How to set up and test a hypothesis?

Types of data and samples. Scale types, relevant measures and their visualisation

Katalin Mády

Pázmány Summer Course in Linguistics for Linguistics Students, Piliscsaba

25 August, 2013

What is this course good for?

- You will learn how to set up an experiment with linguistic data.
- You will learn which statistical method is best suited for your analysis.
- You will be able to understand descriptions of experiments done by others.
- You will get help how to learn to analyse your own data.

☹What this course cannot give you is practical do-it-yourself training in data analysis.

Course website with slides and training material:
`clara.nytud.hu/~mady/courses/statistics/scills2013`

What is statistics good for?

It helps you to answer questions like:

- Are British faster readers than French?
- Does a cocktail diet help people to lose weight after four weeks?
- Do 6-year-old bilingual children have better cognitive abilities than monolinguals?
- Are 22 hesitations significantly more than 18 hesitations?
- Is the planet hotter than it was 100 years ago?

What is statistics not good for?

Many questions cannot be answered by statistics.

- Dark chocolate is more delicious than milk chocolate.
- Rats are the ugliest animals, followed by spiders.
- Women are ideally suited for being housewives and mothers.

Reason: test data must be quantifiable, i.e. they have to be **numbers**.

Quantitative and qualitative data

Quantitative data: countable entities, i.e. some kind of numbers.

Qualitative data: detailed description of observations, e.g. differences between the taste of chocolate types, the reasons why people dislike rats and spiders, or the social situation of women as housewives.

Qualitative data can be quantified in many cases. E.g. the preference for chocolate types can be scored on a scale between 1 and 5, or the satisfaction rates for housewives and female managers can be compared.
Often it is necessary to collect qualitative data first in order to find out the relevant factors for a quantitative analyis (like in sociolinguistics).

Initial observation

The starting point for an experiment is always a tentative observation.

- This year there are more wasps around than in recent years.
- Males prefer to make phone calls, while females prefer to send text messages.
- Females tend to use backchanneling more than males.

Generating theories

Since experiments are done out of scientific interest, you will always have some ideas how to explain that there are more wasps this year (if it turns out to be true).

Potential explanations:

- There were ideal weather conditions in their fertilisation period.
- They have become resistent to current remedies.

Experimental design

In order to test the theories, comparable data have to be collected:

- Number of wasps in a given area with sunny, wet, cold and warm periods in the relevant part of the year.
- Comparison with first years when current remedies were introduced.

Categories that you control for are called **independent variables**, such as weather during the fertilisation period.

The data you are collecting (number of wasps in different regions in the same year) are called **dependent variables**, since they are effected by, i.e. dependent on the conditions they were collected in.

Background

Basic ideas:

- The repeated observation of a fact does not mean that it is the rule or the truth. If you only observe white swans it does not mean that all swans are white.
- Falsification: try to find a swan that is not white! As long as the result is negative (no black swans), the hypothesis "All swans are white" is valid.
- Basic requirement for hypotheses: there must be a way to falsify them. The experimental method has to be chosen so that the hypothesis can be falsified.

Requirements

Operationalisation: the question must be formulated so that it can be answered by empirically observable data.
"People speak less careful nowadays than in former times."
Problem: What measures are appropriate?

Validity: does the measure really measure what it is meant to measure? Physiological measures, reaction times, counts etc. are valid measures. Validity of scores such as in rating scales has to be tested carefully.

Reliability: results must be reproducible with the same experimental design and methods. Important: detailed description of experimental methods so that they can be reproduced in exactly the same way.

Objectivity: results must be independent of experimental environment and researchers. E.g. female vs. male, native vs. non-native experiment leader.

Causality

It is sometimes not clear whether the independent variable explains the effects by itself.

- "Low self-esteem causes dating anxiety." There might be a third reason such as poor social background that explains both factors.
- "Women with breast implants commit suicide more frequently." It is unlikely that silicon has an effect on the psychological constitution.
- "There are less newborn children per year than 30 years ago. Also, there are less storks around." It is not clear whether these facts have identical or different explanations, but it is unlikely that less storks can deliver less children.

Goal: to rule out all other potential explanations and to show that their presence or absence has no effect on the observations.

Hypothesis

Hypothesis: in general: an assumption. Here: a preliminary answer to a scientific question.

Experimental hypothesis: a statement on the relationship of the variables.

Statistical or stochastic hypothesis: a certain event will occur under certain circumstances with a certain probability.

Procedure of hypothesis testing

"Adult males are taller than adult females.".

Alternate hypothesis (H_1): Adult men are taller with a certain probability than adult women.

Null hypothesis (H_0): Adult men and women are equally tall. Why? "Look for non-white swans!"

Statistical procedure: testing the null hypothesis. If H_0 is more probable than a pre-defined threshold $\rightarrow$ H_0 is to be kept as the current hypothesis. If H_0 has low probability $\rightarrow$ it is refused and we assume that H_1 is true.

Basic terms

Population: the sum of elements to be investigated, finite or infinite. It is seldom that all elements can be studied.

Representative sample: it reflects the population according to its relevant features. E.g. a sample of 1000 students where the proportion of disciplines reflects those for all students of that country.

Random sample: each element has the same chance to be chosen. Not always the case in linguistic experiments where most subjects are university students of linguistics...

Variables

- Qualitative: a feature or characteristic (born in February, female, Georgian, verb etc.).
- Discrete: countable, finite (number of errors in a test, age in years).
- Continuous: any real number within a given interval.
- Categories or groups: merged countable variables (e.g. age between 25 and 34 years). It often allows for easier handling, but means information loss.

Scale types

Nominal scale: the values of the variable can be distinguished, but there is no relationship between them. (Gender, religion, hair colour, parts of speech.)

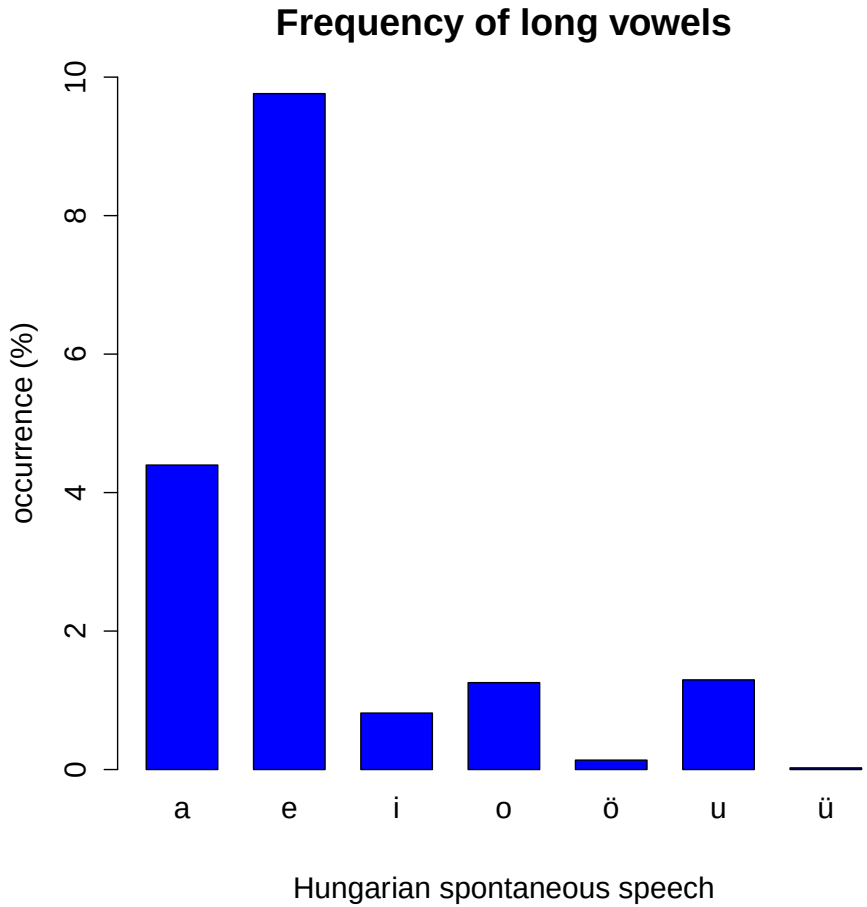
Ordinal scale: values can be ordered, but the distance between units is not equal or not interpretable. (Educational degrees, school grades, many types of scores.)

Metrical scales: the multiples of a given measure. Both proportions and multiples of the values can be interpreted, thus distances are comparable and interpretable.
Interval scale: zero is set arbitrarily (e.g. Celsius). Difference between values can be interpreted, but not their ratio. 20 °C is not twice as warm as 10 °C.
Ratio scale: zero is set in physical terms, rations can be interpreted. (Distance, weight, energy, Kelvin).

Central tendencies: mode

The most frequent category in the sample.

Frequency of long vowels in Hungarian.



Mode is relevant for all scale types.

Central tendencies: median

The middle element in an ordered sequence. If sample n has an even number of elements, then the mean of the two middle elements.

How many Facebook friends do my Facebook friends have?

Random selection of 11 Facebook friends.
Number of their friends:

546 388 724 269 113 467 682 178 149 382 196

Scores in ascending order:

113 149 178 196 269 382 388 467 546 682 724

Middle score: 6th element = 382.
If sample has an even number: mean of two middle scores.

No median can be calculated for nominal data, even if they are encoded by numbers.

Central tendencies: mean

Average of all values, i.e. sum divided by number of all cases.
Mean of Facebook friends of my friends:

```
mean = (546+388+724+269+113+467+682+
+178+149+382+196)/11 = 382.1818
```

Mean is a statistical model and does not necessarily mean realistic data. Nobody has got a 0.1818 friend.

Important: since the mean is based on equidistant data (= all distances are equal), it can only be applied to metrical data. Ranking scales from 5 steps are often regarded as equidistand and thus metrical.

Thus, averaging marks from school is illegal in statistics.

Median or mean?

Imagine one of your friends has just joined Facebook yesterday and has only 11 friends. Another friend is a famous actress and has 5439 friends.
In this case, the mean is

```
mean =
(11+149+178+196+269+382+388+467+546+682+5429)/11 =
791.5455
```

If you look at a larger number of Facebook friends, you will find that very few people have as few as 11 friends, and that 5429 friends are also uncommon.

It is useful to check whether you have such improbable or untypical values by visualising the data (see later today) or just comparing the mean with the median.

The median of these data is still 382 which is still a realistic number – at least for friends in my age...

Description of data

Relevant measures:

- frequency of certain classes,
- distribution,
- central value,
- deviation.

These measures can be illustrated by two-dimensional figures.

Frequency

- Raw value (if samples are of the same size),
- ratio (number of certain cases/all cases), percentage (ratio*100) useful if sample sizes differ,
- cumulative frequency: occurrences below a given value.

Frequency is often calculated for groups. Instead of 21 years, 23 years, 35 years, 43 years 20–29 years, 30–39 years etc.

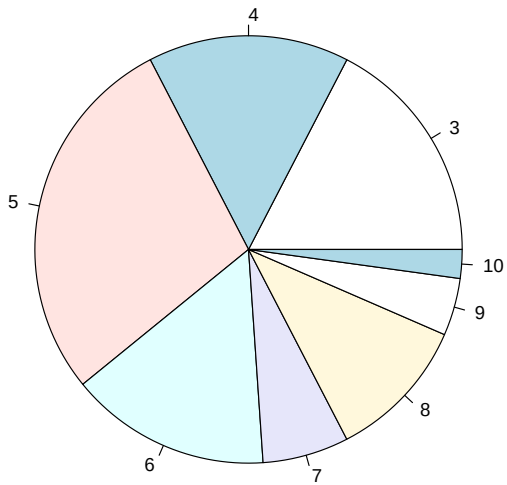
An example

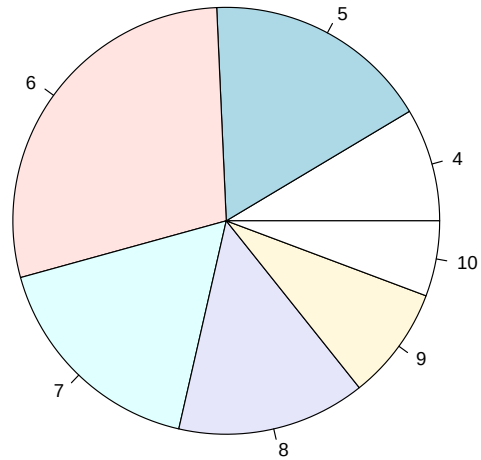
Length of words for animals and plants in English, expressed by the number of letters.

type	sample size
animals	46
plants	35

Pie chart

Frequency of 3-letter words, 4-letter words etc.
Right: animals, left: plants.

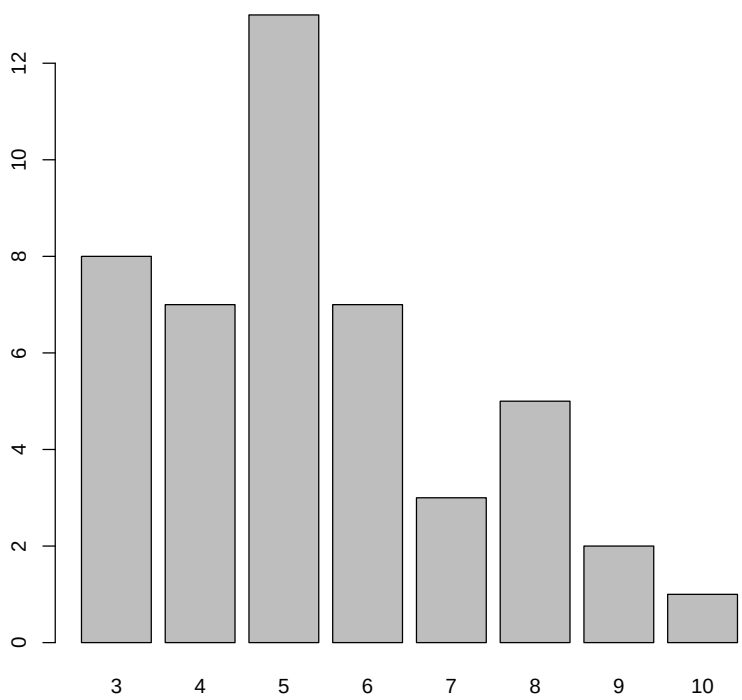


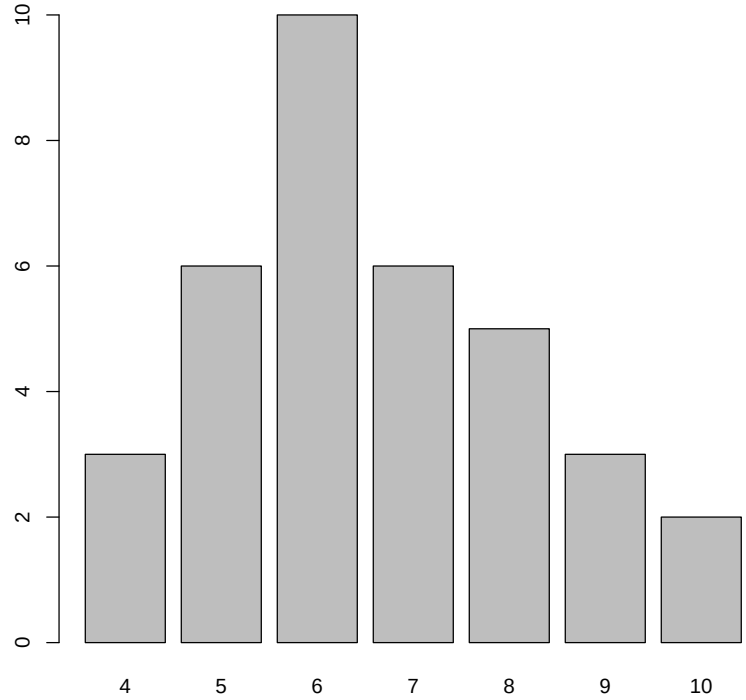


Advantage: the pie is always 100%, thus samples of different size can be compared.
Disadvantage: no possibility to compare each category directly.

Barplot

Length of animal (left) and plant (right) names in raw numbers

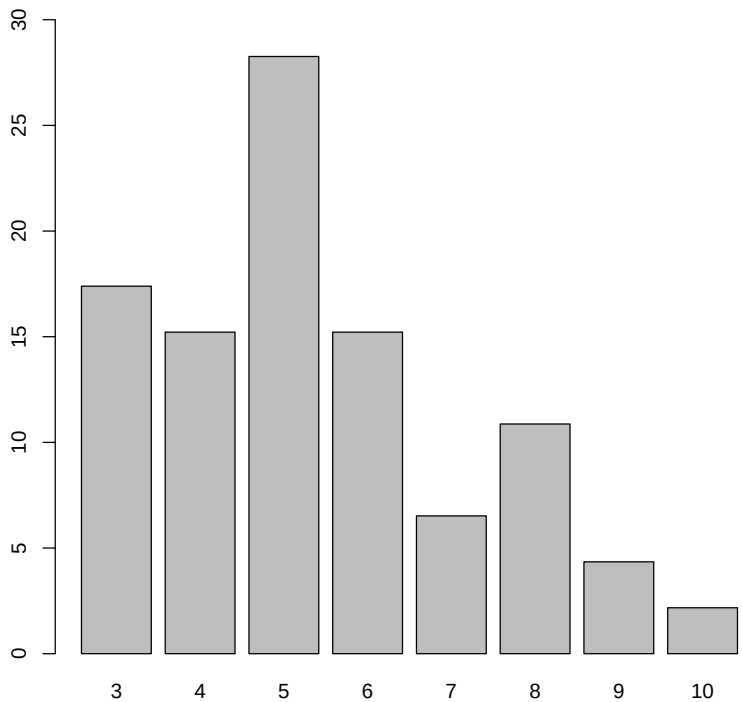


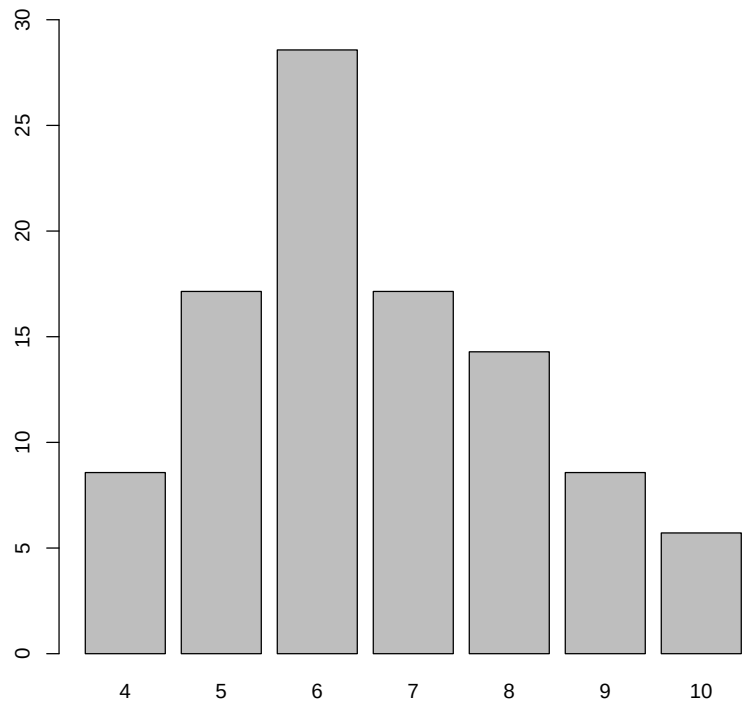


Usage: for nominal data, ordinal discrete data, categorised data.
Problem: size of samples is different ($n_a = 46$, $n_n = 35$).

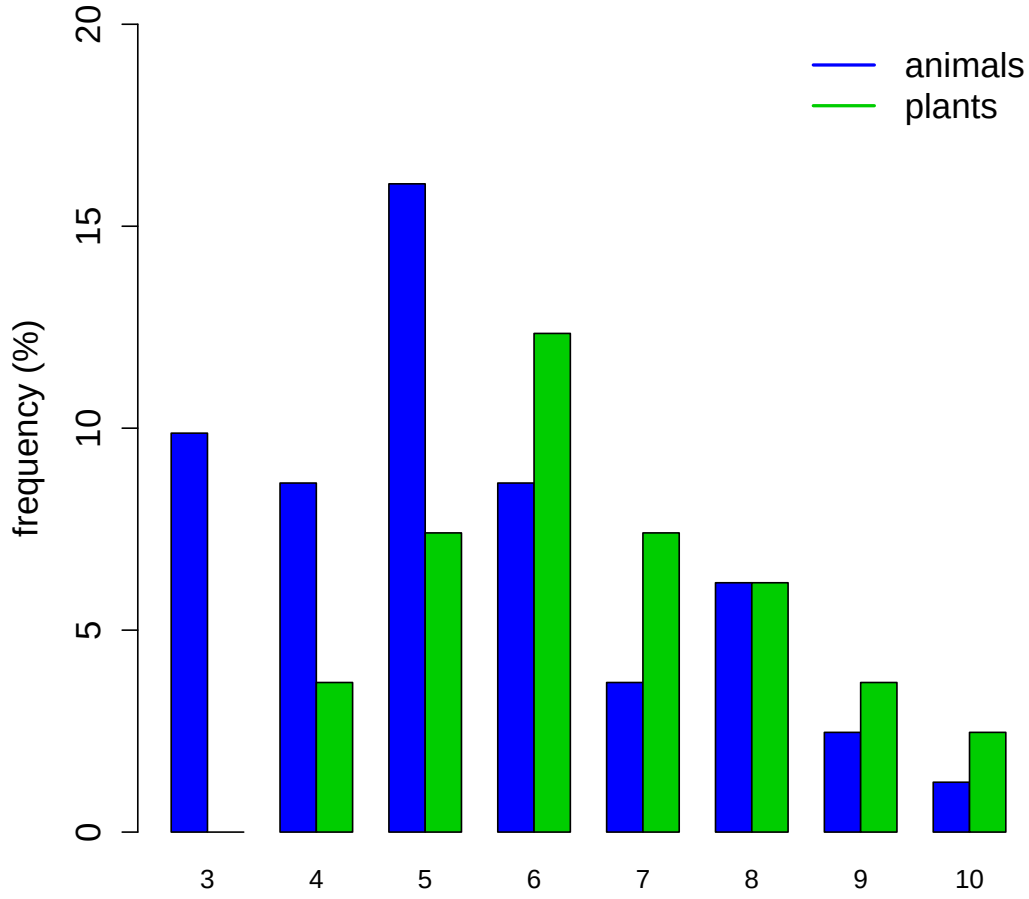
Barplot with percentages

Length of animal (left) and plant (right) names in percentage





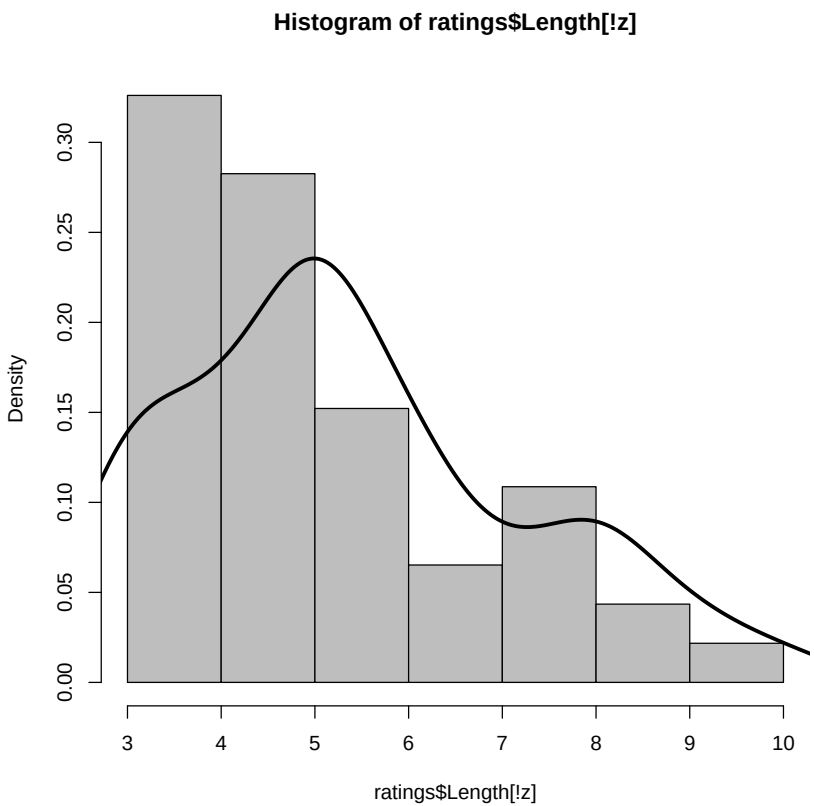
One barplot with two samples

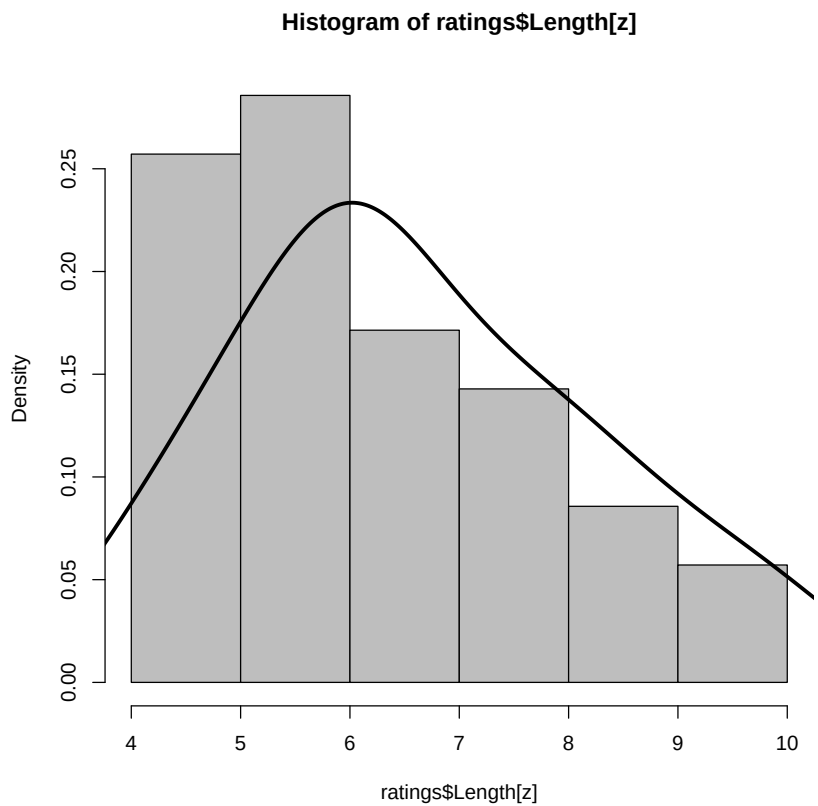


Advantage: direct comparison of groups is possible.

Histogram

Histogram and density of length of animal (left) and plant (right) names





Usage: data have to be ordinal or metrical.

Distribution

- Definition: how often elements occur in a scale.
- Usage: for ordinal and metrical data.
- Procedure: interpolation between continuous or discrete values.
- Importance: basis for statistics based on probability calculations.

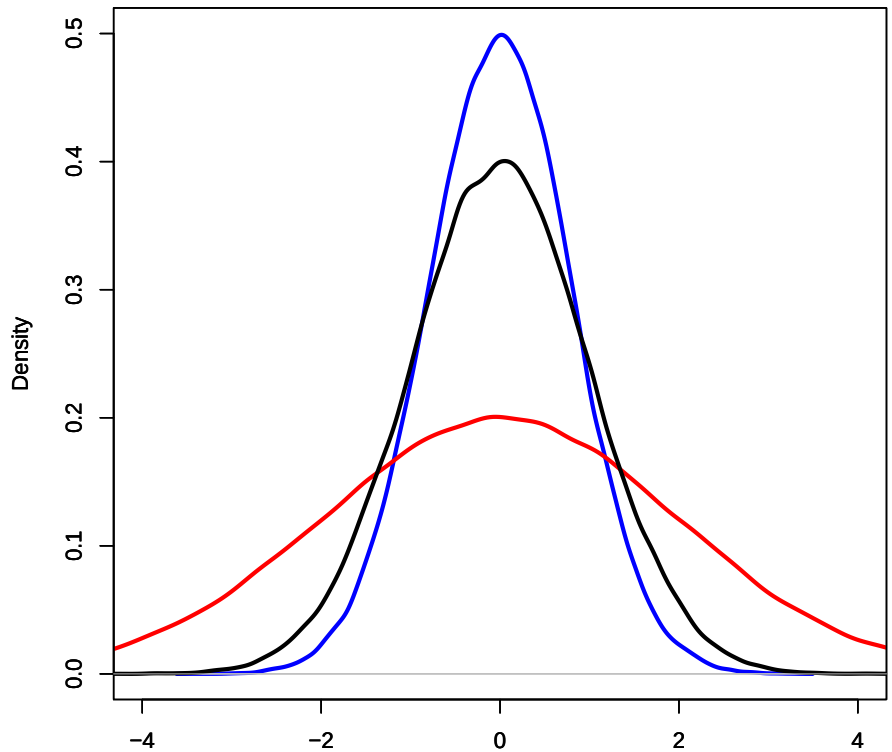
Distribution types

Even distribution

Number of pips when throwing with a die

Distribution types

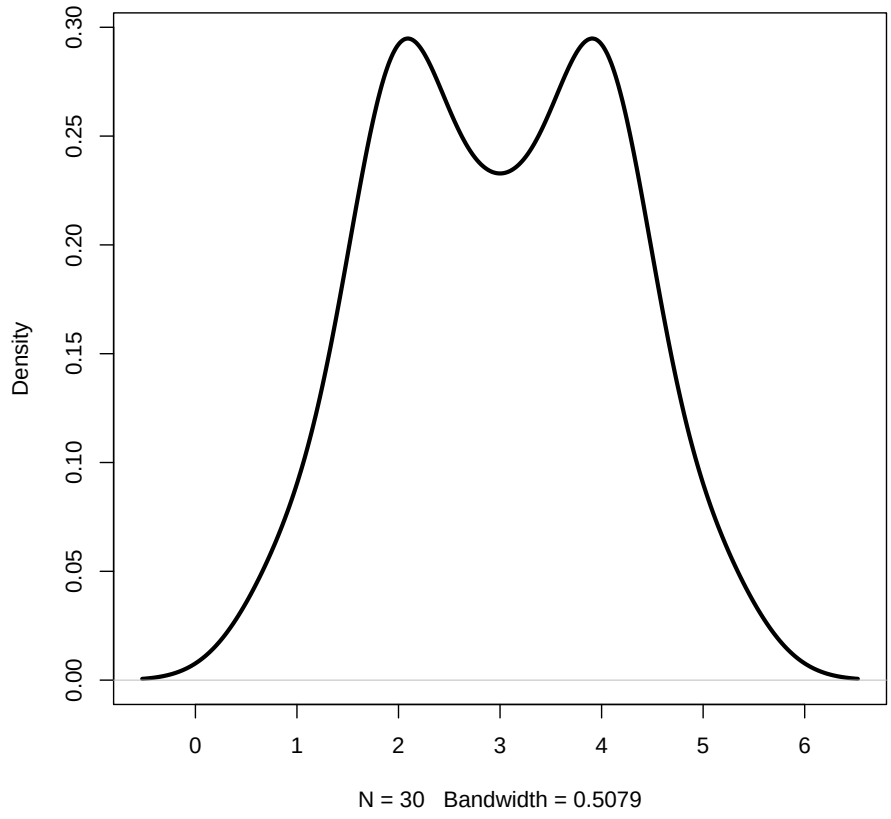
Unimodal: there is only one mode.



It can be symmetric or asymmetric, and the flatness of the curve can be different.

Distribution types

Bimodal: there are two modes.



A bimodal distribution indicates that the sample actually contains two samples. Most statistical tests can only be performed on unimodal data.

Dispersion: range

Dispersion: how the data are spread. It gives information on the width of the distribution. E.g. the read line on slide 31 has a larger dispersion than the blue one.

Range: difference between minimum and maximum value. Applicable to ordinal and metrical scales, but sensitive to extreme values.

In the first case of the Facebook friends example:
`range = 724-113 = 611`

In the second case:
`range = 5439-11 = 5428`

Problem: the first value gives a better estimation of the range, since friends with less than 50 or more than 1000 friends are probably uncommon.

Dispersion: interquartile range

Lower and upper 25% of the scale might contain extreme scores. A way around: cut off these scores.

- Quartiles: they split data into four equal parts.
- Second quartile: median. First/lower quartile: median of the lower 50% of data. Third/upper quartile: median of the upper 75% of data.
- Interquartile range: mid range of scale where 50% of sample is located, i.e. 25% of data lower and higher than median.

Two Facebook sets:
1st set: 113 149 178 196 269 382 388 467 546 682 724
2nd set: 11 149 178 196 269 382 388 467 546 682 5439

$\rightarrow$ interquartile range gives a more reliable description of dispersion than range.

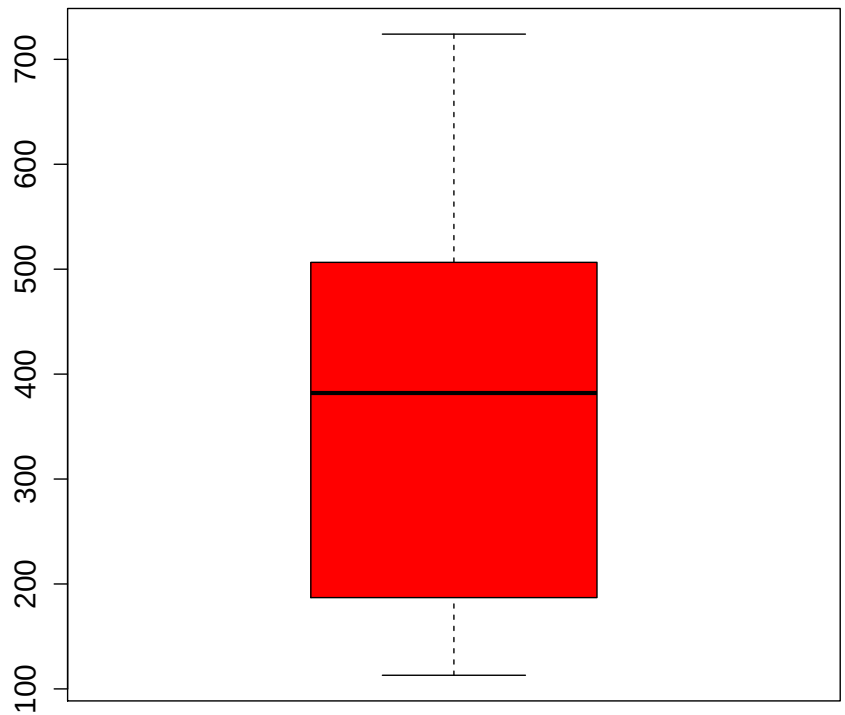
Boxplot

Box-and-whisker plot.

Box: interquartile range range. Whiskers: minimum and maximum values.

1st set of Facebook friends:

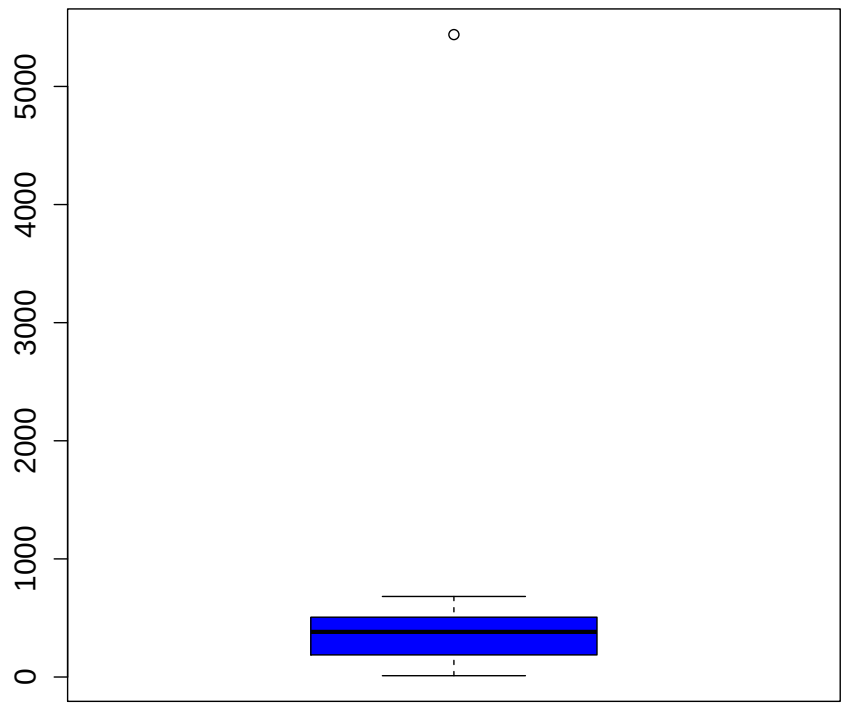
113 149 178 196 269 382 388 467 546 682 724



Boxplot

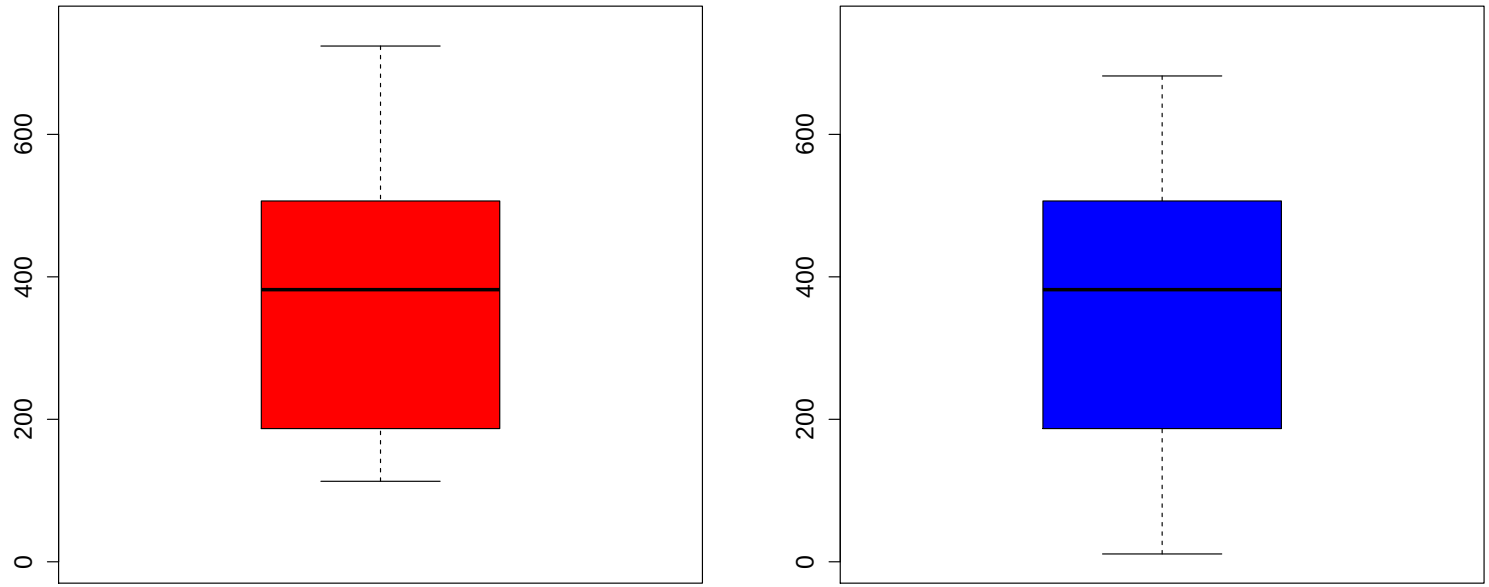
2nd set of Facebook friends:

11 149 178 196 269 382 388 467 546 682 5439



Point: extreme value or outlier. Quartiles for boxplots are in fact calculated by a probability model.

How are the whiskers calculated?



The lowest value within 1.5 interquartile range of the lower quartile, the highest value within 1.5 interquartile range of the upper quartile.

Interquartile range: `546-178 = 368, *1.5 = 552`

Values smaller than 178-552 and larger than 546+552 are ignored when calculating the whisker.

Useful books on statistics

For the special needs of linguists:

Baayen, R. Harald (2008): *Analyzing linguistic data: a practical introduction to statistics using R.* Cambridge: University Press. `http://www.ualberta.ca/~baayen/publications/baayenCUPstats.pdf`

Johnson, Keith (2011): *Quantitative methods in linguistics.* Blackwell: Oxford. `http://203.128.31.71/articles/_QXlGPndikK5L.pdf`

Gries, Stefan Th. (2009): *Quantitative methods in linguistics.* International Encyclopedia of the Social and Behavioral Sciences Amsterdam: Elsevier.

Useful books on statistics

A large variety of statistics books for students of psychology, sociology, economics or medicin are available: funny ones, detailed ones, talkative ones etc. Find the one you like most in your languages.

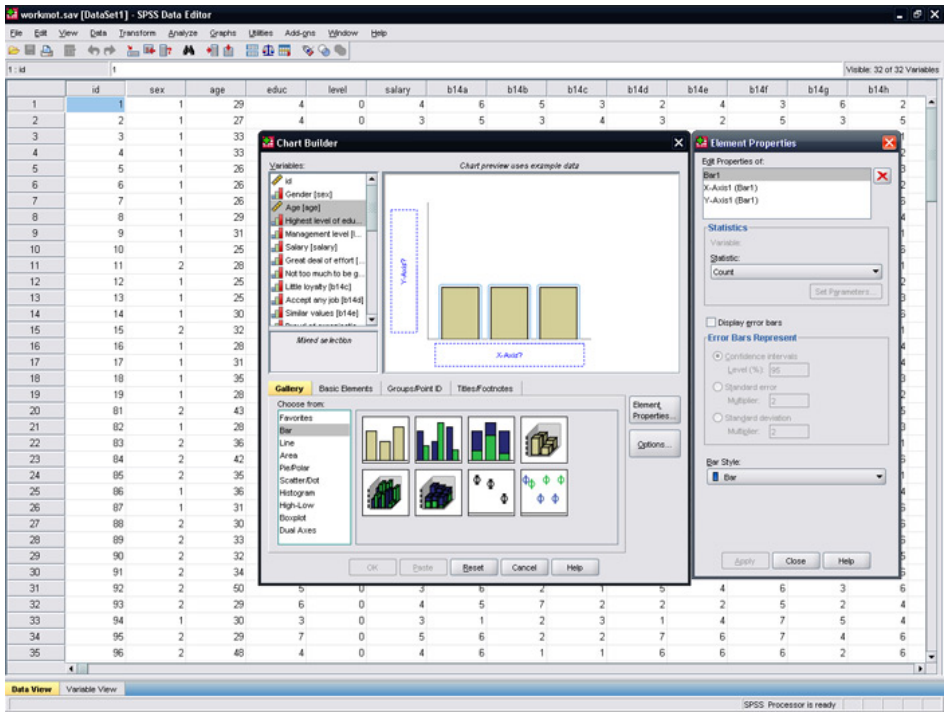
Two recommendations:

Dalgaard, Peter (2008): *Introductory statistics with R, 2nd edition.* New York: Springer.

Field, Andy, Miles, Jeremy, & Field, Zoë (2012): *Discovering statistics using R.* London: SAGE.

Softwares for statistics with graphical user interface

- SPSS: most widespread. Single license extremely expensive ($\sim$ 4000 euros), but many universities have cheap licenses for students.
- Others: Statistica, S+, SAS, Stata, etc.



Open source statistics softwares

With graphical user interface:
PSPP: freeware and platform-independent, but limited functionality. `http://www.gnu.org/software/pspp/get.html`

Without GUI:
R: an extremely powerful software for all kinds of statistical analyis with lots of linguistics-relevant packages.
Disadvantage: it requires the usage of a scripting language rather than clicking on menu points and icons.
But: most statistics books for linguists use R because it saves you a lot of money plus allows for a great flexibility. These books teach you how to use it. `http://www.r-project.org/`

The images for this course were prepared in R. You can find the scripts for them on the course website, along with the tests we discuss here.