# Index of 14-L

(new slides only)

# Practical Information

Major news:

- last home assignment returned Friday (lab 1-4pm),
- you need to notify me about your choice on midterm (by Monday),
- last lab review later today (1-2pm; usual room).

Topics today:

- Exam: Thursday 19/4, 9am-12pm, AVC 278N,
  - exam practical remarks,
  - exam questions (types, calculations),
- Evaluation:
  - I'll leave at 10:40 to let you do the official evaluation,
  - another evaluation form on Friday (mine!),
- Review topics — next slide,
- Sample problems — pick from these:
  - final 2011: 2 (ANOVA),
  - final 2012: 1 (1-sample), 2 (two-way tables),
  - final 2013: 1 (midterm type question),
  - final 2014: 1 (descriptive stats/two-sample inference), 2 (two-way table/proportion data), 3 (regression) — Q1+Q3 have already been posted for lab sessions.
- Your questions...

# Topics for Lecture

Review topics: pick from these!

- outline of statistical analysis (partly new slides: 14L−6/7),
- use of graphs (new slide: 14L−8),
- statistical models (new slides: 14L−9,10),
- overview 1-way & 2-way ANOVA (new slide: 14L−11),
- how to find $P$-values and percentiles (6L−13),
- use of percentiles (review slide: 11L−16),
- degrees of freedom (review slide: 11L−17),
- interpretation of $P$-values (review slide: 12L−16),
- general statistical tests (review slide: 13L−16),
- completely randomized design and block design (2L−9/10),
- binomial setting (4L−10),
- one- or two-sided (6L−6),
- testing by confidence interval (6L−7),
- inference for proportions — overview/single/two (7L−11,12/13/15),
- nonparametric (distribution-free) methods (8L−2),
- sign test (8L−3),
- sample size based on estimation accuracy/power (8L−14/16),
- errors of type I−II and power (8L−15),
- two-way tables: models, estimation and hypotheses (9L−8/9),
- after the ANOVA table: LSD & Bonferroni (10L−12/13),
- use of residuals for model check (12L−2/3; 13L−13).

# Exam Practical Remarks

Start time and no. questions depends on your midterm choice:

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

All aids (books and notes and calculators) are allowed, — except a computer or computer-like device (tablet or smartphone).

The 2 or 3 questions have almost equal weight (with points indicated) — use your time reasonably! (no extra time!)

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

- layout: essential requirements are
  * readability,
  * 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).

Calculations by hand (calculator):

- manageable without data entry into calculator,
- no large summations,
- examples of possible calculations:
  - simple probabilities (e.g., $1-p$, $(1-p)^n$, simple binomial),
  - standardization in normal distribution,
  - normal approximation for binomial,
  - $t$-test and CIs (given suitable estimates),
  - backtransformation of estimates to original scale,
- examples of too complex calculations:
  - statistical analysis of normal models without calculation help (given statistics or computer output),
  - probability plots, residual plots,
  - $X^2$-tests and rank-based nonparametric tests,
  - power calculations.

# Typical Formulas for Exam

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

1L: suspected outlier rule,

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

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

5L: standard error (for mean),

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

7L: $t$-test and CI for 2-sample normal; CI and test for 1 and 2 proportion(s),

8L: sign test; sample size for precision (1-sample normal/ proportion),

9L: expected value (cell) for $X^2$-test,

10L: equal variance guideline; LSD-value; ANOVA table; $t$-test and CI for mean/difference between means,

11L: $t$-test and CI for regression parameters; ANOVA table; prediction,

12L: $t$-test for correlation,

13L: same as for 10L (one-way ANOVA).

# Outline of Statistical Analysis

Data description:

- descriptive statistics: plots, tables, simple statistics,
- purpose:
  - 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 parameters (fixed, unknown numbers describing a population) 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 (though *never* equal to true values),

- all estimates should be accompanied by a measure of uncertainty, such as standard error or confidence interval.

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*[1],

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 confidence *against null hypothesis*, which we may *reject* (low $P$) or *not reject* (high $P$).

Final Model:

- simplest possible after model reduction,
- if necessary, re-estimate model parameters (+ 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).

[1] 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 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 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).

# About Statistical Models

What is a statistical model?

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

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

- formal notation with equations and explicit parameters, e.g.
  i) 1-sample: $X_i$'s are i.i.d. and $\sim \mathrm{N}(\mu, \sigma)$,
  ii) regression: $Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$, with $\varepsilon_i \sim \mathrm{N}(0, \sigma)$,
- descriptive statement, e.g.
  i) one sample from a normal distribution with unknown mean and standard deviation,
  ii) a linear relationship: $Y = \beta_0 + \beta_1 \cdot x +$ error, where the errors are normally distributed with mean 0 and the same standard deviation,
- 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[2] or explanatory/predictor variable?
- continuous or categorical (including binary) variable?
- which variables/groupings/classifications should enter into the model? — some examples:
  - 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).

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

# Overview 1-Way & 2-Way ANOVA

- data description:
  - statistics: group[3] means + standard deviations,
  - graphs: box-plot for groups[3], interaction plot (2-way),
- statistical model:

$$\begin{aligned} \text{1-way}: \quad X_{ij} &= \mu + \alpha_i + \varepsilon_{ij}, \\ \text{2-way (repl.)}: \quad X_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \\ \text{2-way (no repl.)}: \quad X_{ij} &= \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \end{aligned}$$

- model checks — residuals plots and/or
  - equal variance: i) max/min rule, ii) variance test,
  - normality: normal plots/tests,[4]
- statistical analysis:
  - estimation of pooled (or error) standard deviation,
  - hypothesis testing of overall effects → ANOVA table,
  - (2-way repl.): interaction significant/"substantial"?:
    - *yes*: interaction plot, 1-way ANOVA for groups[3],
    - *no*: row and column factors assessed separately,
  - pairwise comparisons between group means for significant effects (based on CI, LSD or $t$-tests): adjusted/unadjusted to simultaneous error level of 0.05,
- presentations: group means with SE or CI + sign. indic.

[3] In a 2-way design with replication, groups refer to the combined groups formed by row and column factors.

[4] With ample replication: observations within groups[3]; With limited/no replication: (standardized) residuals across all groups.