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

General recipe of hypothesis testing

(3) Choose/compute a test statistic

(4) Compare statistic to null distribution

General recipe of hypothesis testing

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

(6) Conclude (reject/fail to reject)

Level = probability of condemning an innocent

Fix level α before the experiment.

Choose small α (typical value is 5%)

Reject $\mathscr{H}_0$ if p-value less than α

F-test for one way ANOVA

Global null hypothesis

No difference between treatments

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

Building a statistic

Denote

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

Decomposing variability into bits

$$\underbrace{\sum_i \sum_k (y_{ik} - \widehat{\mu})^2}_{\text{total sum of squares}} = \underbrace{\sum_i \sum_k (y_{ik} - \widehat{\mu}_k)^2}_{\text{within sum of squares}} + \underbrace{\sum_k n_i (\widehat{\mu}_k - \widehat{\mu})^2}_{\text{between sum of squares}} .$$

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 $\mathsf{F}(K-1, n-K)$.

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:

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

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

The standard error $\mathsf{se}(\widehat{\mu}_j - \widehat{\mu}_k)$ 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

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

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

- 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 reproved (group D) of the `arithmetic` study.

- Sample averages are $\widehat{\mu}_C = 27.4$ and $\widehat{\mu}_D = 23.4$
- 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 se$(\hat{\delta}_{CD}) = 1.6216$

Example

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

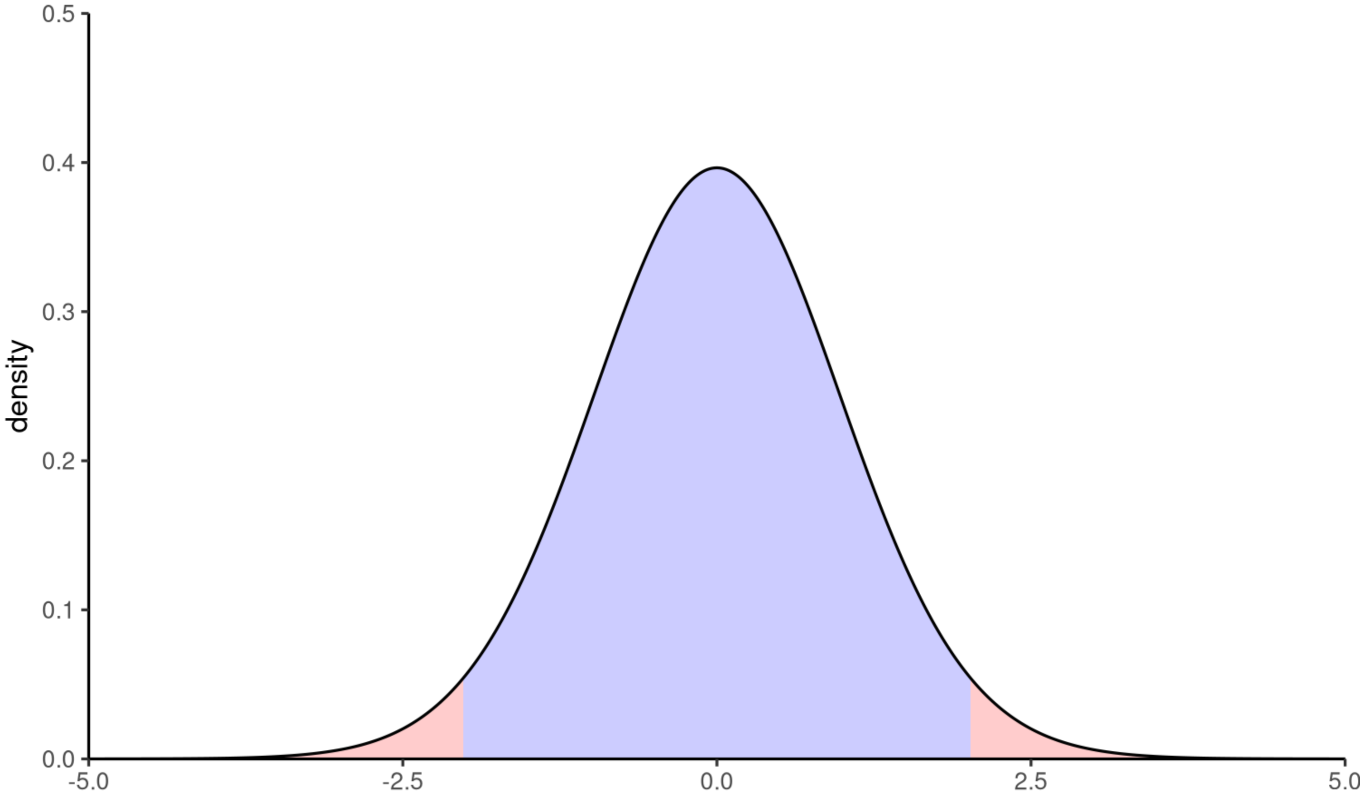
$$t = \frac{\hat{\delta}_{CD} - 0}{\text{se}(\hat{\delta}_{CD})} = \frac{4}{1.6216} = 2.467$$

- 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 $\mathscr{H}_0$.

All values of t falling in the red aread lead to rejection at level 5%.



Critical values

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

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

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

- We call $\mathfrak{t}_{n-k}(1-\alpha/2)$ 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 $\mathsf{se}(\hat{\delta}_{jk})$ the estimated standard error.

The region for which we fail to reject the null is

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

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

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

Interpretation of confidence intervals

The reported confidence interval is

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

Each bound is of the form

$$\text{estimate} + \text{critical value} \times \text{standard error}$$

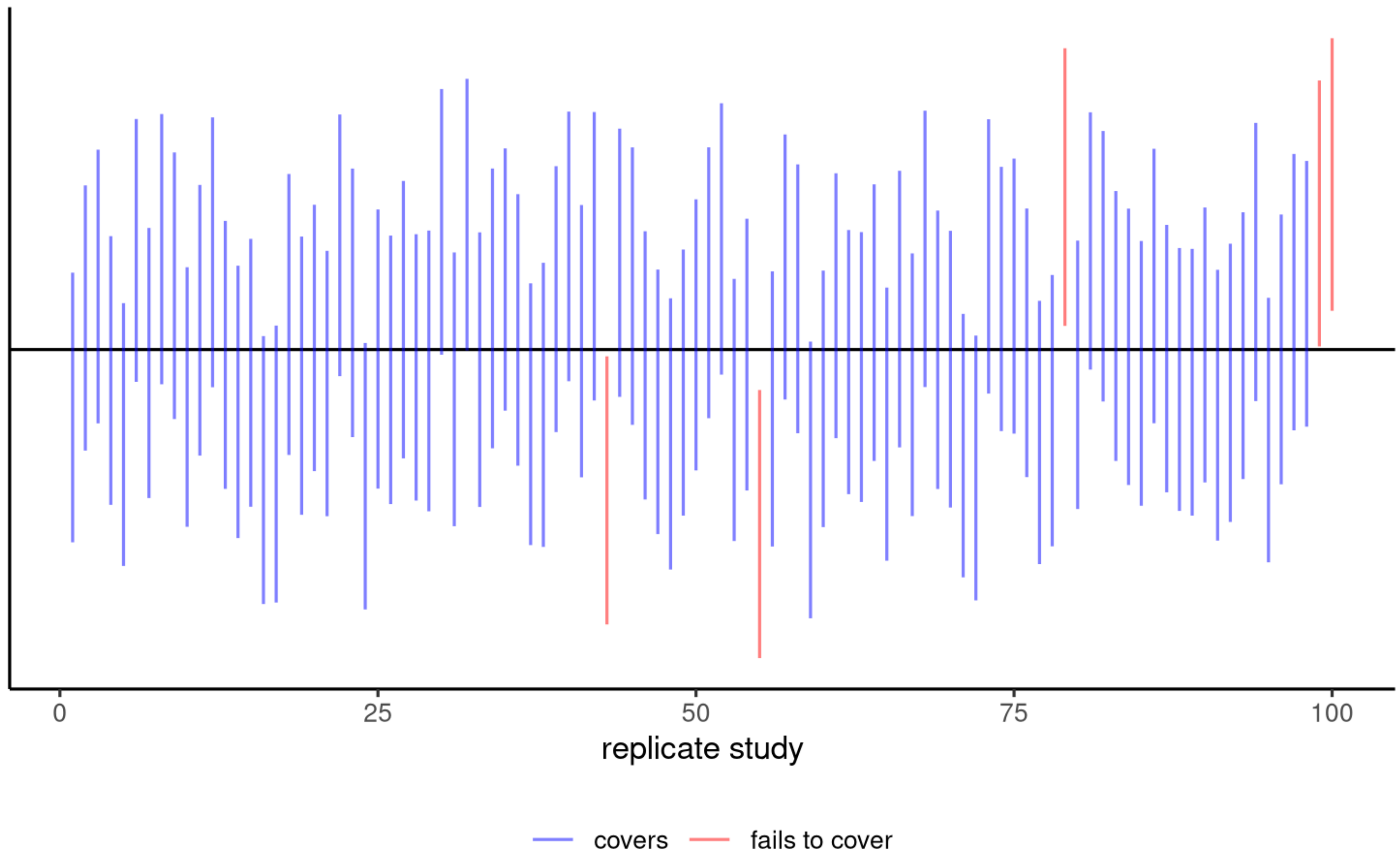
confidence interval = [lower, upper] units

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 α 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
  - `qt(0.975, df = 45 - 5)`
- Since $|t| > 2.021$, reject $\mathscr{H}_0$: 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 $\mathscr{H}_0$.

Pairwise differences in R

```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: $n_1, \ldots, n_K$, are known.
- sample average of each treatment group: $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$.

K means = K parameters

Overall mean is

$$n\widehat{\mu} = n_1\widehat{\mu}_1 + \cdots + n_K\widehat{\mu}_K$$

Parametrization 2: contrasts

In terms of differences, relative to a baseline category j

- Intercept = sample mean $\widehat{\mu}_j$
- Coefficient for group $k \neq j$: $\widehat{\mu}_k - \widehat{\mu}_j$
  - 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 $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$

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

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

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

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

Comparison for the arithmetic example

group	mean	contrasts	sum-to-zero
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	group 2	group 3
(control)	(control)	(praise, reprove, ignore)

Hypothesis of interest

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

Contrasts

With placeholders for each group, write $\mathscr{H}_{01}: \mu_{\text{praised}} = \mu_{\text{reproved}}$ 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 $\mathscr{H}_{02}: \frac{1}{2}(\mu_{\text{control}_1} + \mu_{\text{control}_2}) = \mu_{\text{praise}}$

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

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

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

Contrasts in R

```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 $\alpha = 0.1$?