Statistics for EMCL week 7

Tamás Biró

Humanities Computing

University of Groningen

t.s.biro@rug.nl

This week:

- Power of tests, types of errors.
- Non-parametric tests (M&M 15).
- Summary of tests.

General model of significance tests

- Population: null hypothesis on parameter.
  - Parameter in one population = value?
  - Parameter in two (or more) populations equal?
- Sample → statistic.
- Our level of confidence in a statistical procedure: “What would happen if we used the inference method many times?”

General model of significance tests

- Sampling distribution of statistic (may depend on shape of distribution: criteria on use of tests).

- p-value: the probability of drawing a sample whose statistic is *as extreme as, or more extreme than* the statistic calculated from our sample.

- If p-value lower than significance level α, reject null hypothesis (“it is unlikely that we have had such a bad luck”).

Tests, statistics, distributions

Test	**Statistic**	**Sampling distribution**
one-sample z	z	Normal (Gaussian)
one-sample t	t	$t(\mathrm{df} = n - 1)$
two-sample t	t	$\approx t(\mathrm{df} = \min{(n_1, n_2)} - 1)$
χ^2	χ^2	$\chi^2(\mathrm{df} = (c - 1)(r - 1))$
ANOVA	$\frac{\mathrm{MSG}}{\mathrm{MSE}}$	F(DFG,DFE)
equality of variance	...	F

and many-many more...

Power of test

- The **power** of the test to detect the alternative: probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true.

Types of error

- Type I error: reject H_0 while H_0 is true.
- Type II error: fail to reject H_0 while H_0 is false.
- p-value: probability of Type I error.
- Power: probability of no Type II error.

Non-parametric tests

When classical statistical tests fail:

- Far from Normal distribution & few data:
  - Remove outliers?
  - Non-linear transformation?
  - Tests based on other frequent distributions?
- Non-numeric data.

Bootstrap methods/permutation tests: M&M 16.

Ordinal scale

Ordered (ranked), but differences not meaningful:

- rank listing of job candidates;
- alphabetical order; sonority scale;
- months; years;
- some test scores, marks of satisfaction/agreement.

Likert scales

On a scale of 1-7, this course was
very useful (1) _____ completely useless (7).

- Use at least 5 points.

- Allow a “centre” (use 1 through odd numbers).

- Be consistent: “positive” answers always same side. School grade effect.

- Compare mean using t-test or ANOVA.

Wilcoxon Rank Sum Test

- Data: Group A: 2, 4, 7, 9; Group B: 3, 7, 8, 10, 11

- **Rank transformation:**

Data	2	3	4	7	7	8	9	10	11
Sample	A	B	A	A	B	B	A	B	B
Rank	1	2	3	4.5	4.5	6	7	8	9

- **Wilcoxon rank sum W:** sum of ranks for sample A.

- A new statistic again, with its sampling distribution. p-values for W: from table, software or Normal approximation.

Wilcoxon Rank Sum Test

- Comparing two distributions:

  Null hypothesis:

  two populations have same distribution.

  Alternative hypothesis: for all a,

  $P(X_1 > a) \geq P(X_2 > a)$; and $>$ for some a.

- Comparing medians: supposing that two distributions have same shape.

Non-parametric tests

Normal tests	Non-parametric fallbacks
One-sample t	Wilcoxon signed rank test (15.2) Sign test (M&M 7.1)
Two-sample t	Wilcoxon rank sum W (15.1) Mann-Whitney U
ANOVA	Kruskal-Wallis test (15.3)

Based on sum of ranks, except sign test.

Sign test

When all else fails...

- Divide data into classes $+$, $-$ and 0.
- Compare proportion of positive vs. negative.
- H_0: no weighting toward $+$ or $-$.
- H_a: bias towards one of the signs.
- Refer to binomial distributions (M&M 5.1).

General remarks

Research article/thesis

- Explain background theory. Earlier studies. Explain novelty of your approach, contrast, but fair to others.

- Assumptions. Predictions of theory.

- Describe experiment: sample size, drop outs, control group (group assignment randomly).

- Statistics: population vs. sample, choice of test (requirements met?). Note significance level.

- Data available on ftp server. Visualization, tables.

- Interpret results. Discuss “failed hypotheses”. Alternative explanations, conclusion and future work.

Summary

- Real world is less orderly than statistics textbooks imply (M&M p. 220).
- **Garbage in, garbage out:** statistics can't help if experiment poorly designed, data poorly collected, argumentation flawed.
- Plan statistical procedure *before* data collection: necessary sample size? avoid posthoc

manipulations.

- Visualize data, look at them carefully, before any statistical procedure.
- Before running statistical test, check if criteria of application apply.
- Effect size vs. sample size: large effect significant at small samples; small effects significant at large samples.

Decision tree for tests

- Distribution of non-numeric variables? χ^2.
- Correlation between two variables: correlation r.
- Compare mean of single numeric variable:
  - Population σ known: z-test.
  - Sample vs. value: one-sample t-test.
  - Two values per subject: paired t-test.
  - Two groups: two-sample t-test.
  - More groups: ANOVA.

For each case: non-parametric fallbacks.

Next week:

- Q&A on Wednesday
  Prepare questions!

- Website: JN's slides, Q&A, assignment feedback, paper-and-pen assignments...

- Me: prepare sample test, list of "what to learn", correct assignments.

- Exam: November 27.