STAT305: Introduction to Inference **Fall 2021-22**

Assignment 5

Due date: see Piazza post.

Instructions

Academic integrity policy: We encourage you to discuss verbally with other students about the assignment. However, you should write your answers by yourself (in particular, without ever putting your eyes on someone else's answer or draft answer). If you use online resources (including asking questions on math exchange or similar), you *have* to cite them. If you have not done se already, read:
`http://learningcommons.ubc.ca/guide-to-academic-integrity/` Please also make sure to read the audit process described during lecture 1.

Instructions: read this right away!

1. Due date: refer to the Piazza post announcing the assignment.
2. Justify all your answers. Use precise mathematical notation. Define all symbols you introduce. Marks will be deducted for imprecise, incorrect or insufficient justification.
3. Write your answers in order, and clearly label the question you are answering, for example 1.1 for part 1 of question 1.
4. We only accept submissions via Gradescope. Please make sure at least 24 hours in advance that you are confortable with the submission process, to give us enough time to fix any problems with the submission process.
5. See slide in lecture 1 for late assignment policy.

1 Generalized likelihood ratio test - contingency tables!

You surveyed $n = 150$ students to see which colour of red, green and blue they preferred, and which day they like better, Monday or Friday. You obtained results as follows:

	Monday	Friday
Red	3	47
Green	5	45
Blue	17	33

Suppose the k^{th} student falls into row r column c with probability $\theta_{r,c}$, i.e., let $P(R_k = r, C_k = c) = \theta_{r,c}$ for row $r = 1, 2, 3$ and column $c = 1, 2$, where $\sum_{r,c} \theta_{r,c} = 1$.

Are the two preferences dependent? More formally, your null hypothesis for independence is $H_0 : \theta_{r,c} = \phi_r \rho_c$, where both ϕ_r and ρ_c are in $(0, 1)$ for all r, c, representing the probability a student falls into row r and column c respectively. Note $\sum_r \phi_r = \sum_c \rho_c = 1$.

1. The likelihood is $L(\theta) = n! \prod_{r,c} \frac{\theta_{r,c}^{n_{r,c}}}{n_{r,c}!}$, where $n_{r,c}$ is the number of responses for row r column c. Write down the log likelihood.

[2 mark(s)]

2. Show the maximum likelihood estimates are

$$\hat{\phi}_r = \frac{n_{r\cdot}}{n} \tag{1}$$
$$\hat{\rho}_c = \frac{n_{\cdot c}}{n} \tag{2}$$

under the null hypothesis, where $n_{r\cdot}$ is the row sum of row r, and $n_{\cdot c}$ is the column sum of column c. **Hint:** use Lagrange mutlipliers; MLE for multinomial!

[6 mark(s)]

3. Justify why the dimension of the parameter space (number of free parameters) for H_0 is 3.

[2 mark(s)]

4. Consider the alternative hypothesis H_1, where preferences are dependent, i.e., the parameters are $\theta_{r,c}$ for $r = 1, 2, 3$ and $c = 1, 2$. The MLE under H_1 is $\hat{\theta}_{r,c} = \frac{n_{r,c}}{n}$. Justify why the dimension of the parameter space for H_1 is 5.

[2 mark(s)]

5. Compute Wilk's statistic for this set of data. Should you reject the null hypothesis at $\alpha = 0.05$?

[2 mark(s)]

Question total: [14 mark(s)]