# *Quantitative methodology: Inferential statistics 101*

*Jan Vanhove (jan.vanhove@unifr.ch)*

*30 September 2019*

This handout explains the basic logic behind $p$-value-based inferential statistics. It does so by explicitly linking the computation of $p$-values to the random assignment of participants to conditions in experimental research. If you have ever taken an introductory statistics class, chances are $p$-values were explained to you in a different fashion, presumably by making reference to the Central Limit Theorem. The way they are explained here, however, has the advantage that it is less math-intensive while permitting one to illustrate several key concepts about inferential statistics.[1]

[1] Ähnliche Erklärung auf Deutsch: `http://janhove.github.io/statintro.html`, Kapitel 12.

The goal of this handout is for you to **understand conceptually** what statistical tests attempt to achieve, not for you to be able to use them yourself. As a matter of personal opinion, statistical tests are **overused** in quantitative research. In your own research, focus on **describing** your data (e.g., by means of appropriate graphs) rather than running umpteen significance tests.

## *An example: Does alcohol intake affect speech rate in a second language?*

*Research question*: Does moderate alcohol consumption affect verbal fluency in an L2?

*Method*: Ten students (L1 German, L2 English)[2] are **randomly** assigned to either the control or the experimental condition (five each); they don't know which condition they're assigned to. Participants in the experimental condition drink one pint of ordinary beer; those in the control condition drink one pint of alcohol-free beer.

[2] Ten participants is obviously a very low number of participants, but it keeps things more tractable here.

Afterwards, they watch a video clip and relate what happens in it in English. This description is taped, and two independent raters who don't know which condition the participants were assigned to count the number of syllables uttered by the participants during the first minute. The mean of these two counts serves as the verbal fluency/speech rate variable.

*Results*: The measured speech rates are shown in the plot on the right. On average (mean), the participants in the *with alcohol* condition uttered 4.3 syllables/second, compared to 3.7 syllables/second in the *without alcohol* condition.

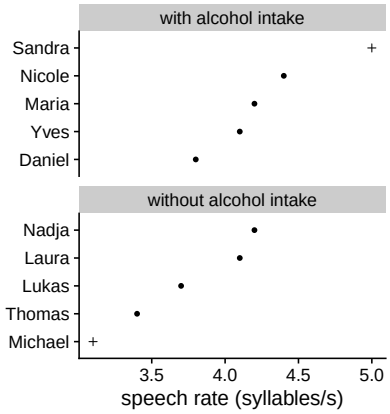


Figure 1: Individual results of a randomised experiment.

### *The basic question in inferential statistics*

While we found a mean difference between the two conditions (4.3 vs. 3.7), this difference could have come about through **chance**. We are, then, faced with two types of accounts for this mean difference:

- **Null hypothesis** (or $H_0$): The difference between the means is due *only* to chance.
- **Alternative hypothesis** ($H_A$): The difference between the means is due to chance *and* systematic factors.

*Assuming the $H_0$ is true*, the participants' results aren't affected by the condition (alcohol vs. no alcohol) they were assigned to. For instance, Sandra was assigned to the *with alcohol* condition and her speech rate was measured to be 5.0, but had she been assigned to the *without alcohol* condition, her speech rate would also have been 5.0. Assuming the $H_0$ is true, then, the difference between the two must be due solely to the random assignment of participants to conditions, due to which more fluent talkers ended up in the *with alcohol* condition. Another roll of the dice could have assigned Sandra to the control condition instead of Michael, and since under the $H_0$, the speech rate of neither is influenced by the condition, this would have produced a slower speech rate in the *with alcohol* condition than in the *without alcohol* one (3.9 vs. 4.1; see Fig. 2).

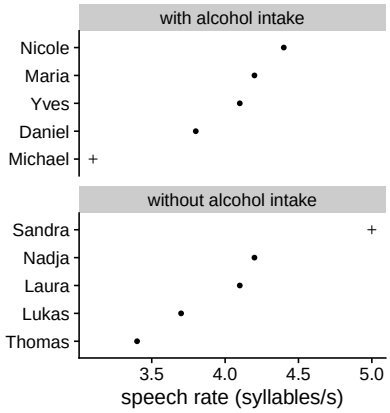


Figure 2: If only chance were at play, Michael's (3.1) and Sandra's (5.0) results would be unaffected by the experimental condition and the outcome might equally well have looked like this (swapping Michael and Sandra).

**Frequentist inferential statistics** seeks to quantify how surprising the results would be if we assume that only chance is at play. To do so, it attempts to answer the following key question: *How likely is it that a difference at least this large would've come about if chance alone were at play?*

If it's pretty unlikely that chance alone would give rise to at least the difference observed, then this can lead one to revisit the assumption that the results are due only to chance—perhaps some systematic factors are at play after all. By tradition, the threshold between 'pretty likely' and 'still too likely' is 5%, but there is nothing special about this number.[3] Before discussing this further, let's see how you can compute how often you would observe a mean difference of at least $4.3 - 3.7 = 0.6$ if chance alone were at play.

### *Testing the null hypothesis by exhausitive re-randomisation*[4]

With 10 participants in two equal-sized groups, there were 252 possible assignments of participants to conditions, each of which was equally likely to occur. To see how easily a difference as least as large as the one observed (4.3 vs. 3.7) could occur due to random assign-

[3] If the $p$-value associated with a difference is below this threshold, the difference is said to be 'statistically significant'. This is just a phrase, however, and arguably a poorly chosen one: statistical 'significance' doesn't tell you anything about a result's practical or theoretical import.

[4] Further reading: Blog post *Explaining key concepts using permutation tests*: goo.gl/o7zFHx. Blog post *A purely graphical explanation of p-values*: goo.gl/UcO1mQ. Stats booklet `http://janhove.github.io`, chapter 12 (in German).

ment alone, we can re-arrange the participants' speech rates into each of these 252 combinations and see for each combination what the difference between the *with* and *without alcohol* condition means is (Figure 3).

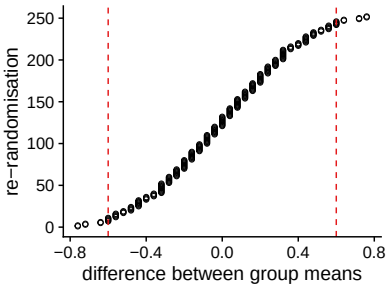


Figure 3: There are 252 different ways in which the 10 participants could have been split up into two groups. The are the differences between the means for all 252 possibilities.

In 22 out of 252 cases, the re-arrangement of participants to conditions produced a difference at least as large in magnitude (positive or negative) as the difference we actually observed. In other words, the probability with which we would observe a difference of at least 0.6 between the conditions if chance alone (random assignment) is at play is $\frac{22}{252} = 0.087$ (8.7%). This is the infamous ***p*-value**.

Since $p = 0.087 > 0.05$, one would typically conclude that a difference of 0.6 or more is still likely enough to occur under the null hypothesis of chance alone and that, hence, there is little need to revisit the assumption that the results may be due to chance alone. **Crucially, this doesn't mean that we have shown $H_0$ to be true.** It's just that $H_0$ would account reasonably well for these data.

## *On 'rejecting' null hypotheses*

Researchers will often say that they 'reject' the null hypothesis in favour of the alternative hypothesis if $p < 0.05$. While this practice is subject to often heated debate, it's important to realise that $p$ can be $< 0.05$ even if the null hypothesis is true,[5] and that $p > 0.05$ even if the alternative hypothesis is true. Consequently, researchers who are in the business of 'rejecting' null hypotheses can make two types of errors, depending on whether $H_0$ or $H_A$ is actually true.

[5] In theory, in fact, $p < 0.05$ in 5% of the studies in which $H_0$ actually is true by definition. In practice, however, things aren't so simple. We'll return to this in a later class.

| | $H_0$ is actually true | $H_A$ is actually true |
|---|---|---|
| $p > 0.05$ | Fine—we didn't reject $H_0$ | Wrong conclusion |
| $p < 0.05$ | Wrong conclusion | Fine—we rejected $H_0$ in favour of $H_A$ |

On the basis of a single study, we can't really know whether '$p < 0.05$' represents an error or a true finding. **(!)**

Note, furthermore, that the $H_A$ stipulates that the results are due to a combination of chance and systematic factors. It doesn't stipulate *which* systematic factors, though. What we would like to conclude is that the main systematic factor at play is our experimental manipulation, but expectancy effects and the like are also systematic factors. What is more, the experimental manipulation may exert a systematic effect on the results, but for different reasons than we think they do.[6]

[6] For instance, a systematic difference between the *with* and *without alcohol* conditions needn't be due to alcohol intake per se but may be related to the taste of the beers in question instead.

## *Analytical short-cuts*

Exhausitive re-randomisation is cumbersome for larger samples and more complex research designs; analytical short-cuts (e.g., the *t*-test,

$\chi^2$-test, ANOVA etc.) and their generalisations usually produce similar results and are used instead. The $p$-values etc. that these procedures return have the same interpretation and are subject to the same caveats as those above.

### *Statistical power*

A study's statistical power is the probability with which its significance test will yield $p < 0.05$. In studies in which one group is compared to a different group, this probability depends on three factors:

1. The size of the difference in the outcome that the intervention causes. Even if the intervention does *not* cause a difference, it is possible to obtain a statistically significant difference due to chance (see table above).
2. The number of observations.
3. The variability in the outcome variable within each group.

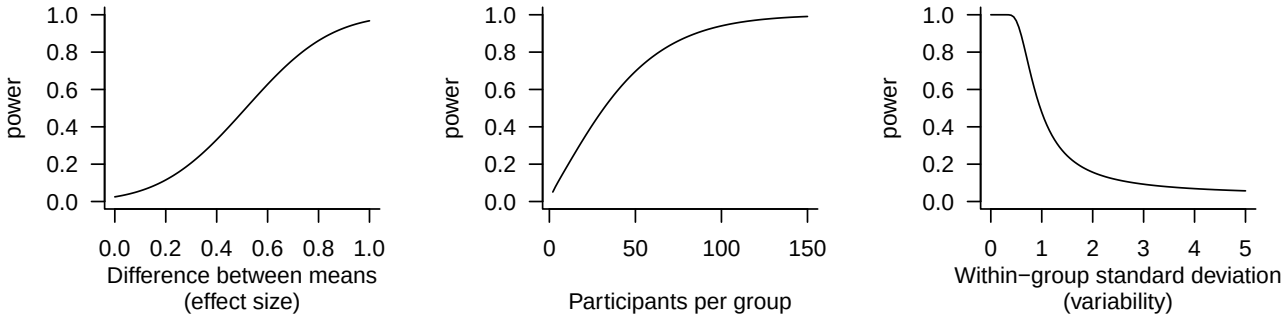


Figure 4: These three graphs show how the statistical power of a study varies with the effect size (left), the number of observations per group (middle) and the variability in the outcome variable within each group (right). The precise numbers along the y-axis aren't too important; what's relevant is the direction and the shape of the curves.

### *Exercise*

Take a look at the graphs above and answer the following questions:

1. How do the effect size, the number of observations and the within-group variability in the outcome affect the probability that a study will yield a statistically significant result?

2. Other things equal, what yields a greater improvement in a study's power: 10 additional participants per group when each group already consists of 10 participants, or 20 additional participants per group, when each group already consists of 50 participants?

3. How could researchers reduce the within-group variability in the outcome variable?