Duration: **130 minutes**
Aids Allowed: **Non-programmable calculator**

Student Number: |_|_|_|_|_|_|_|_|_|_|_|_|

Family Name(s): _____

Given Name(s): _____

Lecture Section: ☐ STA303H1
☐ STA1002H1

---

*Do **not** turn this page until you have received the signal to start.*
In the meantime, please read the instructions below *carefully*.
Please write your name on the front **and** the back of the midterm test.

This term test consists of 8 questions on 30 pages (including this one), printed on both sides of the paper. *When you receive the signal to start, please make sure that your copy of the test is complete, fill in the identification section above, and write your name on the back of the last page.*

Answer each question directly on the test paper, in the space provided, and use the reverse side of the pages for rough work. If you need more space for one of your solutions, use the reverse side of a page and *indicate clearly the part of your work that should be marked.*

Write up your solutions carefully! Even where they are not required, clear and concise explanations of what you're trying to achieve may help us mark your answers, and part marks *might* be given for showing that you understand how to approach the problem, even if your solution is incomplete.

MARKING GUIDE

1: _____/ 25
2: _____/ 10
3: _____/ 15
4: _____/ 10
5: _____/ 10
6: _____/ 10
7: _____/ 10
8: _____/ 10

TOTAL: _____/100

Good Luck!

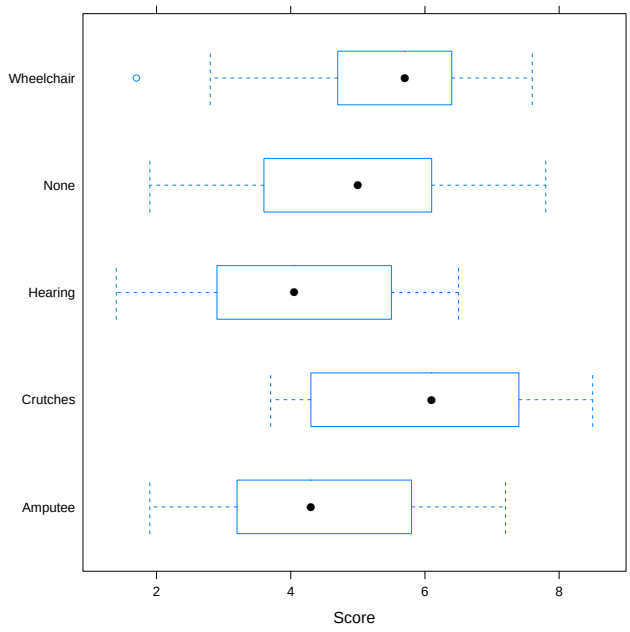
OVER...

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 1. [25 MARKS]

An experiment was conducted to determine whether interviewers display bias against persons with disabilities. An actor recorded 5 different videos of a job interview, with identical scripts, except that in the first video, he appeared in a wheelchair; in the second video, he appeared on crutches; in the third video, he appeared hearing-impaired; in the fourth video, he appeared to have one leg amputated; and in the fifth video, he appeared to have no disabilties.

For each video, 14 different undergraduate students watched it, and rated the job applicant on his qualifications (70 students in total participated in the experiment, each providing one score/rating.) The scores are summarized in a boxplot below.



Part (a) [5 MARKS]

What are the conditions under which we can perform a One-Way ANOVA F-test? List all the conditions, state whether they seem to be satisfied for the discrimination experiment data based on the boxplot, and briefly state why. If you cannot determine whether a condition was satisfied, say so and briefly explain why.

Use this page for rough work—clearly indicate any section(s) to be marked.

Here is some R output for the discrimination experiment

```
> lm(Score~Handicap, data=disc)

Call:
lm(formula = Score ~ Handicap, data = disc)

Coefficients:
       (Intercept)     HandicapAmputee    HandicapCrutches     HandicapHearing
            4.9000             -0.4714              1.0214             -0.8500
HandicapWheelchair
            0.4429
```

Part (b) [3 MARKS]

Write down the formula for predicting the rating for a new video of an interview, where it is known what disability (if any) the actor in the video appears to have. Define all variables.

Part (c) [2 MARKS]

What was the average rating for the videos where it appeared that the actor is an amputee? Show your work.

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (d) [10 MARKS]

Here is some R output for the discrimination experiment, with some of the values omitted

```
> anova(lm(Score~Handicap, data=disc))
Analysis of Variance Table

Response: Score
          Df  Sum Sq Mean Sq F value  Pr(>F)
Handicap   A  30.521  B        C          D      *
Residuals  E 173.321  G
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Compute all the omitted values. You may use functions such as `qt`, `qf`, `pf`, `pnorm`, etc. where necessary; if you use them, you don't have to provide a numerical final answer. Briefly show how you obtained the answers.

A =

B =

C =

D =

E =

G =

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (e) [5 MARKS]

What conclusion can you draw from the F-test for which the p-value is computed in the ANOVA table? (Note the asterisk in the table, which indicates that the p-value was smaller than 0.05.) Be precise, and state the conclusion without referring to the model – just use English.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 2. [10 MARKS]

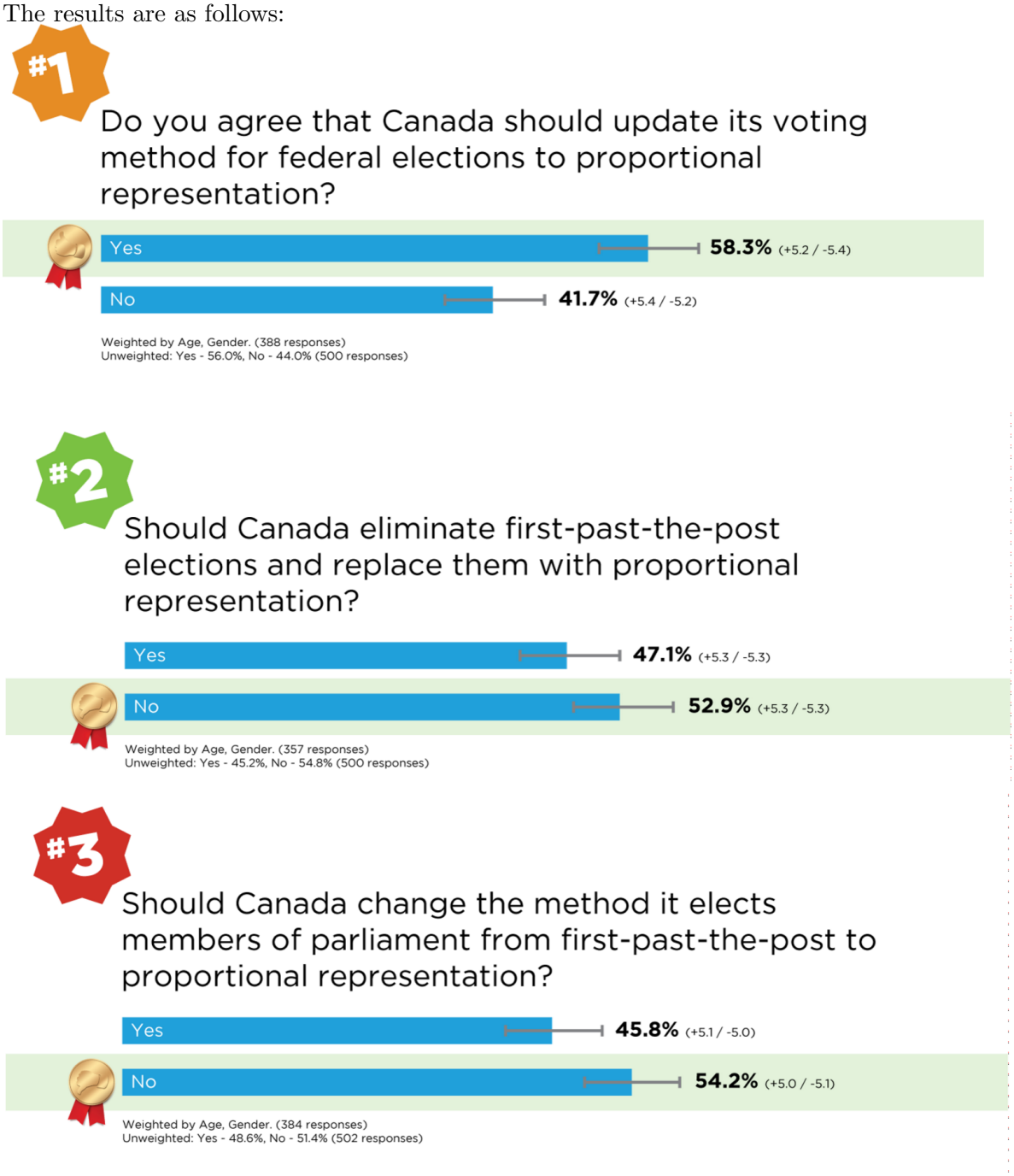
Suppose the experiment described in Question 1 is repeated. State one way in which the ANOVA assumptions may be violated for the data obtained, and provide a plausible scenario (i.e., a story about the actor, raters, etc.) which would lead to the violation of the ANOVA assumptions.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 3. [15 MARKS]

The following is a description of a survey that was recently conducted by `prvote.com`: "We used Google Consumer Surveys to poll 1500 respondents on one of three different phrasings of a possible referendum question, resulting in 500 responses to each question."

The results are as follows:



Use this page for rough work—clearly indicate any section(s) to be marked.

For the purposes of the following questions, you can ignore the fact that the results were weighted by age and gender.

Part (a) [5 MARKS]

Error bars (and the precise numerical size of the bars) are shown in the figure on the previous page. Do they represent 70% or 95% confidence intervals? Explain precisely how you arrived at the answer based on the numbers shown in the graph.

Circle one: 70% CI 95% CI

Part (b) [10 MARKS]

After conducting the polls, `prvote.com` analyzed the data, and claimed that Question 1 is different from Question 3, and Questions 2 is different from Question 3 (i.e., the probability of saying "yes" was different for the different questions). At 95% confidence, state precisely how you would evaluate the claims **based on the graph and numbers on the previous page** using only R functions such as `pnorm`, `qnorm`, etc. Your answer need not be a valid R program, but it should be very clear how to evaluate the claim in R.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 4. [10 MARKS]

Part (a) [3 MARKS]

Using only `rnorm`, write R code to generate a sample from $\chi^2(3)$.

Part (b) [7 MARKS]

Suppose that you observe the measurements $X_1, X_2, ..., X_{10}$. Assume that $X_i \sim N(\mu, \sigma^2)$ and that the observations are independent. Write R code to test the Null Hypothesis that $\sigma = 3$.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 5. [10 MARKS]

Suppose that $Y_1, Y_2, ..., Y_{10}$ are i.i.d and $Y_i \sim Bernoulli(\theta)$.

Part (a) [2 MARKS]

What is the likelihood function of θ given $Y_1, Y_2, ..., Y_{10}$?

Part (b) [8 MARKS]

Prove that the Maximum Likelihood Estimate of θ is $\bar{Y}$.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 6. [10 MARKS]

Explain how leave-one-out cross-validation is used in order to select the best model for the data. In your answer, describe at least two cost functions, and state precisely how to compute them and use them in the context of leave-one-out cross-validation.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 7. [10 MARKS]

Recall that in the Pygmalion Effect dataset, we had platoons, each of which belonged to a company, and each of which either had or had not the Pygmalion Effect treatment. In a context of a dataset similar to the Pygmalion Effect dataset, describe a scenario (i.e., a story about the data) where the MSR (Mean Square Regression) is expected to be larger than the MSE (Mean Square Error). You may use any reasonable simplification (for example, you can assume that there are only two companies.) Write code to randomly generate a dataset where you expect the MSR to be larger than the MSE. The ANOVA assumptions must still hold.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 8. [10 MARKS]

Recall that in the Titanic dataset, we predict the survival based on characteristics like sex and class. Recall that the log-odds in Logistic Regression are defined as

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_1^{(i)} + ... + \beta_k x_k^{(i)}$$

Consider the R output below:

```
> m1 <- glm(survived~sex, family=binomial, data=titanic)
> m1

Call:  glm(formula = survived ~ sex, family = binomial, data = titanic)

Coefficients:
(Intercept)      sexmale
      1.112       -2.467

Degrees of Freedom: 1045 Total (i.e. Null);  1044 Residual
Null Deviance:     1415
Residual Deviance: 1102  AIC: 1106
> m2 <- glm(survived~sex+pclass, family=binomial, data=titanic)
> m2

Call:  glm(formula = survived ~ sex + pclass, family = binomial, data = titanic)

Coefficients:
(Intercept)      sexmale       pclass
     3.0043      -2.5278      -0.8575

Degrees of Freedom: 1045 Total (i.e. Null);  1043 Residual
Null Deviance:     1415
Residual Deviance: 1014  AIC: 1020
```

Part (a) [2 MARKS]

Without using `anova` or similar functions, write R code to perform a Likelihood Ratio Test to determine whether pclass is useful for predicting survival. Make it as clear as possible how you would use the output of the code to draw concusions.

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (b) [3 MARKS]

What is the interpretation of the coefficient that corresponds to pclass?

Part (c) [5 MARKS]

There are 658 men and 388 women in the dataset. How many people in the dataset survived? Show your work.

Use this page for rough work—clearly indicate any section(s) to be marked.

Additional page for answers

On this page, please write nothing except your name.

Family Name(s): _____

Given Name(s): _____