Analysis of Variance

ANOVA — **AN**alysis **O**f **VA**riance

- “generalized t-test”
- compares **means** of more than two groups
- fairly robust
- based on F distribution, compares **variance**
- two versions
  - single ANOVA
    compare groups along 1 dim., e.g. school classes
  - multiple ANOVA
    compare groups along > 1 dim., e.g. school classes and sex

RuG

Analysis of Variance

Typical applications

- single ANOVA
  compare time needed for lexical recognition in healthy adults, patients with Wernicke’s aphasia, patients with Broca’s aphasia
- multiple ANOVA
  compare lexical recognition time in male and female in same three groups

RuG

Comparing Multiple Means

for **two** groups: **t-test**

testing at $p = 0.05$ 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., $\binom{7}{2} = 21$ pairs

caution: several tests (on same data) run the risk of finding significance through sheer chance

Phony Significance through Multiple Tests

Example: Suppose you run three tests, always seeking a result significant at 0.05. The chance of finding this in one of the three is Bonferroni's **family-wise α-level**

$$\begin{aligned} \alpha_{FW} &= 1 - (1-\alpha)^n \\ &= 1 - (1 - .05)^3 \\ &= 1 - (.95)^3 \\ &= 1 - .857 \\ &= 0.143 \end{aligned}$$

to guarantee a family-wise alpha of 0.05, divide this by number of tests

Example: $0.05/3 = 0.017$ (set α at 0.017) —note: $0.983^3 \approx 0.95$

ANOVA indicated, takes group effects into account

Analysis of Variance

based on F distribution

F distribution —Moore & McCabe, § 7.3, pp.435-445

measures difference between two **variances** (variance $= \sigma^2$)

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

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

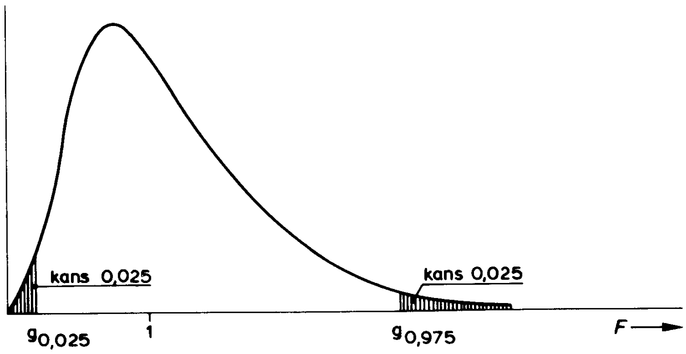
F-Test vs. F Distribution

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

- used in F-test
  H_0: samples from same distribution ($s_1 = s_2$)
  H_a: samples from diff. distribution ($s_1 \neq s_2$)
  - value 1 indicates same variance
  - values near 0 or $+\infty$ indicate diff.
- 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 $p = 0.05$



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

F-test$*$

Example: height

group	sample size	mean	std. dev.
boys	16	180cm	6cm
girls	9	168	4

is the *difference* in std. dev. significant? ($\alpha = 0.05$)

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

degrees of freedom: s_{boys} $16 - 1$
s_{girls} $9 - 1$

F-test Critical Area (for 2-Tailed Test)$*$

$$
\begin{aligned}
P(F(15,8) > f) &= \tfrac{\alpha}{2} = \tfrac{0.05}{2} = 0.025 \\
P(F(15,8) < f) &= 1 - 0.025 \\
P(F(15,8) < \underline{4.1}) &= 0.975 \text{ from M\&M, Tbl. E, p.706} \\
& \text{(no values directly for } P(F(\mathrm{df}_1, \mathrm{df}_2) > f))
\end{aligned}
$$

$$
\begin{aligned}
P(F(15,8) < x) &= \tfrac{\alpha}{2}(= 0.025) \\
&= P(F(8,15) > x') = \tfrac{\alpha}{2} | x' = \tfrac{1}{x} \\
&= P(F(8,15) > x') = 0.025 | x' = \tfrac{1}{x} \\
&= P(F(8,15) > \underline{3.2}) \text{(tables)} \\
P(F(15,8) < \tfrac{1}{3.2}) &= 0.025 \\
P(F(15,8) < \underline{0.31}) &= 0.025
\end{aligned}
$$

Reject H_0 if $F < 0.31$ or $F > 4.1$
Here, $F = \frac{6^2}{4^2} = 2.25$ (no evidence of diff. in distribution)

ANOVA

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

- several types
  - single “one-way”
  - multiple, i.e., “two-, three-, ...n-way”
- examines variation
  - “between-groups” —sex, age,...
  - “within-subject”, “within-groups” —overall
- automatically corrects for looking at several relationships (like Bonferroni correction)
- uses F test, where $F(n, m)$ fixes n typically at number of groups (less 1), m at number of subjects (data points) (less number of groups)

One-Way ANOVA to Analyse Function of Reviews

Example: Gisela Redeker identified three roles for literary book reviews in newspapers *Taalbeheersing* 21(4), 1999, 295-310:

- communicate emotional reactions, subjective opinions
- communicate expert opinion, objective facts
- motivate reading and purchasing of book

She investigated whether different reviewers emphasized different roles: Tom van Deel (*Trouw*), Arnold Heumakers (*de Volkskrant*), and Carel Peeters (*Vrij Nederland*)

stylistic elements indicate one of the three functions, e.g., *ik, maar nee, ben ik bang, lijkt, eerlijk gezegd, ik bedoel,...* indicate **subjective opinions**; logical connectives *want, temeer dat, ...* and quotes indicate an **objective** point of view; etc.

N.b. **validating** link between style and perspectives is important (see Redeker)

Redeker on Literary Criticism's Functions

Gisela Redeker investigates role of lit. criticism, asking whether different critics did not differ in the degree to which they emphasize one or another role.

Sample: reviews of the same books (by Hermans, Heijne and Mulisch), all published 1989-92. Similar in length.

Data: relative frequency of, e.g., **reader-oriented elements** (per $1,000$ words). We are comparing three averages, asking whether their is a difference.

Because she compared more than two averages ANOVA is needed.

Relative Frequency of Reader-Oriented Elements

elements	Critic		
	van Deel	Heumakers	Peeters
evocative	12.4	10.8	15.1
questions	3.2	0	6.1
dram.quotes	6.9	12.9	8.2
intensifiers	25.9	30.1	38.2
ref. to reader	3.6	5.7	11.7
Totals	26.0	29.8	39.7

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

Results: statistically significant difference ($p < 0.02$)

Similar comparisons for “subjective” style, and “argumentative” style (differences present, not statistically significant)

One-Way ANOVA

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

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

i.e., alternative: at least one group has different mean

for the question of whether any particular pair is the same, the t-test is appropriate

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

we want to apply 1-way ANOVA

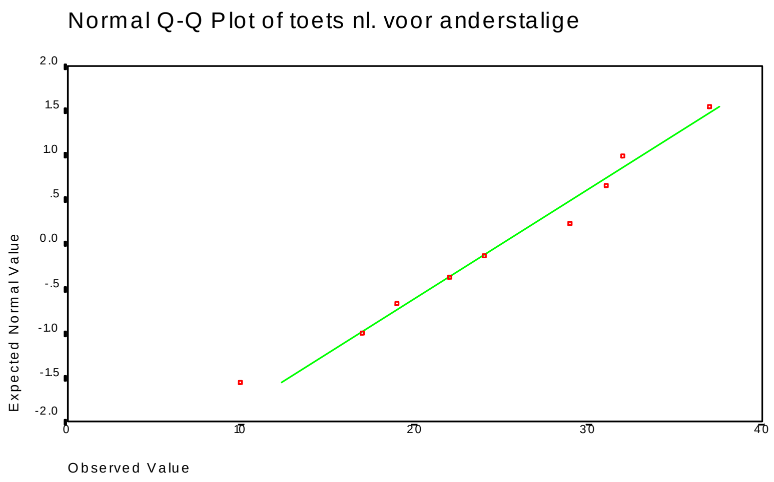
Data: Dutch Proficiency of Foreigners

Four groups of ten each:

	Groups			
	Eur.	*Amer.*	*Afri.*	*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

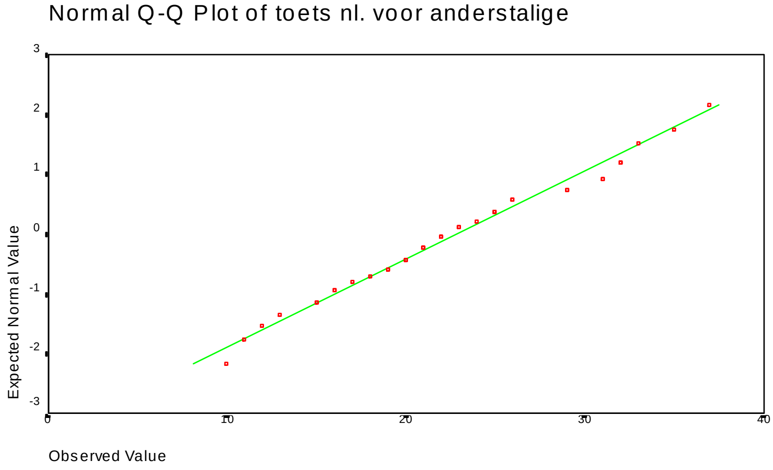
Anova Conditions

Assumption: normal distribution per group, check with normal quantile plot, e.g., for Europeans (and to be repeated for every group):



Anova Conditions

Normal Quantile plot for all values:



Inf. Stats

Anova Conditions

ANOVA assumptions:

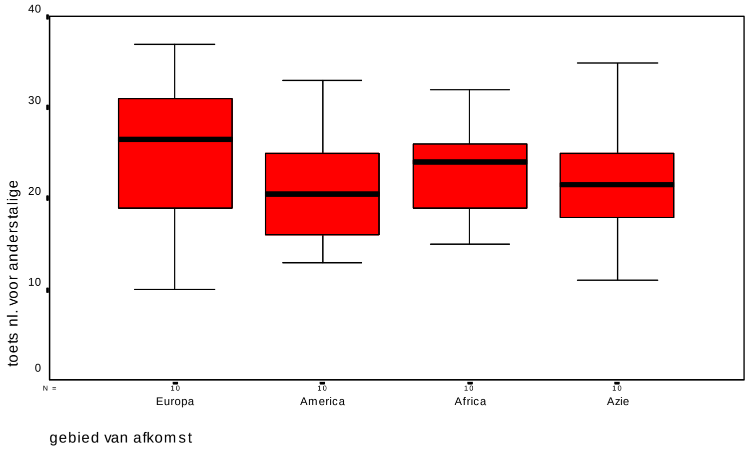
- normal distribution per subgroup
- same variance in subgroups: least sd $>$ one-half of greatest sd
- **independent** observations: watch out for test-retest situations!

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	

Visualizing Anova

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



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

Sketch of Anova

	Groups			
	1	2	3	$4 = I$
	Eur.	Amer.	Afri.	Asia
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$x_{1,j}$	$x_{2,j}$	$x_{3,j}$	$x_{4,j}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Sample Mean	$\bar{x}_1$	$\cdots$	$\bar{x}_i$	$\cdots$

I – number of groups

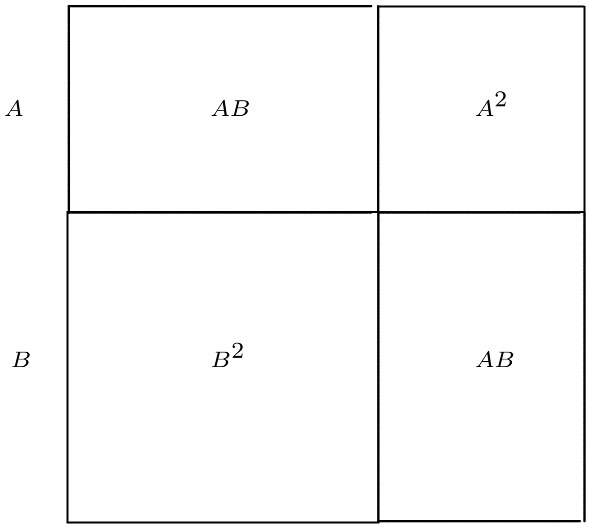
For any data point $x_{i,j}$

$$\begin{array}{ccccc} (x_{i,j} - \bar{x}) & = & (\bar{x}_i - \bar{x}) & + & (x_{i,j} - \bar{x}_i) \\ \text{total residue} & = & \text{group diff} & + & \text{“error”} \end{array}$$

ANOVA question: is it sensible to include the group ($\bar{x}_i$)?

Two Variances$*$

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



Two Variances$*$

$$
\begin{aligned}
(a+b)^2 &= a^2 + b^2 + 2ab \\
(x_{i,j} - \bar{x}) &= (\bar{x}_i - \bar{x}) + (x_{i,j} - \bar{x}_i) \\
(x_{i,j} - \bar{x})^2 &= (\bar{x}_i - \bar{x})^2 + (x_{i,j} - \bar{x}_i)^2 + 2(\bar{x}_i - \bar{x})(x_{i,j} - \bar{x}_i)
\end{aligned}
$$

Sum over elements in i-th group:

$$
\begin{aligned}
\sum_{j=1}^{N_i} (x_{i,j} - \bar{x})^2 &= \sum_{j=1}^{N_i} (\bar{x}_i - \bar{x})^2 + \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)^2 \\
&\quad + \sum_{j=1}^{N_i} 2(\bar{x}_i - \bar{x})(x_{i,j} - \bar{x}_i)
\end{aligned}
$$

Two Variances*

Note that this term must be zero:

$$\sum_{j=1}^{N_i} 2(\bar{x}_i - \bar{x})(x_{i,j} - \bar{x}_i)$$

Since:

$$\sum_{j=1}^{N_i} 2(\bar{x}_i - \bar{x})(x_{i,j} - \bar{x}_i) = 2(\bar{x}_i - \bar{x})\sum_{j=1}^{N_i}(x_{i,j} - \bar{x}_i) \text{ and}$$

$$\sum_{j=1}^{N_i}(x_{i,j} - \bar{x}_i) = 0$$

Sketch of Anova

$$\begin{aligned}\sum_{j=1}^{N_i}(x_{i,j} - \bar{x})^2 &= \sum_{j=1}^{N_i}(\bar{x}_i - \bar{x})^2 + \sum_{j=1}^{N_i}(x_{i,j} - \bar{x}_i)^2 \\ &\qquad (+ \sum_{j=1}^{N_i} 2(\bar{x}_i - \bar{x})(x_{i,j} - \bar{x}_i) \quad = 0)\end{aligned}$$

Therefore:

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

Anova Terminology∗

For any data point $x_{i,j}$

$(x_{i,j} - \bar{x})$	$=$	$(\bar{x}_i - \bar{x})$	$+$	$(x_{i,j} - \bar{x}_i)$
total residue	$=$	group diff	$+$	“error”
$(x_{i,j} - \bar{x})^2$	$=$	$(\bar{x}_i - \bar{x})^2$	$+$	$(x_{i,j} - \bar{x}_i)^2$
SST		SSG	$+$	SSE
Total **S**um of **S**quares	$=$	**G**roup **S**um of **S**quares	$+$	**E**rror **S**um of **S**quares
and				
DFT		DFG	$+$	DFE
$(n - 1)$	$=$	$(I - 1)$	$+$	$(n - I)$
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 Mean

Note that

$\frac{(x_{i,j}-\bar{x})^2}{n-1}$ is a variance, and likewise

SST/DFT SSG/DFG (=MSG) & SSE/DFE (=MSE)

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

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

If this is **large**, differences between groups overshadow differences within groups

Two Variances∗

$H_0: \ \mu_1 = \mu_2 = \mu_3 = \mu_4$

ANOVA: calculate MSG (σ^2 between groups) and MSE (σ^2 within groups), i.e. we measure

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

If this is **large**, differences between groups overshadow differences within groups

Two Variances∗

1. estimate **pooled variance** of population (MSE)

$$\begin{aligned} \frac{\sum_{i \in G} dF_i \cdot s_i^2}{\sum_{i \in G} dF_i} &= \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+(n_3-1)s_3^2+(n_4-1)s_4^2}{(n_1-1)+(n_2-1)+(n_3-1)+(n_4-1)} \\ &= \frac{9 \cdot 66.22+9 \cdot 43.66+9 \cdot 34.99+9 \cdot 47.57}{9+9+9+9} \\ &= \frac{595.98+392.94+314.91+428.13}{36} = 48.11 \end{aligned}$$

   estimates variance in groups (using dF), aka **within-groups estimate** of variance
2. suppose H_0 true
   (a) then group have *sample means* μ, variance $\sigma^2/10$, (& sd $\sigma/\sqrt{10}$)
   (b) 4 means, $25.0, 21.9, 23.1, 21.3$, where $s = 1.63$, $s^2 = 2.66$
   (c) s^2 estimate of $\sigma^2/10$, i.e., $10 \times s^2$ is estimate of $\sigma^2 (\approx \ s^2 = 26.6)$
   (d) this is **between-groups** variance (MSG)

Interpreting Estimates via F

if H_0 true, then we have two variances:

- between-groups estimate s_b^2 (26.6) *and*
- within-groups estimate s_w^2 (48.11)

and their ratio $\frac{s_b^2}{s_w^2}$ follows an F distribution with
$(|\text{groups}| - 1)dF$ from s_b^2 (3),
$(n-4)dF$ from s_w^2 (36)

in this case, $\frac{26.62}{48.11} = 0.55$

$P(F(3,30) > 2.92) = 0.05$, (see tables), so no evidence of nonuniform behavior

ANOVA Summary

Summary to-date (exam results for *NL voor anderstalige*)

Source	dF	SS	MSS	F
between-g	3	79.9	26.6	F(3,36) = 0.55
within-g	36	1731.9	48.1	
Total	39	1811.8		

$P(F(3,30) > \underline{2.92}) = 0.05$, (see tables)

so no evidence of nonuniform 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
```


Other Questions

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

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

Distinguish (in reporting results):

- **prior** contrasts
  questions asked **before** data collected and analyzed

- **post-hoc** (posterior) questions
  questions **after** collection and analysis
  “data-snooping” is exploratory, cannot contribute to hypothesis testing

Prior Contrasts

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

Another formulation:

$$\text{is } H_a : \quad \mu_{\text{Eur}} \neq (\mu_{\text{Am}} = \mu_{\text{Afr}} = \mu_{\text{Asia}})$$

$$\text{where } H_0 : \mu_{\text{Eur}} = (\mu_{\text{Am}} + \mu_{\text{Afr}} + \mu_{\text{Asia}})$$

Reformulation:

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

SPSS requires this (reformulated) version

Prior Contrasts in SPSS

- (the mean of) every group gets a coefficient
- sum of coefficients is zero
- a t-test is carried out, & 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** collection and analysis

Post-hoc Questions

Assume H_0 rejected: which means are distinct?

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

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

We are asking whether $m_1 - m_2$ is significantly larger, we apply a variant of the t-test

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

SD in Post-hoc ANOVA Questions

N.B. SD (among diff. in groups i and j):

$$\text{sd}_\delta = \sqrt{\frac{\text{MSE} \times \frac{N_i+N_j}{N}}{N_i + N_j}} = \sqrt{\frac{48.1 \times \frac{10+10}{40}}{10+10}} = \sqrt{\frac{\frac{48.1}{2}}{20}} = \sqrt{\frac{24.05}{20}} = 4.9/\sqrt{20}$$

and the t value is calculated as p/c where p is the desired significance level and c is 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 ways in which F ratio increases (becomes more significant)

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

1. MSG increases, differences in means grow larger
2. MSE decreases, overall variation grows smaller

Two Models for Grouped Data

$$
\begin{aligned}
x_{i,j} &= \mu + \epsilon_{i,j} \\
x_{i,j} &= \mu + \alpha_i + \epsilon_{i,j}
\end{aligned}
$$

First model

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

Second model

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

ANOVA: is there sufficient evidence for α_i?

Next: Two-way ANOVA