STAT4006 Final Exam 2020-21, 9th December 2020, 1730-2000

Before you begin, you **MUST**

- download the attached spreadsheet named "STAT4006 Final Exam 2020-21 Tables"
- type in your student ID number (all 10 digits) into cell D2. The values in Table 1 and Table 2 will change.

Failure to enter your student ID number in cell D2 will result incorrect answers. Double check you have entered it correctly. Now you can proceed to answer these questions, in any order you like.

Q1. (35%) Refer to **Table 1** in the spreadsheet named "STAT4006 Final Exam 2020-21 Tables". Assume the data has come from a Poisson sampling method

a) Invent where the data has come from and explain why it is indeed an example of Poisson sampling.
b) Test whether the Poisson distribution is a good fit for the data in Table 1.
c) What is overdispersion? Does the data in Table 1 demonstrate overdispersion? Why might it occur? How and why should you handle overdispersion if you observe it in your count data?

Q2. (40%) A group of people were asked "Which is the best football team in North London?" with possible responses "Tottenham" and "Arsenal". The subjects were also asked their gender (variable G: "male" or "female"), favourite colour (variable C: "white", "red", "other") and favourite style of music (variable M: "classical", "punk", "experimental jazz", "other").

The model for π, the probability of response "Tottenham" is

$$logit(\pi) = \alpha + \beta_h^G + \beta_i^C + \beta_j^M + \beta_{ij}^{CM}$$

Refer to **Table 2** in the spreadsheet named "STAT4006 Final Exam 2020-21 Tables. You will see

- The coding for each level of the explanatory variables
- The parameter estimates

Please answer the following questions:

a) What is the loglinear model that is equivalent to this model?
b) Interpret how responding "Arsenal" depends on gender.
c) Identify which group
    i. Is least likely to respond "Tottenham"
    ii. Is most likely to respond "Tottenham"
d) Find the estimated probabilities of a "Tottenham" response for each of the groups you gave in c)i) and c)ii).
e) If we changed the redundancy of the explanatory dummy variables so that the sums of parameters across categories were zero, then what would the $\hat{\beta}_j^M$ equal?
f) Label the response variable F. State the equivalent loglinear and logit models for which
    i. F is jointly independent of G, C and M
    ii. There are main effects of M on F, but F is conditionally independent of C and G, given M

Q3. (25%) Create the toughest "True or False" question you can think of. The topic of the question must be something we have covered in the course. You should:

- Provide the correct answer;
- Explain why it is correct; and
- Explain why you think the question is so difficult.

END OF QUESTIONS