Experimental Humanities II (HUMB002) 2016
STATISTICAL ANALYSIS

HYPOTHESES TESTING

Lecture 5

Pavla Linhartová

The lectures and exercises are based on the lectures from the subject PSY117 – Statistical analysis by Stanislav Ježek and Jan Širůček from Department of Psychology, Faculty of Social Studies MU Brno

Statistical hypotheses

- Examples:
  - H: $\mu = 100$ — Population mean is 100.
  - H: $\sigma = 10$ — Population SD is 10.
  - H: $\mu_1 - \mu_2 = 0$ — Mean in population 1 and mean in population 2 are the same. OR There is no (zero) difference between means in the two population (e.g. between patients and healthy people).
  - H: $\rho_{xy} = 0$ — Variable X and variable Y don’t correlate.
- Let’s take the first hypothesis and confront it with data:
  - In a sample of 1000 randomly sampled adults we measure IQ mean 105 with SD = 14.

Principles of statistical hypothesis testing

- Hypothesis testing is based on probability
  - If we know probability distribution of a statistics, we can infer, how probable is some sampling statistics with regard to hypothesis: ***P (D |H)***
  - ***Example:***
    - Data: m = 105
    - Hypothesis: μ =100
    - *P* (*D* |*H*) is *P* (*m*=105 | μ =100)
    - If this probability is relatively high, the hypothesis is supported by this.
    - If this probability is relatively low, the hypothesis unlikely.
  - **…** What probability is needed to support / reject the hypothesis?

Principles of statistical hypothesis testing

- Fisher, Popper: Falsification principle – the hypothesis can't be confirmed, only rejected
- But we want to confirm our hypotheses, not reject them…
- Principle of hypothesis testing is that we formulate opposite hypothesis (null hypothesis) to our research hypothesis
- If we can reject the null hypothesis, we take it as support for our research hypothesis
- We reject the null hypothesis if: $P(D \mid H_0) < \mathbf{0,05}$ (**0,01**; 0,001; 0,0001)

Results dichotomization

	H_0 kept $P(D \mid H_0) \geq \alpha$	**H_0 rejected** $P(D \mid H_0) \leq \alpha$
H_0 true (no effect)	OK	Type I error (false positive) α *(its probability)*
H_0 false (effect)	Type II error (false negative) β	OK Test power ($1-\beta$)

The lower is α, the higher is β. The exact form of the relationship depends on the test that was used. α and β can be both low only in samples with high N.

Statistical hypotheses

- H_0: null hypothesis (cz: nulová, testová hypotéza)
  - logical negation of alternative hypothesis
- H_1: alternative, scientific, research hypothesis (cz: alternativní, vědecká, výzkumná hypotéza)
  - the one we're interested in

$P(D \mid H_0)$ when we reject H_0:

  - is denoted as **p** or **Sig.**
  - probability of incorrect rejection of H_0 = type I error (cz: chyba prvního typu)
  - if we state it in advance: level of statistical significance (cz: úroveň/hladina statistické významnosti), α, often in % - 5%, 1% etc.
    - error rate we are willing to tolerate in our results

- One-tailed vs. two-tailed hypotheses (cz: jednostranné vs. oboustranné hypotézy)
  - one-tailed – directional: $\mu \geq 23$, $\mu \leq 0$, we usually avoid them
  - two-tailed: $\mu = 23$

Hypothesis testing process

1. **Formulate null hypothesis**, which you're going to try to reject
   (e.g. H_0: $\mu = 0$, nebo H_0: $\mu = 6$)
2. **Choose level of significance**, that is probability that type I error occurs
   (e.g. $\alpha = 0{,}05$)
3. We are looking for probability of obtaining our sampling statistics or more extreme value given that H_0 is actually true: ***P*(*D*|*H*₀), p, Sig.**
   - we go through probability distribution of the statistics (we have to know it)
   - e.g. m = 0.5, H1: $\mu \neq 0$, Ho: $\mu = 0$; then we are looking for: $P\,(|m| \geq 0{,}5 \,|\, \mu = 0)$
   - usually we need to tranform raw statistics to test statistics (e.g. *t* or *z*) for which we know the probability distribution

5. **We reject or keep the null hypothesis:**
   - if $P(D|H_0) < \alpha$, we reject H_0
   - if $P(D|H_0) \geq \alpha$, we don't reject H_0

Example: One-sample t-test

- We are testing a therapy for problematic behaviour.
  - Difference before and after therapy: m=2.7; s=3.5; N=10
  - H1: The therapy is effective ($\mu \neq 0$) – two-tailed hypothesis
1. H0: The therapy is not effective: $\mu = 0$
2. We take the usuall level of significance (in social sciences): $\alpha = 0{,}05$
3. P ($|m| \geq 2{,}7 | \mu=0$) = ?
   - we have to transform raw statistics to test statistic, in t-test the tests statistics is **t,** because we work with Student's t-distribution with df=N-1 (if we knew σ, we would work with normal disribution and use z-test instead of t-test)
     - we compute standard error for mean: $s_m = s / \sqrt{N} = 3.5 / \sqrt{10} = 1.1$
     - $t = (m - \mu) / s_m = 2.7/1.1 = 2.45$
     - t_{krit} = T.INV.2T(p;df) = T.INV.2T(0.05;9) = 2.26
     - P ($|t| \geq 2{,}45 | \tau = 0$) = T.DIST.2T(x;df) = T.DIST.2T(2.45;9) = 0,04
4. P ($|m| \geq 2{,}7 | \mu=0$) < 0,05, thus we reject the null hypothesis
5. With result m=2,7 it is very unlikely that the true difference is 0, and this is our support for statement that there actually is true difference

One-tailed tests

- We usually use them only if the opposite result than the one we're expecting would be nonsense, non-interpretable
- We usually consider one-tailed hypotheses, but we test their two-tailed forms.

Test of Pearson's correlation significance

- H0: $\rho = 0$, $r = 0.4$, $N = 100$
- We have to transform Pearson's r to Fisher Z with normal sampling distribution and $s_Z=1/\sqrt{(n-3)}$
  - =FISHER(0.4) = 0.42
- We compute the standard error: $s_Z=1/\sqrt{(100-3)}=0.1$
- We compute our test statistics: $Z/s_Z = 0.42/0.1 = 4.2$
- $P(D|H_0)=2*(1 - \text{NORM.S.DIST}(Z/s_Z;1) = 0.00003$
- $P(D|H_0) < 0.05$ (and 0.01, 0.001), thus we reject the null hypothesis
- The correlation is considered significant

Problems in statistical hypothesis testing

- Results dichotomization:
  - the same effect size give different results for H0
  - with very high sample sizes even even very small difference can result as significant (even difference with no practical significance)
  - on the other hand, we need sufficient sample size to reject the null hypothesis
- Interpretation problem:
  - p= $P(D \mid H_0)$ a nikoli $P(H \mid D)$
- Always indicate a measure of effect size (Cohen d, r, R^2, η^2, ω^2)
- Always use interval estimates
- Hypothesis testing should be rather supplementary information