Repeated measures and multivariate analysis of variance

Session 9

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

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

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 $S_i \sim \text{No}(0, \sigma_s^2)$ and $\varepsilon_{ij} \sim \text{No}(0, \sigma_e^2)$ are random variables.

The errors and random effects are independent from one another.

Variance components

The model **parameters** are μ, α_j's, σ_s^2 and σ_e^2.

- The global average is μ.
- The variance of the response Y_{ij} is $\sigma_s^2 + \sigma_e^2$.
- The **intra-class correlation** between observations in group i is $\sigma_s^2/(\sigma_s^2 + \sigma_e^2)$.
  - 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)

This approach has a drawback: variance estimates can be negative...

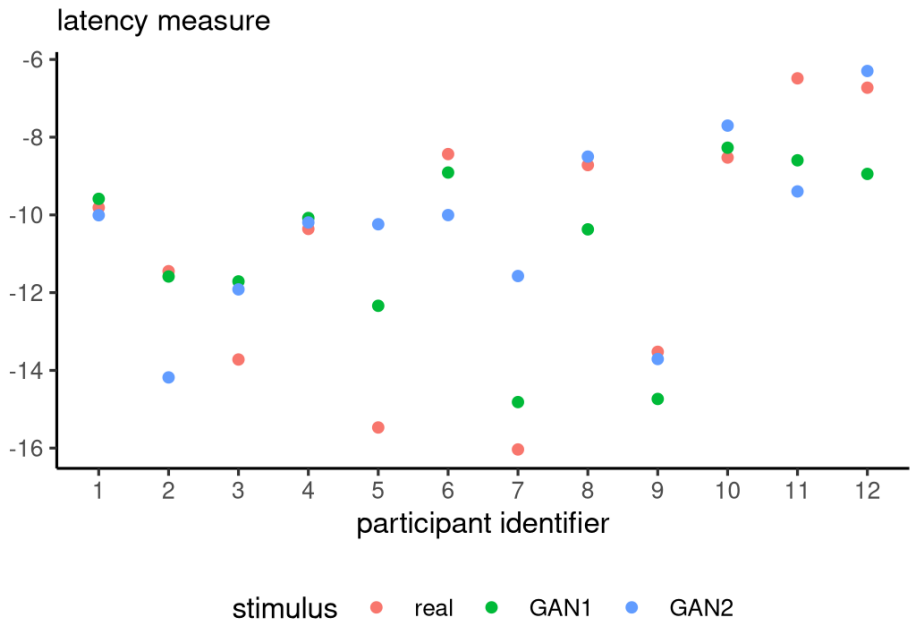
We use the `afex` package to model the within-subject structure.

Load data | Graph | ANOVA

```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))
```

Load data | Graph | ANOVA

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



Load data | Graph | **ANOVA**

```
model <- afex::aov_ez(
  id = "id",            # subject id
  dv = "latency",       # response
  within = "stimulus", # within-subject
  data = AA21,
  fun_aggregate = mean)
anova(model, correction = "none")
```

```
## Anova Table (Type 3 tests)
##
## Response: latency
##           num Df den Df    MSE      F
## stimulus       2     22 1.9557 0.4962 0.00
```

- 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
- 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, people test this assumption (using e.g., Mauchly's test of sphericity)

- if statistically significant, use a correction
- if no evidence, proceed with F tests as usual

Corrections for sphericity

Box suggested to multiply both degrees of freedom of F statistic by $\epsilon < 1$.

- Three common correction factors ϵ:
  - Greenhouse-Geisser
  - Huynh-Feldt (more powerful, but can be larger than 1)
  - lower bound with ν_1, giving $\mathsf{F}(1, \nu_2/\nu_1)$.

Another option is to go fully multivariate.

Sphericity tests with `afex`

```
summary(model)
```

```
Mauchly Tests for Sphericity

          Test statistic p-value
stimulus         0.67814 0.14341

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
```

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, reduces 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:

$$\boldsymbol{Y}_{ij} \sim \text{No}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), \qquad j = 1, \dots J$$

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

We assume multivariate measurements are independent of one another, with

- the same distribution
- same covariance matrix $\boldsymbol{\Sigma}$
- same mean vector $\boldsymbol{\mu}_j$ within each $j = 1, \dots, J$ experimental groups.
  - (randomization)

The model is fitted using multivariate linear regression.

Model assumptions

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

Same as ANOVA, with in addition

- The data follow a multivariate normal distribution
  - Shapiro–Wilk test, univariate QQ-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.

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)
  - less power due to additional covariance parameter estimation
- if correlation is too strong, redundancy
  - don't use Likert scales that measure a similar dimension

Only combine elements that theoretically or conceptually make sense together.

Testing equality of mean vectors

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

With $J = 2$ (bivariate), the MANOVA test finds the best composite score with weights for Y_{i1} and Y_{i2} that maximizes the value of the t-test.

The null distribution is Hotelling's T^2, but a modification of the test statistic can be approximated by a 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 $J = 2$.

Sample size for MANOVA

The number of observations must be sufficiently large.

You can use the software G*Power for power calculations.

To achieve a power of 80%, need the following number of replicates **per group**.

	3 groups				4 groups				5 groups			
effect size \ p	2	4	6	8	2	4	6	8	2	4	6	8
very large	13	16	18	21	14	18	21	23	16	21	24	27
large	26	33	38	42	29	37	44	48	34	44	52	58
medium	44	56	66	72	50	64	74	84	60	76	90	100
small	98	125	145	160	115	145	165	185	135	170	200	230

Laüter, J. (1978), Sample size requirements for the T^2 test of MANOVA (tables for one-way classification). *Biometrical Journal*, **20**, 389-406.

Post hoc testing

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!

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.