Repeated measures

Session 6

MATH 80667A: Experimental Design and Statistical Methods
HEC Montréal

Outline

Unbalanced designs

Repeated measures

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 sample allocation (easier to detect true differences in mean) by minimizing variability.
2. Simplicity of interpretation and calculations: the interpretation of the F test in a linear regression is unambiguous

Finding power in balance

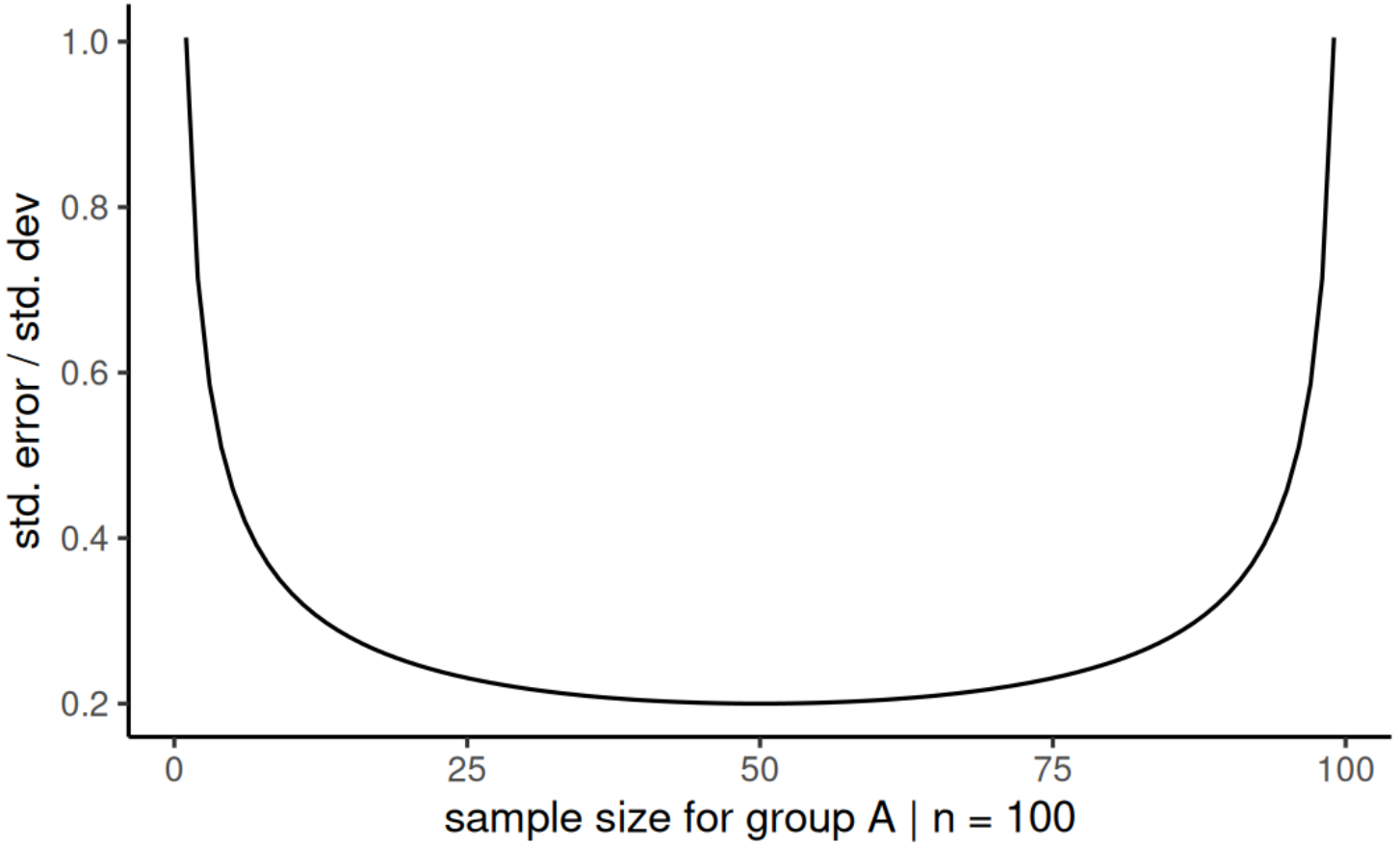
Consider a t-test for assessing the difference between treatments A and B with equal variability

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

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.

Example - loading data

- 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.

Checking balance

```
summary_stats <-
  RS_unb |>
  group_by(condition) |>
  summarize(nobs = n(),
            mean = mean(likelihood))
```

Summary statistics

condition	nobs	mean
unprepared	2192	4.606
prepared	2241	4.060

Marginal means

```
# Enforce sum-to-zero parametrization
options(contrasts = rep("contr.sum", 2))
# Anova is a linear model, fit using 'lm'
# 'aov' only for *balanced data*
model <- lm(
  likelihood ~ gender * condition,
  data = RS_unb)
library(emmeans)
emm <- emmeans(model,
               specs = "condition")
```

Marginal means for condition

condition	emmean	SE
unprepared	4.504	0.0540
prepared	4.022	0.0535

Note unequal standard errors.

Explaining the discrepancies

Estimated marginal means are based on equiweighted groups:

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

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

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

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

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 2×2 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 $\mathsf{SS}(A, B) - \mathsf{SS}(A)$
- for balanced design, $\mathsf{SS}(A, B) = \mathsf{SS}(A) + \mathsf{SS}(B)$ (factorization).

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

Analysis of variance - Type 1 (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 $\mathsf{SS}(A)$
  - $B \mid A$, based on $\mathsf{SS}(A, B) - \mathsf{SS}(A)$
  - $AB \mid A, B$ based on $\mathsf{SS}(A, B, AB) - \mathsf{SS}(A, B)$

Ordering matters

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

Analysis of variance - Type 2

Impact of

- $A \mid B$ based on $\mathsf{SS}(A, B) - \mathsf{SS}(B)$
- $B \mid A$ based on $\mathsf{SS}(A, B) - \mathsf{SS}(A)$
- $AB \mid A, B$ based on $\mathsf{SS}(A, B, AB) - \mathsf{SS}(A, B)$
- the first tests are not of interest if there is an interaction.
- In **R**, use `car::Anova(model, type = 2)`

Analysis of variance - Type 3

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
- does not respect the marginality principle.
- 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 1)

	Df	Sum Sq	F value
gender	1	164.94	29.1
condition	1	332.34	58.7
gender:condition	1	36.55	6.5
Residuals	4429	25086.33	

ANOVA (type 2)

	Df	Sum Sq	F value
gender	1	166.33	29.4
condition	1	332.34	58.7
gender:condition	1	36.55	6.5
Residuals	4429	25086.33	

ANOVA (type 3)

	Df	Sum Sq	F value
gender	1	167.71	29.6
condition	1	227.88	40.2
gender:condition	1	36.55	6.5
Residuals	4429	25086.33	

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 1)

	Df	Sum Sq	F value
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 2)

	Df	Sum Sq	F value
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 3)

	Df	Sum Sq	F value
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 2 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

Repeated measures ANOVA

Beyond between-designs

Each subject (experimental unit) assigned to a single condition.

- individuals (subjects) are **nested** within condition/treatment.

In many instances, it may be possible to randomly assign multiple conditions to each experimental unit.

Benefits of within-designs

Assign (some or) all treatments to subjects and measure the response.

Benefits:

- Each subject (experimental unit) serves as its own control (greater comparability among treatment conditions).
- Filter out effect due to subject (like blocking):
  - increased precision
  - increased power (tests are based on within-subject variability)

Impact: need smaller sample sizes than between-subjects designs

Drawbacks of within-designs

Potential sources of bias include

- Period effect (e.g., practice or fatigue).
- Carryover effects.
- Permanent change in the subject condition after a treatment assignment.
- Loss of subjects over time (attrition).

Minimizing sources of bias

- Randomize the order of treatment conditions among subjects
- or use a balanced crossover design and include the period and carryover effect in the statistical model (confounding or control variables to better isolate the treatment effect).
- Allow enough time between treatment conditions to reduce or eliminate period or carryover effects.

One-way ANOVA with a random effect

As before, we have one experimental factor A with n_a levels, with

$$\underset{\text{response}}{Y_{ij}} = \underset{\text{global mean}}{\mu} + \underset{\text{mean difference}}{\alpha_j} + \underset{\text{random effect for subject}}{S_i} + \underset{\text{error}}{\varepsilon_{ij}}$$

where $S_i \sim \textbf{normal}(0, \sigma_s^2)$ and $\varepsilon_{ij} \sim \textbf{normal}(0, \sigma_e^2)$.

The errors and random effects are independent from one another.

Variance components

The model **parameters** includes two measures of variability σ_s^2 and σ_e^2.

- The variance of the response Y_{ij} is $\sigma_s^2 + \sigma_e^2$.
- The **intra-class correlation** between observations in group i is $\rho = \sigma_s^2/(\sigma_s^2 + \sigma_e^2)$.
  - observations from the same subject are correlated
  - observations from different subjects are independent

This dependence structure within group is termed **compound symmetry**.

Example: happy fakes

An experiment conducted in a graduate course at HEC gathered electroencephalography (EEG) data.

The response variable is the amplitude of a brain signal measured at 170 ms after the participant has been exposed to different faces.

Repeated measures were collected on 12 participants, but we focus only on the average of the replications.

Experimental conditions

The control (`real`) is a true image, whereas the other were generated using a generative adversarial network (GAN) so be slightly smiling (`GAN1`) or extremely smiling (`GAN2`, looks more fake).

Research question: do the GAN image trigger different reactions (pairwise difference with control)?

Models for repeated measures

If we average, we have a balanced randomized blocked design with

- `id` (blocking factor)
- `stimulus` (experimental factor)

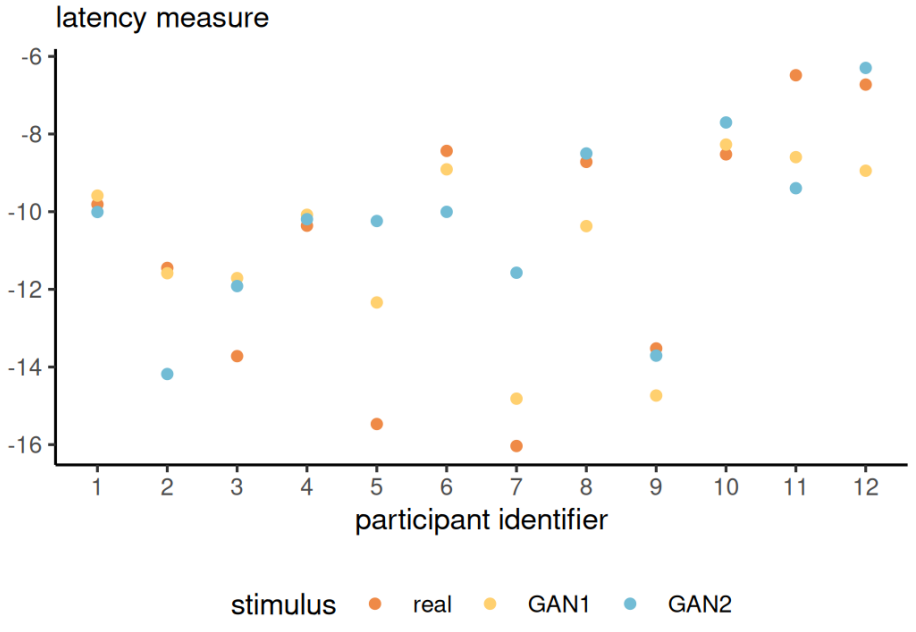
We use the `afex` package to model the within-subject structure.

Load data

```r
# Set sum-to-zero constraint for factors
options(contrasts = c("contr.sum", "contr.poly"))
data(AA21, package = "hecedsm")
# Compute mean
AA21_m <- AA21 |>
  dplyr::group_by(id, stimulus) |>
  dplyr::summarize(latency = mean(latency))
```

Graph

```
library(ggplot2)
ggplot(data = AA21_m,
       aes(x = id,
           colour = stimulus,
           y = latency)) +
  geom_point()
```



ANOVA

```
model <- afex::aov_ez(
  id = "id",            # subject id
  dv = "latency",       # response
  within = "stimulus", # within-subject
  data = hecedsm::AA21,
  fun_aggregate = mean)
# aggregate to one measure per id/conditic
anova(model, # mixed ANOVA model
      correction = "none", # sphericity
      es = "none") # effect size
```

```
# Anova Table (Type 3 tests)
#
# Response: latency
#           num Df den Df   MSE     F Pr(>F)
# stimulus      2     22 1.955 0.496 0.6155
```

- No detectable difference between conditions.

- Residual degrees of freedom: $(n_a - 1) \times (n_s - 1) = 22$ for $n_s = 12$ subjects and $n_a = 3$ levels.

Model assumptions

The validity of the F null distribution relies on the model having the correct structure.

- Same variance per observation $\mathsf{H}_0 : \sigma^2_{\texttt{real}} = \sigma^2_{\texttt{GAN1}} = \sigma^2_{\texttt{GAN2}}$
- equal correlation between measurements of the same subject (*compound symmetry*) $\mathsf{H}_0 : \rho_{\texttt{real,GAN1}} = \rho_{\texttt{real,GAN2}} = \rho_{\texttt{GAN1,GAN2}}$
- normality of the **random** (subject specific) effect

Sphericity

Since we care only about differences in treatment, can get away with a weaker assumption than compound symmetry.

Sphericity: variance of difference between treatment is constant.

In our example, this means:

$$\mathscr{H}_0 : \mathsf{Va}(\widehat{\mu}_{\texttt{real}} - \widehat{\mu}_{\texttt{GAN1}}) = \mathsf{Va}(\widehat{\mu}_{\texttt{real}} - \widehat{\mu}_{\texttt{GAN2}}) = \mathsf{Va}(\widehat{\mu}_{\texttt{GAN1}} - \widehat{\mu}_{\texttt{GAN2}})$$

against the hypothesis that at least one is different from the rest.

Sphericity only applies with more than 2 groups (otherwise, there is a single correlation for the pair).

Mauchly's test of sphericity

Typically, Mauchly's test of sphericity is used to test this assumption

- if statistically significant, use a correction.
- if no evidence, proceed with F tests as usual with $\mathsf{F}(\nu_1, \nu_2)$ benchmark distribution.

Sphericity tests with `afex`

```
summary(model) #truncated output
```

```
Mauchly Tests for Sphericity

          Test statistic p-value
stimulus         0.67814 0.14341
```

- p-value for Mauchly's test is large, no evidence that sphericity is violated.

Corrections for sphericity

If we reject the hypothesis of sphericity (small p-value for Mauchly's test), we need to change our reference distribution.

Box suggested to multiply both degrees of freedom of F statistic by $\epsilon < 1$ and compare to $\mathsf{F}(\epsilon\nu_1, \epsilon\nu_2)$ distribution instead. This yields conservative tests (size greater than α).

- Three common correction factors ϵ:
  - Greenhouse–Geisser
  - Huynh–Feldt (more powerful)
  - take $\epsilon = 1/\nu_1$, giving $\mathsf{F}(1, \nu_2/\nu_1)$.

Another option is to go fully multivariate (MANOVA tests).

Corrections for sphericity tests with `afex`

The estimated corrections $\hat{\epsilon}$ are reported by default with p-values. Use only if sphericity fails to hold, or to check robustness.

```
summary(model) # truncated output
```

```
Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

           GG eps Pr(>F[GG])
stimulus 0.75651     0.5667

             HF eps Pr(>F[HF])
stimulus 0.8514944  0.5872648
```

Note: $\hat{\epsilon}$ can be larger than 1, replace by the upper bound 1 if it happens

Reporting

> We conducted a within-subject one-way ANOVA for the effect of stimulus on latency. Mauchly's test, $p = 0.143$, gives no evidence against the assumption of sphericity. There is no overall effect of GAN, $F(2, 22) = 4.96, p = 0.6155$.

If there were differences in variances, write instead something like

> Mauchly's test shows strong evidence of departure from sphericity. Consequently, we used Huynh-Feldt correction $F(2\hat{\epsilon}, 22\hat{\epsilon}) = 4.96$, $\hat{\epsilon} = 0.85$, adjusted $p = 0.587$.

Contrasts

In within-subject designs, contrasts are obtained by computing the contrast for every subject. Make sure to check degrees of freedom!

```
# Set up contrast vector
cont_vec <- list("real vs GAN" = c(1, -0.5, -0.5))
model |> emmeans::emmeans(spec = "stimulus", contr = cont_vec)
```

```
## $emmeans
##  stimulus emmean    SE df lower.CL upper.CL
##  real      -10.8 0.942 11    -12.8    -8.70
##  GAN1      -10.8 0.651 11    -12.3    -9.40
##  GAN2      -10.3 0.662 11    -11.8    -8.85
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast    estimate    SE df t.ratio p.value
##  real vs GAN   -0.202 0.552 11  -0.366  0.7213
```

Recap

- Repeated measure ANOVA pools data from participants to keep a single measurement (average) per experimental condition.
- It treat individuals as a factor (but no interaction).
  - **within-subject** factors are those which are completed by everyone.
  - **between-subject** factors are mutually exclusive (different individual in those subgroups).
- Correlated data allows us to compare measurements within same individuals:
  - measurements from different individuals are still assumed independent of one another.
  - the more the variability is due to individuals (higher correlation), the smaller the sample size needed for the study.

Recap

- Since observations are correlated, we need to account for the correlation within individuals.
  - We can assume equicorrelation of measurements from same individuals.
  - A weaker statement, sphericity, can be used if we only care about pairwise differences.
  - Sphericity only applies for $J > 2$ groups
  - If sphericity is violated, we need to change our reference benchmark and use a correction, or go fully multivariate.