Index of Lecture 6: ANOVA, Contrasts, Multiple comparisons

Page	Title
1	Practical information
2	One-way ANOVA model and parametrization
3	ANOVA versus regression
4	1-way ANOVA – analysis (recap)
5	How to proceed after the ANOVA test?
6	Multiple comparisons: Overview
7	Bonferroni method
8	Holm method
9	Multiple comparisons: Rat data example
10	More about contrasts
11	Scheffè's method
12	Contrasts: Rat data example
13	Contrasts for 1-way ANOVA with quantitative groups
14	Beyond one-way ANOVA
15	Additional software notes

Practical Information

Today's lecture: expanded review of ANOVA (and regression) with some new topics: contrasts and multiple comparisons.

Notes on for textbook reading (i.e., the GO text):

- Chapter 3 on one-way ANOVA: mostly well-known,[1]
- Chapter 4 on contrasts: short chapter, skip Section 4.4,[2]
- Chapter 5 on multiple comparisons: more detailed than our ambition level, don't focus on mathematical details and read cursorily from Section 5.4.2 onwards.[2]

News/Schedule:

- logistic regression home assignment due soon (February 17),
- lab session this Friday (tomorrow!) 1-4pm (no lab on Monday),
- no mid-semester break next week,
- time to start thinking about data for your project.

[1] Skip Sections 3.9 and 3.11, and discussion of P(p) on p. 49.
[2] See also articles on use of contrasts and multiple comparisons at media page.

One-way ANOVA Model and Parametrization

Rat data example (GO Exercise 3.1, p. 60):

- rat liver weights in percent of body weight following four diets (labelled 1-4) randomly allocated to rats,

$$y_{ij} = \text{rat liver weight for } j\text{th rat in diet group } i,$$
$$i = 1, \ldots, g \ (g=4); \ j = 1, \ldots, n_i \ (n_1=7, n_2=n_4=8, n_3=6),$$

- purpose: assess impact of diets on liver weight.

Statistical model:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \ i = 1, \ldots, g; \ j = 1, \ldots, n_i,$$

where the ε_{ij} are i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$.

Model parameters (referring to underlying population):

- group means $\mu_1, \ldots, \mu_4$, and common within-group/error standard deviation σ.

Alternative formulations of same model (recap from 2aL−4):

$$y_i = \mu_{\text{diet}(i)} + \varepsilon_i, \quad i = 1, \ldots, 29 \ (\sim \text{observation number}),$$
$$y_i = \beta_0 + \beta_1 1_{\text{diet2}(i)} + \beta_2 1_{\text{diet3}(i)} + \beta_3 1_{\text{diet4}(i)} + \varepsilon_i,$$
$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (\text{or, } y_i = \mu + \alpha_{\text{diet}(i)} + \varepsilon_i), \quad \text{with restrictions on } \alpha\text{'s.}^3$$

[3] Restrictions on (α_i): either $\alpha_1 = 0$ (Stata; Minitab `Regression` (default); R), $\alpha_4 = 0$ (SAS), or $\alpha_1 + \ldots + \alpha_4 = 0$ (Minitab `General Linear Model` (default)).

ANOVA versus Regression

= two different frameworks for analyzing the same model and presenting the results.[4]

Advantages of ANOVA framework:

- no reliance on an, often artificial, reference category,[5]
- extra tools for exploring multiple samples and/or multiple factors, in particular for balanced data.

Advantages of regression framework:

- easier to include continuous predictors (the equivalent of "analysis of covariance" (ANCOVA), which no longer plays any prominent role in ANOVA methods),
- full range of model checking and diagnostic tools (although some of these are of questionable value for categorical predictors; e.g. VIF and leverage).

Minitab vs. Stata: (see 6L−15 for notes on R and SAS)

- more complete regression and ANOVA facilities in Stata,
- more easily accessible facilities in Minitab.

[4] From Oehlert (p. 44): "Strictly speaking, ANOVA is an arithmetic procedure for partitioning the variability in a data set [...], however [...] we sometimes speak of testing via ANOVA although the test is not really part of the ANOVA." Other authors (e.g., Christensen 1996, p. 132) use ANOVA "as a name for the entire package of techniques used to compare more than two samples".

[5] A common mistake within the regression framework is to explore only comparisons with the reference category, cf. 4bL−10.

1-Way ANOVA – Analysis (Recap)

Estimation: (g groups, N observations)

- $\hat{\mu}_i = \bar{y}_{i\cdot}$ ($\mathrm{Var}(\hat{\mu}_i) = \sigma^2/n_i$, $\mathrm{SE}(\hat{\mu}_i) = s/\sqrt{n_i}$),
- $\hat{\sigma}^2 = s^2 = \sum_i (n_i - 1)s_i^2/[N-g] = [\sum_{ij}(y_{ij} - \bar{y}_{i\cdot})^2]/[N-g] = \mathrm{SS}_E/\mathrm{DF}_E = \mathrm{MS}_E$,
  – weighted average of the group sample variances s_i^2,
- confidence intervals and tests: "4-step procedure" (L1a–5).

ANOVA table: (g groups, N observations)

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Groups/Treatments	$\mathrm{DF}_{\mathrm{Trt}} = g-1$	$\mathrm{SS}_{\mathrm{Trt}} = \sum_i n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$\mathrm{MS}_{\mathrm{Trt}} = \mathrm{SS}_{\mathrm{Trt}}/\mathrm{DF}_{\mathrm{Trt}}$	$\mathrm{MS}_{\mathrm{Trt}}/\mathrm{MS}_E$
Error	$\mathrm{DF}_E = N-g$	$\mathrm{SS}_E = \sum_{ij}(y_{ij} - \bar{y}_{i\cdot})^2$	$\mathrm{MS}_E = \mathrm{SS}_E/\mathrm{DF}_E$	
Total	$\mathrm{DF}_T = N-1$	$\mathrm{SS}_T = \sum_{ij}(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$\mathrm{MS}_T = \mathrm{SS}_T/\mathrm{DF}_T$	

- F-test in table is for hypothesis H_0: $\mu_1 = \ldots = \mu_g$ (all groups equal, homogeneity between groups) against alternative hypothesis H_a: some μ's differ,
- P-value (for F-test) $= \Pr(F \geq F_{\mathrm{obs}})$; $F \sim F(\mathrm{DF}_{\mathrm{Trt}}, \mathrm{DF}_E)$,
- $\mathrm{E}(\mathrm{MS}_E) = \sigma^2$ and $\mathrm{E}(\mathrm{MS}_{\mathrm{Trt}}) = \sigma^2 + \sum_i n_i(\mu_i - \bar{\mu})^2/(g-1)$,
- (technical) the ANOVA decomposition is based on the equation

$$(y_{ij} - \bar{y}_{\cdot\cdot}) = (y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}).$$

How to Proceed after the ANOVA Test?

We rejected overall H_0: $\mu_1 = \ldots = \mu_g$, but what relations exist between μ_i's (which differ)?

- estimation of parameters: $\hat{\mu}_i = \bar{y}_{i\cdot}$ and of derived parameters such as contrasts:

$$w(\{\mu_i\}) = \sum_{i=1}^{g} w_i \mu_i = \sum_{i=1}^{g} w_i \alpha_i, \quad \text{with } \sum_i w_i = 0 \text{ (required!)},$$

  — examples (for $g = 3$, i.e. 3 groups):

  * $w(\{\mu_i\}) = \mu_1 - \mu_2$ (i.e., $w_1 = 1$, $w_2 = -1$, $w_3 = 0$),
  * $w(\{\mu_i\}) = \frac{1}{2}(\mu_2 + \mu_3) - \mu_1$ (i.e., $w = (-1, \frac{1}{2}, \frac{1}{2})$),

- confidence intervals/tests for interesting parameters, e.g., using $t^* = t(1 - \frac{\alpha}{2}, \mathrm{DF}_E)$:

$$\mu_i \ : \ \bar{y}_{i\cdot} \pm t^* \sqrt{\mathrm{MS}_E} \sqrt{(1/n_i)},$$
$$\mu_i - \mu_{i'} \ : \ \bar{y}_{i\cdot} - \bar{y}_{i'\cdot} \pm t^* \sqrt{\mathrm{MS}_E} \sqrt{(1/n_i) + (1/n_{i'})},$$ [6]

- diagram, e.g. $\hat{\mu}_i$'s with error bars (interval or margins plot),
- problems with choice of contrasts/pairwise comparisons:
  * many hypotheses; if each test has error of 5%, then total error is $\gg 5\%$,
  * above methods apply only to preplanned hypotheses, not to hypotheses suggested by the data,

  — therefore always an advantage to have any other hypotheses (in addition to the overall H_0) defined prior to analysis.

[6] Note: the margin of error equals the LSD (least significant difference) for unadjusted comparisons between groups.

Multiple Comparisons: Overview

Some terminology and basic facts:

- type I (error) probability: probability of rejecting H_0, when H_0 is actually true, [6]
- per comparison or individual error rate: type I probability for each test,
- simultaneous or experimentwise or familywise error rate: type I probability for all tests, i.e., for rejection of any test in a set ("family") of tests carried out; larger than individual error rate(s),
- strong familywise error rate: probability of rejecting any true null hypotheses (but no impact of false null hypotheses/true rejections),
- multiple comparison procedures reduce the type I prob. and increase type II prob. — a trade-off,
- doing (very) many pairwise *t*-tests (LSD or Fisher method) is (very) liberal, i.e. likely to have false significances.

(Relatively) simple methods (in this course):

- Bonferroni & Holm corrections for preplanned or all comparisons,
- Scheffé's method for contrasts suggested by the data.

Other methods exist (in abundance), and can be used to control specific error rates (GO Display 5.2):

- many require balanced data (Tukey, Duncan) for exact inference,
- some are for special cases (e.g., Dunnett for comparison with control),
- some assume independent tests (e.g., Benjamini & Hochberg's FDR[7] method).

[6] Conversely, the type II (error) probability is for not rejecting H_0, when it is false.

[7] The false discovery rate (FDR) is the proportion of false rejections out of all rejections, and thus includes the true rejections.

Bonferroni Method

Idea: if we make K comparisons (tests), we can achieve the simultaneous type I probability for all tests to be $\leq \epsilon$, by taking the type I probability for each test equal to ϵ/K,[8]

- either by changing the significance level (to ϵ/K), or by multiplying P-values for individual tests by K (while keeping the significance level unchanged),
- K may be the number of preplanned comparisons, or for unplanned comparisons[9]:

$$K = \text{total number of comparisons} = \binom{g}{2} = g\,(g-1)/2.$$

Notes for Bonferroni method:

- gives also simultaneous confidence intervals,[10]
- is conservative (i.e., may lead to too few hypotheses rejected) for controlling the strong familywise error rate for unplanned comparisons,
- is available for ANOVA in Minitab only via `General Linear Model` followed by `Comparisons`; in Stata, available by `pwcompare` and `test` commands,
- is flexible: applies to a wide range of settings/models, and can be applied to a subset of comparisons.

[8] The (mathematical) justification for the Bonferroni method was explained in VHM 801 (Lecture 9).
[9] Unplanned comparisons include comparisons suggested by the data, e.g. involving the lowest/highest groups.
[10] The probability that all CIs simultaneously cover their true value is $\geq 1-\epsilon$.

Holm Method

Steps of this sequential (also called step-down) procedure:

1) sort the K unadjusted P-values as:

$$P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(K)},$$

2) for the test corresponding to the ith ordered P-value, compute the adjusted P-value: $P^H_{(i)} = P_{(i)} \times (K - i + 1)$, for $i = 1, \ldots, K$,

3) rules for significance:

(i) if $P^H_{(i)} > \epsilon \Rightarrow$ non-significant (at ϵ),

(ii) if $P^H_{(i)} \leq \epsilon$ and also all $P^H_{(j)} \leq \epsilon$ for all $j = 1, \ldots, i$, $\Rightarrow$ significant (at ϵ).

Notes for Holm method:

- controls the strong familywise error rate, and is less conservative for this than the Bonferroni method,
- does not provide simultaneous confidence intervals,
- is not available in Minitab, but can be carried out manually (by the recipe above),
- adjusted P-values (i.e., the $P^H_{(i)}$ above) are available in Stata (`test` command), but the sequential rule (ii) must be checked manually,
- is also flexible: applies to the same wide range of settings/models.

Multiple Comparisons: Rat Data Example

- group means: $\hat{\mu}_1 = 3.75,\ \hat{\mu}_2 = 3.58,\ \hat{\mu}_3 = 3.60,\ \hat{\mu}_4 = 3.92,$
- assume no preplanned hypotheses or treatment (diet) structure of interest,
- a total of $K = 4 \cdot (4-1)/2 = 6$ pairwise comparisons, shown in the table:

P-value		Multiple comparison method		
pair	order	unadjusted	Bonferroni	Holm
2 vs 4	1	.0025	.015	.015
3 vs 4	2	.0068	.041	.034
1 vs 4	3	.106	.633	.422
1 vs 2	4	.128	.768	.384
1 vs 3	5	.205	1	.409
2 vs 3	6	.869	1	.869

* same conclusions by all methods: only 4 vs 2,3 significant
* Holm $P <$ Bonferroni P (except for first P)

Significance letter coding (groups with same letter not significantly different; available in Minitab `General Linear Model` and Stata `pwcompare`):

- order group means from highest to lowest,
- designate letter a to highest group + all groups not significantly different from it,
- designate letter b to next group in the same way (but drop if same pattern as for a),
- continue through all groups,
- Rat data coding: $4^a\ 1^{ab}\ 3^b\ 2^b$.

More about Contrasts

- constructed to reflect specific (ideally, pre-defined) hypotheses,
- inference by 4-step approach from formulas for estimates and SEs, for $w = w(\{\mu_i\})$:

$$\widehat{w} = \sum_i w_i \bar{y}_{i\cdot} \quad \text{and} \quad \mathrm{SE}(\widehat{w}) = \sqrt{\mathrm{MS}_E}\sqrt{\sum_i w_i^2/n_i}\ .$$

- variation explained by contrast (out of $\mathrm{SS}_{\mathrm{Trt}}$ for the grouping/model):

$\mathrm{SS}(w) = \widehat{w}^2/(\sum_i w_i^2/n_i) = \mathrm{MS}_E \times t_w^2$, where t_w = the t-statistic for testing $w=0$,

- orthogonal contrasts:
  - idea: contrasts that explain different parts of the variation, to allow independent interpretation,[11]
  - definition: $w = \sum_i w_i \mu_i$ and $w^* = \sum_i w_i^* \mu_i$ are orthogonal if: $\sum_i w_i w_i^*/n_i = 0$,
  - fact: there exist at most $(g-1)$ pairwise orthogonal contrasts among g groups; these are not unique,
  - example: (3 groups, equal n_i's)

$$w = \mu_1 - \tfrac{1}{2}(\mu_2+\mu_3), \quad \text{and} \quad w^* = \mu_2 - \mu_3,$$

  - main result: for orthogonal contrasts $w^{(1)}, \ldots, w^{(g-1)}$, it holds that

$$\mathrm{SS}_{\mathrm{Trt}} = \mathrm{SS}(w^{(1)}) + \mathrm{SS}(w^{(2)}) + \ldots + \mathrm{SS}(w^{(g-1)}),$$

  — splitting (decomposing) $\mathrm{SS}_{\mathrm{Trt}}$ into contrast parts.

[11] In practice, it is not always easy to find useful orthogonal contrasts, in particular in unbalanced designs.

Scheffé's Method

- corrects for examining non-preplanned contrasts[12], and therefore "allows" to test contrasts suggested by the data,
- not available in Minitab/Stata $\Rightarrow$ manual calculation.

Idea: use same procedure as with preplanned contrasts, but replace the reference distribution:

$$\text{not} \quad \frac{\widehat{w}-w}{\text{SE}(\widehat{w})} \sim t(\text{DF}_E), \quad \text{but} \quad [\frac{\widehat{w}-w}{\text{SE}(\widehat{w})}]^2/(g-1) \sim F(g-1, \text{DF}_E),$$ [13]

for example,

- test of H_0: $w = 0$ by $F = [\frac{\widehat{w}}{\text{SE}(\widehat{w})}]^2/(g-1) \sim F(g-1, \text{DF}_E)$,
- 95% CI for w: $\widehat{w} \pm \sqrt{(g-1)F(.95, g-1, \text{DF}_E)}\,\text{SE}(\widehat{w})$.

Properties:

- method can never give stronger result than overall F-test ($\text{H}_0: \mu_1 = \ldots = \mu_g$),
- there always exists a contrast to give exactly same result as the overall F-test (but that contrast is usually not interesting),
- method is conservative.

[12] Method should not be used for pairwise comparisons, because in this situation it will be very conservative.
[13] Based on the mathematical relation: $[(\widehat{w}-w)/\text{SE}(\widehat{w})]^2/(g-1) \leq \text{MS}_{\text{Trt}}/\text{MS}_E$.

Contrasts: Rat Data Example

- group means: $\hat{\mu}_1 = 3.746,\ \hat{\mu}_2 = 3.580,\ \hat{\mu}_3 = 3.598,\ \hat{\mu}_4 = 3.923,$
- choice of contrasts guided by group means (in absence of biological hypotheses),
- table of estimates and statistics:

Contrast	Weights				Estim.	Stand. err.	Variation	Tests	
definition	w_1	w_2	w_3	w_4	$\hat{w}$	SE($\hat{w}$)	SS($\hat{w}$) (%)	Wald t	Scheffé F
$(\mu_1+\mu_2+\mu_3)/3-\mu_4$	1/3	1/3	1/3	−1	−0.281	0.085	0.456 (78.9%)	−3.32	3.67
$(\mu_2+\mu_3)/2-\mu_1$	−1	1/2	1/2	0	−0.157	0.094	0.114 (19.6%)	−1.66	0.92
$\mu_2-\mu_3$	0	1	−1	0	−0.018	0.110	0.001 (0.2%)	−0.17	0.01

- t-tests are assessed in $t(\mathrm{DF}_E) = t(25)$ with $t(0.975, 25) = 2.06$;
  F-tests are assessed in $F(g-1, \mathrm{DF}_E) = F(3, 25)$ with $F(0.95, 3, 25) = 2.99$.

Interpretations:

- contrast between last and first three groups $\sim$ 80% of variation, and is significant both with Wald test ($\sim$ pre-planned) and Scheffé test,
- other contrasts are far from significant (in particular with F-test),
- due to the unequal group sizes, the contrasts are not orthogonal (seen for example by their proportions out of $\mathrm{SS}_{\mathrm{Trt}}$ not adding up to 100%).

Contrasts for 1-Way ANOVA with Quantitative Groups

Resin data example: log failure time y_i of unit i subjected to temperature x_i, where $x_i \in \{175, 194, 213, 231, 250°\text{C}\}$ and $i = 1, \ldots, 37$.

Orthogonal polynomial contrasts for 1-way ANOVA model:

- $\sim$ model reductions between polynomial regression models,
- split SS_{Trt} from 1-way ANOVA into interpretable terms,
- coefficients (w_i) listed in textbook Appendix Table D.6.[14]

Illustration: polynomial regression model hierarchy:

Model for y_i	Inter-pretation	Contrast estim. (SE)	Contrast SS	Model SS	Model DF	Error SS
$\mu_{\text{temp}(i)} + \varepsilon_i$	ANOVA			3.538	5	0.294
$\updownarrow$		(same model!)				
$\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i$	4^{th} order regr.			3.538	5	0.294
$\downarrow$		-0.038 (.289)	0.000			
$\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$	cubic regr.			3.538	4	0.294
$\downarrow$		-0.007 (.112)	0.000			
$\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$	quadratic regr.			3.538	3	0.294
$\downarrow$		0.400 (.133)	0.083			
$\beta_0 + \beta_1 x_i + \varepsilon_i$	linear regr.			3.459	2	0.372
$\downarrow$		-2.19 (.115)	3.332			
$\beta_0 + \varepsilon_i$	i.i.d. sample			0	1	3.831

Conclusion: linear and quadratic contrasts significant, others ≈ 0

$\Rightarrow$ quadratic regression model gives best model fit.

[14] The coefficients are valid for equidistant x's and equal group sizes.

Beyond One-way ANOVA

Methods reviewed for one-way ANOVA are generalisable to varying extent:

- in multiple regression models, each categorical predictor can/should be assessed separately unless part of interactions (some methods also important for interaction terms $\Rightarrow$ later lectures),
- construction and assessment of contrasts work for "all" regression models, except that proportion of variation explained is limited to linear models where the factor in question is "unaffected by" (orthogonal to) other effects,
- multiple comparisons are relevant in "all" regression models, but not all methods apply:
  - Bonferroni and Holm methods generally applicable,
  - many methods limited to balanced ANOVAs, and only few methods extend to GLMs,
  - Scheffé's method can be applied for Wald-type z-statistics in GLMs by comparing z^2 to a $\chi^2(g-1)$ distribution, where g = number of groups,
  - principles/ideas behind adjustment (e.g. the distinction between different error rates) for multiple comparisons are general.

Additional Software Notes

R analysis of one-way ANOVA and beyond:

- `oneway.test()` and `pairwise.t.test()` for one-way ANOVA with multiple comparisons,
- `lm()` and `glm()` functions for fitting linear and generalized linear models (incl. logistic regression), respectively,
- `coef()` and `vcov()` functions extract estimates and the variance-covariance matrix, respectively; further manipulation requires vector/matrix programming (e.g. using `se.contrast()` function) or pre-developed package interface,
- `multcomp` package offers wide variety of multiple comparison procedures, see documentation for use with `lm` and `glm` model fits.

SAS analysis of one-way ANOVA and beyond:

- `proc ANOVA`: one-way and multiple ANOVA,
  - limited to balanced designs,
  - includes multiple comparison methods (`means` statement),
- `proc glm`: linear models without any restrictions,
  - includes multiple comparison methods (`lsmeans` statement),
  - includes contrasts (`contrast` and `estimate` statements),
- `proc logistic` (logistic regression) and `proc genmod` (generalized linear models),
  - include contrasts (`contrast` statement), but no multiple comparisons,
- `proc multtest`: general multiple testing procedure, for linear models and import of set of unadjusted P-values.