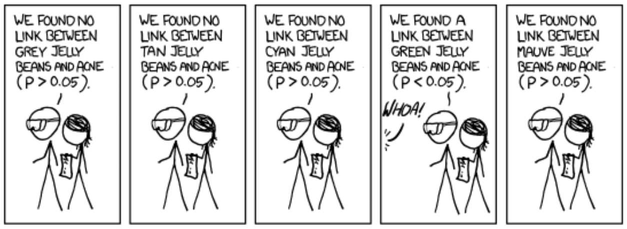
Testing: what is it and why do we care?

Tom Glinnan: Quant for Qual, LSE

October 29, 2018


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).

0.1 So what is testing, anyway?

T-statistics are really hard to get your head around, but they're crucial for understanding how we rule out dumb luck. Quite a few people had questions about how they works so I wrote up these notes for them, but there's **no need to read the whole thing**, just make sure you really get the concept.

I'm going to motivate how testing works with an example from the earliest days of statistics. The inventor of the t-statistic used to work at a beer factory, and where he worked there was a machine which used to deposit beer into (let's say) 500ml bottles. However, one day, he pulled out a bottle for quality control and found that the bottle wasn't quite completely full up. The factory obviously cared about if the bottles were being systematically underfilled: they could be sued, or could lose business from customers who realise they're being shortchanged.

But then he looked at the machine which was depositing beer, and because of the way that machines work, it doesn't spit out exactly 500ml every time. It's sometimes above and sometimes below. The question that came to his mind, and that the company is interested in, is this: was the fact that he pulled out one bottle which was filled to (let's say) 450ml enough evidence that the machine is playing up, or just due to the chance that he pulled out a bottle which the machine happened to put small amounts of beer into? More generally: when do we have enough evidence that the machine is systematically underfilling bottles? If we pulled out 20 bottles and all were filled to 100ml, we'd be sure the machine was broken, but what about if we pulled out 10 bottles which were filled to 495ml on average? This is exactly what it means to have these results by dumb luck... and fortunately for us, we can put a number on how bad this problem is.

The important thing with this area of stats is to get the intuition of how we test whether or not the machine is broken, without talking about maths. The key philosophy is this:

Let's presume that the machine *is* working exactly as we suspect (ie, outputting bottles which are on average 500ml, with some slight variation). Then, let's take some number of beers from the machine at random. *Under the assumption* that the machine is working properly, let's work out the probability that we saw what we just did[1]: in other words, that the bottles were filled in the way that we see that they were. For example, let's say we drew 10 bottles which were all filled to 495ml: we would work out that if the machine was filling bottles to 500ml on average, what would be the probability that our 10 bottles would all be filled to 495ml. If the probability of seeing what we just did, assuming that the machine is working perfectly, is very unlikely (let's say that the probability of this happening is less than one in a million), then our assumption that the machine was working properly was probably untrue.

This takes quite a lot of thinking to get used to, but once you've got it, you've got the essence of how this works: we work out the probability that we saw what we did under the assumption that the world was working as we thought it was. If this probability is really low, then the world probably wasn't working as we thought it was! We think of any numbers (whether Usain Bolt's running times, or people taken into hospital for drug use) as being spit out by some hypothetical machine. Testing helps us answer how this machine works in general, not just for the bits we got to see. This is testing's real power - it allows us to make an inference about how the world *actually exists*, even when we only get to see a small part of it. We use testing all time thing because we can think about pretty much anything as a process where a machine spits out numbers in an analogy to our example: where 'nature' decides a value for us, according to some rules which we use to model the rough situation[2].

All of the components in the beer factory story have technical names which we use when we talk about this in general. The most important ones are:

- The assumption that the machine is outputting beers which are 500ml on average is called the **null hypothesis**. It is the thing that we assume is true in the first place. Another, less formal, name for it, is the thing which we are 'testing'.
- The probability that we saw what we just did is known as the **p-value** of the test[3]. Another way to state how the test works is 'if the p value is very

[1] Very technically, it is the probability that we saw what we just did or something more 'extreme'. However, this is a technical detail which you shouldn't worry about untl you have the main idea nailed down

[2] For example, Usain Bolt's 100m times are very unlikely to be above 13 seconds, and definitely not going to be below 0!

[3] The 'p' stands for 'probability'

low, assuming the null hypothesis, then the null hypothesis is probably not true' - stare at that statement until it makes sense

- The **t-statistic** is the way in which we actually calculate the p-value. To be honest, it's a technical detail and not the most impotant thing to understand in order to get why this works. Intuitively, it compares how far away the average value (eg how full the beers were) of what we observe, to our hypothesis (500ml). This is weighted by how spread out the data is: the question of whether pulling out a beer of 495ml is enough to conclude that the machine is broken depends of course on how accurately the machine fills bottles to 500ml in the first place.

However despite all of this technical terminology, the most important thing is the philosophy of testing itself. Read through it until you have a good feel for what's going on.

0.2 Where are we going with this?

The reason why testing is worth knowing is that it helps us quantify dumb luck - one of the reasons which may cause us to not be able to identify the right causal effect. Let's say that we wish to identify whether A causes B. The fancy tricks that we're going to cover in the next couple of weeks can help us (somewhat) rule out whether B causes A, or whether an external factor C causes both A and B. Let's say that we do that and eliminate these two effects, so we're left with a correlation between A and B. We might be pretty sure that the assosciation isn't due to dumb luck if this correlation is 0.999, but what about if it is 0.1? Is that evidence of a weak correlation between A and B, or is this just dumb luck based on which people ended up taking our survey? In other words, this problem is exactly like our beer machine: it's a question of whether or not we have enough evidence that this correlation actually is 0.1, or whether it's just due to the chance of who we picked. In Econometrics[4] we are often interested in talking about whether something passes the test for being different from 0 or not[5]. Spoiler alert: we call it **statistically significant** if it is, and this will be a major theme in the next few weeks.

0.3 Extra details [not needed for Q4Q]

This section is not needed for Quant for Qual going forward, but helps round off some of the details of testing that we didn't have time to explain. If you want to read it, it'll give you a better view of how testing works and some of the things to be careful of:

[4] The subfield of Economics which is all about establishing causality - what Quant for Qual is all about!

[5] eg if we are confident that a correlation is positive or negative

0.3.1 Significance Levels

When we were doing testing, the whole philosophy is that we say that the null hypothesis probably isn't true if the probability of seeing what we just did see is 'sufficiently small'. But what counts as 'sufficient'? In short: it's up to you! The key thing is that *before*[6] you do the test, you decide what is sufficiently unlikely and then say that the null hypothesis is probbaly false (known as 'rejecting' the null hypothesis') is the p-value is smaller than it. This level that you choose is called the significance level and is denoted by the greek letter alpha (α). By convention, α is usually 1, 5 or 10%. However, it is ultimately down to how sure you need to be that the null hypothesis is false. If, for example, it would cost the beer company in our lots of money to buy a new beer machine, they might want to be *really* sure that it was broken before concluding that it was. This means that they would set α really low - in other words, you would only reject the null hypothesis (and conclude that the machine was broken) if your sample of beers was so incredibly unlikely to happen under the null hypothesis (for example, you took out 5 beers and they were filled to 100ml on average). Being so sure that the null is false is great, but it leads to a problem...

0.3.2 Power

If you make α really low, then you will only conclude that the machine is broken if you are very sure that it is. However, in setting α so low, you run the risk that the machine might still be broken, but you fail to conclude that it is. For example, the machine might be broken and be producing bottles of mean 400ml, but you may conclude that it is not broken because you needed to see bottles filled up to 100ml on average to conclude that it was. There is therefore a tradeoff between the significance level of a test and it's *power* - the test's ability to tell you that the machine is broken when it actually is. This tradeoff is what you use to define your significance level: it is based on how much you care about getting potentially the wrong answer in each direction[7] At this point, you might be wondering if there is anything we can do to improve this tradeoff. This leads us onto the final concept you might have heard of...

0.3.3 One- and Two-Sided Tests

Let's say for argument's sake that the beer company does not care about underfilling beers. Maybe they'e done some market research and found that people don't care when there is slightly less beer in the bottle. However, they still

[6] You have to choose it before because if you chose it afterwards then you could make any test go any way you want, which is not exactly good science!

[7] Some extra details: If you've done some statistics before, you might have heard about *Type I* and *Type II* errors. A Type I error is the probability that you reject the null given that the null is actually true (ie conclude that the machine is broken when it actually isn't); a Type II error is when you don't reject the null when you should have done (ie don't conclude that the machine was broken when it actually was). The significance level is the probability of a Type I error occuring, and the power is defined as the probability of a Type II error *not* occuring. See if you can see why that makes sense.

really care about if they overfill beers as this costs the company money. Is there any way we can use this information to improve the α/Power tradeoff? As you might be able to guess given the fact that I just asked that question, the answer is yes. If we don't care whether the machine is deviating in one direction from the null hypothesis, then we can pretend that the machine is definitely not doing that. This means that for any given α, there is less chance that the machine is actually broken when we conclude that it is probably not[8]. The test where we only care about deviations on *one side* of the null hypothesis is appropriately called a *One-Sided Test*, and the test where we care about both is *Two-Sided*. We say that a one-sided test is more *Powerful* than a two-sided one because it has this higher power for any level of α, so there is not as much of a tradeoff in choosing a lower α. Therefore if you are ever in a situation where you only care about deviations from one side, then use one! More data also makes this tradeoff more bearable.

If you've got to this point then you probably know more about statistics than most people pick up after spending hours of lectures studying. Testing is one of the most important ideas in all of statistics, and you can't do serious data work without understanding how it works. Time to put it all to practice in week 7!

[8] As we pretend that the machine isn't producing bottles which are below 500ml on average, there are fewer places where the broken machine could be and still pass our test!