Research Methodology

Part 2: Quantitative Methods and Inferential Statistics

Christian Hilbe, christian.hilbe@it-u.at
Version January 28, 2026

1 Introduction and motivation

Remark 1.1 (**A brief summary of yesterday's session)**
Among other things, Jie introduced you to the the scientific method, a scheme of how research should be done based on testable hypotheses. It follows the following steps:

1. Define a general question
2. Gather information and resources (e.g., literature review, theory building)
3. Form a hypothesis
4. Perform an appropriate experiment and collect data
5. Analyze the data to derive the results of your study
6. Interpret the results and draw conclusions (this may serve as a starting point for a new hypothesis)
7. Publish results

Today we will focus on steps (3) – (5); how to test hypotheses and how to draw conclusions from data.

Remark 1.2 (**What to take away from this class)**

1. *Practical relevance.* Primarily, today's class should make you (more) familiar with some key quantitative concepts. In particular, it should help you understand the jargon that is used when doing experimental/empirical research, such as:

   > *Hypothesis testing; p values; statistical significance; false positives and false negatives; statistical power; confidence intervals*

   Those working with empirical data should be able to apply these concepts to their own work.

2. *Allowing you to be an informed consumer of science.* It may well be that some concepts you will hear about have no immediate relevance for your own research. Yet, as a scientist, and as a consumer of science, it is still important to have an overview on different methods that allow people to generate knowledge. This comment seems to be particularly important at an interdisciplinary university as ours. Knowledge of each others' methods also makes it easier to start collaborations.

3. *Philosophical aspects.* In courses like this, it is useful to stress that the concepts that will be taught are human constructs. In this case, they are even relatively recent constructs. For example, modern

hypothesis testing and many of the associated statistical methods have only been developed a bit more than a hundred years ago (e.g., Galton, Fisher, Pearson all lived around the turn of the previous century). Even the idea that (controlled) experiments are a viable way to generate knowledge about the world is fairly recent (one early pioneers is Robert Boyle who discovered that at a given temperature, the volume of an ideal gas and the respective pressure are inversely related. This was in the 17th century).

4. *Gets you thinking.* Even if you are already familiar with the concepts of hypothesis testing, perhaps the class makes you ponder about some aspects in more detail. For example: Why is it the case that we usually first formulate a Null hypothesis which describes *the opposite* of what we are hoping to find? Or: Why do larger sample sizes help? Or: Why is it sufficient to sample a few hundred people to make sweeping statements about large populations?

Discussion 1.3 (**Some introduction round**) 1. To which extent are people comfortable with mathematics? To which extent do people know probability theory or inferential statistics? To which extent do people know, say, the *central limit theorem*?

2. For the different student groups, could you briefly describe your ongoing projects? In particular, do they test specific hypotheses, and if so, which? What kinds of (descriptive or inferential) statistics have you been using?

Remark 1.4 (**Plan for this class**)
This class will give you a brief overview on the two main areas of inferential statistics, hypothesis testing (first) and estimation theory (by the end).

2 Basics of hypothesis testing

Remark 2.1 (**Basic idea of inferential statistics**)
When collecting data for scientific research, you often like to draw inferences beyond the specific participants that happened to be in your sample. For example, you might want to know people's voting preferences from asking a sample of $n = 500$ Austrians. Or you may want to know the efficacy of some new medical treatment, after having treated $n = 500$ patients. In each case, you would like to make statements that go beyond those 500 people: statements that apply to the entire Austrian population, or to all future patients that may get this treatment. We would thus like to ask: How is it possible that by asking just 500 people, you can draw inferences about 6.3 million people? (all Austrians eligible to vote). Or more generally, to which extent can the results of a random sample represent the whole population? To this end, it is useful to ask the converse question: How likely is it that your random sample is misleading, by producing very extreme or unrepresentative results?

Example 2.2 (**Illustration of the problem: Estimating happiness**)
Suppose you have the suspicion that Austrians are, on average, are happier than Germans. So you recruit a representative sample of 100 Austrians and 100 Germans, and ask them about to which extent they are satisfied with their everyday life. Participants who answer 'Fully satisfied' or 'Rather satisfied' are classified as happy. Now suppose with this method, you observe 57 happy Germans in your sample, and

68 happy Austrians. This is how the situation represents itself to you:

Population level	Within your sample
German proportion p_D	Proportion in German sample: $f_D = 57/100$
Austrian proportion p_A	Proportion in Austrian sample: $f_A = 68/100$
(This is what you would like to compare)	(This is what you have)

To interpret this situation, keep in mind there are two possible ways how the data of your sample could have come about:

1. In reality, Austrians are indeed more happy than Germans (i.e., $p_A > p_D$), and your sample correctly captures this trend.

2. On average, Austrians are just about as happy, or even less happy than Germans (i.e., $p_A \leq p_D$). It just so happened that when taking our sample, we randomly happened to pick more of the happy ones.

Idea: If we can compute how likely a case like (2) can happen, and if that case would be very unlikely, this would support interpretation (1).

Remark 2.3 (**Hypothesis testing)**
Hypothesis testing is a procedure to evaluate empirical evidence.

1. You start by formulating two competing statements: A Null hypothesis (H_0) suggesting there is no effect, and an alternative hypothesis (H_1) suggesting that there is some effect [as a researcher, you typically 'hope' H_1 to be true]
2. You sample data and you compute a *test statistic* (a function from your measured sample values)
3. Under the assumption that the null hypothesis is true, you know the distribution of the test statistic. This allows you to compute how likely it is that you obtain a value as extreme or even more extreme as in your sample (this likelihood is the p *value*
4. If this likelihood is below your chosen significance threshold α, you 'reject' H_0 (typical significance thresholds are $\alpha = 0.05$, $\alpha = 0.01$, or $\alpha = 0.001$).

Example 2.4 (**Revisiting happiness)**
Suppose you would like to use that scheme to address the question whether Austrians are indeed happier than Germans. In that case,

1. As the null hypothesis, we assume equality, $H_0 : p_D = p_A$. The alternative hypothesis is then $H_1 : p_D \neq p_A$.
2. After doing the sampling you observe $x_D = 57$ happy Germans and $x_A = 68$ happy Austrians, based on two samples of size $n = 100$ each. By looking at your favorite statistics source (or the internet), you learn that a relevant test statics to compare proportions is

$$z = \frac{f_D - f_A}{\sqrt{2f(1-f)/n}},$$

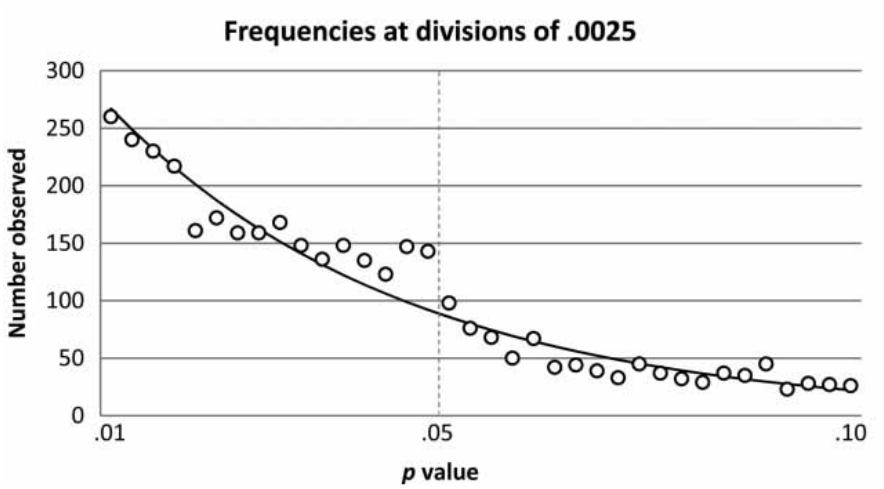

Figure 1: According to Masicampo & Lalande (Q. J. Exp. Psych. 2012), the published literature in three psychology journals shows a curious prevalence of p values just below 0.05.

where $f_D = x_D/n$ is the proportion in the German sample, $f_A = x_A/n$ is the proportion in the Austrian sample, and $f = (x_D + x_A)/(2n)$ is the pooled proportion. In our case, $z = -1.6067$.

3. Under the null hypothesis, this test statistic follows a standard normal distribution. For this distribution, the likelihood to have a z-value as extreme as this one is $p = 0.1074$ (using a *two-tailed test*).
4. Because this p value is not below $\alpha = 0.05$, you are not sufficiently confident that the result could not have happened by chance alone. Hence, you do *not* reject the null hypothesis.

Remark 2.5 (**Which statistical test – which test statistics – to use?**)

1. In our case, the test statistics essentially is due to the central limit theorem – for large n, binomial distributions can be approximated by normal distributions, and the difference of two normal distributions is normally distributed. Hence, if we normalize appropriately, we get a standard normal distribution.
2. In general, however, there are many different statistical tests (test statistics) out there. Which one applies to your case depends on what kind of data you have (e.g., ordinal vs metric) and what kinds of quantities you compare (in our case: proportions).
3. However, the general scheme outlined above is always the same, just the proper test statistics differs.

Discussion 2.6 (**Wait a minute**)

1. What can we learn from a result like $p = 0.1074$? (e.g., does it imply the Null hypothesis is true)?
2. Why do some fields have much stricter significance levels? For example, in physics, researchers only declare a discovery if $p < 5\sigma$, meaning $p < 3 \cdot 10^{-7}$.
3. Why do we need to invoke a null hypothesis in the first place – isn't that a bit roundabout?

Remark 2.7 (**A critical evaluation of the p-value**)

The p value needs to be interpreted with caution. In particular:

1. Just because an effect is statistically significant does not mean the effect is strong (e.g., a weight-loss injection that reduces weight by 0.1kg in all 100 participants of a sample)
2. If we test multiple hypotheses in parallel within a study, the appropriate significance level α needs to be corrected appropriately, to avoid spurious results (*Bonferroni correction*)
3. The current system treats $p = 0.049$ and $p = 0.051$ very differently $\rightarrow$ there is a detectable over-representation of statistical results that barely happen to satisfy $p < 0.05$.
4. Because p values smaller than 0.05 are the gold standard, there are incentives to p-hack (e.g., by stopping participant recruitment once an effect happens to be significant, or by introducing arbitrary data exclusion criteria ex post).

Remark 2.8 (**Ways to be wrong**)
Based on the above-sketched method of hypothesis testing, there are two possible ways to be wrong:

		Population reality	
		H_0	H_1
Decision based on test	H_0	correct	Type II Error *False Negative*
	H_1	Type I Error *False Positive*	correct

The likelihood of a Type I error, we can control; it's given by the significance threshold α we use. By reducing α (i.e., by being more *conservative*), we can make the corresponding risk arbitrarily small. However, doing so tends to increase the risk of a Type II error (often denoted by β).

Remark 2.9 (**Power**)
Statistical power is the probability $1-\beta$ of detecting an effect given it is there (i.e., the probability of rejecting H_0 given that H_1 is in fact true). For fixed α and a given estimated effect size (e.g., given by the literature or through preliminary trials), one can calculate the sample size n required to have a desired power $1-\beta$. A common threshold used for the respective *power analysis* is $\beta = 0.2$ (i.e., if the alternative hypothesis is true, we want to detect it with 80% probability). [Provide a graphical sketch of how to illustrate power, and how it changes with increasing sample size]

Discussion 2.10 (**What does it mean to be underpowered?**)
One can sometimes read that a 'study was underpowered' or 'overpowered'. What does it mean and what might be problematic about it?

3 Estimation theory and confidence intervals

Remark 3.1 (**Estimation theory**)
In addition to hypothesis testing, another aim of inferential statistics is to provide parameter estimates. Estimation theory aims to provide estimators (function that take the sample values as an input, and give a

parameter estimate as the output). The aim is to derive estimators that have certain appealing properties. For example, estimators might be required to be 'unbiased' (i.e., in expectation, the estimator gives the correct parameter). As an example, an unbiased estimator for the proportion of happy Austrians is to take the proportion of happy Austrian participants in the sample $\hat{p}_A = f_A$. One common method to derive estimators is the *maximum likelihood estimation*. It asks: Among all possible parameter values, which is the value that is most likely to produce the realized sample?

Remark 3.2 (**Confidence interval)**
The above estimator for proportions produces a *point estimate*. In science, however, researchers are often required to quantify the uncertainty in their estimates, and to provide an *interval estimate* instead. For example, a 95% confidence interval is a range that would contain the true value in 95% of repeated samples (in practice, we interpret it as the range of plausible values). In Figures, confidence intervals are often indicated as *error bars.*

Example 3.3 (**Confidence interval of a proportion)**
When estimating confidence intervals of a proportion, one can again use the fact that the binomial distribution can be approximated by a Normal distribution. As a result, a standard method to obtain a 95% confidence interval for a proportion is to take $\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Note that for $\hat{p}=0.68$ and $n=100$, we obtain a confidence interval of [0.589,0.771].

Discussion 3.4 (**Bonus Discussion)**

- Why do larger sample sizes help? Or: Why is it sufficient to sample a few hundred people to make sweeping statements about large populations?
- Also, what is it that is approximated by a Normal Distribution in the Central Limit Theorem? For example, income in Austria is not normally distributed; yet when we apply the Central Limit Theorem to the income distribution in Austria *something* is normally distributed; what's this *something*?