Methodological skills

rMA linguistics, week 3

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

Topics today

Parameter of the population. Statistic of the sample.

- Re: descriptive statistics
- Data collection: research design and sampling methods.
- Formulating a research question
- Student projects. Distributing the articles.
- SPSS-lab.

Descriptive statistics (sorry, again)

Properties: Descriptive statistics

Data compression:

Parameter of the population. **Statistic** of the sample.

- Accidental vs. main characteristics.
- Visualization: overall pattern.

  Outliers: errors or not? remove from data set or don't?
- Main information: shape, centre and spread.

  Relationships: correlation, etc. (not today).

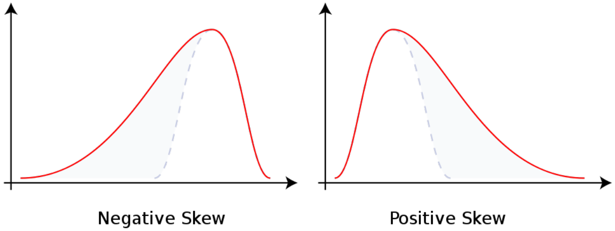
Data types: qualitative or quantitative?

- Discrete variables vs. continuous variables.
- *Categorical* scales:
  - *Nominal*: categories (binary or n-ary).
  - *Ordinal*: ordered categories.
- *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.



http://en.wikipedia.org/wiki/File:Skewness_Statistics.svg

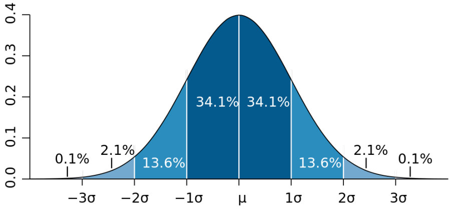
- Symmetric vs. skewed.

  Positive skew: skewed to the right = tail to the right.

  Negative skew: skewed to the left = tail to the left.

Gaussian (Normal) distribution

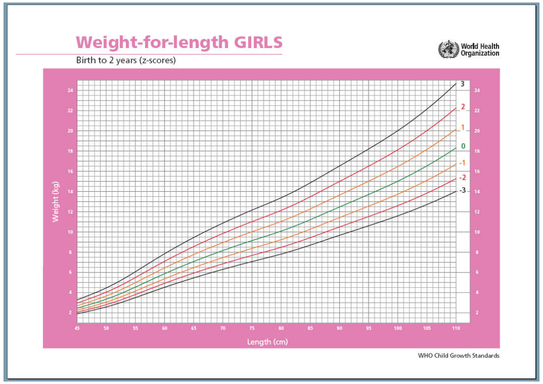
A famous distribution:



Gaussian (a.k.a. Normal, bell-shaped).

http://en.wikipedia.org/wiki/File:Sta

http://bcs.whfreeman.com/ips6e/content/cat_050/ips6e_table-a.pdf



http://sportsandfitness1.com/height-weight-chart-growth-of-8-to-12-months

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-quartile range*: IQR = Q3 - Q1.

  Semi-interquartile range: (Q3 - Q1)/2.

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

- $\textit{Deviation}$: distance from mean: $x_i - \overline{x}$.

- $\textit{Variance}$: 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.)

Descriptive statistics: summary

Data compression: distribution reduced to a few numbers

- Basic info: number of cases, minimum, maximum, sum.
- Position: mean, median, mode.
- Spread: range, (semi-)IQR, standard deviation, variance.
- Shape: kurtosis, skewness.

Data collection: research design

Beginning a research

- **Anecdotal evidence**: haphazardly selected individual cases.
- **Available data**: data produced in the past for some other reason.
- **Pilots**
- Sample survey (observational study), vs.

  Experiment (includes intervention): controlled, but artificial.

Designing data collection, define:

- **Population** and the **parameter(s)** you are interested in.
- **Units**: individuals/subjects/cases on which experiment/survey is done.
- **Explanatory variable(s)** (factors): what are their **levels**?
- **Response/dependent variable**: what are its **levels**?
- What do you want to know about these variables?

Factors

“Comparison between” or “effect of” different factors:

- Sex, age group, etc.
- Different treatments, or no treatment at all (‘control group’): Comparative experiment: treatment → response.
- Controlling factors: randomization or matching.
- **Lurking variables**: not included, but influencing responses.
- **Biased** design: systematically favours certain outcomes.

Principles of Experimental Design

(Moore & McCabe, ed. 6, pp. 183-4)

- **Compare** two or more treatments. This will control the effects of lurking variables on the response.
- **Randomize**: use impersonal chance to assign experimental units to treatments.

  Use software for randomization, or *random digit table*:

  `http://bcs.whfreeman.com/ips6e/content/cat_050/ips6e_table-b.pdf`
- **Repeat** treatment on many units to reduce chance variation in the results. To find a *statistically significant* effect.

- Likert-scale
- Double-blind
- Lack of realism: results cannot be generalized.
  E.g., campus students, WEIRD people.
- Matched pairs design: compare two treatments.
- Block design:

  block = group of units that are known before the experiment to be similar in some way (e.g., men vs. women). Random assignment within a block, then compare the blocks.

Data collection: sampling methods

Sampling methods

- Census (e.g., elections) vs. sample survey.
- **Voluntary response sample**: response rate, response bias.
- Probability sampling vs. non-probability sampling.
- Ideal for simple statistic techniques: SRS = **simple random sample**.
- **Stratified random sample**. Multistage sampling.

Sampling methods (cont'd)

- **Probability sampling methods**:

  SRS, stratified sampling, quota sampling, etc.

- **Non-probability sampling methods**:

  Convenience (haphazard, accidental) sampling; judgmental sampling; deviant case; snowball sampling; etc.

Sampling methods (cont'd)

- **Representativeness**? Representative for some controllable factors, but we can't know whether representative for dependent factors.

- **Undercoverage**: some groups in the population left out of the process of sampling.

- **Nonresponse**: individual chosen for the sample can't be contacted or does not cooperate.

Ethics

- Ethical committee
- Informed consent
- Confidentiality (not necessarily anonymity)
- No physical danger.

  Embarrassing or anxiety-inducing situations?
- Deception: misleading participants? only temporarily!

Formulating a research question

Formulating a research question

- What interests me? What motivates my investigation? What do I conjecture from informal observations, anecdotal evidence? What does theory predicts?

- That's the goal. First, need to operationalize the research question:

- What is the population, variables?

- Exactly what do I want to know about them?

  → data collection: experiment or systematic observation.

To prepare for next week:

Formulate your research question:

- Email to me in subject-line (by Tuesday, February 28)
- One-minute presentations, with one slide.
- Email to me presentation (ppt or pdf; by Wednesday, February 29, noon)

Next week:

- Sampling distribution.

  Data collection: reliability and validity.

  Read: Bachman 2004, chapters 4-5 (on Blackboard).

- Presenting research questions.

- Presenting articles.

SPSS lab

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

See you next week!