Homework 4
Categorical Data Analysis

1. A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at a university. One question asked which major within the business program the student had chosen. Here are the data from the students who responded:

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

This is an example of a single sample classified according to two categorical variables (gender and major).

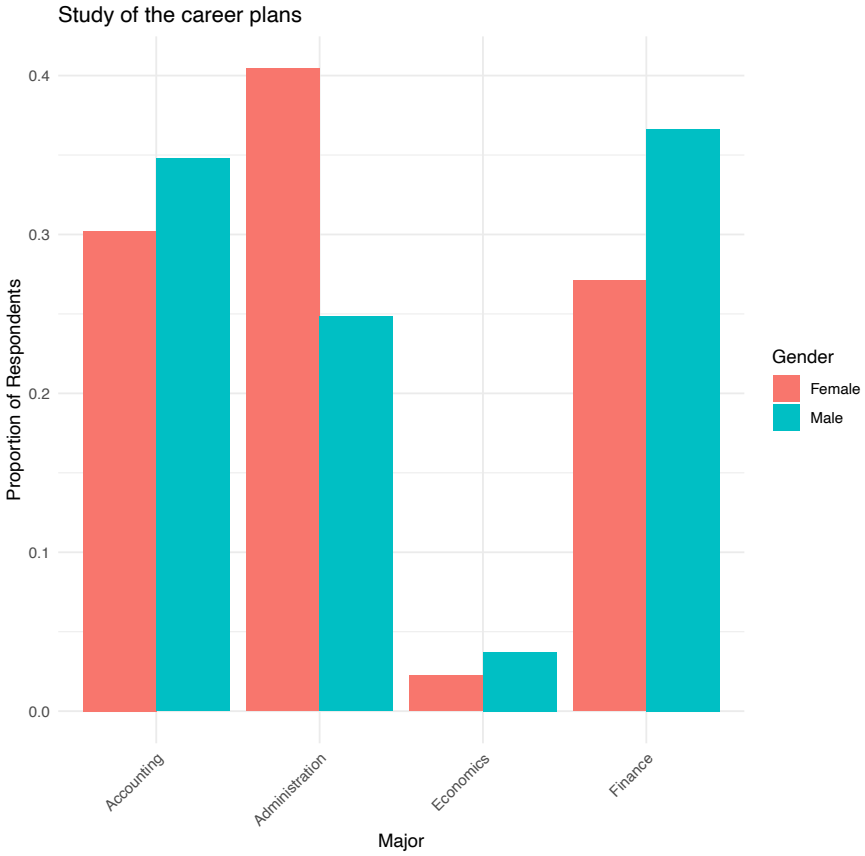
a. Test the null hypothesis that there is no relationship between the gender of students and their choice of major. Give a P-value and state your conclusion.
Ans：根據結果，p-value 為 0.0127 小於顯著水準 0.05，因此可以拒絕 null hypothesis，性別和專業之間是存在關係且不獨立。

```
> print(data_test)

        Pearson's Chi-squared test

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

b. Verify that the expected cell counts satisfy the requirement for use of chi-square.
Ans： 可以看到超過 80%的 cell counts 是超過 5 的，滿足使用 chi-square 的要求。

```
> print(data_expected)
                Female      Male
Accounting      72.279793 51.720207
Administration 76.360104 54.639896
Economics       6.411917  4.588083
Finance        69.948187 50.051813
```

c. Describe the differences between the distributions of majors for women and men with percents, with a graph, and in words.
Ans： 可以看出女性在 Administration 的百分比是唯一超過男性的，而男女比例差 Administration > Finance > Accounting > Economics。



d. Which two cells have the largest terms of the chi-square statistic? How do the observed and expected counts differ in these cells? (This should strengthen your conclusions in (b).)

Ans : 最大的兩個 cells 是在 Administration 列，Female 在 Administration 中，實際觀察值超過期望值但在其它都低於期望值，而 Male 則是在 Administration 中，實際觀察值低於期望值但其他都超過期望值。

```
> result = data_test$observed - data_test$expected
> result
                Female       Male
Accounting     -4.279793   4.279793
Administration 14.639896 -14.639896
Economics      -1.411917   1.411917
Finance        -8.948187   8.948187
```

e. What percent of the students did not respond to the questionnaire? The nonresponse weakens conclusions drawn from these data.

Ans : 共有約 47%沒有參與問卷，這樣會弱化結果。

```
> respondents = sum(rowSums(data))
> total = 722
> percent_nonrespondents = ( 1 - respondents / total ) * 100
> paste("Percent of nonrespondents: ", percent_nonrespondents, "%")
[1] "Percent of nonrespondents:  46.5373961218837 %"
```

2. Wabash Tech has two professional schools, business and law. Here are two-way tables of applicants to both schools, categorized by gender and admission decision. (Although these data are made up, similar situations occur in reality.)

	Admit	Deny
Male	480	120
Female	180	20

Business School

	Admit	Deny
Male	10	90
Female	100	200

Law School

a. Make a two-way table of gender by admission decision for the two professional schools together by summing entries in these tables.
Ans :

```
> data
       Admit Deny
Male     490  210
Female   280  220
```

b. From the two-way table, calculate the percent of male applicants who are admitted and the percent of female applicants who are admitted. Wabash admits a higher percent of male applicants.
Ans : 由結果可以看出 Wabash 的男性的錄取率高於女性。

```
> cat("Male/Female admit rate: ", male_admit*100, "% /", female_admit*100, "%")
Male/Female admit rate:  70 % / 56 %
```

c. Now compute separately the percents of male and female applicants admitted by the business school and by the law school. Each school admits a higher percent of female applicants.
Ans : 由結果可以看出無論是商學院還是法學院，女性的錄取率都高於男性。

```
> cat("Business Male/Female admit rate: ", Business_male_admit*100, "% /", Business_female_admit*1
00, "%")
Business Male/Female admit rate:  80 % / 90 %
> cat("Law Male/Female admit rate: ", Law_male_admit*100, "% /", Law_female_admit*100, "%")
Law Male/Female admit rate:  10 % / 33.33333 %
```

d. This is Simpson's paradox: both schools admit a higher percent of the women who apply, but overall Wabash admits a lower percent of female applicants than of male applicants. Explain carefully, as if speaking to a skeptical reporter, how it

can happen that Wabash appears to favor males when each school individually favors females.

Ans : 由整體來看男性較易錄取，但分別觀察兩個子群體時卻是女性較易錄取，這可能是因為女性傾向申請錄取率低的學院，所以整體分析會顯示男性較易錄取，而女性整體的錄取率偏低。

3. Most students in a large statistics course are taught by teaching assistants (TAs). One section is taught by the course supervisor, a senior professor. The distribution of grades for the hundreds of students taught by TAs this semester was

Grade	A	B	C	D/F
Probability	0.32	0.41	0.20	0.07

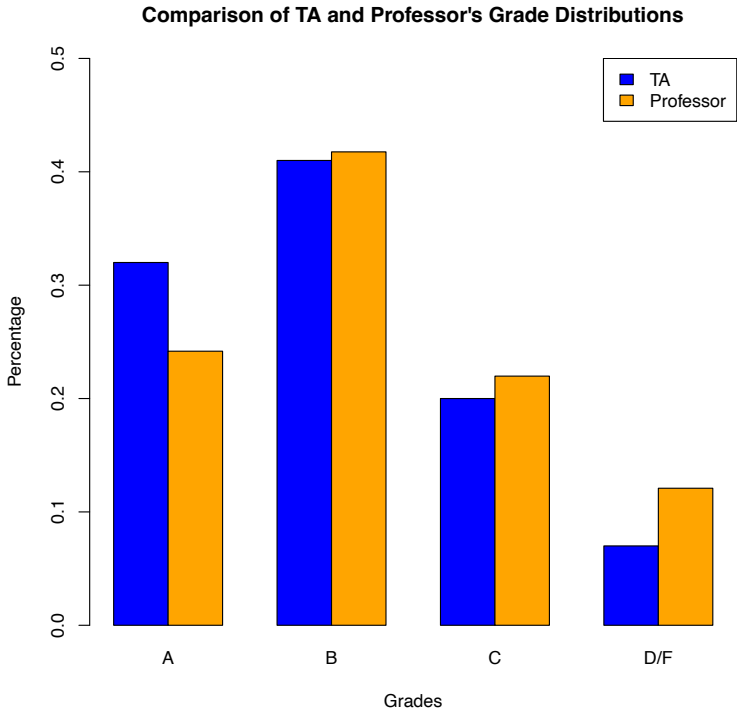
The grades assigned by the professor to students in his section were

Grade	A	B	C	D/F
Count	22	38	20	11

(These data are real. We won't say when and where, but the professor was not the author of this book.)

a. What percents of each grade did students in the professor's section earn? In what ways does this distribution of grades differ from the TA distribution?
Ans： 可以看到在 A 級分的 professor 低於 TA，其餘級分比例皆高於 TA。

```
Professor's percents A: 24.17582 % B: 41.75824 % C: 21.97802 % D/F: 12.08791 %
```



b. Because the TA distribution is based on hundreds of students, we are willing to regard it as a fixed probability distribution. If the professor's grading follows this distribution, what are the expected counts of each grade in his section?
Ans： 以下是 professor’s expected counts of each grade。

A : 29.12, B : 37.31, C : 18.20 and D/F : 6.37。

```
> expected_count = Probability * total_student
> expected_count
[1] 29.12 37.31 18.20  6.37
```

c. Does the chi-square test for goodness of fit give good evidence that the professor follows a different grade distribution? (Give the test statistic, its P-value, and your conclusion.)

Ans : 假設為教授的評分分佈與 TA 的分佈相同，由下圖結果由於 p-value 為 0.1513 大於顯著水準 0.05，所以不能拒絕假設，表示教授與 TA 的成績分佈相同。

```
> chisq.test(Count, p=Probability)

        Chi-squared test for given probabilities

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