Chapter 23

The Chi-Square Test

In this chapter, we will discuss the following topics:

- We will plot the chi-square density function using the R-function **dchisq**.
- We will find the critical values of the chi-square distribution using the R-function **qchisq**.
- We will test for goodness of fit using the R-function **chisq.test** which is a measure for whether a categorical variable has a specified distribution.
- We will test whether there is a relationship between two categorical variables using the R-function **chisq.test**.

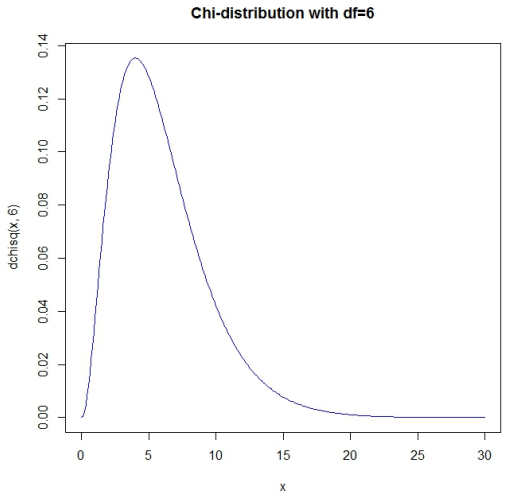
The Chi-Square Distribution

We will denote the critical value of the Chi-square distribution with k degrees of freedom and significance level α by $\chi^2_\alpha(k)$, where there is a probability α to the right for the point $\chi^2_\alpha(k)$ under the Chi-square density curve.

Problem. Plot the Chi-square density function with degrees of freedom, $df = 6$.

Solution. We will plot the function from $x = 0$ to $x = 30$.

```
> x=seq(0,30,0.1)
> plot(x,dchisq(x,6),main="Chi-distribution with df=6",type="l",col="blue")
```



Explanation. The code can be explained as follows:

- The function **dchisq(x,6)** returns the Chi-square density function with $df = 6$.

Problem. Find the critical value, $\chi^2_\alpha(k)$, of the Chi-square distribution with 6 degrees of freedom and significance level $\alpha = 0.05$.

Solution. We obtain:

```
> qchisq(0.05,6,lower.tail=F)
[1] 12.59159
```

Thus $\chi^2_{0.05}(6) = 12.59159$.

Explanation. The code can be explained as follows:

- The function **qchisq(0.05,6,lower.tail=F)** returns the critical value q for which $P(X \geq q) = 0.05$, where X follows a Chi-square distribution with $df = 6$. That is, it returns the value q for which the area under the density curve to the right for q is 0.05.

The Chi-Square Test for Goodness of Fit

In 1900, the British statistician Karl Pearson introduced a chi-square test for measure of goodness of fit which has had such enormous impact on categorical data analysis that it named after him, called the Pearson's chi-squared test. His original goal was to analyze whether each outcome on a roulette wheel in a Monte Carlo casino had equal chance. [1] and [4].

Suppose that a categorical variable have k possible outcomes each with probability p_j, $j = 1, 2, ..., k$. That is, p_j is the probability that the outcome is in category j. Suppose that we have n independent observations of this categorical variable. Let Y_j denote the number of observations in category j. It follows that

$$n = \sum_{j=1}^{k} Y_j.$$

The joint probability mass function of Y_1, Y_2,..., Y_k is the multinomial distribution. For this distribution we have $E(Y_j) = np_j$ and $V(Y_j) = np_j(1 - p_j)$ for $j = 1, 2, ..., k$. Notice that for each category j, Y_j follows the binomial distribution, that is, the marginal distributions of Y_j are the binomial distributions.

We would like to test whether p_j is equal to a known probability p_{j0} for $j = 1, 2, ..., k$. The null hypothesis is:

$$H_0 : p_j = p_{j0} \text{ for all } j = 1, 2, ..., k.$$

The alternative hypothesis is:

$$H_a : p_j \neq p_{j0} \text{ for at least one } j = 1, 2, ..., k.$$

The expected values of $\{Y_j\}$ are called expected frequencies or counts and are given by $\mu_j = np_{j0}$ if the null hypothesis is true for $j = 1, 2, ..., k$. The Pearson's chi-square statistics is given by:

$$\chi^2 = \sum_{j=1}^{k} \frac{(Y_j - \mu_j)^2}{\mu_j} = \sum_{j=1}^{k} \frac{(Y_j - np_{j0})^2}{np_{j0}}.$$

For large enough samples, χ^2 has approximately a chi-square distribution with $(k-1)$ degrees of freedom. We reject H_0 at significance level $\alpha = 0.05$ if $\chi^2 \geq \chi^2_\alpha(k-1)$. The Chi-Square test statistics is a measure of the closeness of the observed counts from the expected counts. If this difference is small, we do not reject H_0.
This test is called the test for *goodness of fit* when it is testing whether a categorical variable has a specified distribution.

As a rule of thumb, the distribution of the statistics χ^2 can be approximated by the chi-square distribution when all cells have expected counts 1 or greater and when no more than 20% of the cells have expected counts less than 5 [3].

In 1936, R.A. Fisher applied Pearson's test to analyze Mendel's result in genetics about his theories of natural inheritance [1]. Mendel experimented with pea plants and crossed pea plants of pure green strain with plants of pure yellow strain. He predicted that the seeds would be 75% yellow and 25% green.

Problem. In one of Mendel's experiment there were $n = 8023$ seeds. Mendel obtained the following result in this particular experiment.

plant	frequency
green	2001
yellow	6022

Test the null hypothesis that the probability, p_1, for green seed is equal to $p_{10} = 0.25$ and the probability, p_2, for yellow seed is equal to $p_{20} = 0.75$, that is,

$$H_0 : p_1 = 0.25, \;\; p_2 = 0.75.$$

Solution. Using the Chi-square test, we obtain:

```
> chisq.test(c(2001,6022),p=c(0.25,0.75))

        Chi-squared test for given probabilities

data:  c(2001, 6022)
X-squared = 0.015, df = 1, p-value = 0.9025
```

Here $\chi^2 = 0.015$ with $df = 1$ and a p-value of 0.90 so Mendel's hypothesis is not rejected.

Explanation. The code can be explained as follows:

- The argument of the function **chisq.test** is here the vector of observed values $c(2001, 6022)$ and the corresponding vector of success probabilities for the null hypothesis, $p = c(0.25, 0.75)$.

Notice that if we do the calculation by hand, we obtain:

$$\begin{aligned}\chi^2 &= \frac{(Y_1 - np_{10})^2}{np_{10}} + \frac{(Y_2 - np_{20})^2}{np_{20}} \\ &= \frac{(2001 - 8023 * 0.25)^2}{8023 * 0.25} + \frac{(6022 - 8023 * 0.75)^2}{8023 * 0.75} = 0.015\end{aligned}$$

with $\chi_{0.05}(1) = 3.841459$. We see that $\chi^2 < \chi_{0.05}(1)$ so the null hypothesis is not rejected. Mendel performed several such experiments and in 1936 Fisher summarized Mendel's result. When Fisher tested Mendel's result, he obtained a summary Chi-squared statistics equal to 42 with $df = 84$ and a p-value of 0.99996 [1]. According to Fisher, this p-value is so small that the result is too good to be true. Fisher thought that Mendel had been tricked by a gardening assistant who did not understand the concept of random chance. Mendel's result is now considered controversial.

The Chi-Square Test and Contingency Tables

In this chapter we will consider two categorical variables displayed in a two-way table. We will use the chi-square test to test the null hypothesis

`Ho: no relationship between two categorical variables`

arising from two different situations:

- The sample total is fixed and individuals are classified according to both the row and column variables. In this case, we would wish to test the hypothesis of independence. This means that we are testing whether the probability of being classified into a particular cell is equal to the product of its marginal probabilities.
- The rows totals are fixed so we are testing whether the column distributions are the same for each row. That is, we are testing whether two or more multinomial distributions are the same. In the other way around, the column totals could be fixed.

Test for Independence

We will now consider a random experiment that results in an outcome classified by two categorical variables, which we denote by A and B. Suppose that we have a total of n independent observations and that A has k categories and B has h categories. We can construct a rectangular table with k rows for A and h columns for B where the kh possible outcomes are represented by the cells of the table. If the cells contain the frequency count of a sample, we call this table a contingency table.

	B_1	B_2	...	B_h
A_1				
A_2				
...				
A_k				

Let p_{ij} denote the joint probability that the outcome is in row i and column j. Let Y_{ij} denote the frequency count in cell ij where $i = 1, 2, ..., k$ and $j = 1, 2, ...h$. Denote the marginal distributions given by the row totals as $\{p_{i.}\}$, where

$$p_{i.} = \sum_{j=1}^{h} p_{ij}$$

is the sum across columns for fixed row i. Denote the marginal distribution given by column totals as $\{p_{.j}\}$ where

$$p_{.j} = \sum_{i=1}^{k} p_{ij}$$

is the sum across rows for fixed column j. Notice that

$$\sum_i p_{i.} = \sum_j p_{.j} = 1. \tag{0.1}$$

We will test the hypothesis that the joint probability that the outcome is in row i and column j is equal to the product of the two marginal probabilities; that is,

$$H_0 : p_{ij} = p_{i.}p_{.j} \;\; i = 1, 2, ..., k \;\; j = 1, 2, ..., h$$

which is the test for independence of the attributes A and B. If $\{p_{i.}\}$ and $\{p_{.j}\}$ are unknown, we estimate $p_{i.}$ by

$$\hat{p}_{i.} = \frac{y_{i.}}{n},$$

where $y_{i.} = \sum_{j=1}^{h} y_{ij}$ is the observed frequency of the i^{th} category of A. Similarly, we estimate

$$\hat{p}_{.j} = \frac{y_{.j}}{n},$$

where $y_{.j} = \sum_{i=1}^{k} y_{ij}$ is the observed frequency of the j^{th} category of B. Then the expected value of Y_{ij} is given by $\mu_{ij} = np_{i.}p_{.j}$ if the null hypothesis is true. We estimate μ_{ij} by

$$\hat{\mu}_{ij} = n\hat{p}_{i.}\hat{p}_{j.} = \frac{y_{i.}y_{.j}}{n}.$$

It follows that the test statistics is given by

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(Y_{ij} - \frac{Y_{i.}Y_{.j}}{n})^2}{\frac{Y_{i.}Y_{.j}}{n}}. \tag{0.2}$$

If H_0 is true, this distribution follows approximately a Chi-square distribution with $df = (k-1)(h-1)$. We reject H_0 at $\alpha = 0.05$ if $\chi^2 \geq \chi^2_\alpha(k-1)(h-1)$.

The degrees of freedom can be explained as follows: Since the contingency table has kh categories, the $df = (kh - 1)$. However, since we needed to estimate $\{p_{i.}\}$ and $\{p_{.j}\}$ each of which satisfies the constraint in (0.1), it follows that we need to estimate $k - 1 + h - 1$ parameters. Thus, $df = kh - 1 - (k - 1 + h - 1) = (k-1)(h-1)$.

Notice that the expected count $\hat{\mu}_{ij} = \frac{y_{i.}y_{.j}}{n}$ can be described as:

$$\text{expected count} = \frac{\text{row total in } \; i^{th} \; \text{row} \;\; \times \text{column total in } \; j^{th} \; \text{column}}{\text{table total}}.$$

The χ^2 statistics can be described as:

$$\chi^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}},$$

where the sum is over all cells in the table. For example, for $k = h = 2$, the contingency table looks like:

	B_1	B_2	Row total
A_1	Y_{11}	Y_{12}	$Y_{1.}$
A_2	Y_{21}	Y_{22}	$Y_{2.}$
Column total	$Y_{.1}$	$Y_{.2}$	n

Problem. A random sample of 50 women were tested for cholesterol and classified according to age and cholesterol level. The results are given in the following contingency table, where the rows represent age and the columns represent cholesterol level:

	< 210	≥ 210	Row total
< 50	18	7	25
≥ 50	8	17	25
Column total	26	24	50

Test the following null hypothesis at significance level $\alpha = 0.05$: age and cholesterol are independent attributes of classification.

Solution. We obtain:

```
> w=matrix(c(18,8,7,17),nrow=2)
> chisq.test(w,correct=FALSE)

        Pearson's Chi-squared test

data:  w
X-squared = 8.0128, df = 1, p-value = 0.004645
```

Since the p-value is 0.004645 we reject the null hypothesis at $\alpha = 0.05$ level. Thus, age and cholesterol are not independent attributes.

Explanation. The code can be explained as follows:

- The command **matrix(c(18,8,7,17),nrow=2)** creates a matrix with two rows and entries 18,8,7,17 filled by columns.
- The argument **correct=FALSE** in the function **chisq.test** turns off the calculation of Yates's continuity correction for 2 by 2 tables. This will be in accordance with the textbook definition for 2 by 2 tables.

Notice that if we do this calculation by hand, we obtain:

$$\chi^2 = \frac{(18-(25*26)/50)^2}{(25*26)/50} + \frac{(7-(25*24)/50)^2}{(25*24)/50} + \frac{(8-(25*26)/50)^2}{(25*26)/50} + \frac{(17-(25*24)/50)^2}{(25*24)/50} = 8.0128$$

with $df = 1$. The critical value is $\chi_{0.05}(1) = 3.841459$ so $\chi^2 \geq \chi^2_{0.05}(1)$.

Test for equality of two or more multinomial distributions
Suppose that we have 2 independent experiments each of which have h possible outcomes. Let p_{1j} be the probability that experiment 1 falls into category j for $j = 1, 2, ..., h$.

Let p_{2j} be the probability that experiment 2 falls into category j for $j = 1, 2, ..., h$.
Let Y_{1j}, $j = 1, 2, ..., h$, be the observed counts for category j for the n_1 independent observations of experiment 1.
Let Y_{2j}, $j = 1, 2, ..., h$, be the observed counts for category j for the n_2 independent observations of experiment 2.
The total observations of the two experiments are $n = n_1 + n_2$. We wish to test whether the two multinomial distributions are equal. This corresponds to testing whether the two experiments have equal probability for each categories, that is,

$$H_0 : p_{1j} = p_{2j} \qquad \text{for} \qquad j = 1, 2, ..., h$$

against

$$H_a : p_{1j} \neq p_{2j} \qquad \text{for at least one} \qquad j = 1, 2, ..., h.$$

Since the probabilities are usually unknown, we estimate them by

$$\hat{p}_j = \frac{Y_{1j} + Y_{2j}}{n_1 + n_2} \qquad \text{for} \qquad j = 1, 2, ..., h - 1.$$

The probability $\hat{p}_h$ is estimated from the constraint that the sum of probabilities is 1. Thus, to estimate each probability, we add the columns counts and divide by the total number of observations. If H_0 is true, the statistics,

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{h} \frac{(Y_{ij} - n_i \hat{p}_j)^2}{n_i \hat{p}_j}$$

follows approximately a Chi-square distribution with $(k - 1)$ degrees of freedom.

We can extend this scenario to k experiments, where the probabilities $\{p_{ij}\}$, for $i = 1, 2, ..., k$, $j = 1, 2, ..., h$ can be estimated by

$$\hat{p}_j = \frac{\sum_{i=1}^{k} Y_{ij}}{\sum_{i=1}^{k} n_i} \text{ for } j = 1, 2, ..., h$$

and $\hat{p}_h = 1 - \sum_{j=1}^{h-1} \hat{p}_j$. If the null hypothesis is true, the statistics,

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{(Y_{ij} - n_i \hat{p}_j)^2}{n_i \hat{p}_j}$$

follows approximately a Chi-square distribution with $(h - 1)(k - 1)$ degrees of freedom.
Notice that the χ^2 statistics can be written exactly as in (0.2). We can see this by recognizing that we can write the total number of observations for experiment i as

$$n_i = \sum_{j=1}^{h} Y_{ij} = Y_{i.}$$

and the total observations in category j as

$$\sum_{i=1}^{k} Y_{ij} = Y_{.j}.$$

Hence,

$$n_i\hat{p}_j = \frac{n_i \sum_{i=1}^{k} Y_{ij}}{\sum_{i=1}^{k} n_i} = \frac{Y_{i.}Y_{.j}}{n}.$$

Problem. Suppose we want to test if online instruction in a math course results in a different grade distribution than a traditional class. From each group, 50 students are randomly selected. At the end of the semester each student is assigned a grade of A,B,C,D, or F. We have the following data:

Instruction	A	B	C	D	F
Online	10	13	16	9	2
Traditional	4	12	15	9	10

Test if the two instructions methods give about the same grades at the 5% significance level.

Solution. We have two experiments, online and traditional, both of which follows a multinomial distribution. We will test if the probability of obtaining an A,B,C,D, and F are the same for these two distributions. Thus, we wish to test the null hypothesis:

H_0 : online- and traditional methods give about the same grade distribution.

We obtain:

```
> w=matrix(c(10,4,13,12,16,15,9,9,2,10),nrow=2)
> colnames(w)=c("A","B","C","D","F")
> rownames(w)=c("online","traditional")
> chisq.test(w)

        Pearson's Chi-squared test

data:  w
X-squared = 7.977, df = 4, p-value = 0.09242

> E=chisq.test(w)$expected
> E
            A    B    C D E
online      7 12.5 15.5 9 6
traditional 7 12.5 15.5 9 6
```

All the expected counts are greater than 5 so we can use the approximation to the Chi-square distribution. The p-value is $0.09242 > 0.05$ so we do not reject the null hypothesis. Thus, we do not have sufficient evidence for the claim that there is a difference between the two instruction methods.

Explanation. The code can be explained as follows:

- The commands **rownames(w)=c(" ")** and **colnames(w)=c(" ")** label the rows and columns of the matrix w, respectively.

- The command **E=chisq.test(w)$expected** returns the expected counts.

Notice that whether we are testing for independence of attributes or equality of multinomial distributions, we obtain the same value of the chi-square statistics and the same number of degrees of freedom. Thus, there are no differences in these tests, and the terminology we are using should depend on the question asked as the next example shows:

Problem. In a Montana study [2] (Montana Economic Outlook Poll) conducted in May 1992, a random sample of Montana residents were asked whether their financial situation was worse, the same, or better compared to a year ago. The residents were classified according to age groups and one of the financial outlook categories, *worse*, *same*, *better*, as given in the table below:

	worse	same	better
34 or younger	21	16	34
35-54	17	23	26
55 or older	22	37	11

Is the personal financial outlook the same across different age groups?

Solution. We will test the null hypothesis:

H_0 : Age and personal financial outlook are independent attributes of classification.

The alternative hypothesis is:

H_a : Age and personal financial outlook are not independent attributes of classification.

We create the contingency table and perform a chi-square test:

```
> w=matrix(c(21,17,22,16,23,37,34,26,11),nrow=3)
> rownames(w)=c("34 or younger","35-54","55 or older")
> colnames(w)=c("worse","same","better")
> chisq.test(w)

        Pearson's Chi-squared test

data:  w
X-squared = 20.6793, df = 4, p-value = 0.0003666
```

Since the p-value is smaller than 0.01, we reject H_0 and conclude that age is not independent of personal financial outlook. The expected counts are all greater than 5 so we can use the approximation to the Chi-square distribution as shown below. We can look at which cells that contribute the most to the observed value of the statistics:

```
> O=chisq.test(w)$observed
> E=chisq.test(w)$expected
> E
                worse     same   better
```

```
34 or younger 20.57971 26.06763 24.35266
35-54         19.13043 24.23188 22.63768
55 or older   20.28986 25.70048 24.00966
> ((O-E)^2)/E
                     worse       same    better
34 or younger 0.008583384 3.88824071 3.8218100
35-54         0.237252964 0.06262568 0.4993969
55 or older   0.144140787 4.96796429 7.0492997
```

We can see that the cells that contribute the most to the statistics are the following cells:

- Age group 55 or older and better. This cell has a higher expected count than observed count.
- Age group 55 or older and same. This cell has a lower expected count than observed count.
- Age group 34 or younger and same. This cell has a higher expected count than observed count.
- Age group 34 or younger and better. This cell has a lower expected count than observed count.

Thus, we find that the youngest age group has a better personal financial outlook than the oldest age group.

Explanation. The code can be explained as follows:

- The command **O=chisq.test(w)$observed** returns the observed counts.
- The command **E=chisq.test(w)$expected** returns the expected counts.
- The command $((O-E)^2)/E$ returns the cell contributions to the statistics.

References

[1] A. Agresti, *Categorical Data Analysis.* Second Edition, Wiley Series in Probability and Statistics John Wiley & Sons, Inc., Hoboken, NJ, 2002.

[2] Bureau of Business and Economic Research, Montana Economic Outlook Poll, University of Montana, May 1992. The data can also be found on DASL at http://lib.stat.cmu.edu/DASL/DataArchive.html.

[3] D. S. Moore, W. I. Notz, M. A. Fligner, R. Scoot Linder. *The Basic Practice of Statistics.* W. F. Freeman and Company, New York, 2013.

[4] S. Stigler, *The History of Statistics: The Measurement of Uncertainty before 1900.* Harvard University Press, Cambridge, MA, 1986

00