Statistics Homework 4

109550121 溫柏萱

Problem 1

(a)

```
> data <- matrix(c(68, 56, 91, 40, 5, 6, 61, 59),
+                nrow = 4,
+                byrow = TRUE,
+                dimnames = list(c("Accounting", "Administration",
"Economics", "Finance"),
+                                c("Female", "Male")))
> data
               Female Male
Accounting         68   56
Administration     91   40
Economics           5    6
Finance            61   59
> chi_square_test <- chisq.test(data)
Warning message:
In chisq.test(data) : Chi-squared approximation may be incorrect
> chi_square_test

  Pearson's Chi-squared test

data:  data
X-squared = 10.827, df = 3, p-value = 0.0127
```

The null hypothesis in this case is that there is no relationship between the gender of students and their choice of major. The p-value for gender and choice of major is $0.0127 < 0.05$, therefore, the null hypothesis can be rejected, indicating that there exist relations between gender and the choice for major.

(b)

```
> chi_square_test$expected
                  Female      Male
Accounting     72.279793 51.720207
Administration 76.360104 54.639896
Economics       6.411917  4.588083
Finance        69.948187 50.051813
>
```

The Chi-square method can be used when the following conditions are satisfied.

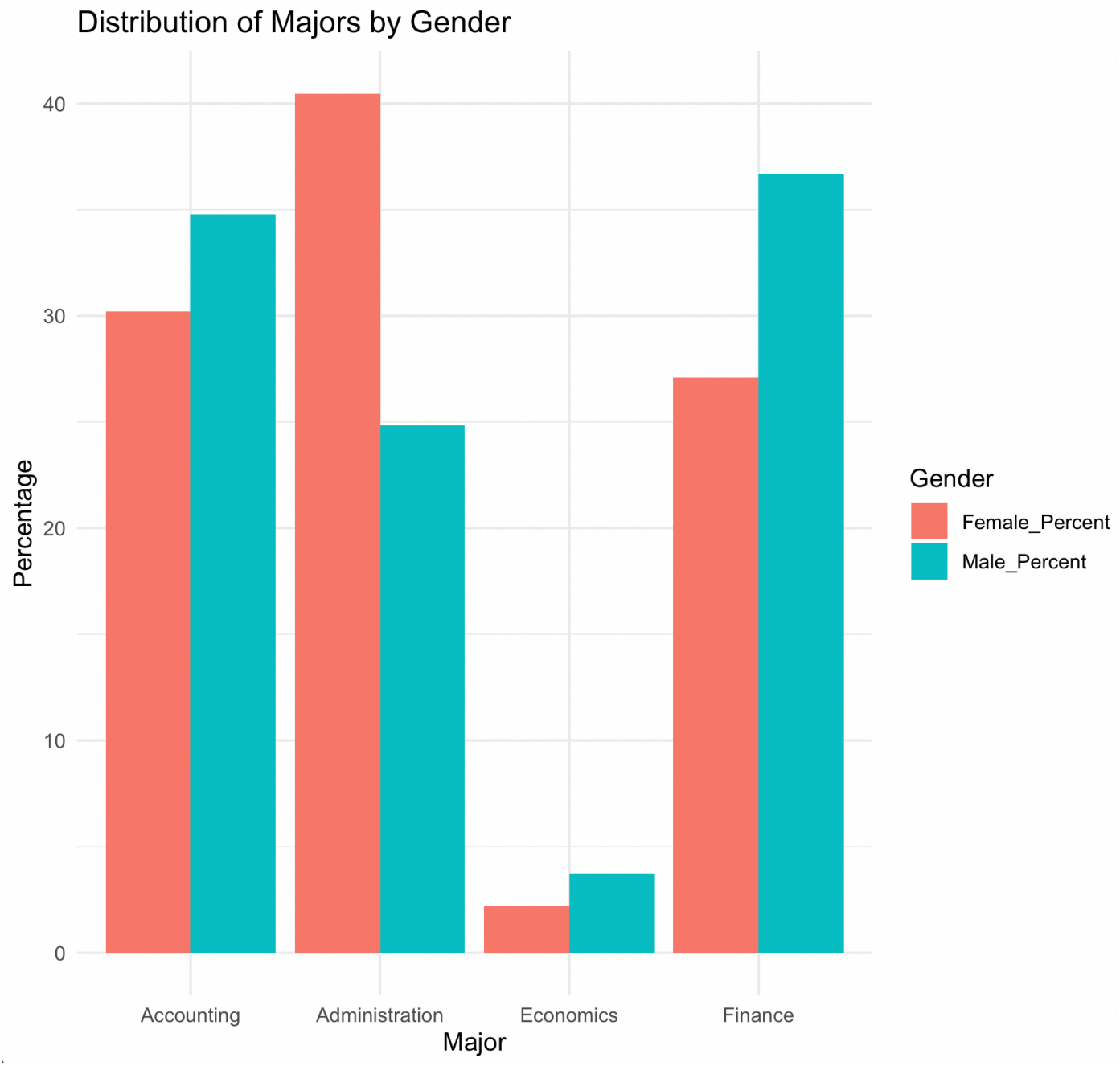
1. There are no empty cells
2. Over 80% cells have expectation greater than 5

There are 8 cells in total, where 7 cells have expectation greater than 5, thus 87.5% cells have expectation greater than 5, the Chi-square method can be used.

(c)

```
> data_df <- as.data.frame(data)
> data_df$Female_Percent <- (data_df$Female / sum(data_df$Female)) * 100
> data_df$Male_Percent <- (data_df$Male / sum(data_df$Male)) * 100
> data_df
               Female Male Female_Percent Male_Percent
Accounting         68   56      30.222222    34.782609
Administration     91   40      40.444444    24.844720
Economics           5    6       2.222222     3.726708
Finance            61   59      27.111111    36.645963
> data_df$Major <- rownames(data_df)
> data_df
               Female Male Female_Percent Male_Percent          Major
Accounting         68   56      30.222222    34.782609     Accounting
Administration     91   40      40.444444    24.844720 Administration
Economics           5    6       2.222222     3.726708      Economics
Finance            61   59      27.111111    36.645963        Finance
> data_melted <- melt(data_df, id.vars = "Major", measure.vars =
c("Female_Percent", "Male_Percent"), variable.name = "Gender", value.name
= "Percentage")
```

```
> ggplot(data_melted, aes(x = Major, y = Percentage, fill = Gender)) +
+   geom_bar(stat = "identity", position = "dodge") +
+   labs(title = "Distribution of Majors by Gender",
+        x = "Major",
+        y = "Percentage",
+        fill = "Gender") +
+   theme_minimal()
>
```



In Accounting, both genders have a significant percentage of students, with 30% female chosing the major and 34.7% male , in Administration, females are significantly greater than male. Both genders have low percentages in Economics with male being slightly higher, and male is significantly greater than female in Finance.

Accounting and Finance are popular among both genders comparing to least-favored choices such as Economics, Administration is particularly favored by females, and Economics is the least chosen major by both.

	Accounting	Administration	Economics	Finance
Female Distribution	30.2%	40.4%	2.2%	27.1%
Male Distribution	34.8%	24.8%	3.7%	36.6%

(d)

```
> observed <- chi_square_test$observed
> expected <- chi_square_test$expected
> contribution <- (observed - expected)^2 / expected
> contribution
                  Female      Male
Accounting     0.2534128 0.3541483
Administration 2.8067873 3.9225288
Economics      0.3109070 0.4344974
Finance        1.1447050 1.5997431
```

The two cells with the largest contributions to the chi-square statistic are Administration (Male) with contribution 3.9225 and Administration (Female) with contribution 2.8068.

Cell	Expected	Observed
Administration (Male)	54.639896	40
Administration (Female)	76.360104	91

The observed number for male is significantly lower than expectation, while the observed number for female is significantly higher than expected, strengthening the conclusion that there is a significant relationship between gender and choice of major and the use of Chi-square method is adequate.

(e)

```
> total_responses <- 68 + 56 + 91 + 40 + 5 + 6 + 61 + 59
> total_students <- 722
> non_responses <- total_students - total_responses
> non_response_percent <- (non_responses / total_students) * 100
> non_response_percent
[1] 46.5374
```

46.53% of the students did not respond to the questionnaire. This high non-response rate weakens the conclusions drawn from the data as the sample may not be fully representative of the entire student population.

Problem 2

(a), (b)

```
> combined_data <- matrix(c(490, 210, 280, 220),
+                         nrow = 2,
+                         byrow = TRUE,
+                         dimnames = list(c("Male", "Female"),
+                                         c("Admit", "Deny")))
> combined_df <- as.data.frame(combined_data)
> combined_df
       Admit Deny
Male     490  210
Female   280  220
```

Gender	Admit	Deny
Male	490	210
Female	280	220

Gender	Acceptence Rate
Male	$\frac{490}{490+210} = 70\%$
Female	$\frac{280}{280+220} = 56\%$

Wabash Tech admits more male applicants than female applicants.

(c)

School	Gender	Acceptence Rate
Business	Male	$\frac{480}{480+120} = 80\%$
	Female	$\frac{180}{180+20} = 90\%$
Law	Male	$\frac{10}{90+10} = 10\%$
	Female	$\frac{100}{100+200} = 33.33\%$

Each school admits a higher percentage of female applicants than male applicants.

(d)

Because the number of male applicants is significantly greater than the number of female applicants, even if female applicants' acceptance rate is higher, it could not overcome the large number of male applicants, also, more males apply to the Business School, which has a higher admission rate, and more females apply to the Law School, which has a lower admission rate, the overall male admission rate appears higher. Therefore, combining different groups can lead to misleading conclusions when not analyzed carefully.

Problem 3

(a)

```
> prof_grades <- c(A = 22, B = 38, C = 20, D_F = 11)
> total_prof_students <- sum(prof_grades)
> prof_percentages <- (prof_grades / total_prof_students) * 100
> prof_percentages
       A        B        C      D_F
24.17582 41.75824 21.97802 12.08791
> ta_probabilities <- c(A = 0.32, B = 0.41, C = 0.20, D_F = 0.07) * 100
> ta_probabilities
  A   B   C D_F
 32  41  20   7
> comparison <- data.frame(Grade = names(prof_percentages),
+                          Professor = prof_percentages,
+                          TA = ta_probabilities)
> comparison
    Grade Professor TA
A       A  24.17582 32
```

```
B        B  41.75824 41
C        C  21.97802 20
D_F    D_F  12.08791  7
```

Grade	Percentage
A	24.17582%
B	41.75824%
C	21.97802%
D	12.08791%

The professor assigned less number of A's than the TAs, the number of B's and C's assigned by the professor and the TAs are similar, the professor fails more students than the TAs.

(b)

```
> prof_grades <- c(A = 22, B = 38, C = 20, D_F = 11)
> total_prof_students <- sum(prof_grades)  # 91
> ta_probabilities <- c(A = 0.32, B = 0.41, C = 0.20, D_F = 0.07)
> expected_counts <- total_prof_students * ta_probabilities
> expected_counts
    A     B     C   D_F
29.12 37.31 18.20  6.37
>
```

If the professor follows the TA distribution, the number of count is expected is as follows.

Grade	Count
A	29
B	37
C	18
D	6

(c)

```
> chi_square_test <- chisq.test(prof_grades, p = ta_probabilities,
rescale.p = TRUE)
> chi_square_test

  Chi-squared test for given probabilities

data:  prof_grades
X-squared = 5.297, df = 3, p-value = 0.1513
```

The null hypothesis assumes that the professor and TAs follow the same distribution.

The p-value $0.1513 > 0.05$, we fail to reject the null hypothesis. There is not enough evidence to suggest that the professor follows a different grade distribution from the TAs. The professor's grade distribution does not differ significantly from the TA distribution based on the given data. Since H_0 cannot be rejected, the professor and TAs follow the same distribution.