BIST5615 Homework 1

Due date: Sunday November 3, 2024

General Instructions

- Use this R Markdown template for homework submission.
- Answer the questions by inserting R code and necessary comments if applicable. Your output must contain the R code (do not use the `echo=FALSE` option) if applicable.
- Save the compiled PDF file under the file name `LastName-FirstName-HW1.pdf` and submit it through HuskyCT by the deadline.

Max Points: 35 points

1. For a 3×3 table, please answer the following questions.

(a) (5 points) Assume that

$$\begin{aligned} P(X=1|Y=1)=0.5,\ P(X=2|Y=1)=0.5,\ P(X=3|Y=1)=0, \\ P(X=1|Y=2)=0,\ P(X=2|Y=2)=0.5,\ P(X=3|Y=2)=0.5, \\ P(X=1|Y=3)=0.5,\ P(X=2|Y=2)=0,\ P(X=3|Y=3)=0.5; \end{aligned}$$

and

$$\begin{aligned} P(Y=1|X=1)=0.5,\ P(Y=2|X=1)=0.5,\ P(Y=3|X=1)=0, \\ P(Y=1|X=2)=0,\ P(Y=2|X=2)=0.5,\ P(Y=3|X=2)=0.5, \\ P(Y=1|X=3)=0.5,\ P(Y=2|X=3)=0,\ P(Y=3|X=3)=0.5. \end{aligned}$$

Can you uniquely determine the joint distribution of X and Y? Please provide the reason if not and find it if so.

(b) (5 points) Assume that

$$\begin{aligned} P(X=1|Y=1)=0.5,\ P(X=2|Y=1)=0.5,\ P(X=3|Y=1)=0, \\ P(X=1|Y=2)=0,\ P(X=2|Y=2)=0.5,\ P(X=3|Y=2)=0.5, \\ P(X=1|Y=3)=0.3,\ P(X=2|Y=3)=0.3,\ P(X=3|Y=3)=0.4; \end{aligned}$$

and

$$\begin{aligned} P(Y=1|X=1)=0.5,\ P(Y=2|X=1)=0.5,\ P(Y=3|X=1)=0, \\ P(Y=1|X=2)=0,\ P(Y=2|X=2)=0.5,\ P(Y=3|X=2)=0.5, \\ P(Y=1|X=3)=0.5,\ P(Y=2|X=3)=0,\ P(Y=3|X=3)=0.5. \end{aligned}$$

Can you uniquely determine the joint distribution of X and Y? Please provide the reason if not and find it if so.

Ans.

a) 0/0 creates the problem. Not necessarrily, it will be always 1. For uniquely getting the joint we see,

$$\nexists x^* \text{ such that } P(X = x^*|Y = y) > 0 \text{ for all } y = 1, 2, 3.$$

Let's say X and Y take values in finite sets M and N, and that their joint mass function $\mathbb{P}[X = x, Y = y]$ is > 0 for all $(x, y) \in M \times N$. Then: Now,

$$\mathbb{P}[X = x \mid Y = y]/\mathbb{P}[Y = y \mid X = x] = \mathbb{P}[X = x]/\mathbb{P}[Y = y],$$

for all $(x, y) \in M \times N$. Summing over x, we find that:

$$1 \cdot \mathbb{P}[Y = y] = \sum_{x \in M} \mathbb{P}[X = x \mid Y = y]/\mathbb{P}[Y = y \mid X = x]$$

is completely determined by the conditional distributions. Hence, so is $\mathbb{P}[Y = y]$ determined. Finally, $\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x \mid Y = y]/\mathbb{P}[Y = y]$ is also determined by the conditional distributions.

with this logic,

for b)

$P(X = x, Y = y)$	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	0	0	$\frac{3}{16}$
$X = 2$	0	$\frac{3}{16}$	$\frac{3}{16}$
$X = 3$	0	$\frac{3}{16}$	$\frac{4}{16}$

2. The sinking of the Titanic is a famous event, and new books are still being published about it. Many well-known facts–from the proportions of first-class passengers to the "women and children first" policy, and the fact that that policy was not entirely successful in saving the women and children in the third class–are reflected in the survival

rates for various classes of passenger. These data were originally collected by the British Board of Trade in their investigation of the sinking. The file titanicdat.csv contains the raw data (2201 observations, 4 variables). The file titanicdat.sas contains both the data and SAS codes. The file titanic.txt is a documentation file containing a brief description of the dataset. These files are posted on the Husky class website.

Data file: 'titanic

```r
hw1q2 <-  read.csv("titanicdat.csv")
str(hw1q2)
```

```
## ’data.frame’:    2201 obs. of  4 variables:
##  $ class   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sex     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survival: int  1 1 1 1 1 1 1 1 1 1 ...
```

```
table(hw1q2)
```

```
## , , sex = 0, survival = 0
##
##      age
## class   0   1
##     0   0   3
##     1   0   4
##     2   0  13
##     3  17  89
##
## , , sex = 1, survival = 0
##
##      age
## class   0   1
##     0   0 670
##     1   0 118
##     2   0 154
##     3  35 387
##
## , , sex = 0, survival = 1
##
##      age
## class   0   1
##     0   0  20
##     1   1 140
##     2  13  80
##     3  14  76
##
## , , sex = 1, survival = 1
##
##      age
## class   0   1
##     0   0 192
##     1   5  57
##     2  11  14
##     3  13  75
```

(a) (5 points) For adults who sailed on the Titanic, use the raw data to reproduce the odds of survival for females was 11.4 times that of males as well as the odds of survival for females equaled 2.9 discussed in Lecture Note 3.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
adult_female_surv <- hw1q2 %>% filter(age==1,sex==0) %>% select(survival)%>% table()
```

```
adult_female_surv
```

```
## survival
##   0   1
## 109 316
```

```
(adult_female_dead <- hw1q2 %>% filter(age==1,sex==0) %>% select(survival)
%>% table() %>% .[1])
```

```
##   0
## 109
```

```
(adult_male_surv <- hw1q2 %>% filter(age==1,sex==1) %>% select(survival)
%>% table() %>% .[2])
```

```
##   1
## 338
```

```
(adult_male_dead <- hw1q2 %>% filter(age==1,sex==1) %>% select(survival)
%>% table() %>% .[1])
```

```
##    0
## 1329
```

The odds of adult female survival is given by

$$\frac{\text{\#adult females survived}}{\text{\#adult females died}} = \frac{\text{adult\_female\_surv}}{\text{adult\_female\_dead}} = \frac{316}{109} = \text{round}\left(\frac{316}{109}, 1\right) = 2.9$$

The odds of adult male survival is given by

$$\frac{\text{\#adult males survived}}{\text{\#adult males died}} = \frac{\text{adult\_male\_surv}}{\text{adult\_male\_dead}} = \frac{338}{1329} = \text{round}\left(\frac{338}{1329}, 1\right) = 0.3$$

Therefore, the odds ratio of survival for females vs. males is given by

$$\frac{\text{\#adult females survived}}{\text{\#adult females died}}, \frac{\text{\#adult males survived}}{\text{\#adult males died}} = \frac{\text{adult\_female\_surv}}{\text{adult\_female\_dead}}, \frac{\text{adult\_male\_surv}}{\text{adult\_male\_dead}} = \frac{316}{109} \frac{338}{1329} =$$

The last answer is 11.4.

(b) (5 points) For adults who sailed on the Titanic, (i) compute the 95% exact confidence interval of the odds ratio of survival for females versus males and (ii) compute a 95% asymptotic confidence interval and the 95% exact confidence interval of γ for gender and survival status.

```r
# Create the contingency table
titanic_data <- matrix(c(316, 109, 338, 1329), nrow = 2, byrow = TRUE)

# Perform Fisher's exact test to get the odds ratio and exact CI
fisher.test(titanic_data)$conf.int
```

```
## [1]  8.831768 14.744632
## attr(,"conf.level")
## [1] 0.95
```

```r
# Calculate the standard error of the log-odds ratio
se_log_or <- sqrt(1/316 + 1/109 + 1/338 + 1/1329)

# Calculate the log-odds ratio
log_or <- log(11.38)

# Calculate the asymptotic CI for the log-odds ratio
```

```r
ci_lower <- log_or - 1.96 * se_log_or
ci_upper <- log_or + 1.96 * se_log_or

# Convert back to the odds ratio scale by exponentiating
exp(ci_lower)
```

```
## [1] 8.877754
```

```r
exp(ci_upper)
```

```
## [1] 14.58752
```

(c) (5 points) Using the entire raw data, determine whether the death rate for the third class passengers was much higher than for the others including the crew members.

```r
# Data: number of deaths and total for each class
deaths <- c(528, 120, 166, 685)
total <- c(709, 320, 285, 899)

# Proportion test comparing third-class death rate to other groups
prop.test(deaths, total)
```

```
##
##  4-sample test for equality of proportions without continuity correction
##
## data:  deaths out of total
## X-squared = 189.77, df = 3, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4
## 0.7447109 0.3750000 0.5824561 0.7619577
```

If the p-value is small (e.g., <0.05), you would conclude that there is a statistically significant difference in death rates, meaning the death rate for third-class passengers is likely much higher than for others.

(d) (5 points) Using the entire raw data, determine whether the death rate for females was much less than for males.

```r
# Calculate the number of deaths and total number of passengers by gender
death_counts <- hw1q2 %>%
    group_by(sex) %>%
    summarise(Deaths = sum(survival == 0, na.rm = TRUE),  # 0 indicates death
              Total = n()) %>%
    mutate(DeathRate = Deaths / Total)

# View death counts and rates
print(death_counts)
```

```
## # A tibble: 2 x 4
##     sex Deaths Total DeathRate
##   <int>  <int> <int>     <dbl>
## 1     0    126   470     0.268
## 2     1   1364  1731     0.788
```

```r
# Extract values for the test
male_deaths <- death_counts$Deaths[death_counts$sex == 1]
female_deaths <- death_counts$Deaths[death_counts$sex == 0]
male_total <- death_counts$Total[death_counts$sex == 1]
female_total <- death_counts$Total[death_counts$sex == 0]

# Calculate pooled proportion
pooled_prop <- (male_deaths + female_deaths) / (male_total + female_total)

# Calculate the test statistic
z <- (male_deaths / male_total - female_deaths / female_total) /
    sqrt(pooled_prop * (1 - pooled_prop) * (1/male_total + 1/female_total))

# Calculate the p-value
p_value <- pnorm(z)

# Display results
cat("Z-value:", z, "\n")
```

```
## Z-value: 21.37461
```

```r
cat("P-value:", 1-p_value, "\n")
```

```
## P-value: 0
```

So we reject null.The conclusion is death rate for females was statistically significantly lower than that for males. (e) (5 points) Using the entire raw data, determine whether the "women and children first" policy was successful in saving the women and children.

```r
titanic_data <- hw1q2 %>%
  mutate(AgeGroup = ifelse(age < 18, "Child", ifelse(sex == 0, "Woman", "Man")))

# Calculate survival rates for Women, Children, and Men
survival_summary <- titanic_data %>%
  group_by(AgeGroup) %>%
  summarise(Survived = sum(survival == 1, na.rm = TRUE),  # 1 indicates survival
            Total = n()) %>%
  mutate(SurvivalRate = Survived / Total)

# View survival rates
print(survival_summary)
```

```
## # A tibble: 1 x 4
##   AgeGroup Survived Total SurvivalRate
##   <chr>       <int> <int>        <dbl>
## 1 Child         711  2201        0.323
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
```

```r
# Perform Chi-square test
# Create a contingency table
contingency_table <- titanic_data %>%
  group_by(AgeGroup, survival) %>%
  summarise(Count = n()) %>%
  pivot_wider(names_from = survival, values_from = Count, values_fill = 0)
```

```
## ‘summarise()‘ has grouped output by ’AgeGroup’. You can override using the
## ‘.groups‘ argument.
```

```r
# Perform Chi-square test
chi_square_test <- chisq.test(contingency_table[, -1])  # Exclude AgeGroup column

# Display the test result
cat("Chi-square statistic:", chi_square_test$statistic, "\n")
```

```
## Chi-square statistic: 275.7115
```

```r
cat("P-value:", chi_square_test$p.value, "\n")
```

```
## P-value: 6.458938e-62
```

```r
# Interpretation
alpha <- 0.05
if (chi_square_test$p.value < alpha) {
    cat("Reject the null hypothesis: There is a significant difference in survival rates
} else {
    cat("Fail to reject the null hypothesis: No significant difference in survival rates
}
```

```
## Reject the null hypothesis: There is a significant difference in survival rates among
```