

# Complete factorial designs

## Session 6

MATH 80667A: Experimental Design and Statistical Methods  
HEC Montréal

# Outline

**Unbalanced designs**

**Multifactorial designs**

# Unbalanced designs

# Premise

So far, we have exclusively considered balanced samples

**balanced = same number of observational  
units in each subgroup**

Most experiments (even planned) end up with unequal sample sizes.

# Noninformative drop-out

Unbalanced samples may be due to many causes, including randomization (need not balance) and loss-to-follow up (dropout)

If dropout is random, not a problem

- Example of Baumann, Seifert-Kessel, Jones (1992):

Because of illness and transfer to another school, incomplete data were obtained for one subject each from the TA and DRTA group

# Problematic drop-out or exclusion

If loss of units due to treatment or underlying conditions, problematic!

Rosensaal (2021) rebuking a study on the effectiveness of hydrochloriquine as treatment for Covid19 and reviewing allocation:

Of these 26, six were excluded (and incorrectly labelled as lost to follow-up): three were transferred to the ICU, one died, and two terminated treatment or were discharged

Sick people excluded from the treatment group! then claim it is better.

Worst: "The index [treatment] group and control group were drawn from different centres."

# Why seek balance?

Two main reasons

1. Power considerations: with equal variance in each group, balanced samples gives the best allocation
2. Simplicity of interpretation and calculations: the interpretation of the  $F$  test in a linear regression is unambiguous

# Finding power in balance

Consider a t-test for assessing the difference between treatments  $A$  and  $B$  with equal variability

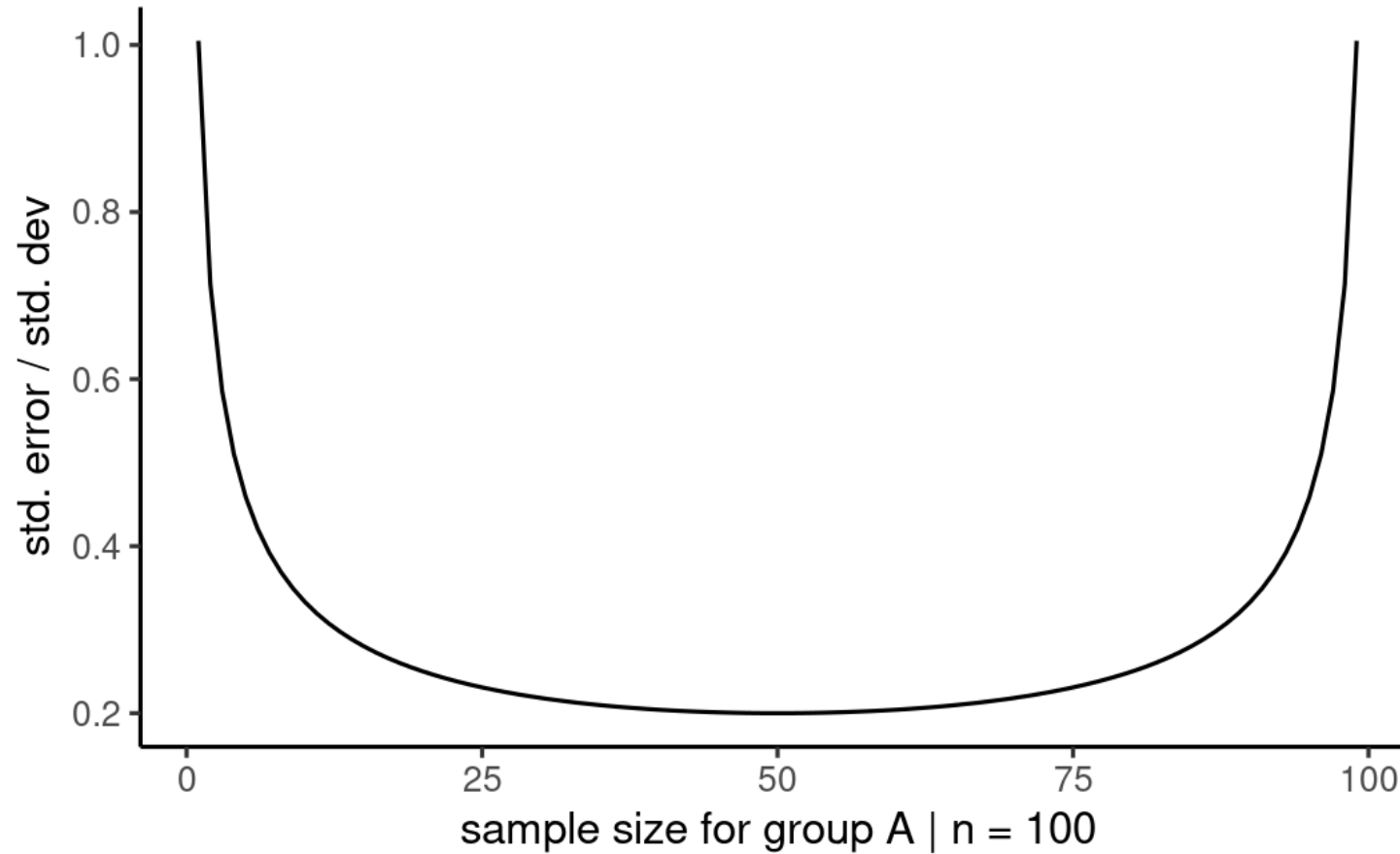
$$t = \frac{\text{estimated difference}}{\text{estimated variability}} = \frac{(\hat{\mu}_A - \hat{\mu}_B) - 0}{\text{se}(\hat{\mu}_A - \hat{\mu}_B)}.$$

The standard error of the average difference is

$$\sqrt{\frac{\text{variance}_A}{\text{nb of obs. in } A} + \frac{\text{variance}_B}{\text{nb of obs. in } B}} = \sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}$$



# Optimal allocation of ressources



The allocation of  $n = n_A + n_B$  units that minimizes the std error is  $n_A = n_B = n/2$ .

# Example: tempting fate

We consider data from Multi Lab 2, a replication study that examined Risen and Gilovich (2008) who

explored the belief that tempting fate increases bad outcomes. They tested whether people judge the likelihood of a negative outcome to be higher when they have imagined themselves [...] tempting fate [...] (by not reading before class) or not [tempting] fate (by coming to class prepared). Participants then estimated how likely it was that [they] would be called on by the professor (scale from 1, not at all likely, to 10, extremely likely).

The replication data gathered in 37 different labs focuses on a 2 by 2 factorial design with gender (male vs female) and condition (prepared vs unprepared) administered to undergraduates.

---

Load data

Check balance

Marginal means

- We consider a 2 by 2 factorial design.
- The response is `likelihod`
- The experimental factors are `condition` and `gender`
- Two data sets: `RS_unb` for the full data, `RS_bal` for the artificially balanced one.

# Explaining the discrepancies

Estimated marginal means are based on equiweighted groups:

$$\hat{\mu} = \frac{1}{4}(\hat{\mu}_{11} + \hat{\mu}_{12} + \hat{\mu}_{21} + \hat{\mu}_{22})$$

where  $\hat{\mu}_{ij} = n_{ij}^{-1} \sum_{r=1}^{n_{ij}} y_{ijr}$ .

The sample mean is the sum of observations divided by the sample size.

The two coincide when  $n_{11} = \dots = n_{22}$ .

# Why equal weight?

- The ANOVA and contrast analyses, in the case of unequal sample sizes, are generally based on marginal means (same weight for each subgroup).
- This choice is justified because research questions generally concern comparisons of means across experimental groups.

# Revisiting the $F$ statistic

Statistical tests contrast competing **nested** models:

- an alternative model, sometimes termed "full model"
- a null model, which imposes restrictions (a simplification of the alternative model)

The numerator of the  $F$ -statistic compares the sum of square of a model with (given) main effect, etc., to a model without.

# What is explained by condition?

Consider the  $2 \times 2$  factorial design with factors  $A$ : gender and  $B$ : condition (prepared vs unprepared) without interaction.

What is the share of variability (sum of squares) explained by the experimental condition?

# Comparing differences in sum of squares (1)

Consider a balanced sample

```
anova(lm(likelihood ~ 1, data = RS_bal),  
      lm(likelihood ~ condition, data = RS_bal))  
# When gender is present  
anova(lm(likelihood ~ gender, data = RS_bal),  
      lm(likelihood ~ gender + condition, data = RS_bal))
```

The difference in sum of squares is 141.86 in both cases.



# Comparing differences in sum of squares (2)

Consider an unbalanced sample

```
anova(lm(likelihood ~ 1, data = RS_unb),  
      lm(likelihood ~ condition,  
          data = RS_unb))  
# When gender is present  
anova(lm(likelihood ~ gender, data = RS_unb),  
      lm(likelihood ~ gender + condition,  
          data = RS_unb))
```

The differences of sum of squares are respectively 330.95 and 332.34.

# Orthogonality

Balanced designs yield orthogonal factors: the improvement in the goodness of fit (characterized by change in sum of squares) is the same regardless of other factors.

So effect of  $B$  and  $B \mid A$  (read  $B$  given  $A$ ) is the same.

- test for  $B \mid A$  compares  $SS(A, B) - SS(A)$
- for balanced design,  $SS(A, B) = SS(A) + SS(B)$  (factorization).

We lose this property with unbalanced samples: there are distinct formulations of ANOVA.

# Analysis of variance - Type I (sequential)

The default method in **R** with `anova` is the sequential decomposition: in the order of the variables  $A, B$  in the formula

- So  $F$  tests are for tests of effect of
  - $A$ , based on  $SS(A)$
  - $B \mid A$ , based on  $SS(A, B) - SS(A)$
  - $AB \mid A, B$  based on  $SS(A, B, AB) - SS(A, B)$

Ordering matters

Since the order in which we list the variable is **arbitrary**, these  $F$  tests are not of interest.

# Analysis of variance - Type II

Impact of

- $A \mid B$  based on  $SS(A, B) - SS(B)$
- $B \mid A$  based on  $SS(A, B) - SS(A)$
- $AB \mid A, B$  based on  $SS(A, B, AB) - SS(A, B)$
- tests invalid if there is an interaction.
- In **R**, use `car::Anova(model, type = 2)`

# Analysis of variance - Type III

Most commonly used approach

- Improvement due to  $A \mid B, AB, B \mid A, AB$  and  $AB \mid A, B$
- What is improved by adding a factor, interaction, etc. given the rest
- may require imposing equal mean for rows for  $A \mid B, AB$ , etc.
  - (**requires** sum-to-zero parametrization)
- valid in the presence of interaction
- but  $F$ -tests for main effects are not of interest
- In **R**, use `car::Anova(model, type = 3)`

# ANOVA for unbalanced data

```
model <- lm(
  likelihood ~ condition * gender,
  data = RS_unb)
# Three distinct decompositions
anova(model) #type 1
car::Anova(model, type = 2)
car::Anova(model, type = 3)
```

ANOVA (type I)

|                  | <b>Df</b> | <b>Sum Sq</b> | <b>F value</b> |
|------------------|-----------|---------------|----------------|
| gender           | 1         | 164.94        | 29.1           |
| condition        | 1         | 332.34        | 58.7           |
| gender:condition | 1         | 36.55         | 6.5            |
| Residuals        | 4429250   | 86.33         |                |

ANOVA (type II)

|                  | <b>Df</b> | <b>Sum Sq</b> | <b>F value</b> |
|------------------|-----------|---------------|----------------|
| gender           | 1         | 166.33        | 29.4           |
| condition        | 1         | 332.34        | 58.7           |
| gender:condition | 1         | 36.55         | 6.5            |
| Residuals        | 4429250   | 86.33         |                |

ANOVA (type III)

|                  | <b>Df</b> | <b>Sum Sq</b> | <b>F value</b> |
|------------------|-----------|---------------|----------------|
| gender           | 1         | 167.71        | 29.6           |
| condition        | 1         | 227.88        | 40.2           |
| gender:condition | 1         | 36.55         | 6.5            |
| Residuals        | 4429250   | 86.33         |                |

# ANOVA for balanced data

```
model2 <- lm(
  likelihood ~ condition * gender,
  data = RS_bal)
anova(model2) #type 1
car::Anova(model2, type = 2)
car::Anova(model2, type = 3)
# Same answer - orthogonal!
```

ANOVA (type I)

|                  | <b>Df</b> | <b>Sum</b> | <b>SqF</b> | <b>value</b> |
|------------------|-----------|------------|------------|--------------|
| condition        | 1         | 141.86     | 24.1       |              |
| gender           | 1         | 121.69     | 20.6       |              |
| condition:gender | 1         | 37.88      | 6.4        |              |
| Residuals        | 2500      | 14733.84   |            |              |

ANOVA (type II)

|                  | <b>Df</b> | <b>Sum</b> | <b>SqF</b> | <b>value</b> |
|------------------|-----------|------------|------------|--------------|
| condition        | 1         | 141.86     | 24.1       |              |
| gender           | 1         | 121.69     | 20.6       |              |
| condition:gender | 1         | 37.88      | 6.4        |              |
| Residuals        | 2500      | 14733.84   |            |              |

ANOVA (type III)

|                  | <b>Df</b> | <b>Sum</b> | <b>SqF</b> | <b>value</b> |
|------------------|-----------|------------|------------|--------------|
| condition        | 1         | 141.86     | 24.1       |              |
| gender           | 1         | 121.69     | 20.6       |              |
| condition:gender | 1         | 37.88      | 6.4        |              |
| Residuals        | 2500      | 14733.84   |            |              |

# Recap

- If each observation has the same variability, a balanced sample maximizes power.
- Balanced designs have interesting properties:
  - estimated marginal means coincide with (sub)samples averages
  - the tests of effects are unambiguous
  - for unbalanced samples, we work with marginal means and type 3 ANOVA
  - if empty cells (no one assigned to a combination of treatment), cannot estimate corresponding coefficients (typically higher order interactions)



# Practice

From the OSC psychology replication

People can be influenced by the prior consideration of a numerical anchor when forming numerical judgments. [...] The anchor provides an initial starting point from which estimates are adjusted, and a large body of research demonstrates that adjustment is usually insufficient, leading estimates to be biased towards the initial anchor.

Replication of Study 4a of Janiszewski & Uy (2008, Psychological Science) by J. Chandler

# Multifactorial designs

# Beyond two factors

We can consider multiple factors  $A, B, C, \dots$  with respectively  $n_a, n_b, n_c, \dots$  levels and with  $n_r$  replications for each.

The total number of treatment combinations is

$$n_a \times n_b \times n_c \times \dots$$

**Curse of dimensionality**

# Full three-way ANOVA model

Each cell of the cube is allowed to have a different mean

$$\begin{array}{ccccc} Y_{ijk r} & = & \mu_{ijk} & + & \varepsilon_{ijk r} \\ \text{response} & & \text{cell mean} & & \text{error} \end{array}$$

with  $\varepsilon_{ijk r}$  are independent error term for

- row  $i$
- column  $j$
- depth  $k$
- replication  $r$

# Parametrization of a three-way ANOVA model

With the **sum-to-zero** parametrization with factors  $A$ ,  $B$  and  $C$ , write the response as

$$\begin{aligned} \underset{\text{theoretical average}}{\mathbf{E}(Y_{ijk})} &= \underset{\text{global mean}}{\mu} \\ &+ \underset{\text{main effects}}{\alpha_i + \beta_j + \gamma_k} \\ &+ \underset{\text{two-way interactions}}{(\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}} \\ &+ \underset{\text{three-way interaction}}{(\alpha\beta\gamma)_{ijk}} \end{aligned}$$



global mean, row, column and depth main effects



row/col, row/depth and col/depth interactions and three-way interaction.

# Example of three-way design

Petty, Cacioppo and Heesacker (1981). Effects of rhetorical questions on persuasion: A cognitive response analysis. Journal of Personality and Social Psychology.

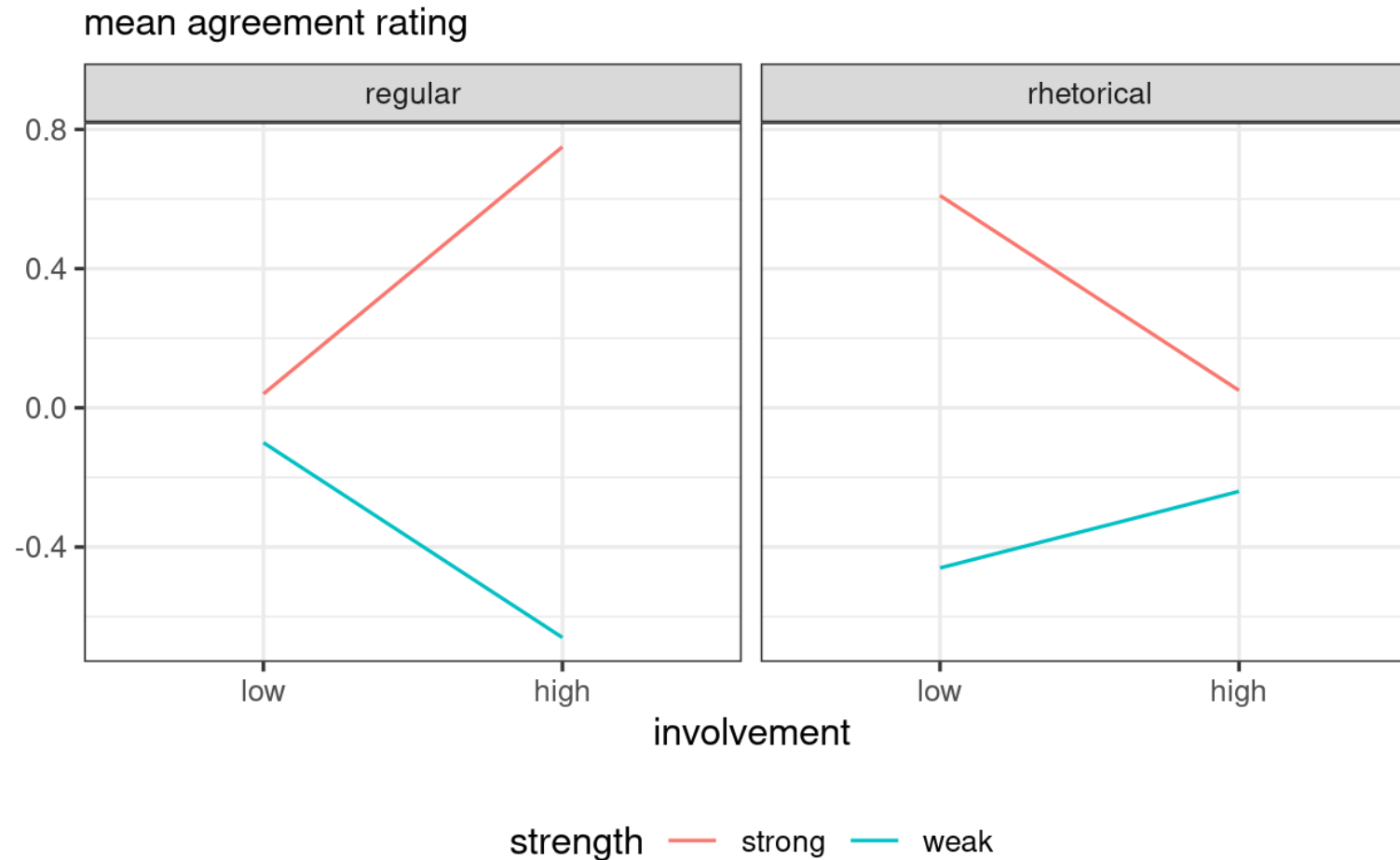
A  $2 \times 2 \times 2$  factorial design with 8 treatments groups and  $n = 160$  undergraduates.

Setup: should a comprehensive exam be administered to bachelor students in their final year?

- **Response** Likert scale on  $-5$  (do not agree at all) to  $5$  (completely agree)
- **Factors**
  - $A$ : strength of the argument (strong or weak)
  - $B$ : involvement of students low (far away, in a long time) or high (next year, at their university)
  - $C$ : style of argument, either regular form or rhetorical (Don't you think?, ...)

# Interaction plot

Interaction plot for a  $2 \times 2 \times 2$  factorial design from Petty, Cacioppo and Heesacker (1981)





# The microwave popcorn experiment

What is the best brand of microwave popcorn?

- **Factors**
- brand (two national, one local)
- power: 500W and 600W
- time: 4, 4.5 and 5 minutes
- **Response:** ~~weight, volume, number~~, percentage of popped kernels.
- Pilot study showed average of 70% overall popped kernels (10% standard dev), timing values reasonable
- Power calculation suggested at least  $r = 4$  replicates, but researchers proceeded with  $r = 2$ ...

---

ANOVA

QQ-plot

R code

Interaction plot

```
data(popcorn, package = 'hecedsr')  
# Fit model with three-way interaction  
model <- aov(percentage ~ brand*power*time,  
             data = popcorn)  
# ANOVA table - 'anova' is ONLY for balanced designs  
anova_table <- anova(model)  
# Quantile-quantile plot  
car::qqPlot(model)
```

Model assumptions: plots and tests are meaningless with ( $n_r=2$ ) replications per group...

# Analysis of variance table for balanced designs

| terms    | degrees of freedom            |
|----------|-------------------------------|
| $A$      | $n_a - 1$                     |
| $B$      | $n_b - 1$                     |
| $C$      | $n_c - 1$                     |
| $AB$     | $(n_a - 1)(n_b - 1)$          |
| $AC$     | $(n_a - 1)(n_c - 1)$          |
| $BC$     | $(n_b - 1)(n_c - 1)$          |
| $ABC$    | $(n_a - 1)(n_b - 1)(n_c - 1)$ |
| residual | $n_a n_b n_c (R - 1)$         |
| total    | $n_a n_b n_c n_r - 1$         |

Analysis of variance table for microwave-popcorn

|                  | Degrees of freedom | Sum of squares | Mean square | F statistic | p-value |
|------------------|--------------------|----------------|-------------|-------------|---------|
| brand            | 2                  | 331.10         | 165.55      | 1.89        | 0.180   |
| power            | 1                  | 455.11         | 455.11      | 5.19        | 0.035   |
| time             | 2                  | 1554.58        | 777.29      | 8.87        | 0.002   |
| brand:power      | 2                  | 196.04         | 98.02       | 1.12        | 0.349   |
| brand:time       | 4                  | 1433.86        | 358.46      | 4.09        | 0.016   |
| power:time       | 2                  | 47.71          | 23.85       | 0.27        | 0.765   |
| brand:power:time | 4                  | 47.33          | 11.83       | 0.13        | 0.967   |

# Omitting terms in a factorial design

The more levels and factors, the more parameters to estimate (and replications needed)

- Costly to get enough observations / power
- The assumption of normality becomes more critical when  $r = 2!$

It may be useful not to consider some interactions if they are known or (strongly) suspected not to be present

- If important interactions are omitted from the model, biased estimates/output!

# Guidelines for the interpretation of effects

Start with the most complicated term (top down)

- If the three-way interaction  $ABC$  is significant:
  - don't interpret main effects or two-way interactions!
  - comparison is done cell by cell within each level
- If the  $ABC$  term isn't significant:
  - can marginalize and interpret lower order terms
  - back to a series of two-way ANOVAs

# What contrasts are of interest?

- Can view a three-way ANOVA as a series of one-way ANOVA or two-way ANOVAs...

Depending on the goal, could compare for variable  $A$

- marginal contrast  $\psi_A$  (averaging over  $B$  and  $C$ )
- marginal conditional contrast for particular subgroup:  $\psi_A$  within  $c_1$
- contrast involving two variables:  $\psi_{AB}$
- contrast differences between treatment at  $\psi_A \times B$ , averaging over  $C$ .
- etc.

See helper code and chapter 22 of Keppel & Wickens (2004) for a detailed example.

# Effects and contrasts for microwave-popcorn

Following preplanned comparisons

- Which combo (brand, power, time) gives highest popping rate? (pairwise comparisons of all combos)
- Best brand overall (marginal means marginalizing over power and time, assuming no interaction)
- Effect of time and power on percentage of popped kernels
- pairwise comparison of time  $\times$  power
- main effect of power
- main effect of time



# Preplanned comparisons using emmeans

Let  $A$ =brand,  $B$ =power,  $C$ =time

Compare difference between percentage of popped kernels for 4.5 versus 5 minutes, for brands 1 and 2

$$\mathcal{H}_0 : (\mu_{1.2} - \mu_{1.3}) - (\mu_{2.2} - \mu_{2.3}) = 0$$

```
library(emmeans)
# marginal means
emm_popcorn_AC <- emmeans(model,
                           specs = c("brand", "time"))
contrast_list <-
  list(
    brand12with4.5vs5min = c(0, 0, 0, 1, -1, 0, -1, 1, 0))
contrast(emm_popcorn_AC, # marginal mean (no time)
         method = contrast_list) # list of contrasts
```

# Preplanned comparisons

Compare all three times (4, 4.5 and 5 minutes)

At level 99% with Tukey's HSD method

- Careful! Potentially misleading because there is a `brand * time` interaction present.

```
# List of variables to keep go in `specs`: keep only time
emm_popcorn_C <- emmeans(model, specs = "time")
pairs(emm_popcorn_C,
      adjust = "tukey",
      level = 0.99,
      infer = TRUE)
```