Methodological skills

rMA linguistics, week 2

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

Statistics need probability theory...

... in order to answer many interesting questions.

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".

Population $\gg$ sample. Representative?

Eventual research questions:

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

Example: Birth age of parents

One row per case. One column per variable.

subject	X = birth father	Y = birth mother	$Z = X - Y$
Anne	1956	1954	2
Bart	1947	1951	-4
...			
Zoe	1951	1960	-9

Are fathers older than mothers? Calculate:

- mean (a.k.a. average) of X vs. mean of Y;
- mean of Z;
- how many positive vs. how many negative values of Z.

Example: Anonymous test

Population = sample (or almost).

Eventual research questions:

- What is the average knowledge of students?
- What is the max and min of their knowledge?
- What is the typical range/spread of their knowledge?

NB: Quantitative vs. qualitative information.

Example: Anonymous test

One row per case. One column per variable.

subject	X = t-test	Y = ANOVA	Z = histogram	...
Anne	4	3	4	...
Bart	2	1	3	...
...				
Zoe	4	1	3	...

- Mean of each case?
- Mean, minimum, maximum and spread of each variable?
- etc.

The general case

One row per **case** (a.k.a. *individual*).

One column per **variable**.

A cell = **value** of the variable.

subject	X	Y	Z	...
Anne	x_1	y_1	z_1	...
Bart	x_2	y_2	z_2	...
...				
Irene	x_i	y_i	z_i	...
...				
Zoe	x_n	y_n	z_n	...

The general case

subject	X	Y	Z	...
Anne	x_1	y_1	z_1	...
...				
Irene	x_i	y_i	z_i	...
...				
Zoe	x_n	y_n	z_n	...

- *Variable* X is **independent** of variable Y if knowing the value of X does not help guessing the value of Y.

- *Case* i is **independent** of case j if knowing the values of i does not help guessing the values of case j.

- Measured variables vs. calculated variables.

Statistics need probability theory

- Population of size N. Sample of size n. ($n \ll N$).
- Interested in variables X, Y, etc. *in the population*: their "average", their spread, their inter-dependence, etc.
- Measuring variables X, Y, etc. *in the sample*: $x_1, x_2, ..., x_i, ..., x_n; y_1, y_2, ..., y_i, ..., y_n; ...$.
- Is it a **random sample**? That is, are the cases independent?
- Measure a *statistic* of the sample:

Statistics need probability theory

- Measure a **statistic** of the sample,

  For example, the $sample\ mean$ of X:

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

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 random sample of size n has yielded value s for the statistic S. What is the most probable property P of the entire population?
- My random 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 random sample of size n and yielding value s for statistic S,
  provided that the entire population has such property P.

Roadmap for the rest of the course

- How to draw a sample as random as possible?
  — next weeks.

- The properties/measures of the sample/population we may be interested in — next slides.

- The answers that experts in mathematical statistics give us
  — later weeks.

- Tools, software: the "expert" installed on your computer
  — SPSS lab.

Properties: Descriptive statistics

Properties: Descriptive statistics

- Data types: qualitative or quantitative?
- Visualization: overall pattern and **outliers**:

  errors or not? remove from data set?
- Shape, centre and spread.
- Relationships: correlation, etc. (not today).

Data types

- Discrete variables vs. continuous variables.

- *Categorical* scales:
  - *Nominal*: categories (binary or n-ary).
  - *Ordinal*: ordered categories.
    For example: *Likert-scales* (odd or even).

- *Quantitative* scales:
  - *Interval*: only difference is meaningful.
  - *Ratio*: difference and ratio are both meaningful.
  - *Logarithmic*: successive intervals multiply in size.

Shape of the distribution

- **Mode**: major peak.

  Unimodal, bimodal etc. distributions.

- Symmetric vs. skewed.

- Famous distribution: Gaussian (a.k.a. Normal, bell-shaped).

Distribution of the data

- *Minimum:* lowest value. *Maximum*: highest value.
- *Median:* half of the cases above, half below.
- *1st quartile:* quarter of the cases below.
- *3rd quartile:* quarter of the cases above.
- *nth percentile:* $n\%$ of the cases below.

Measures of centre

- *Mode:* most frequent element.
- *Median:* half of the cases above, half below.
- *Mean:* arithmetic average:

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

$\overline{x}$ or m = sample mean; μ = population mean.

Measures of spread (1)
The **five-number summary**

- *Five-number summary*: Min, Q1, Med, Q3, Max.
- *Boxplot* or *box-n-whisker*.
- "Suspected outliers": if it fails more than $1.5 \times IQR$ above the third quartile or below the first quartile.

Measures of spread (2)
Median and **ranges**

- (none for non-numeric data)

- $Range$ = maximum - minimum.

- $Inter\text{-}quartile\ range$: IQR = Q3 - Q1.

  $Semi\text{-}interquartile\ range$: (Q3 - Q1)/2.

Measures of spread (3): mean and **standard deviation**

- *Deviation*: distance from mean: $x_i - \overline{x}$.
- *Variance*: average of the squared deviations

$$\sigma^2 = \frac{(x_1 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

NB: divide by n or $n-1$?

- *Standard deviation*: root square of variance

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

σ^2 for population, s^2 for sample.

- Compare to semi-interquartile range ((Q3 - Q1)/2).

- (*Skew*: measures the symmetry of the distribution.

  Kurtosis: measures the flatness/peakedness of the distribution.)

Next week:

- Sampling.
- Research design.

To prepare for next week:

- (Read Judd et al, chapts. 6 and 9, if you haven't.)
- Read *Neuman*, chapter 9 on research design.
- Student presentations.
- (Ongoing: student projects.)

SPSS lab

- http://www.birot.hu/courses/2012-methodology/lab1.html
- PCH 005 (mediatheek)

See you next week!