Repeated measures and MANOVA

Session 9

MATH 80667A: Experimental Design and Statistical Methods
HEC Montréal

Outline

Repeated measures

MANOVA

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 $S_i \sim \textsf{Normal}(0, \sigma_s^2)$ and $\varepsilon_{ij} \sim \textsf{Normal}(0, \sigma_e^2)$ are random variables.

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)

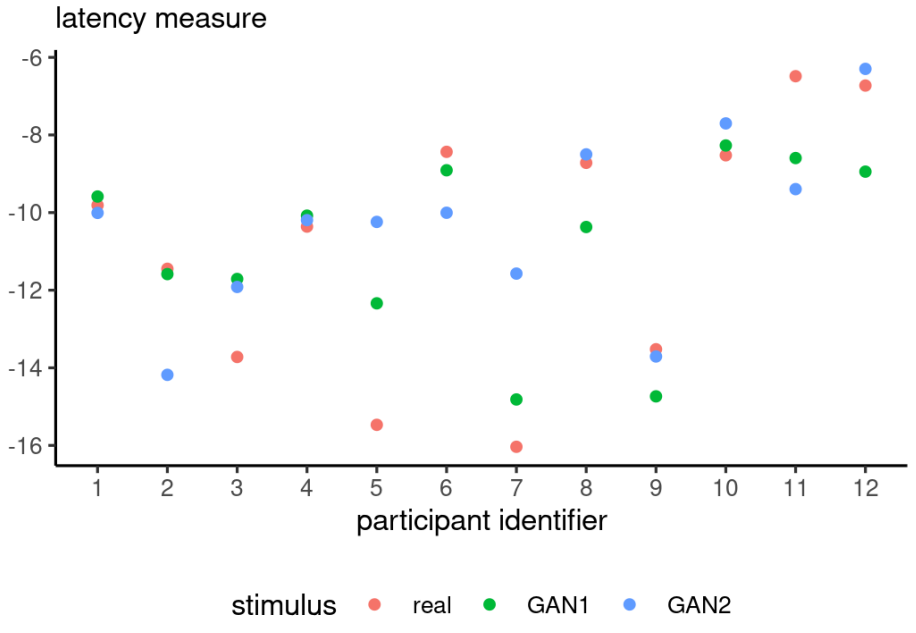
We use the `afex` package to model the within-subject structure.

Load data

```
# 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)
anova(model, # mixed ANOVA model
      correction = "none", # sphericity
      es = "none") # effect size
```

- No detectable difference between conditions.

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

- 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
- equal correlation between measurements of the same subject (*compound symmetry*)
- normality of the random 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.

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

- if statistically significant, use a correction (later)
- 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.
- Report the p-value of the F-test: $F(2, 22) = 0.6155$.

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

- 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

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
```

Multivariate analysis of variance

Motivational example

From Anandarajan et al. (2002), Canadian Accounting Perspective

> This study questions whether the current or proposed Canadian standard of disclosing a going-concern contingency is viewed as equivalent to the standard adopted in the United States by financial statement users. We examined loan officers' perceptions across three different formats

Alternative going-concern reporting formats

Bank loan officers were selected as the appropriate financial statement users for this study.

Experiment was conducted on the user's interpretation of a going-concern contingency when it is provided in one of three disclosure formats:

1. Integrated note (Canadian standard)
2. Stand-alone note (Proposed standard)
3. Stand-alone note plus modified report with explanatory paragraph (standard adopted in US and other countries)

Multivariate response

4. Please circle the pricing you would charge on borrowings under a line of credit *as a spread over your bank's base lending rate ("Prime rate").*

0.25	0.50	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00
3.25		3.50		3.75		4.00		Other	_____	

5. Please circle on the scale shown below your perception of *the ability of the company to service debt.*

LOW ABILITY				**HIGH ABILITY**
1	2	3	4	5

6. Please circle on the scale shown below your perception of the *likelihood that the company can improve its profitability.*

VERY UNLIKELY				**VERY LIKELY**
1	2	3	4	5

Why use MANOVA?

1. Control experimentwise error
   - do a single test instead of J univariate ANOVAs, thereby reducing the type I error
2. Detect differences in combination that would not be found with univariate tests
3. Increase power (context dependent)

Multivariate model

Postulate the following model:

$$\boldsymbol{Y}_{ij} \sim \text{Normal}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), \qquad j = 1, \ldots J$$

Each response $\boldsymbol{Y}_{ij}$ is p-dimensional.

We assume multivariate measurements are independent of one another, with

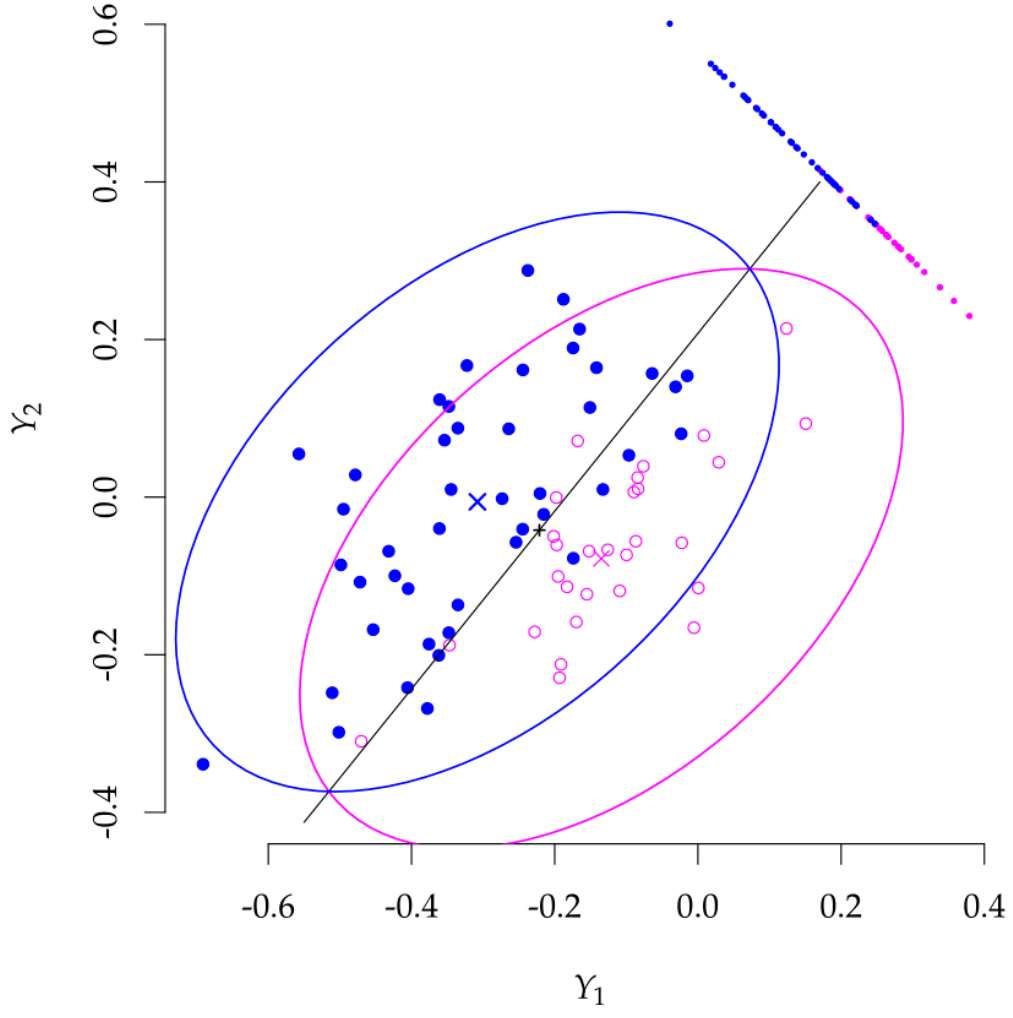
- the same multivariate normal distribution
- same covariance matrix $\boldsymbol{\Sigma}$ (each measurement can have different variance)
- same mean vector $\boldsymbol{\mu}_j$ within each $j = 1, \ldots, J$ experimental groups.

The model is fitted using multivariate linear regression.

In **R**, we fit a model binding the different vectors of response in a matrix with p columns

```
data(AVC02, package = "hecedsm")
# Fit the model binding variables with cbind
# on left of tilde (~) symbol
modMANOVA <- manova(
  cbind(prime, debt, profitability) ~ format,
  data = AVC02)
```

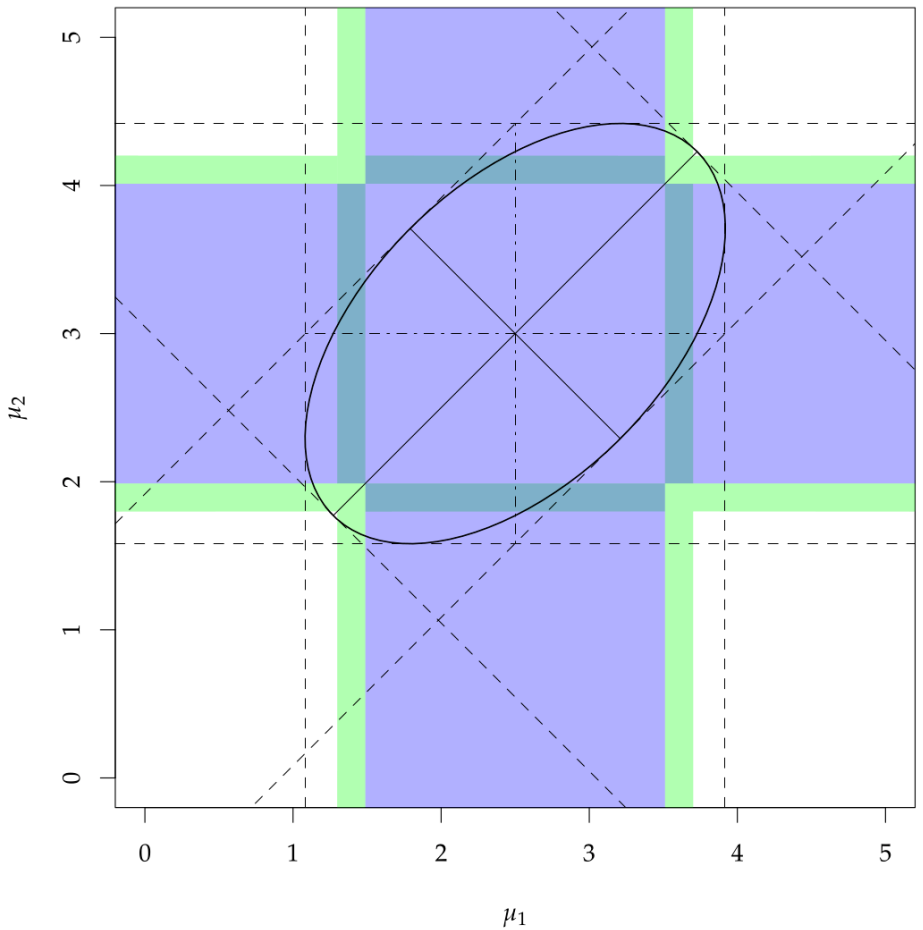
Bivariate MANOVA



Confidence ellipses for bivariate MANOVA with discriminant analysis.

We use the correlation between the p measurements to find better discriminant (the diagonal line is the best separating plane between the two variables).

Confidence intervals and confidence regions



Simultaneous confidence region (ellipse), marginal confidence intervals (blue) and Bonferroni-adjusted intervals (green).

The dashed lines show the univariate projections of the confidence ellipse.

Model assumptions

The more complex the model, the more assumptions...

Same as ANOVA, with in addition

- The response follow a multivariate normal distribution
  - Shapiro–Wilk test, univariate Q-Q plots
- The covariance matrix is the same for all subjects
  - Box's M test is often used, but highly sensitive to departures from the null (other assumptions impact the test)

Normality matters more in small samples (but tests will often lead to rejection, notably because of rounded measurements or Likert scales)

When to use MANOVA?

In addition, for this model to make sense, you need just enough correlation (Goldilock principle)

- if correlation is weak, use univariate analyses
  - (no gain from multivariate approach relative to one-way ANOVAs)
  - less power due to additional covariance parameter estimation
- if correlation is too strong, redundancy
  - don't use Likert scales that measure a similar dimension (rather, consider PLS or factor analysis)

Only combine elements that theoretically or conceptually make sense together.

Testing equality of mean vectors

The null hypothesis is that the J groups have the same mean

- $\mathscr{H}_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_J$ against the alternative that at least one vector is different from the rest.
- The null imposes $(J-1) \times p$ restrictions on the parameters.

The test statistic is Hotelling's T^2 (with associated null distribution), but we can compute using an F distribution.

Choice of test statistic

In higher dimensions, with $J \geq 3$, there are many statistics that can be used to test equality of mean.

The statistics are constructed from within/between sum covariance matrices.

These are

- Roy's largest root (most powerful provided all assumptions hold)
- Wilk's Λ: most powerful, most commonly used
- **Pillai's trace**: most robust choice for departures from normality or equality of covariance matrices

Most give similar conclusion, and they are all equivalent with $J = 2$.

Results for MANOVA

```
summary(modMANOVA) # Pilai is default
```

```
##             Df  Pillai approx F num Df den Df Pr(>F)
## format       2 0.02581  0.55782      6    256 0.7637
## Residuals  129
```

```
summary(modMANOVA, test = "Wilks")
```

```
##             Df   Wilks approx F num Df den Df Pr(>F)
## format       2 0.97424   0.5561      6    254  0.765
## Residuals  129
```

```
summary(modMANOVA, test = "Hotelling-Lawley")
```

```
##             Df Hotelling-Lawley approx F num Df den Df Pr(>F)
## format       2          0.026397  0.55434      6    252 0.7664
## Residuals  129
```

MANOVA for repeated measures

We can also use MANOVA for repeated measures to get away from the hypothesis of equal variance per group or equal correlation

```
model$Anova # for models fitted via 'afex'
```

```
##
## Type III Repeated Measures MANOVA Tests: Pillai test statistic
##             Df test stat approx F num Df den Df    Pr(>F)
## (Intercept)  1   0.95592   238.56      1     11 8.373e-09 ***
## stimulus     1   0.09419     0.52      2     10    0.6098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Less powerful than repeated measures ANOVA because we have to estimate more parameters. Still assumes that the covariance structure is the same for each experimental group.

Follow-up analyses

Researchers often conduct *post hoc* univariate tests using univariate ANOVA. In **R**, Holm-Bonferonni's method is applied for marginal tests (you need to correct for multiple testing!)

```
# Results for univariate analysis of variance (as follow-up)
summary.aov(modMANOVA)
# Note the "rep.meas" as default name
# to get means of each variable separately
emmeans::emmeans(modMANOVA, specs = c("format", "rep.meas"))
```

A better option is to proceed with descriptive discriminant analysis, a method that tries to find the linear combinations of the vector means to discriminate between groups. Beyond the scope of the course.