Part 1
What is statistics?

Natalia Levshina © 2017

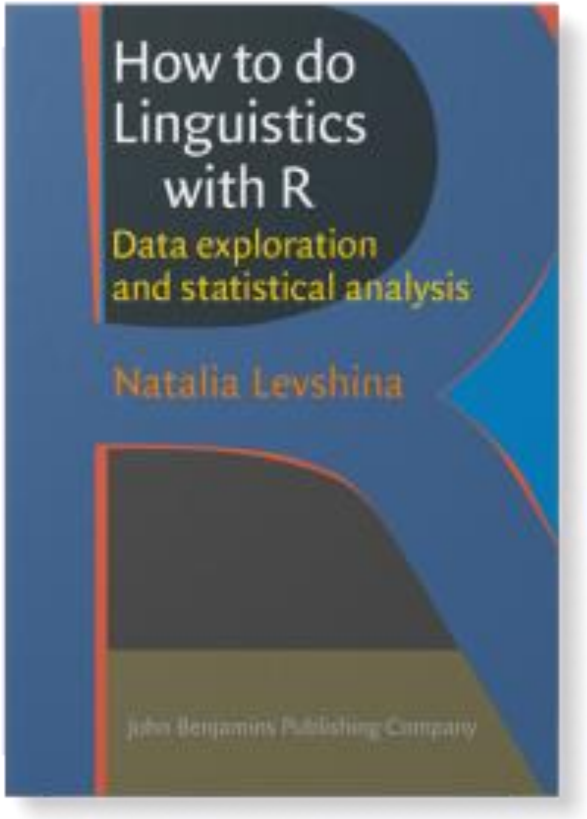
University of Mainz, Germany
June 2017

Practicalities

- The slides (pdf) are downloadable from http://natalialevshina.com/statistics.html

- We will use R, free statistical software. R code can be copied from the slides and pasted into R

More information here:



How to do Linguistics with R

Data exploration and statistical analysis

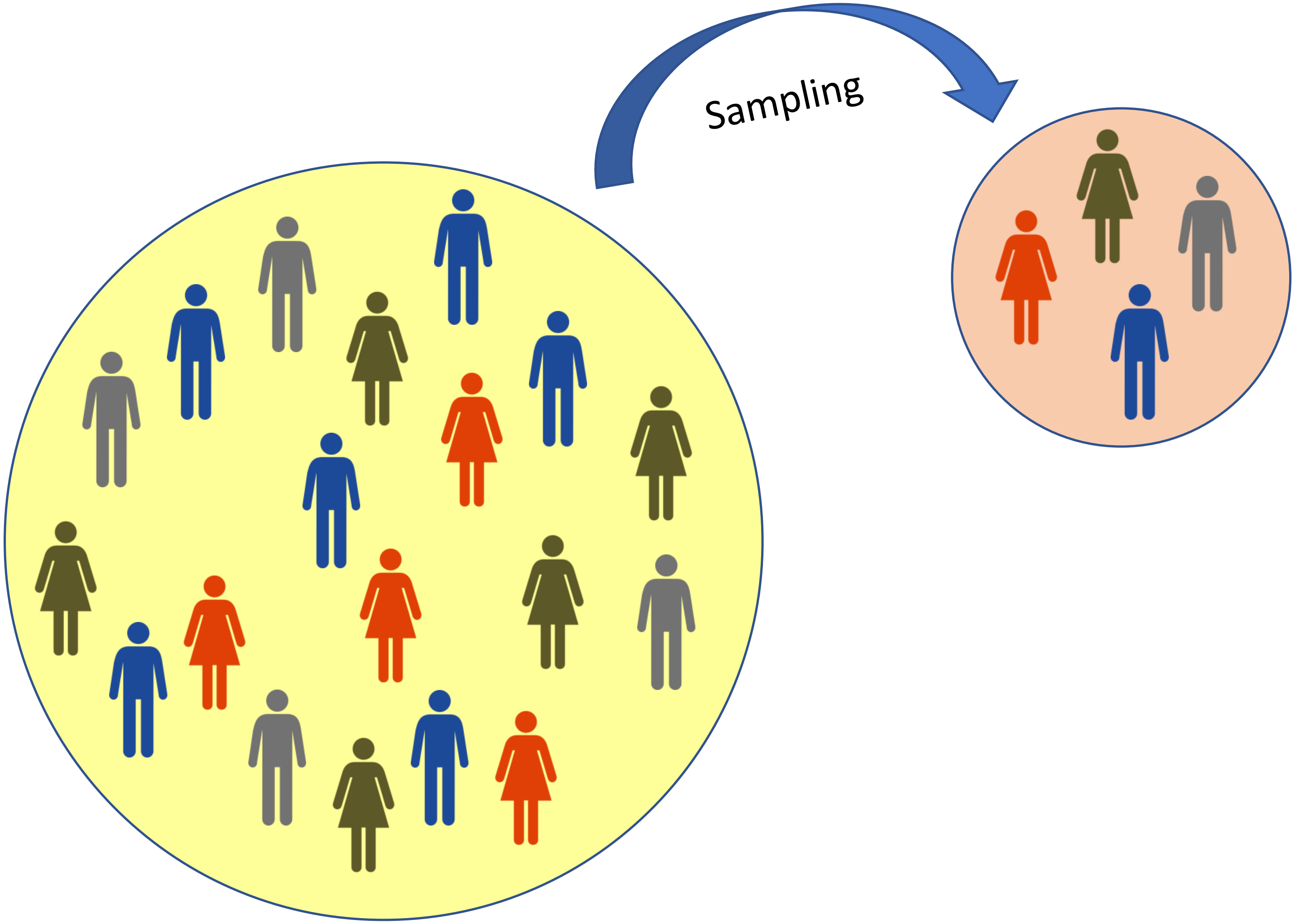
Natalia Levshina

John Benjamins Publishing Company

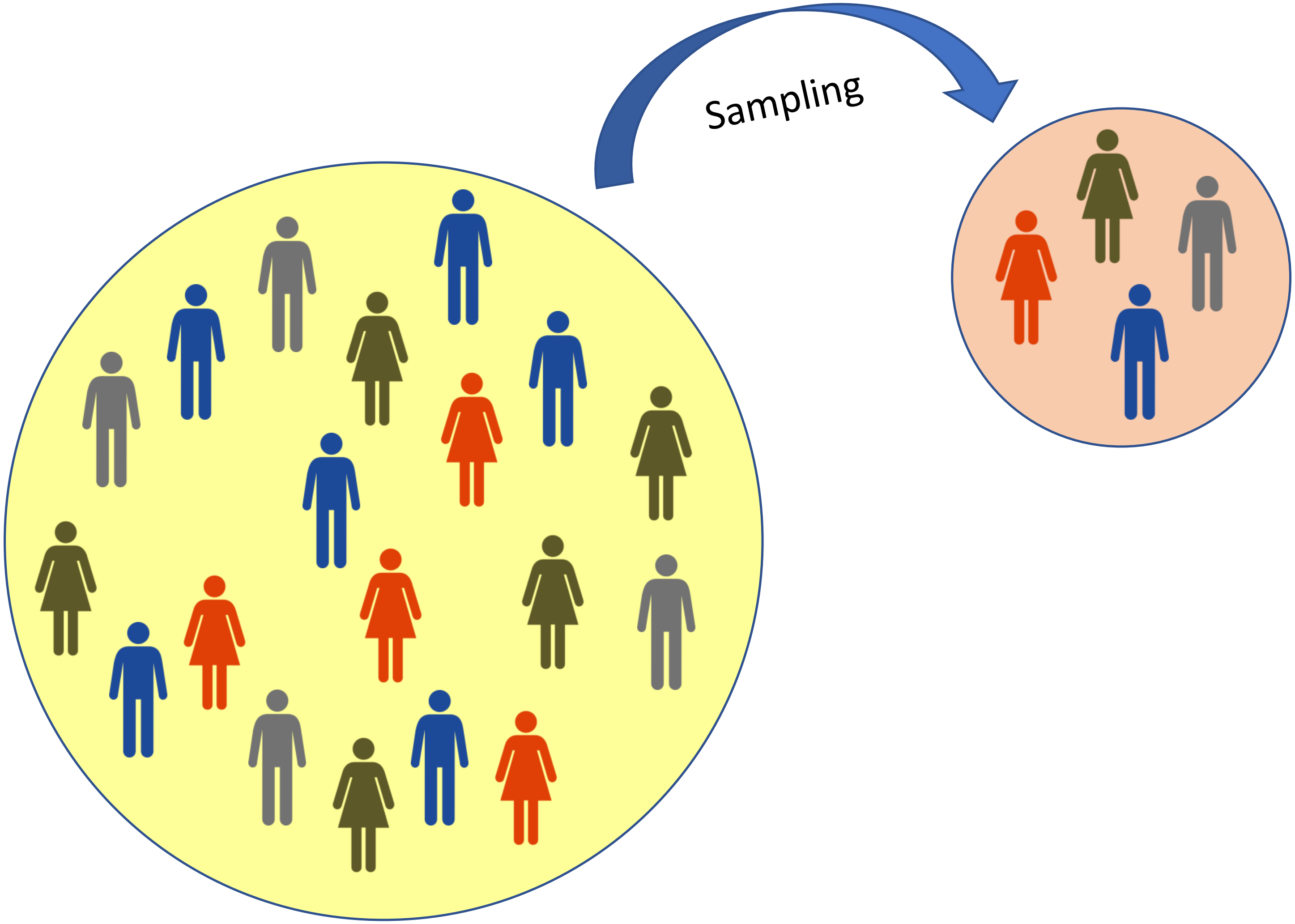
Course outline

1. What is statistics?
2. Introduction to R
3. Basic descriptive statistics and graphs for numeric variables
4. Basic descriptive statistics and graphs for categorical variables
5. Chi-squared test and Fisher exact test
6. Correspondence Analysis
7. Linear regression
8. Conditional inference trees and random forests

Population and Sample



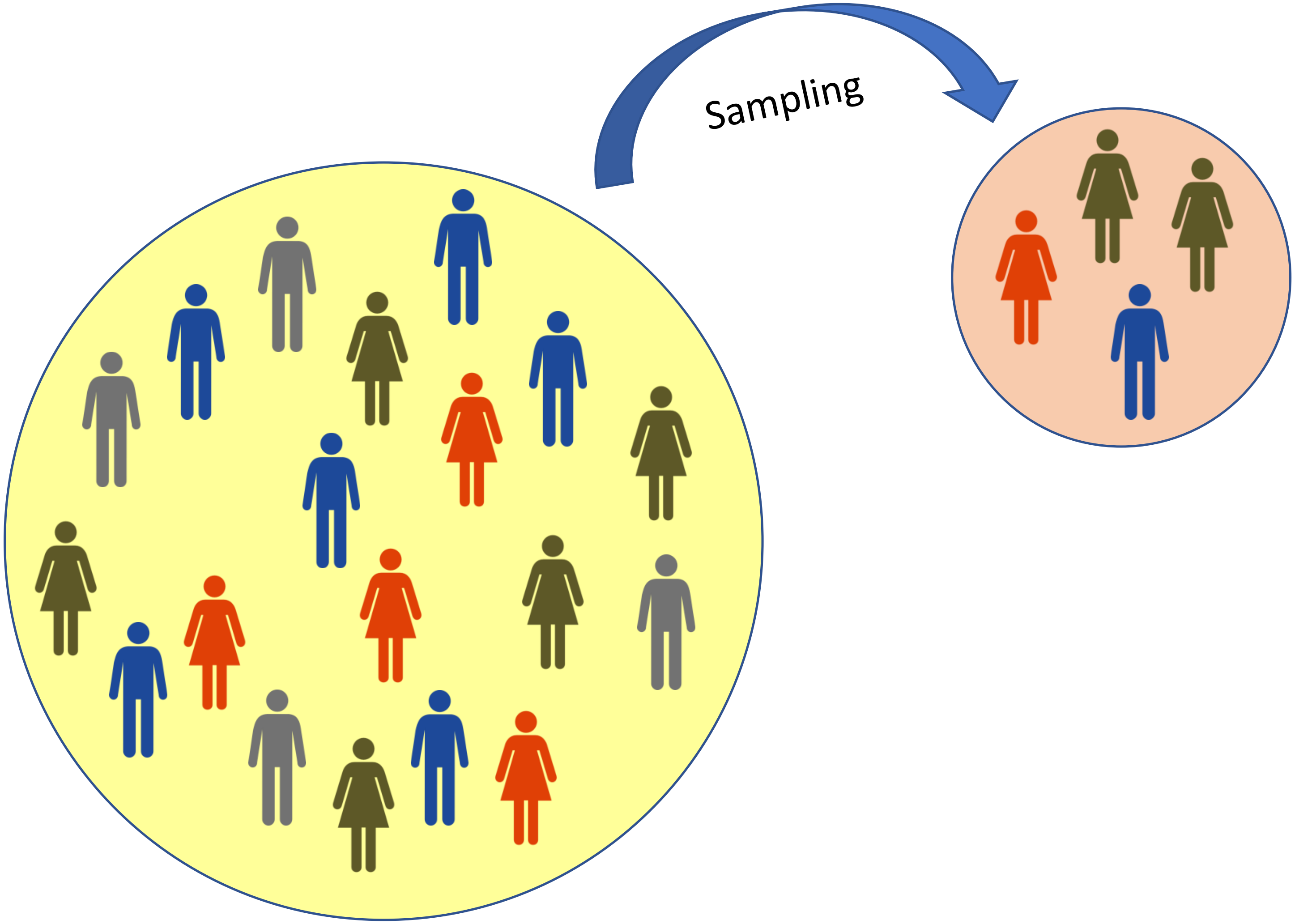
Population and Sample



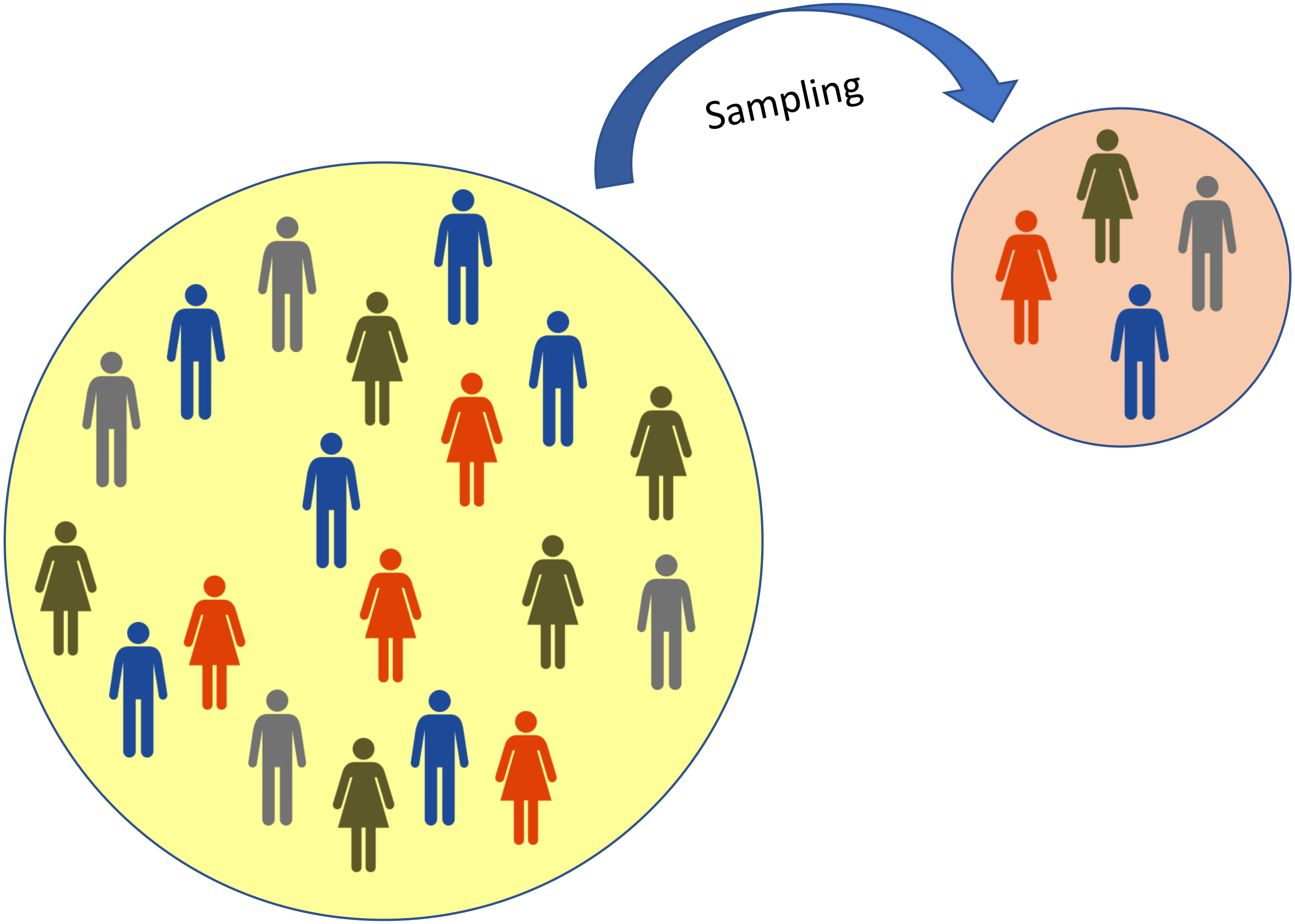
Population parameters (mean, variance, etc.)

Sample statistics (mean, variance, etc.)

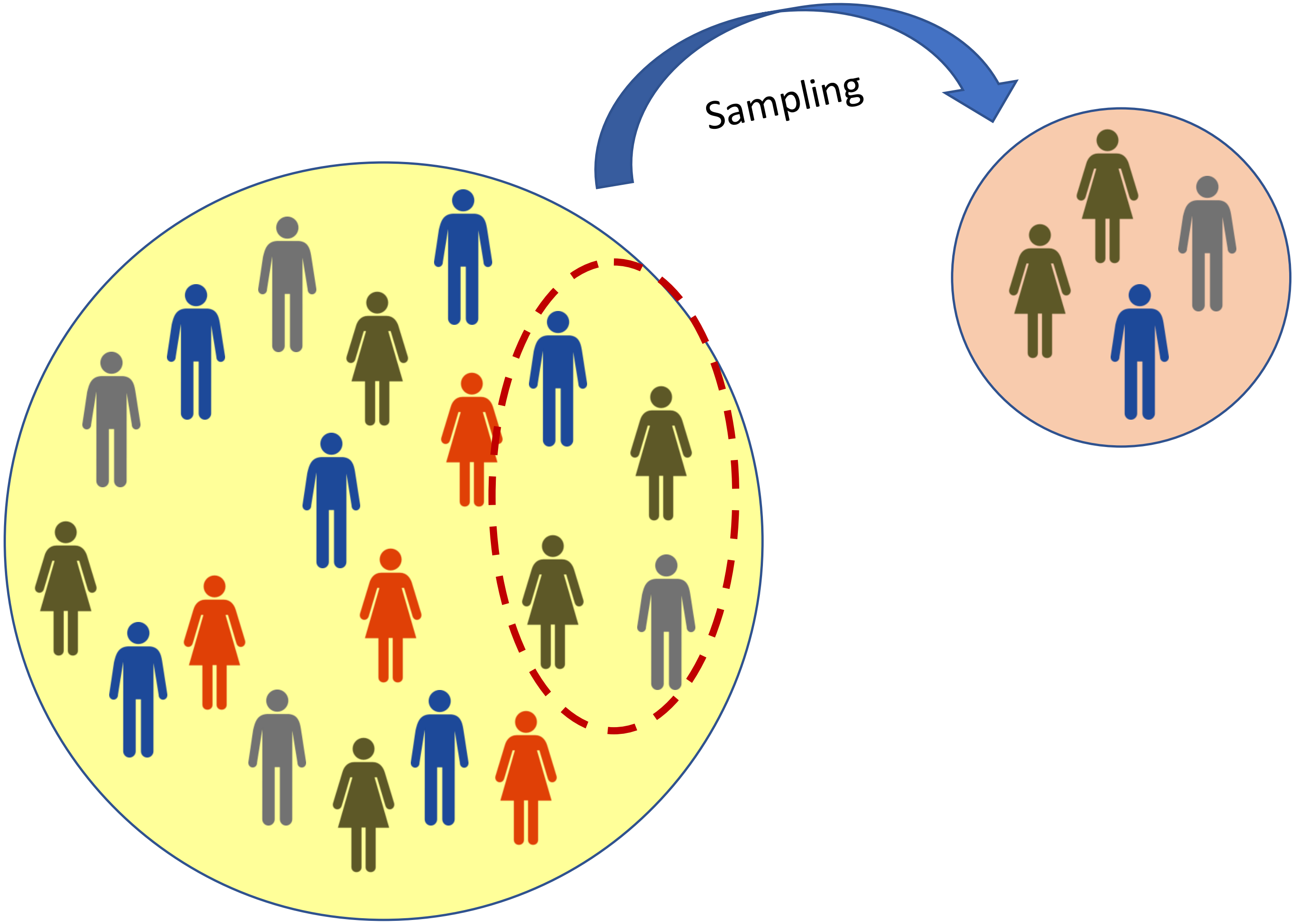
Random sampling



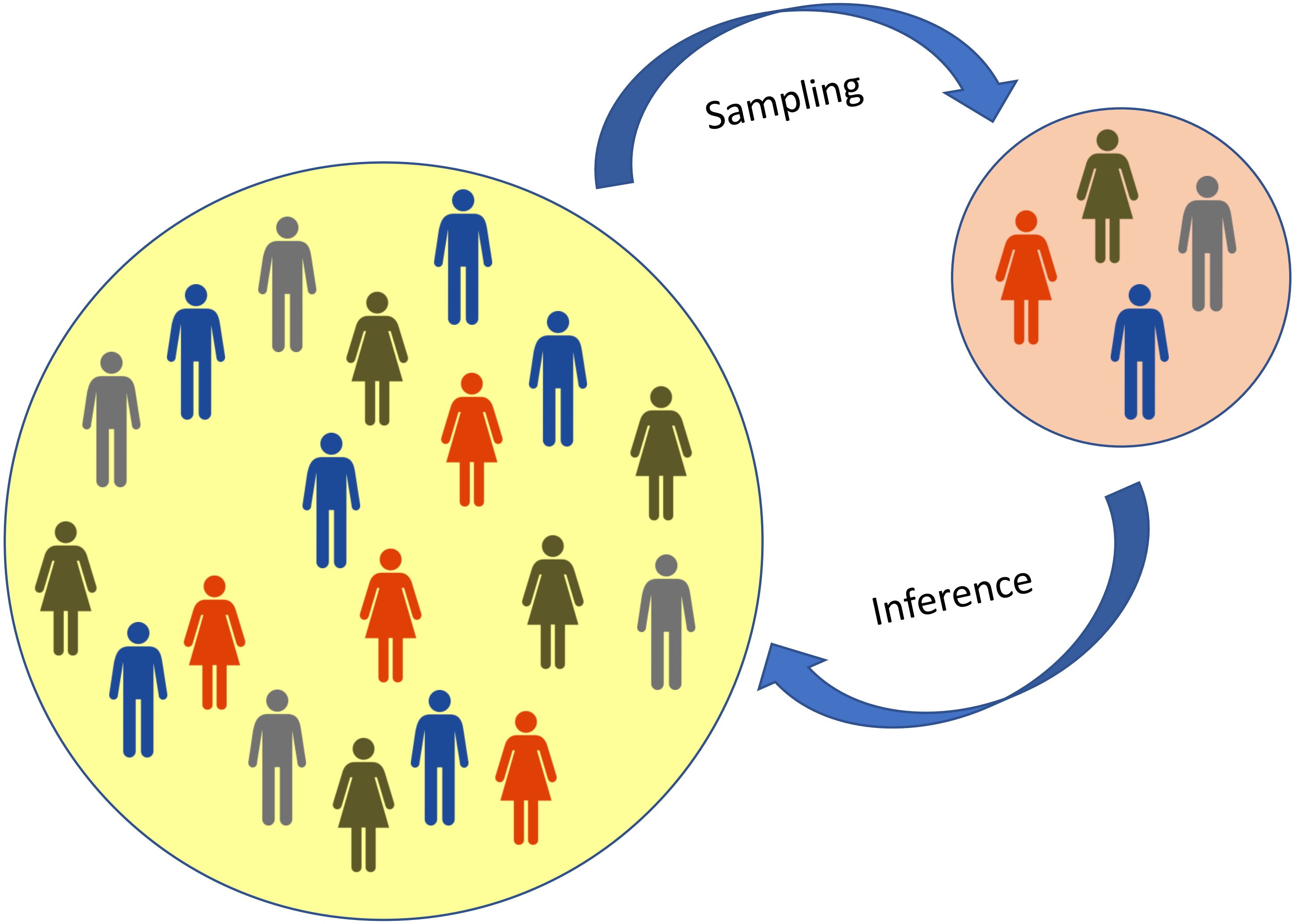
Representative sampling



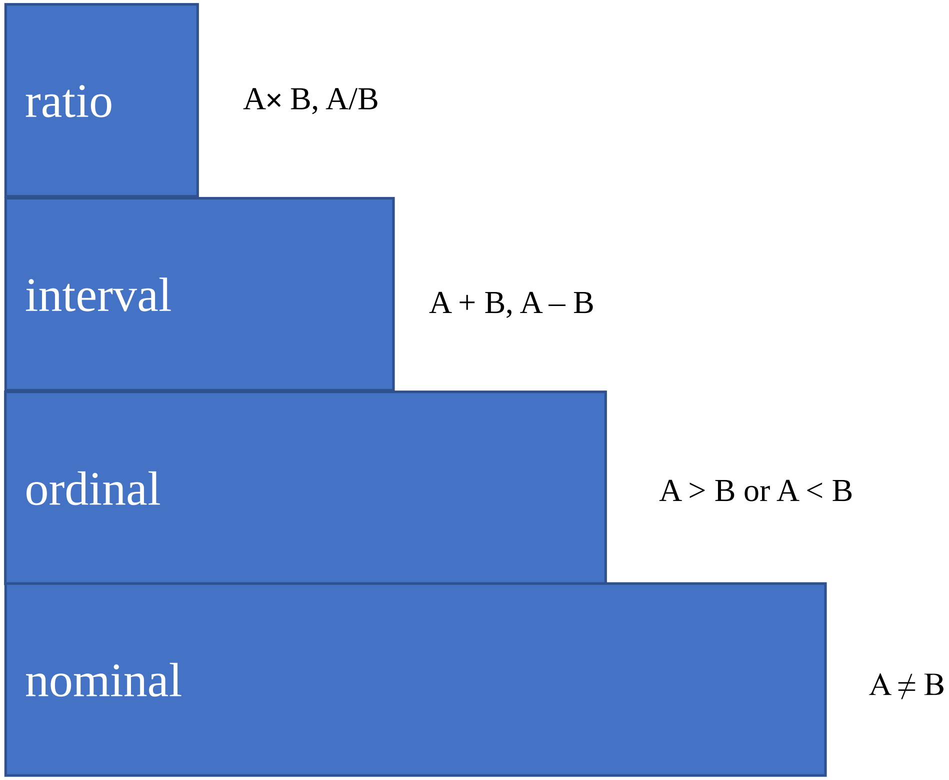
Convenience sampling



Inferential statistics



Scales of measurement



ratio $A \times B, A/B$

interval $A + B, A - B$

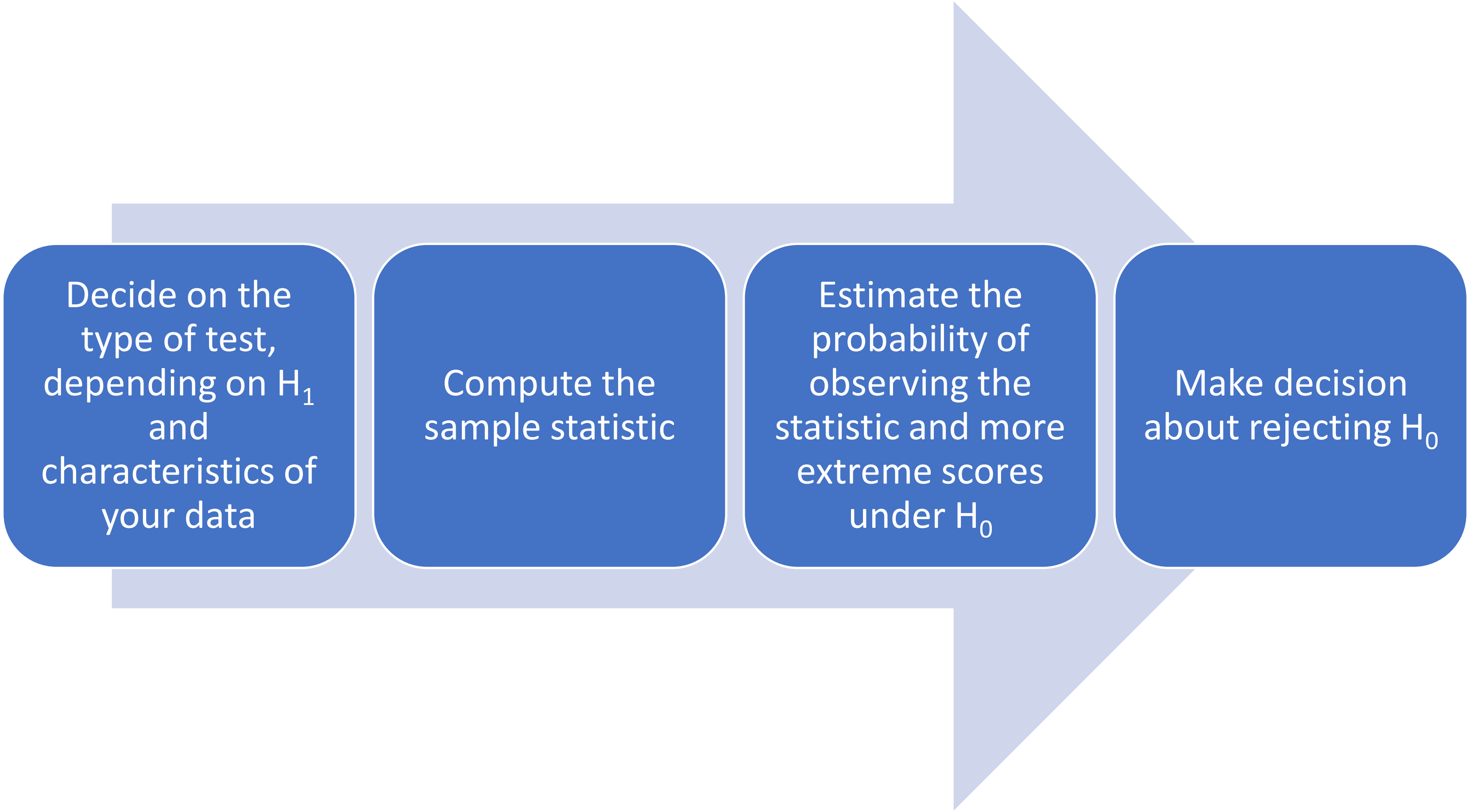
ordinal $A > B$ or $A < B$

nominal $A \neq B$

Exercise

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

Hypothesis testing algorithm



Decide on the type of test, depending on H_1 and characteristics of your data

Compute the sample statistic

Estimate the probability of observing the statistic and more extreme scores under H_0

Make decision about rejecting H_0

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 Eskimo and Yucatec Maya.

H_1 (the alternative hypothesis): There are more lexemes that denote snow in Eskimo than in Yucatec Maya.

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?

Exercise

Think about two research questions and try to formulate

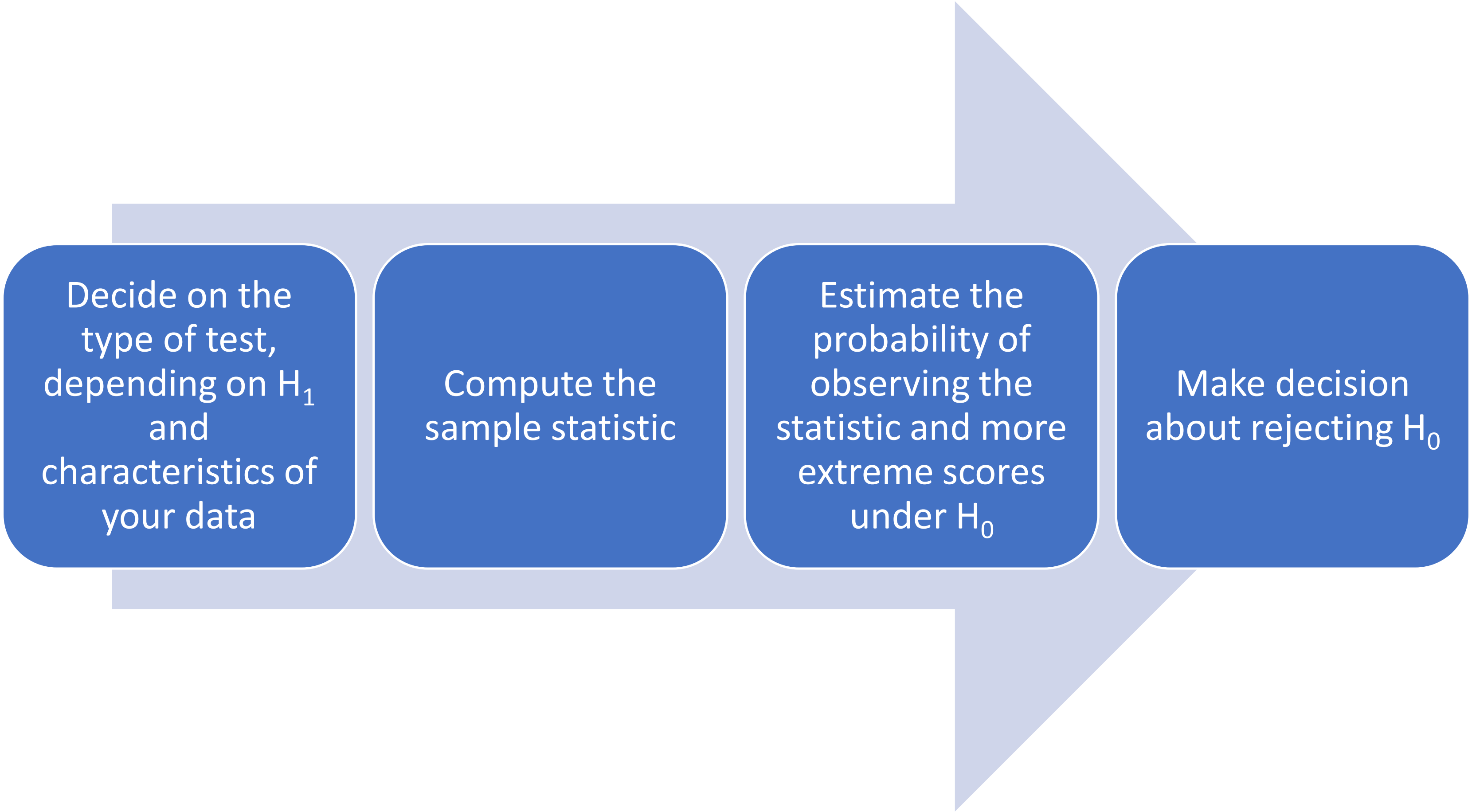
a) a null hypothesis and a non-directional alternative hypothesis;

b) a null hypothesis and a directional alternative hypothesis.

Case study: refugees and media

- You investigate the representation of immigration in mass media. You take a newspaper X and find 10 instances of the word ‘immigrant(s)’ in the newspaper. In 8 cases, the context is negative. In the remaining 2 cases, the context is neutral or positive.
- What can be the null and the alternative hypotheses?
- When a bias of some sort is investigated (e.g. biased coins, biased opinions) and there are two possible outcomes, one can use the binomial test.

Hypothesis testing algorithm



Decide on the type of test, depending on H_1 and characteristics of your data

Compute the sample statistic

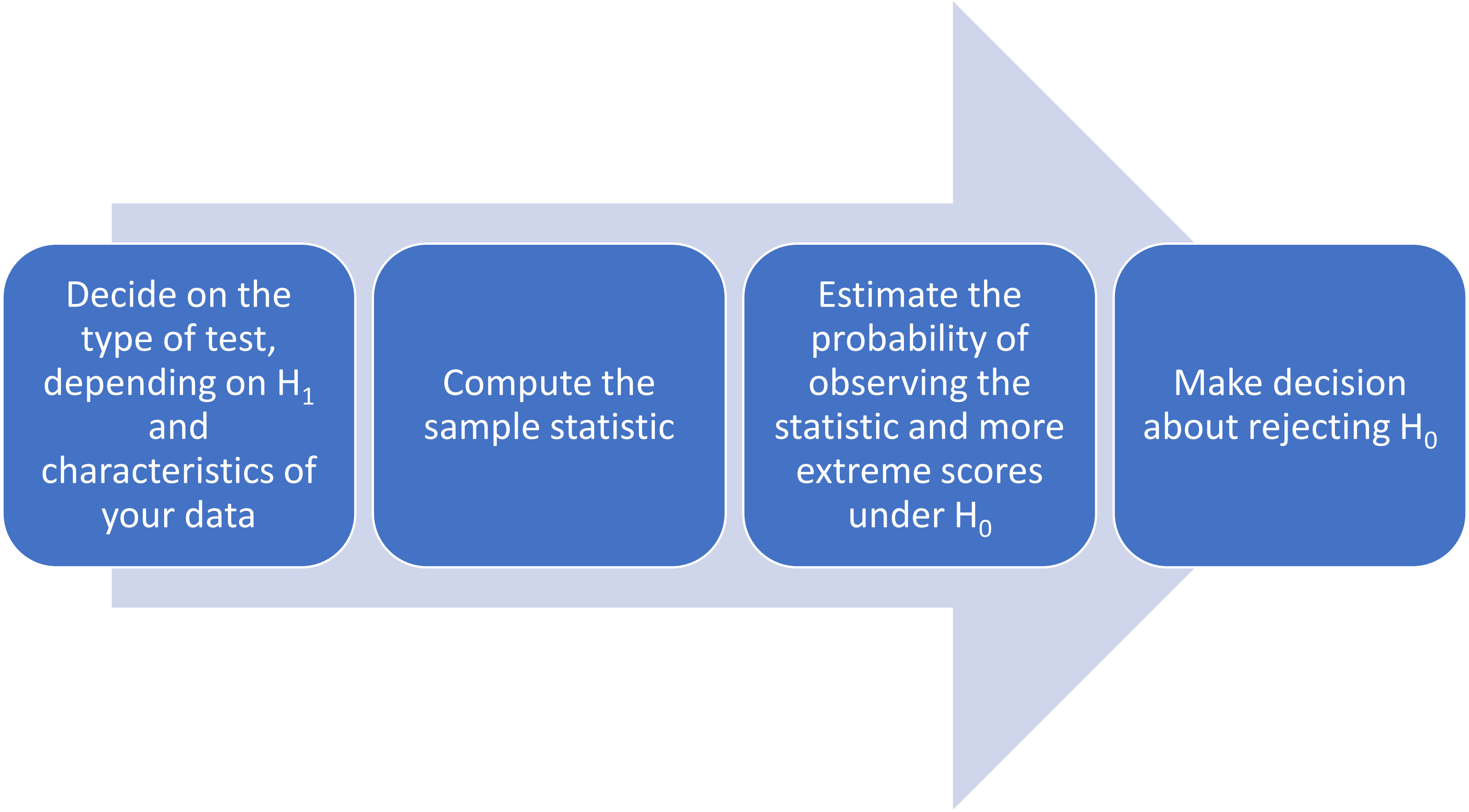
Estimate the probability of observing the statistic and more extreme scores under H_0

Make decision about rejecting H_0

Compute the sample statistics

- The probability of success is 0.8, or 80%.

Hypothesis testing algorithm

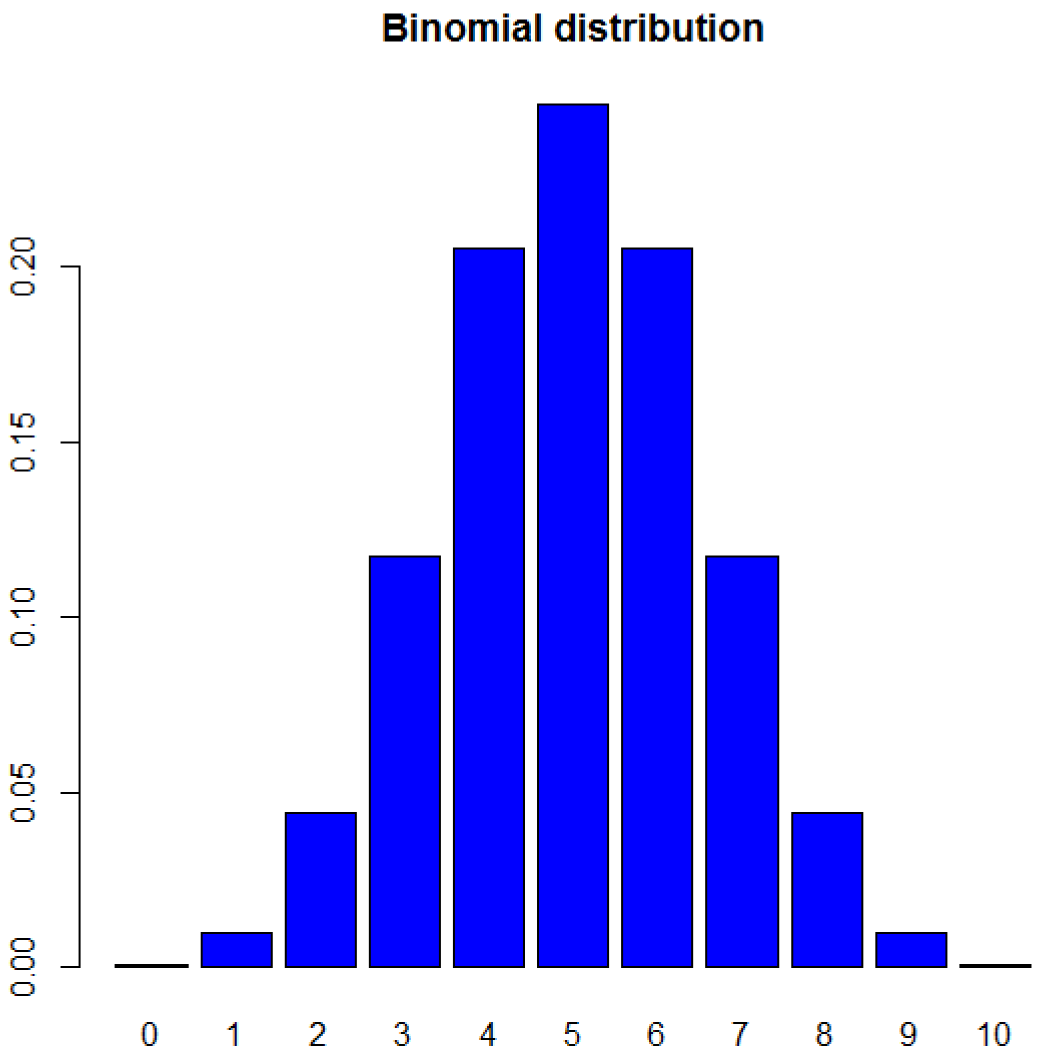


Decide on the type of test, depending on H_1 and characteristics of your data

Compute the sample statistic

Estimate the probability of observing the statistic and more extreme scores under H_0

Make decision about rejecting H_0


Binomial distribution
0.00
0.05
0.10
0.15
0.20
0
1
2
3
4
5
6
7
8
9
10

Estimate the probability under H_0

- If you have a directional alternative hypothesis (finding a negative bias), then you compute the chances of observing 8 negative mentions and more if the results are due to chance alone, i.e. there is no bias.
- It’s called a one-tailed test.

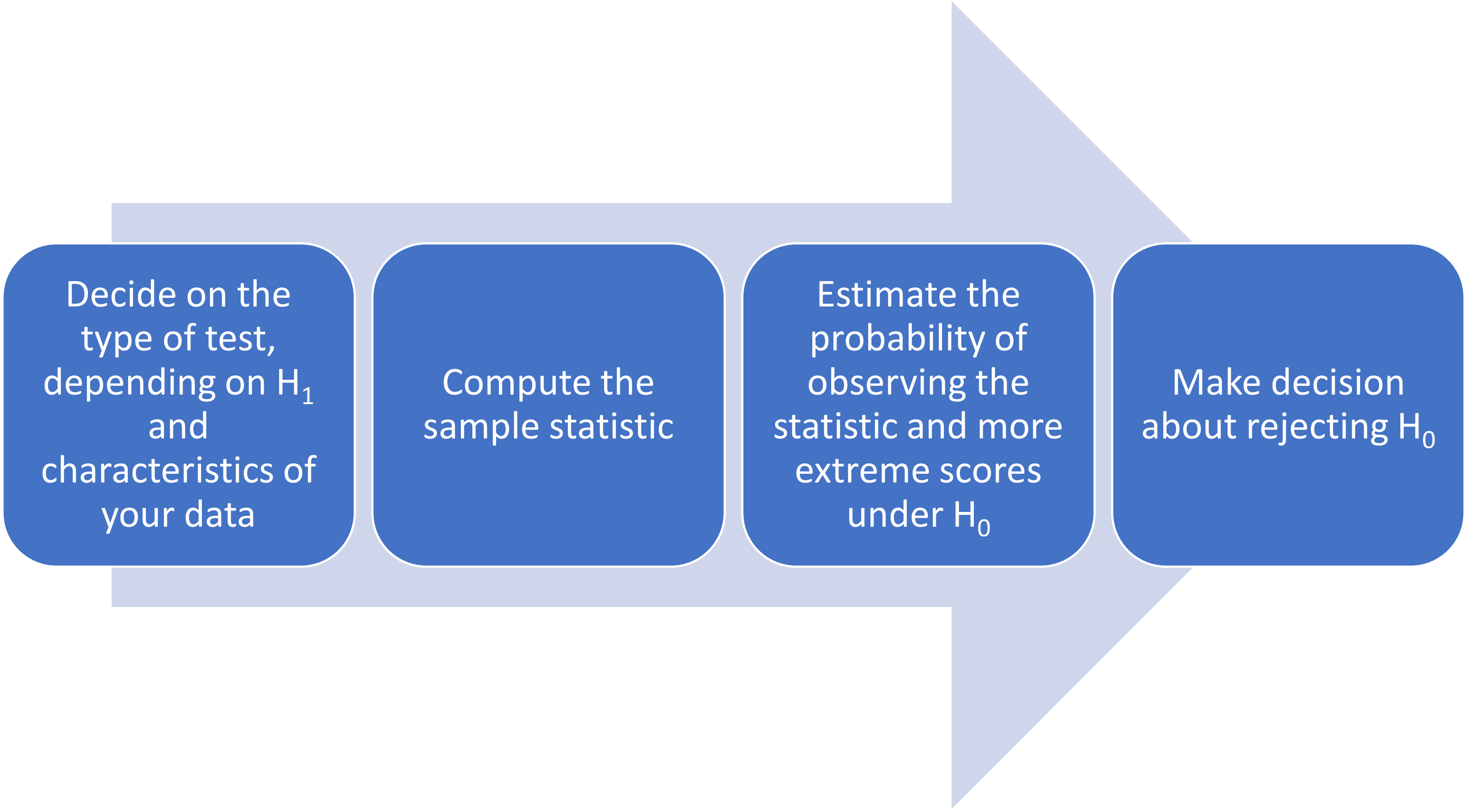
One-tailed test: Computing the probability

- P observing 8 negative mentions = 0.044
- P observing 9 negative mentions = 0.01
- P observing 10 negative mentions = 0.001

- Total P = 0.044 + 0.01 + 0.001 = 0.055

This is the p-value! The p-value of observing 8 and more negative mentions is 0.055. I

Hypothesis testing algorithm



Decide on the type of test, depending on H_1 and characteristics of your data

Compute the sample statistic

Estimate the probability of observing the statistic and more extreme scores under H_0

Make decision about rejecting H_0

One-tailed test: result

- The p-value is somewhat greater than 0.05, the conventional threshold. We cannot reject the null hypothesis.

Two-tailed test

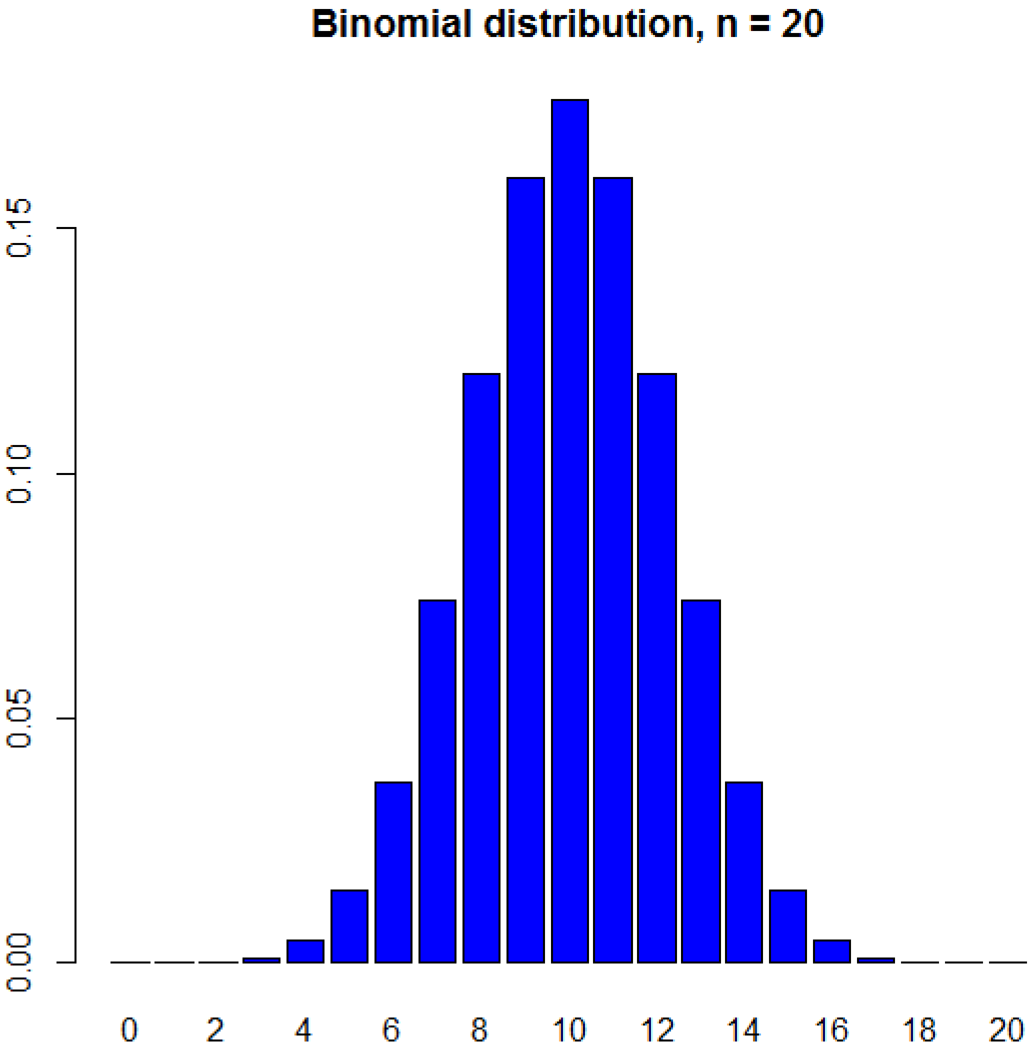
- If your alternative hypothesis was bidirectional: some bias (either negative or positive), then you should sum up the probabilities of all outcomes of 8 and greater and those of 2 and smaller (i.e. 0, 1, 2).
- This is called a two-tailed test because you're looking at two tails of the distribution.
- P observing 8 and greater = 0.055
- P observing 2 and less = 0.055 (mirror image!)
- P observing both = 0.055 + 0.055 = 0.11.
- Again, we cannot reject the null hypothesis.

A word of warning

- You see that it is easier to observe a significant result if the test is one-tailed.
- You shouldn’t change your alternative hypothesis if you don’t like the results of the test (when it’s not significant)!
- Always use the one-directional test? Very risky! It can backfire if you get the direction wrong. For example, if you expect a negative bias, and observe only two negative outcomes, then you get $p = 0.989$.
- The cut-off point (0.05) is called the significance level. It should not be changed (unless you know very well what to do!).

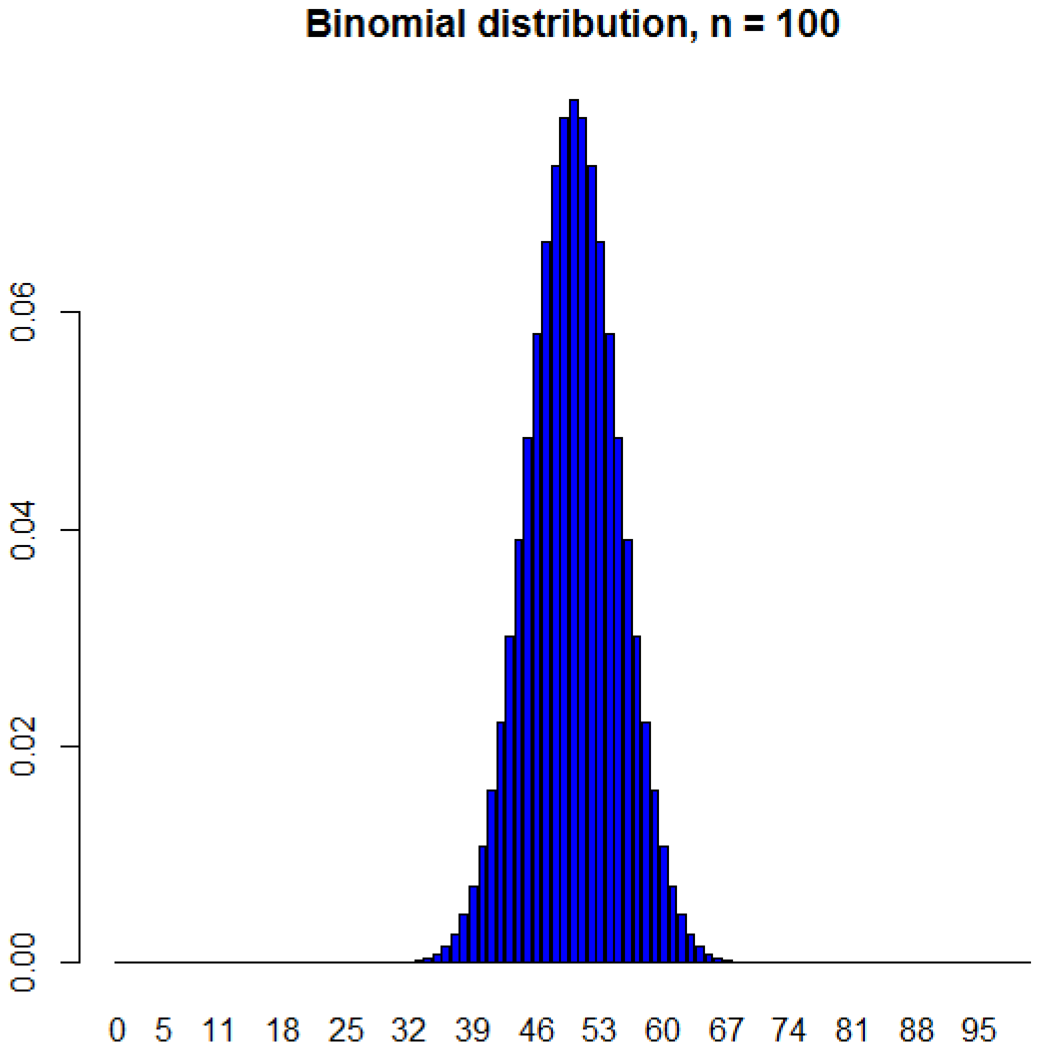
Sample size matters

- Now, imagine that you’ve worked very hard and found 10 more mentions of the word “immigrant”. Now you have 20 in total, 16 negative mentions and 4 neutral or positive mentions.
- The proportion remains the same: 0.8, or 80%.
- However, the p-value will change.
    - One-tailed test: 0.006
    - Two-tailed test: 0.012


Binomial distribution, n = 20
0.00
0.05
0.10
0.15
0
2
4
6
8
10
12
14
16
18
20

Even more data

- You work even harder and find 100 mentions! 80 of them are negative, and 20 are neutral or positive.
- Again, the proportion of negative mentions remains the same: 0.8, or 80%.
- But the *p*-value drops very dramatically:
    - One-tailed: 0.000000001
    - Two-tailed: 0.0000000006


Binomial distribution, n = 100
0.00
0.02
0.04
0.06
0
5
11
18
25
32
39
46
53
60
67
74
81
88
95

Effect size vs. significance

- The proportion (0.8) represents the effect size (how strong the bias is). It does not depend on the sample size.
- The *p*-value is a measure of statistical significance. It reflects how confident we can be that the result we observe is not due to chance alone.
- More data => smaller p-values => more confidence!

Some good news

- You will not have to compute the *p*-values manually. R will do it for you.
- What is more important, is
  - to have a sufficiently large representative sample
  - to formulate your hypothesis
  - to choose the right test
  - to interpret the results correctly