Statistics II
Summary

Çağrı Çöltekin

University of Groningen
Information Science

April 22, 2014

First things first: the exam

Exam date & time: **June 6, 14:00–17:00** room: *A. Jacobshal 01*.

- A mixture of multiple choice and short-answer questions.
- It should take about 90 minutes, but you can use all 3 hours reserved for the exam.
- An example exam is already on Nestor, under ‘course documents’.
- You do not need a calculator, but you are allowed to bring a simple calculator (without network capabilities).

The plan of the day

A summary (with a new/different perspective at times):

Practical stuff
Basics: hypothesis testing, statistical models.
Correlation
Regression
Multiple regression
ANOVA
Factorial ANOVA
Repeated-measures ANOVA
Logistic Regression

...some common problems & your questions.

Unconditional inference: confidence intervals

- The simplest case of ‘inference’ is an unconditional estimate, e.g., the population mean estimated from a sample.
- Reliability/uncertainty associated with such an estimate can be quantified using confidence intervals.
- Confidence intervals are related to hypothesis testing: if the interval does not contain the value expected by the null hypothesis, the result is statistically significant at the corresponding level.

Null-hypothesis significance testing procedure

- Define a null hypothesis (H_0) that expresses when your hypothesis is wrong.
- Define an alternative hypothesis (H_a, or H_1) as what you expect to find. (well...depending on which NHST procedure you follow.)
- Choose a significance level (α-level) at which to reject the H_0. Typical values are 0.05, 0.01, 0.001.
- Apply the appropriate test, say t-test, which will yield a p-value, of obtaining the sample you have, if H_0 was true.
- If $p < \alpha$, we reject the H_0, otherwise, we fail to reject the H_0.

NHST: problems/suggestions

Beware:

- The p-value is not the probability of null-hypothesis being true.
- Not finding a significant difference does not mean there is none: you can never accept the null hypothesis.
- Statistical significance does not warrant practical importance.

Suggestions:

- Whenever you see a p-value insert 'if null hypothesis was true' in your conclusions.
- Report value of the p (not just $p < .05$).
- Always look for effect sizes, interpret along with (confidence) interval estimates around the effect sizes.

Effect sizes: what are they?

A few examples:

- The estimate of the mean.
- The estimate of the difference between two means. Or, *Cohen's d* ($\frac{\bar{x}_1 - \bar{x}_2}{s}$), if you like standardized measures.
- Ratio or percentage of change (say, in a year, or after treatment).
- Correlation coefficient r (or r^2).
- Slope values in a regression analysis.
- Proportion of variance explained by a model: multiple-r^2 (or adjusted-r^2), η^2 (or ω^2).

It is best to interpret effect sizes with respect to the problem studied.

Statistical models

All statistical analyses can be cast into a model:

$$\mathit{response} = \mathit{model} + \mathit{error}$$

- model is what we are interested in.
- error effects the precision (and certainty) of our estimates.
- estimation is about finding a good model that fits/explains the data.
- inference is about assessing uncertainty of our estimates.

What are the models?

- Model of the mean (sometime called the null model):

$$y = \mu + e$$

- Model with multiple group means (like in ANOVA):

$$y = \mu + \delta_1 + \delta_2 + e$$

- Model with a single predictor (regression, but also t-test):

$$y = a + bx + e$$

- Model with a single predictor (regression, ANOVA):

$$y = a + b_1x_1 + b_2x_2 + \ldots + e$$

Correlation

The correlation coefficient (r) is a standardized symmetric measure of covariance between two variables.

- The correlation coefficient ranges between -1 and 1.
  - -1 perfect negative correlation: x decreases as y increase.
  - 0 no relationship.
  - $+1$ perfect positive correlation: x increases as y increase.
- Correlation is symmetric.
- Typically between two numeric variables, but also with binary categorical variables (point biserial correlation).

Correlation: how to do it

- The most common correlation coefficient is Pearson's r,

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

  r indicates the strength and direction of the correlation.

- Inference can be based on t-distribution, the base on the statistic,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- Assumptions are exactly like linear regression (coming soon).
- When the assumptions fail, non-parametric alternatives Spearman's ρ or Kendall's τ can be used.

The simple regression

$$y_i = a + bx_i + e_i$$

y is the *outcome* (or response, or dependent) variable. The index i represent each unit observation/measurement (sometimes called a 'case').

x is the *predictor* (or explanatory, or independent) variable.

a is the intercept.

b is the slope of the regression line.

a and b are called *coefficients*.

$a + bx$ is the *deterministic* part of the model. It is the model's prediction of y ($\hat{y}$), given x.

e is the residual, error, or the variation that is not accounted for by the model. Assumed to be (approximately) normally distributed with 0 mean (e_i are assumed to be i.i.d).

Regression: how to do it

Least-squares regression is the method of determining regression coefficients that minimizes the sum of squared residuals (SS_R).

$$y_i = \underbrace{a + bx_i}_{\hat{y}_i} + e_i$$

Regression: how to do it

Least-squares regression is the method of determining regression coefficients that minimizes the sum of squared residuals (SS_R).

$$y_i = \underbrace{a + bx_i}_{\hat{y}_i} + e_i$$

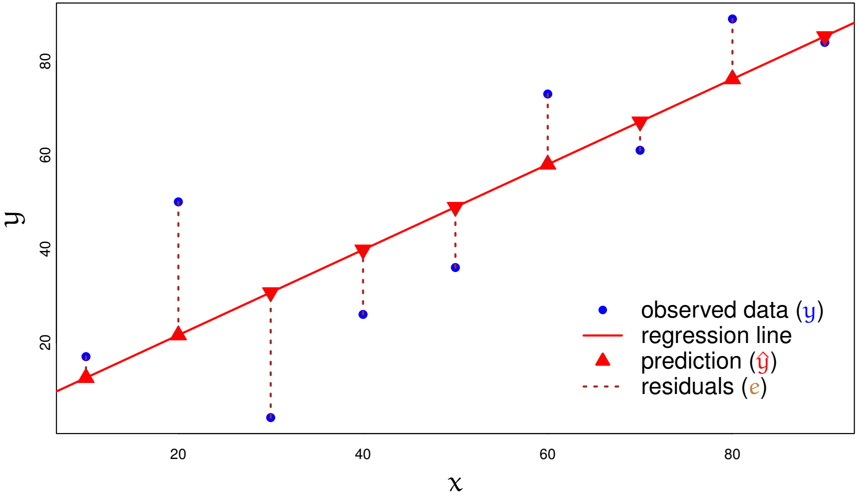
- We try to find a and b, that minimizes the prediction error:

$$\sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (a + bx_i))^2$$

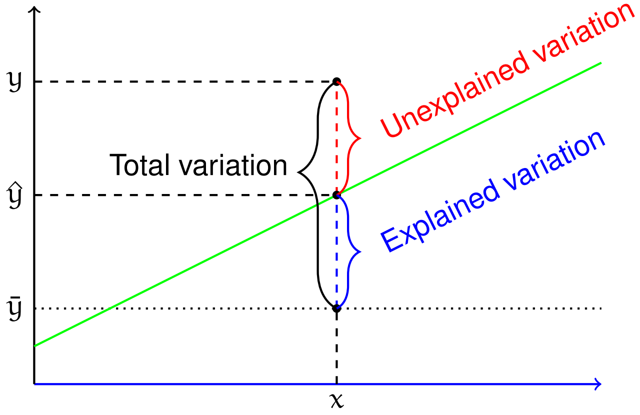
- This minimization problem can be solved analytically, yielding:

$$b = r\frac{\sigma_y}{\sigma_x}$$
$$a = \bar{y} - b\bar{x}$$

Visualization of regression procedure



Variation explained by regression



Total variation	=	Unexplained variation	+	Explained variation
$y - \bar{y}$	=	$y - \hat{y}$	+	$\hat{y} - \bar{y}$

Regression: what to watch out for

linearity: scatter plot of 'y vs. x' or 'residuals vs. fitted'.

normality: (of residuals!) histogram, Q-Q (or P-P) plot.

constant variance: (of residuals!) 'residuals vs. fitted' plot.

outliers: scatter plot of 'y vs. x' together with regression line, residual histogram or box plot.

influential cases: scatter plot of 'y vs. x', 'residuals vs. fitted', or more specialized statistics like *Cook's distance*.

Regression: when things are not as expected

When things fail ...

- **independence** use more complex models (e.g., multilevel/mixed-effect models).
- **linearity** transform the input or the response variable, use non-linear regression.
- **normality** transform the input or the response variable, use GLMs with non-normal error.
- **constant variance** transform the input or the response variable, use GLMs.
- **influential cases** remove the observation (if it is a real outlier), or collect more data.

Regression: important concepts

- Coefficient of determination

$$r^2 = \frac{\text{Explained variance}}{\text{Total variance}} = \frac{\sum_i^n (\hat{y}_i - \bar{y}_i)^2}{\sum_i^n (y_i - \bar{y}_i)^2} = \frac{SS_M}{SS_T}$$

- r^2 is the standardized effect size for regression. Estimates of slope(s) indicate effect sizes of individual predictors.
- Inference for the complete model is based on F distribution with $DF = (k, n - k - 1)$

$$F = \frac{\text{Explained variance}}{\text{Unexplained variance}} = \frac{\frac{1}{k}\sum_i^n (\hat{y}_i - \bar{y}_i)^2}{\frac{1}{n-k-1}\sum_i^n (y_i - \hat{y}_i)^2} = \frac{MS_M}{MS_R}$$

  for n data points and k predictors.
- Inference (confidence intervals or significance testing) for individual coefficients are performed using t-test.

Multiple regression

$$y_i = \underbrace{a + b_1 x_{i,1} + b_2 x_{2,i} + \ldots + b_k x_{k,i}}_{\hat{y}} + e_i$$

a is the intercept (as before).

$b_{1..k}$ are the coefficients of the respective predictors.

e is the error term (residual).

It is a generalization of simple regression with some additional power and complexity.

Multiple regression: issues and difficulties

Multiple regression shares all aspects/assumptions of simple regression, and

- Visual inspection of the data becomes more difficult.
- Multicollinearity causes problems in estimation and interpretation of multiple-regression models.
- Suppression is another possibility, where combination of predictors are more useful than individual predictors.
- Overfitting, occurs when there are large number of predictors.
- Model selection (finding a model that fits the
- Model fit is still measured by r^2 (but, called multiple-r^2). Adjusted-r^2 corrects by-chance increase in multiple-r^2 by adding more predictors.

ANOVA

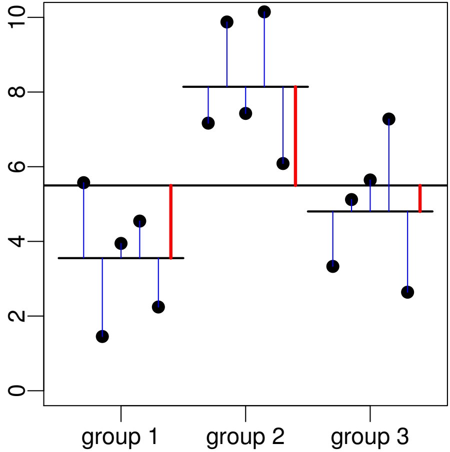
We want to know whether there are **any** differences between the means of k groups.

- If the variance between the groups is higher than the variance within the groups, there must be a significant group effect.
- Between group variance ($MS_{between}$, or MS_M or MS_G) is characterized by variance between the group means.
- Within group variance (MS_{within}, or MS_R or MS_E) is characterized by variance of data round the group means.

Then, the statistic of interest is

$$F = \frac{MS_{between}}{MS_{within}} = \frac{MS_M}{MS_R}$$

ANOVA: visualization



$$F = \frac{MS_M}{MS_R}$$

$$F = \frac{\frac{SS_M}{DF_M}}{\frac{SS_R}{DF_R}}$$

$$DF_M = k - 1$$
$$DF_R = n - k$$

where k is the number of groups, and n is the number of observations.

ANOVA: what to watch out for

normality of response in all groups check with,
- box plots,
- histogram,
- Q-Q (or P-P) plot.

homogeniety of variance among the groups.
- Rule of thumb: no variance twice another group's variance.
- Box plots for visual inspection.
- Formal tests include 'Levene' or 'Bartlett' tests of homogeneity of variances.

ANOVA: when things go wrong

independence
: Use repeated-measures ANOVA, or multilevel/mixed-effect linear models.

normality
: Transform the response variable, or use non-parametric Kruskal–Wallis test.

homogeniety of variance
: Use corrected F-ratios, transform the response variable.

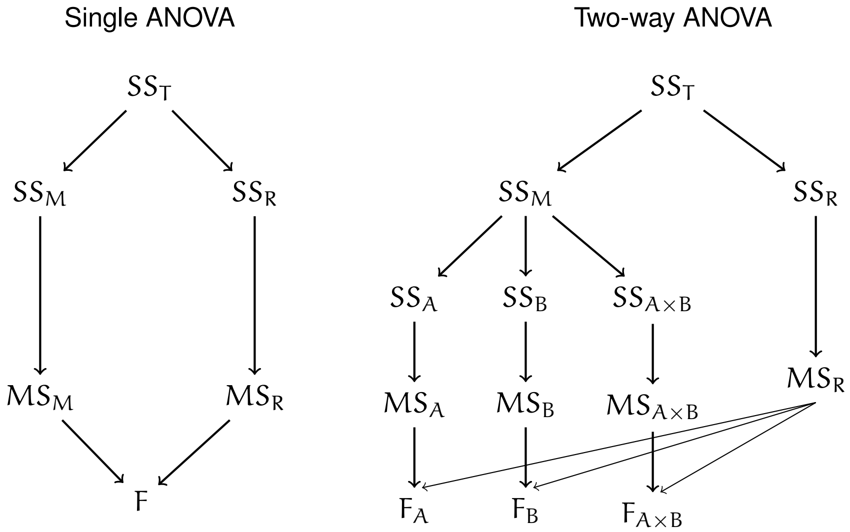
Prior contrasts and post-hoc tests

- ANOVA indicates whether there are **any** differences between any pair of group means.
- A limited set of specific differences (contrasts) can be coded in ANOVA.
- One can also do post-hoc tests for comparing individual group means after ANOVA.
- In exploratory multiple-comparison analysis, you need to adjust your p-values (or your α level), for example using Bonferroni correction.

Factorial ANOVA

- Factorial ANOVA is a generalization of single ANOVA (or t-test).
- Compare groups along more than one dimension.
- Efficient in use of subjects.
- Allows to investigate interaction.
- Same assumptions with single ANOVA.
  - independent observations.
  - all groups are (approximately) normally distributed
  - all groups have (approximately) equal variances

Factorial ANOVA: partitioning the variance



ANOVA: main effects and the interaction(s)

- For two-way ANOVA, with factors A and B, SS_M is partitioned as:

$$SS_{\mathrm{M}} = \underbrace{SS_{\mathrm{A}} + SS_{\mathrm{B}}}_{\text{main effects}} + \underbrace{SS_{\mathrm{A}\times\mathrm{B}}}_{\text{interaction}}$$

- For three-way ANOVA, with factors A, B and C, SS_M is partitioned as:

$$SS_{\mathrm{M}} = \underbrace{SS_{\mathrm{A}} + SS_{\mathrm{B}} + SS_{\mathrm{C}}}_{\text{main effects}} + \underbrace{SS_{\mathrm{A}\times\mathrm{B}} + SS_{\mathrm{A}\times\mathrm{C}} + SS_{\mathrm{B}\times\mathrm{C}}}_{\text{2-way interctions}} + \underbrace{SS_{\mathrm{A}\times\mathrm{B}\times\mathrm{C}}}_{\text{3-way inter.}}$$

Factorial ANOVA: degrees of freedom and F-tests

As in single ANOVA:

$$\begin{array}{rcccc} DF_T & = & DF_M & + & DF_R \\ n-1 & = & k-1 & + & n-k \end{array}$$

If we have k_A levels due to factor A, and k_B levels due to factor B, total number of groups is $k = k_A \times k_B$. We can now further partition the DF_M as,

$$\begin{array}{rcccccc} DF_M & = & DF_A & + & DF_B & + & DF_{A\times B} \\ k-1 & = & k_A-1 & + & k_B-1 & + & (k_A-1)\times(k_B-1) \end{array}$$

For two-way ANOVA we get three F-tests:

$$\begin{array}{rcl} F_A & = & \frac{MS_A}{MS_R} \\ F_B & = & \frac{MS_B}{MS_R} \\ F_{A\times B} & = & \frac{MS_{A\times B}}{MS_R} \end{array}$$

Repeated-measures ANOVA

Essentially, (factorial) ANOVA, with repeated (not independent) measurements.

- A lot more economical in experiment design.
- More powerful, since individual variation is not a problem for RM ANOVA.
- A generalization of paired t-test to multiple groups.

Repeated measures can be,

over time: testing effects of treatment, teaching method or just time. Typically you get more than two pre-tests or post-tests.

not time related. Examples:

- reaction time for different sort of stimuli
- measurements taken in the same city/region/country

RM ANOVA: Between subjects and within subjects variance

- A between subjects variance is the variation you observe due to differences between individuals.
- In independent (single or factorial) ANOVA, all variation observed is between subjects.
- A within subjects variation is due to variation observed in repeated measurement over the same subject.
- In a purely repeated design ANOVA, all experimental effect is confined in within-subjects variance.

Note: measures do not have to be repeated over ‘subjects’, can be other ‘items’ present in the experimental setup.

RM ANOVA: Partitioning the variance

Single ANOVA

$$\begin{tikzcd}
 & SS_T \arrow[dl] \arrow[dr] & \\
SS_M \arrow[d] & & SS_R \arrow[d] \\
MS_M \arrow[dr] & & MS_R \arrow[dl] \\
 & F &
\end{tikzcd}$$

Repeated Measures ANOVA

$$\begin{tikzcd}
 & SS_T \arrow[dl] \arrow[drr] & & & \\
SS_{between} & & & SS_{within} \arrow[dl] \arrow[dr] & \\
 & & SS_M \arrow[d] & & SS_R \arrow[d] \\
 & & MS_M \arrow[dr] & & MS_R \arrow[dl] \\
 & & & F &
\end{tikzcd}$$

RM ANOVA: what to watch out for

- Assumptions
  - Normality of response variable in all experimental conditions.
  - Sphericity: homogeneity of variances of all pairwise differences.
- RM ANOVA is very sensitive to unbalanced designs, missing values.
- Carry-over effects (e.g., learning or fatigue) in experiment sequence.

RM ANOVA: when things fail

normality
: transformation or more complex models (generalized linear multilevel/mixed-effect models) may help.

sphericity
: use adjusted F-values or again complex models (generalized linear multilevel/mixed-effect models) may help.

unbalanced data
: generalized linear multilevel/mixed-effect models, or recollect your data more carefully.

carryover effects
: randomize the order of stimuli during the experiment, or switch to between-subjects designs, do multiple experiments.

ANOVA and effect size

- ANOVA as a model view:
  - η^2 ($= r^2$, same calculation, same interpretation, just different name).

$$\eta^2 = \frac{\text{Explained variance}}{\text{Total variance}} = \frac{SS_M}{SS_T}$$

  - partial-η^2 in factorial ANOVA gives variance explained by each factor (or interaction term).
  - Analogous to adjusted-r^2, ω^2 is adjusts for by-chance increase in η^2. Use/report (partial-)ω^2 when you can.
- ANOVA as hypothesis testing method:
  - Mean differences (or Choen's d) in pairwise comparisons.
  - Coefficients of contrasts.

Logistic regression

Logistic regression is an extension of regression (or a case of generalized linear models) where response variable is binary.
Two important differences:

- Transform the response variable so that estimated values are between 0 and 1.
- Allow non-normal residuals.

$$\underbrace{\mathrm{logit}(p_i)}_{\log \frac{p}{1-p}} = a + b_1 x_{1,i} + \ldots + b_k x_{k,i} + e_i$$

Logistic regression: estimation

- Maximum likelihood estimation (MLE) tries to find the set of model parameters, or coefficients, a, b_1, $\ldots b_k$, which make the data most likely (or minimize the error).
- MLE is an iterative search for the optimum parameter values. There is no exact solution.
- In some cases, MLE may fail to find a solution.
- If errors are normally distributed, MLE is equivalent to least-squares estimation.
- With MLE, r^2 is not the measure of model fit. Instead we use $deviance = -2LogLikelihood$ to measure model fit (lower, better).
- Unlike r^2, deviance is not comparable for models fit on different data.

Logistic regression: what to watch out for

- Overdispersion: when variance diverges from what is expected in binomial data.
- Linear relationship between logit transformed response and predictors.
- MLE related: MLE may fail to find a good fit. In case of
  - complete separation.
  - unevenly distributed data points.
- Otherwise the same as multiple regression.

Logistic regression: when things fail

overdispersion GLMs with quasi-binomial error.

MLE fails Collect more data, or use Bayesian estimation.

independence Same as regression: multilevel (generalized) linear models.

linearity Same as regression: transform predictor/response or use non-linear regression.