Homework 2

Requirement

- To be submitted individually. You are allowed to discuss with your friend, but should finish the tasks on your own. Acknowledge your friend if he/she has provided substantial help to you.
- You can finish the homework in either Chinese or English.
- For programming exercises, you can print out the code, results and figures.
- Hand in to the instructor when you attend the lecture, on or before 6 Apr 2023. Late submissions will be penalized by a 20% deduction in score.

Exercises

Exercise 1:

In the lecture, we learned the generalized likelihood ratio test. A useful test statistic is $G^2 = -2\log(L_0/L_1)$, where L_0 is the maximum likelihood of the model to test (with parameter space ω_0), and L_1 is the maximum likelihood of a bigger model (with parameter space Θ'). An important theorem states that G^2 follows a chi-squared distribution (with degree of freedom $\dim\Theta' - \dim\omega_0$) in the large sample limit.

Problem:

Write down some example applications of this theorem.

Exercise 2:

Enrolment rate (升学率) is a common criterion for evaluating high schools. Suppose high school A has an enrolment rate 45%, while high school B has an enrolment rate 40%.

Problem:

In terms of the enrolment rate, can one simply conclude that high school A is better than B?

(Hint: different school can have different proportions of students studying different subjects, like science, arts and sports.)

Exercise 3:

Job Satisfaction Data

Below there are data from a large scale investigation of blue collar workers in Denmark (1968), which include three (binary) categorical variables:

- Worker job satisfaction (Low, High),

		Bad Management			Good Management			
		Worker's satisfaction				Worker's satisfaction		
		Low	High			Low	High	
Supervisor's satisfaction	Low	103	87	190	Low	59	109	168
	High	32	42	74	High	78	205	283
		135	129	264		137	314	451

Figure 1: Job satisfaction data.

- Supervisor job satisfaction (Low, High),
- Quality of Management (Bad, Good).

The data are summarized in the 3-way contingency table ($2 \times 2 \times 2$) shown in Fig. 1. The task of this exercise is to analyze these data.

Problems:

1. Copy the following code block into `R console` to obtain the contingency table corresponding to the data:

```
## Variable names and levels:
var.levels <- expand.grid(worker = c("Low", "High"),  ## Y direction
                          supervisor = c("Low", "High"), ## X direction
                          management = c("Bad", "Good")) ## Z direction

## The count data:
job.count <- c(103, 87, 32, 42, 59, 109, 78, 205)

## Data frame:
job.freq <- data.frame(var.levels, count = job.count)

## Contingency table:
job.tab <- xtabs(count ~ supervisor + worker + management,
                 data = job.freq)
```

2. Think of what possible association information can be extracted from the data, and use any appropriate method to analyze the data (e.g., hypothesis testing, log-linear model or logistic regression). Plot some graphs to help you get a better idea.