Statistiek II

John Nerbonne
using reworkings by Hartmut Fitz and Wilbert Heeringa

Dept of Information Science
j.nerbonne@rug.nl

February 13, 2014

Course outline

1. One-way ANOVA.
2. Factorial ANOVA.
3. Repeated measures ANOVA.
4. Correlation and regression.
5. Multiple regression.
6. Logistic regression.
7. Hierarchical, or “mixed” models

Lecture outline

Today: One-way ANOVA

1. General motivation
2. *F*-test and *F*-distribution
3. ANOVA example
4. The logic of ANOVA

Short break

5. ANOVA calculations
6. Post-hoc tests

What's ANalysis Of VAriance (ANOVA)?

- Most popular statistical test for numerical data
- Generalized t-test
- Compares means of more than two groups
- Fairly robust
- Based on F-distribution
- compares variances (between groups and within groups)
- Two basic versions:
  a One-way (or single) ANOVA: compare groups along one dimension, e.g., grade point average by school class
  b N-way (or factorial) ANOVA: compare groups along ≥ 2 dimensions, e.g., grade point average by school class and gender

Typical applications

- One-way ANOVA:
  Compare time needed for lexical recognition in
  1. healthy adults
  2. patients with Wernicke's aphasia
  3. patients with Broca's aphasia
- Factorial ANOVA:
  Compare lexical recognition time in male and female in the same three groups

Comparing multiple means

- For **two** groups: use *t*-test
- Note: testing for p-value of 0.05 shows significance 1 time in 20 if there is no difference in population mean (effect of chance)
- But suppose there are 7 groups, i.e., we test $\binom{7}{2} = 21$ pairs of groups
- **Caution**: several tests (on the same data) run the risk of finding significance through sheer chance

Multiple comparison problem

Example: Suppose you run $k = 3$ tests, always seeking a result significant at $\alpha = 0.05$

$\Rightarrow$ probability of getting at least one false positive is given by:

$$\begin{aligned} \alpha_{FW} &= 1 - P(\text{zero false positive results}) \\ &= 1 - (1 - \alpha)^k \\ &= 1 - (1 - 0.05)^3 \\ &= 1 - (0.95)^3 \\ &= 0.143 \end{aligned}$$

Hence, with only 3 pairwise tests, the chance of committing type I error almost 15% (and 66% for 21 tests!)

α_{FW} called Bonferroni family-wise α-level

Bonferroni correction for multiple comparisons

To guarantee a **family-wise** α-level of 0.05, divide α by number of tests.

Example: $0.05/3$ ($= \alpha/\#$ tests) $= 0.017$ (note: $0.983^3 \approx 0.95$)
$\Rightarrow$ set $\alpha = 0.017$ ($=$ Bonferroni-corrected α-level)

- If p-value is less than the Bonferroni-corrected target α: reject the null hypothesis.
- If p-value greater than the Bonferroni-corrected target α: do not reject the null hypothesis.

Analysis of variance

- ANOVA automatically corrects for looking at several relationships (like Bonferroni correction)
- Based on F-distribution: Moore & McCabe, §7.3, pp. 435–445
- Measures the difference between two variances (variance σ^2)

$$F = \frac{s_1^2}{s_2^2}$$

- always positive since variances are positive
- two degrees of freedom interesting, one for s_1, one for s_2

F-test vs. F-distribution

F-value: $F = \frac{s_1^2}{s_2^2}$

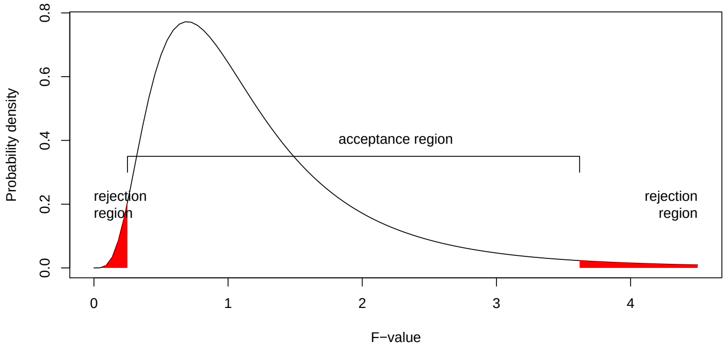
- F-values used in F-test (Fisher's test)
  H_0: samples are from same distribution ($s_1 = s_2$)
  H_a: samples are from different distributions ($s_1 \neq s_2$)
  - value near 1 indicates same variance
  - value near 0 or $+\infty$ indicates difference in variance

- F-test very sensitive to deviations from normal

- ANOVA uses F-distribution, but is different: ANOVA $\neq$ F-test!

F-distribution

Critical area for F-distribution at $p = 0.05$ (df: 12,10)



Note the symmetry: $P(\frac{s_1^2}{s_2^2} < x) = P(\frac{s_2^2}{s_1^2} > \frac{1}{x})$

(because $y < x \Leftrightarrow \frac{1}{y} > \frac{1}{x}$ for $x, y \in \mathbb{R}^+$)

F-test

Example: height

group	sample size	mean	standard deviation
boys	16	180cm	6cm
girls	9	168cm	4cm

Is the difference in standard deviation significant?

Examine $F = \frac{s^2_{\text{boys}}}{s^2_{\text{girls}}}$

Degrees of freedom: $$\begin{aligned} \text{df}_{\text{boys}} &= 16-1 \\ \text{df}_{\text{girls}} &= 9-1 \end{aligned}$$

F-test critical area (for two-tailed test with $\alpha = 0.05$)

$$
\begin{array}{rrcl}
 & P(F(15,8) > x) & = & \frac{\alpha}{2} = 0.025 \\
 & P(F(15,8) < x) & = & 1 - 0.025 \\
 & P(F(15,8) < \underline{4.1}) & = & 0.975 \quad \text{Moore \& McCabe, Table E, p. 706} \\
 & & & (\text{no values directly for } P(F(df_1, df_2) > x)) \\
 & P(F(15,8) < x) & = & 0.025 \\
\Leftrightarrow & P(F(8,15) > x') & = & 0.025 \quad \text{where } x' = \frac{1}{x} \\
\Leftrightarrow & P(F(8,15) > 3.2) & = & 0.025 \quad (\text{tables}) \\
\Leftrightarrow & P(F(15,8) < \frac{1}{3.2}) & = & 0.025 \\
\Leftrightarrow & P(F(15,8) < \underline{0.31}) & = & 0.025
\end{array}
$$

Reject H_0 if $F < 0.31$ or $F > 4.1$

Here, $F = \frac{6^2}{4^2} = 2.25$ (hence no evidence of difference in distributions)

ANOVA

Analysis of Variance (ANOVA) most popular statistical test for numerical data

- several types
  - single, "one-way"
  - factorial, "two-, three-,..., n-way"
  - single/factorial repeated measures
- examines variation
  - "between-groups"—gender, age, etc.
  - "within-groups"—overall
- automatically corrects for looking at several relationships (like Bonferroni correction)
- uses F-distribution, where $F(n, m)$ fixes n typically at the number of groups (minus 1), m at the number of subjects, i.e., data points (minus number of groups)

Detailed example: one-way ANOVA

Question: Are exam grades of **four** groups of foreign students "Nederlands voor anderstaligen" the same? More exactly, are the four averages the same?

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
$H_a : \mu_1 \neq \mu_2$ **or** $\mu_1 \neq \mu_3 \ldots$ **or** $\mu_3 \neq \mu_4$

Alternative hypothesis: at least one group has a different mean

For the question of whether any particular pair is different, the *t*-test is appropriate.

For testing whether all language groups are the same, pairwise *t*-tests *exaggerate* differences (increase the chance of type I error)

We therefore want to apply one-way ANOVA

Data: Dutch proficiency of foreigners

Four groups of ten students each:

	Group			
	Europe	America	Africa	Asia
	10	33	26	26
	19	21	25	21
	⋮	⋮	⋮	⋮
	31	20	15	21
Mean	25.0	21.9	23.1	21.3
Samp. SD	8.14	6.61	5.92	6.90
Samp. Variance	66.22	43.66	34.99	47.57

ANOVA conditions

ANOVA assumptions:

- Normal distribution per subgroup
- Same variance in subgroups: least SD $>$ one-half of largest SD
- **independent** observations: watch out for test-retest situations!

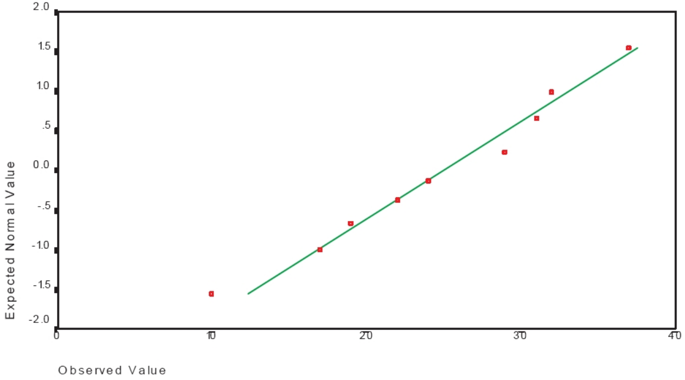
Check differences in SD's! (some SPSS computing)

Variable	Std Dev	Valid N	Label
Europa	8.14	10	
America	6.61	10	
Africa	5.92	10	
Azie	6.90	10	

ANOVA conditions

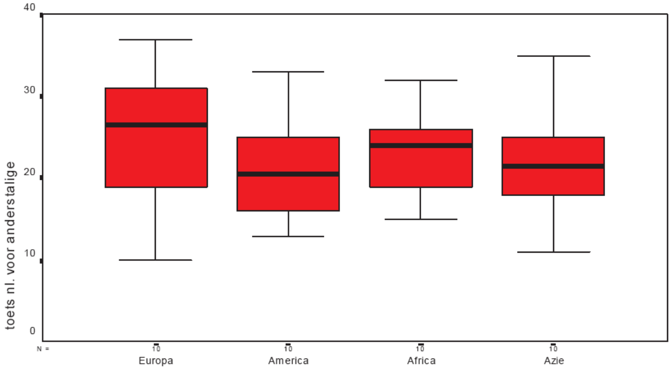
Assumption: normal distribution **per group**, check with normal quantile plot, e.g., for Europeans below (repeat for every group)

Normal Q-Q plot of toets.nl voor anderstalige



Visualizing ANOVA data

Is there a significant difference in the means (of the groups being contrasted)?



Take care that boxplots sketch **medians** not **means**.

Sketch of ANOVA

		Group	
1	2	3	4
Eur.	Amer.	Africa	Asia
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1j}	x_{2j}	x_{3j}	x_{4j}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$

Notation:

Group index: $i \in \{1, 2, 3, 4\}$
Sample index: $j \in N_i$ = size of group i
Data point x_{ij}: ith group, jth observation
Number of groups: $I = 4$
Total mean: $\overline{x}$
Group mean: $\overline{x}_i$

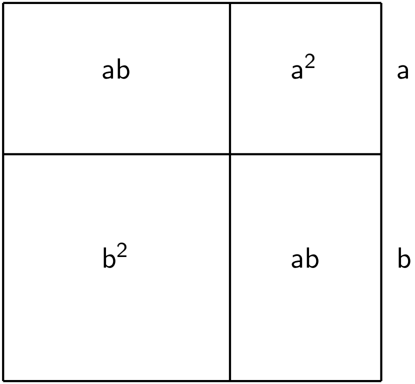
For any data point x_{ij}:

$$\begin{array}{rcccc} (x_{ij} - \overline{x}) & = & (\overline{x}_i - \overline{x}) & + & (x_{ij} - \overline{x}_i) \\ \text{total residue} & = & \text{group diff.} + & & \text{“error”} \end{array}$$

ANOVA question: does group membership influence the response variable?

Two variances

Reminder of high school algebra: $(a+b)^2 = a^2 + b^2 + 2ab$



Two variances

Data point x_{ij}:

$$(x_{ij} - \overline{x}) \quad = \quad (\overline{x}_i - \overline{x}) \quad + \quad (x_{ij} - \overline{x}_i)$$

Want sum of squared deviates for each group:

$$(x_{ij} - \overline{x})^2 \quad = \quad (\overline{x}_i - \overline{x})^2 + (x_{ij} - \overline{x}_i)^2 + 2(\overline{x}_i - \overline{x})(x_{ij} - \overline{x}_i)$$

Sum over elements in ith group:

$$\sum_{j=1}^{N_i}(x_{ij} - \overline{x})^2 \quad = \quad \sum_{j=1}^{N_i}(\overline{x}_i - \overline{x})^2 + \sum_{j=1}^{N_i}(x_{ij} - \overline{x}_i)^2 + \sum_{j=1}^{N_i} 2(\overline{x}_i - \overline{x})(x_{ij} - \overline{x}_i)$$

Two variances

Note that this term must be zero:

$$\sum_{j=1}^{N_i} 2(\overline{x}_i - \overline{x})(x_{ij} - \overline{x}_i)$$

Because:

(a) $$\sum_{j=1}^{N_i} 2(\overline{x}_i - \overline{x})(x_{ij} - \overline{x}_i) \quad = \quad 2(\overline{x}_i - \overline{x}) \underbrace{\sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)}_{0}$$

(b) $$\sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i) = 0 \quad \Leftrightarrow \quad \overline{x}_i = \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i}$$

Two variances

So we have:

$$\sum_{j=1}^{N_i} (x_{ij} - \overline{x})^2 = \sum_{j=1}^{N_i} (\overline{x}_i - \overline{x})^2 + \sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)^2$$
$$\left(+ \sum_{j=1}^{N_i} 2(\overline{x}_i - \overline{x})(x_{ij} - \overline{x}_i) = 0\right)$$

Therefore:

$$\sum_{j=1}^{N_i} (x_{ij} - \overline{x})^2 = \sum_{j=1}^{N_i} (\overline{x}_i - \overline{x})^2 + \sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)^2$$

And finally we can sum over all groups:

$$\sum_{i=1}^{I} \sum_{j=1}^{N_i} (x_{ij} - \overline{x})^2 = \sum_{i=1}^{I} \sum_{j=1}^{N_i} (\overline{x}_i - \overline{x})^2 + \sum_{i=1}^{I} \sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)^2$$

ANOVA terminology

$$
\begin{array}{ccccc}
(x_{ij} - \overline{x}) & = & (\overline{x}_i - \overline{x}) & + & (x_{ij} - \overline{x}_i) \\
\text{total residue} & = & \text{group diff.} & + & \text{"error"}
\end{array}
$$

$$
\begin{array}{ccccc}
\sum_{i=1}^{I} \sum_{j=1}^{N_i} (x_{ij} - \overline{x})^2 & = & \sum_{i=1}^{I} N_i(\overline{x}_i - \overline{x})^2 & + & \sum_{i=1}^{I} \sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)^2 \\
\text{SST} & & \text{SSG} & & \text{SSE}
\end{array}
$$

Total **S**um of **S**quares = **G**roup **S**um of **S**quares + **E**rror **S**um of **S**quares

---

$$
\begin{array}{ccccc}
(n - 1) & = & (I - 1) & + & (n - I) \\
\text{DFT} & & \text{DFG} & & \text{DFE}
\end{array}
$$

Total **D**egrees of **F**reedom = **G**roup **D**egrees of **F**reedom + **E**rror **D**egrees of **F**reedom

Variances are mean squared differences to the mean

Note that

SST/DFT: $\frac{\sum_{i=1}^{I}\sum_{j=1}^{N_i}(x_{ij}-\overline{x})^2}{n-1}$ is a variance, and likewise

SSG/DFG: labelled **MSG** ("Mean square between groups"), and

SSE/DFE: labelled **MSE** ("Mean square error" or sometimes "Mean square within groups")

In ANOVA, we compare MSG (variance between groups) and MSE (variance within groups), i.e. we measure

$$F = \frac{\text{MSG}}{\text{MSE}}$$

If this F-value is large, differences between groups overshadow differences within groups.

Two variances

1) Estimate the pooled variance of the population (MSE):

$$\text{MSE} = \frac{\text{SSE}}{\text{DFE}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{N_i}(x_{ij}-\overline{x}_i)^2}{n-I} \overset{\text{equiv}}{=} \frac{\sum_{i=1}^{I}\text{DF}_i \cdot s_i^2}{\sum_{i=1}^{I}\text{DF}_i}$$

In our example (Nederlands for anderstaligen):

$$\begin{aligned}
\frac{\sum_{i=1}^{I}\text{DF}_i \cdot s_i^2}{\sum_{i=1}^{I}\text{DF}_i} &= \frac{(N_1-1)s_1^2 + (N_3-1)s_2^2 + (N_3-1)s_3^2 + (N_4-1)s_4^2}{(N_1-1)+(N_3-1)+(N_3-1)+(N_4-1)} \\
&= \frac{9 \cdot 66.22 + 9 \cdot 43.66 + 9 \cdot 34.99 + 9 \cdot 47.57}{9+9+9+9} \\
&= \frac{595.98 + 392.94 + 314.91 + 428.13}{36} = 48.11
\end{aligned}$$

Estimates the variance in groups (using DF), aka **within-groups estimate** of variance

Two variances

2) Estimate the **between-groups** variance of the population (MSG):

$$\text{MSG} = \frac{\text{SSG}}{\text{DFG}} = \frac{\sum_{i=1}^{I} N_i(\overline{x}_i - \overline{x})^2}{I-1}$$

In our example (Nederlands for anderstaligen):

We had 4 group means: 25.0, 21.9, 23.1, 21.3, grand mean: 22.8

$$\text{MSG} = \frac{10 \cdot ((25-22.8)^2 + (21.9-22.8)^2 + (23.1-22.8)^2 + (21.3-22.8)^2)}{4-1} = 26.6$$

The **between-groups** variance (MSG) is an aggregate estimate of the degree to which the four sample means differ from one another

Interpreting estimates with F-scores

If H_0 is true, then we have two variances:

- Between-groups estimate: $s^2_{\text{bg}} = 26.62$ and
- Within-groups estimate: $s^2_{\text{wg}} = 48.11$

and their ratio $\frac{s^2_{\text{bg}}}{s^2_{\text{wg}}}$ follows an F-distribution with:

$(\#\text{ groups} - 1) = 3$ degrees of freedom for s^2_{bg} and
$(\#\text{ observations} - \#\text{ groups}) = 36$ degrees of freedom for s^2_{wg}

In our example: $F(3, 36) = \frac{26.62}{48.11} = 0.55$

$\mathrm{P}(F(3, 40) > 2.84) = 0.05$ (see tables), so there is no evidence of non-uniform behavior

SPSS summary

```
- - - - -  O N E W A Y  - - - - -

    Variable  NL_NIVO    toets nl. voor anderstalige
 By Variable  GROUP      gebied van afkomst

                            Analysis of Variance

                            Sum of      Mean        F       F
      Source      D.F.     Squares    Squares    Ratio   Prob.

Between Groups      3        79.9       26.6      .55     .65
Within Groups      36      1731.9       48.1
Total              39      1811.8
```

No evidence of non-uniform behavior

Other questions

ANOVA H_0: $\mu_1 = \mu_2 = \ldots = \mu_n$

But sometimes particular **contrasts** are important—e.g., are Europeans better (in learning Dutch)?

Distinguish (in reporting results):

- **prior** contrasts
  questions asked before data is collected and analyzed
- **post hoc** (posterior) questions
  questions asked **after** data collection and analysis
  "data-snooping" is exploratory, cannot contribute to hypothesis testing

Prior contrasts

Questions asked **before** data collection and analysis—e.g., are Europeans better (in learning Dutch)?

Another way of putting this:

$$
\begin{aligned}
H_0: \quad \mu_{\text{Eur}} &= \frac{1}{3}(\mu_{\text{Am}} + \mu_{\text{Afr}} + \mu_{\text{Asia}}) \\
H_a: \quad \mu_{\text{Eur}} &\neq \frac{1}{3}(\mu_{\text{Am}} + \mu_{\text{Afr}} + \mu_{\text{Asia}})
\end{aligned}
$$

Reformulation (SPSS requires this):

$$
H_0: \quad 0 = -\mu_{\text{Eur}} + 0.33\mu_{\text{Am}} + 0.33\mu_{\text{Afr}} + 0.33\mu_{\text{Asia}}
$$

Prior contrasts in SPSS

- Mean of every group gets a coefficient
- Sum of coefficients is 0
- A *t*-test is carried out and two-tailed *p*-value is reported (as usual):

		Eur	Am.	Afr.	Azie
Contrast	1	-1.0	.3	.3	.3

			Pooled Variance Estimate			
		Value	S. Error	T Value	D.F.	T Prob.
Contrast	1	-2.9	2.53	-1.15	36	.260

No significant difference here (of course)

Note: prior contrasts are legitimate as hypothesis tests as long as they are formulated **before** data collection and analysis

Post-hoc questions

Assume H_0 is rejected: which means are distinct?

Data-snooping problem: in a large set, **some** distinctions are **likely** to be statistically significant

But we can still look (we just cannot claim to have **tested** the hypothesis)

We are asking whether $m_i - m_j$ is significantly larger, we apply a variant of the t-test

The relevant sd is $\sqrt{\frac{\text{MSE}}{n}}$ (differences among scores), but there is a correction since we're looking at a proportion of the scores in any one comparison

SD in post-hoc ANOVA questions

Standard deviation (among differences in groups i and j):

$$\text{sd} = \sqrt{\text{MSE} \times \frac{N_i + N_j}{N}} = \sqrt{48.1 \times \frac{10+10}{40}} = 4.9$$

$$t = \frac{\overline{x}_i - \overline{x}_j}{\text{sd} \cdot \sqrt{\frac{1}{N_i} + \frac{1}{N_j}}}$$

The critical t-value is calculated as $\frac{p}{c}$ where p is the desired significance level and c is the number of comparisons.

For pairwise comparisons: $c = \binom{I}{2}$

Post-hoc questions in SPSS

SPSS post-hoc 'Bonferroni'-searches among **all** groupings for statistically significant ones

```
                    - - - - -  O N E W A Y  - - - - -

      Variable  NL_NIVO    toets nl. voor anderstalige
   By Variable  GROUP      gebied van afkomst

Multiple Range Tests:  Modified LSD (Bonferroni) test w. signif. level .05

The difference between two means is significant if
      MEAN(J)-MEAN(I)  >= 4.9045 * RANGE * SQRT(1/N(I) + 1/N(J))
      with the following value(s) for RANGE: 3.95
- No two groups significantly different at .05 level
 Homogeneous Subsets (highest \& lowest means not sig. diff.)

Group      Azie      America    Africa     Europa
Mean       21.3       21.9       23.1       25.0
```

But in this case there are none (of course)

How to win at ANOVA

Note the ways in which the F-ratio increases (i.e., becomes more significant):

$$F = \frac{\text{MSG}}{\text{MSE}}$$

1. MSG increases: differences in means between groups grow larger
2. MSE decreases: overall variation within groups grows smaller

Two models for grouped data

$$x_{ij} = \mu + \epsilon_{ij}$$
$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

First model:

- no group effect
- each data point represents error (ϵ) around a mean (μ)

Second model:

- real group effect
- each data point represents error (ϵ) around an overall mean (μ), combined with a group adjustment (α_i)

ANOVA asks: is there sufficient evidence for α_i?

Fall backs?

Suppose some cells are non-normal, or some standard deviations too large

Fall backs?

Suppose some cells are non-normal, or some standard deviations too large

- Kruskal-Wallis, non-parametric comparison of > 2 medians;
- apply (monotonic) transformation to reduce SD, perhaps improve fit to normality;
- trim most extreme 1% (or 5%) of data

Always report transformations or "trimming"!

Next week

Next week: factorial ANOVA