Methodological skills

rMA linguistics, week 8

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

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*

Student projects:

Motivation, background: anecdotal evidence, past data.

Precise research question, operationalized.

Units, variables, population, sample.

Sampling distribution of the mean:

The Central Limit Theorem

NB: Sampling distribution of other statistics discussed later.

Central Limit Theorem

Four steps three weeks ago (note colours: red, black, green):

1. An ugly mathematical function with two parameters (μ and σ):

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

2. *Normal distribution:* a distribution that follows such an ugly function.
3. A mathematician will tell you that

   Mean of such a distribution (μ) $=$ first parameter of the function (μ).

   Std. dev. of such a distribution (σ) $=$ 2nd param. of the function (σ).
4. Central Limit Theorem (next slide): $\mu = \mu$ and $\sigma = \sigma/\sqrt{n}$.

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 $\bar{x}$. Sampling distribution of the mean: repeat sampling + averaging many times.

- **Central Limit Theorem**:

  Sampling distribution of $\bar{x}$ (approximately) follows a Normal distribution: $N\left(x|\mu = \mu = \mu, \sigma = \sigma = \frac{\sigma}{\sqrt{n}}\right)$.

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 ('converges' to) a Normal distribution, as n grows larger.

- Therefore: many statistical procedures require:
  - Independence of the cases in the sample.

    and

  - Normality of the population, or
  - close to Normal distribution and larger sample size, or
  - very large sample size (if Normality does not hold).

  Additionally:

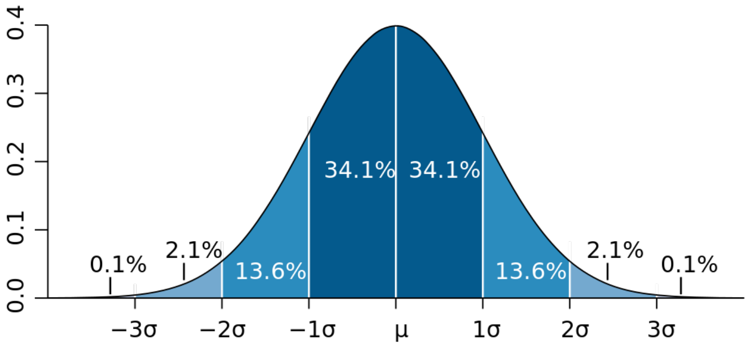
"Normality of the population" can be replaced by
"Normality of the sample".

  Testing Normality of the sample: Normal quantile plots!

Standard Normal (Gaussian) distribution

Normal (Gaussian) distribution

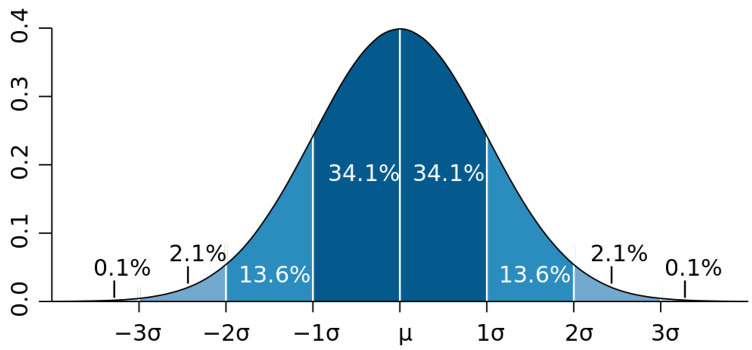
$$N(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



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

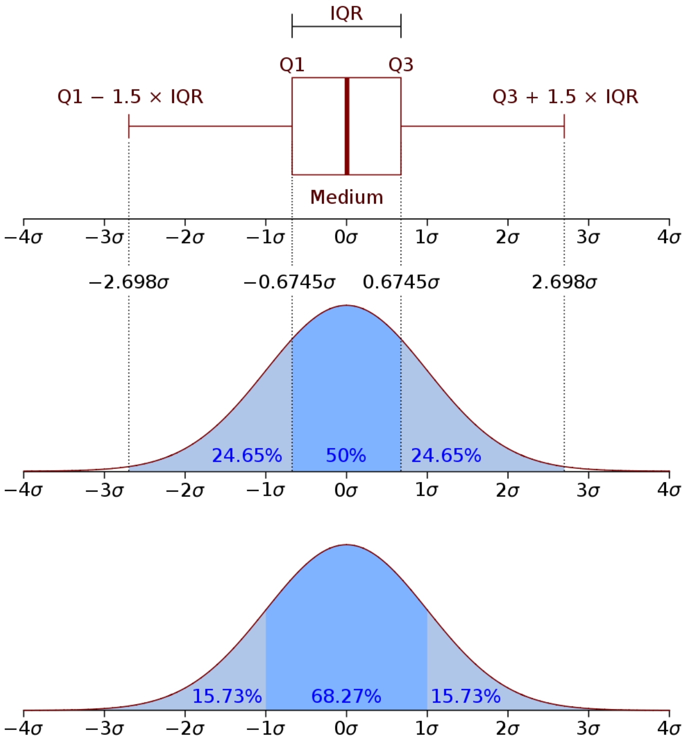
Standard Normal distribution

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



- $e = 2.7182\ldots$. 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 ($x \rightarrow z$) 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?

Transforming variables:

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 that $z = \frac{\overline{x}-3.5}{1.5/\sqrt{9}} > \frac{3.5-3.5}{1.5/\sqrt{9}} = 0$.

- What is the probability that $\overline{x} < 2.5$?
  The same as the probability that $z < \frac{2.5-3.5}{1.5/\sqrt{9}} = -2$.

- What is the probability that $3 < \overline{x} < 4$? The same as the probability that $\frac{3-3.5}{1.5/\sqrt{9}} = -1 < z < \frac{4-3.5}{1.5/\sqrt{9}} = +1$.

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 as 4 (i.e., $\overline{x} \geq 4 = \mu + 0.5$)?
- ... what is the probability of drawing a SRS with an $\overline{x}$ at least as extreme as 4 (i.e., $\overline{x} \geq 4 = \mu + 0.5$ 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 $p < 0.05$.

Cookbook z-test and p-test with one sample

Basic question: $\mu =?$

What is the population mean?

- Draw SRS (simple random sample) of size n.
- Calculate sample mean $\overline{x}$.
- Best guess for population mean: $\mu \approx \overline{x}$.

Confidence interval: $\mu = ?$

- Draw SRS of size n, with sample mean $\overline{x}$.
- Best guess for population mean: $\mu = \overline{x} \pm \text{error margin} =$
$= [\overline{x} - \text{error margin}, ..., \overline{x} + \text{error margin}]$
- If $\overline{x} - \mathrm{SE}_m \leq \mu \leq \overline{x} + \mathrm{SE}_m$,
then it is not very improbable to draw a SRS such as ours.
- Confidence level $C\%$: if we repeat the procedure many times, then in $C\%$ of the cases, the population mean will fall within the confidence interval.

Statistical tests: $\mu =?$

Null hypothesis H_0 vs. alternative hypothesis H_a.

- Draw SRS of size n. Calculate statistic s.
- If H_0 is true, how improbable to draw a SRS such as ours?
- p-value:
  given the sampling distribution of s, and
  provided that H_0 is true,
  what is the probability of drawing a SRS
  with an s at least as extreme as the s of our sample?

Population σ known: z-test

Null hypothesis H_0: population mean $\mu = m$.

Alternative hypothesis H_a: population mean $\mu > m$.

One-sided z-test:

- Calculate z-statistic: $z = \frac{\overline{x}-m}{\sigma/\sqrt{n}}$.
- p-value: probability that the sample's z-statistic $\geq$ our z.

Population σ known: z-test

Null hypothesis H_0: population mean $\mu = m$.

Alternative hypothesis H_a: population mean $\mu < m$.

One-sided z-test:

- Calculate z-statistic: $z = \frac{\overline{x}-m}{\sigma/\sqrt{n}}$.
- p-value: probability that the sample's z-statistic $\leq$ our z.

Population σ known: z-test

Null hypothesis H_0: population mean $\mu = m$.

Alternative hypothesis H_a: population mean $\mu \neq m$.

Two-sided z-test:

- Calculate z-statistic: $|z| = |\frac{\overline{x}-m}{\sigma/\sqrt{n}}|$.
- p-value: probability that the sample's $|z|$-statistic $\geq$ our $|z|$.

Population σ known: z-procedure

What is the population mean μ?

- Draw SRS of size n, with sample mean $\overline{x}$.
- Standard error of the mean: $\mathrm{SE}_m = \frac{\sigma}{\sqrt{n}}$.
- z^*: critical value for *confidence level* $C\%$.
- Best guess for the population mean: $\mu = \overline{x} \pm z^* \cdot \mathrm{SE}_m$.
- If we repeat sampling many times, in $C\%$ of the cases $\overline{x} - z^* \cdot \mathrm{SE}_m \leq \mu \leq \overline{x} + z^* \cdot \mathrm{SE}_m$.

Population σ unknown: Student's t-procedures

Estimate population std.dev. σ with sample std.dev. s $(n-1)$.

	z-procedures	Student's t-procedures
	one-sided z-tests two-sided z-tests conf. interval with z	one-sided t-tests two-sided t-tests conf. interval with t
	population σ	sample s (with $n-1$)
Statistic	$z = \frac{x-\mu}{\sigma/\sqrt{n}}$	$t = \frac{x-\mu}{s/\sqrt{n}}$
Sampling distribution	Normal distribution	Student's t distribution with df $= n-1$

Population σ unknown: t-test

Null hypothesis H_0: population mean $\mu = m$.

Alternative hypothesis H_a: population mean $\mu > m$.

One-sided t-test:

- Calculate t-statistic: $t = \frac{\overline{x}-m}{s/\sqrt{n}}$.
- p-value: probability that the sample's t-statistic $\geq$ our t.

Population σ unknown: t-test

Null hypothesis H_0: population mean $\mu = m$.

Alternative hypothesis H_a: population mean $\mu < m$.

One-sided t-test:

- Calculate t-statistic: $t = \frac{\overline{x}-m}{s/\sqrt{n}}$.
- p-value: probability that the sample's t-statistic $\leq$ our t.

Population σ unknown: t-test

Null hypothesis H_0: population mean $\mu = m$.

Alternative hypothesis H_a: population mean $\mu \neq m$.

Two-sided t-test:

- Calculate t-statistic: $|t| = |\frac{\overline{x}-m}{s/\sqrt{n}}|$.
- p-value: probability that the sample's $|t|$-statistic $\geq$ our $|t|$.

Population σ unknown: t-procedure

What is the population mean μ?

- Draw SRS of size n, with sample mean $\overline{x}$.
- Standard error of the mean: $\mathrm{SE}_m = \frac{s}{\sqrt{n}}$.
- t^*: critical value for *confidence level* $C\%$, with df $= n - 1$.
- Best guess for the population mean: $\mu = \overline{x} \pm t^* \cdot \mathrm{SE}_m$.
- If we repeat sampling many times, in $C\%$ of the cases $\overline{x} - t^* \cdot \mathrm{SE}_m \leq \mu \leq \overline{x} + t^* \cdot \mathrm{SE}_m$.

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.

Basics of inference

(Cf. Cohen's two articles.)

H_0, H_a and H_1

- H_0 (null-hypothesis): effect size ES $= 0$ (most often).

  (Cohen, 'The Earth Is Round ($p < .05$)': "nil hypothesis")

- H_a (alternative hypothesis): there is an effect, ES $\neq 0$.

  Cohen: the "nil hypothesis" is (practically) always true!

- Cf. Cohen, 'A Power Primer':
  H_1: there is a well-defined small/medium/large ES.

Goal: reject H_0 to argue for H_a.

The (in)famous p-value

- H_0 (null-hypothesis): effect size ES $= 0$ (most often).
  Cohen, 'The Earth Is Round ($p < .05$)': "nil hypothesis"

  $\rightarrow$ which (usually) correspond to statistic $= 0$.

  Sampling distribution of the test statistic:
  if H_0 is true, then test statistic is most often close to 0.

- H_a (alternative hypothesis): there is an effect, ES $\neq 0$.
  H_1: the effect is ES (where ES $\neq 0$).

  $\rightarrow$ which (usually) correspond to a statistic $\neq 0$.

The (in)famous p-value

- H_0 (null-hypothesis): effect size ES $= 0$ (most often).
  Cohen, 'The Earth Is Round ($p < .05$)': "nil hypothesis"

  $\rightarrow$ which (usually) correspond to statistic $= 0$.

  Sampling distribution of the test statistic:
  if H_0 is true, then test statistic is most often close to 0.

$p =$ the probability of $($ obtaining a test statistic at least as extreme as the one we have just obtained based on our observations $|$ provided that H_0 is true $)$.

The (in)famous p-value

$p =$ the probability of $($ obtaining a test statistic at least as extreme as the one we have just obtained based on our observations $|$ provided that H_0 is true $)$.

- Low p-value $\rightarrow$ either H_0 is false, or we have bad luck.
- We reject H_0 with **confidence level** α if $p < \alpha$

  — the level of "bad luck" that we hope never to have.
- If statistic from data $>$ critical value corresponding to α, then $p < \alpha$.

The (in)famous p-value

$p =$ the probability of $($ obtaining a test statistic at least as extreme as the one we have just obtained based on our observations $|$ provided that H_0 is true $)$.

- High p-value $\rightarrow H_0$ is either true, or false (e.g., small effect size), or we have bad luck.

- We say we do not have sufficient evidence to reject H_0.

- BIG ERROR: to conclude that H_0 is true!

Example: z-test

- (Suppose we know std. dev. of population is σ.)
- H_0: the population mean is m.
- Sample of size n. Data x_1, x_2,..., x_n.

  Calculate sample mean $\overline{x}$, then z-statistic: $z := \frac{\overline{x}-m}{\sigma/\sqrt{n}}$.
- $P(z = ...|H_0)$: what is the chance of getting such a value for z, supposing H_0 is true?
- Hence, is it probable that H_0 is true?

Example: z-test

- (Known σ.) H_0: the population mean is m.
  Sample of size n. Calculate z-statistic: $z := \frac{\overline{x}-m}{\sigma/\sqrt{n}}$.

- From the Central Limit Theorem we know that
  if H_0 is true, then probability of $|z| > 1.96$ is less then 5%.

  So, critical value for $C = 95\%$ confidence level: $z^* = 1.96$.
  If $z > z^* = 1.96$, then reject H_0 with confidence level $\alpha = 0.05$ (two-tailed).

- Higher n or higher $\frac{\overline{x}-m}{\sigma}$ ('effect size') $\rightarrow$ higher z $\rightarrow$ higher chance to reject H_0, given a simple random sample (SRS).

Some of the problems with inference

(Cf. Cohen's two articles.)

H_0, H_a and H_1

- H_0 (null-hypothesis): effect size ES $= 0$ (most often).

  (Cohen, 'The Earth Is Round ($p < .05$)': "nil hypothesis")

- H_a (alternative hypothesis): there is an effect, ES $\neq 0$.

  Cohen: the "nil hypothesis" is (practically) always true!

- Cf. Cohen, 'A Power Primer':
  H_1: there is a well-defined small/medium/large ES.

Goal: reject H_0 to argue for H_a.

H_0, H_a and H_1

(Cohen, 'The Earth Is Round ($p < .05$)':

A correct, non-probabilistic Aristotelian *modus tollens*:

- If H_0 is correct, then data D cannot occur.
- D has, however, occurred.
- Therefore, H_0 is false.

H_0, H_a and H_1

(Cohen, 'The Earth Is Round ($p < .05$)':

An incorrect probabilistic "*modus tollens*":

- If H_0 is correct, then data D would probably not occur.
- D has, however, occurred.
- Therefore, H_0 is probably false.

Conditional probability

- $P(A|B)$: probability of A, provided that we know that B is true. $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- Researcher interested in $P(H_0|D)$:
  the probability that H_0 is true, given observation D.
- Statistics can only provide $P(D|H_0)$:
  probability of obs. data (and more extreme data), given H_0.

Bayes' theorem:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B) \cdot P(B)}{P(A)}$$

H_0, H_a and H_1

(Cohen, 'The Earth Is Round ($p < .05$)':

Result	normal	schizophrenic	Total
Negative test	949	1	950
Positive test	30	20	50
Total	979	21	1000

Test is "good": most normal people tested as negative, and most schizo people tested as positive. Still, a positive test does not prove schizophrenia ($p = 0.60$), because very low $P(H_0)$.

Type I error and Type II error

Statistical procedure set at conf. level C	H_0 is true in reality	H_1 / H_a is true in reality
Effect-size is	$= 0$	$\neq 0$
rejects H_0	Type I error	
does not reject H_0		Type II error

$\alpha = 1 - C = P(\text{Type I error}|H_0)$; $\beta = P(\text{Type II error}|H_a)$

What interests us: **power** of the statistical test $= 1 - \beta$:
the probability of rejecting H_0 if H_0 is false.

Cohen: power depends on Effect-size, n and C (or α).

SPSS lab

`http://www.birot.hu/courses/2012-methodology/lab2.html`

Next week

- Crosstabs and the χ^2-test.
- Two student presentations.

See you next week!