# Statistically proven? Do you even?

"*It is statistically proven that…*", one of those phrases that *"those who don't understand it use, and those who do, don't"*. One of my biggest pet peeves is hearing people quote this from some article through social media to try and prove a point that they are making. Don't get me wrong, I am not saying that there is a total absence of 'authority' to that phrase. However, if you do want to try and sound 'smart' in front of your peers, I will try my best in this entry to educate you on the underlying implications and models that you think you believe in, every time you use that statement, so you can back your claim up with some substance.

## What is a proof?

To begin, it seems fitting to break the phrase down into its individual words so that we can try to decipher what the actual meaning is from their definitions. If you google the definition of [proof](https://www.google.com.au/search?safe=off&ei=Ow5iWprMNcWk8AXEjquoCg&q=proof+definition&oq=proof+definition&gs_l=psy-ab.3..0l10.28160.31734.0.31910.15.9.0.4.4.0.412.1059.0j2j1j0j1.4.0....0...1c.1.64.psy-ab..7.8.1073...0i7i30k1j0i67k1j0i20i263k1j0i10k1.0.OMte4dPVtL0), you will find that it is defined as evidence or arguments establishing a fact or the truth of a statement. Well, seems simple enough to understand with hardly any ambiguity so far.

## What are statistics?

Also from google, [statistics](https://www.google.com.au/search?q=what+is+statistics&oq=what+is+statistics&aqs=chrome..69i57j69i60l2j69i59j69i60j69i59.1894j0j9&sourceid=chrome&ie=UTF-8) as a noun is defined as the practice or science (art, really) of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

## Putting them together

So, once we put these 2 definitions together into a sentence, it simply means that 'statistical proving' is the act of collecting evidence or arguments from a representative sample to infer some facts or truth about some statements regarding the population? Wait, so if my 'truths' relate only to the smaller sample, does that make that same 'truth' apply for everyone in the exact same way? After all that is the definition of "a fact regarding the population". In my opinion, this oxymoronic nature of the 2 statements when combined together gives people the wrong idea regarding the true underlying meaning of the phrase 'statistically proven'.

## Don't get confused

I am not saying there is no such thing as proofs for statistics. Statistics itself is a broad discipline with many sub-categories such as probability theory, mathematical statistics, stochastic processes and many more, all blurring on the categorical edge of mathematics. However, proofs that are done in a statistics class usually refer to the underlying axioms and some clever algebraic manipulation, which, in essence is actually a mathematical proof. 1 famous example would be the central limit theorem, where the sum of individual distributions can be shown to slowly converge to the normal distribution as the sum tends to infinity. (*A probability distribution function is a mathematical expression that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment*) There is an excellent video showing the convergence of many distributions into the normal distribution found [here](https://www.youtube.com/watch?v=6YDHBFVIvIs).

The formal proof of CLT (*with some assumptions*) is basically a calculus problem which can be [solved](http://mathworld.wolfram.com/CentralLimitTheorem.html) in a few ways, one of which being a Fourier transformation of a gaussian function. Fret not if you did not understand the previous sentence. The point is that the proof relies on the same concepts as why 1 + 1 = 2.

## So, what does “statistically proven” actually mean?

To be frank, it can mean a multitude of things depending on the experiment being done. I will go through the simplest example without including all the jargon, but also highlight the essential characteristics of this concept.

## The premise of the experiment.

In its simplest form, the 'statistical experiment' done can also be called a hypothesis test. This means that I, the tester would make a statement about the whole population that I think is true. For example, I think that 'families in Singapore do not spend $1000 weekly on average'. This means that the opposite argument would be 'families in Singapore spend $1000 weekly on average'. By convention, I would then proceed to gather my data of weekly spending for a smaller group of families in Singapore (*because it is near impossible, or inefficient to collect the data for every single family in Singapore*). Let's say that I managed to collect the data of 5000 families' weekly spending and their average is $800, with a standard deviation (*'average' variation amongst the spending of the 5000 families*) of $200.

## Assumptions that you 'believe' in

Because I am trying to disprove/prove the claim 'families in Singapore spend $1000 weekly on average', I must rely on some sort of framework that would link my sample and the population together. This is where the CLT comes in. I am going to boldly declare that all families in Singapore share identical spending patterns on average and that none of the families affect each other’s' spending at all. (*another way of saying this is that each family’s spending pattern is identically and independently distributed, IID*) This is because, although CLT is in essence a mathematical proof, it requires the underlying assumption of each component in the sum being IID. If you can get past accepting that condition, you can then confidently start to draw the distribution of your population of families, which will look like this:

INSERT PICTURE HERE

Now, because we do not know the standard deviation of the population, we have to make do with the sample standard deviation (more hand-waving) and make some adjustments later on in our calculations. Now, we can add in our observed sample average of $800 in the graph.

INSERT PICTURE HERE

If I continue and do my calculations, I will find that the value of $800 is **approximately** 1 standard deviation (after adjustments) below the average of $1000. This calls for my next assumption, my significance level. The convention for a significance level is 5% (which is the same as a confidence level of 95%), which is totally arbitrary and nonsensical in some non-trivial situations. What this means is that, if I observe my sample average being outside the bounds of the shaded area below (for example $450), I can then say that I have enough confidence to reject the claim that families spend $1000 on average because an average of $450 is significantly lower than my theory of a $1000 average. Who is to say that $800 is not significantly too low as well? No one. In fact, changing the significance level of a test will affect its 'power', which relates to false positives and true negatives. I will not go into any more details on this, but it is similar to the discussion on precision and recall that can be found in the other post. So, in essence, changing some parameter of the test will not change the actual scenario but the way results are interpreted. We humans already have a word for that, and it's called manipulation. If, however, I decide to follow the 5% significance rule blindly, I will finally get to a result of **'I do not have enough evidence to reject the claim that families spend $1000 weekly on average, given a 95% confidence level and assuming CLT applies'**. This statement is perfectly true and precise. The problem only arises when a social media site such as 'Business Insider' posts the article with the title 'Scientists reveal that Singaporean families spend $1000 weekly!' and all my friends that read the title would bring it up as 'it is statistically proven that your family spends $1000 weekly'.

Sun of a beach.

## Another technical issue.

There is 1 more fundamental concept regarding the interpretation of significance levels and confidence intervals. **This section is about to get a tad more technical than the previous ones.** Statistics can be somewhat classified into 2 broad categories:

* Classical statistics
* Bayesian statistics

I will not go into detail the features of both, but the main idea is that classical statistics assumes that the population parameter we are doing tests to find (in this case the average weekly spending) is a deterministic variable, or a constant. On the other hand, a Bayesian model would interpret the population weekly spending as a random variable. The very definition of classical statistics goes against the concept of a 'confidence interval' for the parameter. To put this in perspective, let's say for example we do a statistical experiment to infer the weekly average spending for Singaporean families at a 95% confidence interval. A typical answer in class would look like: The 95% confidence interval for a family's weekly spending is $600 - $1400, with a mean of $1000 for example. So, if according to the classical model the mean is supposed to be a single value, where is all this variation (and thus the confidence intervals) coming from? Turns out, in the classical setting, the term confidence interval actually means 'the frequency of possible confidence intervals that contain the true value of their corresponding parameter'. For example, the size of my interval above is $800 ($1400 minus $600). If I were to choose randomly, many $800 sized intervals through the whole spectrum of my distribution ($0 - $800, $900 - $1700, $500 - $1300, $450 - $1250, $100 $900 ...), 95% of the time, the actual spending (unknown to us) would be contained within that interval. This is not an issue at all for Bayesian models because, as mentioned above, Bayesian models assume that the parameter itself is random, which means that it has its own set variation, where a confidence interval would make sense.

## Finally.

Statistics, despite all the arbitrary inputs and hand-waving is still an extremely powerful and necessary tool in the world for philosophical reasons I may talk about in another post. I hope this gives you a little more insight into what assumptions and implications actually go into the term 'statistically proven'. Perhaps the next time you want to use this phrase to prove your point, trying thinking of:

1. The underlying assumption of the population distribution (not everything is bell-shaped)
2. Does the IID assumption really make sense for that experiment or was it sensible to assume it?
3. What was the significance level (confidence level) used for the claim? Did it make sense?
4. Was the model following a classical or Bayesian framework?
5. What does the confidence interval mean in that case?

If, however you cannot get the answer to all these questions, maybe it's best to educate yourself before saying anything at all.