UNIVERSITÄT LEIPZIG

erc European Research Council

What is statistics?

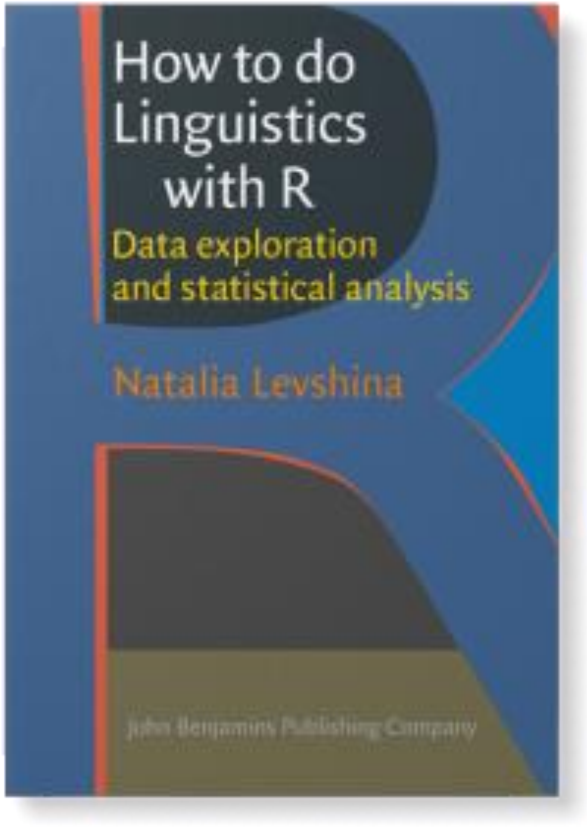
Natalia Levshina © 2019

Tallinn University
May 14-18 2019

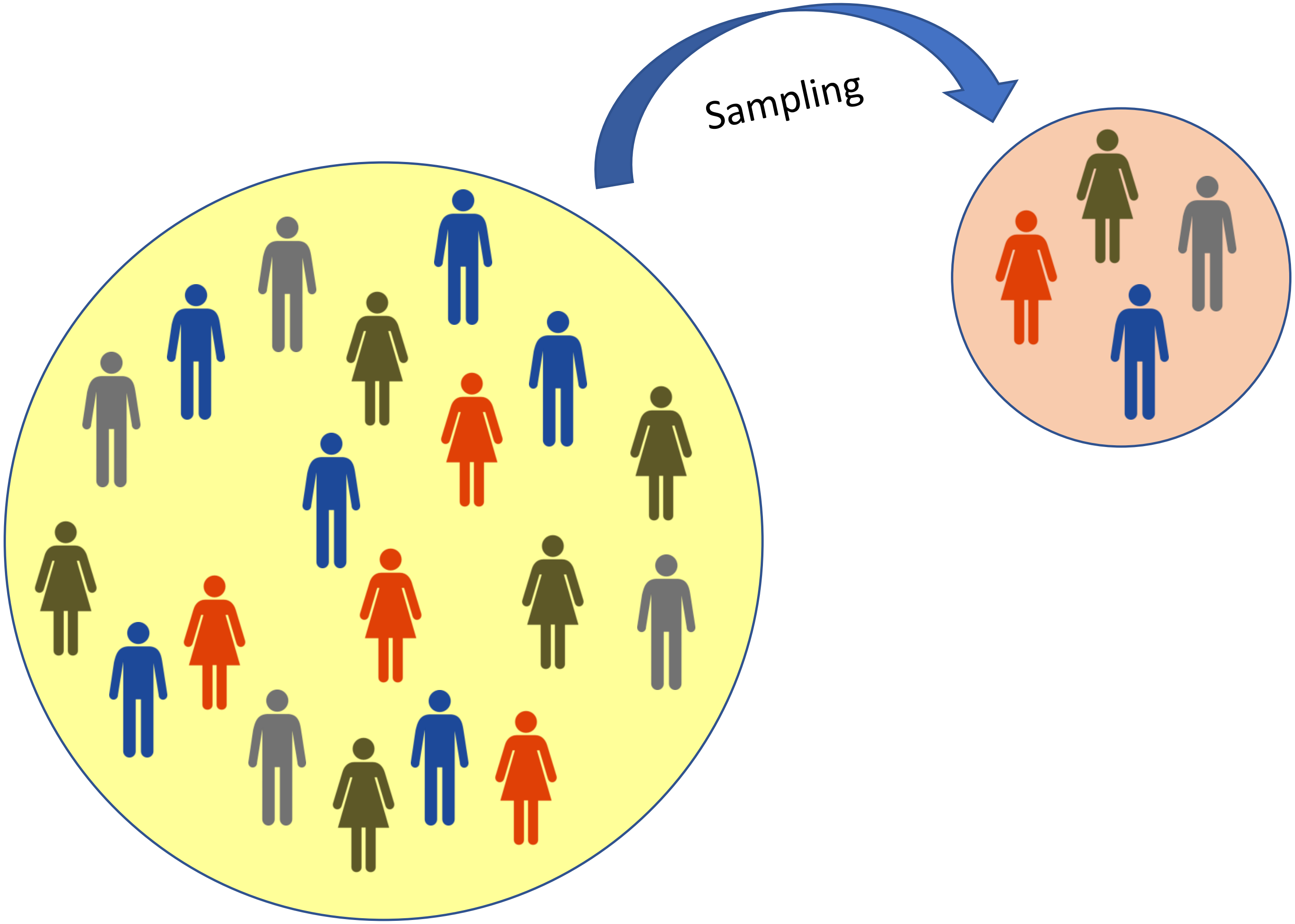
Practicalities

- The slides (ppt and pdf) are downloadable from http://github.com/levshina/Tallinn
- We will use R, free statistical software, and Rstudio.
- To install R: CRAN (Comprehensive R Archive Network) http://cran.r-project.org/
- To install RStudio: https://www.rstudio.com/

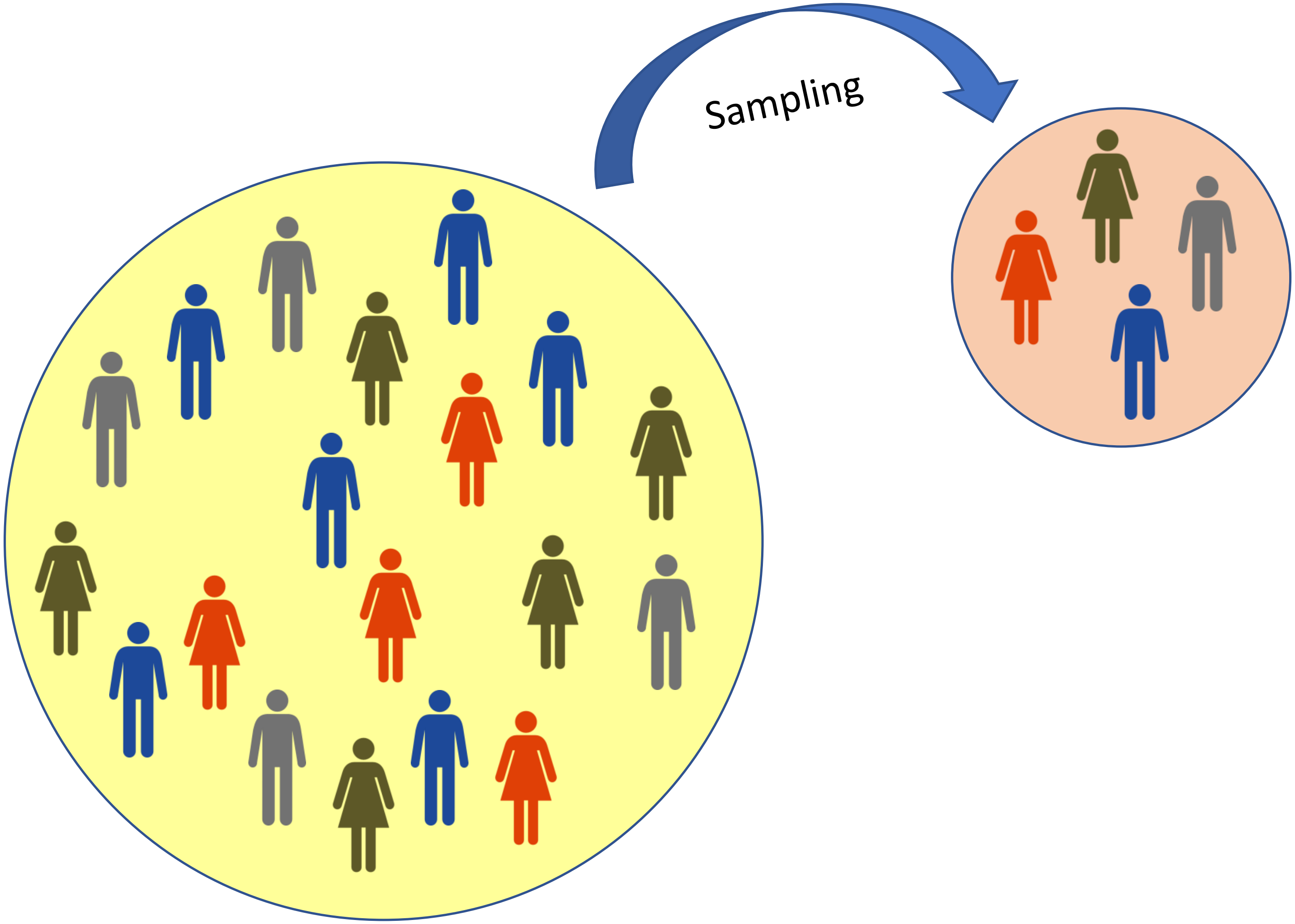
More information here:


How to do Linguistics with R
Data exploration and statistical analysis
Natalia Levshina
John Benjamins Publishing Company

Population and Sample



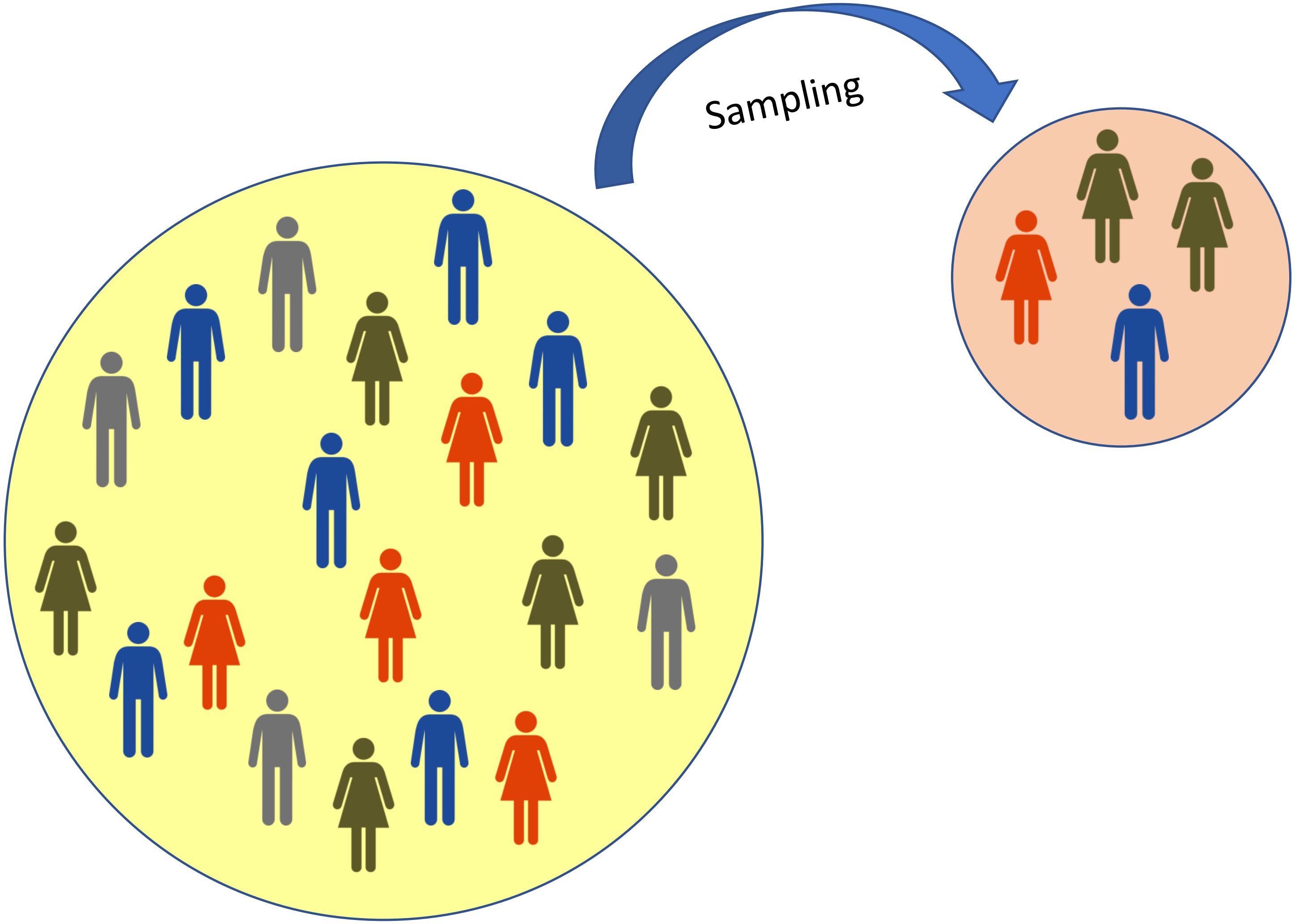
Population and Sample



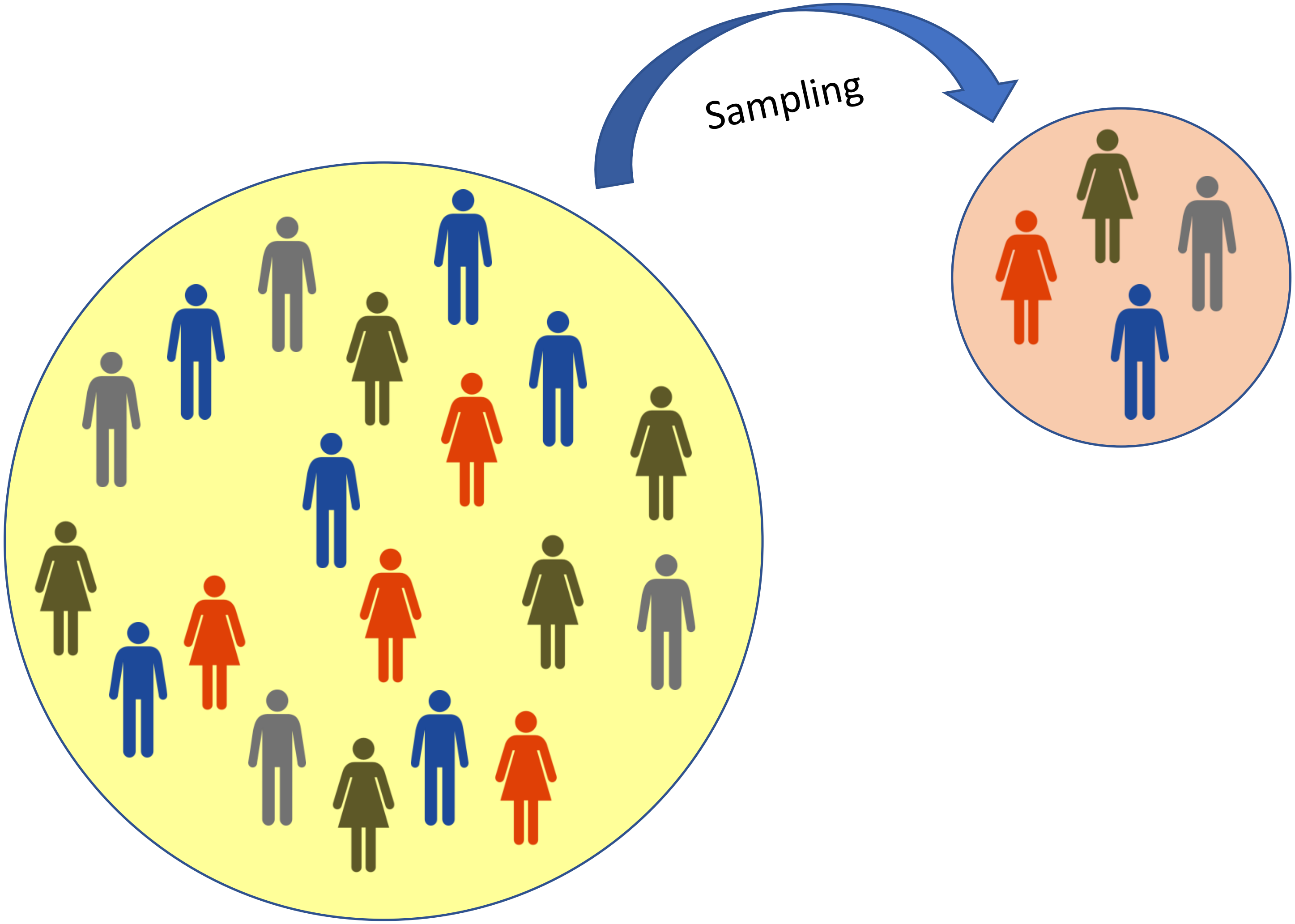
Population parameters (mean, variance, etc.)

Sample statistics (mean, variance, etc.)

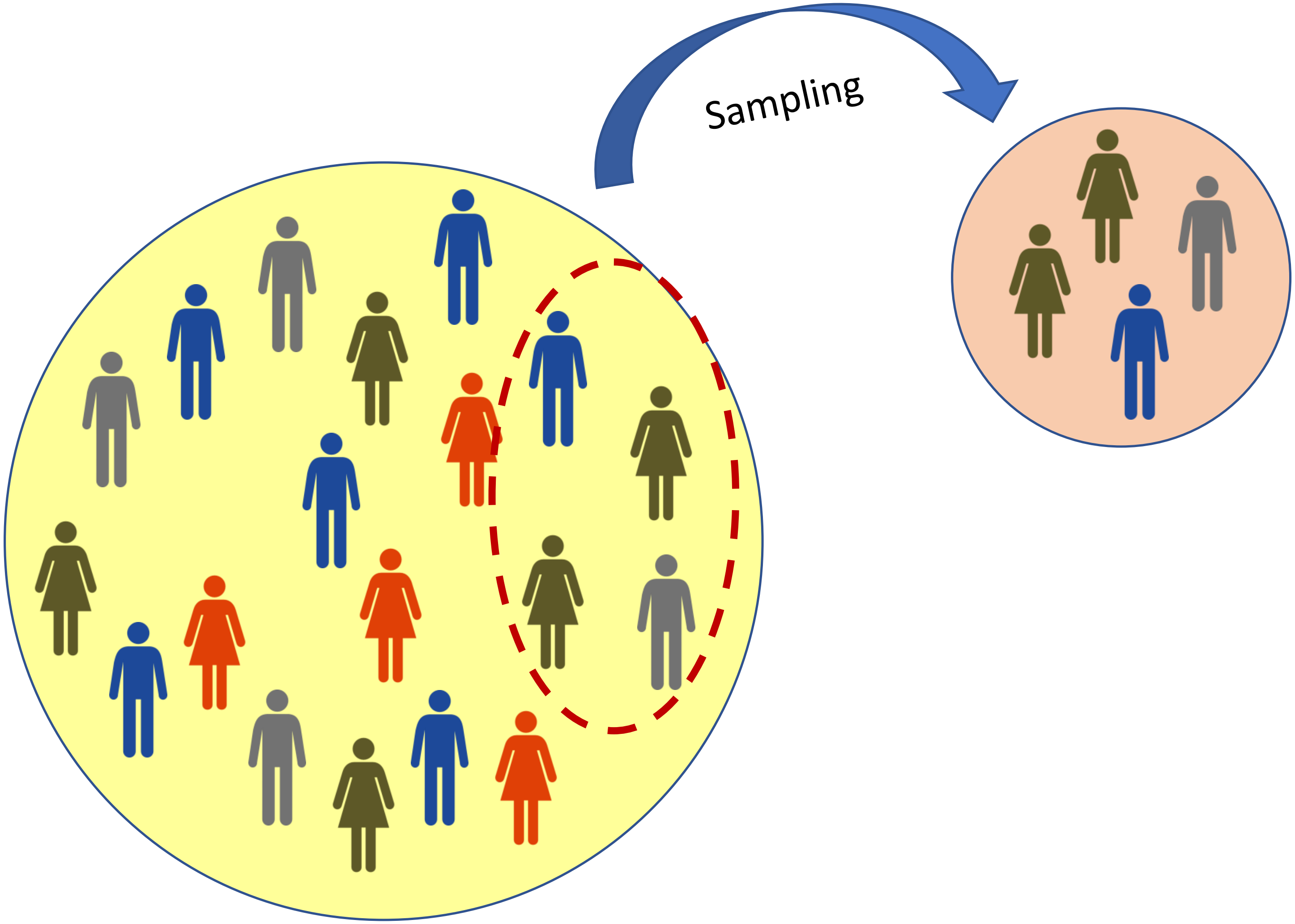
Random sampling



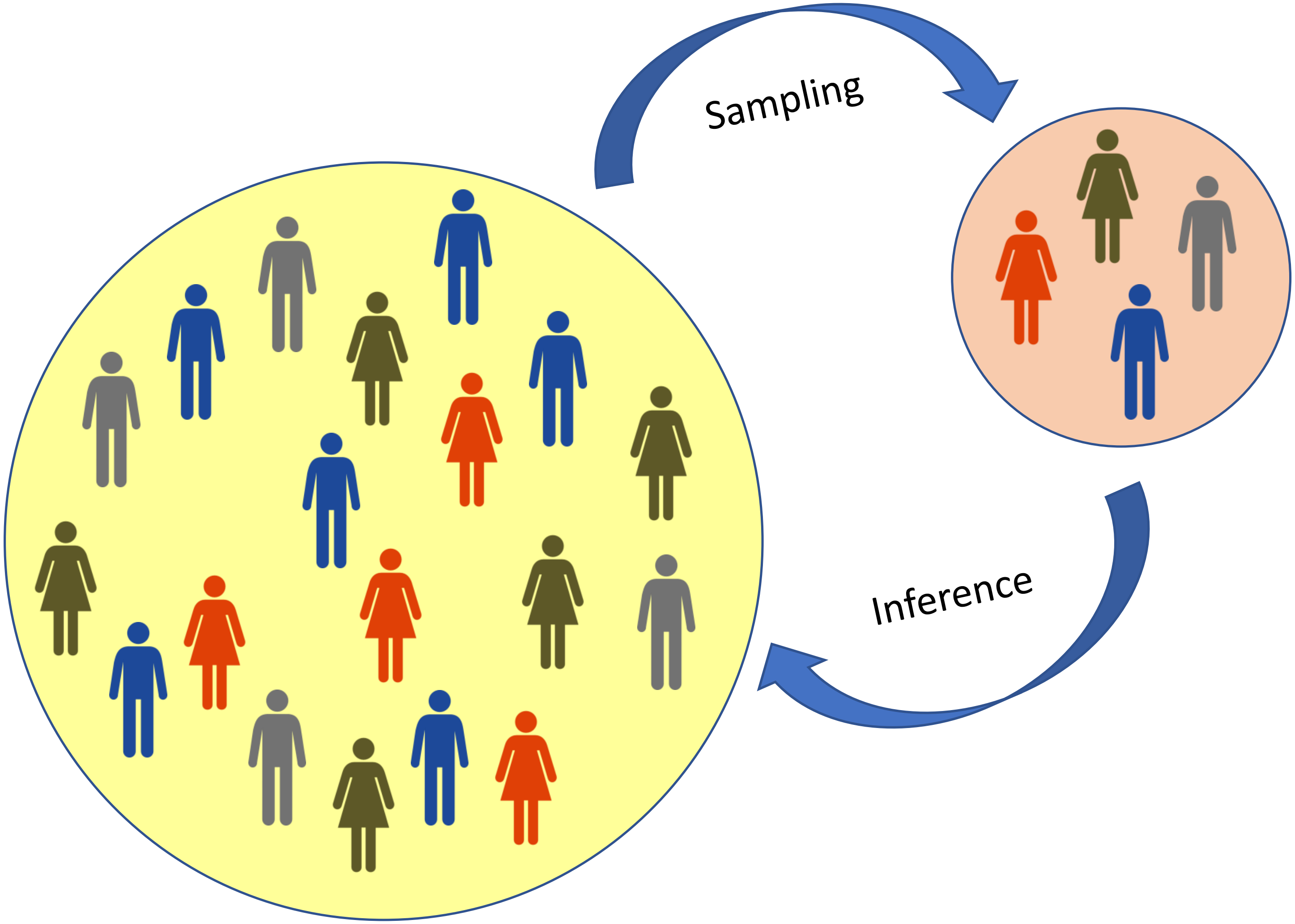
Representative sampling



Convenience sampling



Inferential statistics



WEIRD subjects

- Experimental linguistics has a strong bias towards so-called WEIRD subjects. Who are they?

WEIRD

- Western
- Educated
- Industrialized
- Rich
- Democratic countries

- At the same time, there are indications that WEIRD subjects are indeed cognitively and behaviorally ‘weird’, i.e. quite different from the rest of the species, which makes them the least representative for making generalizations about humans (Henrich et al. 2010).

Reproducibility of your results

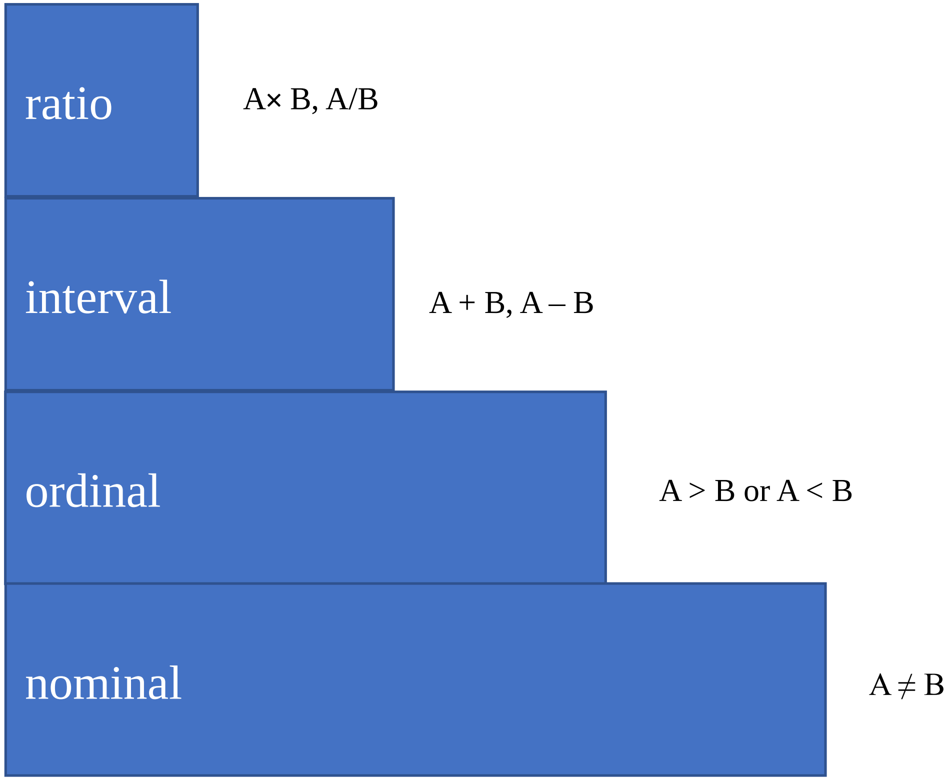
- **Methods reproducibility**: providing sufficient amount of detail about the data and methods so that they same procedures could be exactly repeated.

- **results reproducibility** (often referred to as replicability): obtaining the same results as a result of an independent study, where the procedure of the original study is followed as closely as possible.

- **inferential reproducibility:** drawing similar conclusions from a replicated study or reanalyzed original study.

Goodman et al. (2016)

Scales of measurement



ratio $A \times B$, A/B

interval $A + B$, $A - B$

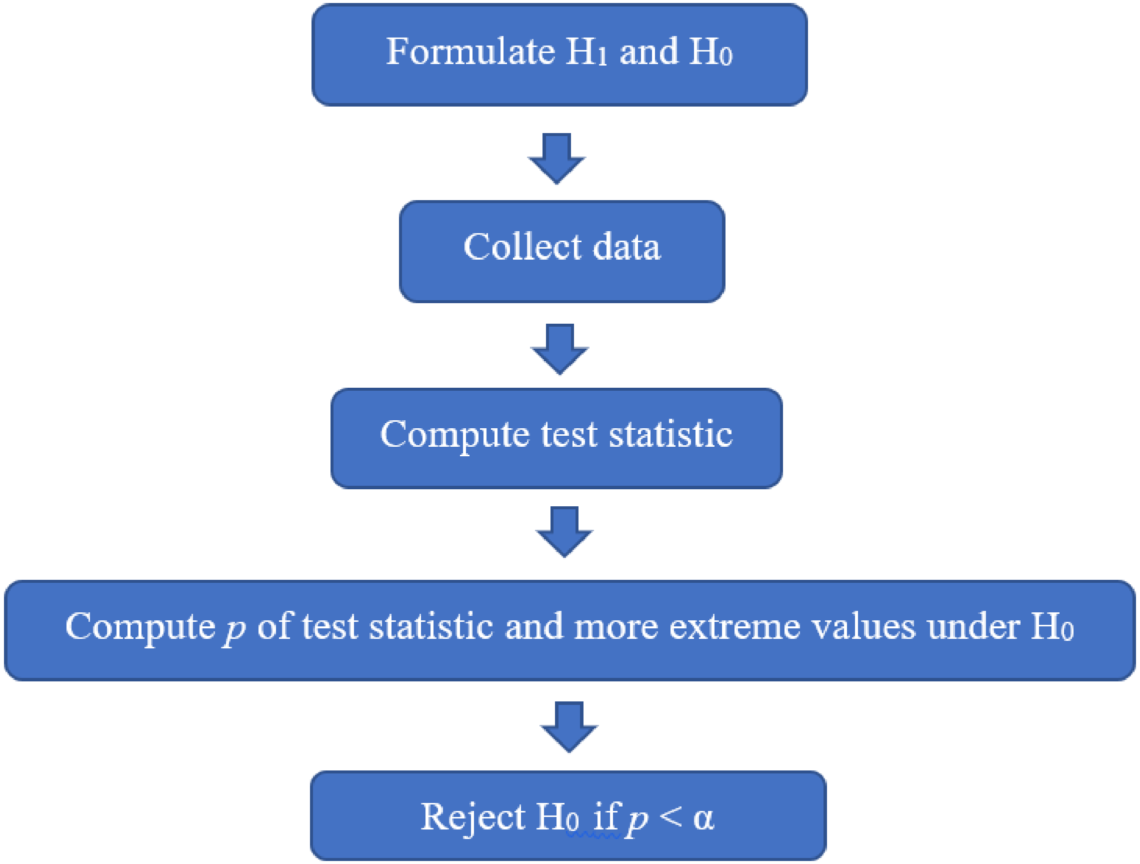
ordinal $A > B$ or $A < B$

nominal $A \neq B$

Exercise

Give examples of variables on the nominal, ordinal, interval and ratio scale of measurement.

The logic of hypothesis testing



Formulate H_1 and H_0

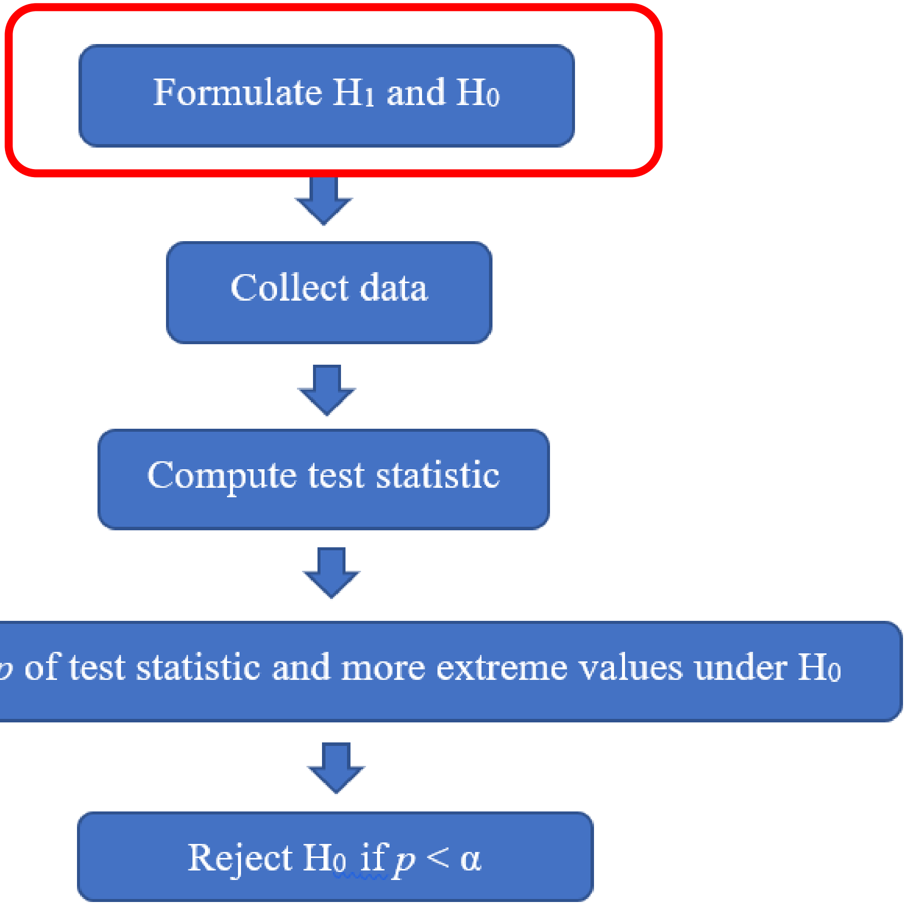
Collect data

Compute test statistic

Compute p of test statistic and more extreme values under H_0

Reject H_0 if $p < \alpha$

Alternative and null hypotheses



Formulate H_1 and H_0

Collect data

Compute test statistic

Compute p of test statistic and more extreme values under H_0

Reject H_0 if $p < \alpha$

Alternative vs. null hypothesis

- Alternative hypothesis (your research idea: difference between groups, association between variables)

    - directional (e.g. group 1 is GREATER/LESS than group 2; there is a POSITIVE/NEGATIVE correlation between variables A and B)

    - non-directional (some difference, some association)

- Null hypothesis (no difference between groups, no association between variables, etc.)

Example 1

H_0 (the null hypothesis): There is no difference in the number of lexemes that denote snow in languages spoken in hot and cold climates.

H_1 (the alternative hypothesis): There are more lexemes that denote snow in languages spoken in a cold climate.

Is H_1 directional or non-directional?

Example 2

H_0 (the null hypothesis): there is no relationship between the frequency of a word and how fast it is recognized in a lexical decision task.

H_1 (the alternative hypothesis): the more frequent a word, the faster it is recognized in a lexical decision task.

Is H_1 directional or non-directional?

Example 3

H_0 (the null hypothesis): there is no difference in the relative frequencies of metaphoric expressions used by men and women when they speak about sex.

H_1 (the alternative hypothesis): there is a difference in the relative frequencies of metaphoric expressions used by men and women when they speak about sex.

Is H_1 directional or non-directional?

Exercises

1. Consider an alternative hypothesis: Heterosexual men lower their voice pitch when they speak to women who they find sexually attractive, in comparison with the voice pitch they use with other interlocutors. Is it directional or non-directional? What is the corresponding null hypothesis?

2. Think about two research questions and try to formulate

a) a null hypothesis and a non-directional alternative hypothesis;

b) a null hypothesis and a directional alternative hypothesis.

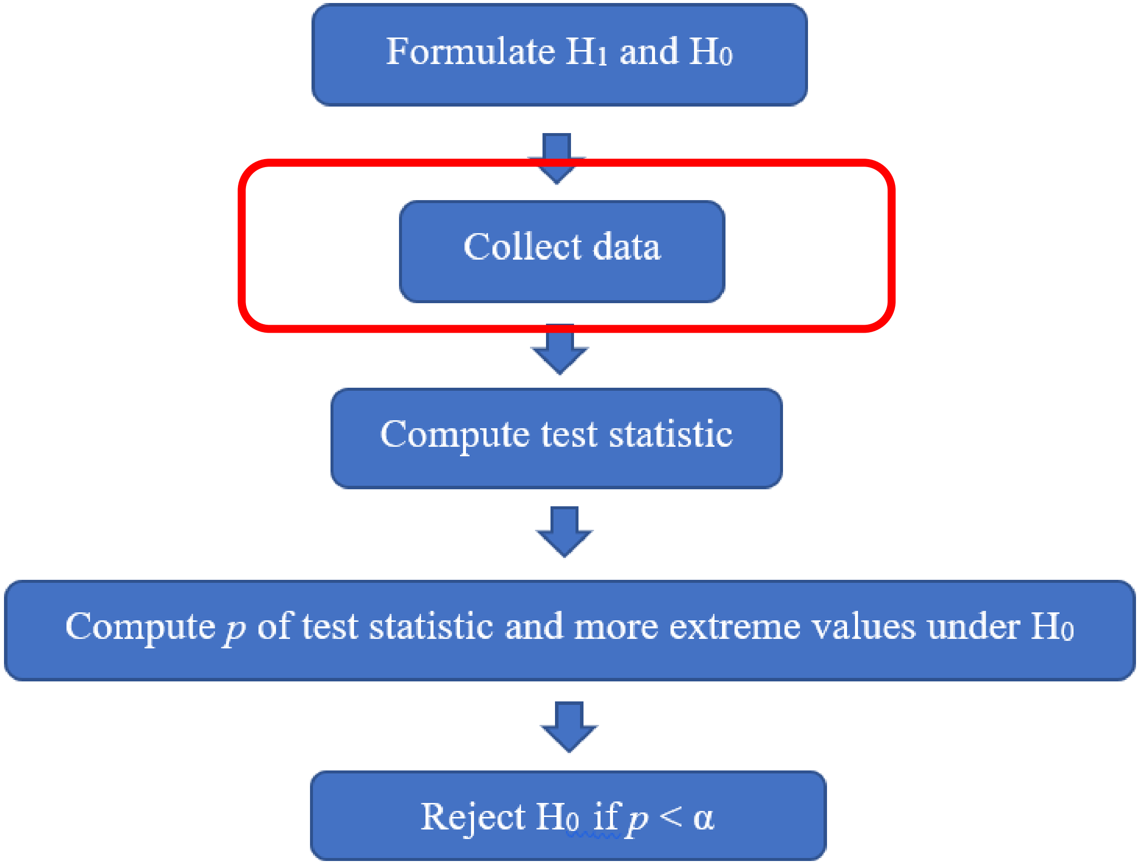
Testing the null hypothesis

- Traditional statistics in fact tests the null hypothesis, not the alternative one.
- It tries to reject the null.
- If it fails, the null hypothesis can be either true or false.

Another example

- Null hypothesis: There is no difference between the test scores of English learners who can play a musical instrument and the scores of those who cannot.
- What do you think might be the alternative hypothesis?

Collect data



Formulate H_1 and H_0

Collect data

Compute test statistic

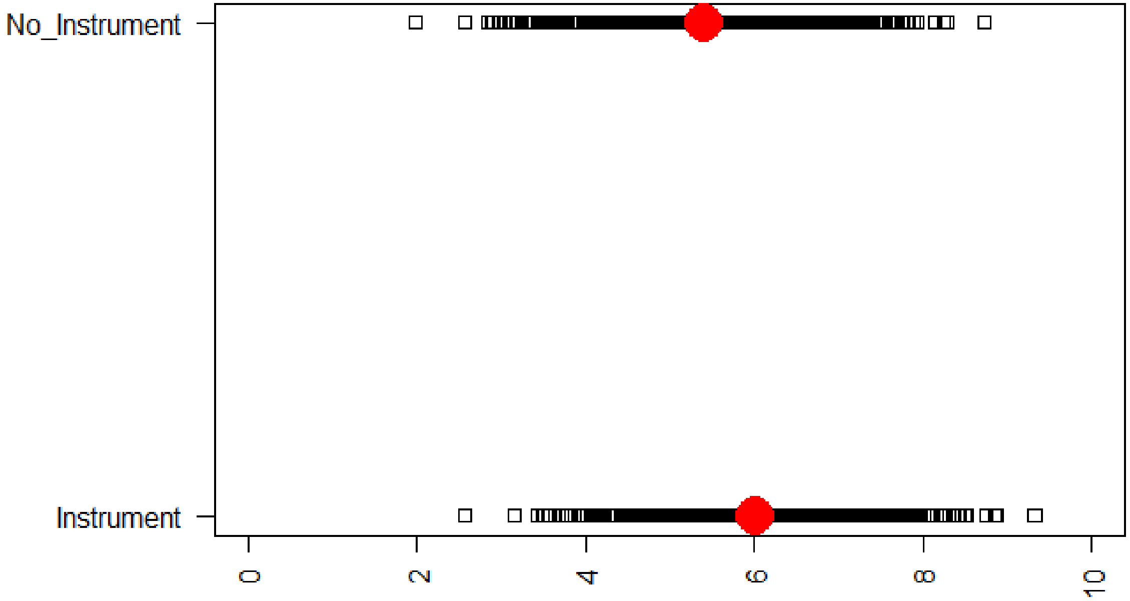
Compute p of test statistic and more extreme values under H_0

Reject H_0 if $p < \alpha$

Distribution of scores

- Let us imagine that we have collected the language test scores of 1,000 students all over the country who can play an instrument, and 1,000 scores of those who do not. These are all students in the same year. The scores were from 0 to 10.
- The test scores can be represented then as points on the horizontal axis ranged from the smallest to the largest ones (there's some overlap).

Descriptive statistics: the mean

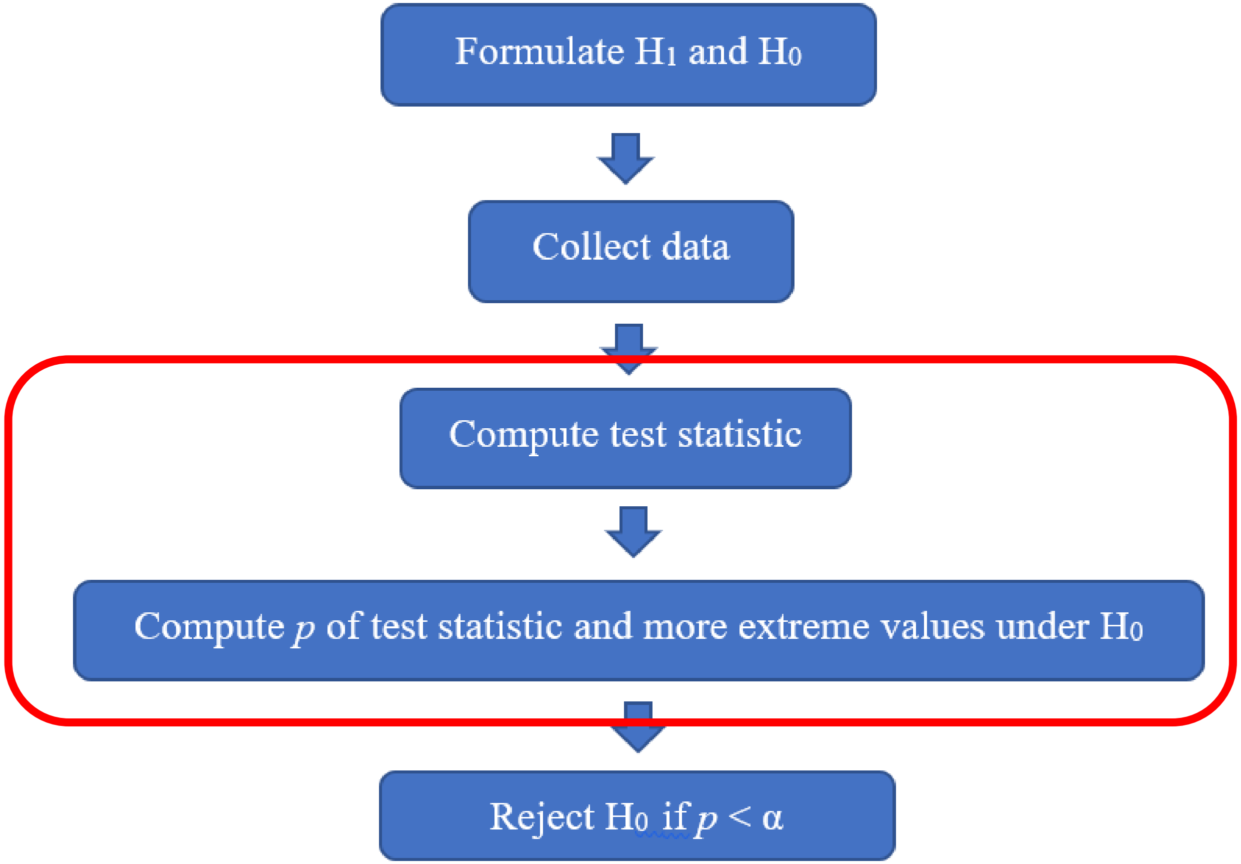


Mean Instrument = 6, Mean No Instrument = 5.4

Effect size vs. statistical significance

- The difference between the mean scores is *d* = 6 – 5.4 = 0.6.
- The students who play a musical instrument have on average higher scores.
- This meets our expectations! Time to party?
- Not yet. Is the difference **statistically significant**? That is, if we repeat the procedure, will we do find that students who play an instrument have higher scores?
- In contrast, the difference *d* shows the **effect size**.
- Important: an effect can be statistically significant, but not practically significant. Example: linguistic relativity studies (McWhorter, *The Language Hoax*).

Compute test statistic and p-value



Formulate H_1 and H_0

Collect data

Compute test statistic

Compute p of test statistic and more extreme values under H_0

Reject H_0 if $p < \alpha$

But which test statistic to use?

- There exist many different tests: t-test, chi-squared test, F-test, etc.
- The choice depends on the data and research question.
- For example, the t-test can test the differences between two groups.

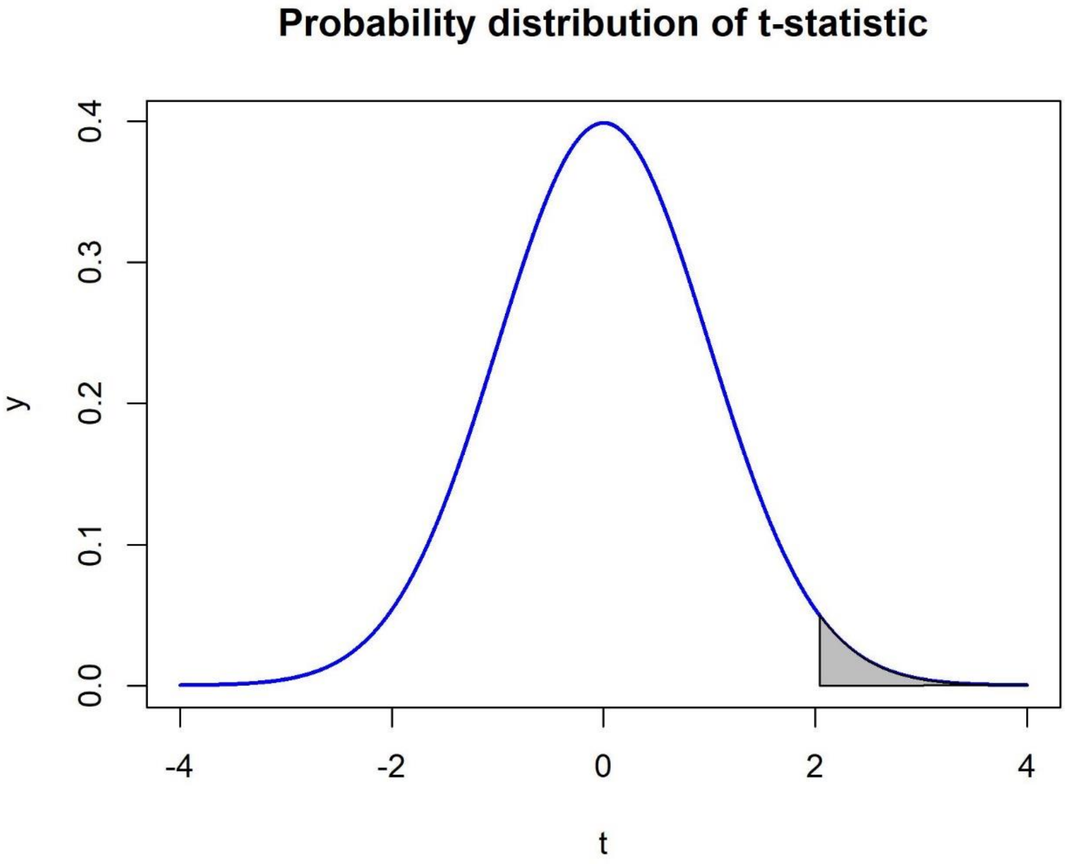
t-test in R

```
> t.test(s, s2, alternative = "greater")

	Welch Two Sample t-test

data:  s and s2
t = 13.69, df = 1998, p-value < 2.2e-16
alternative hypothesis: true difference in means is
greater than 0
95 percent confidence interval:
 0.5282617       Inf
sample estimates:
mean of x mean of y
 6.008368  5.407930
```

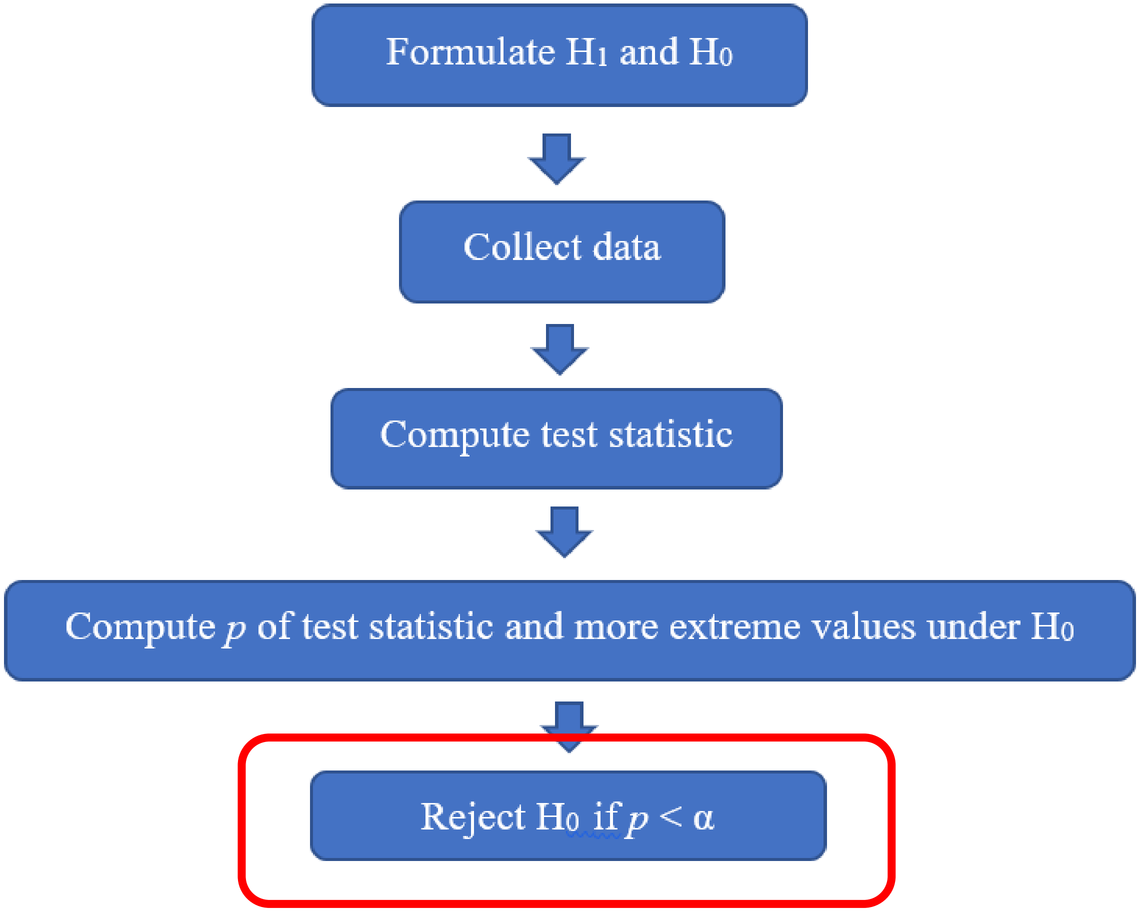
Where does the p-value come from?



The meaning of the p-value

the probability of finding the observed, or more extreme, test statistic (here: t-statistic) when the null hypothesis (here: no difference) is true.

Reject or not reject the null hypothesis



Formulate H_1 and H_0

Collect data

Compute test statistic

Compute p of test statistic and more extreme values under H_0

Reject H_0 if $p < \alpha$

Reject or not to reject the null?

- If a *p*-value is smaller the **significance level** (usually 0.05, or 5%), then the null hypothesis is rejected, and one has grounds to believe that the result is not due to chance. Therefore, one can conclude that there is indeed a difference between the groups, association between the variables, etc., depending on the research hypothesis and statistical test.
- If the *p*-value is larger than this conventional value, then the null hypothesis cannot be rejected, and you can conclude that there is no sufficient evidence that the groups are different (or the variables are correlated, associated, etc.).
- In our case, the p-value is extremely small. Therefore, we can reject the null hypothesis of no difference.

Presumption of innocence

- The logic of null hypothesis testing is similar to a judicial procedure based on the presumption of innocence (Feinberg 1971).
- The presumption of innocence can be expressed in statistical terms as the null hypothesis that a defendant is innocent.
- This null hypothesis is discarded if there is compelling evidence that it is false.
- If there is reasonable doubt, the defendant must be acquitted.
- Importantly, the jury comes with the verdict “Guilty” or “Not Guilty”, but the latter does not mean “Innocent”. “Not guilty” means only that guilt has not been proven.

Type I and Type II errors

	Real world: the null hypothesis is false	Real world: the null hypothesis is true
Statistical Model: reject the null hypothesis	correct rejection	Type I error
Statistical Model: do not reject the null hypothesis	Type II error	correct non-rejection

Exercise

- Using the analogy with a jury trial, what are the four outcomes for the defendant? Which correspond to Type I and Type II?
- What are the four outcomes for our case study of language and music?

Significance level and power

- The significance level of 0.05 is our tolerance for Type I error.
- Type II error determines the **power** of a statistical test (the chances that the false null will be rejected). It depends on sample size (the more the better) and effect size (the larger the difference, the easier it is to detect it).

What statistics cannot do for you

- Correlation does not imply causation
  - e.g. spurious correlations, https://tylervigen.com/spurious-correlations)
- Choice of appropriate methods
  - E.g. Chen (2013) reported a correlation between the presence of a grammatical future tense in languages and their speakers' propensity to save, while controlling for numerous economic and demographic factors. Languages which grammatically distinguish the present and the future may bias their speakers to distinguish them psychologically.
    - For example, languages like English and Spanish influence their speakers in the Worfian way to less future-oriented decision making because the future is perceived as something different from the present moment. In contrast, languages where the present tense is used to express future events, e.g. Mandarin Chinese or Finnish, do not have this contrast between now and some remote future, which makes the speakers save money.
  - But this effect disappears if one controls for the genealogical and geographical relationships between the languages (Roberts et al. 2015).