PC9 – Chi-square tests

November 5 2018

Exercise 1: Chi square goodness of fit test for cross-bredding plants

Two different types of plants are cross-bred. They differ by two traits ; the first trait can be A or a, the second trait can be B or b. The first generation of cross-bredding plants is homogeneous: all the plants have the following genotype $AaBb$. We question the following model:

- A is dominant and a is recessive,
- B is dominant and b is recessive.

By the Mendelian inheritance, this model would lead to a second generation for which the 4 phenotypes AB (genotype $AABB$, $AaBB$, $AABB$ or $AaBb$), Ab (genotype $AAbb$ or $Aabb$), aB (genotype $aaBB$ or $aaBb$) and ab (genotype $aabb$) would have the following respective frequencies $9/16$, $3/16$, $3/16$ and $1/16$.

Yet, with a sample of 160 plants, we observe 100 phenotypes AB, 18 phenotypes Ab, 24 phenotypes aB, Ab and 18 phenotypes ab.

1. Give the statistical model.
2. Write the likelihood in this model.
3. Propose an estimator for the parameter of your model using the method of moments.
4. Test the considered model at level $\alpha = 0.05$.
5. What can you say about the p-value associated with the observed result? In other words below which level, this result does not lead to rejection at the considered level? You can use the R function qchisq to obtain the χ^2 distribution quantiles.
6. Check that you obtain the same result using the function **chisq.test** of R.
7. Answer Question 2 in the case where the sample has 80 plants, and we observe 50 phenotypes AB, 9 phenotypes Ab, 12 phenotypes aB and 9 phenotypes ab (i.e. with the same proportion as in the previous data).

Solutions of Exercise 1

1. Let x_i be the phenotype of the i-th plant, $i \le n := 160$. We denote

$$x_i = \begin{cases} 1 & \text{if the } i\text{-th plant has phenotype } AB \\ 2 & \text{if the } i\text{-th plant has phenotype } Ab \\ 3 & \text{if the } i\text{-th plant has phenotype } aB \\ 4 & \text{if the } i\text{-th plant has phenotype } ab \end{cases}$$

so $x_i \in \{1,2,3,4\}$. We have observed that $\sum_{i=1}^n \mathbb{1}_{x_i=1} = 100$, $\sum_{i=1}^n \mathbb{1}_{x_i=2} = 18$, $\sum_{i=1}^n \mathbb{1}_{x_i=3} = 24$ and $\sum_{i=1}^n \mathbb{1}_{x_i=4} = 18$. We assume that $(x_1,\ldots,x_n)$ is a realization of the random vector $Z = (X_1,\ldots,X_n)$ with values in $\{1,2,3,4\}^n$ (with sigma-field $\mathcal{P}(\{1,2,3,4\}^n)$). The distribution of a phenotype X_i depends on the parameter $\theta = p \in \Delta_3 := \{p \in (\mathbb{R}_+)^4 : \ p_1+p_2+p_3+p_4 = 1\}$. The distribution of Z is denoted P_p. Under P_p, X_i are i.i.d. and $P_p(X_i = c) = p_c$ for $c \in \{1,2,3,4\}$. Finally the statistical model is

$$\big(\{1,2,3,4\}^n, \mathcal{P}(\{1,2,3,4\}^n), \{P_p : p \in (\mathbb{R}_+)^4 : \ p_1+p_2+p_3+p_4 = 1\}\big)$$

2. The likelihood is

$$\mathcal{L}(Z,p) = \prod_{i=1}^n p_{X_i} = p_1^{N_1} p_2^{N_2} p_3^{N_3} p_4^{N_4},$$

where $N_c = \#\{i \in \{1,\dots,n\} : \ X_i = c\}$, $c \in \{1,2,3,4\}$.

3. For all $c \in \{1,2,3,4\}$,

$$p_c = P(X_i = c) = E(1_c(X_i))$$

so that

$$\hat{p}_c = \frac{1}{n}\sum_{i=1}^{n} 1_c(X_i) = \frac{N_c}{n}$$

is an estimator of p_c using the method of moments.

4. Let $p^* = (9/16, 3/16, 3/16, 1/16)$. We want to test

$$H_0 : \ p = p^* \ \text{ against } \ H_1 : \ p \neq p^* \,.$$

It is a **goodness of fit test**; we use the following statistics:

$$S = \sum_{j=1}^{4} \frac{(N_j - e_j)^2}{e_j},$$

where $e_j = (p^*)_j n$. Under H_0, S is approximatively distributed as a $\chi^2(3)$. We reject H_0 if S is large, i.e. $S \geq c_\alpha$ for some constant c_α. Indeed, under H_0, S should be close to 0 and S is always non-negative. We now determine c by controlling the type I error:

$$\alpha = P_{p^*}\left(S \geq c_\alpha\right) \sim P_{Y\sim\chi^2_3}(Y \geq c_\alpha),$$

so we choose $c_\alpha = \chi^2_{3,1-\alpha}$, i.e. in the case where $\alpha = 0.05$, $c_{0.05} = 7.81$.

```
qchisq(0.95,df=3)
```

```
## [1] 7.814728
```

and the test is

$$\phi_\alpha(X_1,\dots,X_n) = \begin{cases} 1 & \text{if } \sum_{j=1}^{4} \frac{(N_j - e_j)^2}{e_j} \geq c_\alpha \\ 0 & \text{otherwise.} \end{cases}$$

We have

j	1	2	3	4
$N_j(\omega)$	100	18	24	18
e_j	90	30	30	10

and

$$S(\omega) = \frac{(100-90)^2}{90} + \frac{(18-30)^2}{30} + \frac{(24-30)^2}{30} + \frac{(18-10)^2}{10} \simeq 13.51.$$

```
p <- c(9/16,3/16,3/16,1/16)
N <- c(100,18,24,18)
n <-sum(N)
E <- n*p
S <- sum(((N-E)^2)/E)
S
```

```
## [1] 13.51111
```

We reject H_0 (since $13.51 > 7.81$).

5. Then the p-value is

$$\hat{\alpha} = P_{p^*}\left(S \geq 13.51\right),$$

```r
1-pchisq(S,df=3)
```

```
## [1] 0.003652111
```

The p-value is smaller than the chosen level (0.05) so we indeed reject H_0.

In R:

```r
S
```

```
## [1] 13.51111
```

```r
1-pchisq(S,df=3)
```

```
## [1] 0.003652111
```

```r
chisq.test(x=N, p=p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  N
## X-squared = 13.511, df = 3, p-value = 0.003652
```

6. With the new data

j	1	2	3	4
$N_j(\omega_2)$	50	9	12	9
e_j	45	15	15	5

and

$$S(\omega_2) = \frac{(50-45)^2}{45} + \frac{(9-15)^2}{15} + \frac{(12-15)^2}{15} + \frac{(9-5)^2}{5} \simeq 6.76.$$

```r
p <- c(9/16,3/16,3/16,1/16)
N2 <- c(50,9,12,9)
n2 <-sum(N2)
E2 <- n2*p
S2 <- sum(((N2-E2)^2)/E2)
S2
```

```
## [1] 6.755556
```

With these data, $6.76 \leq c_{0.05}$ and we would not reject H_0 at level $\alpha = 0.05$. We don't have enough evidence against H_0.

In R:

```r
p <- c(9/16,3/16,3/16,1/16)
N2 <- c(50,9,12,9)
n2 <-sum(N2)
E2 <- n2*p
S2 <- sum(((N2-E2)^2)/E2)
S2
```

```
## [1] 6.755556
```

```r
1-pchisq(S2,df=3)
```

```
## [1] 0.08011092
```

```r
chisq.test(x=N2, p=p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  N2
## X-squared = 6.7556, df = 3, p-value = 0.08011
```

Exercise 2: Chi square independence test for burgers

The Mac Burger's society launches its new burger FolBurger in the US and in Europe. It does a survey by asking for a feedback (bad, correct and good) to customers living in four cities. They obtain the following answers:

	Bad	Correct	Good
Los Angeles	29	124	87
Chicago	74	278	208
Madrid	114	277	87
Paris	182	417	123

The society wants to test the dependence of feedbacks with the place of residence of the customer.

1. Give the statistical model.
2. Enter the data in a matrix called "tab" and give names to rows and columns. Compute the sums by rows and columns and merge it with the previous matrix in a new matrix called tab2.
3. Explain what is displayed in R when you execute the following instructions:
   - "prop.table(tab)",
   - "prop.table(tab,margin=1)",
   - "prop.table(tab,margin=2)".
4. Display the six barplots of the feedbacks of customers living in Los Angeles, Chicago, the US, Madrid, Paris and Europe.
5. Comment these histograms.
6. Give a test at level $\alpha = 0.01$ to answer the following questions.
   (a) Do the feedbacks on FolBurger depend on the place of residence of the customer?
   (b) Are the feedbacks of customers living in Los Angeles different from those living in Chicago?
   (c) Are the feedbacks of customers living in Madrid different from those living in Paris?
7. Do the same tests using the R function chisq.test.
8. What do you conclude?

Solution of Exercise 2

1. The surveyed population is all the customers living in Los Angeles, Chicago, Madrid or Paris who have eaten at least one FolBurger. Let $x_i \in \{1, 2, 3, 4\}$ be the city where the i-th surveyed people live and $y_i \in \{B, C, G\}$ its feedback for $i \leqslant n = 2000$.

We assume that $((x_1, y_1), \cdots, (x_n, y_n))$ is a realization of a random vector $((X_1, Y_1), \cdots, (X_n, Y_n))$ where (X_i, Y_i) are i.i.d. from a distribution P_p. The distribution P_p is parameterized by $p \in \Theta := \{p \in (\mathbb{R}_+)^{4\times 3} : \sum_{c=1}^{4} \sum_{j\in\{B,C,G\}} p_{c,j} = 1\}$ and defined as $P_p(X_i = c, Y_i = j) = p_{c,j}$ for $j \in \{1, 2, 3\}$.

Finally the statistical model is

$$(\{1, 2, 3, 4\} \times \{1, 2, 3\}, \mathcal{P}(\{1, 2, 3, 4\} \times \{1, 2, 3\}), \{P_p : p \in \Theta\}) .$$

2. For $c \in \{1,2,3,4\}$ and $j \in \{B,C,G\}$, let $N_{c,j} = \text{card}\{i \leqslant n : \; X_i = c \text{ and } Y_i = j\}$ be the number of feedbacks with value j from a customer living in city c. We have observed the following $n_{c,j}$:

```
tab <- matrix(c(29,124,87,74,278,208,114,277,87,182,417,123),
              ncol=3,byrow = T)
rownames(tab)= c("Los Angeles","Chicago","Madrid","Paris")
colnames(tab)=c("Bad","Correct","Good")
tab
```

```
##             Bad Correct Good
## Los Angeles  29     124   87
## Chicago      74     278  208
## Madrid      114     277   87
## Paris       182     417  123
```

Let $N_{c,\cdot} = \sum_{j\in\{B,C,G\}} N_{c,j}$ be the total number of surveyed custommers living in city c. Let $N_{\cdot,j} = \sum_{c=1}^{4} N_{c,j}$ be the total number of feedbacks with value j. Then $n = \sum_{c=1}^{4}\sum_{j\in\{B,C,G\}} N_{c,j} = \sum_{c=1}^{4} N_{c,\cdot} = \sum_{j\in\{B,C,G\}} N_{\cdot,j} = 2000$. The observed values of $N_{c,\cdot}$ are displayed in "tot_cities" and the observed values of $N_{\cdot,j}$ are displayed in "tot_feedbacks".

```
tot <- sum(tab)
tot
```

```
## [1] 2000
```

```
tot_cities <- rowSums(tab)
tot_cities
```

```
## Los Angeles     Chicago      Madrid       Paris
##         240         560         478         722
```

```
tot_feedbacks <- colSums(tab)
tot_feedbacks
```

```
##     Bad Correct    Good
##     399    1096     505
```

```
tab2 <- rbind(tab,tot_feedbacks)
tab2 <- cbind(tab2,c(tot_cities,tot))
rownames(tab2)= c("Los Angeles","Chicago","Madrid","Paris","TOTAL")
colnames(tab2)=c("Bad","Correct","Good","TOTAL")
tab2
```

```
##             Bad Correct Good TOTAL
## Los Angeles  29     124   87   240
## Chicago      74     278  208   560
## Madrid      114     277   87   478
## Paris       182     417  123   722
## TOTAL       399    1096  505  2000
```

4. "prop.table(tab)" displays the proportion of feedbacks j from customers living in city c among the n feedbacks:

$$\text{prop.table(tab)}_{c,j} = N_{c,j}/n.$$

```
prop.table(tab)
```

```
##                Bad Correct   Good
## Los Angeles 0.0145  0.0620 0.0435
## Chicago     0.0370  0.1390 0.1040
## Madrid      0.0570  0.1385 0.0435
```

```
## Paris       0.0910  0.2085 0.0615
```

```r
tab/tot
```

```
##                 Bad Correct   Good
## Los Angeles 0.0145  0.0620 0.0435
## Chicago     0.0370  0.1390 0.1040
## Madrid      0.0570  0.1385 0.0435
## Paris       0.0910  0.2085 0.0615
```

"prop.table(tab,margin=1)" displays the proportion of feedbacks j among the $N_{c,\cdot}$ feedbacks of customers living in city c for all cities:

$$\text{prop.table(tab,margin=1)}_{c,j} = N_{c,j}/N_{c,\cdot}.$$

```r
prop.table(tab,margin=1)
```

```
##                   Bad   Correct      Good
## Los Angeles 0.1208333 0.5166667 0.3625000
## Chicago     0.1321429 0.4964286 0.3714286
## Madrid      0.2384937 0.5794979 0.1820084
## Paris       0.2520776 0.5775623 0.1703601
```

```r
tab/cbind(tot_cities,tot_cities,tot_cities)
```

```
##                   Bad   Correct      Good
## Los Angeles 0.1208333 0.5166667 0.3625000
## Chicago     0.1321429 0.4964286 0.3714286
## Madrid      0.2384937 0.5794979 0.1820084
## Paris       0.2520776 0.5775623 0.1703601
```

"prop.table(tab,margin=2)" displays the proportion of feedbacks of customers living in city c among the $N_{\cdot,j}$ feedbacks with value j for all possible j:

$$\text{prop.table(tab,margin=2)}_{c,j} = N_{c,j}/N_{\cdot,j}.$$

```r
prop.table(tab,margin=2)
```

```
##                   Bad   Correct      Good
## Los Angeles 0.0726817 0.1131387 0.1722772
## Chicago     0.1854637 0.2536496 0.4118812
## Madrid      0.2857143 0.2527372 0.1722772
## Paris       0.4561404 0.3804745 0.2435644
```

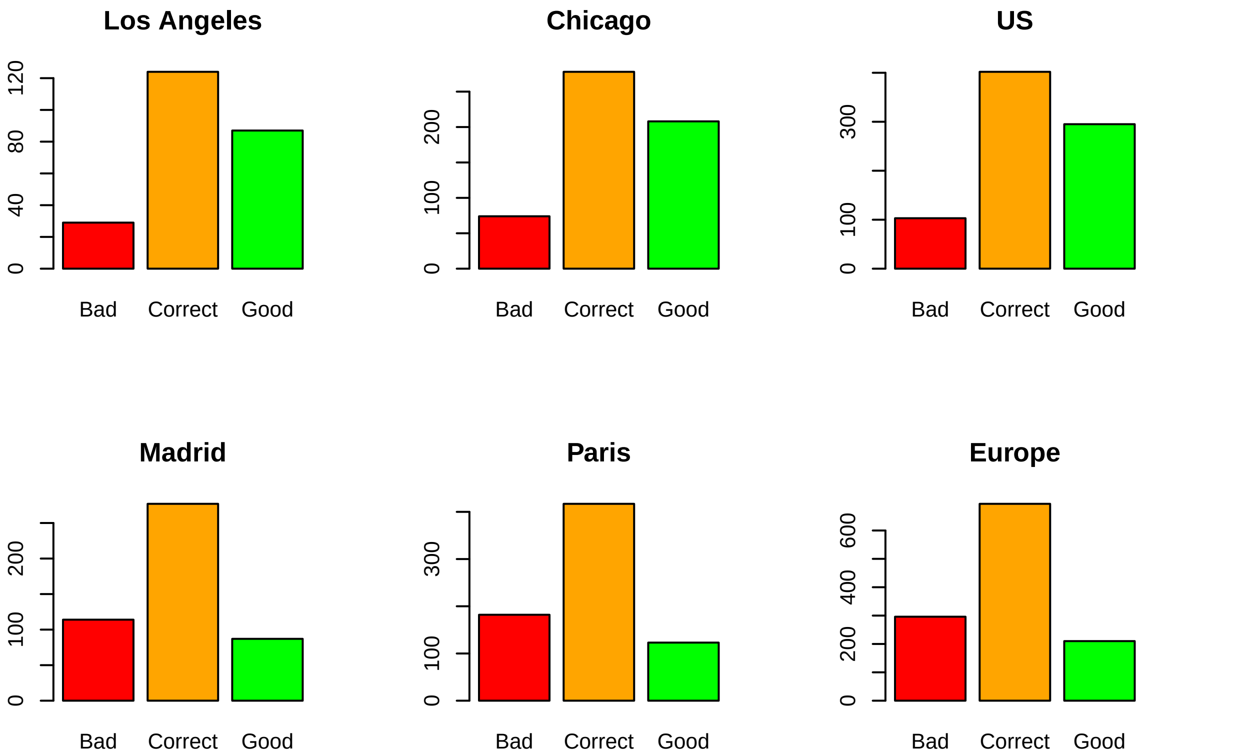
```r
tab/rbind(tot_feedbacks,tot_feedbacks,tot_feedbacks,tot_feedbacks)
```

```
##                   Bad   Correct      Good
## Los Angeles 0.0726817 0.1131387 0.1722772
## Chicago     0.1854637 0.2536496 0.4118812
## Madrid      0.2857143 0.2527372 0.1722772
## Paris       0.4561404 0.3804745 0.2435644
```

5.

```r
par(mfrow=c(2,3))
color <- c("red","orange","green")
barplot(tab[1,],main="Los Angeles",col=color)
barplot(tab[2,],main="Chicago",col=color)
barplot(tab[1,]+tab[2,],main="US",col=color)
barplot(tab[3,],main="Madrid",col=color)
```

```r
barplot(tab[4,],main="Paris",col=color)
barplot(tab[3,]+tab[4,],main="Europe",col=color)
```



From the barplots, it looks like the feedbacks depend on the place of residence of the customer, more precisely if the cusomer lives in Europe or in the US. But it seems that feedbacks from customers living in Los Angeles and Chicago are similar and feedbacks from customers living in Paris and Madrid are also similar. We now have to check these speculations by hypothesis testings.

6.a. We want to test if the variable X and Y are independent. If these variables are independent then for all c and j $p_{c,j} = p_{c,\cdot}p_{\cdot,j}$ for some $p_{c,\cdot}$ and $p_{\cdot,j}$.

$$H_0: \ p_{c,j} = p_{c,\cdot}p_{\cdot,j} \text{ for all } c,j \text{ for some } p_{c,\cdot} \text{ and } p_{\cdot,j} \quad \text{against} \quad H_1: \ p_{c,j} \neq p_{c,\cdot}p_{\cdot,j} \text{ for some } c,j \text{ for all } p_{c,\cdot}\,.$$

It is an independence test. In the lecture, the following test statistic was proposed:

$$S((X_1,Y_1),\cdots,(X_n,Y_n)) = n\sum_{c=1}^{4}\sum_{j\in\{B,C,G\}}\frac{(N_{c,j}/n - N_{\cdot,j}N_{c,\cdot}/n^2)^2}{N_{\cdot,j}N_{c,\cdot}/n^2}.$$

Under H_0, $N_{c,j}/n$, which is an estimator of $p_{c,j}$, should be close to $N_{\cdot,j}N_{c,\cdot}/n^2$, which is an estimator of $p_{c,\cdot}p_{\cdot,j}$ and H_0, $S((X_1,Y_1),\cdots,(X_n,Y_n))$ should be close to zero. Besides S is always nonnegative. So we reject H_0 when S is large, i.e. when $S \geqslant s_c$ for some positive critical value s_c.

Moreover under H_0, S is asymptotically distributed from a chi square distribution $\chi^2((4-1)\times(3-1)) = \chi^2(6)$ This critical value is chosen to control the type I error:

$$\mathbb{P}_{H_0}(S \geqslant s_c) \simeq \mathbb{P}_{Z\sim\chi^2(6)}(Z \geqslant s_c) = 1 - F_{\chi^2(6)}(s_c).$$

Then we choose $s_c = \chi^2_{6,0.99} \simeq 16.81$, where $\chi^2_{6,0.99}$ is the 0.99 quantile of the $\chi^2(6)$ distribution.

```r
qchisq(0.99,df=6)
```

```
## [1] 16.81189
```

Finally, we reject H_0 when $S \geqslant \chi^2_{6,0.99}$.

Our observed value of S is $s_{obs} \simeq 110.56$ which is larger than our critical value, so we reject H_0.

```
pcpj_hat <- rbind(tot_feedbacks,tot_feedbacks,tot_feedbacks,tot_feedbacks)*cbind(tot_cities,tot_cities,t
rownames(pcpj_hat)=c("Los Angeles","Chicago","Madrid","Paris")
tot*pcpj_hat
```

```
##                 Bad Correct    Good
## Los Angeles  47.880 131.520  60.600
## Chicago     111.720 306.880 141.400
## Madrid       95.361 261.944 120.695
## Paris       144.039 395.656 182.305
```

```
s <- tot*sum(sum(  ((tab/tot - pcpj_hat )^2)/ pcpj_hat ))
s
```

```
## [1] 110.5614
```

The p-value is

$$\mathbb{P}_{H_0}\left(S \geqslant s_{obs}\right) \simeq \mathbb{P}_{Z\sim\chi^2(6)}\left(Z \geqslant s_{obs}\right) = 1 - F_{\chi^2(6)}(s_{obs})$$

very small.

```
1-pchisq(s,df=6)
```

```
## [1] 0
```

7.a

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 110.56, df = 6, p-value < 2.2e-16
```

6.b. We only consider the n_2 observations for which $X = 1$ or $X = 2$. We want to test if the variable X and Y are independent. If these variables are independent then for all $c \in \{1, 2\}$ and j $p_{c,j} = p_{c,\cdot}p_{\cdot,j}$ for some $p_{c,\cdot}$ and $p_{\cdot,j}$.

H_0 : $p_{c,j} = p_{c,\cdot}p_{\cdot,j}$ for all $c \in \{1, 2\}, j$ for some $p_{c,\cdot}$ and $p_{\cdot,j}$ against H_1 : $p_{c,j} \neq p_{c,\cdot}p_{\cdot,j}$ for some c, j for all $p_{c,\cdot}$ and $p_{\cdot,j}$.

It is an independence test again. Similarly, we choose the following test statistic:

$$S_2((X_1, Y_1), \cdots, (X_{n_2}, Y_{n_2})) = n_2 \sum_{c=1}^{2} \sum_{j\in\{B,C,G\}} \frac{(N_{c,j}/n_2 - N_{\cdot,j}N_{c,\cdot}/n_2^2)^2}{N_{\cdot,j}N_{c,\cdot}/n_2^2},$$

where $(X_i, Y_i)_{i=1}^{n_2}$ denotes the observations corresponding to the US. We reject H_0 when S_2 is large, i.e. when $S_2 \geqslant s_c$ for some positive critical value s_c.

Moreover under H_0, S_2 is asymptotically distributed from a chi square distribution $\chi^2((2-1)\times(3-1)) = \chi^2(2)$ This critical value is chosen to control the type I error:

$$\mathbb{P}_{H_0}\left(S_2 \geqslant s_c\right) \simeq \mathbb{P}_{Z\sim\chi^2(2)}\left(Z \geqslant s_c\right) = 1 - F_{\chi^2(2)}(s_c).$$

Then we choose $s_c = \chi^2_{2,0.99} \simeq 9.21$, where $\chi^2_{2,0.99}$ is the 0.99 quantile of the $\chi^2(2)$ distribution.

```
qchisq(0.99,df=2)
```

```
## [1] 9.21034
```

Finally, we reject H_0 when $S_2 \geqslant \chi^2_{2,0.99}$.

Our observed value of S_2 is $s_{2,obs} \simeq 0.34$ which is smaller than our critical value, so we don't reject H_0.

```
tab2 <- tab[1:2,]
n2 <- sum(tab2)
tot_feedbacks2 <-colSums(tab2)
tot_cities2 <-rowSums(tab2)
pcpj_hat2 <- rbind(tot_feedbacks2,tot_feedbacks2)*cbind(tot_cities2,tot_cities2,tot_cities2)/(n2^2)
rownames(pcpj_hat2)=c("Los Angeles","Chicago")
n2*pcpj_hat2
```

```
##              Bad Correct  Good
## Los Angeles 30.9   120.6  88.5
## Chicago     72.1   281.4 206.5
```

```
s2 <- n2*sum(sum(  ((tab2/n2 - pcpj_hat2 )^2)/ pcpj_hat2 ))
s2
```

```
## [1] 0.3401518
```

The p-value is

$$\mathbb{P}_{H_0}\left(S \geqslant s_{2,obs}\right) \simeq \mathbb{P}_{Z\sim\chi^2(2)}\left(Z \geqslant s_{2,obs}\right) = 1 - F_{\chi^2(2)}(s_{2,obs}) \simeq 0.84.$$

```
1-pchisq(s2,df=2)
```

```
## [1] 0.8436008
```

7.b.

```
chisq.test(tab2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab2
## X-squared = 0.34015, df = 2, p-value = 0.8436
```

6.c. We only consider the n_3 observations for which $X = 3$ or $X = 4$. We want to test if the variable X and Y are independent. If these variables are independent then for all $c \in \{3,4\}$ and j $p_{c,j} = p_{c,\cdot}p_{\cdot,j}$ for some $p_{c,\cdot}$ and $p_{\cdot,j}$.

H_0 : $p_{c,j} = p_{c,\cdot}p_{\cdot,j}$ for all $c \in \{3,4\}, j$ for some $p_{c,\cdot}$ and $p_{\cdot,j}$ against H_1 : $p_{c,j} \neq p_{c,\cdot}p_{\cdot,j}$ for some c, j for all $p_{c,\cdot}$ and $p_{\cdot,j}$.

It is an independence test again. Similarly, we choose the following test statistic:

$$S_3((X_1,Y_1),\cdots,(X_{n_3},Y_{n_3})) = n_3 \sum_{c=3}^{4} \sum_{j\in\{B,C,G\}} \frac{(N_{c,j}/n_3 - N_{\cdot,j}N_{c,\cdot}/n_3^2)^2}{N_{\cdot,j}N_{c,\cdot}/n_3^2},$$

where $(X_i,Y_i)_{i=1}^{n_2}$ denotes the observations corresponding to Europe. We reject H_0 when S_3 is large, i.e. when $S_3 \geqslant s_c$ for some positive critical value s_c.

Moreover under H_0, S_3 is asymptotically distributed from a chi square distribution $\chi^2((2-1)\times(3-1)) = \chi^2(2)$ This critical value is chosen to control the type I error:

$$\mathbb{P}_{H_0}\left(S_3 \geqslant s_c\right) \simeq \mathbb{P}_{Z\sim\chi^2(2)}\left(Z \geqslant s_c\right) = 1 - F_{\chi^2(2)}(s_c).$$

Then we choose $s_c = \chi^2_{2,0.99} \simeq 9.21$, where $\chi^2_{2,0.99}$ is the 0.99 quantile of the $\chi^2(2)$ distribution.

```r
qchisq(0.99,df=2)
```

```
## [1] 9.21034
```

Finally, we reject H_0 when $S_3 \geqslant \chi^2_{2,0.99}$.

Our observed value of S_3 is $s_{3,obs} \simeq 0.44$ which is smaller than our critical value, so we don't reject H_0.

```r
tab3 <- tab[3:4,]
n3 <- sum(tab3)
tot_feedbacks3 <-colSums(tab3)
tot_cities3 <-rowSums(tab3)
pcpj_hat3 <- rbind(tot_feedbacks3,tot_feedbacks3)*cbind(tot_cities3,tot_cities3,tot_cities3)/(n3^2)
rownames(pcpj_hat2)=c("Madrid","Paris")
n3*pcpj_hat3
```

```
##                      Bad  Correct   Good
## tot_feedbacks3 117.9067 276.4433  83.65
## tot_feedbacks3 178.0933 417.5567 126.35
```

```r
s3 <- n3*sum(sum(  ((tab3/n3 - pcpj_hat3 )^2)/ pcpj_hat3 ))
s3
```

```
## [1] 0.4399826
```

The p-value is

$$\mathbb{P}_{H_0}\left(S \geqslant s_{3,obs}\right) \simeq \mathbb{P}_{Z\sim\chi^2(2)}\left(Z \geqslant s_{3,obs}\right) = 1 - F_{\chi^2(2)}(s_{3,obs}) \simeq 0.80.$$

```r
1-pchisq(s3,df=2)
```

```
## [1] 0.8025258
```

7.c

```r
chisq.test(tab3)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab3
## X-squared = 0.43998, df = 2, p-value = 0.8025
```

Exercise 3: Empirical study of the asymptotics of square goodness of fit tests

Let X_i be i.i.d. random variables with values in $\{1, 2, \ldots, k\}$. Under our model, the distribution of X_i depends on the parameter $\theta = p \in \Delta_{k-1} := \{p \in (\mathbb{R}_+)^k : \ p_1 + p_2 + \cdots + p_k = 1\}$. The distribution of X_i is denoted P_p. Under P_p, the X_i are i.i.d. and $P_p(X_i = c) = p_c$ for $c \in \{1, 2, \cdots, k\}$. So the statistical model is

$$\left(\{1,2,\cdots,k\}^n, \mathcal{P}(\{1,2,\cdots,k\}^n), \{P_p : p \in (\mathbb{R}_+)^k : \ p_1 + p_2 + \cdots + p_k = 1\}\right).$$

Let p^* be in Δ_{k-1}. We want to test

$$H_0 : \ p = p^* \ \text{ against } \ H_1 : \ p \neq p^*.$$

It is a test of goodness of fit. In the lecture, the following test statistic was proposed:

$$S(X_1, \cdots, X_n) = \sum_{j=1}^{4} \frac{(N_j - e_j)^2}{e_j},$$

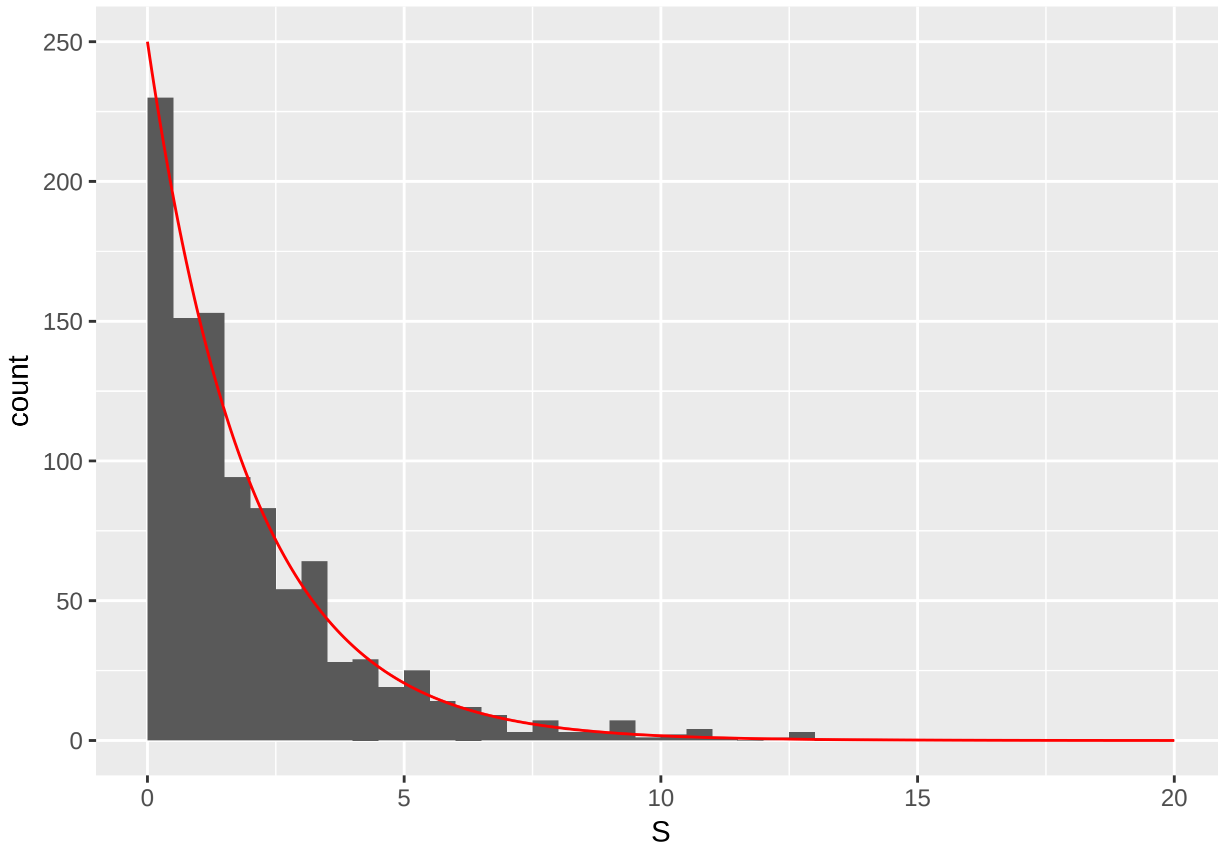
where $e_j = (p^*)_j n$ and $N_j(X_1, \ldots, X_n) = \#\{i \in \{1, \ldots, n\} : \ X_i = j\}$, $j \in \{1, \ldots, k\}$.

1. Under H_0, S is asymptotically distributed as a $\chi^2(k-1)$. Check numerically this property with R, when $p^* = (0.2, 0.3, 0.5)$ and $n = 100$. First create a function in R with inputs p^*, x and n and output $S(x)$.
2. What happens when n is not large enough. Check the distribution of S when $p^* = (0.002, 0.3, 0.5, 0.198)$ and $n = 100$.

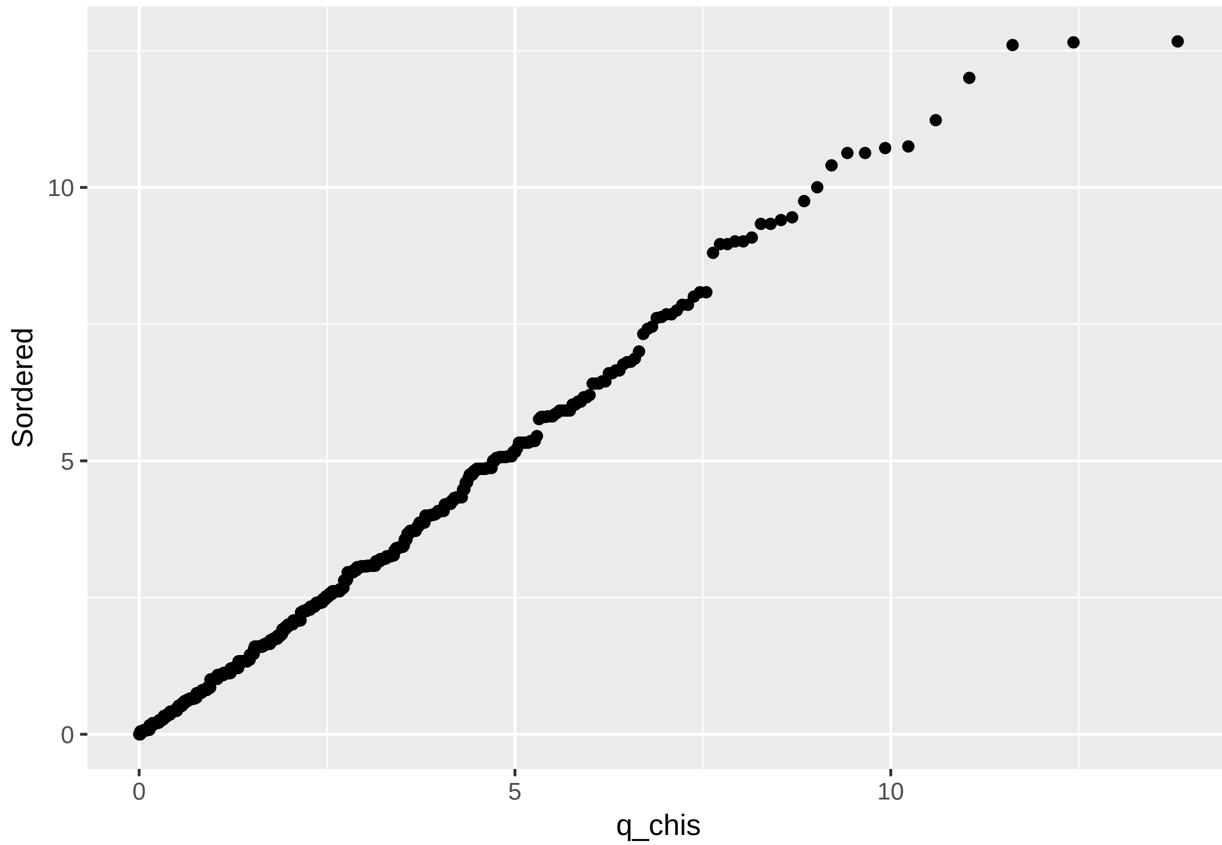
Solution of Exercise 3

```
statistic_qui_squared = function(p0,X,n){
  k <- length(p0)
  N <- sapply(1:k, function(x) sum(X==x))
  E <- n*p0
  S <- sum(((N-E)^2)/E)
  S
}
```

```
p0 <- c(0.2,0.3,0.5)
k <- length(p0)
n <- 100

I <- 1000
X <- lapply(1:I, function(i) sample(x = 1:k, n, replace = T, prob = p0))
S <- sapply(X, function(x) statistic_qui_squared(p0,x,n))
data <-data.frame(S=S, Sordered=sort(S), q_chis= qchisq((1/(I+1)*(1:I)),df=k-1))

x <- seq(0,20,by=0.01)
y <- dchisq(x,df=k-1)
data_dens_chisq <- data.frame(x=x, y=y)
bw <- 0.5

library(ggplot2)
pl <-ggplot(data=data)+geom_histogram(aes(x=S),boundary=0, binwidth = bw)
pl + geom_line(data=data_dens_chisq, aes(x=x,y=y*I*bw), col ="red")
```

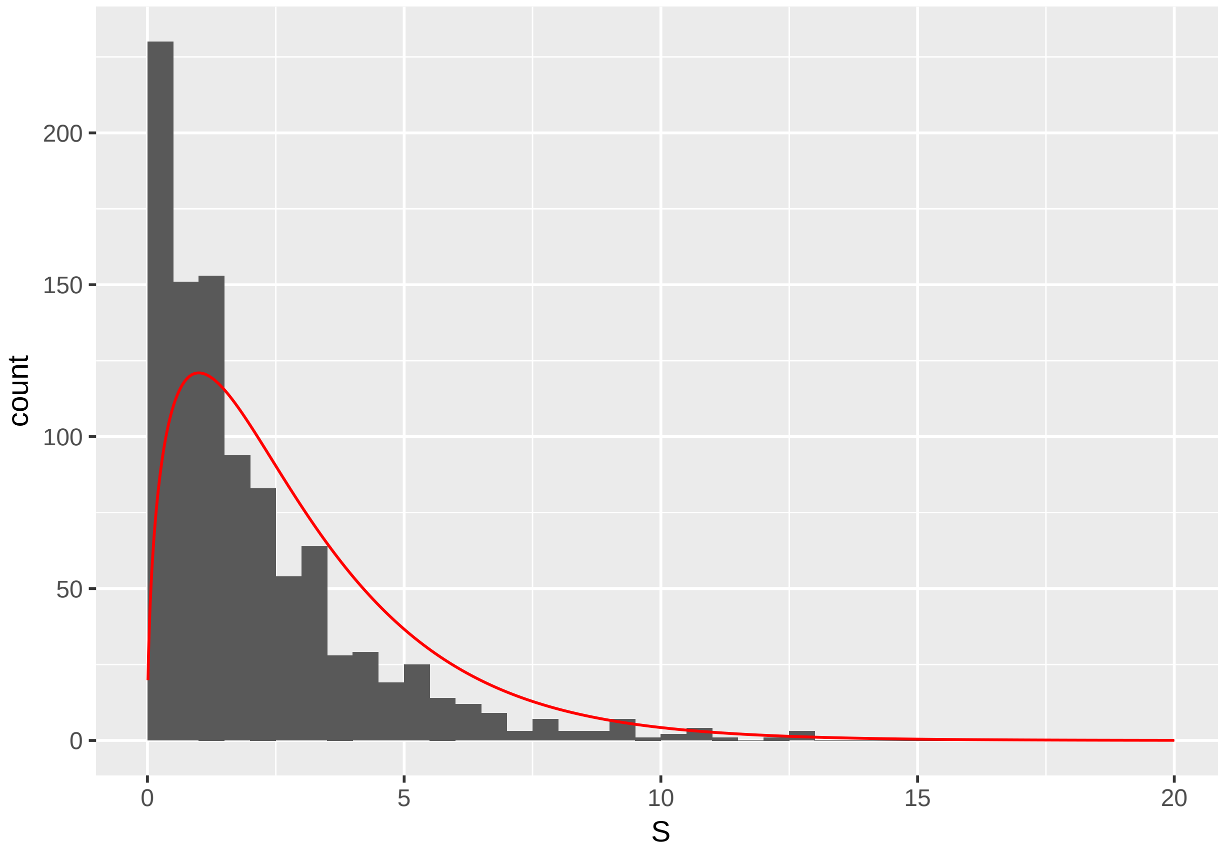



```
ggplot(data=data)+geom_point(aes(x=q_chis,y=Sordered))
```

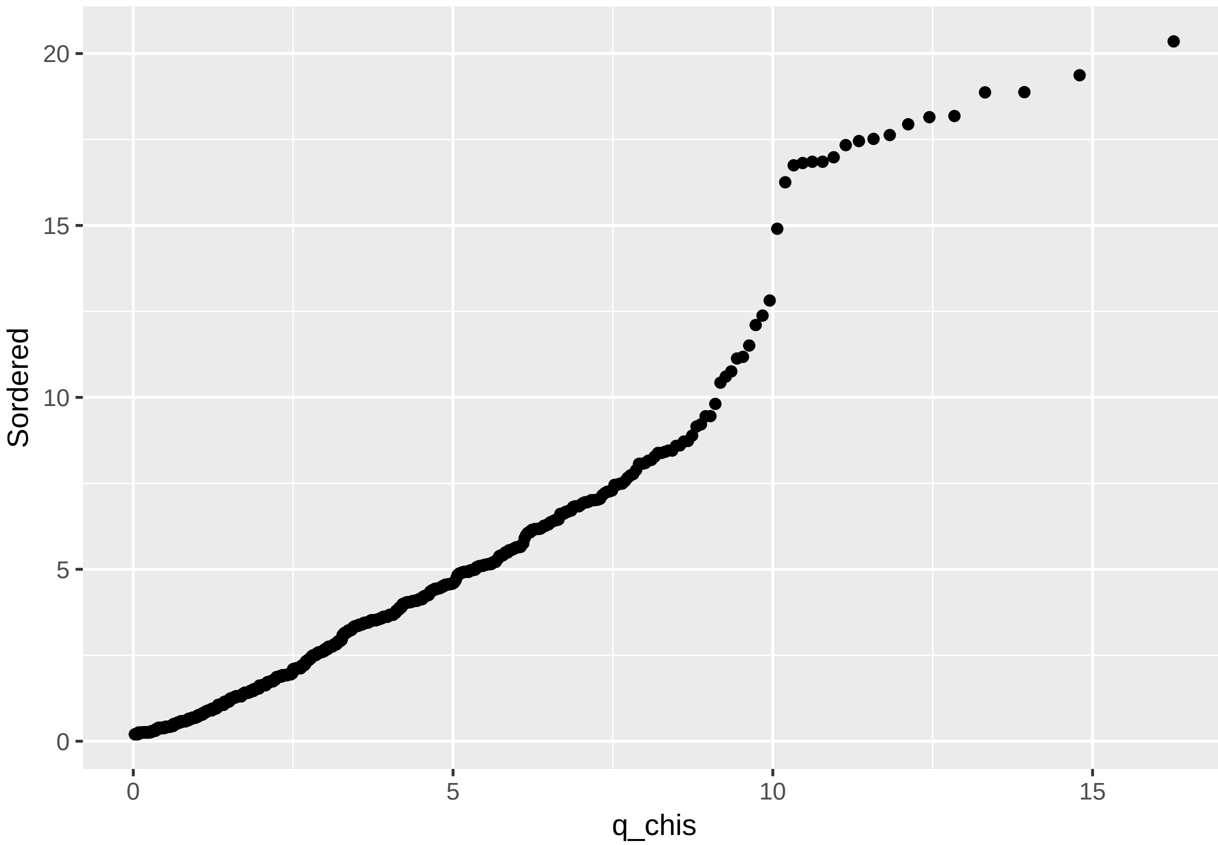


2.

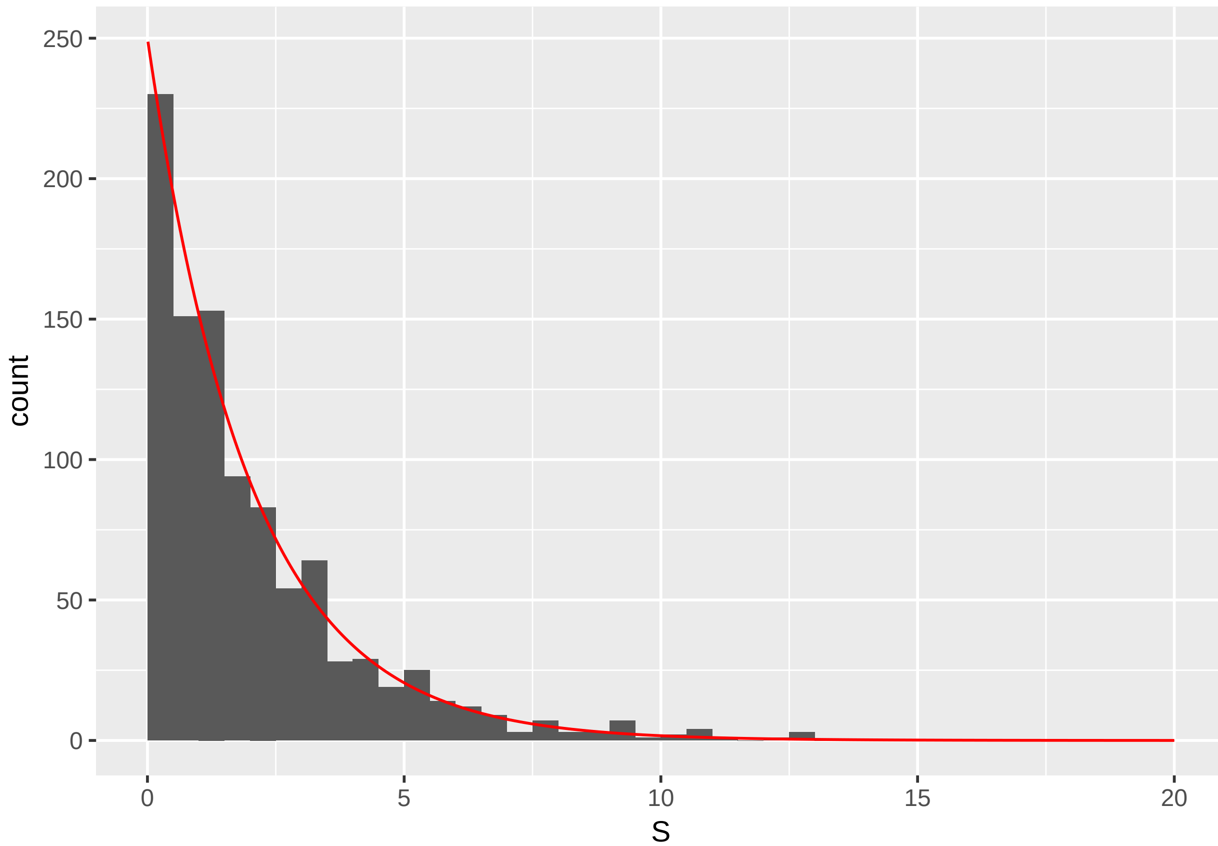
```
p0 <- c(0.002,0.3,0.5,0.198)
k <- length(p0)
n <- 100

I <- 1000
X <- lapply(1:I, function(i) sample(x = 1:k, n, replace = T, prob = p0))
S <- sapply(X, function(x) statistic_qui_squared(p0,x,n))
data2 <-data.frame(S=S, Sordered=sort(S), q_chis= qchisq((1/(I+1)*(1:I)),df=k-1))

x2 <- seq(0.01,20,by=0.01)
y2 <- dchisq(x2,df=k-1)
data_dens_chisq2 <- data.frame(x=x2, y=y2)
bw <- 0.5

library(ggplot2)
pl <-ggplot(data=data)+geom_histogram(aes(x=S),boundary=0, binwidth = bw)
pl + geom_line(data=data_dens_chisq2, aes(x=x,y=(y*I*bw)), col ="red")
```



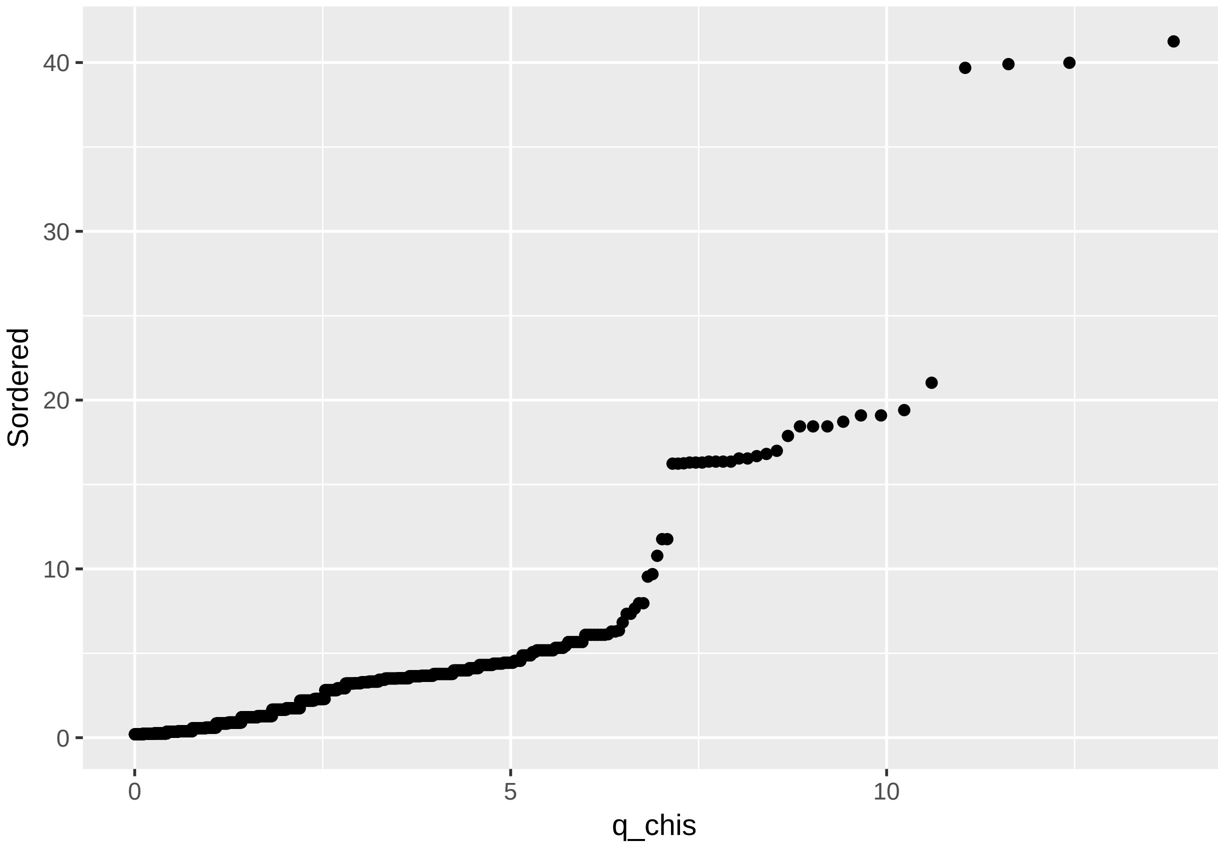

```
ggplot(data=data2)+geom_point(aes(x=q_chis,y=Sordered))
```




```
p0 <- c(0.002,0.4,0.598)
k <- length(p0)
n <- 100

I <- 1000
X <- lapply(1:I, function(i) sample(x = 1:k, n, replace = T, prob = p0))
S <- sapply(X, function(x) statistic_qui_squared(p0,x,n))
data2 <-data.frame(S=S, Sordered=sort(S), q_chis= qchisq((1/(I+1)*(1:I)),df=k-1))

x2 <- seq(0.01,20,by=0.01)
y2 <- dchisq(x2,df=k-1)
data_dens_chisq2 <- data.frame(x=x2, y=y2)
bw <- 0.5

library(ggplot2)
pl <-ggplot(data=data)+geom_histogram(aes(x=S),boundary=0, binwidth = bw)
pl + geom_line(data=data_dens_chisq2, aes(x=x,y=(y*I*bw)), col ="red")
```




```
ggplot(data=data2)+geom_point(aes(x=q_chis,y=Sordered))
```



So when the number of classes (k) is increasing or some of the p_i^* are small, one has to increase the number of samples (n).