

## 10. Experiments

AEA 2026

One-way ANOVA

Post-hoc analysis

# Analysis of Variance (ANOVA)

*Analysis of Variance (ANOVA)* explores variation among the observed responses.

- ▶ Used to test the claim that **three or more population means are equal**.
- ▶ An extension of the two independent samples t-test.

*One-way (one-factor) ANOVA*: a quantitative outcome and a single categorical explanatory variable with any number of levels

*Two-way (two-factor) ANOVA*: for studying the effects of two factors at the same time

## ANOVA: examples

a. Students from different colleges take the same exam. You want to see if one college outperforms the other.

- ▶ dependent variable: the performance at the exam
- ▶ independent variable: the colleges

b. Study if an alcoholic support group and individual counseling combined is the most effective treatment for lowering alcohol consumption.

- ▶ dependent variable: #alcoholic beverages consumed per day
- ▶ independent variable: groups of participants (medication only, medication and counseling, counseling only)

## ANOVA: examples

c. An experiment is conducted in which a crop yield is compared for 3 different levels of pesticide spray and 3 different levels of anti-fungal seed treatment. There are 4 replications of the experiment at each level combination.

Do the different levels of pesticide spray and anti-fungal treatment effect crop yield and is there a significant interaction?

# Multiple t-tests

Pairwise comparison

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_1 = \mu_3$$

$$H_0 : \mu_2 = \mu_3$$

$\alpha = 0.05$  Type I error rate at 95% confidence

The error COMPOUNDS with each t-test

$$(.95)(.95)(.95) = .857$$

$$\alpha = 1 - .857 = .143!$$

# One-way ANOVA

- ▶ A **categorical** explanatory variable (**independent**), with  $k$  levels
- ▶ A **quantitative** outcome (**dependent**)
- ▶ The outcomes for each group have **mean** parameters  $\mu_i$ ,  $i = 1, \dots, k$ . The **variances** for the outcome for each of the  $k$  groups are  $\sigma^2$ . For each group  $i$ : the distribution  $N(\mu_i, \sigma^2)$ .

# One-way ANOVA

Assumptions:

- ▶ the outcomes for each group are normally distributed with a common variance  $\sigma^2$
- ▶ the "errors" (deviations of individual outcomes from the population group means) are independent

The **null hypothesis**:  $H_0 : \mu_1 = \dots = \mu_k$  all of the population means are equal.

The **alternative hypothesis**:  $H_A : \exists i, j : \mu_i \neq \mu_j$  at least one of the  $k$  population means differs from all of the others.

## The F statistic (ratio)

- ▶ For a group  $i$ ,  $SS_i = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$  sum of squared deviations from the mean,  $df_i = n_i - 1$  degrees of freedom.

Mean squares  $MS = SS/df$ .

- ▶  $MS_{within}$  a good (unbiased) estimate of  $\sigma^2$  if it is defined as:

$$MS_{within} = SS_{within}/df_{within}$$

$$SS_{within} = \sum_{i=1}^k SS_i$$

$$df_{within} = \sum_{i=1}^k df_i = \sum_{i=1}^k (n_i - 1) = N - k$$

# Within-group deviations and between-group deviations

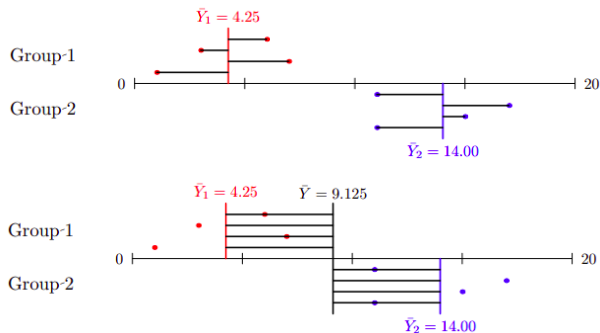


Figure: a) Deviations for **within-group** sum of squares b) Deviations for **between-group** sum of squares

## The F statistic (ratio)

$SS_{between}$  is the sum of the  $N$  squared between-group deviations, where the deviation is the same for all subjects in the same group.

$$SS_{between} = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{\bar{Y}})^2$$

$\bar{\bar{Y}}$  the **grand mean**.

$$df_{between} = k - 1.$$

$MS_{between}$  is a good estimate of  $\sigma^2$  only if  $H_0$  is true. Otherwise it tends to be larger.

# The F statistic ratio

Look at the ratio  $F = \frac{MS_{between}}{MS_{within}}$  to evaluate  $H_0$ .

The values of F-statistic tend to fall around 1.0 when  $H_0$  is true, and are *bigger* when  $H_A$  is true because:

- ▶ the numerator is either an estimate of  $\sigma^2$  (under  $H_0$ ) or is inflated (under  $H_A$ )
- ▶ the denominator is always an estimate of  $\sigma^2$ .

Large values of  $F$  argue for rejection of  $H_0$ .

# Null sampling distribution of the F statistic

The null sampling distribution of the F-statistic: the F-distribution with numerator degrees of freedom  $df_{between} = k - 1$  and denominator degrees of freedom  $df_{within} = N - k$ .

- ▶ We can quantify "large" for the F-statistic by comparing it to its null sampling distribution.
- ▶ The F-distribution is skewed to the right; have some tiny probability no matter how large F gets.

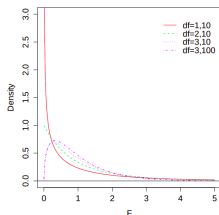


Figure: A variety of F-distribution pdfs

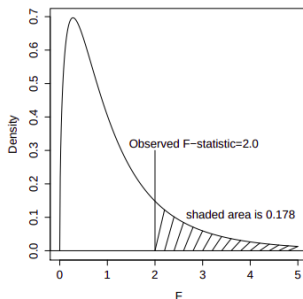
## Inference: hypothesis testing

- ▶ Calculate the observed F-statistic  $MS_{between}/MS_{within}$
  - ▶ Compare it to F-critical
    - ▶ if the statistic is smaller than the critical value, accept  $H_0$
    - ▶ if it is equal to or bigger than the critical value, reject  $H_0$
- OR
- ▶ Calculate the  $p$ -value (the area under the null sampling distribution of F that is bigger than the observed F-statistic).  
Reject the null hypothesis if  $p \leq \alpha$ ,  $\alpha$  the significance level.

# Inference: hypothesis testing

- a. The null sampling distribution of  $F$  can be used to calculate  $p$ -value.

Example: The  $F(3,10)$  pdf,  $F=2.0$

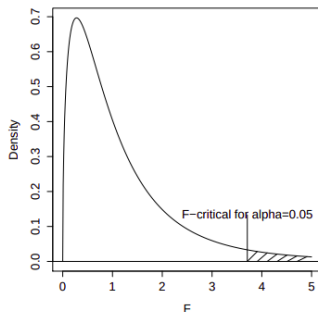


$p$ -value: the chance of getting an  $F$ -statistic  $\geq 2.0$  when  $H_0$  is true (the shaded area 0.178).

## Inference: hypothesis testing

b. The null sampling distribution of  $F$  can be used to find the **F-critical** value.

Example: The  $F(3,10)$  pdf,  $F=2.0$



**F-critical value** (3.71): the  $F$  value above which  $100\alpha\%$  of the null sampling distribution occurs, for a given significance level ( $\alpha = 0.05$ ).

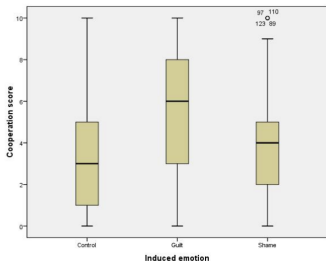
# Example 1: moral sentiments

Goal: investigate the effects of emotion on cooperation.

Explanatory variable: “induced emotion” (a nominal categorical variable with 3 levels: control, guilt, shame).

Outcome var: cooperation (#coins offered).

If we do see evidence that “cooperation” differs among the groups, we can claim that induced emotion *causes* different degrees of cooperation.



Induced emotion		Statistic	Std.Error
Cooperation score	Control	Mean	3.49
		95% Confidence Interval for Mean	2.48
	Lower Bound	4.50	
	Upper Bound	3.00	
	Median	3.11	
	Std. Deviation	0	
	Minimum	10	
	Maximum	0.57	
	Skewness	0.38	
	Kurtosis	-0.81	
Guilt	Guilt	Mean	5.38
		95% Confidence Interval for Mean	4.37
	Lower Bound	6.39	
	Upper Bound	6.00	
	Median	3.25	
	Std. Deviation	0	
	Minimum	10	
	Maximum	-0.19	
	Skewness	0.36	
	Kurtosis	-1.17	
Shame	Shame	Mean	3.78
		95% Confidence Interval for Mean	2.89
	Lower Bound	4.66	
	Upper Bound	4.00	
	Median	2.95	
	Std. Deviation	0	
	Minimum	10	
	Maximum	0.71	
	Skewness	0.35	
	Kurtosis	-0.20	

## Example 1: moral sentiments

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	86.35	2	43.18	4.5	0.013
Within Groups	1181.43	123	9.6		
Total	1267.78	125			

$p = 0.013 < 0.05 \rightarrow$  reject  $H_0$ .

Conclude that differences in mean cooperation are caused by the induced emotions, and that among control, guilt, and shame, at least two of the population means differ.

One-way ANOVA

Post-hoc analysis

# Post Anova Analysis

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

- ▶ if  $H_0$  is not rejected, no further analysis is required
- ▶ if  $H_0$  is rejected, further analysis is required

If  $k = 2$  and the  $H_0$  is rejected, look at the sample means.

# Contrasts and custom hypothesis

Contrast null hypothesis compares two population means or combinations of population means.

- ▶ **Simple contrast hypothesis** compares two population means, e.g.  $H_0 : \mu_1 = \mu_5$ . The alternative hypothesis:  $H_1 : \mu_1 \neq \mu_5$ .
- ▶ **Complex contrast hypothesis** has multiple population means on either or both sides of the equal sign.

$$\text{Ex: } \frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4 + \mu_5}{3}$$

# Tukey's method for pairwise comparison

- ▶ Also called *Honestly Significant Differences (HSD)* test
- ▶ Compare pairs of the sample means, using their absolute differences

Example 2:

	Type A	Type B	Type C
	77	83	80
	79	91	82
	87	94	86
	85	88	85
	78	85	80
Sample mean	81.2	88.2	82.6

$$|\bar{x}_1 - \bar{x}_2|, |\bar{x}_1 - \bar{x}_3|, |\bar{x}_2 - \bar{x}_3|$$

# Tukey criterion

The Tukey criterion is defined as

$$T = q_{\alpha(c, n-c)} \sqrt{\frac{MSE}{n_i}}$$

- ▶  $q_{\alpha(c, n-c)}$  the Studentized range distribution, based on  $c$  and  $n - c$  df
  - ▶  $c$  #groups (i.e. #columns)
  - ▶  $n$  total sample size
- ▶ MSE mean square error (from ANOVA table)
- ▶  $n_i$  sample size of the group with smallest no of observations

## Tukey criterion: Example 2

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups ( <i>Treatment</i> )	137.20	2	68.60	4.3145	0.0387	3.8853
Within Groups ( <i>Error</i> )	190.80	12	15.90			
Total	328.00	14				

Reject  $H_0$

$\alpha = 0.05$ . From the *Studentized Range Distribution* table ([http://davidmlane.com/hyperstat/sr\\_table.html](http://davidmlane.com/hyperstat/sr_table.html))

$$q_{\alpha(c, n-c)} = q_{0.05(3, 15-3)} = q_{0.05(3, 12)} = 3.773$$

Substituting and solving:

$$T = q_{\alpha(c, n-c)} \sqrt{\frac{MSE}{n_i}} = 3.773 \sqrt{\frac{15.9}{5}} = 6.73$$

## Tukey criterion: Example 2

Absolute values of the paired means are:

- ▶  $|\bar{x}_1 - \bar{x}_2| = |81.2 - 88.2| = 7.0 > 6.73$  {difference btw 1 and 2 is sig at  $\alpha=0.05$ }
- ▶  $|\bar{x}_1 - \bar{x}_3| = |81.2 - 82.6| = 1.4 < 6.73$  {difference btw 1 and 3 is not sig}
- ▶  $|\bar{x}_2 - \bar{x}_3| = |88.2 - 82.6| = 5.6 < 6.73$  {difference btw 2 and 3 is not sig}

# The Scheffe procedure

- ▶ It is one of the most popular *post-hoc* procedure
- ▶ Like the Tukey procedure, it compares two means at a time (pairwise comparison), but it has less statistical power

Example 3: Compare the average time to relief of headache pain under 3 medications (drug A, B, and C). 15 patients who suffer from headaches are randomly selected and 5 subjects are randomly assigned to each treatment. The times to relief after taking the assigned drug:

Drug A	Drug B	Drug C
30	25	15
35	20	20
40	30	25
25	25	20
35	30	20

## The Scheffe procedure: Example

Source of variation	SS	df	MS	F
between	423.329	2	211.66	10.1598
within	250	12	20	
total	673.329	14		

$F > F_{crit}$ , so there is evidence at 0.05 significance to conclude that at least two drugs differ.

	$n_1 = 5$	$n_2 = 5$	$n_3 = 5$
Summary statistics by treatment	$\bar{x}_1 = 33$	$\bar{x}_2 = 26$	$\bar{x}_3 = 20$
	$\bar{s}_1 = 5.7$	$\bar{s}_2 = 4.2$	$\bar{s}_3 = 3.5$

Which drugs were significantly different?

# The Scheffe procedure

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

$$F_0 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{MS_{within}(\frac{1}{n_1} + \frac{1}{n_2})}$$

$$F_0 = \frac{(33-26)^2}{20.833(1/5+1/5)} = 5.88$$

$$F_c = (k - 1)F_{0.05,2,12} = 2(3.89) = 7.78$$

$F_0 < F_c$  Not reject the null hypothesis

## The Scheffe procedure

$$H_0 : \mu_1 = \mu_3$$

$$H_a : \mu_1 \neq \mu_3$$

$$F_0 = \frac{(33-20)^2}{20.833(1/5+1/5)} = 20.28 > F_c \text{ Reject the null hypothesis}$$

$$H_0 : \mu_2 = \mu_3$$

$$H_a : \mu_2 \neq \mu_3$$

$$F_0 = \frac{(26-20)^2}{20.833(1/5+1/5)} = 4.32 < F_c \text{ Not reject the null hypothesis}$$