A quick and dirty introduction to R for linguists

Alessandra Zarcone
GK SFB 732
21/01/2011

Overview

- ☐ WHY? why statistics? why R?
- ☐ WHO? who needs statistics anyway?
- ☐ WHAT? descriptive statistics (part 1), inferential statistics (part 2)
- ☐ HOW? mini-tutorial on R on descriptive statistics (part 1) and inferential statistics (part 2)

WHY?

why statistics?

Why statistics?

- “You’re not thinking statistically, Martin!”
- Properly design your experiment / study
- Choosing the right tool to analyze your data
- Presenting / comparing / publishing results (don’t let the reviewers bite!)

Why R?

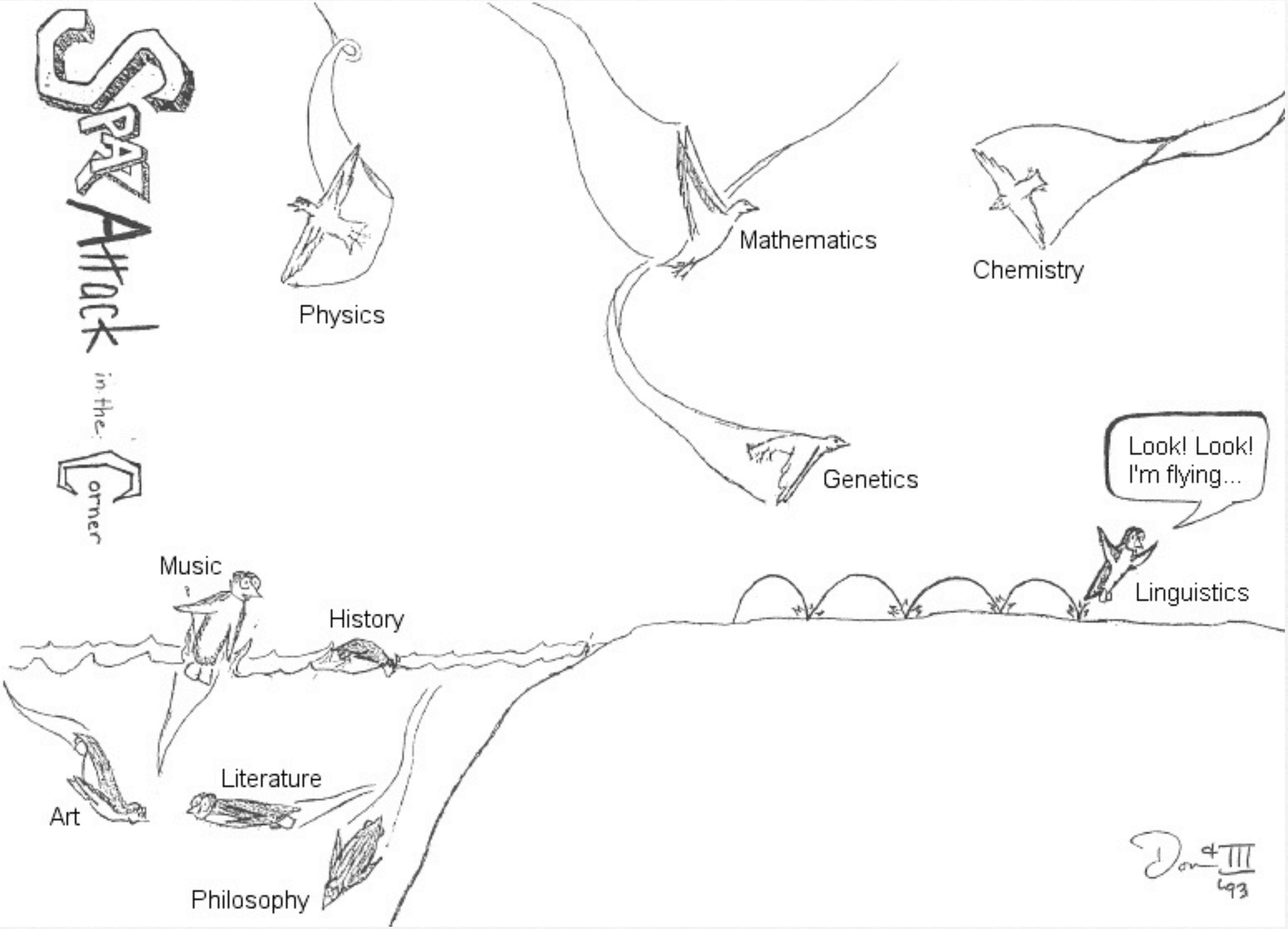
- interactive programming environment
- Free (as in “free beer”) and
  free (as in “free speech”)
- Invites you to THINK about your data
- Powerful
- Almost a standard

WHO?

who needs statistics anyway?

Who needs statistics?

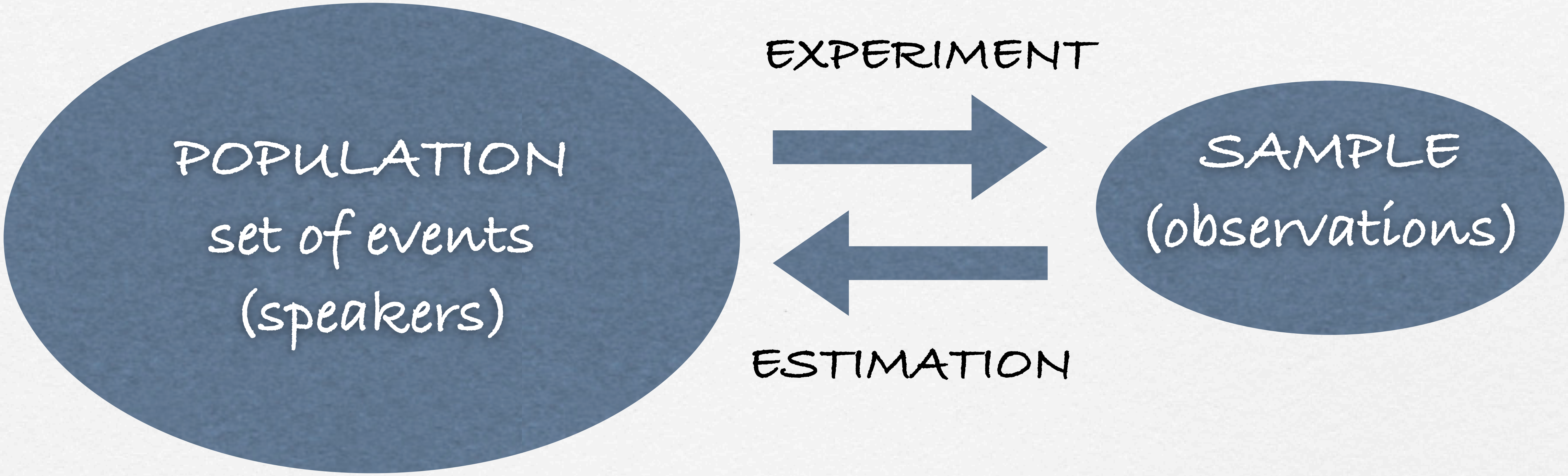
- Cinzia, theoretical linguistics: grammatical judgement studies
- Lukas, computational linguistics: plausibility judgements
- Kerstin, corpus linguistics: corpus frequencies
- Kati, phonetics: acoustic data, continuous and categorical
- Alessandra, psycholinguistics: RT data


Spaz Attack in the Corner
Physics
Mathematics
Chemistry
Genetics
Look! Look! I'm flying...
Linguistics
Music
History
Art
Literature
Philosophy

WHAT?

the ABC

What is statistics?



<u>DESCRIPTIVE statistics:</u>

how to visualize and
present your data
how to show patterns

<u>INFERENTIAL statistics:</u>

how to estimate characteristics
of the population based on the
sample: do the events in the
sample data occur by chance?

The ABC: Variables

- what we measure or manipulate in the study
- DEPENDENT variable (DV): measured / registered
- INDEPENDENT variable (IV): controlled / manipulated - it causes a change in the DV
- e.g. in a priming experiment we want to compare the RT of "chair" after "table" (related prime) and after "bread" (non-related prime)

The ABC: Variables

- CONTINUOUS variables take any value
  e.g. RT, corpus frequencies
- DISCRETE variables take only a small set of possible values
  e.g. gender, valency, voice

The ABC: Factors and Levels

- a discrete IV is often called FACTOR
- the possible values of a factor are called LEVELS

  e.g. in a priming experiment we want to compare the RT of "chair" after "table" (related prime) and after "bread" (non-related prime)

  - prime is a FACTOR

  - related/non-related are its LEVELS

The ABC: Type of data

- DISCRETE variables:
  - NOMINAL scale: if an item belongs to a category (male/female, trans./intrans.)
  - ORDINAL scale: rankings (1st most grammatical sentence, 2nd, 3rd...)
- CONTINUOUS variables:
  - INTERVAL scale: rankings on a scale (this item has plausibility 1, 2, 5..)
  - RATIO scale: the scale has a 0 point (e.g. temperatures)

HOW?

getting started with R

Getting started with R

☐ Open R

☐ Open file “GK210111.R” with R

☐ Import “heid”.. and play around

WHAT?

descriptive statistics

Descriptive statistics

- the "identity card" of your data:
  how do they look like?
  (measures of central tendency and variability)
- how to visualize and present your data (graphics)

Measures of central tendency

- MEAN: sum of the values divided by number of observations
- MEDIAN: if we rank our observations in numerical order, the median is the middle value

A measure of variability

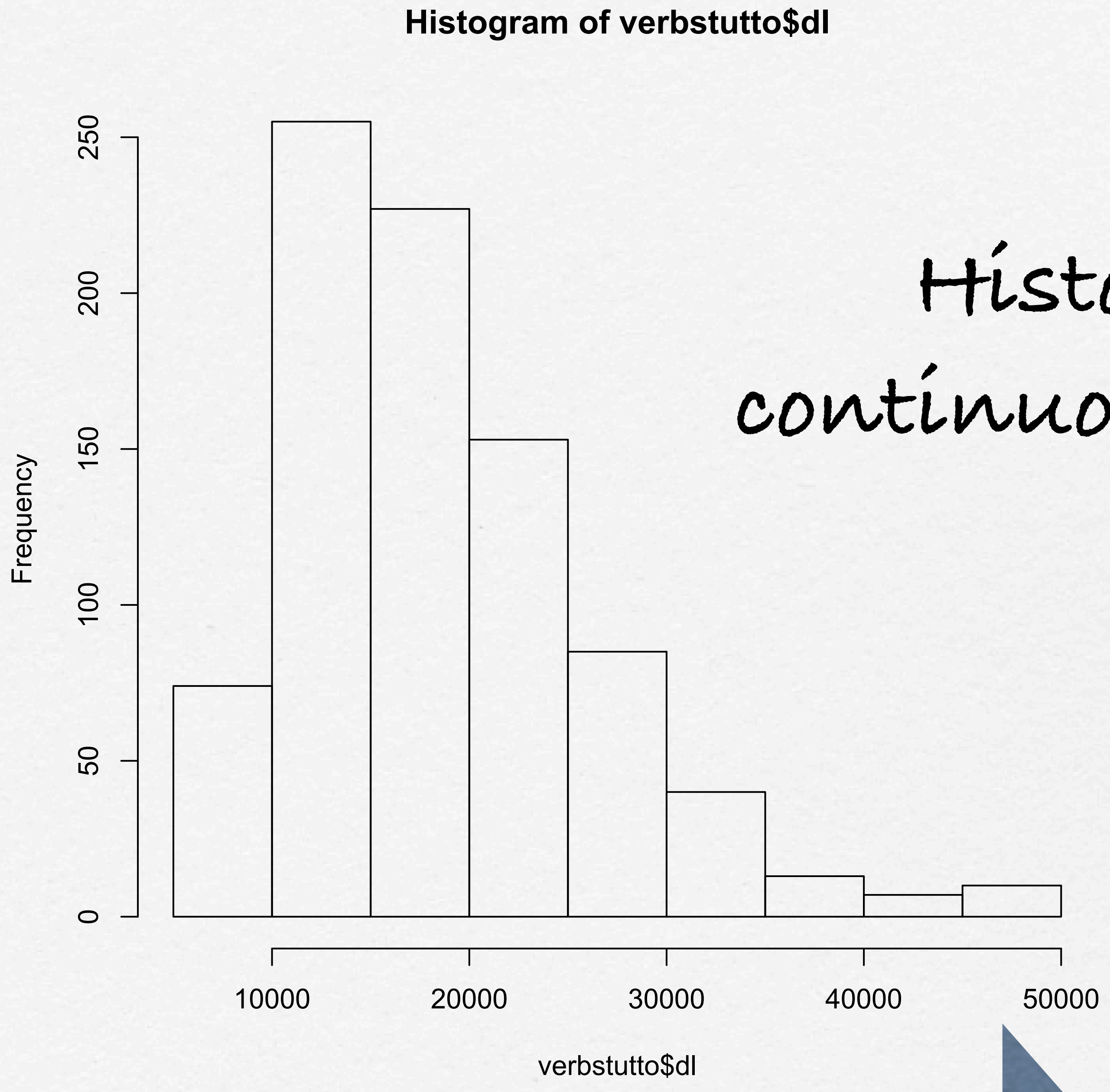
- STANDARD DEVIATION: how much "dispersion" there is from the mean
    - low SD: the data points tend to be very close to the mean
    - high SD: the data points are spread out over a large range of values

Distributions

Participant	Word	log(RT)
pp1	basaalheid	6,69
pp1	markantheid	6,81
pp1	ontroerdheid	6,51
pp1	contentheid	6,58
pp1	riantheid	6,86
pp1	tembaarheid	6,35
...	...	...

A table: not very easy to read

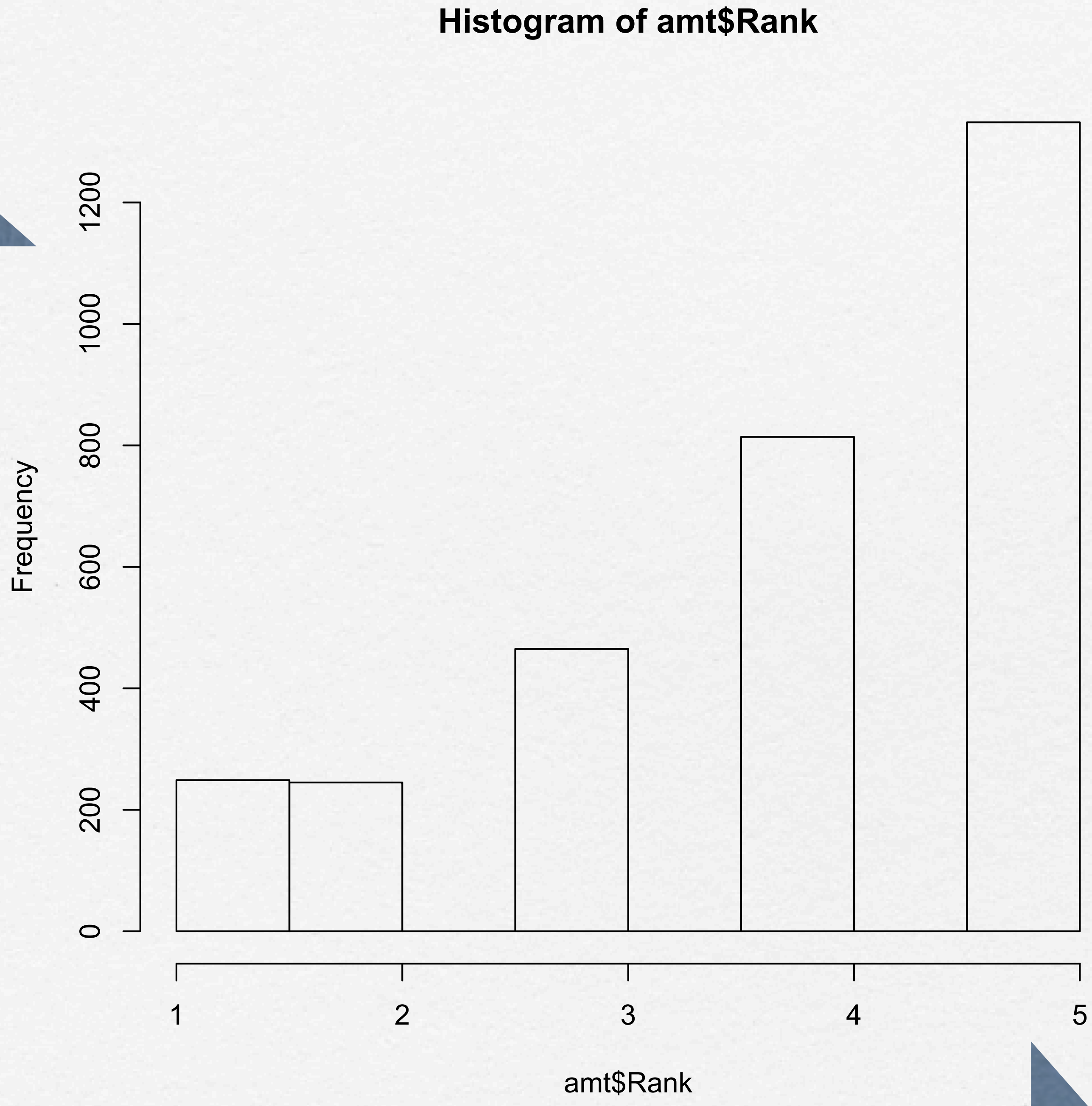
Histogram: continuous variable



Frequency (how many values)

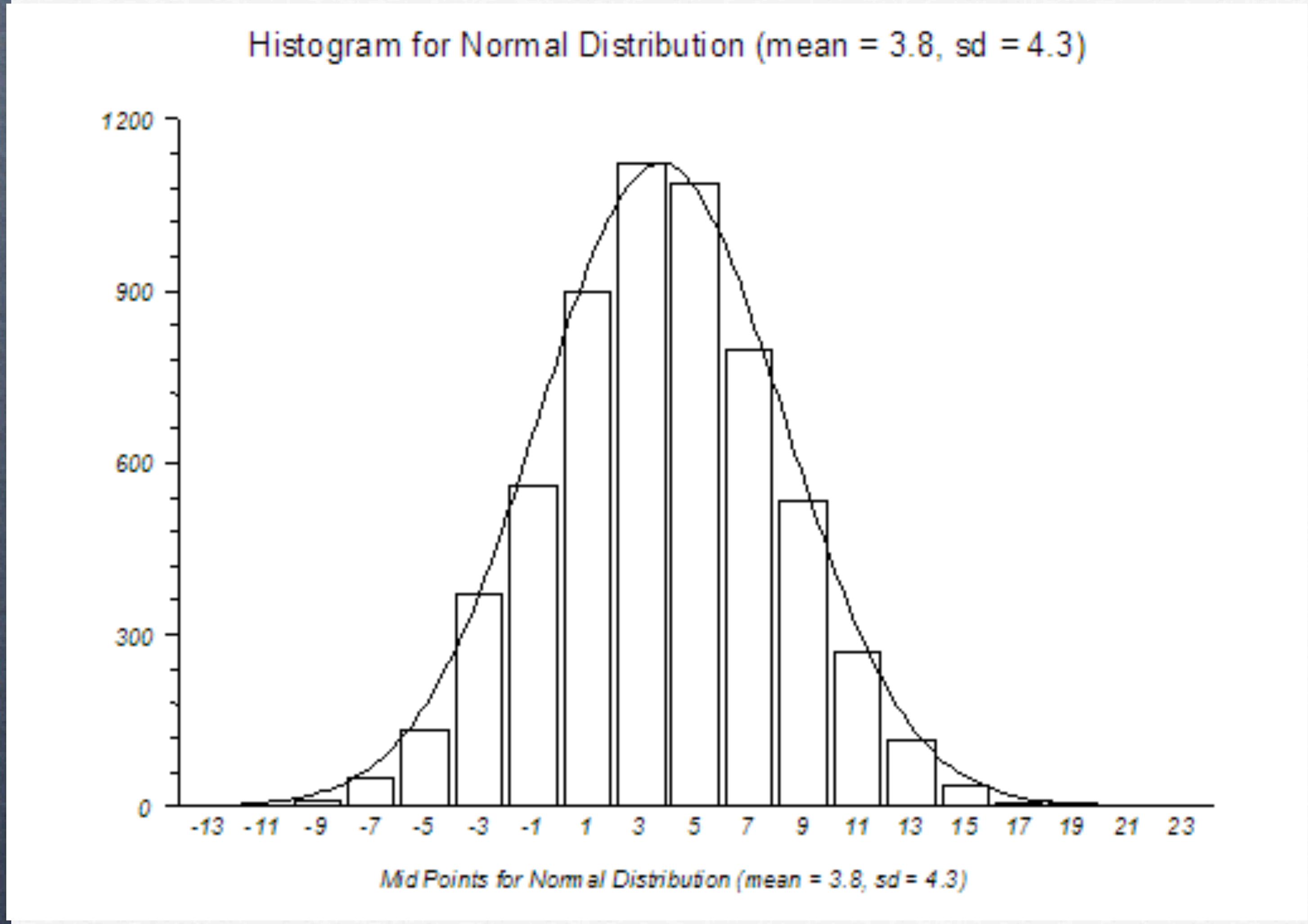
Values (e.g. RT)

Histogram:
discrete variable

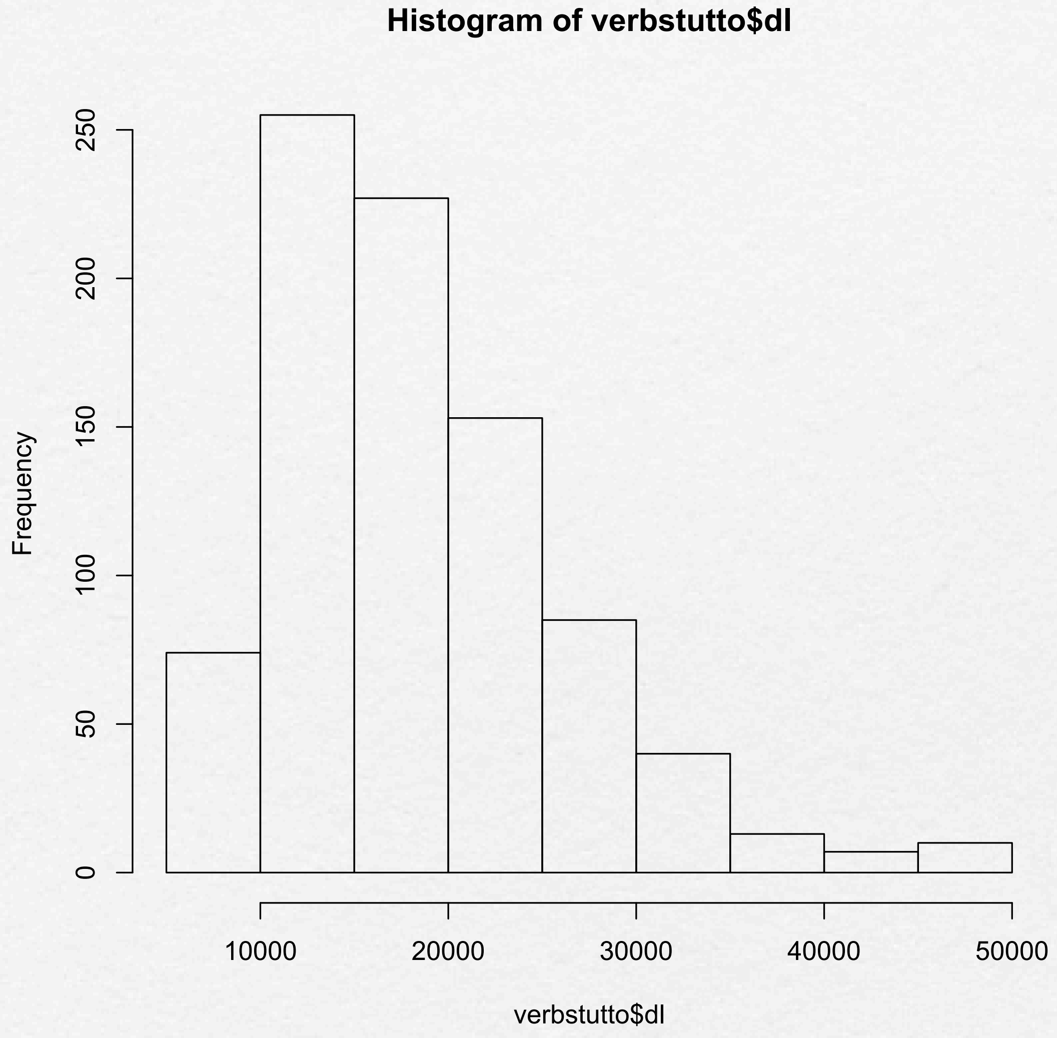


Frequency (how many values)

values (e.g. rankings)



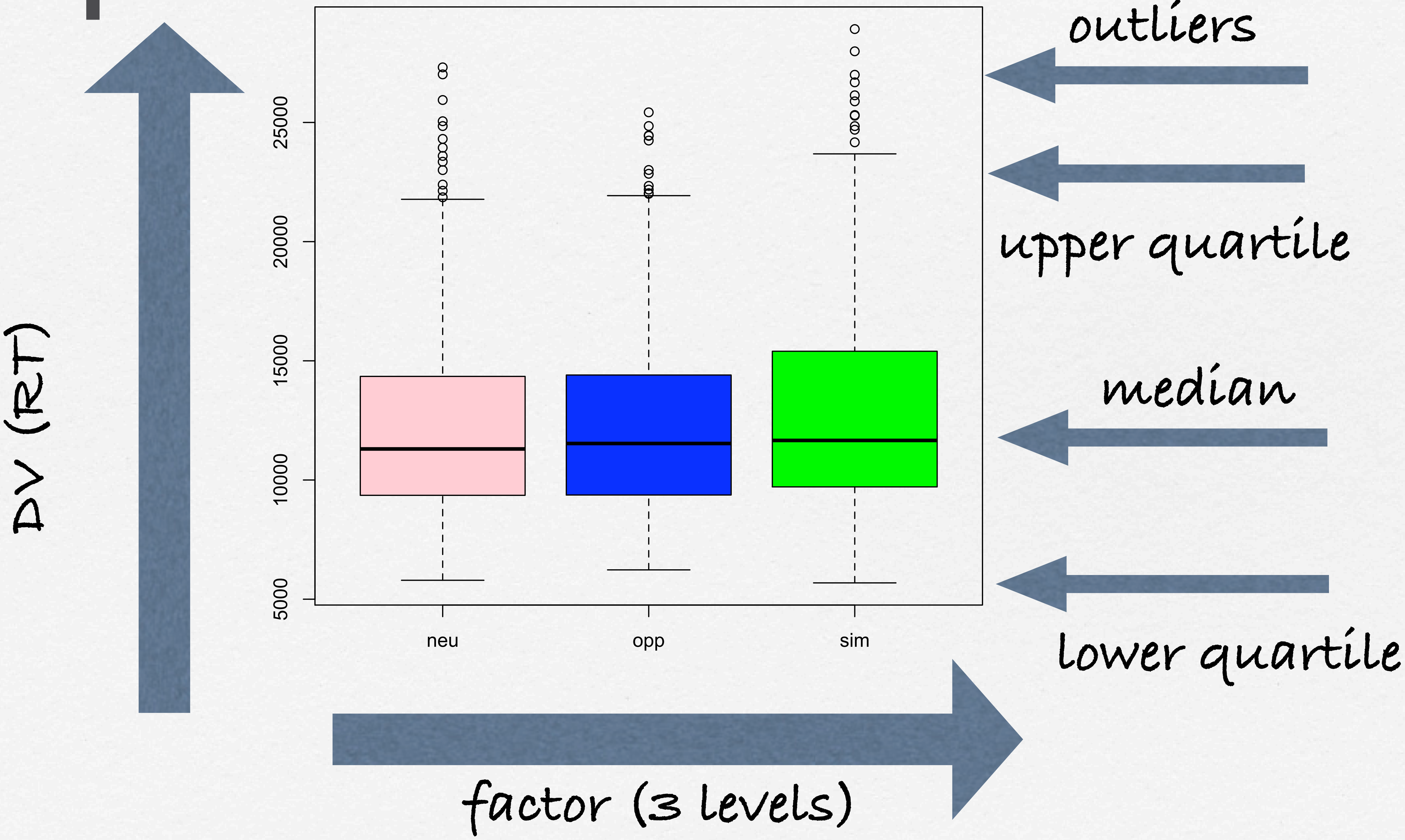
NORMAL
distribution



SKEWED
distribution

Graphics

Box and whiskers plot (continuous DV, one factor)



outliers

upper quartile

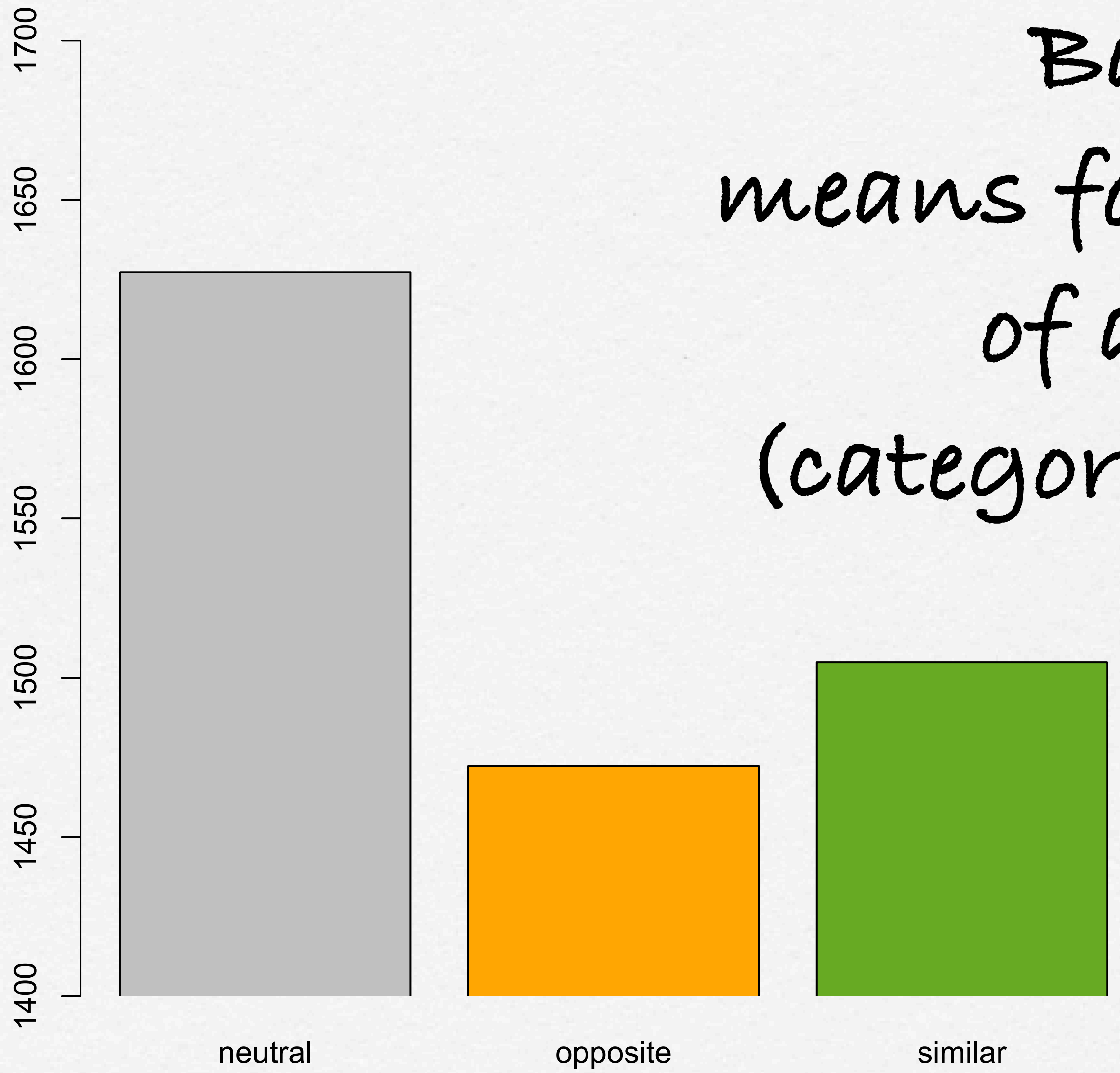
median

lower quartile

DV (RT)

factor (3 levels)

Graphics



Barplot:
means for three levels
of a factor
(categorial variable)

HOW?

descriptive statistics with R

WHAT?

inferential statistics

Inferential statistics

- do the events in the sample data occur by chance?
- do two samples come from the same population?
- how to estimate characteristics of the population based on the sample

The Null Hypothesis (H_0)

- Inferential Statistics, testing H_0: do two samples come from the same population?
  - H_0: There is no difference between samples
  - H_1: There is a difference between samples
  - REJECTING H_0

An example

- Priming experiment:
we want to compare the RT of "chair" after "table" (related prime) and after "bread" (non-related prime)
    - sample 1: RT with related prime
    - sample 2: RT with non-related prime
- REJECTING H_0: we want to reject the hypothesis that the two samples belong to the same population

Statistical test and significance

- A STATISTICAL TEST tells us if we can reject H_0
- Its significance level is a probability of wrongly rejecting H_0, if it is in fact true
- e.g. $p = 0.001$
  the probability of wrongly rejecting H_0 is 0.001

Small is good

- the smaller the better:
  the smaller the p,
  the more likely it is to replicate the result
- human sciences: $p < 0.5$ is usually good

To each its own test

- There are a lot of different tests to use
- The differences can be difficult to grasp,
  but learning to identify the most appropriate test
  for your data is time worth spending!

Decisions...

- Is my distribution normal? (Parametric vs. Non Parametric tests)
- Continuous vs. Discrete DV
- One, more DV
- Correlations, differences

Correlations

- E.g. in an experiment where the participant have to estimate the size and the weight of an item (e.g. "apricot", "elephant"), is there a correlation between estimated size and estimated weight?
- the correlation is the (linear) relationship between two random variables, in range -1 +1 (-/+1 strong correlation, 0 no correlation)

HOW?

inferential statistics with R

correlation

T-test

- One-sample t-test:
  we have a sample of RT (one-sample), we want to know if it differs from the population mean
  (e.g. log(RT) = 6.7)
- Two-sample t-test:
  we have two samples of RT (e.g. RT with related prime, RT with), we want to know if they differ significantly
- Paired two-sample t-test (repeated measures)
- Watch out: we are assuming
  continuous DV, normal distribution

Predictors

- Linear regression models
  one predictor (IV), one dependent variable
  $y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$
  e.g. prime and RT, length and frequency, etc.
- Multilinear regression models
  multiple predictors (IVs), one dependent variable
  $y_i = \beta_0+\beta_1 x_i+\ldots+ \beta_k x_k+\varepsilon_i$
  e.g. prime and valency on RT, etc.

HOW?

inferential statistics with R

More statistical tests

- if the DV is not continuous but discrete...
- if the continuous DV is not normally distributed...

90 minutes are not enough but...

Help with R

- http://cran.r-project.org/doc/contrib/Short-refcard.pdf
- http://rseek.org

Further readings:

- Baayen, R. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge University Press.
- Gries, S.T. (2009). Quantitative corpus linguistics with R. Routledge.