientists and psychologists tend to report a positive result if there are less than 5/100 chances of it being wrong. This is a perfectly valid thing to do given the nature of biology, which means that it is much harder to collect large amounts of quantitative data and when collected it tends to be much more variable. They do their science in a way that takes account of the lower degree of trust you can invest in any particular experimental result – and it works for them and the field still advances. Using the 5/100 threshold means that they can do experiments that are individually less expensive. Some of these turn out to be wrong in the end, but when they get results that might look interesting they do more focussed experiments. On the whole, experience shows, it works best for their way of doing science.

Our 3 in 1000 is conventional for a lot of physics. We find it easier to collect large amounts of numerical data and the experiments often need such expensive and complicated apparatus (e.g. the Large Hadron Collider) that in practice other scientists would find it difficult to easily replicate the results. Hence, we need to be that much more confident before we report discoveries to avoid wasting other scientists' time.

There is another consequence of having a large amount of data to analyse, known as the "Look Elsewhere Effect". Let us suppose that our day/night analysis had not produced a positive result first time. So we might go on to analyse data from a different week, and if that does not work we do another week, and so on. Sooner or later a positive result will turn up, just by chance. Keep tossing the dice and you will get six sixes eventually.

In reality, the Look Elsewhere effect is a lot harder to spot. Having failed to detect a day/night effect, we might have looked for a morning/evening effect, correlation with phases of the Moon, England winning at cricket and so on. You did not get a positive result? Look elsewhere! Sooner or later you *will* get a positive result, just by chance. Unfortunately rather too much of this probably does occur in professional science, because scientific journals only ever publish positive results. Reports detailing all the experiments that did not produce positive results rarely get into print.

If you want claim the discovery of a new sub-atomic particle at the Large Hadron Collider you would need to use a discovery threshold of better than 1 in 3.5 million because they have mountains of raw data that gets statistically processed in lots and lots of different ways. Hence, they worry a lot about the Look Elsewhere Effect, and it also would matter so much more if we got it wrong because no one else can do the experiment again – there is only one LHC. (At the very least someone would get a Nobel Prize they did not deserve!)

I therefore suspend judgement on our discovery of day/night differences in cosmic ray arrival times. I would need to see the result replicated on data from other weeks, and I would also like to see whether effects such as changes in temperature or pressure are better explanations (always remembering that "Look Elsewhere" might still fool us, so we need to be really sure).

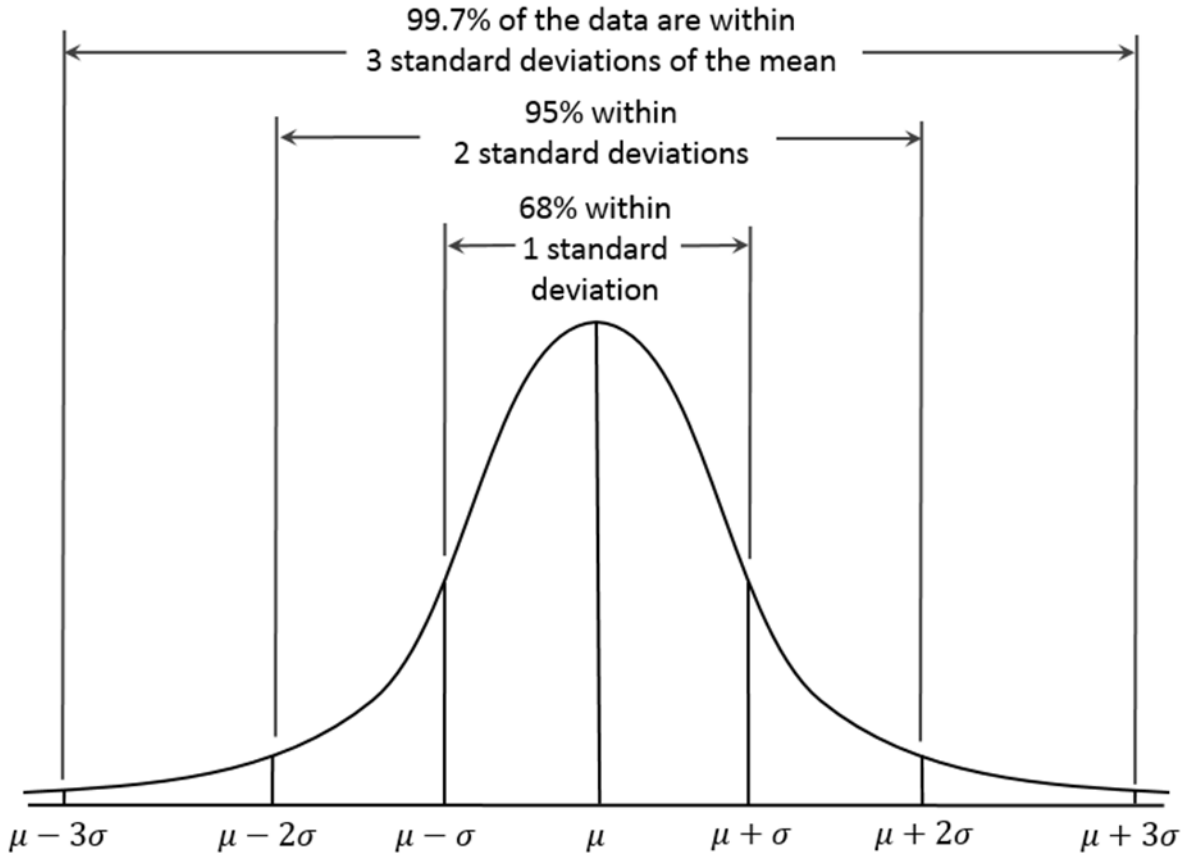
Remember! Nature is out to fool you! (Well, no. But it does not care if you fool yourself.)

A lot of statistics analysis is based on the principle that we should first assume the opposite of what we wish to believe and then we should calculate the chance that the data that we claim supports our

expectation is actually the result of chance. We only deserve to get excited and run around shouting "Eureka!" if the statistics are more significant than our 3/1000 threshold. If we pass this threshold, our fellow scientist will accept that we have a pretty good argument that we are not allowing ourselves to be fooled – but they still remember the 3/1000 possibility that we are wrong.

If there is a day/night difference in cosmic ray arrival rates (which is not inherently implausible) then a large effect should show up over a relatively small period of data collection. We found a result – though we also heard from Maria that most schools do not get a positive result. Maybe we were fooled by our data. Sooner or later one school would get a positive result just by chance and it may have been us. Alternative, it may be that there is a real, but small effect, and one week's worth of data is not enough to show it up reliably. Most schools do not see it – it stays below the threshold. In our case maybe the combination of a small effect and chance just pushed it above the threshold.

Why do we settle on the 3/1000 figure? Why is the figure not 5/1000? It turns out that a lot of random data tends to be distributed in a "bell" curve, generally known to statisticians as the *Normal Distribution* because it *normally* turns up almost everywhere. There are good reasons for this.



Most random variability in real world data is actually the combined effect of several independent random influences. For example, the average height of people is almost certainly the result of the combined effect of a number of different genes, and also environmental influences, like the amount of food at certain stages of growth. It turns out that this type of situation when you add the separate effects together always leads to a bell curve. This even works if the individual effects are much more on-off – such as *this* gene variant always boosts height by 10cm if present, but never by 2cm or 5cm. (The full explanation of this almost miraculous, non-obvious but *highly* convenient result – the

Central Limit Theorem, which is one of the most important results in statistics - is university level maths, so please just accept it for now.)

The width of this curve around the mean of the data (μ in the diagram above) is measured by a parameter called the "standard deviation" (denoted by σ in the diagram above). Our 3/1000 is actually the chance that a measurement will turn up that is more than 3 standard deviations away from the mean – the central point of the curve.

You will therefore hear physicists talking about a "3σ" result when they mean less than 3/1000 probability of the result being due to chance, or "2σ" when they mean less than 5/100 probability of the result being due to chance. The high energy physicists at the LHC require a "5σ" result (1/3,500,000) before they get to claim their Nobel Prize.

In my own opinion there are three levels at which we use statistics:

1. We can plot graphs and if we are lucky we get a crystal clear correlation between two variables. We do not bother to work out a probability of being fooled because it is obviously going to be well beyond any reasonable threshold of significance. This is known as qualitative, or visual statistics and some very influential statisticians say that visual statistics should be the first and best choice if they can produce a clear result.
2. We can do quick and rough estimates to see if we are likely to be well within some claimed significance level. Every scientist who uses statistics knows the $\sqrt{n}$ rule which turns up all over statistics. In our case, if the average number of counts that we would expect in a certain period is "n", we would expect to see this vary up and down in individual periods by about $\sqrt{n}$. So 1250 counts per hour on average gives about =/- 35, and we would expect to see about 70% of the data within this range. We do not get excited unless we see much bigger differences than this between different data collection periods. As a general rule, any estimate you make from "n" observations will have some uncertainty which varies as roughly $\sqrt{n}$ (which means that reducing the uncertainties by a factor of two needs four times as much data).
3. When the rough estimate does not produce a crystal clear result (but maybe it suggests that there are some real differences) we have to do proper statistical tests, where we take care to get the mathematics right. We normally begin studying the proper methods at A-level maths (and continue at university) but we may have to look at some of the easier methods to analyse our cosmic ray data. We will hope to find that the simpler methods produce clear results. The day/night analysis we did in Maria's workshop was the very simplest of the "better than rough estimate" tests: valid because we got a very clear result. Sometimes one finds that simple methods do not give a clear yes/no and we have to try something more complicated.

The general rule, therefore, is to use the simplest method that produces a clear yes/no result. There is no point in wasting time doing complicated maths if we do not need to. The best experiments are those that do not need heavy weight statistics in order to understand the results.

Making a good statistics argument is not easy, but it is very important indeed for professional scientists. It is very easy to fool yourself in science. Statistics is all about not fooling yourself.

Calculating the Standard Deviation

Let us assume that we already know that the average value of all the data we collect should be distributed around some mean value, μ. (Bear with me – this is not usually the case, but the explanations are easier this way.)

We further assume that we have a collection of *n* data readings, which we will call x_i where the index *i* will vary from 1 to n and each x_i is a unique reading.

The *variance, v,* of the data around this known mean value, μ, is then given by:

$$V = \frac{1}{n}\sum_{i=1}^{i=n}[x_i - \mu]^2 \qquad (1)$$

This means: "sum over the all squares of the deviations from the mean". The *standard deviation* is then defined to be:

$$SD = \sqrt{V} \qquad (2)$$

The problem is that usually we do not know in advance the value of the mean, μ . We normally also have to estimate this from the data. This estimate is, of course easy:

$$m = \frac{1}{n}\sum_{i=1}^{i=n} x_i \qquad (3)$$

This is our best evidence of the value of the mean, μ, but it is unlikely to be *exactly* the same as the real underlying mean of the random effects controlling the data. Can we replace μ with m in the calculation of standard deviation? Not quite. The fact that m is not the same as the real mean, μ, means that our calculation of the variation is a slight under-estimate. We have to use the equation:

$$V = \frac{1}{n-1}\sum_{i=1}^{i=n}[x_i - m]^2 \qquad (4)$$

We can see this makes some sense, because if we had only one data point that would define the value of the mean but we should not be able to tell anything about the amount of variation in the data. If you want the proper explanation of this, the full derivation of the modified formula, you will have to take maths beyond GCSE. You will find that Microsoft Excel gives both options for calculating the standard deviation. The one you will most likely wish to use is STDEV.S() for standard deviation based on the sample of data. (The other one is STDEV.P() which means using population information – i.e. the known value of the mean.) Note that it does not make much difference if you have very large data sets. It matters for experimenters dealing with, say, less than 20 readings (which often occurs in biology, so biologists *must* take care to use the right formula).

If you are not using Excel, there is a well-known short-cut to calculating V with equation (4) useful when working with a basic calculator and small data sets.

$$V = \frac{1}{n-1}\left[\sum_{i=1}^{i=n}[x_i]^2 - n.\left(\sum_1^n x_i\right)^2\right]$$

This should give exactly the same answer as (4) – but watch out! This sometimes leads to calculating a small difference between two large numbers and there may be big rounding errors.