Inferential Statistics

Prerequisites:

- academic level of analytic thinking
- descriptive statistics,
- interest in humanities
  - communication;
  - **or** language, incl. afasiology **or**
  - humanities computing

Non-prerequisites:

- mathematical sophistication
- facility with statistical software (SPSS) — taught here!

RuG

Inferential Statistics

Course Emphases

- practical use of statistics, incl. software, graphics
- reasoning behind statistics
- typical application areas in linguistics, communications and humanities computing

Formal Requirements

- Weekly lecture (**attendance required**)
- Five exercises with SPSS (labs)
- One laboratory test (Exercise to be graded).
- One exam

Grades

- Lectures (10%)
  Attendance at least five times.
- SPSS Labs (20%)
  Complete/Incomplete
- Exam (70%)

Why Statistics in Humanities?

Lots of humanities issues are EMPIRICAL and VARIABLE

empirical — involving matters of fact, not purely conceptual

variable — issues that may be decided in different ways for different individual cases

Examples of empirical, variable issues:

- sex is related to verbal fluency
- web sites with banners get more attention
- grammatical structure influences language processing

Statistical analysis needed for EMPIRICAL, VARIABLE issues.

RuG

Hypothesis Testing

We begin with a research question, which we try to formulate as a **hypothesis**

- sex is related to verbal fluency
- web sites with banners get more attention
- grammatical structure influences language processing

Normally, we need to translate this to a concrete form before statistics are useful

- men and women score differently on tests of verbal fluency
- web sites with banners are revisited more often
- object relative clauses (i.e., those in which relative pronouns are grammatical objects) take longer to read than subject relative clauses

Abstraction

Given a research question, translated into a concretely testable hypothesis

- web sites with banners are revisited more often

= "**all** web sites with banners are revisited more often than web sites without banners"?

—probably not. The data is variable, there are other factors:

- amount of information (library system)
- value of information (Centraal Bureau voor Statistiek)
- changeability of data (weather, flight arrivals)

We normally need statistics to abstract away from the variability of the observations.

- web sites with banners are revisited more often **on average**

Subject Matter

- web sites with banners are revisited more often **on average**

We **must** study this on the basis of a limited number of web sites — a SAMPLE. But we're interested in the larger class of all web sites — the POPULATION.

The hypothesis concerns the population, which is studied through a representative sample.

- **men and women** differ in verbal fluency (study based on 30 men and 30 women)
- **web sites** with banners are revisited to more often (studied on the basis of 30 web sites)
- **object relative clauses** take longer to read than **subject relative clauses** (studied on the basis of 30 people's reading of 20 relative clauses of each type).

Analysis

Given a research question, translated into a concretely testable hypothesis, expressed abstractly

- web sites with banners are revisited more often **on average**

You measure rates of revisiting for a randomly selected group of sites, with and without banners.

Will any difference in averages (in the right direction) be proof?

—probably not. Very small differences might be due to **chance**.

We normally need statistics to analyse results.

- STATISTICALLY SIGNIFICANT results are those unlikely to be due to chance.

Hypothesis Testing

A **statistical hypothesis** concerns a population about which a hypothesis is made involving some statistic

- population (all web sites)
- parameter (statistic) (rate of revisiting)
- hypothesis (ave. rate of revisiting higher when banners used)

- **always** about populations, not just about samples
- sampling statistic identified
  - mean
  - frequencies
  - ...

RuG

Identifying Hypotheses

ALTERNATIVE HYPOTHESIS (= original hypothesis) is contrasted with NULL HYPOTHESIS — hypothesis that nothing out of the ordinary is involved.

- H_a: (ave. rate of revisiting is higher when banners used)

contrasts with NULL HYPOTHESIS:

- H_0 (null hypothesis): (banners make no difference in ave. rate of revisiting)

Logically, H_0 should imply $\neg H_a$

Quantifying Significance

- STATISTICALLY SIGNIFICANT results are those unlikely to be due to chance.

We quantify significance by estimating how likely it is that results could be due to chance.

Concretely: if the null hypothesis were true, how likely would the sample statistic be?

Example: If in fact banners make **no** difference in how often web sites are revisited, how likely is it that a sample of 20 web sites with and without banners would show that 18% of the visitors return to the former and only 13% to the latter?

p-VALUE is the chance of sample given H_0

A low p-value is evidence against H_0, and for H_a

“Significant at the 0.05 level”

We normally determine in advance which significance level is required for (probabilistic) proof.

For example, we may agree that any result with a p-value less that 0.05 is sufficient proof against the H_0 (and therefore for the H_a) that we will be convinced.

The p-value that is determined to be sufficient for the rejection of H_0 is referred to as the α-LEVEL

We may then report the results of the experiment as “significant at the $p \leq 0.05$-level” or “significant at the 0.05-level”.

RuG

Other Significant Levels

Sometimes, α is determined to be 0.01, sometimes 0.001

α is threshold of REGION OF REJECTION — score needed to reject H_0 (and accept H_a)

—low values unlikely if H_0 is true, likely if H_a true

Size of region (α) inversely proportional to acceptable risk (of wrongly accepting H_a)

RuG

Other Significant Levels — Example

Example: You have aphasia test, with known μ (mean), σ (standard deviation) from US, & may wish to use it in the Netherlands

H_0: $\mu_{\text{US}} = \mu_{\text{NL}}$ (same population, therefore same μ)

H_a: $\mu_{\text{US}} \neq \mu_{\text{NL}}$ (different populations, maybe due to language dependencies)

region of rejection: 0.05

—you reject H_0 even though results would be consistent 5% of time

Region of rejection variable

- perhaps new test very expensive
- perhaps this aspect of diagnosis not essential

RuG

Interpreting Results

1. Take a sample (of n aphasic patients), administer test, determine μ_{NL}.
2. Determine z score for sample of n.

$$z = \frac{\mu_{\mathsf{NL}} - \mu_{\mathsf{US}}}{\sigma/\sqrt{n}}$$

3. Use tables to determine chance of z score, $P(z)$. This is the p-value, the chance of the sample if $\mu_{\mathsf{NL}} = \mu_{\mathsf{US}} (= H_0)$
4. If sample statistic is in rejection region, e.g., $p < 0.01$, reject H_0 in favor of H_a (statistically significant)
5. If sample statistic *not* in rejection region, then either accept H_0 or suspend judgement

Possible Errors

You could, of course, be wrong.
The selection of the sample could be unlucky (unrepresentative). Possibilities:

H_0	true	false
accepted	correct	type II error
rejected	type I error	correct

Type I Errors — focus of hypothesis testing
p-value – chance of a type I error
α-level: boundary of acceptable level of type I error

Formulating Results

H_0	true	false
accepted	correct	type II error
rejected	type I error	correct

Note that results with $p = 0.06$ aren't very different from $p = 0.05$, but we need to specify a boundary. 0.05 is low because the "burden of proof" is on the alternative.

In these cases we certainly don't feel that we've **proven** H_0, only that we've failed to show convincingly that it's wrong.
We speak of "retaining H_0" ("H_0 *handhaven*").

Type II Errors (null hypothesis accepted by false)
β —probability of type II error
$1 - \beta$ —"power of statistical test" (no further mention in this course)

Degrees of Freedom

Degrees of Freedom (dF) — the number of ways in which data could vary (and still yield same result).

Example: 5 data points, mean

If mean & 4 data points known, fifth is determined

Mean 6, data is $4, 5, 7, 8$ and one unknown

□ fifth $= 6$

There are **four** degrees of freedom in this set.

In general, with n numbers, $n - 1$ degrees of freedom (for the mean).

Most hypothesis-tests require that this be specified

Degrees of Freedom