Übung Ultimate attainment in SLA

Statistics 01

Laura Becker

FAU Erlangen-Nürnberg

SS 2020

Research

This is how research usually (ideally?) works:

❶ you observe something interesting

↓

❷ this gives you an interesting idea

↓

❸ you come up with a theory to explain your observation

↓

❹ you formulate your hypothesis

↓

❺ you design a study/experiment to test your hypothesis (what and how?)

↓

❻ you collect data

↓

❼ you analyse the data you collected:
(description, explanation, prediction)

Formulating a hypothesis

A scientific hypothesis that you can use for research has to fulfill the following criteria:

- it is a general statement that is concerned with more than just a singular event
- it is a statement that at least implicitly has the structure of a conditional sentence (*if …, then …*, or *the …, the …*) or can be paraphrased as one
- it is potentially falsifiable, which means it must be possible to think of events or situations that contradict the statement
- it is testable (for practical, financial, ethical, etc. reasons)

Formatting your data

- In order to actually test your hypothesis, you need to think about what to measure and how.
- This means, you also need to think about your data structure.
- Keeping your data organized in a good structure will make your life **A LOT** easier later!
- Usually, a good way to structure your data is the following:
  each observations gets its own line, with all the relevant variables in columns

	var 1	**var 2**	**var 3**	...	**var n**
obs 1	0	A	1.5	...	23
obs 2	1	B	1.7	...	40
obs 3	1	A	0.9	...	55
...	...	...	...	...	...
obs n	0	C	0.3	...	62

Types of variables

What are variables?

- some object that can vary (i.e. take different values)
- In research, we usually want to associate an outcome (variable) with another, independent predictor (variable).
- Those two types of variables are usually referred to as:
  - predictor / independent variable
  - response / dependent variable

What types of variables can we distinguish?

- categorical
  - nominal / categorical variables (green, yellow, blue)
  - binary variables (often: 1 “presence of X” vs. 0 “absence of X”)
  - ordinal variables (strongly disagree < disagree < neutral < agree < strongly agree)
- numerical
  - interval with arbitrary zero point (5, 20, 25, ...)
  - ratio with meaningful zero point (1,2,5,10, ...)
  - continuous vs. discrete (1.34, 0.4241, vs. 2, 1)

Experimental designs

For a psycholinguistic experiment this roughly translates into *who does what*?

between-subjects design / independent-measures design

- individuals see only one of the possible levels of an experimental condition

within-subjects design / repeated-measures design

- every individual sees each of the experimental conditions consecutively, and their responses to each level are measured

counterbalancing

- randomizing or reversing the order of conditions among subjects
- this is useful in a repeated-measures design to ensure that the order of conditions does not influence the results of the experiment

Between-subjects design

- Let's imagine a toy study in which we want to find out whether modality (written vs. spoken stimuli) has an effect on grammaticality judgments of L2 speakers of English.
  - our participants are 4 L2 speakers of English
  - the items are 4 English sentences: 2 grammatical, 2 ungrammatical ones
  - we test them in either the written or the spoken modality

N obs	participant	item	modality
1	part_1	item_1_G	written
2	part_1	item_2_G	written
3	part_1	item_3_U	written
4	part_1	item_4_U	written
5	part_2	item_1_G	written
6	part_2	item_2_G	written
7	part_2	item_3_U	written
8	part_2	item_4_U	written
9	part_3	item_1_G	spoken
10	part_3	item_2_G	spoken
11	part_3	item_3_U	spoken
12	part_3	item_4_U	spoken
13	part_4	item_1_G	spoken
14	part_4	item_2_G	spoken
15	part_4	item_3_U	spoken
16	part_4	item_4_U	spoken

Within-subjects design

- Let's modify our toy study:
  - to make things simple, our participants are 2 L2 speakers of English
  - the items are 4 English sentences: 2 grammatical, 2 ungrammatical ones
  - we test both participants in both the written and the spoken modality, but we switch the order ($\rightarrow$ counterbalancing)

N obs	participant	item	modality
1	part_1	item_1_G	written
2	part_1	item_2_G	written
3	part_1	item_3_U	written
4	part_1	item_4_U	written
5	part_1	item_1_G	spoken
6	part_1	item_2_G	spoken
7	part_1	item_3_U	spoken
8	part_1	item_4_U	spoken
9	part_2	item_1_G	spoken
10	part_2	item_2_G	spoken
11	part_2	item_3_U	spoken
12	part_2	item_4_U	spoken
13	part_2	item_1_G	written
14	part_2	item_2_G	written
15	part_2	item_3_U	written
16	part_2	item_4_U	written

Populations and samples

Population

- a group that represents all objects of interest
  (e.g. **all** native speakers of a given language, or **all** L2 learners of a given language)

Sample

- in probably most cases, we cannot collect and analyze the data from an entire population

☞ we collect and analyse data from a representative subset of the population

- ideally, the observed distribution from the sample will also hold for the real, underlying distribution in the population

Statistic(s)

- a set of techniques and tools for describing and analysing data

⚠ a statistic (SG) is a measure obtained from the sample, e.g. the average score of L2 speaker participants in a grammaticality judgment task

- **descriptive**: summarizes some characteristics of the sample
- **inferential**: allows us to use the characteristics of a sample in order to draw conclusions about the population

Null hypothesis significance testing (NHST)

Null and alternative hypothesis

❓ Why do we need a null hypothesis?

- We actually do not need a null hypothesis.
  (☞ Bayesian statistics is an alternative to that)

- However, null hypothesis testing is one of the two most commonly used approaches in statistics.

- It uses statistic measures that cannot prove the actual hypothesis, but it can be used to reject the null hypothesis

Null and alternative hypothesis testing

The idea is the following:

- you formulate a null hypthesis that states the opposite of what your actual hypothesis states
- your actual hypothesis is the so-called alternative hypothesis
- you assess how likely your data is based on the null hypothesis
- if the the data is very unlikely (<5%), we (the scientists) agree that you can reject the null hypothesis, which is supporting evidence in favor of your actual hypothesis

☞ you can accept the alternative hypothesis

⚠ if the data is not very unlikely (>5%) you cannot reject the null hypothesis, but that does not mean that you could accept the alternative hypothesis

- This is a counter-intuitive but crucial point that researchers often ignore or forget:
- If you do null hypothesis testing, your test statistic does **not tell you anything about the probability of the alternative hypothesis directly** given your data.

4 steps for null-hypothesis testing

1. You define a significance level $p_{critical}$, which is usually set to 0.05 (5%). This represents the threshold value for rejecting or retaining H_0.
2. You analyze your data by computing an effect e using some test statistic that makes sense for your data.

   This is where the p-value comes in:
3. You compute the so-called probability of error p how likely it is to find e or something that deviates from H_0 even more in your sample when, in the population, H_0 is true.
4. You compare $p_{critical}$ and p and decide:

- If $p < p_{critical}$ (0.05), then you can reject H_0 and accept H_1

☞ your result is statistically significant

- Otherwise, your result is statistically not significant and you must retain H_0 (which does not help you at all with your alternative hypothesis, the one you are actually interested in).

Important interim summary

What some researchers think that null hypothesis testing and p-values show:

- the probability of the research ("alternative") hypothesis given the data $p(H|D)$

What null hypothesis testing and p-values **actually** show:

- **the probability of the data given the null hypothesis** $p(D|H_0)$

- "The p-value shows the **probability** of **obtaining** a given test **statistic value** or more extreme values if the **null hypothesis** is **true**." (Levshina 2015: 11)

- "When we collect data to test theories we have to work in these terms: we cannot talk about the null hypothesis being true or the experimental hypothesis being true, we can only talk in terms of the **probability** of **obtaining a particular set of data** if, hypothetically speaking, the **null hypothesis** was **true**." (Field et al. 2012: 28)

Significance levels

Early Fisher (1935)

- The level of significance has to be determined *before* conducting a test (in the sense of a convention, e.g. $\alpha = 5\%$). Thus, the level of significance is a property of the *test*.

Late Fisher (1956)

- The exact level of significance has to be calculated *after* a test is conducted (p-value). Here, the level of significance is a property of the *data*. An arbitrarily determined convention is no longer required.

Neyman and Pearson

- α and β have to be determined *before* conducting a test. α and β are the relative frequencies of an error of the first or second kind and are therefore properties of the *test*. Yet, to determine α and β no convention is required, but rather a cost-benefit estimation of the severity of the two kinds of error.

(Haller & Kraus 2002)

Null-hypothesis testing vs. Bayesian statistics

Bayesian statistics allow us to do more intuitive hypothesis testing:

- it shows the probability of the research hypothesis given the data and prior assumptions $p(H|D)$
- it allows for a much more intuitive interpretation and tests directly what we want to test anyway
- however, even though the methods are old, the computation of many Bayesian statistics is computationally very demanding and was not possible for practical reasons until very recently

Type I & type II errors

Type I error: false positive
Type II error: false negative

	true state of the world	
	H_0 is **false**	H_0 is **true**
reject H_0	correct decision	Type I error
retain H_0	Type II error	correct decision

What statistics can and cannot do

"All models are wrong, but some are useful." (Box 1978)

"Correlation does not imply causation."

What statistics can do

- ☺ statistics can help us to quantify and test scientific hypotheses
- ☺ it provides the tools to test a hypothesis on a sample and allows us to measure our certainty about whether or not the hypothesis also holds for the entire population

What statistics cannot do

- ☹ interpret or make sense of results
- ☹ We as humans / researchers have to think about the result, about what they really mean.
- ☹ If we find a statistically significant association between our predictor and the outcome, is the predictor really the cause? Or are these two variables just correlated by chance? Is there a third, confounding factor that we may have overlooked?