Methodological skills

rMA linguistics, week 1

Tamás Biró
ACLC
University of Amsterdam
t.s.biro@uva.nl

Methodological + Skills

"Methodological"

- lectures: methodology, statistics (incl. probability theory), etc.
- student presentations of selected articles.

"Skills"

- minor research projects: continuously discussed + final paper.
- SPSS-lab.

Methodological + Skills

Requirements for credit:

- Eventual weekly assignments.
- Presentation of an article (10%).
- Research report on a small study (30%).

  Write article, in which you use SPSS.
- A 3-hour-long exam (60%)

  on which you have to score a minimum of 5.5.

Material of the course:

- On Blackboard.
- `http://www.birot.hu/courses/2012-methodology/`.

  Username: “meth”. Password: “skills”.

And now:

`http://www.let.rug.nl/~birot/courses/2008-stat/files/histogram.php`

The history of linguistics in a nutshell

Period	aims to understand	language as a ... phenomenon	language belongs to
"Philological" linguistics	text	literary	a book/author
Historical linguistics	history	historical	a nation
Structuralist linguistics	societies and sign systems	social and semiotic	a speaker community
Generative linguistics	brain or mind	biological neurological	an individual or a species

Sociolinguistics. Psycho- and neuro-linguistics.

Nativist vs. emergentist vs. functionalist approaches.

Combination of historical, social and biological aspects.

Methodologies in linguistics

- Look up texts: literary text; historical text; etc.

  Modern times: look up corpora.

- Observations. Field work. Controlled experiments.

- Self-reflection:

  prohibited by behaviorism, dominates Chomskyan linguistics.

Methodologies in linguistics

- The importance of cross-linguistic typologies: as broad and *representative* sample as possible.
- The importance of in-depth fieldwork: the data are never so simple.
- Theory building.

  Computer experiments: extreme control, oversimplification, cheap.

Methodologies in linguistics

Questions:

- Interested in *langue* or *parole*?

  in *competence* or *performance*?

  in actual, measurable facts or something more abstract?

- Hence, theory driven data collection?

  Raw data vs. interpreted data.

- Data $\rightarrow$ theory. Theory $\rightarrow$ data.

Methodologies in science in general

"The research loop/cycle":

Theory $\rightarrow$ research question

$\rightarrow$ data collection $\rightarrow$ data visualization

$\rightarrow$ data interpretation: interpretation of the raw data

$\rightarrow$ data interpretation: inferences beyond the raw data

$\rightarrow$ feedback to theory $\rightarrow$ new research question.

In this course

- Formulating a research question.

  Planning data collection.

- Data visualization: e.g., using Excell and SPSS.

- Interpretation of raw data: *descriptive statistics*.

- Inference from raw data: *inferential statistics*.

Research questions in science

- *How many? How long? When? Where? Who? What? Which colour? How frequent? What level? ...* — survey research.
- *Why?* — depends on theory, on approach, on paradigm.
- Hypothesis testing: falsifiable claims

  "boys do better than girls", "SVO languages have more X than SOV languages", "treated patients perform better than untreated".

Intro to Probability Theory

Intro to Probability Theory

- What is the chance to get a 6 when throwing a die?
- What is the chance of getting more than 3 with a die?
- What is the chance to get an even number with a die?

Intro to Probability Theory

- What is the chance to get a 6 when throwing a die?

  1 out of 6: if I repeat the experiment many times, I get a 6 in approximately $1/6$ of the cases (supposing a fair die).

  $\rightarrow$ the traditional interpretation of *probability*, a.k.a. the lay formulation of the *law of large numbers*.

- What is the chance of getting more than 3 with a die?

  $1/2$, because I will get a 4 in roughly $1/6$ of the time, a 5 in another roughly $1/6$ of the time, and a 6 in yet another $1/6$ of the time. No overlap, so I can sum up these cases.

Intro to Probability Theory

- *Random variable* X: the outcome of some "experiment".

  head/tail; a number; a part-of-speech; an error type;...
- *Probability Distribution Function*:

  $Pr(X = n)$: probability of getting n.

  For a fair die: a *uniform distribution*.
- *Cumulative Distribution Function*:

  $Pr(X \leq n)$: the probability of getting n or less.

$$Pr(X \leq n) = \ldots + Pr(X = n-2) + Pr(X = n-1) + Pr(X = n)$$
$$= \sum_{i=-\infty}^{n} Pr(X = i)$$

For a fair die: a step function (linear, if you look at integers only).

Intro to Probability Theory

- Two dice = two random variables, X_r and X_b.
- Many possible events, for instance:
  - $X_r = 3$: "the red die gives 3".
  - $X_r = 5 \cap X_b = 2$: "red die gives 5 and the blue one 2"
  - $X_r > 4 \cup X_b > 2$: "red gives 5 or 6, or blue gives > 1".
  - X_b is prime.

What is $Pr(X_r = 6 \cap X_b = 6)$?

What is $Pr(X_r = 6 \cup X_b = 6)$?

Probability axioms and further basic facts

1 The probability of an event E is a non-negative real number. $0 \leq P(E) \leq 1$.

2 The sum of the probability of the elementary events in the entire *sample space* is 1. $\sum_{E_i \in \Omega} P(E_i) = 1$.

3 E_1 and E_2 are *independent events*, if the probability of E_1 happening does not influence the probability of E_2 happening, and vice versa.

If E_1 and E_2 are *independent events*, then the probability that E_1 *and* E_2 happens is: $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$.

Probability axioms and further basic facts

3 E_1 and E_2 are *independent events*, if the probability of E_1 happening does not influence the probability of E_2 happening, and vice versa.

If E_1 and E_2 are *independent events*, then the probability that E_1 *and* E_2 happens is: $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$.

4 E_1 and E_2 are *disjoint events*, if E_1 cannot happen when E_2 happens, and vice versa. $P(E_1 \cap E_2) = 0$.

If E_1 and E_2 are *disjoint events*, then the probability that E_1 *or* E_2 happens is: $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Otherwise, $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.

Statistics need probability theory

The key question of statistics

Population: composed of many (N) individuals.
N is too large to test everyone.

Sample: composed of much fewer (n) individuals.

Key question: can we use the information gained from the sample to say something about the population?

- How to choose the sample?
  What to measure on the sample?

- *Inference*: How reliable is the measurement on the sample regarding the entire population?

Example: birth age of parents

Population: the parents of the current UvA students.

Sample: parents of the students of "Methodological skills".

Eventual research questions:

- What is the average age of the UvA students' parents?
- Is the average age of the fathers higher than the average age of the mothers?
- Do students have "on average" a father older than their mother?

Statistics need probability theory

- Population of size N. Sample of size n. ($n \ll N$).
- Random variable X_i: the result of the experiment on member i of the sample. ($1 \leq i \leq n$.) The value of X_i is x_i.
- Sample is random: X_i and X_j are *independent random variables*, that is, the value of X_i does not influence the value of X_j (if $i \neq j$).
- Measure a *statistic* of the sample:

  For example, the *sample mean* of X_i:

Statistics need probability theory

- Measure a *statistic* of the sample, e.g., the *sample mean*:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Is the *sample mean* equal to the *population mean*?
"Most probably" not.

Is the *sample mean* a good predictor of the *population mean*?
"Most probably" yes.

If the *sample mean* is 3.14, then what are the "most probable" values of the *population mean*? 3.14 $\pm$ some margin.

Statistics need probability theory

Questions to an expert in mathematical statistics:

- My sample of size n has yielded value s for the statistic S.
  What is the most probable property P of the entire population?
- My sample of size n has yielded value s for the statistic S. In which interval is property P of the entire population most likely to be?
- etc.

To answer the question, the expert will calculate:

- The probability of drawing a sample of size n and yielding value s for statistic S,
  provided that the entire population has such property P.

Statistics need probability theory

This has been very abstract.

Do worry: this is only the introduction to statistics.

Do not worry: we shall repeat it, before going further.

Do not worry: you do not need to understand the details in order to use “statistics cookbooks”.

Do worry: you must understand the basic concepts in order to use statistics in a meaningful way.

If this is your first encounter with this topic, and you are completely lost, then read *Moore and McCabe*, chapters 4-5.

SPSS-demonstration

Entering data in SPSS.

Next week:

- Lecture: Descriptive statistics.

  Recommended for the beginners: Moore and McCabe, chapt. 1.

- Lecture: Sampling (types of sampling, representativeness, confidence intervals)

  Required reading: Judd et al., chapts. 6 and 9.

- Student presentations: distributing the articles.

  Student projects: discussing the ideas.

- SPSS-lab: Data visualization.

To prepare for next week:

- First, digest probability theory.

  If needed, read Moore and McCabe, chapts. 1, 4 and 5.

- Read Judd et al., chapts. 6 and 9.

- Prepare idea for student project.

To prepare for next week:

Student projects:

- Goal: write an "article" with the following structure:

  (1) background, (2) research question, (3) methodology, (4) results, (5) discussion.

- Choose a simple question, not necessarily linguistic one. No need for elaborate theory, no need for lengthy experiments.

- Individually or in pairs.

  If in pairs, then two related "articles" to be submitted.

See you next week!