Statistics 2

Analysis of Variance (ANOVA)

Casper Albers & Jorge Tendeiro

Lecture 10, 2019 – 2020

university of groningen

Overview

Literature for this lecture

Read:

Agresti, Section 12.3

Comparing two means

Regression with code variables (recall Lecture 8).

Two groups:

- $\text{Group}_1 \sim \mathcal{N}(\mu_1, \sigma)$
- $\text{Group}_2 \sim \mathcal{N}(\mu_2, \sigma)$
- Same σ assumed
- Sample size n_1 and n_2
- $\mathcal{H}_0 : \mu_1 = \mu_2$

Coding: 0s for Group 1; 1s for Group 2.
Thus:

$$\mu_1 = \beta_0$$

$$\mu_2 = \beta_0 + \beta_1$$

This implies $\beta_1 = \mu_2 - \mu_1$.

Comparing two means

- $\beta_1 = \mu_2 - \mu_1$.
- $\mathcal{H}_0 : \mu_1 = \mu_2$ is equivalent to $\mathcal{H}_0 : \beta_1 = 0$.

Testing $\mathcal{H}_0$ is done through

$$t = \frac{\bar{y}_2 - \bar{y}_1}{SE_{b_1}}$$

with $n_1 + n_2 - 2$ degrees of freedom.

Conceptually:

$$t = \frac{\text{Distance between groups}}{\text{Variability within groups}}$$

Comparing two means

Another approach to compare two means: The t-test.

Test statistic:

$$t = \frac{\overline{y}_2 - \overline{y}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Conceptually:

$$t = \frac{\text{Distance between groups}}{\text{Variability within groups}}$$

$$SE_{b_1} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

so both t-tests are the same.

$\boxed{\text{$t$-test and regression with a dummy variable are the same!}}$

Comparing two means

Example

- Data from Moore, McCabe, & Craig.
- Reading performance in two groups of pupils:
  One with and one without 'directed reading activities'.
- Sample sizes $n_1 = 21$, $n_2 = 23$.

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	51.476	3.175		16.211	$< .001$
Group	-9.954	4.392	-0.330	-2.267	0.029

	t	df	p
Group	2.267	42.00	0.029

Apart from sign issues, due to irrelevant coding choices, both approaches are mathematically equivalent.

Comparing means: More than two groups

What to do when more than two groups need to be compared?

Again, there are two approaches:

Regression with multiple dummy variables

ANOVA: ANalysis Of VAriance

Approaches mathematically equivalent.

Both are common within social sciences, thus important to be able to work with both.

ANOVA

- **Principle:**
  Study differences in the means of g independent groups.

- **Test:**

$$\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_g.$$

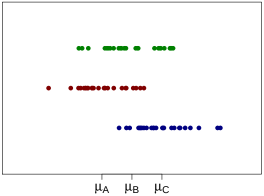
versus

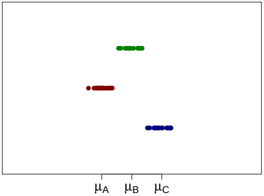
$$\mathcal{H}_a : \text{ Not all } \mu\text{'s are equal.}$$

- **Procedure:**
  Compare the between and the within group variances using the F-test.

ANOVA

Why call it Analysis of Variance when comparing means?





Distance between groups: relative to distance within groups.

$$F = \frac{\text{Variability between groups}}{\text{Variability within groups}}$$

The t-test is a special kind of ANOVA:

- Two-sample t test:

$$t = \frac{\overline{y}_2 - \overline{y}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

- Look at t^2 (and assume $n_1 = n_2 = n$ for simplicity):

$$t^2 = \frac{\frac{n}{2}(\overline{y}_1 - \overline{y}_2)^2}{s_p^2} = F \sim F(1, n_1 + n_2 - 2)$$

variation between means

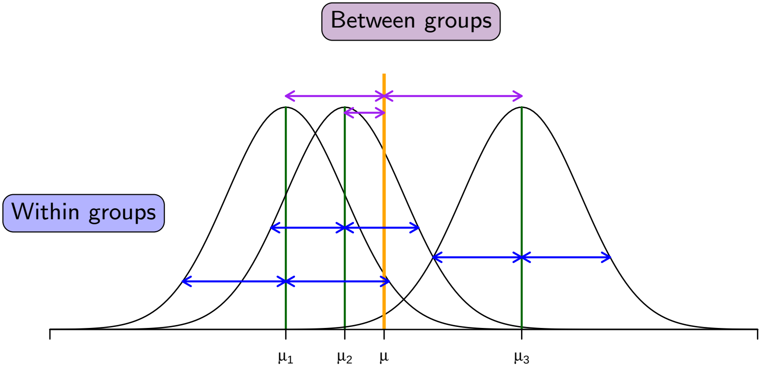
estimator of the variance between groups

pooled variance of both groups

estimator of σ^2, the variance within groups

ANOVA: General principle

Split total variance: **Between** groups and **within** groups



ANOVA: Partitioning the variance

Split the total variance in two parts:

- A part that can be explained by differences between groups.
- A part that remains unexplained within groups.

$$
\begin{array}{ccccc}
(y_{ij} - \overline{y}) & = & (\overline{y}_i - \overline{y}) & + & (y_{ij} - \overline{y}_i) \\
\sum_{ij}(y_{ij} - \overline{y})^2 & = & \sum_{ij}(\overline{y}_i - \overline{y})^2 & + & \sum_{ij}(y_{ij} - \overline{y}_i)^2 \\
\downarrow & & \downarrow & & \downarrow \\
\boxed{\text{Total SS}} & & \boxed{\text{Group SS}} & & \boxed{\text{Residual SS}}
\end{array}
$$

$$\boxed{\text{TSS} = \text{GSS} + \text{RSS}}$$

SS = sum of squares

ANOVA: Notation

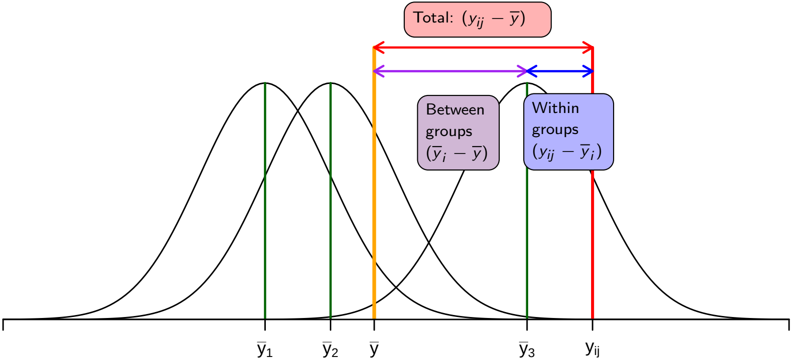
$$\sum_{ij}(y_{ij} - \overline{y})^2 = \sum_{ij}(\overline{y}_i - \overline{y})^2 + \sum_{ij}(y_{ij} - \overline{y}_i)^2$$

- i indexes the groups.
- j indexes the persons in a group.

- $y_{ij} =$ observation of person j in group i.
- $\overline{y}_i =$ mean of the DV y in group i.
  (i.e., over all persons in group i)
- $\overline{y} =$ overall, or grand, mean of y.
  (i.e., over all persons in all groups)

ANOVA: Partitioning the variance

Variance: $\frac{1}{n-1}\sum_{ij}(y_{ij} - \overline{y})^2$



ANOVA: Partitioning the variance

With g groups:

$$\underbrace{\sum_{ij}(y_{ij} - \overline{y})^2}_{\text{TSS}} = \underbrace{\sum_{ij}(\overline{y}_i - \overline{y})^2}_{\text{GSS}} + \underbrace{\sum_{ij}(y_{ij} - \overline{y}_i)^2}_{\text{RSS}}$$

Convert SS's in variances: Divide by degrees of freedom (df)

Mean Squares (MS)

	Total	Group	Residual
SS	TSS	GSS	RSS
df	$\text{df} = n - 1$	$\text{df}_1 = g - 1$	$\text{df}_2 = n - g$
MS	TMS=TSS/df	GMS=GSS/df_1	RMS=RSS/df_2

Variance in y!

s_p^2 = pooled variance!

ANOVA: F-test

- **Hypotheses:**

  $\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_g$.

  $\mathcal{H}_a$: Not all μ's are equal.

- **Test statistic:**

$$F = \frac{\text{GMS}}{\text{RMS}} = \frac{\text{GSS}/\text{df}_1}{\text{RSS}/\text{df}_2}.$$

- $\boxed{\text{If } \mathcal{H}_0 \text{ holds: } F \approx 1.}$

  Q: Why?

  A: Because:

  1. RMS $= s_p^2$ always estimates σ^2, the common group variance.
  2. Under $\mathcal{H}_0$, GMS also estimates σ^2.
  3. Hence, under $\mathcal{H}_0$, the F ratio is ≈ 1.

- $\boxed{\text{If } \mathcal{H}_0 \text{ does not hold: } F > 1.}$

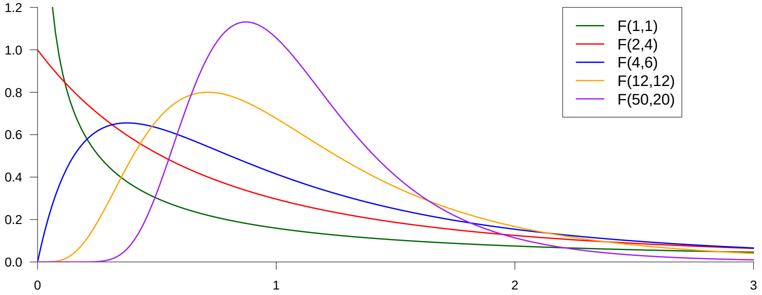
ANOVA: F-test

Conclusion: Reject $\mathcal{H}_0$ is F is too large (i.e., $F \gg 1$).

Q: But how large need statistic F be?

A: Use the sampling distribution.

$$F \sim F(g-1, n-g)$$



ANOVA: Table

Source	SS	df	MS	F
Group	$\sum_{ij}(\overline{y}_i - \overline{y})^2$	$g-1$	GSS/df$_1$	GMS/RMS
Residual	$\sum_{ij}(y_{ij} - \overline{y}_i)^2$	$n-g$	RSS/df$_2$	
Total	$\sum_{ij}(y_{ij} - \overline{y})^2$	$n-1$		

Example: Directed Reading Activities

Cases	Sum of Squares	df	Mean Square	F	p
group	1088	1	1087.8	5.137	0.029
Residual	8893	42	211.7		

Reject $\mathcal{H}_0 : \mu_1 = \mu_2$

Recall: t-test provided $t = 2.267$, $p = .029$.

Indeed, $\sqrt{5.137} = 2.267$: t and F-test equivalent.

Example with more than 2 groups

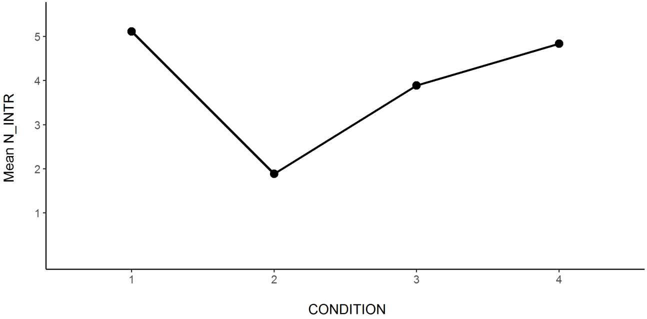
James et al. (2015) studied whether playing a computer game (Tetris) could prevent intrusive memories (flashbacks) related to a traumatic event from occurring, via a reactivation-reconsolidation mechanism[1].

- Dependent variable
  y = N_INTR = Number of intrusive memories over the next seven days
- Factor
  CONDITION, with four levels:
  - 1 = No-task control
  - 2 = Reactivation + Tetris
  - 3 = Tetris only
  - 4 = Reactivation only

[1] See Lecture 8

Example: Preventing flashbacks

	N_INTR			
	1	2	3	4
Valid	18	18	18	18
Mean	5.111	1.889	3.889	4.833
Std. Deviation	4.227	1.745	2.888	3.330



Question: Is there an effect of CONDITION on N_INTR?

In week 8, we studied this using code variables. Today: ANOVA

Example: Preventing flashbacks

Recall from Lecture 8: Regression with 3 code variables.

Regression output

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	4.833	0.749		6.457	$< .001$
z1	0.278	1.059	0.036	0.262	0.794
z2	-2.944	1.059	-0.382	-2.781	0.007
z3	-0.944	1.059	-0.123	-0.892	0.375

- These tests are indicative of specific contrasts
- Overall model fit assessed through R^2
- JASP provided $R^2 = .149$
- $F = (R^2/g)/((1 - R^2)/(n - g)) = 3.795$ with $p = .014$

Example: Preventing flashbacks

ANOVA-approach

$\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$

ANOVA table

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	114.8	3	38.27	3.795	0.014
	Residual	685.8	68	10.09		
	Total	800.7	71			

Conclusions:

- Reject $\mathcal{H}_0$.
- Both the F-test in ANOVA and in regression with code variables are equivalent (whatever the choice of dummy coding!).

The Kruskal-Wallis test

In ANOVA one makes three assumptions:

1. Independent observations.
2. Homogeneity: Variance in each group is equal.
3. Normality: Each group is normally distributed.

In Lecture 7 we learned how to check these assumptions and the consequences of violations.

Non-parametric alternative to ANOVA: The Kruskal-Wallis test.
Does not assume normality nor homogeneity.
You don't need to know technical details of KW, just be able to work with it.

$\mathcal{H}_0$: The distribution of observations in each group is identical.
$\mathcal{H}_1$: The distribution of observations in each group is not identical.

Factor	Statistic	df	p
Condition	13.56	3	0.004

$p = .004$ thus reject H_0. Significant differences between groups.

ANOVA: Computing CIs for a group's mean

This is not in the textbook but it is important!!

Two ways to compute CIs for group means:

- Based on the pooled SD, s_p.

  (ideal when homoscedasticity is met)

$$\text{CI for group } i = \bar{y}_i \pm t^*_{n-g} \frac{s_p}{\sqrt{n_i}}$$

$n =$ total sample size

- Based on the groups SD, s_i.

  (when homoscedasticity is violated)

$$\text{CI for group } i = \bar{y}_i \pm t^*_{n_i-1} \frac{s_i}{\sqrt{n_i}}$$

$n_i =$ group sample size

For the next lecture

Contents:

- Analysis of Variance (ANOVA):
  Two-way ANOVA

Read: Section 12.4.