1: Fundamentals of Null-Hypothesis Significance Testing (Part 1)

Bernhard Angele

7 February 2018

Why are we doing this?

Hello everyone and welcome to our first "real" lecture on Advanced Statistics. In our introductory session, I have already explained to you a bit about my idea of what "advanced" means in this context. In our undergraduate statistics units, we try to give students the absolute minimum in statistical education acceptable. This may surprise you, but given the resistance that many undergraduate students have to statistics (I'm not saying that this is the students' fault – far too many of them – and you –have internalised the misconception that you can't do maths), the lack of mathematical skills (compounded by the resistance to anything mathematical), and the limited time, this is the best we can do. We give you a small toolkit of step-by-step instructions to perform basic statistical tests without making catastrophic mistakes. If we are really good, we give you some degree of understanding of the basic ideas behind null-hypothesis significance testing (NHST). This is the kind of statistical education I have received myself and that you probably received as well (if you didn't receive any statistical education at all, don't worry: you were spared all the downsides of this type of training and can easily catch up on the practical step-by-step parts). This kind of undergraduate education is a necessity – we have to turn out students who can be trusted to do basic experiments and who can understand the statistical tests reported in research papers. However, I don't consider "advanced" statistics to be a continuation of teaching the "click here in SPSS" ritual for more complex tests that few people ever use (such as MANOVA). If you really understand what statistics is about and why we use them, understanding those complex tests (should you ever need them) will not be too difficult. What I want to achieve in this unit is to get you to understand why we do statistics, and how to interpret statistical findings responsibly. This is much more important than knowing how to do a MANOVA in SPSS by heart. Indeed, the current replication crisis in Psychology may be rooted in many cases in misunderstandings about statistics that even established psychologists have: misconceptions and confusions that were established in undergraduate education and that were never corrected even during years of postgraduate study and subsequent research work. These individuals certainly are highly skilled at scientific writing, at experiment design, and at using statistical software, but in some cases, they have "painted themselves into a corner" by consistently misinterpreting their results and by engaging in behaviour that (probably unwittingly) increased the probability of false positives. Now they have built a career on theories and findings that are possibly not replicable and for which maybe much less evidence exists than they thought. Note that this is not scientific fraud. I firmly believe that the individuals concerned performed their research to the best of their ability and conscience. However, if we want to move past this issue as a discipline, we have to improve our understanding of what we are doing when we are doing statistical tests. These first lectures will address this question, and in order to do that, we have to start with the Philosopht of Science. This part of the lecture will be following Zoltan Dienes' book "Understanding Psychology as a Science" fairly closely. I highly recommend that you buy it – it is not very expensive at all.

To sum this introductory part up, as MSc students, you are studying for a science degree. Even if you are not looking to have a career in academia, your employers and colleagues will consider you to have scientific training and will look to you for making science-based decisions. In order to make these decisions responsibly, you MUST understand the ideas and assumptions underlying the methods that we are using in Psychology. To not teach you this would be a disservice to you and the field.

Case study: Power posing

At this point, please read the New York Times article and watch the videos on the Power Posing controversy that I put on myBU. It is just one example of the issues that are really at stake. Then, continue with the rest of this video.

Philosophy of Science

We will start with a fundamental question: What can we know? How can we be sure about something we think we know? We know that our senses and our minds can be tricked, as we have all made errors before. For example, just now I was sure that I had left my phone on the table, when actually it had fallen behind the sofa. If I can be wrong about something this trivial, how can I ever make conclusions about important things, such as how the brain works, or, if you are more practically minded, about whether a patient with anxiety will benefit from a new intervention, or whether a child would be better off being taken out of his or her family? If anything, those latter issues are even more critical than lofty pursuits such as understanding the human mind, as I can always correct a misconception I might have had about the visual system, but a child failed by the system may suffer damage for life. Philosophers have been asking themselves these questions for a long time. As this is a class on statistics, and not philosophy, we will quickly cover the very basics of Karl Popper's ideas, and then jump right into the part about Neyman and Pearson and null-hypothesis significance testing (Chapter 3). I do recommend that you read the first two chapters that deal with Popper, Kuhn, and Lakatos in more detail as well, as their thinking underlies much of how we think about science.

Karl Popper: Falsification and severe tests

Philosophers have been thinking about the question of what we can know for certain, for a very long time. There are some philosophers, like Hume, who thought that really nothing we think we know is certain. Everything could be an illusion. Most philosophers don't go this far. In the early 20th century, the "Vienna Circle" of logical positivists (e.g. Kurt Gödel, Rudolf Carnap, and Carl Hempel) proposed a distinction between statements that are definitions (like "a triangle has three sides"), which are necessarily true, statements that are verifiable through observation of the world ("my desk is three foot tall"), and statements that are unverifiable ("The world is an illusion"). The logical positivists thought that unverifiable statements were essentially meaningless metaphysics, and we should focus on definitions and verifiable statements (i.e. logic, mathematics, and science). How do you verify a statement based on empirical observations? For "my desk is three foot tall", that is easy, you just get a ruler and measure it, but what about generalisations like "all swans are white"? For this, the logical positivists proposed a process called induction: inferring universal rules given particular observations:

> Swan 1 is white
>
> Swan 2 is white
>
> ...
>
> Swan 999 is white
>
> Conclusion: All swans are white?

The problem is that this conclusion isn't necessarily true: There is always the possibility that the next swan you see is black even if the last 999 were white (apparently, this actually happened when the British went to Australia!). Karl Popper was an Austrian philosopher who thought a great deal about this problem. His conclusion was: You cannot actually use inference to prove a theory about the world to be true. However, a single counterexample is enough to prove a theory false! A single black swan causes us to reject the hypothesis that all swans are white. Science is critical discussion: Everyone is always making up new theories and then trying to disprove both their own and others' theories. Theories that cannot be falsified are non-science or metaphysics. Theories that survive criticism (for a while, at least) are not "proven", they are "corroborated".

Falsifiable and non-falsifiable theories

What is a falsifiable theory? Consider psychoanalysis: A psychoanalyst proposes that a patient's fear of flying is caused by an Oedipus complex (i.e. sexual attraction to his mother in his infancy). If the patient confirms this, the psychoanalyst will take this as evidence for his theory. If the patient strongly denies this (which is probably more likely), the psychoanalyst will conclude that the Oedipus complex must be repressed by the patient. The patient's denial is then also evidence for the psychoanalyst's theory. Essentially, psychoanalytic theories are not falsifiable, therefore Popper would say that they are not a scientific.

As Psychologists and scientists, we must ensure that our theories actually yield testable, that is, falsifiable hypotheses.

Dienes (2008) gives a good example (Box 1.5, p. 9):

> Consider the following two factor theory of liking: Factor 1: We are preogrammed to like familiar things (e.g. foods, people, animals, tools, etc.) because our knowledge and skills are likely to apply to them. They are not dagerous, we can deal with them. This, there is a mechanism that automatically makes us like things as we come across them more often. Factor 2: But we also get bored with familiar things, because there is little to learn from them and we have a drive to learn. These two factors act in opposition to each other. So increasing people's exposure to a new thing can
>
> 1. Increase people's liking because the familiarity means it is safe (first factor operating)
> 2. Decrease people's liking because they get bored (second factor operating)
> 3. First increase then decrease liking because the first factor operates initially before boredom becomes stronger.
> 4. First decrease then increase liking becasue boredom operates initially before the first factor becomes stronger.
>
> The theory is a good one because it explains all these outcomes. Discuss.

Think about this example now. We will discuss it later in class.

Degrees of falsifiability

Are there some theories that are easier to falsify than others? Essentially, the more observations can potentially contradict the theory, the more falsifiable is. For example, for many psychoanalytic theories, there is essentially no observation that could possibly contradict it. But also think of Psychology: Often, the most basic hypothesis is a very simple one of difference:

> Participants in Condition A will perform differently from those in Condition B.

There are not very many ways to falsify this hypothesis since the conditions can differ in many possible ways. A hypothesis like

> Participants Condition A will perform better than those in Condition B

is better, since it can be falsified by 50% of the possible outcomes. A hypothesis like

> The performance in Condition A will be 30% better than that in Condition B

is even better, since it can be falsified by a lot of possible outcomes. Similar considerations apply to hypotheses about correlation:

> There is a correlation between Factor A and Factor B

can only be falsified by a very specific outcome, while

> There is a positive correlation between Factor A and Factor B

can be falsified by 50% of possible outcomes.

This issue is directly related to the "multiverse" or "garden of forking paths" issue described by Andrew Gelman in the talk I have asked you to watch. Do your choices in analysing the data make it harder to falsify your hypothesis?

Popper's work is of course applicable to Psychology. However, we face (along with some other sciences) a particular problem in terms of how distinguish between observations that reflect a real property of the world and observations that are merely the result of random, unrelated noise. In some fields like physics, the noise can often be reduced considerably until the signal (the effect we want to observe) becomes quite clear. In contrast, in Psychology, we measure either human behaviour or the correlates of human behaviour. Since humans are extremely complex systems, our effects are at all times influenced by a lot of processes acting at the same time. Therefore, we need to use statistical methods to distinguish between signal and noise. We will work out the logic of null-hypothesis significance testing using the simple dice problem you did in our last lecture.

Neyman and Pearson: Null-hypothesis significance testing

In the 1920s and 30s, Jerzy Neyman and Egon Pearson established a consistent logical procedure for significance testing (although the term "significance" itself comes from Fisher, as does the suggestion of using .05 as the cutoff). To understand what they proposed, we first have to understand about some fundamentals about probability. First, it's important to distinguish between the subjective and the objective interpretation of probability. The subjective interpretation is of probability as a degree of certainty: If I think that there is a probability of .5 that it will rain tomorrow, that would mean that I'm uncertain whether it will rain or not, while a probability of .9 would indicate that I am rather convinced that it will rain. This interpretation is the foundation of Bayesian statistics, which we will talk about later. The objective interpretation of probability is that probabilities are not degrees of conviction, but properties of the world: According to this view, the probability applies to a collective of random events such as a coin toss. If the probability of a coin landing heads is .5, that means that, in the long run, half of the times I will toss a coin it will land heads. This is a property of the coin, not of my conviction about the coin. A singular event has no probability: We cannot say that there is a probability that the null hypothesis is true, because, in our universe at least, the null hypothesis is either true or untrue. There is no randomness about the null hypothesis in this interpretation. On the other hand, Bayesians can talk about the probability of a hypothesis being true without any problems, as to them, "probability" just means their degree of conviction about something. The problem arises when we start confusing the two. In null-hypothesis significance testing, we always have to use the objective interpretation.

The basic logic of hypothesis testing

Remember, according to the objective interpretation of probability, a hypothesis is either true or false, depending on the true properties of the world. In making conclusions about the true properties of the world based on our data, we can either be correct or we can be wrong. We cannot avoid making errors sometimes, but the basic idea behind NHST is to define decision rules that will ensure that we can control the error rate so that it is acceptable to us. In order to do this, we set up two hypotheses, a null hypothesis and an alternative hypothesis. Usually, the null hypothesis is a hypothesis of no difference or no relation, but it doesn't have to be. Hypotheses are defined in terms of populations, while our data is composed out of a sample from the larger population. Numerical population properties are called parameters and are usually specified using greek letters such as $/mu$ for the mean and $/sigma$ for the standard deviation, while sample properties are called statistics and are symbolized by Roman letters. The idea is, then, to simulate running an experiment given that the null hypothesis is true, and to determine how unusual or "extreme" the relevant sample statistic we observed is given that the null hypothesis is true. We can set a cutoff for the sample statistic at which we decide to reject the null hypothesis. For example, if we decide to set the cutoff at a sample statistic that would only be observed 5% of the time if the null hypothesis were true, this would result

in us rejecting the null hypothesis falsely 5% of the time. This is called the /*alpha*. Fisher suggested setting this at .05 so that only 1 in 20 rejections of the null hypothesis would be erroneous. Of course, we can also commit an error by NOT rejecting the null hypothesis even though it is actually false. We will talk about that next week when we consider statistical power. Now, we will work through an example of NHST based on your dice experiment. Don't be too taken aback by the maths – they really only symbolise I have just explained.