

# One way ANOVA

## Session 3

MATH 80667A: Experimental Design and Statistical Methods  
for Quantitative Research in Management  
HEC Montréal

# Outline

**Hypothesis tests for ANOVA**

**Parametrizations and interpretation**

**Planned comparisons and *post-hoc* tests**

# Hypothesis tests for ANOVA

# General recipe of hypothesis testing

**(1) Define variables**

**(2) Write down hypotheses**

**(3) Choose/compute a test statistic**

**(4) Compare statistic to null distribution**

**(5) Compute  $p$ -value / confidence interval**

**(6) Conclude (reject/fail to reject)**

Level = probability of condemning an innocent

**Fix level  $\alpha$  before the experiment.**

**Choose small  $\alpha$  (typical value is 5%)**

**Reject  $\mathcal{H}_0$  if p-value less than  $\alpha$**

# F-test for one way ANOVA

## Global null hypothesis

No difference between treatments

- $\mathcal{H}_0$  (null): all of the  $K$  treatment groups have the same average  $\mu$
- $\mathcal{H}_a$  (alternative): at least two treatments have different averages

# Building a statistic

Denote

- $y_{ik}$  is observation  $i$  of group  $k$
- $\hat{\mu}_1, \dots, \hat{\mu}_K$  the sample average of groups  $1, \dots, K$
- $\hat{\mu}$  is overall sample mean

## Decomposing variability into bits

$$\sum_i \sum_k (y_{ik} - \hat{\mu})^2 = \sum_i \sum_k (y_{ik} - \hat{\mu}_k)^2 + \sum_k n_i (\hat{\mu}_k - \hat{\mu})^2 .$$

total sum of squares                  within sum of squares                  between sum of squares

null model

alternative model

added variability

# Degrees of freedom

The parameters of the null distribution are called **degrees of freedom**

- $K - 1$  is the number of constraints imposed by the null
- $n - K$  is the number of observations minus number of mean parameters estimated under alternative



# F-test statistic

## Omnibus test

With  $K$  groups and  $n$  observations, the statistic is

$$F = \frac{\text{between sum of squares} / (K - 1)}{\text{within sum of squares} / (n - K)}$$

The null distribution (benchmark) is  $F(K - 1, n - K)$ .

# Intuition behind $F$ -test

Idea of  $F$ -statistic: under the null, both numerator and denominator are estimators of the variance.

- the  $F$  ratio should be approximately one on average
- the numerator is more variable (so skewed null distribution)...

# Pairwise differences and $t$ -tests

The pairwise differences ( $p$ -values) and confidence intervals for groups  $j$  and  $k$  are based on the  $t$ -statistic:

$$t = \frac{\text{estimated} - \text{postulated difference}}{\text{uncertainty}} = \frac{(\hat{\mu}_j - \hat{\mu}_k) - (\mu_j - \mu_k)}{\text{se}(\hat{\mu}_j - \hat{\mu}_k)}$$

which has a Student- $t$  null distribution, denoted  $\text{St}(n - k)$ .

The standard error  $\text{se}(\hat{\mu}_j - \hat{\mu}_k)$  uses the pooled variance estimate (based on all groups).

# $t$ -tests

If we postulate  $\delta_{jk} = \mu_j - \mu_k = 0$ , the test statistic becomes

$$t = \frac{\hat{\delta}_{jk} - 0}{\text{se}(\hat{\delta}_{jk})}$$

The  $p$ -value is  $p = 1 - \Pr(-|t| \leq T \leq |t|)$  for  $T \sim \text{St}_{n-k}$ .

- probability of statistic being more extreme than  $t$

The larger the values of  $t$  (positive or negative), the more evidence against the null hypothesis.

# Example

Consider the pairwise average difference in scores between the praised (group C) and the reprovved (group D) of the `arithmetic` study.

- Sample averages are  $\hat{\mu}_C = 27.4$  and  $\hat{\mu}_D = 23.4$
- The estimated pooled standard deviation for the five groups is 1.15
- The estimated average difference between groups  $C$  and  $D$  is  $\hat{\delta}_{CD} = 4$ .
- The standard error for the difference is  $\text{se}(\hat{\delta}_{CD}) = 1.6216$

# Example

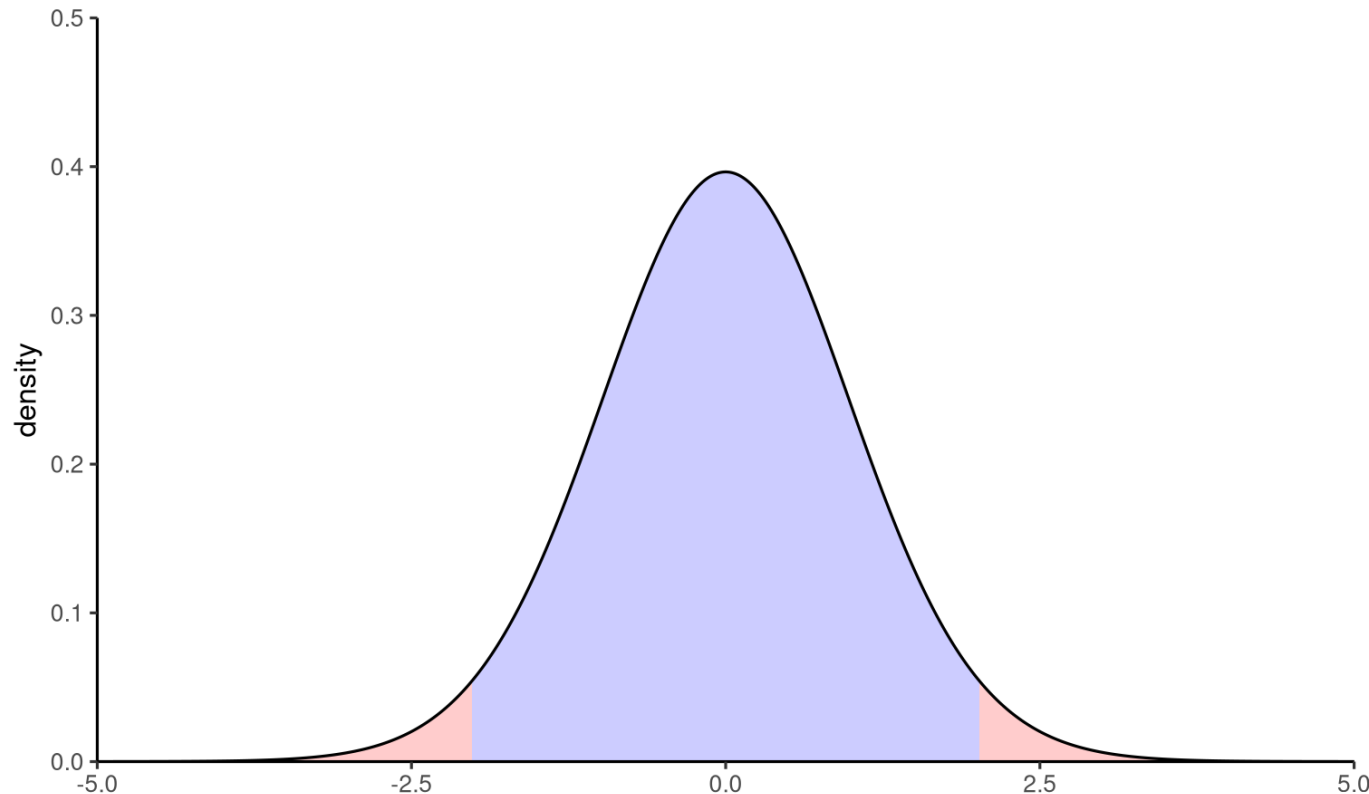
- If  $\mathcal{H}_0 : \delta_{CD} = 0$ , the  $t$  statistic is

$$t = \frac{\hat{\delta}_{CD} - 0}{\text{se}(\hat{\delta}_{CD})} = \frac{4}{1.6216} = 2.467$$

- The  $p$ -value is  $p = 0.018$ .
- We reject the null at level  $\alpha = 5\%$  since  $0.018 < 0.05$ .
- Conclude that there is a significant difference at level  $\alpha = 0.05$  between the average scores of subpopulations  $C$  and  $D$ .

# Null distribution

The blue area defines the set of values for which we fail to reject null  $\mathcal{H}_0$ .  
All values of  $t$  falling in the red area lead to rejection at level 5%.



# Critical values

For a test at level  $\alpha$  (two-sided), fail to reject all values of the test statistic  $t$  that are in interval

$$t_{n-k}(\alpha/2) \leq t \leq t_{n-k}(1 - \alpha/2)$$

Because of symmetry around zero,  $t_{n-k}(1 - \alpha/2) = -t_{n-k}(\alpha/2)$ .

- We call  $t_{n-k}(1 - \alpha/2)$  a **critical value**.
- in **R**, `qt(1-alpha/2, df = n - k)` where  $n$  is the number of observations and  $k$  the number of groups



# Confidence interval

Let  $\delta_{jk} = \mu_j - \mu_k$  denote the population difference,  $\hat{\delta}_{jk}$  the estimated difference (difference in sample averages) and  $\text{se}(\hat{\delta}_{jk})$  the estimated standard error.

The region for which we fail to reject the null is

$$t_{n-k}(\alpha/2) \leq \frac{\hat{\delta}_{jk} - \delta_{jk}}{\text{se}(\hat{\delta}_{jk})} \leq t_{n-k}(1 - \alpha/2)$$

which rearranged gives the  $(1 - \alpha)$  confidence interval for the (unknown) difference  $\delta_{jk}$ .

$$\hat{\delta}_{jk} - \text{se}(\hat{\delta}_{jk})t_{n-k}(1 - \alpha/2) \leq \delta_{jk} \leq \hat{\delta}_{jk} + \text{se}(\hat{\delta}_{jk})t_{n-k}(1 - \alpha/2)$$

# Interpretation of confidence intervals

The reported confidence interval is

$$[\hat{\delta}_{jk} + \mathbf{se}(\hat{\delta}_{jk})t_{n-k}(\alpha/2), \hat{\delta}_{jk} + \mathbf{se}(\hat{\delta}_{jk})t_{n-k}(1 - \alpha/2)].$$

Each bound is of the form

estimate + critical value  $\times$  standard error

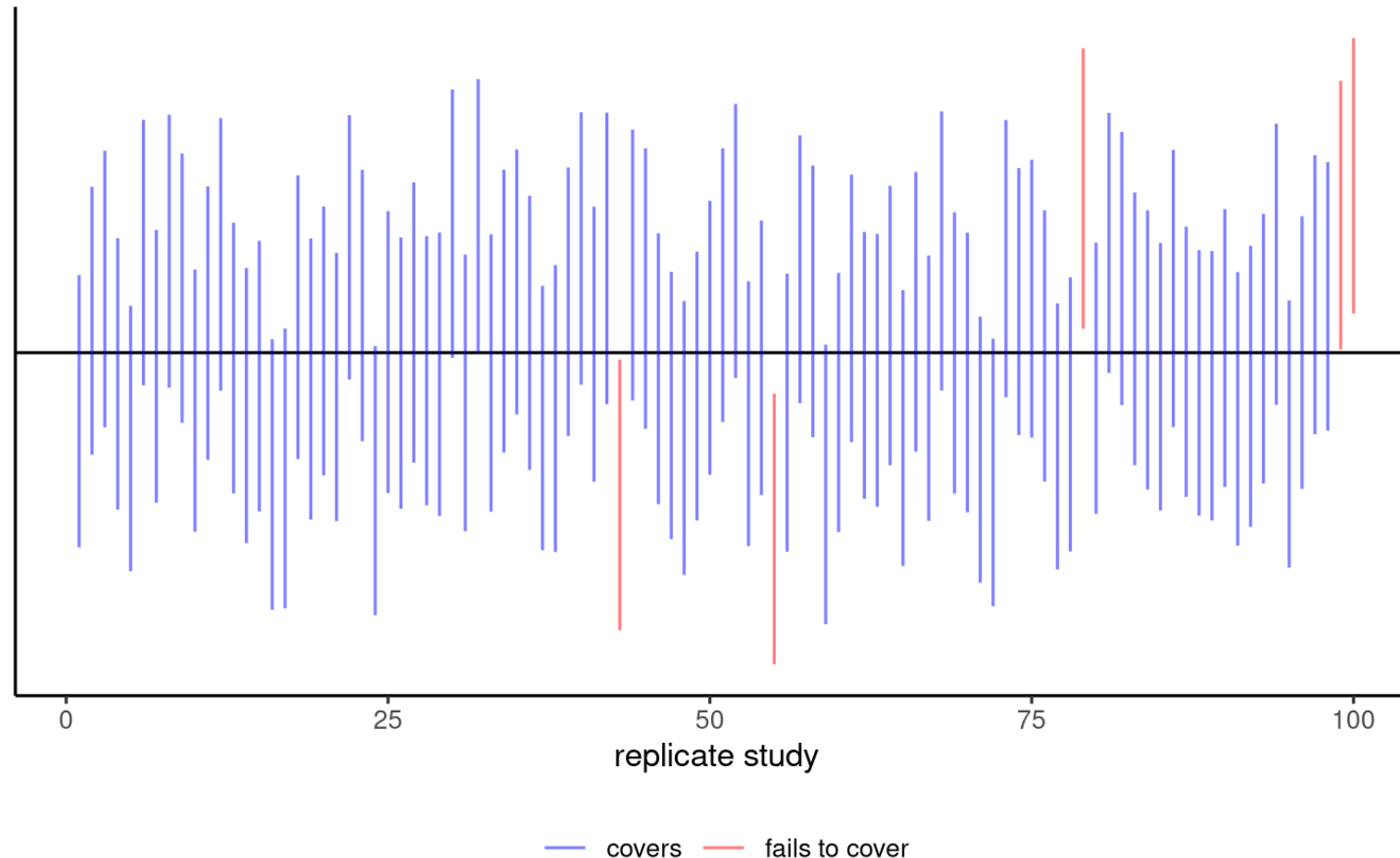
**confidence interval = [lower, upper] units**

If we replicate the experiment and compute confidence intervals each time

- on average, 95% of those intervals will contain the true value if the assumptions underlying the model are met.

# Interpretation in a picture: coin toss analogy

Each interval either contains the true value (black horizontal line) or doesn't.



# Why confidence intervals?

Test statistics are standardized,

- Good for comparisons with benchmark
- typically meaningless (standardized = unitless quantities)

Two options for reporting:

- $p$ -value: probability of more extreme outcome if no mean difference
- confidence intervals: set of all values for which we fail to reject the null hypothesis at level  $\alpha$  for the given sample

# Example

- Mean difference of  $\hat{\delta}_{CD} = 4$ , with  $\text{se}(\hat{\delta}_{CD}) = 1.6216$ .
- The critical values for a test at level  $\alpha = 5\%$  are  $-2.021$  and  $2.021$ 
  - $\text{qt}(0.975, \text{df} = 45 - 5)$
- Since  $|t| > 2.021$ , reject  $\mathcal{H}_0$ : the two population are statistically significant at level  $\alpha = 5\%$ .
- The confidence interval is

$$[4 - 1.6216 \times 2.021, 4 + 1.6216 \times 2.021] = [0.723, 7.28]$$

The postulated value  $\delta_{CD} = 0$  is not in the interval: reject  $\mathcal{H}_0$ .

# Pairwise differences in R

```
library(tidyverse) # data manipulation  
library(emmeans) # marginal means and contrasts  
url <- "https://edsm.rbind.io/data/arithmetic.csv"  
# load data, define column type (factor and integer)  
arithmetic <- read_csv(url, col_types = "fi")  
# fit one-way ANOVA model  
model <- lm(score ~ group, data = arithmetic)  
# Compute average of groups with model specification  
margmeans <- emmeans::emmeans(model, specs = "group")  
# Contrasts (default to pairwise comparisons) - no adjustment  
contrast(margmeans, adjust = 'none', infer = TRUE)  
#infer = TRUE for confidence intervals
```

# Parametrizations and interpretation

# Parametrization 1: sample averages

Most natural parametrization, not useful for test

- Sample sizes in each group:  $n_1, \dots, n_K$ , are known.
- sample average of each treatment group:  $\hat{\mu}_1, \dots, \hat{\mu}_K$ .

*$K$  means =  $K$  parameters*

Overall mean is

$$n\hat{\mu} = n_1\hat{\mu}_1 + \dots + n_K\hat{\mu}_K$$



# Parametrization 2: contrasts

In terms of differences, relative to a baseline category  $j$

- Intercept = sample mean  $\hat{\mu}_j$
- Coefficient for group  $k \neq j$ :  $\hat{\mu}_k - \hat{\mu}_j$ 
  - difference between averages of group  $k$  and baseline

In **R**, the baseline is the smallest alphanumerical value.

```
lm(response ~ group)
```

# Parametrization 3: sum-to-zero

In terms of differences, relative to average of  $\hat{\mu}_1, \dots, \hat{\mu}_K$

- Intercept =  $(\hat{\mu}_1 + \dots + \hat{\mu}_K) / K$
- Coefficient for group  $k$ :  $\hat{\mu}_k$  minus intercept

In **R**, the last factor level is dropped by default.

```
lm(response ~ group, contrasts = contr.sum(group))
```

Warning: Intercept  $\neq \hat{\mu}$  unless the sample is balanced.

# Comparison for the arithmetic example

<b>group</b>	<b>mean</b>	<b>contrasts</b>	<b>sum-to-zero</b>
intercept		19.66	21.00
control 1	19.66		-1.33
control 2	18.33	-1.33	-2.66
praised	27.44	7.77	6.44
reproved	23.44	3.77	2.44
ignored	16.11	-3.55	

# Planned comparisons and post-hoc tests

# Planned comparisons

Oftentimes, we are not interested in the global null hypothesis.

- Can formulate planned comparisons *at registration time* for effects of interest

**What is the scientific question of interest?**

# Arithmetic example

## Setup

group 1

(control)

group 2

(control)

group 3

(praise, reprove, ignore)

## Hypothesis of interest

- $\mathcal{H}_{01}: \mu_{\text{praise}} = \mu_{\text{reproved}}$  (attention)
- $\mathcal{H}_{02}: \frac{1}{2} (\mu_{\text{control}_1} + \mu_{\text{control}_2}) = \mu_{\text{praised}}$  (encouragement)

# Contrasts

With placeholders for each group, write  $\mathcal{H}_{01} : \mu_{\text{praised}} = \mu_{\text{reproved}}$  as

$$0 \cdot \mu_{\text{control}_1} + 0 \cdot \mu_{\text{control}_2} + 1 \cdot \mu_{\text{praised}} - 1 \cdot \mu_{\text{reproved}} + 0 \cdot \mu_{\text{ignored}}$$

The sum of the coefficients,  $(0, 0, 1, -1, 0)$ , is zero.

**Contrast = sum-to-zero constraint**

Similarly, for  $\mathcal{H}_{02} : \frac{1}{2}(\mu_{\text{control}_1} + \mu_{\text{control}_2}) = \mu_{\text{praise}}$

$$\frac{1}{2} \cdot \mu_{\text{control}_1} + \frac{1}{2} \cdot \mu_{\text{control}_2} - 1 \cdot \mu_{\text{praised}} + 0 \cdot \mu_{\text{reproved}} + 0 \cdot \mu_{\text{ignored}}$$

The entries of the contrast vector  $(\frac{1}{2}, \frac{1}{2}, -1, 0, 0)$  sum to zero.

Equivalent formulation is obtained by picking  $(1, 1, -2, 0, 0)$

# Contrasts in R

```
library(emmeans)
linmod <- lm(score ~ group, data = arithmetic)
linmod_emm <- emmeans(linmod, specs = 'group')
contrast_specif <- list(
  controlvspraised = c(0.5, 0.5, -1, 0, 0),
  praisedvsreproved = c(0, 0, 1, -1, 0)
)
contrasts_res <-
  contrast(object = linmod_emm,
           method = contrast_specif)
# Obtain confidence intervals instead of p-values
confint(contrasts_res)
```



# Post-hoc tests

Maybe there is some difference between groups?

Unplanned comparisons: go fishing...

Comparing all pairwise differences =  $\binom{K}{2}$  tests

With  $K = 5$  groups, we get 10 pairwise comparisons.

```
emmeans(modlin, pairwise ~ group)
```

If there were no differences between the groups, how many do we expect to find significant by chance with  $\alpha = 0.1$ ?