STAT 961: Homework 5

Name

Due Friday, December 10 at 11:59pm

1 Instructions

Setup. Pull the latest version of this assignment from Github and set your working directory to `stat-961-fall-2021/homework/homework-5`. Consult the getting started guide if you need to brush up on `R`, `LaTeX`, or `Git`.

Collaboration. The collaboration policy is as stated on the Syllabus:

> "Students are permitted to work together on homework assignments, but solutions must be written up and submitted individually. Students must disclose any sources of assistance they received; furthermore, they are prohibited from verbatim copying from any source and from consulting solutions to problems that may be available online and/or from past iterations of the course."

In accordance with this policy,

Please list anyone you discussed this homework with:

Please list what external references you consulted (e.g. articles, books, or websites):

Writeup. Use this document as a starting point for your writeup, adding your solutions between `\begin{sol}` and `\end{sol}`. See the preparing reports guide for guidance on compilation, creation of figures and tables, and presentation quality. Show all the code you wrote to produce your numerical results, and include complete derivations typeset in LaTeX for the mathematical questions.

Programming. The `tidyverse` paradigm for data manipulation (`dplyr`) and plotting (`ggplot2`) are strongly encouraged, but points will not be deducted for using base `R`.

```
library(tidyverse)
```

Grading. Each sub-part of each problem will be worth 3 points: 0 points for no solution or completely wrong solution; 1 point for some progress; 2 points for a mostly correct solution; 3 points for a complete and correct solution modulo small flaws. The presentation quality of the solution for each problem (as exemplified by the guidelines in Section 3 of the preparing reports guide) will be evaluated out of an additional 3 points.

Submission. Compile your writeup to PDF and submit to Gradescope.

Problem 1. Conditional independence testing in $J \times K \times L$ tables.

Suppose that

$$(x_1, x_2, x_3) \in \{0, \dots, J-1\} \times \{0, \dots, K-1\} \times \{0, \dots, L-1\}$$

are jointly distributed discrete random variables, with

$$\mathbb{P}[x_1 = j, x_2 = k, x_3 = l] = \pi_{jkl}, \quad \text{where} \quad \sum_{j,k,l} \pi_{jkl} = 1.$$

In this problem, we will explore tests of the conditional independence null hypothesis

$$H_0 : x_1 \perp\!\!\!\perp x_2 \mid x_3$$

based on data

$$y_{jkl} \overset{\text{ind}}{\sim} \text{Poi}(\mu_{jkl}); \quad \mu_{jkl} = \mu_0 \pi_{jkl}.$$

(a) Suppose we parametrize

$$\log \mu_{jkl} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13} + \beta_{kl}^{23} + \beta_{jkl}^{123},$$

constraining

$$\beta_0^1 = \beta_0^2 = \beta_0^3 = \beta_{j0}^{12} = \beta_{0k}^{12} = \beta_{j0}^{13} = \beta_{0l}^{13} = \beta_{k0}^{23} = \beta_{0l}^{23} = \beta_{jk0}^{123} = \beta_{j0l}^{123} = \beta_{0kl}^{123} = 0.$$

Show that

$$H_0 : x_1 \perp\!\!\!\perp x_2 \mid x_3 \quad \Longleftrightarrow \quad H_0 : \beta_{jk}^{12} = \beta_{jkl}^{123} = 0 \text{ for all } j, k, l. \tag{1}$$

(b) What are the maximum likelihood estimates $\widehat{\mu}_{jkl}$ under H_0? What is the relationship between these estimates and the maximum likelihood estimates under independence in $J \times K$ tables? Intuitively, why is this the case?

(c) What is the Pearson X^2 statistic for testing H_0, and what is its asymptotic null distribution? How do this statistic and its asymptotic null distribution relate to the corresponding quantities for testing independence in $J \times K$ tables?

The test derived in part (c) is asymptotic, and may not be well-calibrated when the counts in the table are small. Ideally we would find a way to exactly calibrate the test statistic in finite samples. Maybe we can generalize Fisher's exact test to this more complicated setting? Let's try to do so in parts (d) and (e) below.

(d) Let's first recall Fisher's exact test for $H_0 : x_1 \perp\!\!\!\perp x_2$, where $J = K = 2$. Letting y_{jk} ($j, k \in \{0, 1\}$) be the counts in the 2×2 contingency table (Figure 1a), Fisher's exact test is based on the null hypergeometric distribution

$$\mathbb{P}[y_{11} = t \mid y_{11} + y_{01} = v, y_{11} + y_{10} = m_1, y_{00} + y_{01} = m_0] = \frac{\binom{m_1}{t}\binom{m_0}{v-t}}{\binom{m_1+m_0}{v}}. \tag{2}$$

A different way of obtaining the null distribution of y_{11} is by considering all permutations of the x_2 column in Figure 1b (the long-format equivalent of Figure 1a). Prove that the permutation-based null distribution of y_{11} is the same as the hypergeometric null distribution (2).

(e) For general $J \times K$ tables, both the test statistic and its sampling distribution under the null are harder to work with analytically. Nevertheless, propose an exact permutation-based test inspired by the result in part (d) to generalize Fisher's exact test to $J \times K$ tables. Returning to the $J \times K \times L$ problem, generalize the above test to test conditional independence. This exact test is called the *conditional permutation test*.

Solution 1.

	$x_2 = 1$	$x_2 = 0$	total
$x_1 = 1$	y_{11}	y_{10}	m_1
$x_1 = 0$	y_{01}	y_{00}	m_0
total	v	$m_1 + m_0 - v$	$m_1 + m_2$

(a) 2×2 table format

x_1	x_2
0	0
...	...
0	0
0	1
...	...
0	1
1	0
...	...
1	0
1	1
...	...
1	1

(b) Long format

Figure 1: Two representations of cross-tabulated binary data.

Problem 2. Testing for association between income and job satisfaction, given gender.

Consider the job satisfaction data below, which cross-tabulate income, job satisfaction, and gender:

```
job_satisfaction = read_tsv("../../data/job_satisfaction.tsv")
print(job_satisfaction, n = 5)

## # A tibble: 32 x 4
##   Income      Job.Satisfaction   Gender Count
##   <chr>       <chr>              <chr>  <dbl>
## 1 <5000       Very Dissatisfied  Female     1
## 2 5000-15000  Very Dissatisfied  Female     2
## 3 15000-25000 Very Dissatisfied  Female     0
## 4 >25000      Very Dissatisfied  Female     0
## 5 <5000       A Little Satisfied Female     3
## # ... with 27 more rows
```

We'd like to test whether there is a relationship between income and job satisfaction, conditional on gender.

(a) Create a plot to visualize the relationship between income and job satisfaction for males and females, making sure to respect the natural orderings of the income and job satisfaction variables. Comment on the trends you observe in this plot.

(b) Implement the test from Problem 1c on the `job_satisfaction` data. You may use the `glm()` function but not more specialized functions for conditional independence testing. What p-value do you obtain, and what is the corresponding conclusion?

(c) Why may the test implemented in part (b) be underpowered? Propose a test statistic that may be able to more sensitively pick up the relationship between income and job satisfaction.

(d) Note that the conditional permutation test from Problem 1e can be applied to calibrate *any* test statistic, not just the Pearson X^2 statistic. Implement the conditional permutation test to calibrate the statistic you proposed in part (c). What is the resulting p-value? What is your conclusion about the relationship between income and job satisfaction, controlling for gender?

(e) Suppose you applied a regular permutation test to search for association between income and job satisfaction, ignoring the gender variable. Would this result in a valid p-value for testing independence between income and job satisfaction, conditional on gender? Why or why not?

Solution 2.

Problem 3. Bradley-Terry model for the NBA.

The NBA has 30 teams, and each team plays a total of 82 games during the regular season. If the game between team j and team k takes place in team j's arena, then team j is said to be the "home" team and team k is said to be the "away" team. Suppose that each team $j \in \{1, \dots 30\}$ has an associated parameter β_j representing how good the team is, and β_0 is a parameter representing "home court advantage." Then, a simple model for the outcome of the match between team j and team k is

$$\text{logit}(\mathbb{P}[\text{home team } j \text{ beats away team } k]) = \beta_0 + \beta_j - \beta_k. \tag{3}$$

This model is called the *Bradley-Terry model*. In this problem, we'll be fitting the Bradley-Terry model to data from the NBA 2017-2018 season:

```
nba_data = read_tsv("../../data/nba_data.tsv")
print(nba_data, n = 5)

## # A tibble: 2,460 x 5
##    game team  conference location result
##   <dbl> <chr> <chr>      <chr>    <chr>
## 1     1 BOS   East       Away     Loss
## 2     1 CLE   East       Home     Win
## 3     2 HOU   West       Away     Win
## 4     2 GS    West       Home     Loss
## 5     3 CHA   East       Away     Loss
## # ... with 2,455 more rows
```

(a) (Identifiability) This model suffers from an identifiability issue. State the issue and how you could restrict the parameters to resolve this issue. A more subtle identifiability issue would arise if the teams could be split into two groups such that games were played only within groups. Discuss why the latter issue would arise and check whether this issue is a concern in the given data.

(b) (Model fitting) Reformulate the Bradley-Terry model as an ungrouped logistic regression model (i.e. what are the predictors and the response)? Based on this reformulation, transform the data into the format expected by `glm()`, print the resulting tibble (no need for a kable table), and then call `glm()` on this tibble to obtain fitted model parameters.

(c) (Home court advantage) Produce a point estimate and Wald confidence interval for the factor by which the odds of winning increase with home court advantage. Comment on the direction and magnitude of the effect. Produce a scatter plot of home game win percentage versus away game win percentage (this plot should contain 30 points, one per team), and comment on what this plot says about home court advantage.

(d) (Team rankings) Produce a table with 30 rows and four columns: team, win percentage for home games, win percentage for away games, and estimated value of β_j. Order the columns in descending order of $\widehat{\beta}_j$. Comment on the degree of concordance or discordance between the columns in this table. Intuitively, what might be a reason why a team has a relatively higher win percentage but a relatively lower value of $\widehat{\beta}_j$?

(e) (NBA finals prediction) The NBA finals in 2017-2018 were between the Golden State Warriors (GS) and the Cleveland Cavaliers (CLE). The first two games were in Oracle Arena (Oakland,

California) while the last two games were in Quicken Loans Arena (Cleveland, Ohio). Golden State won all four of these games and the NBA championship. For each of these four games, what are point estimates and Wald confidence intervals for the probability that Golden State won?

Solution 3.