Methodological skills

rMA linguistics, week 6

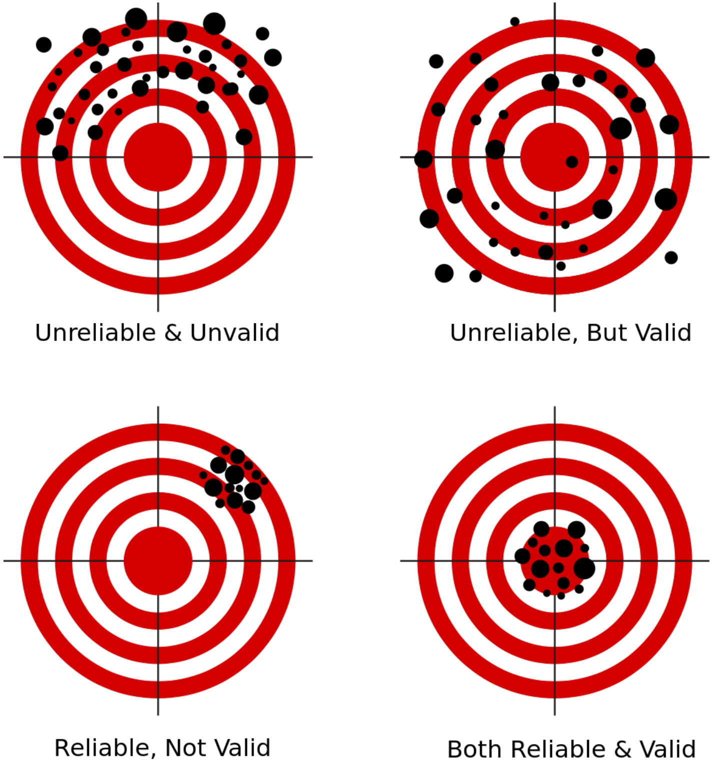
Tamás Biró
ACLC
University of Amsterdam
`t.s.biro@uva.nl`

Throwing a die

- **Population:** the outcomes of all throws in NL in 2012.
- Distribution of the population: approx. uniform distribution with *parameters* $\min = 1$, $\max = 6$, $\mu = 3.5$, etc.
- **Sample**: *observations* $x_1, ..., x_n$, $\rightarrow$ *statistics* $\overline{x}$, s_{n-1}, etc.
- **Sampling distribution** of the mean for samples of size n.

  The lesson of the Excel experiment last week:
  Mean of this sampling distribution $\approx \mu$.
  Std. dev. of sampling distribution decreases, as n increases.

Reliability and validity



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

Reliability and validity

- **Reliability** of the procedure:

  Procedure is *reliable* if sampling distribution has small spread — given our procedure.

  Repeating the experiment will yield similar results.

- **Validity** of the procedure:

  Unbiased statistic: if mean of sampling distribution is targeted parameter — given our procedure.

  The experiment answers our question.

Sampling distribution

- To reduce bias, achieve validity:

  use random sampling!

- To reduce variability, achieve reliability :

  use larger sample! Central Limit Theorem!

- Population size N (if much larger than sample size n) does not matter.

Sampling distribution of the mean:

The Central Limit Theorem

NB: Sampling distribution of other statistics discussed later.

Central Limit Theorem

- Given population with any distribution.

  Population mean is μ. Population standard deviation is σ.

- Draw a *simple random sample* (SRS) of size n.

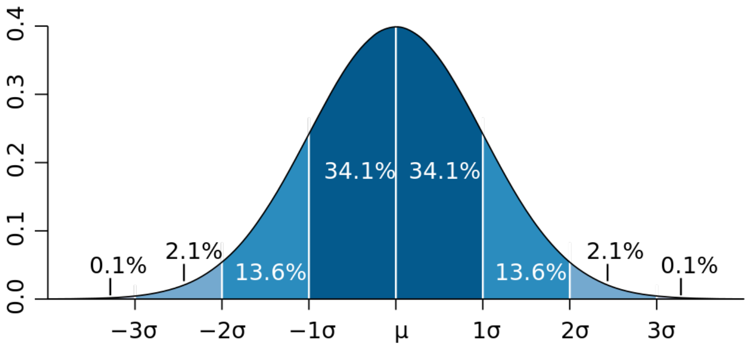
  Calculate sample mean $\bar{x}$.

- **Central Limit Theorem**:

  sampling distribution of $\bar{x}$ is (approx.) Normal: $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Normal (Gaussian) distribution

$$N(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



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

- $e = 2.7182....$ Mean: μ. Standard deviation: σ.
- Area under curve is 1.

Central Limit Theorem

- Even if we do not know the distribution the entire population,

  we know the behaviour of the means $\bar{x}$ of large samples:

- Sampling distribution of the mean is distributed around the mean μ of the population, and

- follows a Normal distribution of mean μ and st. dev. $\frac{\sigma}{\sqrt{n}}$.

  The larger the sample size n, the narrower the distribution.

- Hence, infer μ from $\overline{x}$, for *large* and *random* samples.

Central Limit Theorem

- **Central Limit Theorem** (version 1):

  sampling distribution of $\bar{x}$ is Normal: $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

- This theorem is only approximately true if original population is not Normal, but n is large. (Not true if n is small.)

- **Central Limit Theorem** (version 2):

  The sum (and, hence, the mean) of *independent* random variables X_1, X_2,...,X_n approaches a Normal distribution, as n grows large.

- Therefore: many statistical procedures require:
  - Independence of the cases in the sample.
  - Normality of the population, or
  - close to Normal distribution and larger sample size, or
  - very large sample size (if Normality does not hold).

Additionally:

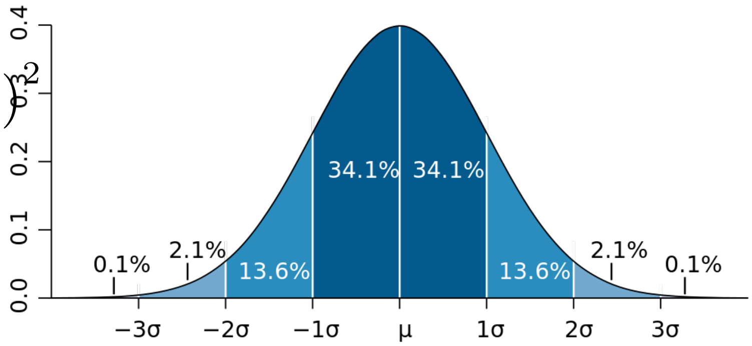
"Normality of the population" can be replaced by
"Normality of the sample".

Testing Normality of the sample: Normal quantile plots!

Standard Normal (Gaussian) distribution

Normal (Gaussian) distribution

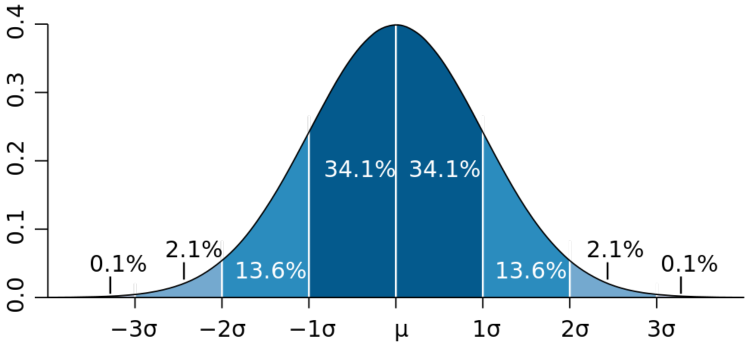
$$N(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



- $e = 2.7182\ldots$. Mean: μ. Standard deviation: σ.
- Area under curve is 1.
- 68–95–99.7 rule: area within $1/2/3$ σ from μ.

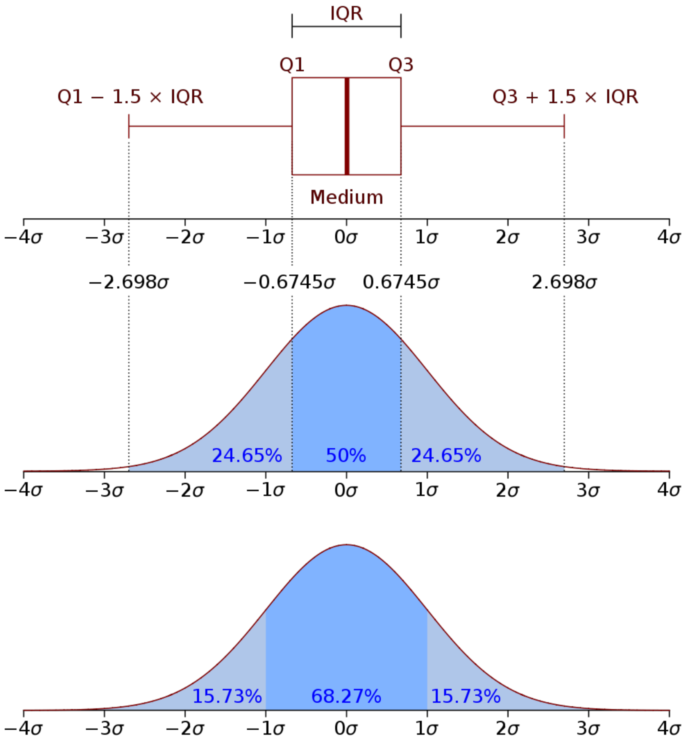
Standard Normal distribution

$$N(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



- $e = 2.7182...$. Mean: $\mu = 0$. Standard deviation: $\sigma = 1$.
- Area under curve is 1.
- 68–95–99.7 rule: area within 1/2/3 from 0.

Standard Normal distribution



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

Standard Normal distribution

A **Standard Normal Table**: *cumulative proportions*

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

- Normal distribution is a **continuous distribution:**

  Probability $P(a < X \leq b)$ of the random variable X having a value between a and b
  is equal to the area under the *probability density* curve between a and b.

- Value for b in the Standard Normal table: $P(-\infty < X \leq b)$, the area between $-\infty$ and b.

Standard Normal distribution

A **Standard Normal Table**: *cumulative proportions*

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

- Probability $P(a < X \leq b)$ of the random variable X having a value between a and b is the difference of the value for b and the value for a: $P(-\infty < X \leq b) - P(-\infty < X \leq a)$.

- Symmetry of the Standard Normal Distribution: $P(-\infty < X \leq b) = P(-b \leq X < +\infty)$.

- $P(|X| \geq |a|) = 2 \cdot P(-\infty < x \leq -|a|)$.

Normal calculations, inverse Normal calculations

- Calculate area *right* to $z = 1.47$.
- Find area from $z = -1.82$ to $z = 0.93$.
- What is z if left to it you find area 0.300?
- Similar questions with any other Normal distribution: normalize it ($x \to z$) first.

Normal calculations,
inverse Normal calculations

And now, you:

- For what z is 95% of area between $-z$ and z?
- For what z is 5% of area right of z?

Standardizing observations

Standardizing observations

μ: population mean of variable X.

σ: population standard deviation of variable X.

Cases	X	Y	...	Z $= X$ *standardized*
case 1	x_1			$z_1 = \frac{x_1 - \mu}{\sigma/\sqrt{n}}$
case 2	x_2			$z_2 = \frac{x_2 - \mu}{\sigma/\sqrt{n}}$
...				
case i	x_i			$z_i = \frac{x_i - \mu}{\sigma/\sqrt{n}}$
...				
case n	x_n			$z_n = \frac{x_n - \mu}{\sigma/\sqrt{n}}$
sample mean	$\overline{x}$			$\overline{z} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$
sample std. dev.	s			

Standardizing observations

- μ: population mean of variable X.
  σ: population standard deviation of variable X.

- Transform each data point: $z_i = \frac{x_i - \mu}{\sigma/\sqrt{n}}$.

- Averaging over the entire sample: $z := \overline{z} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$.

- z-statistic: a new statistic that we measure on the sample.

- Sampling distribution of $\overline{x}$ is $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
  Thus, the sampling distribution of the z-statistic is $N(0,1)$.

Toward the inference for the mean

Suppose $\mu = 3.5$ and $\sigma = 1.5$.
You draw a random sample of size $n = 9$, and calculate $\overline{x}$.

- What is the probability that $\overline{x} > 3.5$?
  The same as the probability of $z = \frac{\overline{x}-3.5}{1.5/\sqrt{9}} > 0$.

- What is the probability that $\overline{x} < 2.5$?
  The same as the probability of $z < -2$.

- What is the probability that $3 < \overline{x} < 4$?
  The same as the probability that $-1 < z < +1$.

Inference for the mean: z-test and p-scores

Suppose you know that $\sigma = 1.5$. You have drawn a Simple Random Sample (SRS) of size $n = 9$. You have got $\overline{x} = 4$.

Your null-hypothesis H_0 is that $\mu = 3.5$. Supposing H_0 is true,

- (... what is the probability of drawing a SRS with $\overline{x} = 4$?)
- ... what is the probability of drawing a SRS with an $\overline{x}$ at least as extreme 4: $\overline{x} \geq 4 = \mu + 0.5$? or $\overline{x} \leq \mu - 0.5$?
- ... what is the probability of drawing a SRS with an $\overline{x}$ at least as extreme 4: $\overline{x} \geq 4 = \mu + 0.5$? or $\overline{x} \leq 3 = \mu - 0.5$?

Inference for the mean: confidence interval

Suppose you know that $\sigma = 1.5$. You have drawn a Simple Random Sample (SRS) of size $n = 9$. You have got $\overline{x} = 4$.

- What is the best guess you can give for μ?
- Find an interval such that
  if μ falls within that interval,
  then the probability of drawing a SRS
  with $\overline{x}$ not more extreme than 4
  is less than $p < 0.05$.

Normal quantile plots

Do data follow Normal distribution?

- Arrange observed data values from smallest to largest. Record what percentile a value occupies.

- Normal score: z value of a percentile in the Standard Normal distribution. The value that the corresponding percentile should have, if the distribution were really Normal.

- Plot data against corresponding Normal score.

If data follow Normal distribution, then plotted points lie close to a straight line.

Student projects:

structure, variables

Student projects

- Intro: General problem
  $\rightarrow$ *anecdotal evidence* and *available data*.

- Precise research question:
  Hypothesis to be tested/rejected (H_0 and H_a).

- How to proceed?
  Sample survey (observation) or experiment (intervention)?

- Pilot vs. "the real stuff".

Student projects:

Define research question, in terms of what is your:

- Motivation? General problem? Operationalized research question?
- Population?

  Parameter(s) of the population that interests you?
- Units?

  Sample and sampling method?
- Explanatory variables, response/dependent variables?

  Levels of the variables?

Types of the explanatory variables

$\times$ type of the dependent variable

Scale of the explanatory variable(s) is	categorical (nominal, ordinal)	quantitative (interval, ratio, logarithmic)
Dependent variable with categorical scale	*crosstabs*	*logistic regression*
Dependent variable with quantitative scale	*t-test, ANOVA*	*correlation, regression*

Article presentation

Bakeman and Gottman (1997). Chapter 4: 'Assessing observer agreement' (presented by Simone).

Next week

- z-test vs. t-test.
- Power.
- Relationships.
- χ^2-test.
- SPSS-lab.

Read for next week:

Significance testing and power:

Jacob Cohen (1992), A power primer.

Jacob Cohen (1994), The earth is round (p ¡ .05).

See you next week!