Index of 13-L

Practical Information

Major news:

- last home assignment to be returned to you in Monday's lab 1-4pm (requests for review topics still welcome),
- course evaluations (official SOTS on Moodle, my survey: Monday/at website),
- you need to notify me about your choice on midterm (by Wednesday, December 8th), unless already done (thanks!).

Today's lecture:

- Exam: Monday 13/12, 9am-12pm, AVC Lecture Theatre A,
  - exam practical remarks,
  - exam questions (types, calculations),
- review slides: new or partly new slides, 13L−4/. . . /14,
- also a list of review slides from previous lectures (appendix, 13L−15),
- review of exam questions from final exam 2015:
  1. descriptive statistics/two-sample inference,
  2. proportion data/sample size,
  3. ANOVA/regression.

Exam Practical Remarks

Start time and number of questions depend on your midterm choice:

- use midterm: start at 10am, 2 questions — 35%,
- drop midterm: start at 9am, 3 questions — 50%.

Exam rules: same as for midterm (see instructions at webpage), in particular:

- open book (all aids are allowed), except a computer-like device, although cell phones and tablets can be used (in flight mode):
  - as a calculator (with basic calculation functions, no statistical software),
  - to access course notes and electronic textbook material.

Some hints and advices: (to use or not...)

- layout — essential requirements: readability, and a clear distinction between what is *in* your answer and what is not — don't write first a draft and then a final version,
- conclusions should be part of all analyses,
- statistical model should be part of all data analysis,
- explicit calculations may prevent loss of points due to typing errors (or the like),
- errors: if you realize an error and do not have time to correct it: write what is wrong, what should have been done and how the error would affect the result.

Exam Questions

2 or 3 major questions, with subquestions — possible types:

- choice of statistical model and analysis:
  - carry out analysis when calculations manageable (see below),
  - or base analysis on Minitab print + extra calculations,
  - or outline analysis if calculations not manageable,
- probability calculations manageable (see below),
- multiple choice (one or several correct answers).
- brief essay (10 lines) to explain statistical concept, to interpret result, or to explain why something is wrong.

Calculations by complexity — some examples (not all cases covered)

Possible calculations	Too complex calculations
simple probabilities (e.g., $1-p$, $(1-p)^n$, simple binomial)	statistical analysis of normal distrib. models without calculation help (given statistics or computer output)
standardization in normal distribution	probability plots, residual plots
t-test and CI (given suitable estimates)	prediction intervals in regression
backtransformation of means to original scale	rank-based nonparametric tests
proportions (with CI)	X^2-tests
sample size for precision	power calculations

Typical Formulas for Exam

List of formulas (non-exhaustive!) of potential relevance for exam:

1L: suspected outlier rule,

3L: probability rules (e.g.: addition, multiplication); means and variances for random variables,

4L: normal distrib.: calc. of probabilities + percentiles (incl. standardization, z-score); binomial distrib.: calc. of simple probabilities, mean, stand.dev.,

5L: standard error (for mean); z-test and CI for 1-sample normal,

6L: t-test and CI for 1-sample and 2-sample normal,

7L: CI and test for 1 and 2 proportion(s); sign test

8L: expected value (cell) for X^2-test,

9L: one-way ANOVA: equal variance guideline; LSD-value; ANOVA table; t-test and CI for mean/difference between means,

10L: t-test and CI for regression parameters; ANOVA table; prediction; t-test for correlation,

11L: two-way ANOVA: same as for 9L,

12L: sample size for precision (1-sample normal/ proportion, detection of disease).

Outline of Statistical Analysis

Data description:

- descriptive statistics: plots, tables, simple statistics — for the purposes:
  - overview of the data,
  - detect errors / "different" observations (outliers)
  - focus attention on what's relevant.

Statistical model: we formulate statistical models containing theoretical distributions and unknown parameters, in order to

- make clear the assumptions (and utilize them),
- let (statistical) parameters[1] represent issues of interest,

— "All models are approximations, but some are useful" (Box).

Estimation: our information about the parameters from the observed data is summarized in statistics or estimates:

- quantities calculated from the data to give possible parameter values,
- aim of statistical methodology: obtain estimates as close to true values as possible[2],
- all estimates should be accompanied by a measure of uncertainty, such as standard error or confidence interval.

[1] In our statistical setup, parameters are fixed, unknown numbers describing a population.
[2] In most cases, it is unlikely and implausible that an estimate would be exactly equal to the true value.

Model check:

- comparison of observed distribution and assumed theoretical distribution (using estimated parameters),
- methods: graphical (plots) or numerical (tests),
- if model unsatisfactory, start over with new model[3],

Hypothesis testing:

- formulate, using model parameters, null hypothesis H_0 (model simplification) and alternative hypothesis H_a,
- the test statistic and associated P-value summarize our evidence against null hypothesis, which we may reject (low P) or not reject (high P).

Final Model:

- simplest possible after model reduction(s),
- if necessary, re-estimate model parameters (with CI).

Conclusion / Presentation:

- summary of test results,
- illustrations of implications of the final model, e.g. prediction; also, presentation of estimates (with SE or CI).

[3] For less serious violations of model assumptions, it may alternatively be reasonable to proceed with caution.

Graphs in Statistical Analysis

Graphs have different purposes:

- Descriptive: show the shape of the distribution in a dataset,
  - dotplot and stemplot (raw data),
  - histogram (grouped raw data),
  - boxplot (schematic for descriptive statistics),
  - scatterplot (raw data, 2 variables),

  note: small datasets are best illustrated by raw data plots, and more data $\Rightarrow$ more schematic/grouped plots,
- Model check: graphical assessment of one or more model assumptions,
  - normal probability plots (check normal distribution within groups/samples),
  - residual plots (check different assumptions in normal distribution models),
- Presentation: graphical display of estimates (possibly with measure of uncertainty) from (complex) analysis,
  - group mean plots with error bars, based on model standard deviation ($\sqrt{\text{MSE}}$), (ANOVA models),
  - fitted line plots (regression).
  - interaction plots (2-way ANOVA).

About Statistical Models

What is a statistical model?

- a formal statement/description of the assumptions made for specific statistical inference, such as statistical hypothesis test and confidence (or prediction) interval,
- the assumptions quantify the random variability in the data relative to model parameters (= the constants representing the underlying population).

How to state/write a statistical model? (for the exam) — choose between:

- formal notation with equations and explicit parameters, e.g.

  i) 1-sample: $\boldsymbol{X}_i$'s are i.i.d. and $\sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$,

  ii) regression: $\boldsymbol{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \cdot \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i$, with $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\sigma})$,

- or a descriptive (but detailed) statement, e.g.

  i) sample from a normal distribution with unknown mean and standard deviation,

  ii) a linear relationship: $\boldsymbol{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \cdot \boldsymbol{x} +$ error,
  where the errors are normally distributed with mean 0 and the same standard deviation,

- or any mix (or variation) of these, where assumptions are explicitly stated or clearly understood.

Choice of Statistical Model

Some useful questions to ask about the data:

- purpose of study?
- what is the observational unit/experimental unit/subject?
- response/outcome[4] or explanatory/predictor variable?
- continuous or categorical (including binary) variable?
- which variables/groupings/classifications should enter into the model? — e.g.,
  - a single sample (normal, binomial),
  - two independent samples (normal, binomial, multinomial),
  - several independent samples (one-way ANOVA, two-way table of counts),
  - paired observations $\Rightarrow$ single sample for differences,
  - two-way classification (two-way table of counts),
- continuous variable (explanatory or response) to be used for prediction of another variable? (regression)
- two continuous response variables with no wish to predict one from the other? (correlation),
- transformation? (to achieve normal distribution, homogeneity of variance, linear relation).

[4] What defines a response variable is that it has (real) random variation.

Multi-Purpose Statistical Tests

Some statistical tests are specific to a single situation/use, e.g. tests for normality and rank-based tests. However, tests named after a probability distribution usually have multiple uses:[5]

Name	Use(s) in course	Instances/Versions
z-test	normal distrib. inference with known σ	1-sample z-test
		2-sample# z-test (barely covered)
	binomial distrib. inference	z-test for 1 and 2 proportions
t-test ("Student" t)	normal distrib. inference with unknown σ	1-sample and 2-sample# t-tests
		t-test for regression parameters
		t-test for correlation coefficient
		t-test for (contrasts and) pairwise comparisons
χ^2-test (chi-square)	inference for counts	2-way table tests for homogeneity and independence
	rank-based tests	Kruskal-Wallis test
F-test	effects in normal distrib. models	linear regression (slope)
		factorial effects (ANOVA)

2 independent samples

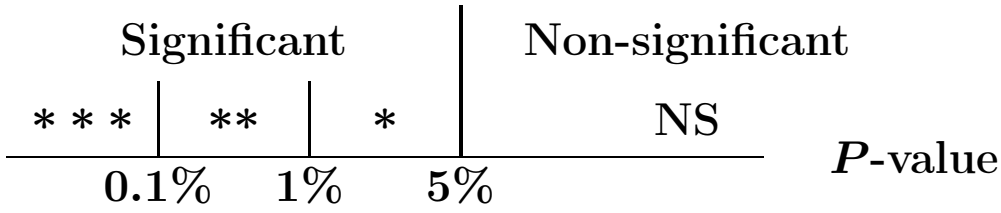
Relations between tests:

- t-test (df) with very large df $\approx$ z-test,
- z-test squared $\sim \chi^2(\text{df}=1)$-test,
- t-test (df) squared $\sim F(\text{df}_1=1, \text{df}_2=\text{df})$-test.

[5] The table shows examples from VHM 801 only, many more exist for other statistical models/procedures!

Review: *P*-values

Statistical significance, as determined from *P*-value[6]:



Interpretation and suggested wording (personal preferences, not banning significance):

- *P*-values above 0.05 are usually denoted non-significant and interpreted as indicating that the result (observed value of the test statistic) could have occurred by coincidence if H_0 is true, so that H_0 cannot be rejected,
- *P*-values just around 0.05 (no matter if just above or below) are close to significant, and give indication that H_0 may not be true,
- formally, a *P*-value below 0.05 is evidence against H_0,
- *P*-values at 0.01 or below are clearly significant and give clear indication that H_0 is false, so that H_0 should probably be rejected,
- *P*-values at 0.001 or below are strongly significant and show that the data are incompatible with H_0, so that H_0 should be rejected, unless other explanations exist.

[6] Warning ! — do not confuse *P*-values and values for *p*'s:
- *P*-values are values (certain probabilities) computed to interpret test results,
- *p*'s are (probability) parameters in particular models, e.g. $B(n, p)$.

Review: Another Look at t and t^*

From a (hypothetical) confused individual: what are all those different t's:
t^*, $t(\text{df})$, $t_{1-\alpha}(\text{df})$, t_{obs}, t, ...

For confidence intervals, e.g. with confidence level 95% $(1-\alpha)$ and error level 5% (α), we need in the 1-sample formula:

$$\mu: \ \bar{X} \pm t^* \, s/\sqrt{n}, \quad t^* = t_{1-\alpha/2}(\text{df}),$$

a number, $t^* = t_{1-\alpha/2}(\text{df})$, from a t distribution with df degrees of freedom:

- determines the middle 95% of the t distrib., between 2.5% and 97.5% ($\alpha/2$ and $1-\alpha/2$),
- equals the 97.5% percentile in the t distribution,
- can be found in a statistical table (under conf. level 95% or tail area probability 2.5%),
- Minitab: `Probability Distribution Plot` with Right Tail Probability = 0.025.

For a test of H_0: $\mu = \mu_0$ (where μ_0 is a known value), we compute the observed value of our t-statistic from the 1-sample formula:

$$t_{\text{obs}} = (\bar{X} - \mu_0) \ / \ (s/\sqrt{n}),$$

and the P-value is calculated from $\text{P}(t(\text{df}) > |t_{\text{obs}}|)$,[7]

- it is a tail area probability, and can be evaluated as $<$ or $>$ specific probabilities in a table, based on the table values below or above $|t_{\text{obs}}|$,
- Minitab: `Probability Distribution Plot` with Right Tail X Value = $|t_{\text{obs}}|$.

[7] Generally, $t(\text{df})$ refers to the t distribution with df degrees of freedom, sometimes just t when the df are understood.

Review: Degrees of Freedom

What are those "degrees of freedom" (DF or df) really?

- values to distinguish between different distributions of the same type (distribution types: t, χ^2, F),
- always positive integer numbers: 1,2,3,...,
- in normal models, a large DF for the estimate of σ corresponds to good precision (many observations),
- part of ANOVA tables (e.g., use DFE for t^*),
- specific formulas exist for standard situations:[8]

Model	Data	Purpose	DF
single sample	$X_1, \ldots, X_n$	conf. int. t-test	$n-1$
two samples (same σ)	$X_1, \ldots, X_{n_1}$ $Y_1, \ldots, Y_{n_2}$	conf. int. t-test	n_1+n_2-2
1-way ANOVA	X_{ij}, $i = 1, \ldots, I$ $j = 1, \ldots, n_i$	conf. int. t/F-test	error: $\sum_i n_i - I$ groups: $I-1$
2-way ANOVA	X_{ijk}, $i = 1, \ldots, I$ $j = 1, \ldots, J$ $k = 1, \ldots, n_{ij}$	conf. int. t/F-test	error: $\sum_{ij} n_{ij} - IJ$ groups: $I-1$ & $J-1$ interac: $(I-1)(J-1)$
simple linear regression	(Y_i, x_i) $i = 1, \ldots, n$	conf. int. t-test	$n-2$
2-way table	N_{ij}, $i = 1, \ldots, I$ $j = 1, \ldots, J$	chi-square test	$(I-1)(J-1)$

[8] (technical) All DFs can be interpreted as the difference in number of (free) parameters between two models: a comprehensive model and a reduced model.

Review: General Principles of Tests and CIs

- Data $X_1, \ldots, X_n$,
- Statistical model containing a parameter μ, typically a mean parameter.
- Estimate $\hat{\mu}$ for μ, based on $X_1, \ldots, X_n$.
- Standard error $\text{SE}(\hat{\mu})$, either
  - estimated from the data, or
  - known value (rarely realistic in practice),

  note: in normal models (with error standard deviation σ) we often have

  $$\text{Var}(\hat{\mu}) = A\,\sigma^2 \quad \text{and} \quad \text{SE}(\hat{\mu}) = \sqrt{A}\,\sigma,$$

  where A is a constant determined by the form of $\hat{\mu}$,
- Reference distribution of $\dfrac{\hat{\mu} - \mu}{\text{SE}(\hat{\mu})} \sim$ percentiles t_p; note: in normal models with estimated $\text{SE}(\hat{\mu})$ the reference distribution is usually a t-distribution,
- Confidence interval $(1-\alpha)$ for μ : $\hat{\mu} \pm t^* \,\text{SE}(\hat{\mu})$, $\quad (t^* = t_{1-\alpha/2})$
- Test of H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$,

  $$\text{test statistic:} \quad t = \frac{\hat{\mu} - \mu_0}{\text{SE}(\hat{\mu})}, \quad P\text{-value:} \quad P = 2 \times \text{P}(t \geq |t_{\text{obs}}|),$$

  where $t \sim$ the reference distribution.

Appendix: List of Potential Review Slides for Course

Selected slides on statistical analysis across all sessions (Summary notes may also be useful):

- completely randomized design and block design (2L−9/10),
- binomial setting (4L−10),
- one- or two-sided (5L−18),
- testing by confidence interval (5L−19),
- how to find P-values and percentiles (6L−7),
- inference for non-normal data (6L−9),
- inference for proportions — overview/single/two (7L−2/3/4/6),
- nonparametric (distribution-free) methods (7L−8/11),
- sign test (7L−9),
- two-way tables: models, estimation and hypotheses (8L−8/9),
- after the ANOVA table: LSD & Bonferroni (9L−13/14),
- use of residuals for model check (10L−11/12; 11L−19),
- correlation vs. regression (10L−15),
- overview 1-way & 2-way ANOVA (11L−20),
- statistical testing issues (12L−3),
- sample size based on estimation accuracy/power (12L−10/11/13).