Chi-squared tests for qualitative data

1. Objectives

- Goodness-of-fit test
- Test of independence

2. Exercises

Exercise 1 (12.6 in Textbook 1): The American Bankers Association collects data on the use of credit cards, debit cards, personal checks, and cash when consumers pay for in-store purchases (The Wall Street Journal, December 16, 2003). In 1999, the following usages were reported.

In-Store Purchase	Percentage
Credit card	22
Debit card	21
Personal check	18
Cash	39

A simple random sample taken in 2003 found that for 220 in-stores purchases, 46 used a credit card, 67 used a debit card, 33 used a personal check, and 74 used cash. At $\alpha = 0.01$, can we conclude that a change occurred in how customers paid for in-store purchases over the four-year period from 1999 to 2003?

a. Write the hypotheses.
b. Check the assumptions for the test: are the assumptions satisfied?
c. Report the test statistic with its degrees of freedom as well as the p-value; make decisions and write the conclusion.

Instructions:

In this exercise, there is only one qualitative variable (What is name of that variable?)

Let p_1, p_2, p_3, p_4 represent the population proportion of in-store purchases paid by credit card, debit card, personal check and cash in 1999, respectively.

H_0: $p_1 = 0.22$, $p_2 = 0.21$, $p_3 = 0.18$, $p_4 = 0.39$

H_a: At least one p_i is not equal to its specified value

Note: the sum of all p_is in Ho must be 1

Question 1: Why are p_is the population proportions of 1999, not 2003?

For this exercise, we do not need to import any data file. We are going to use R to run the chi-squared goodness-of-fit test. The format of the necessary code is given below:

```
> observed <- c(O1, O2, …, Ok)   #observed frequencies

> expected <- c(p1, p2,…,pk)    #expected proportions

> chisq.test(x = observed, p = expected) #run the chi-squared test
```

Explanations : `c(O1, O2, …, Ok)` is a vector that contains the observed frequencies. After we construct this vector, we assign it to the variable "observed". `c(p1, p2, …, pk)` is a vector that contains the proportions specified in H_0. This vector is assigned to the variable "expected". Remember that the category of `Oi` must match the category of `pi`. For example, in our exercise, p_1 is the population proportion of in-store purchases paid by **credit card**, then O_1 must be 46, the observed frequency of **credit card** in the sample for 2003.

Let's now write the code to run the chi-squared test for this exercise. You are expected to produce the same output as the following:

```
Chi-squared test for given probabilities

data:  observed
X-squared = 12.206, df = 3, p-value = 0.006709
```

If you want to see the expected frequencies, you should do the following:

```
> result <- chisq.test(x = observed, p = expected)

> result$expected
```

Question 2: Why is the degrees of freedom equal to 3? Answer all the questions specified in the exercise.

Exercise 2 (12.7 in Textbook 1):

The Wall Street Journal's Shareholder Scoreboard tracks the performance of 1000 major U.S. companies (The Wall Street Journal,March 10, 2003). The performance of each company is rated based on the annual total return, including stock price changes and the reinvestment of dividends. Ratings are assigned by dividing all 1000 companies into five groups from A (top 20%), B (next 20%), to E (bottom 20%). Shown here are the one-year ratings for a simple random sample of 60 of the largest companies. Do the largest companies differ in performance from the 1000 companies in the Shareholder Scoreboard? Use $\alpha = .05$.

A	B	C	D	E
5	8	15	20	12

R output for the chi-squared test is given below:

```
Chi-squared test for given probabilities

data:  observed
X-squared = 11.5, df = 4, p-value = 0.02148
```

R output for expected frequencies:

```
[1] 12 12 12 12 12
```

Exercise 3 (12.13 in Textbook 1):

With double-digit annual percentage increases in the cost of health insurance, more and more workers are likely to lack health insurance coverage (USA Today, January 23, 2004). The following sample data provide a comparison of workers with and without health insurance coverage for small, medium, and large companies. For the purpose of this study, small companies are companies that have fewer than 100 employees. Medium companies have 100 to 999 employees, and large companies have 1000 or more employees. Sample data are reported for 50 employees of small companies, 75 employees of medium companies, and 100 employees of large companies.

	Health Insurance		
Size of Company	Yes	No	Total
Small	36	14	50
Medium	65	10	75
Large	88	12	100

a. Conduct a test of independence to determine whether employee health insurance coverage is independent of the size of the company. Use $\alpha = 0.05$. What is the p-value, and what is your conclusion?
b. The USA Today article indicated employees of small companies are more likely to lack health insurance coverage. Use percentages based on the preceding data to support this conclusion.

Instructions:

In this exercise, there are two qualitative variables (what are their names?). We want to analyze the relationship between these variables.

Write the H_0, H_a.

The R codes with explanations are provided below.

```
#Vectors containing rows of observed frequencies in the
contingency table.

> R1 <- c(36, 14)

> R2 <- c(65, 10)

> R3 <- c(88,12)

> rows <- 3 #number of rows is 3


# Constructing a matrix from the 3 vectors. To perform
a chi-squared test of independence, a matrix of
observed frequencies is required by the chisq.test()
function.

> Cont_Table <- matrix(c(R1, R2, R3), nrow=rows,
byrow=TRUE)

> Cont_Table

#Naming the rows and columns is optional

> rownames(Cont_Table) <- c("Small", "Medium", "Large")

> colnames(Cont_Table) = c("Yes", "No")

#Check the matrix

> Cont_Table

#Perform the chi-squared test

> chisq.test(Cont_Table)
```

R output:

The Cont_Table is printed below. It should be the same as the table in the question.

```
       Yes No
Small   36 14
Medium  65 10
Large   88 12
```

R output for the chi-squared test of independence:

```
Pearson's Chi-squared test

data:  Cont_Table
X-squared = 6.9444, df = 2, p-value = 0.03105
```

To check whether the rule of five is satisfied, you can obtain the expected frequencies by:

```
> result <- chisq.test(Cont_Table)
> result$expected
```

Exercise 4 (12.17 in Textbook 1) (Homework). The National Sleep Foundation used a survey to determine whether hours of sleeping per night are independent of age (Newsweek, January 19, 2004). The following show the hours of sleep on weeknights for a simple random sample of individuals age 49 and younger and for a sample of individuals age 50 and older.

	Hours of Sleep				
Age	Fewer than 6	6 to 6.9	7 to 7.9	8 or more	Total
49 or younger	38	60	77	65	240
50 or older	36	57	75	92	260

a. Conduct a test of independence to determine whether the hours of sleep on weeknights are independent of age. Use $\alpha = 0.05$. What is the p-value, and what is your conclusion?
b. What is your estimate of the percentage of people who sleep fewer than 6 hours, 6 to 6.9 hours, 7 to 7.9 hours, and 8 or more hours on weeknights?

Run R and answer the questions above. You should be able to obtain the following R ouput for the test.

```
Pearson's Chi-squared test

data:  Cont_Table
X-squared = 4.007, df = 3, p-value = 0.2607
```