MATH 80667A *Experimental Design and Statistical Methods*

Practice final examination

Examiner: Léo Belzile

Instructions: The time allotted for the examination is 180 minutes. You may answer in either English or French. No written material may be brought into the examination, but a simple (non-programmable) calculator may be used.

There are a total of 50 marks available in the exam paper, the distribution of which can be found in the right margin. The number in the box indicates the total per question, the numbers between square brackets the points of each sub-question.

Last name:

First name:

STUDENT ID:

Question:	1	2	3	4	5	Total
Points:	10	6	6	15	13	50
Score:						

Question 1. Short questions **10**

For each question, provide a brief justification, explanation, or counterexample.

1.1 If the variance of the response differs by experimental condition, would the p-value for the F-test be reliable? [2]

> **Solution:** Possibly; the extent to which it is depends on the group heterogeneity. Under the null hypothesis (and equal variance), there is a probability of α of rejecting the null when there is no difference between groups. We could reject more often (by chance) if the variance is heterogeneous, so potentially unreliable.

1.2 Would it be valid to compare differences between treatment level (main effects) if there was an interaction between experimental factor and the covariate (i.e., the covariate is a moderator)? [2]

> **Solution:** No, it would be misleading. We would need to compare the non-parallel slopes at different values of the moderator.

1.3 What is the purpose of including a covariate in an analysis of covariance? [2]

> **Solution:** Explaining part of the variability to reduce the residual error and increase power to detect differences that are due to the experimental manipulation.

1.4 What is the main drawback of online panels (e.g., Amazon MTurk or Prolific)? [2]

> **Solution:** We have no idea of the population of participants on these plateforms, nor guarantees about the truthfullness of answers.

1.5 Should the treatment be modelled as a random effect if we are interested in testing for differences between treatment levels in a linear mixed model? [2]

> **Solution:** No, we normally would use random effects for factors that do not exhaust the population and for whose variability we are interested in (use fixed effects for difference in means).

Question 2. Guidelines on reporting results **6**

The following quote is taken from the *Strategic Management Journal* guidelines and addresses reporting of results of statistical analyses:

> *Authors of submitted papers should not search databases for statistically significant coefficients with the intention of subsequently formulating hypotheses that fit the significant coefficients. Authors also should not adapt experimental designs with the primary intention of producing statistically significant results. In addition, authors of submitted papers should address the material significance (magnitude) of the results, in addition to statistical significance.*

Explain the quote in the context of the replication crisis, addressing

- *post hoc* formulation of hypothesis
- use of 'statistical significance' (e.g., $p < 0.05$) for assessing results
- material relevance of results

Solution:

- Hypotheses should be dictated by (scientific) research question
- Fishing, harking, etc. leads to nonreplicable results and spurious findings. More chance of type I mistake (the more you look, the more you find).
- $p < 0.05$ is arbitrary as a threshold, and p-values should be adjusted to account for multiple testing. There is nothing magical about this cutoff.
- The p-value gets smaller as the sample size increase, regardless of whether differences are practically relevant or not (thus look at effect sizes).

Question 3. Statistical fallacies and the file-drawer problem **6**

The replication crisis is in part due to selective reporting of studies. Indeed, many studies which fail to reach "statistical significance" at the 5% are not published. The purpose of this question is to study the implications of this so-called file-drawer problem.

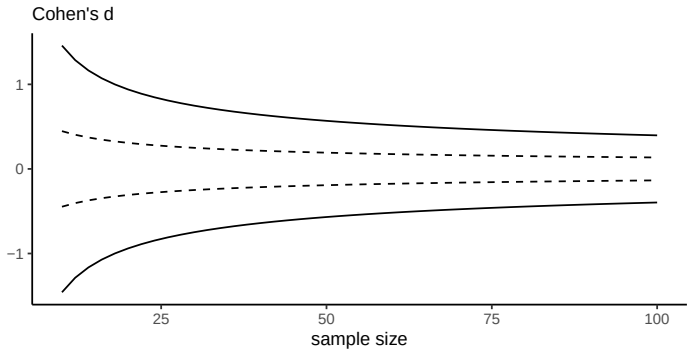


Figure 1: Funnel plot, depicting 95% (full line) and 50% (dashed) lines intervals for the distribution fo Cohen's d effect size as a function of the sample size, assuming balanced samples.

Consider the replication of a published study and sample size calculation to replicate the estimated effect size with a power of 90%.

3.1 What is the impact on power calculations of assuming larger effect size? [2]

Solution: Larger power.

3.2 Many of the studies that fail to replicate have small sample size. Explain how this is problematic, referring to Figure 1. [2]

Solution: Effect size estimates are more noisy in small samples than their counterparts from large-sample studies. Coupled with the file drawer problem, this suggests that studies with small samples which have statistically significant effects have more of a chance of having inflated effect sizes. Thus many are flukes and fail to replicate.

3.3 Would the apply to meta-analysis, which pool the results from multiple published studies? Why or why not? [2]

Solution: This reduces the noise (average less noisy than individual effect sizes), but does not address the bias due to selective reporting.

Question 4. Consistency of product evaluation **15**

A study of Lee and Choi (2019) considered the impact on product evaluations (Y, `prodeval`) of `consistency` of product description, denoted X, a binary factor with two levels, either consistent (0, reference) or inconsistent (1) when the image did not match the description — see Figure 2. Because the authors expect people familiar with the brand (as evidenced by higher `familiarity` scores, Z) to have higher reviews, they included the latter as a covariate in an **analysis of covariance** (ANCOVA).

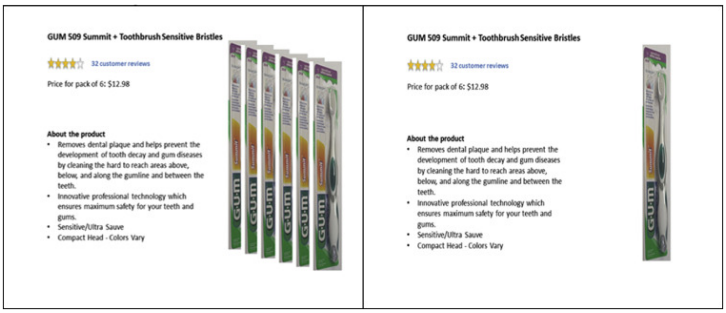


Figure 2: Depiction of consistent (left) and inconsistent (right) description and image for a pack of six toothbrushs.

4.1 Would the model conclusions for the effect of the experimental manipulation be valid if we only fitted an analysis of variance model for Y (`prodeval`) based on X (`consistency`)? [2]

> **Solution:** Yes. ANCOVA is used to increase power, but since X is randomly allocated we can draw conclusions nevertheless.

4.2 You fit a linear regression model with an interaction between `familiarity` and `consistency`. The ANOVA table gives a F value of 0.023 with a p-value of 0.88 for the interaction term. What do you conclude and how does it impact your conclusions for the difference in product evaluations? [2]

> **Solution:** If the interaction is not significant, we fail to reject the null hypothesis that the slopes of Z are the same, and thus that the mean difference for each group is the same for a given value of Z (regardless of the latter).

4.3 The authors report the estimated pairwise difference between groups in Table 2. What do we conclude? [2]

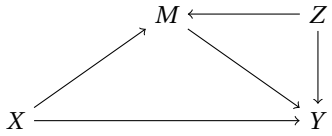
> **Solution:** Products with a consistent description are rated more favorably than inconsistent; the mean difference (std. error) is 0.57 (0.26) and the p-value is 0.03, so we reject the null of no effect at the 5% level.

4.4 They authors also evaluated the potential mediating effect of `fluency` (M) using a linear mediation analysis. The linear models fitted to the data have mean specification [2]

$$\mathsf{E}(\texttt{fluency}) = \alpha_0 + \alpha_1 \texttt{consistency} + \alpha_2 \texttt{familiarity}$$
$$\mathsf{E}(\texttt{respeval}) = \beta_0 + \beta_1 \texttt{consistency} + \beta_2 \texttt{familiarity} + \beta_3 \texttt{fluency}$$

Draw a directed acyclic graph corresponding to the linear mediation analysis. Your diagram should include nodes X, Y, Z and M.

Solution: Z affects both M and Y, but there are no interactions/moderation.



4.5 Comment on the results of the mediation analysis report in Table 4. [2]

Solution: Half of the effect is seemingly due to the pathway $X \to M \to Y$. The indirect effect (ACME) is negative (for higher fluency, the score decreases more for inconsistent (versus baseline consistent). The interval includes zero, but the p-value is very close to 5%

4.6 Can the authors successfully claim mediation considering the study uses an experimental design and randomly allocates experimental condition (`consistency`)? Why or why not? [2]

Solution: No. There is no assessment of whether M causes Y, or if they are due to another variable. No check was conducted to assess model specification, presence of confounders, etc.

4.7 In a follow-up study, Lee and Choi manipulated both image and text description. Table 5 gives the counts of expected outcome for the delivery based on the image (either six or a single toothbrush). Pearson's chi-square test statistic has a value of 2.92, with a p-value of 0.23. [3]

(a) What is the null hypothesis of this test and the conclusion we can draw based on the results?
(b) Report the degrees of freedom of the χ^2 statistic.

Solution: The null hypothesis of 'independence' or lack of interaction. There is no evidence against the null that the expected proportion of items depends on the image (p-value of 0.23). There are two additional parameters in the saturated model, so $\nu = 2$.

term	sum of squares	df	F	p-value
`(Intercept)`	1074.79	1	564.37	$< 10^{-4}$
`familiarity`	9.14	1	4.80	0.03
`consistency`	9.21	1	4.84	0.03
residuals	209.48	110		

Table 1: Analysis of covariance: type III sum of square decomposition

contrast	estimate	std. error	df	statistic	p-value
`consistent` − `inconsistent`	0.57	0.26	110	2.2	0.03

Table 2: t-test for the average difference in `consistency`. The average (std. error) of product evaluations are 7.25 (0.18) for `consistent` and 6.68 (0.18) for `inconsistent`.

term	estimate	std. error	statistic	p-value
(Intercept)	5.70	0.26	22.07	$< 10^{-4}$
`familiarity`	0.08	0.06	1.41	0.16
`consistency`	−0.48	0.24	−2.01	0.05
`fluency`				

estimate	std. error	statistic	p-value
3.32	0.56	5.97	$< 10^{-4}$
0.09	0.05	1.67	0.10
−0.29	0.22	−1.29	0.20
0.60	0.09	6.81	$< 10^{-4}$

Table 3: Coefficients of the linear regression models for the mediation (left) and response model (right).

	estimate	lower	upper	p-value
ACME (indirect effect)	−0.2858	−0.5907	−0.02	0.040
ADE (direct effect)	−0.2873	−0.7330	0.16	0.207
total effect	−0.5731	−1.0861	−0.06	0.028
proportion mediated	0.4987	−0.0373	1.75	0.058

Table 4: Causal mediation analysis: nonparametric bootstrap 95% confidence intervals (percentile method) and p-values obtained with $B = 10,000$ replications

	image	
expected	one	six
not sure	9	13
one	54	44
six	35	45

Table 5: Table of counts for expected response based on image

Question 5. Peace prospects **13**

Study 5 of Halevy and Berson (2022) aimed to demonstrate that events in the distant future rather than the near future influenced the prospect of peace. The experimental design is a

> *2 (current state: war vs. peace) by 2 (predicted outcome: war vs. peace) by 2 (temporal distance: next year vs. twenty years into the future) mixed design*

with current state and predicted outcome as between-subject factors and temporal distance as within-subject factor. The response is the estimated likelihood of each outcome on a 7-point Likert scale ranging from extremely unlikely (1) to extremely likely (7). The question asked was

> *There is currently [war/peace] between the two tribes in Velvetia. Thinking about [next year/in 20 years], how likely is it that there will be [war/peace] in Velvetia?*

	predout	
cstate	peace	war
peace	148	164
war	118	124

Table 6: Repartition of the individuals by subgroup

predout	curstate	tempdist	estimate	std. error
peace	peace	1 yr	5.58	0.17
war	peace	1 yr	3.24	0.16
peace	war	1 yr	2.69	0.19
war	war	1 yr	5.50	0.18
peace	peace	20 yr	4.51	0.16
war	peace	20 yr	4.95	0.16
peace	war	20 yr	5.14	0.18
war	war	20 yr	4.16	0.18

Table 7: Estimated marginal means for each of the eight subgroups

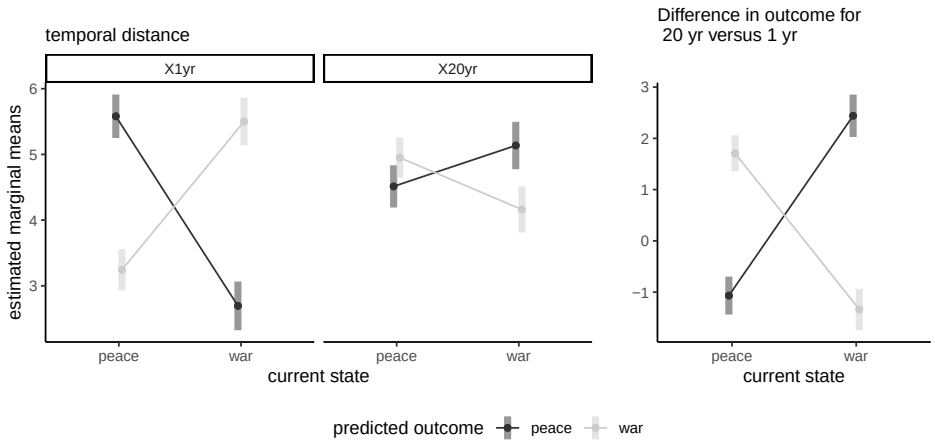


Figure 3: Interaction plots with estimated marginal means and 95% confidence intervals for the $2 \times 2 \times 2$ mixed design (left) and for the 2×2 between design obtained by looking at difference between 20 years and 1 years (right).

term	num. df	den. df	F statistic	p-value
`curstate`	1	273	1.96	0.16
`predout`	1	273	0.01	0.90
`tempdist`	1	273	19.73	$< 10^{-4}$
`curstate:predout`	1	273	42.87	$< 10^{-4}$
`curstate:tempdist`	1	273	1.39	0.24
`predout:tempdist`	1	273	6.56	0.01
`curstate:predout:tempdist`	1	273	279.38	$< 10^{-4}$

Table 8: ANOVA type 3 sum of square decomposition and F statistics for the 2^3 mixed design.

`predout`	`curstate`	estimate	std. error	lower CI	upper CI
peace	peace	−1.07	0.19	−1.44	−0.70
war	peace	1.71	0.18	1.36	2.06
peace	war	2.44	0.21	2.03	2.86
war	war	−1.34	0.21	−1.74	−0.93

Table 9: Estimated marginal means, std. errors and 95% confidence intervals for the individual differences in scores for predicted likelihood for 20 years minus 1 year.

5.1 You wish to marginalize over temporal distance (i.e., averaging the results for 20 year and 1 year for each of the four pairs of predicted outcome/current state). Based on Table 8, is this sensible? [2]

> **Solution:** No, since there is a significant three-way interaction (last row of Table 8), as seen in also in Figure 3. Averaging would lead to misleading conclusions.

5.2 You test for equality of variance in each of the subgroups. Levene's test returns a p-value of 0.013. What do you conclude and how does it affect your conclusions? [2]

> **Solution:** The variance differs in the eight-subgroup, and this may affect the validity of our inference (there is nevertheless overwhelming evidence for the interaction, and this is unlikely to change that conclusion).

5.3 Table 7 gives the estimated marginal means for each of the eight categories. Using the same order as in Table 7, what do the following contrasts test? [2]

$$C_1 : (1,0,0,-1,0,0,0,0); \qquad C_2 : (0,0,0,0,0.5,-0.5,-0.5,0.5)$$

> **Solution:**
> - C_1: status quo in 1 year, for state at peace vs at war
> - C_2: change vs no change in state, in 20 years

We consider next changes in predictions from future (20 years) versus short term (1 year), but computing individual differences. The marginal means for this two-way between subject ANOVA are reported in Table 9. We estimate the variance terms separately for each subcategory to account for potential heterogeneity in responses.

5.5 Is the study design balanced? [1]

> **Solution:** No, the number in each subcategory is unequal as showcased in Table 6.

5.6 Write the vector of contrasts for the weights of each of the four subcategories for testing: [2]

- differences in scores for the status-quo (same predicted outcome as current state).
- difference in outcome if the current state in Velvetia is war.

Use the same order as in Table 9.

> **Solution:** Any non-zero multiple of C_3: $(1,0,0,-1)$ and C_4: $(0,0,1,-1)$.

5.7 Could we use Tukey's honest significant difference to control the family-wise error rate (FWER) for the contrasts? Why or why not? [2]

> **Solution:** No, because these are not pairwise differences. We could use Scheffé's method, or Holm–Bonferroni if the number of contrasts is specified in advance.

5.8 Assume we want to control the FWER at level α using Bonferroni's correction for the contrasts. Circle the correct statement. [2]

(a) When applying Bonferroni's correction, we will reject more null hypotheses relative to the situation with no correction.

(b) The Bonferroni correction consists in testing the individual hypotheses at level αm.

(c) We can only apply the Bonferroni correction if the tests are independent.

(d) The Bonferroni correction can be performed by multiplying the p-values obtained from the individual tests by m and using the same level α.

See the practice exams from Dr. Lukas Meier (ETHZ) for additional examples of data analysis questions (clickable link).