Chi Square Tests

2020

Chi Square Tests

The **chi-squared tests** considered in this topic **compare the observed frequencies in different groups with their corresponding expected values given H_0.** They use the well-known $\mathscr{X}^2$**-distribution**.

Chi Square Distribution

It is the distribution of the sum of squared standard normal random variables. More precisely if $Z_i \in N(0,1)$, $i = 1, 2, \ldots, m$ i.i.d. and $X = Z_1^2 + Z_2^2 + \ldots + Z_m^2$, then

$$X \in \mathscr{X}^2(m)$$

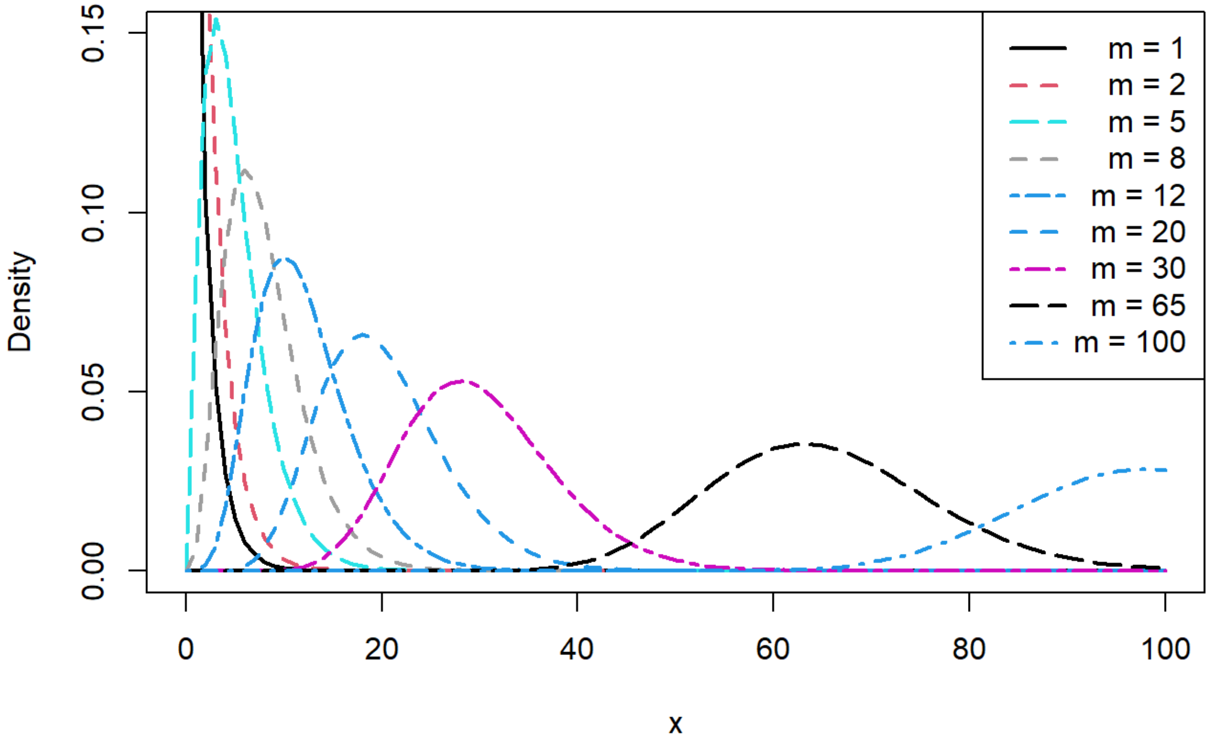
where m is the degrees of freedom.

The $\mathscr{X}^2(m)$ distribution peaks at $x = m - 2$ (its mode is equal to $m - 2$). Its mean is equal to m and its variance is equal to $2m$.

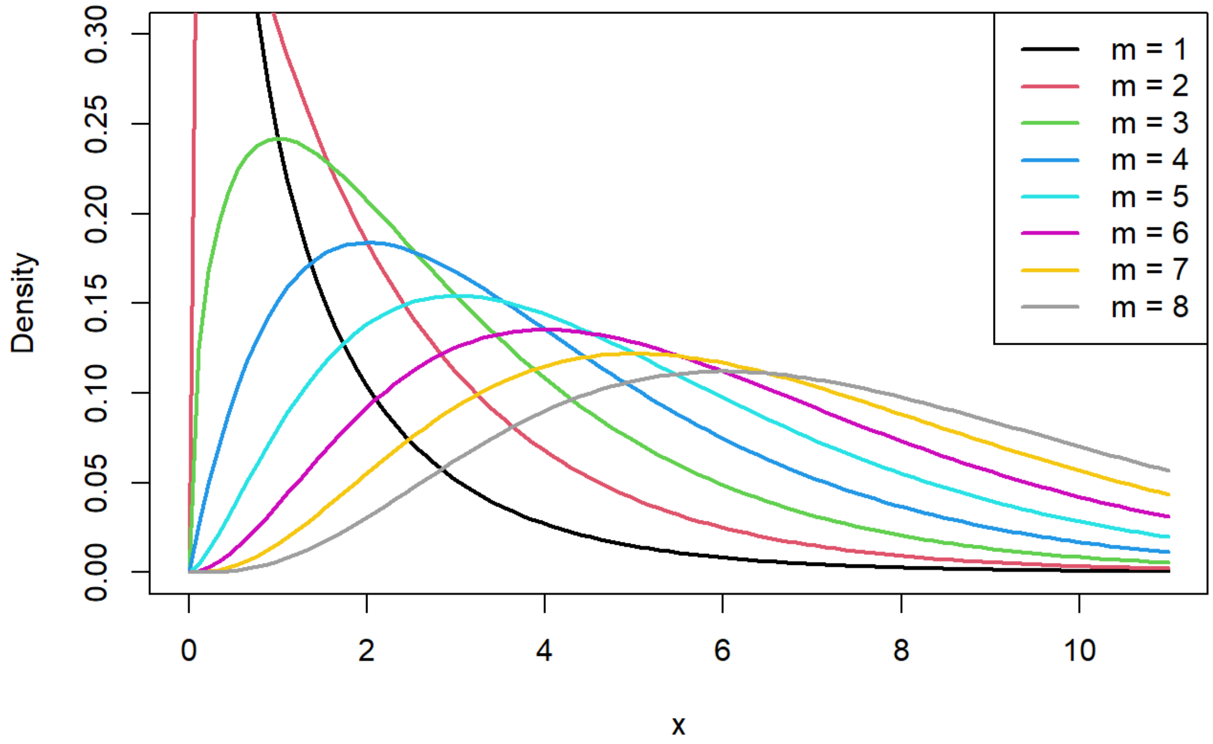
It is important to remind that for a small number of degrees of freedom this distribution is very skewed. The larger the degrees of freedom, the more symmetric is the distribution.

```
> curve(dchisq(x, 1), from = 0, to = 100, col = "1", lwd
= 2, ylim = c(0,0.15), type = "l", ylab = "Density", main
= "")
> m <- c(2, 5, 8, 12, 20, 30, 65, 100)
> for(i in m)
+   curve(dchisq(x, i), add = TRUE, lwd = 2, col = i, lty
= i)
> temp <- legend("topright", legend = rep(" ", 9),
text.width = 10, lty = c(1, m), lwd = rep("2", 9), col =
c(1, m), xjust = -1, yjust = 1)
```

```
> text(temp$rect$left + temp$rect$w, temp$text$y, c("m =
1", "m = 2",  "m = 5", "m = 8", "m = 12", "m = 20", "m =
30", "m = 65", "m = 100"), pos = 2)
```



```
> curve(dchisq(x, 1), from = 0, to = 11, col = "1", lwd =
2, ylim = c(0,0.3), type = "l", ylab = "Density", main =
" ")
> for(i in 2:8)
+   curve(dchisq(x, i), add = TRUE, lwd = 2, col = i)
> temp <- legend("topright", legend = rep(" ", 8), lwd =
rep("2", 8), text.width = 1, lty = rep(1, 8), col =
c(1:8), xjust = -1, yjust = 1)
> text(temp$rect$left + temp$rect$w, temp$text$y, c("m =
1", "m = 2", "m = 3", "m = 4", "m = 5", "m = 6", "m = 7",
"m = 8"), pos = 2)
```



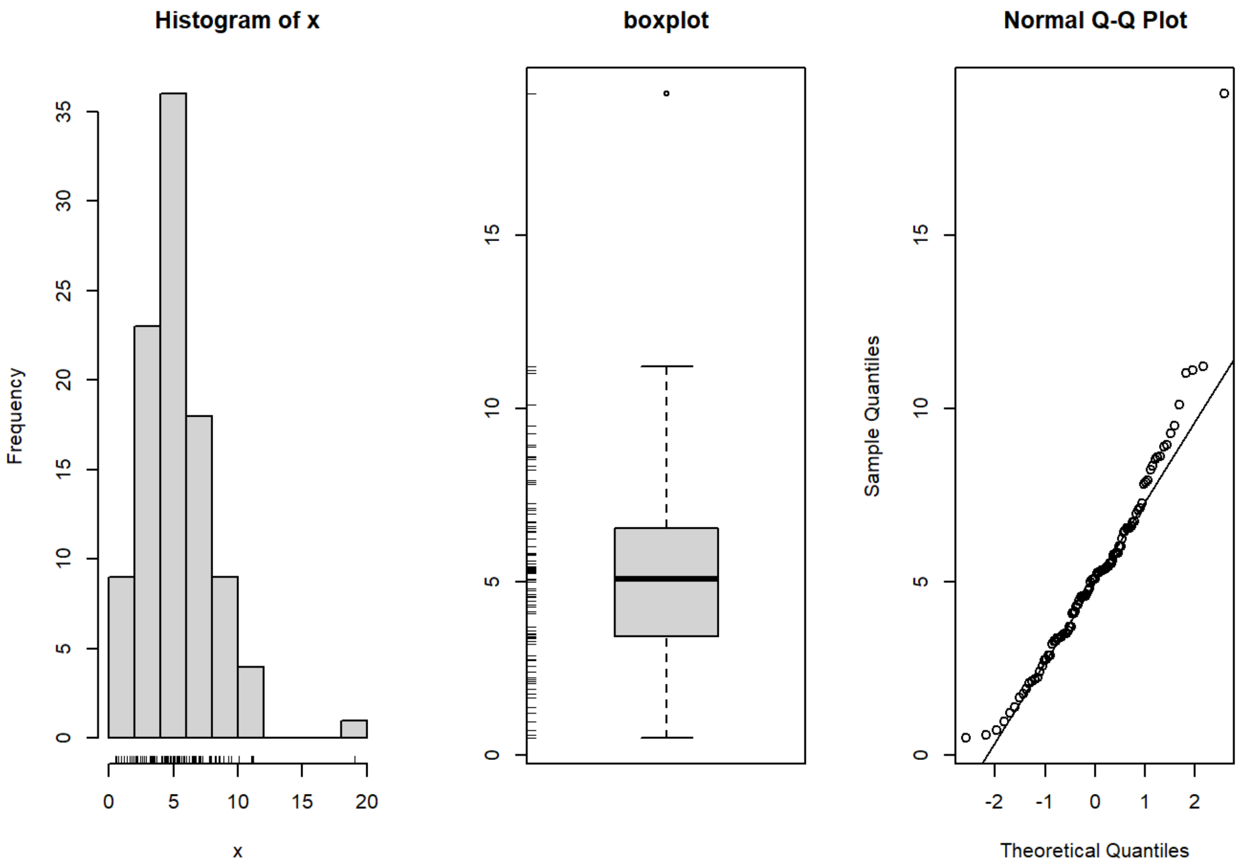
Let us perform **Shapiro test** for normality on a sample of 100 observations on $X \in \mathscr{X}^2(5)$.

```
> library(UsingR)
Warning: package 'UsingR' was built under R version 4.0.3
Loading required package: MASS
Loading required package: HistData
Loading required package: Hmisc
Loading required package: lattice
Loading required package: survival
Loading required package: Formula
Loading required package: ggplot2

Attaching package: 'Hmisc'
The following objects are masked from 'package:base':

    format.pval, units

Attaching package: 'UsingR'
The following object is masked from 'package:survival':
    cancer
> x <- rchisq(100, 5)
> simple.eda(x)
```



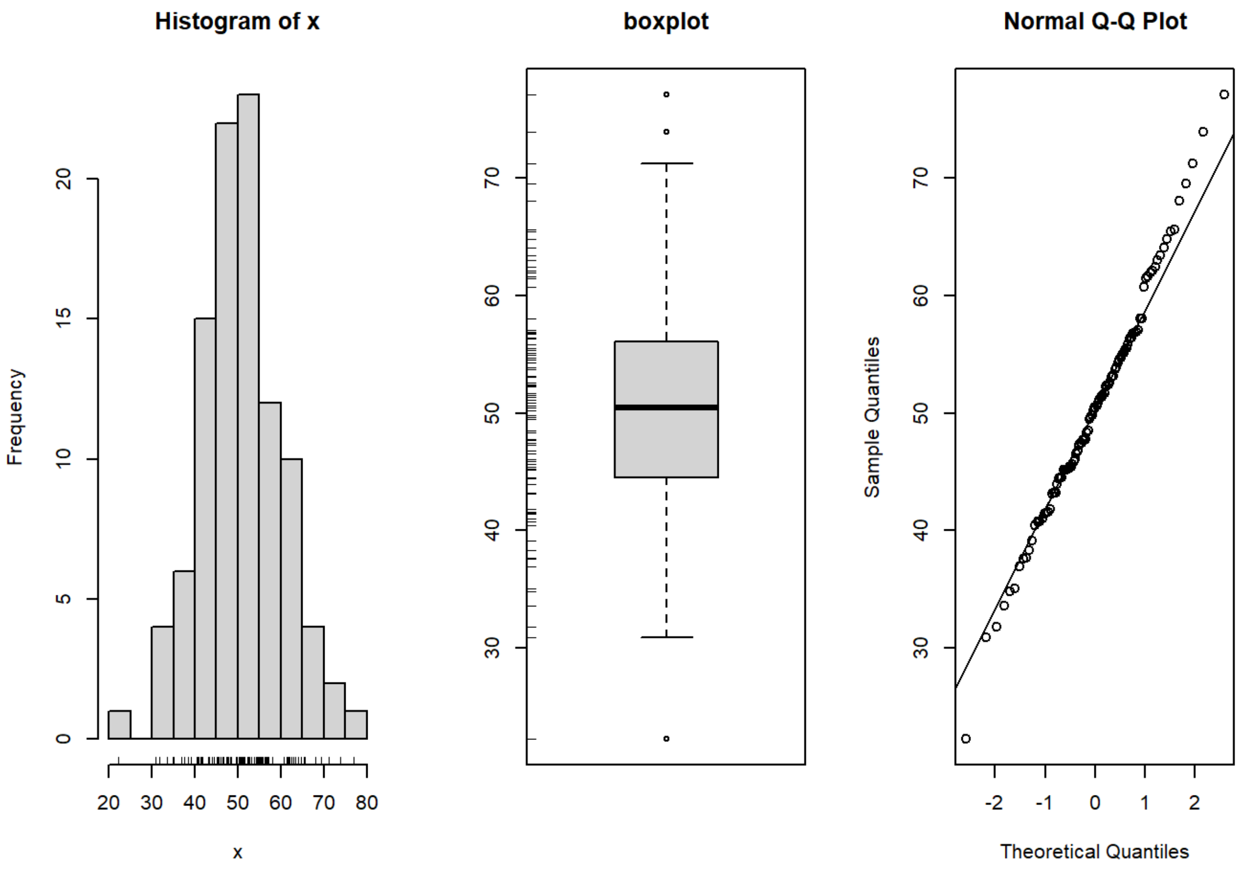
```
> shapiro.test(x)

    Shapiro-Wilk normality test

data:  x
W = 0.91967, p-value = 1.353e-05
```

Let us now perform **Shapiro test** for normality on a sample of 100 observations on $X \in \mathscr{X}^2(50)$

```
> x <- rchisq(100, 50)
> simple.eda(x)
```



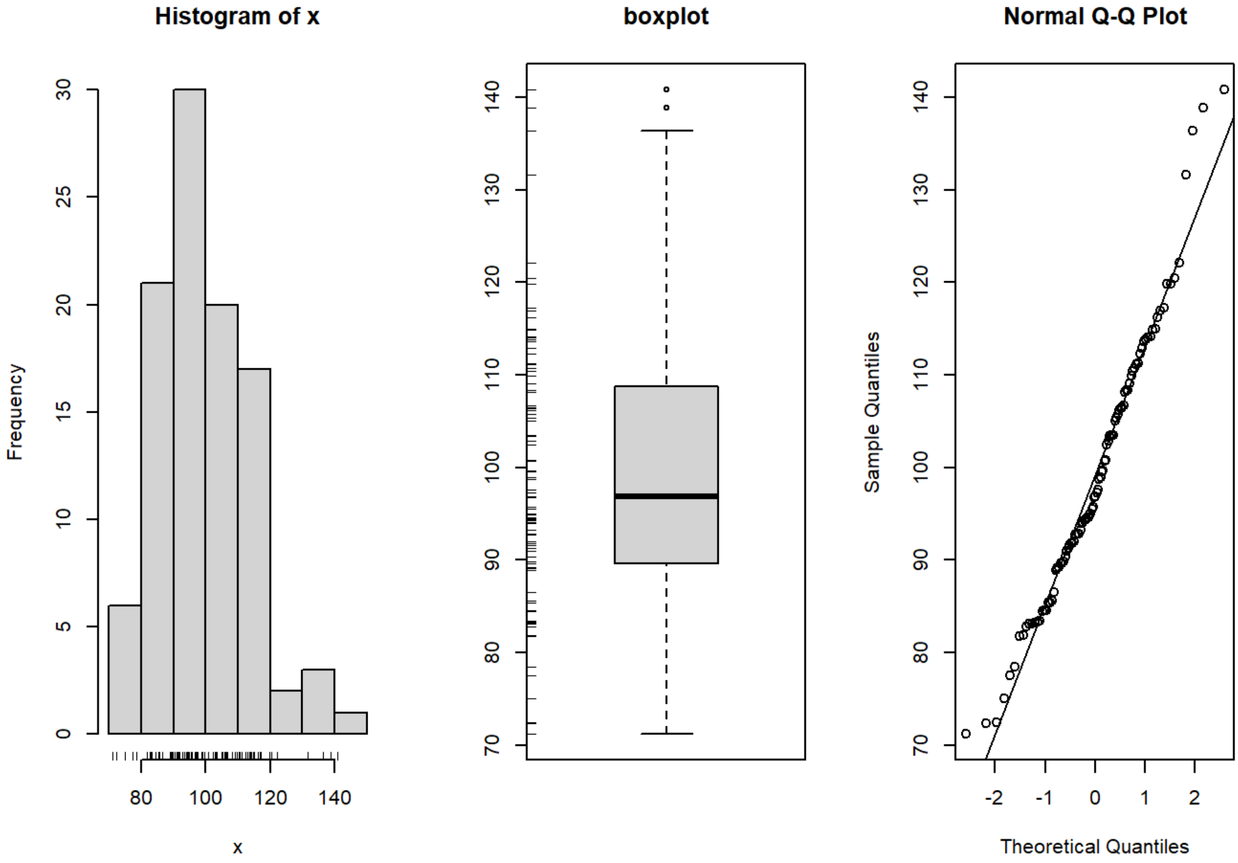
```
> shapiro.test(x)

	Shapiro-Wilk normality test

data:  x
W = 0.99287, p-value = 0.8799
```

Let us now perform **Shapiro test** for normality on a sample of 100 observations on $X \in \mathscr{X}^2(100)$

```
> x <- rchisq(100, 100)
> simple.eda(x)
```



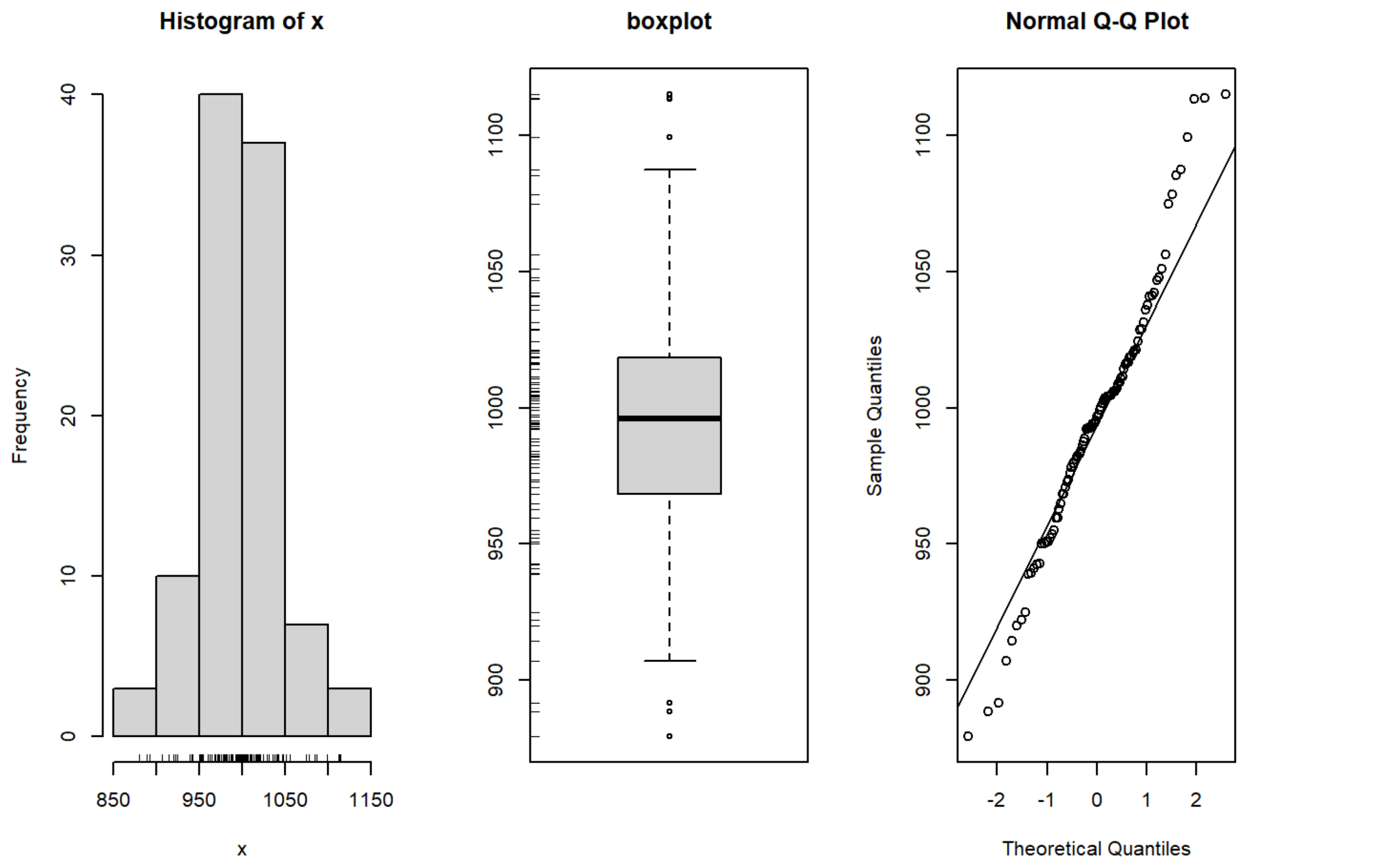
```
> shapiro.test(x)

    Shapiro-Wilk normality test

data:  x
W = 0.97342, p-value = 0.04061
```

Let us now perform **Shapiro test** for normality on a sample of 100 observations on $X \in \mathscr{X}^2(1000)$

```
> x <- rchisq(100, 1000)
> simple.eda(x)
```



```
> shapiro.test(x)

    Shapiro-Wilk normality test

data:  x
W = 0.97808, p-value = 0.09413
```

We observe that when the degrees of freedom are large enough the difference between $\mathscr{X}^2$ distribution and the normal one are not statistically significant.

Chi-squared Goodness of Fit Tests

Goodness of fit tests check if the data come from some specific population. These tests are based on asymptotic theorems, therefore, they work better with large samples.

The **chi-squared goodness of fit tests** allows us to check if data come from a distribution described in H_0. As far as if this distribution is continuous we have better tests. The **chi-squared goodness of fit**

tests are useful mainly if we compare empirical and discrete theoretical distribution. Moreover as far as the possible values of the observed random variable does not participate in the computations these tests are applicable also on categorical observed variables. Actually it reduces to the test that the observed distribution is **multivariate binomial** or the so, called **multinomial distribution**.

The tested hypothesis is H_0 : The observed distribution coincides with

Possible values	**Category 1**	**Category 2**	**...**	**Category k**	**Total**
Theoritical probability	p_1	p_2	...	p_k	1

If the sample size is n, then the number of observations ν_i in the i-th group, given H_0 would be $(\nu_i \,|\, H_0) \in Bi(n, p_i)$, $i = 1, 2, \ldots, k$. Therefore, the expected number of observations in each group, given H_0 would be

Possible values	**Category 1**	**Category 2**	**...**	**Category k**	**Total**
$\mathbb{E}(\nu_i \,|\, H_0)$	np_1	np_2	...	np_k	n

The expected count of category i according to H_0 is usually denoted by $e_i = np_i$.

Let us suppose that the tested distribution in H_0 has r unknown parameters estimated from the sample.

The alternative is usually one-sided although when the number of groups is large enough two-sided alternatives are also possible.

H_A: The observed distribution does not coincide with this theoretical one.

We chose level of significance α.

As a measure of discrepancy between the observed and the theoretical distribution described in H_0, this test uses the statistics

$$\sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}$$

If the data is i.i.d., if H_0 is correct, if the sample is large enough, if for all $i = 1, 2, \ldots, k$, $np_i \geq 1$ (as far as we can not divide to 0) and if $80\,\%$ of np_i, $i = 1, 2, \ldots, k$ are bigger than 5, then,

$$\left(\sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i} \,|\, H_0 \right) \overset{d}{\rightarrow} \mathcal{X}(k - r - 1)$$

Therefore, if the sample is large enough the one sided critical area is

$$W_\alpha = \left\{ \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i} \geq x_{1-\alpha, \mathcal{X}^2(k-r-1)} \right\}$$

Example 1

John have tossed a die 150 times and found that it has the following distribution

Face	**1**	**2**	**3**	**4**	**5**	**6**
Number of rolls	22	21	22	27	22	36

Help him to check if the die is regular.

If the dice is regular then we would expect the six categories (faces) to have the same chance to happen.

In $n = 150$ rolls we would expect each face to have about $(\nu_i \,|\, H_0) \in Bi\left(n, \frac{1}{6}\right)$. Therefore, H_0 : The observed distribution coincides with

Possible values	Category 1	Category 2	...	Category 6	Total
Theoretical probability	$\frac{1}{6}$	$\frac{1}{6}$	...	$\frac{1}{6}$	1

and the expected number of observations in each group, given H_0 would be

Possible values	Category 1	Category 2	...	Category 6	Total
$\mathbb{E}(\nu_i \| H_0)$	$np_1 = \frac{150}{6}$	$np_2 = \frac{150}{6}$	...	$np_6 = \frac{150}{6}$	$n = 150$

```
> 150/6
[1] 25
```

25 appearances.

Possible values	Category 1	Category 2	...	Category 6	Total
$\mathbb{E}(\nu_i \| H_0)$	25	25	...	25	$n = 150$

John obtains 6 points 36 times. Is this coincidence or perhaps something else?

The answer to this question is to look at **how far off the data is from the expected**. We denote by

f_i - the observed frequency of category i in the sample and by

e_i - the expected count of category i according to H_0.

$$H_0 : e_i = \frac{1}{6}, i = 1, 2, 3, 4, 5, 6$$

$$H_A : \exists i : e_i \neq \frac{1}{6}, i = 1, 2, 3, 4, 5, 6$$

In our case $k = 6$. The sample size is $n = 150$ - relatively large. We have not estimated parameters of the distribution in H_0 from the sample, therefore $r = 0$. The data is i.i.d., the conditions $e_i \geq 1$, $i = 1, 2, \ldots, k$, and the one to have at least $80\,\%$ e_i bigger than 5 are satisfied.

Therefore,

$$W_\alpha = \left\{ \sum_{i=1}^{6} \frac{(\nu_i - 25)^2}{25} \geq x_{1-\alpha, \mathcal{X}^2(6-0-1)} \right\}$$

Now we chose $\alpha = 0.05$ and compute $x_{1-\alpha, \mathcal{X}^2(5)} = 11.0705$.

```
> qchisq(0.95,5)
[1] 11.0705
```

Therefore,

$$W_\alpha = \left\{ \sum_{i=1}^{6} \frac{(\nu_i - 25)^2}{25} \geq 11.0705 \right\}$$

Now we can check if the sample is in the critical area W_α for H_0. Therefore, we replace the random variables ν_i, $i = 1, 2, \ldots, 6$ correspondingly with $f_1, f_2, \ldots, f_6$ observed in the sample

$$x_{emp} = \sum_{i=1}^{6} \frac{(f_i - 25)^2}{25} =$$

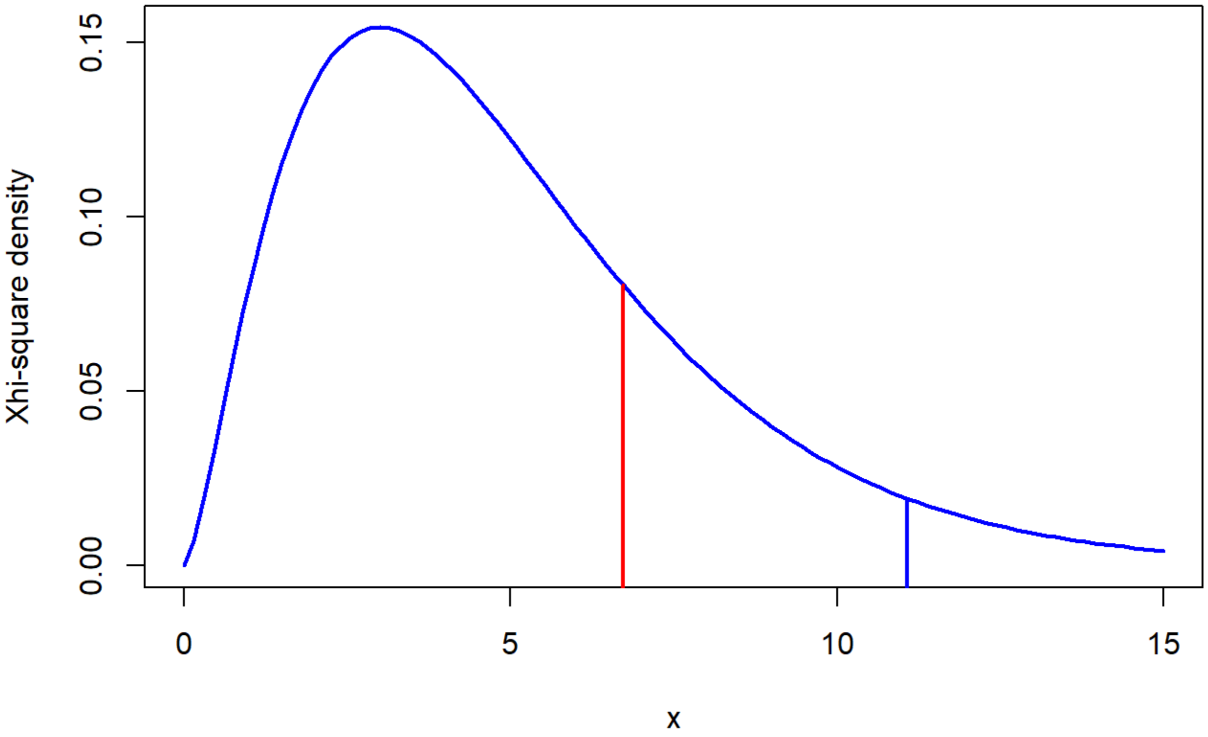
$$= \frac{(22-25)^2}{25} + \frac{(21-25)^2}{25} + \frac{(22-25)^2}{25} + \frac{(27-25)^2}{25} + \frac{(22-25)^2}{25} + \frac{(36-25)^2}{25} = 6.72$$

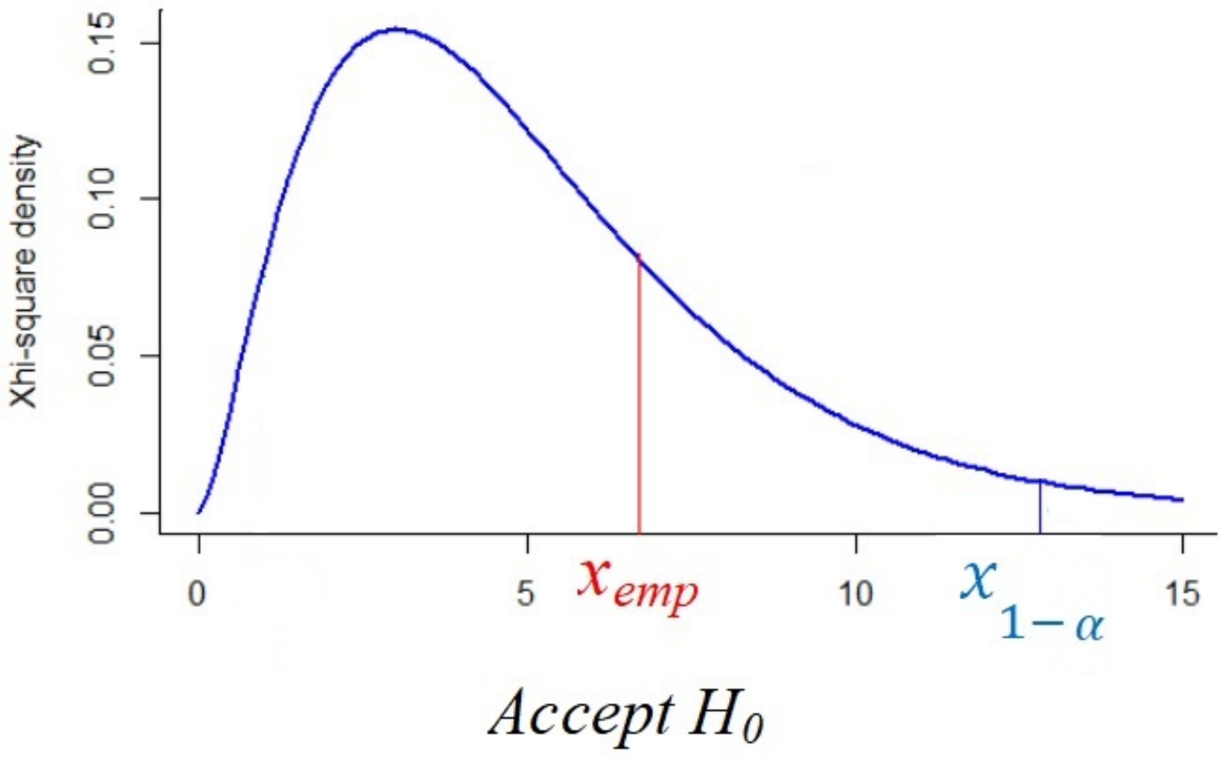
```
> f <- c(22, 21, 22, 27, 22, 36)
> e <- sum(f) * 1/6
> xemp <- sum((f-e)^2/e); xemp
```

```
[1] 6.72
```

Now, we compare $x_{emp=6.72}$ with $x_{1-\alpha,\mathscr{X}^2(5)} = 11.0705$

```
> df <- length(f) - 1
> alpha <- 0.05
> curve(dchisq(x, df), from = 0, to = 15, col = "blue",
lwd = 2, ylab = "Xhi-square density")
> x <- qchisq(1 - alpha, df)
> segments(x, -1, x, dchisq(x, df), col = "blue", lwd =
2)
> segments(xemp, -1, xemp, dchisq(xemp, df), col = "red",
lwd = 2)
```



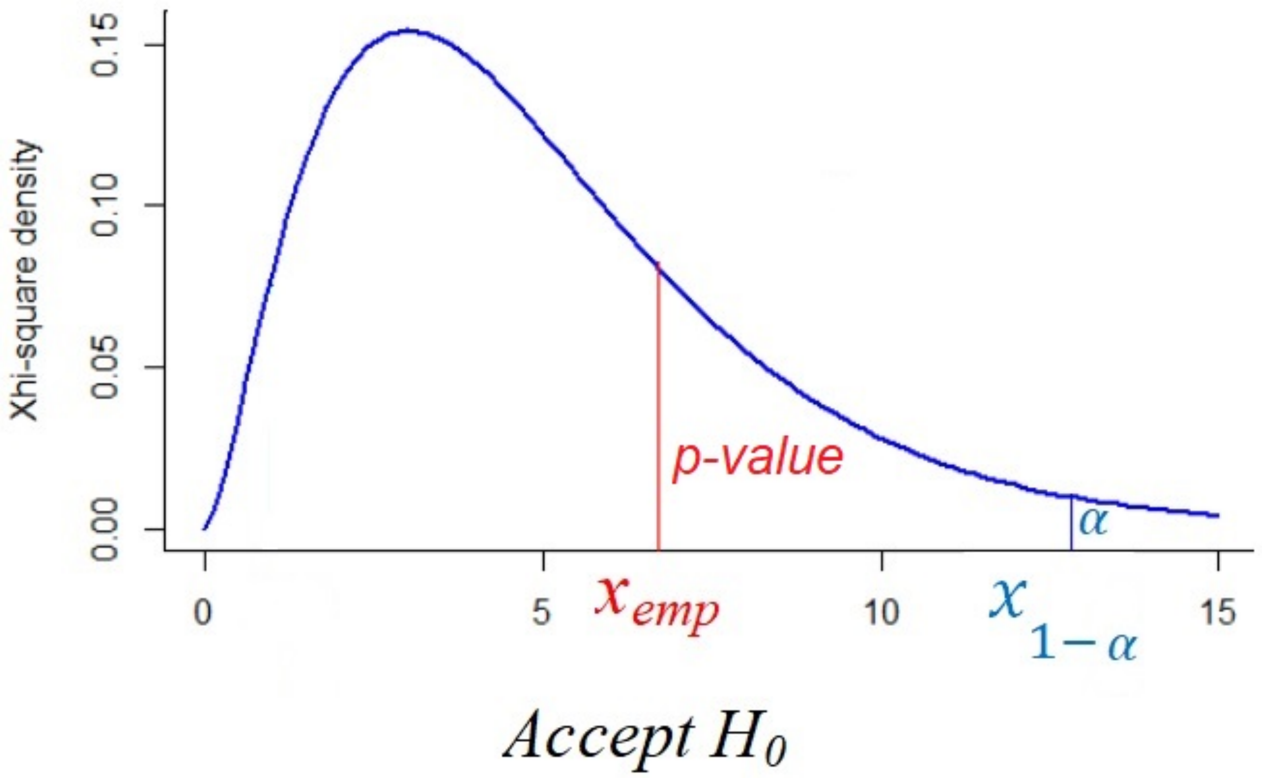


and conclude that the sample is not in the critical area for H_0. Therefore, we have no evidence to reject H_0.

Another way to make the same conclusion is via the $p-value$. It is the probability of the event that a $\mathscr{X}^2$ random variable having 5 df is bigger than x_{emp}

$$\mathbb{P}\left(\sum_{i=1}^{6}\frac{(\nu_i-25)^2}{25}>x_{emp}\right)=\mathbb{P}(\eta>6.72)\approx 0.2423, \eta\in\mathscr{X}^2(5)$$

```
> pchisq(xemp, df, lower.tail = FALSE)
[1] 0.2423109
```



The $p-value = 0.2423109 > 0.05 = \alpha$, so we have no evidence to reject H_0.

The easiest way to make the same (only in cases when the alternative is once sided) is to use the build in function `chisq.test`

```
> freq <- c(22, 21, 22, 27, 22, 36)
> prob <- rep(1/6, 6)
> chisq.test(freq, p = prob)

    Chi-squared test for given probabilities

data:  freq
X-squared = 6.72, df = 5, p-value = 0.2423
```

The $p-value = 0.2423 > 0.05 = \alpha$, so we have no evidence to reject H_0. We can assume that the die is fair.
The $X - squared = x_{emp} = 6.72$ and $df = 5$.

.

Example 2

Let us now suppose that we have the original data (not only the frequencies in the groups). More precisely in order to trow a fair die 150 times let us generate 150 observations on a Discrete Uniform random variable variable.

```
> set.seed(1)
> y <- sample(1:6, 150, replace = TRUE)
> freq <- table(y); freq
y
 1  2  3  4  5  6 
28 25 23 23 22 29 
```

H_0 : The observed distribution coincides with

Possible values	Category 1	Category 2	...	Category 6	Total
Theoretical probability	$\frac{1}{6}$	$\frac{1}{6}$	...	$\frac{1}{6}$	1

and the expected number of observations in each group, given H_0 would be

Possible values	Category 1	Category 2	...	Category 6	Total
$\mathbb{E}(\nu_i \mid H_0)$	$np_1 = \frac{150}{6}$	$np_2 = \frac{150}{6}$	...	$np_6 = \frac{150}{6}$	$n = 150$

```
> 150/6
[1] 25
```

25 appearances.

Possible values	Category 1	Category 2	...	Category 6	Total
$\mathbb{E}(\nu_i \mid H_0)$	25	25	...	25	$n = 150$

$H_A : \exists i : e_i \neq \frac{1}{6}, i = 1, 2, 3, 4, 5, 6$

Again $k = 6$, the sample size is $n = 150$. We assume that it is large. We have not estimated parameters of the distribution in H_0 from the sample, therefore $r = 0$. The data is i.i.d., the conditions $e_i \geq 1$, $i = 1, 2, \ldots, k$, and the one to have at least $80\,\%$ e_i bigger than 5 are satisfied.

Therefore,

$$W_\alpha = \left\{ \sum_{i=1}^{6} \frac{(\nu_i - 25)^2}{25} \geq x_{1-\alpha,\mathcal{X}^2(6-0-1)} \right\}$$

Now we chose $\alpha = 0.05$ and compute $x_{1-\alpha,\mathcal{X}^2(5)} = 11.0705$.

```
> qchisq(0.95,5)
[1] 11.0705
```

Therefore, the critical area is the same

$$W_\alpha = \left\{ \sum_{i=1}^{6} \frac{(\nu_i - 25)^2}{25} \geq 11.0705 \right\}$$

Now we can check if the sample is in the critical area W_α for H_0. Therefore, we replace the random variables ν_i, $i = 1, 2, \ldots, 6$ correspondingly with $f_1, f_2, \ldots, f_6$ observed in the sample

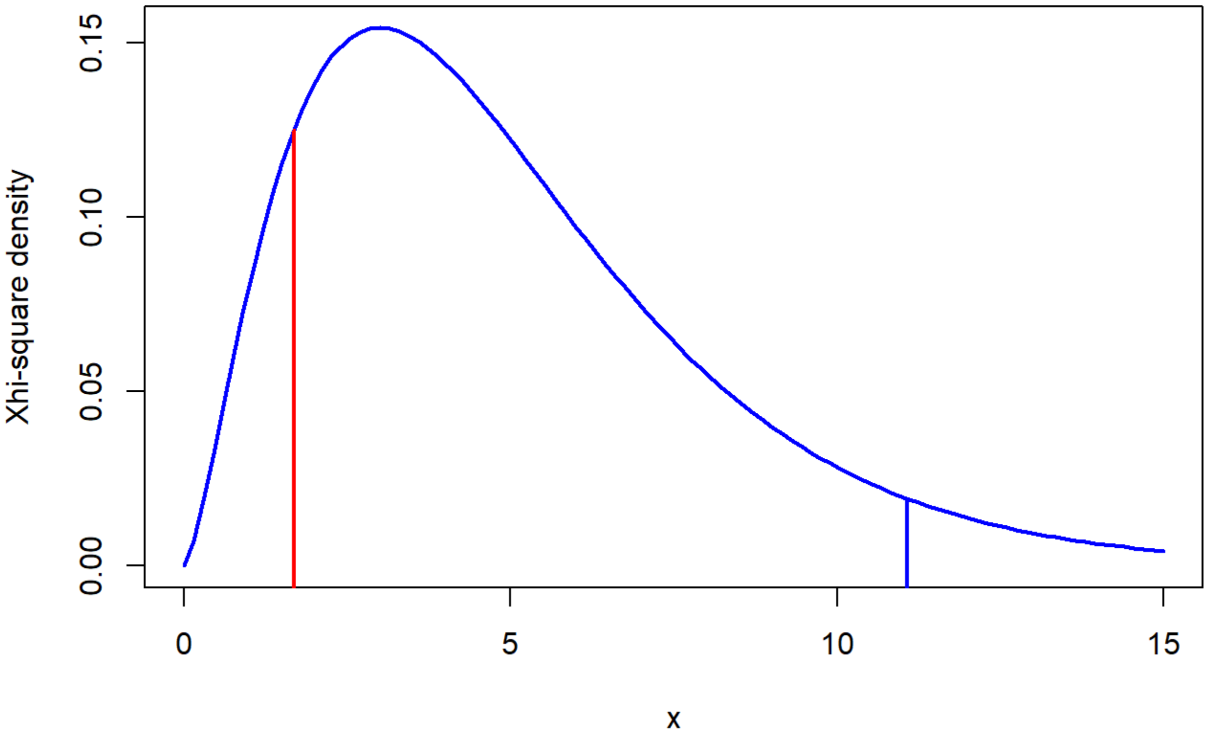
$$x_{emp} = \sum_{i=1}^{6} \frac{(f_i - 25)^2}{25} =$$

$$= \frac{(28-25)^2}{25} + \frac{(25-25)^2}{25} + \frac{(23-25)^2}{25} + \frac{(23-25)^2}{25} + \frac{(22-25)^2}{25} + \frac{(29-25)^2}{25} = 1.68$$

```
> f <- c(28, 25, 23, 23, 22, 29)
> e <- sum(f) * 1/6
> xemp <- sum((f-e)^2/e); xemp
[1] 1.68
```

Now, we compare $x_{emp} = 1.68$ with $x_{1-\alpha,\mathscr{X}^2(5)} = 11.0705$

```
> df <- length(f) - 1
> alpha <- 0.05
> curve(dchisq(x, df), from = 0, to = 15, col = "blue",
lwd = 2, ylab = "Xhi-square density")
> x <- qchisq(1 - alpha, df)
> segments(x, -1, x, dchisq(x, df), col = "blue", lwd =
2)
> segments(xemp, -1, xemp, dchisq(xemp, df), col = "red",
lwd = 2)
```



and conclude that the sample is not in the critical area for H_0. Therefore, we have no evidence to reject H_0.

The $p-value$ is the probability of the event that a $\mathscr{X}^2$ random variable having 5 df is bigger than x_{emp}

$$\mathbb{P}\left(\sum_{i=1}^{6}\frac{(\nu_i-25)^2}{25}>x_{emp}\right)=\mathbb{P}(\nu>1.68)\approx 0.8914097,$$
$\eta \in \mathscr{X}^2(5)$

```
> pchisq(xemp, df, lower.tail = FALSE)
[1] 0.8914097
```

The $p-value = 0.8914097 > 0.05 = \alpha$, so we have no evidence to reject H_0.

Let us now use the `chisq.test` function

```
> f <- c(28, 25, 23, 23, 22, 29)
> prob <- rep(1/6, 6)
> chisq.test(f, p = prob)

	Chi-squared test for given probabilities

data:  f
X-squared = 1.68, df = 5, p-value = 0.8914
```

The $p-value$, so we have no evidence to reject H_0. We can assume that the die is fair. The $X-squared = x_{emp} = 1.68$ and $df = 5$.

Example 3 (The sample size is not large enough)

Let's try with different frequencies. The last means that John trows an unfair die 150 times. In order to see the results let us
generate 150 observations on a Discrete Uniform random variable with probability mass function

Possible values	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Theoretical probability	$\frac{0.75}{6}$	$\frac{0.75}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1.25}{6}$	$\frac{1.25}{6}$	1

and to see the performance of the same test.

```
> set.seed(1)
> y <- sample(1:6, 150, replace = TRUE,
+              prob = c(0.75, 0.75, 1, 1, 1.25, 1.25)/6)
> freq <- table(y); freq
y
 1  2  3  4  5  6 
16 19 27 30 26 32 
> probs <- c(1, 1, 1, 1, 1, 1)/6
```

We formulate the same

H_0 : The observed distribution coincides with

Possible values	Category 1	Category 2	...	Category 6	Total
Theoretical probability	$\frac{1}{6}$	$\frac{1}{6}$	...	$\frac{1}{6}$	1

and the expected number of observations in each group, given H_0 would be

Possible values	Category 1	Category 2	...	Category 6	Total
$\mathbb{E}(\nu_i \mid H_0)$	$np_1 = \frac{150}{6}$	$np_2 = \frac{150}{6}$	...	$np_6 = \frac{150}{6}$	$n = 150$

```
> 150/6
[1] 25
```

25 appearances.

Possible values	Category 1	Category 2	...	Category 6	Total
$\mathbb{E}(\nu_i \mid H_0)$	25	25	...	25	$n = 150$

$$H_A : \exists i : e_i \neq \frac{1}{6}, \, i = 1, 2, 3, 4, 5, 6$$

Again $k = 6$, the sample size is $n = 150$. We assume that it is large. We have not estimated parameters of the distribution in H_0 from the sample, therefore $r = 0$. The data is i.i.d., the conditions $e_i \geq 1, 2, \ldots, k$, and the one to have at least $80\,\%$ e_i bigger than 5 are satisfied.

Therefore,

$$W_\alpha = \left\{ \sum_{i=1}^{6} \frac{(\nu_i - 25)^2}{25} \geq x_{1-\alpha, \mathcal{X}^2(6-0-1)} \right\}$$

Now we chose $\alpha = 0.05$ and compute $x_{1-\alpha,\mathcal{X}^2(5)} = 11.0705$.

```
> qchisq(0.95,5)
[1] 11.0705
```

Therefore, the critical area is the same

$$W_\alpha = \left\{ \sum_{i=1}^{6} \frac{(\nu_i - 25)^2}{25} \geq 11.0705 \right\}$$

Now we can check if the sample is in the critical area W_α for H_0. Therefore, we replace the random variables ν_i, $i = 1, 2, \ldots, 6$ correspondingly with $f_1, f_2, \ldots, f_6$ observed in the sample

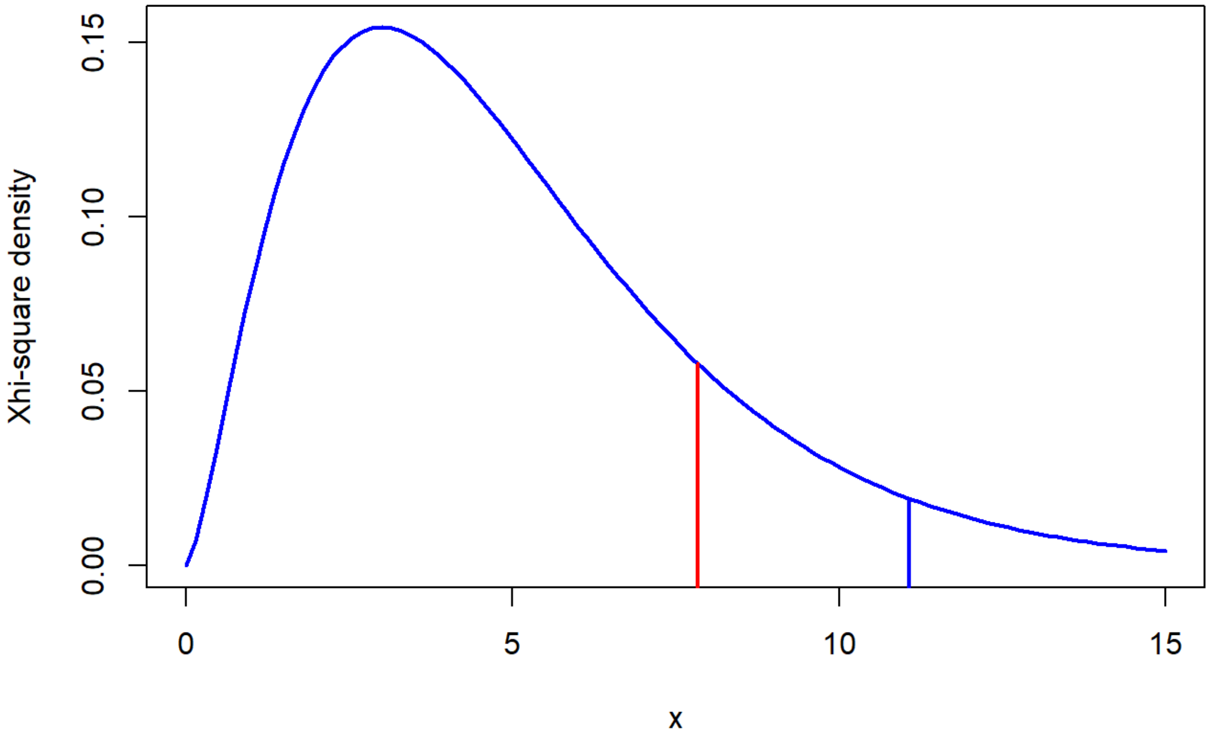
$$x_{emp} = \sum_{i=1}^{6} \frac{(f_i - 25)^2}{25} =$$

$$= \frac{(16-25)^2}{25} + \frac{(19-25)^2}{25} + \frac{(27-25)^2}{25} + \frac{(30-25)^2}{25} + \frac{(26-25)^2}{25} + \frac{(32-25)^2}{25} = 7.84$$

```
> f <- c(16, 19, 27, 30, 26, 32)
> e <- sum(f) * 1/6
> xemp <- sum((f - e)^2 / e); xemp
[1] 7.84
```

Now, we compare x_{emp} with $x_{1-\alpha,\mathcal{X}^2(5)} = 11.0705$

```
> df <- length(f) - 1
> alpha <- 0.05
> curve(dchisq(x, df), from = 0, to = 15, col = "blue", 
lwd = 2, ylab = "Xhi-square density")
> x <- qchisq(1 - alpha, df)
> segments(x, -1, x, dchisq(x, df), col = "blue", lwd = 
2)
> segments(xemp, -1, xemp, dchisq(xemp, df), col = "red", 
lwd = 2)
```



and conclude that the sample is not in the critical area for H_0. Therefore, we have no evidence to reject H_0.

Here we observe that when the sample size is not large enough we can obtain wrong results.

From computational point of view we can proceed further on.
The $p-value$ is the probability of the event that a $\mathscr{X}^2$ random variable having 5 df is greater than x_{emp}

$$\mathbb{P}\left(\sum_{i=1}^{6}\frac{(\nu_i-25)^2}{25} > x_{emp}\right) = \mathbb{P}(\eta > 7.84) \approx 0.1652768,$$

$$\eta \in \mathscr{X}^2(5)$$

```
> pchisq(xemp, df, lower.tail = FALSE)
[1] 0.1652768
```

The $p-value = 0.1652768 > 0.05 = \alpha$, so we have no evidence to reject H_0.

Let us now use the `chisq.test` function

```
> f <- c(16, 19, 27, 30, 26, 32)
> prob <- rep(1/6, 6)
> chisq.test(f, p = prob)

    Chi-squared test for given probabilities

data:  f
X-squared = 7.84, df = 5, p-value = 0.1653
```

The $p-value = 0.1653 > 0.05 = \alpha$, so we have no evidence to reject H_0. We can assume that the die is fair.
The $X-squared = x_{emp} = 7.84$ and $df = 5$.

Example 4

Let's us now increase the sample size. Suppose that John trows an unfair die 1500 times. In order to see the results let us generate 1500 observations on a Discrete Uniform random variable with probability mass function

Possible values	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Theoretical probability	$\frac{0.75}{6}$	$\frac{0.75}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1.25}{6}$	$\frac{1.25}{6}$	1

and to see the performance of the same test.

```
> set.seed(1)
> y <- sample(1:6, 1500, replace = TRUE,
+             prob = c(0.75, 0.75, 1, 1, 1.25, 1.25)/6)
> freq <- table(y); freq
y
  1   2   3   4   5   6 
192 178 242 232 326 330 
```

We formulate the same

H_0 : The observed distribution coincides with

Possible values	**Category 1**	**Category 2**	**...**	**Category 6**	**Total**
Theoretical probability	$\frac{1}{6}$	$\frac{1}{6}$	...	$\frac{1}{6}$	1

and the expected number of observations in each group, given H_0 would be

Possible values	**Category 1**	**Category 2**	**...**	**Category 6**	**Total**
$\mathbb{E}(\nu_i \mid H_0)$	$np_1 = \frac{1500}{6}$	$np_2 = \frac{1500}{6}$	...	$np_6 = \frac{1500}{6}$	$n = 1500$

```
> 150/6
[1] 25
```

25 appearances.

Possible values	Category 1	Category 2	...	Category 6	Total
$\mathbb{E}(\nu_i \mid H_0)$	250	250	...	250	$n = 1500$

$$H_A : \exists i : e_i \neq \frac{1}{6}, i = 1, 2, 3, 4, 5, 6$$

Again $k = 6$, the sample size is $n = 1500$. We assume that it is large. We have not estimated parameters of the distribution in H_0 from the sample, therefore $r = 0$. The data is i.i.d., the conditions $e_i \geq 1$, $i = 1, 2, \ldots, k$, and the one to have at least $80\,\%$ e_i bigger than 5 are satisfied.

Therefore,

$$W_\alpha = \left\{ \sum_{i=1}^{6} \frac{(\nu_i - 25)^2}{25} \geq x_{1-\alpha, \mathcal{X}^2(6-0-1)} \right\}$$

Now we chose $\alpha = 0.05$ and compute $x_{1-\alpha, \mathcal{X}^2(5)} = 11.0705$

```
> qchisq(0.95, 5)
[1] 11.0705
```

Therefore, the critical area is the same

$$W_\alpha = \left\{ \sum_{i=1}^{6} \frac{(\nu_i - 25)^2}{25} \geq 11.0705 \right\}$$

Now we can check if the sample is in the critical area W_α for H_0. Therefore, we replace the random variables ν_i, $i = 1, 2, \ldots, 6$ correspondingly with $f_1, f_2, \ldots, f_6$ observed in the sample

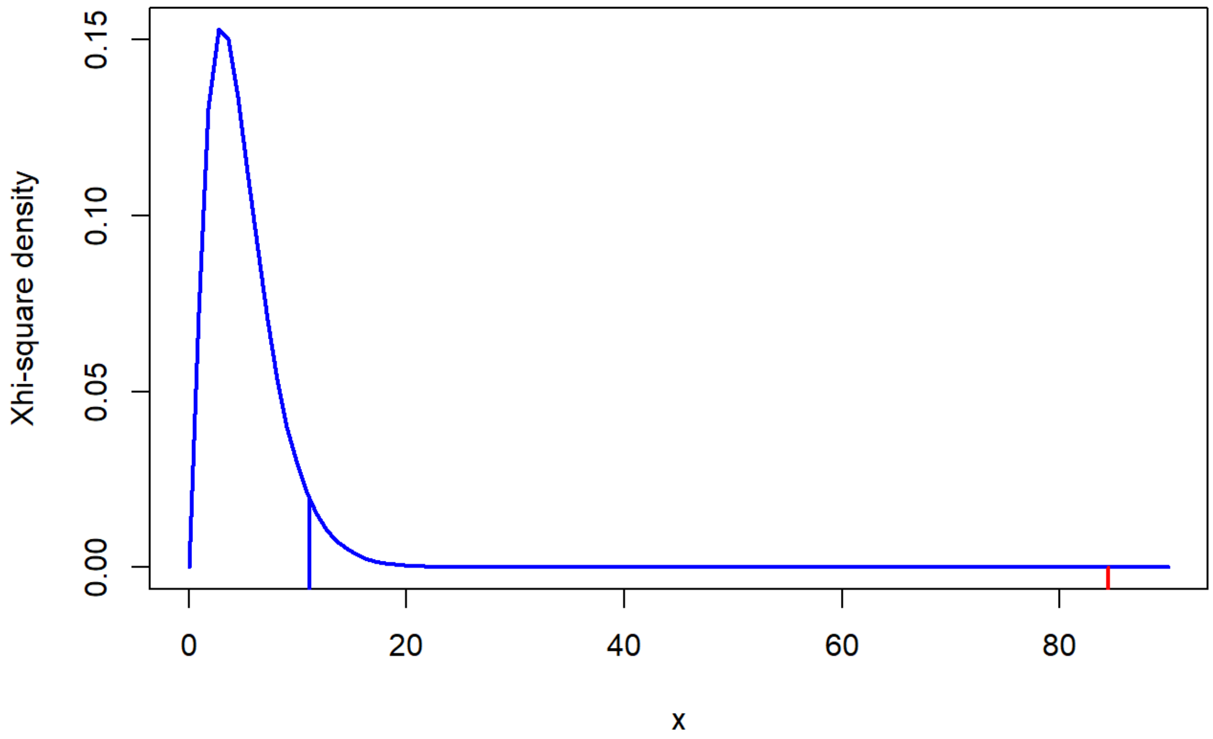
$$x_{emp} = \sum_{i=1}^{6} \frac{(f_i - 25)^2}{25} =$$

$$= \frac{(192-250)^2}{250} + \frac{(178-250)^2}{250} + \frac{(242-250)^2}{250} + \frac{(232-250)^2}{250} + \frac{(326-250)^2}{250} + \frac{(330-250)^2}{250} = 84.448$$

```
> f <- c(192, 178, 242, 232, 326, 330)
> e <- sum(f) * 1/6
> xemp <- sum((f-e)^2/e); xemp
[1] 84.448
```

Now, we compare $x_{emp} = 84.448$ with $x_{1-\alpha,\mathscr{X}^2(5)} = 11.0705$

```
> df <- length(f) - 1
> alpha <- 0.05
> curve(dchisq(x, df), from = 0, to = 90, col =
"blue",lwd=2, ylab="Xhi-square density")
> x <- qchisq(1 - alpha, df)
> segments(x, -1, x, dchisq(x, df), col = "blue", lwd =
2)
> segments(xemp, -1, xemp, dchisq(xemp, df), col = "red",
lwd = 2)
```



and conclude that the sample is in the critical area for H_0. Therefore, we reject H_0. Now the result is correct as far as the sample size is already large enough.

The $p-value$ is the probability of the event that a $\mathscr{X}^2$ random variable having 5 df is greater than x_{emp}

$$\mathbb{P}\left(\sum_{i=1}^{6}\frac{(\nu_i-250)^2}{250}>x_{emp}\right)=\mathbb{P}(\eta>84.448)\approx 9.826244e-17,\ \eta\in\mathscr{X}^2(5)$$

```
> pchisq(xemp, df, lower.tail = FALSE)
[1] 9.826244e-17
```

The $p-value = 9.826244e-17 < 0.05 = \alpha$, so we reject H_0.

Let us now use the build in function `chisq.test`

```
> f <- c(192, 178, 242, 232, 326, 330)
> prob <- rep(1/6, 6)
> chisq.test(f, p = prob)
```

```
    Chi-squared test for given probabilities

data:  f
X-squared = 84.448, df = 5, p-value < 2.2e-16
```

The $p-value < 2.2e-16 < 0.05 = \alpha$, so we reject H_0. We cannot assume that the die is fair.
The $X-squared = x_{emp} = 84.448$ and $df = 5$.

Example 5

Company printed baseball cards. It claimed that:

- $30\,\%$ of its cards were rookies;
- $60\,\%$ were veterans but not All-Stars;
- and $10\,\%$ were veteran All-Stars.

Random sample is taken of 100 cards:

- 50 of the cards were rookies;
- 45 were veterans but not All-Stars;
- and 5 were veteran All-Stars.

See whether our sample distribution differed statistically significantly from the distribution claimed by the company? Use significance level 0.05.

We formulate

Possible values	**rookies**	**veterans but not All-Stars**	**veteran All-Stars**	**Total**
Theoretical probability	0.3	0.6	0.1	1

and the expected number of observations in each group, given H_0 would be

Possible values	rookies	veterans but not All-Stars	veteran All-Stars	Total
$\mathbb{E}(\nu_i \mid H_0)$	$np_1 = 100 \times 0.3 = 30$	$np_2 = 60$	$np_3 = 10$	$n = 100$

H_A : There exists statistically significant difference between tested and the observed distribution.

Now $k = 3$, the sample size is $n = 100$. We assume that it is large. We have not estimated parameters of the distribution in H_0 from the sample, therefore $r = 0$. The data is i.i.d., the conditions $e_i \geq 1$, $i = 1, 2, 3$, and the one to have at least $80\,\%$ bigger than 5 are satisfied.

Therefore,

$$W_\alpha = \left\{ \sum_{i=1}^{3} \frac{(\nu_i - np_i)^2}{np_i} \geq x_{1-\alpha, \mathcal{X}^2(3-0-1)} \right\}$$

Now we chose $\alpha = 0.05$ and compute $x_{1-\alpha, \mathcal{X}^2(2)=5.991465}$.

```
> qchisq(0.95,2)
[1] 5.991465
```

Therefore, the critical area is

$$W_\alpha = \left\{ \sum_{i=1}^{3} \frac{(\nu_i - np_i)^2}{np_i} \geq 5.991465 \right\}$$

Now we can check if the sample is in the critical area W_α for H_0. Therefore, we replace the random variables ν_i, $i = 1, 2, 3$ correspondingly with f_1, f_2, f_3 observed in the sample

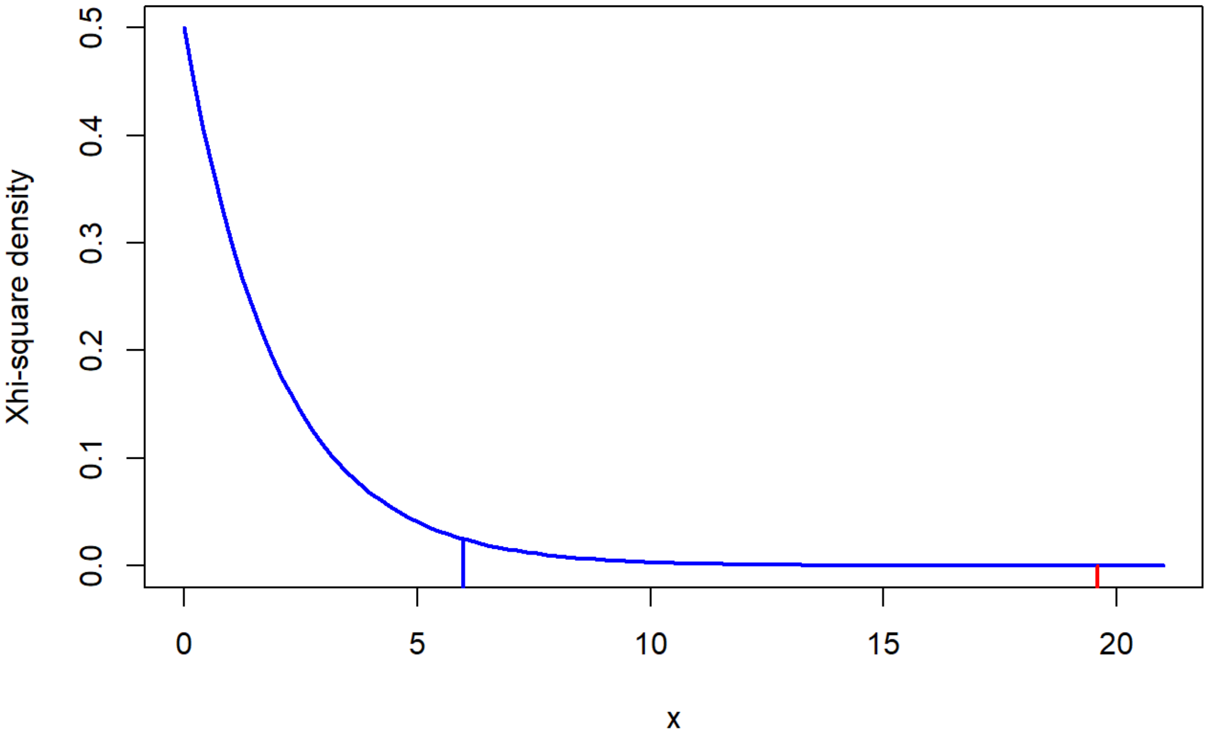
$$x_{emp} = \sum_{i=1}^{3} \frac{(f_i - np_i)^2}{np_i} =$$

$$= \frac{(50-30)^2}{30} + \frac{(45-60)^2}{60} + \frac{(5-10)^2}{10} = 19.58333$$

```
> f <- c(50, 45, 5)
> e <- c(30, 60, 10)
> xemp <- sum((f - e)^2 / e)
> xemp
[1] 19.58333
```

Now, we compare $x_{emp} = 19.58333$ with $x_{1-\alpha,\mathcal{X}^2(2)} = 5.991465$

```
> df <- length(f) - 1
> alpha <- 0.05
> curve(dchisq(x, df), from = 0, to = 21, col = "blue",
lwd = 2, ylab = "Xhi-square density")
> x <- qchisq(1 - alpha, df)
> segments(x, -1, x, dchisq(x, df), col = "blue", lwd =
2)
> segments(xemp, -1, xemp, dchisq(xemp, df), col = "red",
lwd = 2)
```



and conclude that the sample is in the critical area for H_0. Therefore, we reject H_0.

The $p-value$ is the probability of the event that a $\mathscr{X}^2$ random variable having 2 df is greater than x_{emp}

$$\mathbb{P}\left(\sum_{i=1}^{3}\frac{(\nu_i - np_i)^2}{np_i} > x_{emp}\right) = \mathbb{P}(\eta > 19.68333) \approx 5.5915634e-5,$$
$$\eta \in \mathscr{X}^2(4)$$

```
> pchisq(xemp, df, lower.tail = FALSE)
[1] 5.591563e-05
```

The $p-value = 5.591563e-5 < 0.05 = \alpha$, so we reject H_0.

Let us now use the `chisq.test` function

```
> f <- c(50,45,5)
> prob <- c(30,60,10)/100
> chisq.test(f, p = prob)

    Chi-squared test for given probabilities

data:  f
X-squared = 19.583, df = 2, p-value = 5.592e-05
```

The $p-value = 5.592e-05 < 0.05 = \alpha$, so we reject H_0. We cannot say that the company is correct. The sample distribution differed statistically significantly from the distribution claimed by the company.

Example 6

Letter distribution.

The 6 most popular letters in the English language are E, T, A, N, R, O and their frequencies are:

- $E - 12\,\%$
- $T - 9\,\%$
- $A - 8\,\%$
- $N - 7\,\%$
- $R - 7\,\%$
- $O - 7\,\%$
- $Another - 50\,\%$

A text is analyzed and the number of E, T, A, N, R, O's and others are counted:

- $E - 100$
- $T - 110$
- $A - 90$
- $N - 80$
- $R - 55$
- $O - 14$
- $Another - 600$

Is this text from the English language?

$H_0 : p_E = 0.12, p_T = 0.09, p_A = 0.08, p_N = 0.07, p_R = 0.07, p_O = 0.07,$ $p_{Another} = 0.5$

According to H_0 the text is written in English. The differences between the theoretical and empirical frequencies are not statistically significant.

H_A : At least one category doesn't have this specified probability.

Note: chi-squared test requires independence of each letter, so this is not quite appropriate, but let's suppose the letters are independent.

```
> f <- c(100, 110, 90, 80, 55, 14, 600)
> probs = c(12, 9, 8, 7, 7, 7, 50) / 100
> chisq.test(f, p = probs)
```

```
    Chi-squared test for given probabilities
```

```
data:  f
X-squared = 72.516, df = 6, p-value = 1.245e-13
```

$p-value = 1.245e-13 < 0.05 = \alpha$, so we reject H_0. The text is not written in English. The differences between the theoretical and empirical frequencies are statistically significant.

Example 7

The 6 most popular letters in the English language are E, T, A, N, R, O and as we see their frequencies are:

- $E - 12\,\%$
- $T - 9\,\%$
- $A - 8\,\%$
- $N - 7\,\%$
- $R - 7\,\%$
- $O - 7\,\%$
- $Another - 50\,\%$

Take some part from Verzani and calculate the number of appearances of E, T, A, N, R, O and the number of other characters. You can use this site. By using $\mathscr{X}^2$ test check if this text is an English text.

Let us consider page 8, the text in Section 3 before the title “Categorical data”.

According to the above letter count calculator the total number of characters is $n = 1671$, where:

- $E - 215$
- $T - 165$
- $A - 149$
- $N - 103$
- $R - 100$
- $O - 113$

- $Another - 826$

The hypothesis are the same

$H_0 : p_E = 0.12, p_T = 0.09, p_A = 0.08, p_N = 0.07, p_R = 0.07, p_O = 0.07,$
$p_{Another} = 0.5$

According to H_0 the text is written in English. The differences between the theoretical and empirical frequencies are not statistically significant.

H_A : At least one category doesn't have this specified probability.

```
> f <- c(215, 165, 149, 103, 100, 113, 826);
> probs = c(12, 9, 8, 7, 7, 7, 50) / 100
> chisq.test(f, p = probs)

    Chi-squared test for given probabilities

data:  f
X-squared = 8.5939, df = 6, p-value = 0.1977
```

The $p-value = 0.1977 > 0.05 = \alpha$. We have no evidence to reject H_0. The text is written in English. The differences between the theoretical and empirical frequencies are not statistically significant.

Example 8

There are sewing machines in an industrial unit. Everyday at 6 o'clock the machines are checked and the number of the machines that have to be repaired is determined. The manufacturer has observed 450 days. The results from the observation are given in the following table.

The number of damaged machines	The number of days
0	149
1	141
2	74
3	37
4	32
More	15

By using the $\mathscr{X}^2$ test check if the observed random variable is **Poisson distributed**.

The Poisson distribution have one **parameter**. It is its **mean**. Therefore, we are going to estimate the parameter λ via the average of the damaged sewing machines within a day.

In order to compute the mean we replace “more” with 5.

```
> n <- 450
> x <- c(0:5)
> f <- c(149, 141, 74, 37, 32, 15)
> lambda <- sum(x * f) / n
> lambda
[1] 1.34
```

On average there are 1.34 damaged machines daily in this industrial unit.

H_0 : The number of the damaged machines within a day is Poisson distributed with parameter $\lambda = 1.34$

We take the theoretical probabilities $p_k = \dfrac{\lambda^k}{k!}e^{-\lambda}$,

$k = 0, 1, 2, 3, 4$ from the probability mass function of a Poisson distributed random variable with parameter $\lambda = 1.34$. p_5 is the probability Poisson distributed random variable with parameter $\lambda = 1.34$ to be bigger than or equal to 5. When we multiply these probabilities by 450 we obtain the **expected number of the damaged machines within a day** in this industrial unit.

According to H_0 the differences between these numbers and the empirical frequencies are not statistically significant.

H_A At least one category doesn't have this specified probability.

In this case we have $k = 6$ groups. In order to formulate H_0 we have estimated one-parameter λ from the sample. Therefore, $r = 1$. The degrees of freedom of the corresponding $\mathcal{X}^2$ distribution are $k - r - 1 = 6 - 1 - 1 = 4$. This is not taken into account in the function `chisq.test` in R.

```
> probs <- c(dpois(0:4, lambda), ppois(4, lambda,
lower.tail = FALSE))
> chisq.test(f, p = probs)

    Chi-squared test for given probabilities

data:  f
X-squared = 55.608, df = 5, p-value = 9.786e-11
```

Therefore, we will perform the test step by step.

Let us first determine the critical area

$$W_\alpha = \left\{ \sum_{i=1}^{3} \frac{(\nu_i - np_i)^2}{np_i} \geq x_{1-\alpha, \mathcal{X}^2(6-1-1)} \right\}$$

Now we chose $\alpha = 0.05$ and compute $x_{1-\alpha, \mathcal{X}^2(4)} = 9.487729$.

```
> qchisq(0.95, 4)
[1] 9.487729
```

Therefore, the critical area is

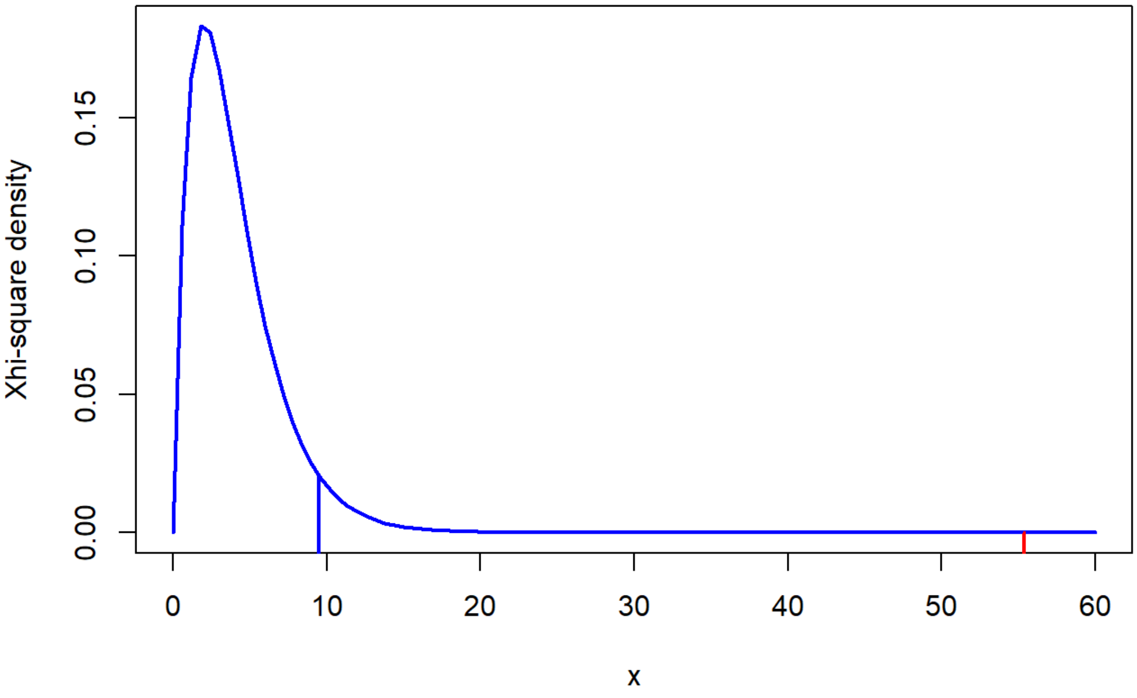
$$W_\alpha = \left\{ \sum_{i=1}^{3} \frac{(\nu_i - np_i)^2}{np_i} \geq 9.487729 \right\}$$

Now we can check if the sample is in the critical area W_α for H_0. Therefore, we replace the random variables ν_i, $i = 1, 2, \ldots, 6$ correspondingly with $f_1, f_2, \ldots, f_6$ observed in the sample

```
> e <- n * probs
> xemp <- sum((f - e)^2 / e)
> xemp
[1] 55.3697
```

Now, we compare $x_{emp} = 55.3697$ with $x_{1-\alpha,\mathcal{X}^2(4)} = 9.487729$

```
> df <- length(f) - 1 - 1
> alpha <- 0.05
> curve(dchisq(x, df), from = 0, to = 60, col = "blue",
lwd = 2, ylab = "Xhi-square density")
> x <- qchisq(1 - alpha, df)
> segments(x, -1, x, dchisq(x, df), col = "blue", lwd =
2)
> segments(xemp, -1, xemp, dchisq(xemp, df), col = "red",
lwd = 2)
```



and conclude that the sample is in the critical area for H_0. Therefore, we reject H_0.

The $p-value$ is the probability of the event that a $\mathscr{X}^2$ random variable having 4 df is greater than

$$\mathbb{P}\left(\sum_{i=1}^{6}\frac{(\nu_i - np_i)^2}{np_i} > x_{emp}\right) = \mathbb{P}(\eta > 54.84679) \approx 3.498212e-11$$

, $\eta \in \mathscr{X}^2(4)$.

```
> pchisq(xemp, df, lower.tail = FALSE)
[1] 2.718164e-11
```

The $p-value = 2.718164e-11 < 0.05 = \alpha$, so we reject H_0. The distribution of the damaged sewing machines is not Poisson.

Chi-square test for independence

Now let us study how to test independence **between the observed random variables** . In case when both random variables are numeric there are more appropriate tests. Chi-square tests for independence are usually used when at least one of the observed variables is categorical.

These test work with large sample tests as far as they use asymptotic distribution in order to determine the critical area for H_0.

Let us assume that we have independent observations on two random variables X and Y. The results are summarized in the following table

X\Y	Category Y_1	Category Y_2	...	Category Y_c	Total:
Category X_1	f_{11}	f_{12}	...	f_{1c}	f_1^X
Category X_2	f_{21}	f_{22}	...	f_{2c}	f_2^X
...	...	...	...	...	...
Category X_r	f_{r1}	f_{r2}	...	f_{rc}	f_r^X
Total:	f_1^Y	f_2^Y	...	f_c^Y	n

where f_{ij} is the number of observations having i-th category of X and j-th category of Y. f_1^X is the number of observations in the first category of X, and so on to f_r^X. f_1^Y is the number of observations in the first category of Y, and so on to f_c^Y. n is the sample size. Then,

$$\sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij} = n, \ \sum_{j=1}^{c} f_{ij} = f_i^X, \ \sum_{i=1}^{r} f_{ij} = f_j^Y.$$

If the observed X and Y are independent, then for all $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$,

$$\mathbb{P}(X = X_i, Y = Y_i) = \mathbb{P}(X = X_i)\mathbb{P}(Y = Y_i).$$

The estimator of $\mathbb{P}(X = X_i)$ is $\dfrac{f_i^X}{n}$.

The estimator of $\mathbb{P}(Y = Y_i)$ is $\dfrac{f_j^Y}{n}$.

Therefore, if X and Y are independent the expected values e_{ij} would be such that

$$p_{ij} = \frac{f_i^X f_j^Y}{n^2} = \frac{e_{ij}}{n}$$

and if we multiply by n we obtain

$$e_{ij} = \frac{f_i^X f_j^Y}{n} = np_{ij}$$

The sums in rows in the table with the expected values coincide with the corresponding one in the initial table

$$\sum_{i=1}^{r} \sum_{j=1}^{c} e_{ij} = n, \ \sum_{j=1}^{c} e_{ij} = f_i^X, \ \sum_{i=1}^{r} e_{ij} = f_j^Y.$$

X\Y	**Category** Y_1	**Category** Y_2	**...**	**Category** Y_c	**Total:**
Category X_1	e_{11}	e_{12}	...	e_{1c}	f_1^X
Category X_2	e_{21}	e_{22}	...	e_{2c}	f_2^X
...	...	...	...	...	...
Category X_r	e_{r1}	e_{r2}	...	e_{rc}	f_r^X
Total:	f_1^Y	f_2^Y	...	f_c^Y	n

H_0 : X and Y are independent. More precisely for all $i = 1. 2, \dots, r$ and $j = 1, 2, \dots, c$,

$$\mathbb{P}(X = X_i, Y = Y_i) = \mathbb{P}(X = X_i)\mathbb{P}(Y = Y_i)$$

The alternative is

H_A : There exist $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$ such that

$$\mathbb{P}(X = X_i, Y = Y_i) \neq \mathbb{P}(X = X_i)\mathbb{P}(Y = Y_i)$$

Which is the same as X and Y are dependent.

We chose level of significance α.

In this case the critical area is

$$W_\alpha = \left\{ \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - np_{ij})^2}{np_{ij}} \geq x_{1-\alpha,\, \mathcal{X}^2((r-1)(c-1))} \right\}$$

Example 9

Explorers would like to test the hypothesis if the car belts influence the injury level. They have made a sample of $n = 86769$ independent observations and have obtained the following results

Car crush \ Injury level	None	Minimal	Minor	Major	Toral
yes belt	12813	647	359	42	
no belt	65963	4000	2642	303	
Total:					

Are the two variables independent? Does the seat belt make a statistically significant difference in injury levels?

One way to solve this is:

```
> r <- 2
> c <- 4
> df <- (r - 1) * (c - 1)
> None <- c(12813, 65963)
> Minimal <- c(647, 4000)
> Minor <- c(359, 2642)
> Major <- c(42, 303)
> car <- data.frame(None, Minimal, Minor, Major)
> row.names(car) <- c("yes belt", "no belt")
> car
          None Minimal Minor Major
yes belt 12813     647   359    42
no belt  65963    4000  2642   303
> AllNone <- sum(car[, 1])
> AllMinimal <- sum(car[, 2])
> AllMinor <- sum(car[, 3])
> AllMajor <- sum(car[, 4])
> AllYes <- sum(car[1, ])
> AllNo <- sum(car[2, ])
> All <- sum(car)
> f <- c(None, Minimal, Minor, Major)
> e11 <- (AllYes * AllNone) / All
> e12 <- (AllYes * AllMinimal) / All
> e13 <- (AllYes * AllMinor) / All
> e14 <- (AllYes * AllMajor) / All
> e21 <- (AllNo * AllNone) / All
> e22 <- (AllNo * AllMinimal) / All
> e23 <- (AllNo * AllMinor) / All
> e24 <- (AllNo * AllMajor) / All
> e <- c(e11, e21, e12, e22, e13, e23, e14, e24)
> Xemp <- sum((f - e)^2 / e)
> Xemp
[1] 59.22397
```

$x_{emp} = 59.22397$ is the probability that a chi-square statistic having $(2-1)(4-1) = 3$ degrees of freedom is more extreme than 59.22397

$$p-value = \mathbb{P}(\eta > 59.22397), \eta \in \mathscr{X}^2(3)$$

```
> pchisq(Xemp, df, lower.tail = FALSE)
[1] 8.610376e-13
```

The $p-value = 0.0000000000008610376 < 0.05 = \alpha$, so we reject H_0. There is a relationship between the wearing belts and injury levels.

Another way to solve this is by using the function `chisq.test`

```
> yesbelt <- c(12813, 647, 359, 42)
> nobelt <- c(65963, 4000, 2642, 303)
> chisq.test(data.frame(yesbelt, nobelt))

    Pearson's Chi-squared test

data:  data.frame(yesbelt, nobelt)
X-squared = 59.224, df = 3, p-value = 8.61e-13
```

The $p-value = 0.000000000000861 < 0.05 = \alpha$, so we reject H_0. There is a relationship between the wearing belts and injury levels.

Example 10

A government would like to test the hypothesis if gender influences the voting preferences. They have made a sample of $n = 1000$ independent observations and have obtained the following results

Gender \ Voting Preferences	Republican	Democrat	Independent	Total:
Male	200	150	50	400
Female	200	300	50	600
Total:	450	450	100	1000

Do the mens' voting preferences differ significantly from the womens' preferences?

H_0 : Gender and voting preferences are independent.

H_A : Gender and voting preferences are not independent.

One way to solve this is:

```
> r <- 2
> c <- 3
> df <- (r - 1) * (c - 1)
> Republican <- c(200, 250)
> Democrat <- c(150, 300)
> Independent <- c(50, 50)
> voting <- data.frame(Republican, Democrat, Independent)
> row.names(voting) <- c("Male", "Female")
> voting
       Republican Democrat Independent
Male          200      150          50
Female        250      300          50
> AllRepublican <- sum(voting[, 1])
> AllDemocrat <- sum(voting[, 2])
> AllIndependent <- sum(voting[, 3])
> AllMale <- sum(voting[1, ])
> AllFemale <- sum(voting[2 ,])
> All <-sum(voting)
> f <- c(Republican, Democrat, Independent)
> e11 <- (AllMale * AllRepublican) / All
> e12 <- (AllMale * AllDemocrat) / All
> e13 <- (AllMale * AllIndependent) / All
> e21 <- (AllFemale * AllRepublican) / All
> e22 <- (AllFemale * AllDemocrat) / All
> e23 <- (AllFemale * AllIndependent) / All
> e <- c(e11, e21, e12, e22, e13, e23)
> Xemp <- sum((f - e)^2 / e)
> Xemp
[1] 16.2037
```

$x_{emp} = 16.2037$ is the probability that a chi-square statistic having $(2-1)(3-1) = 2$ degrees of freedom is more extreme than 16.2037

$$\mathbb{P}(\eta > x_{emp}) = \mathbb{P}(\eta > 16.2037) = 0.0003029781$$

```
> pchisq(Xemp, df, lower.tail = FALSE)
[1] 0.0003029775
```

The $p-value = 0.0003 < 0.05 = \alpha$, so we reject H_0. There is a relationship between gender and voting preference.

Another way to solve this is by using function `chisq.test`

```
> Male <- c(200, 150, 50)
> Female <- c(250, 300, 50)
> chisq.test(data.frame(Male, Female))
```

```
	Pearson's Chi-squared test

data:  data.frame(Male, Female)
X-squared = 16.204, df = 2, p-value = 0.000303
```

The $p-value = 0.000303 < 0.05 = \alpha$, so we reject H_0.

Chi-square test for homogeneity

We frequently need to test if the **distributions of two populations coincides in practice**. Again the method works for any kind or random variables. However, as far as for numerical random variables there are better tests, usually they are applied for categorical random variables.

Suppose we have m observations on random variables X and s observations on a random variable Y, any of them with k categories. Let us denote the relative frequencies in the categories correspondingly with

$$\hat{p}_{X_1}, \hat{p}_{X_2}, \ \ldots, \hat{p}_{X_k} \text{ and } \hat{p}_{Y_1}, \hat{p}_{Y_2}, \ \ldots, \hat{p}_{Y_k}$$

Assume that the samples are large enough, and in any of these groups there are at least 5 observations.

$H_0 : X$ and Y are homogeneous, i.e. $X \overset{d}{=} Y$. The differences in the samples are not statistically significant.

The alternative is

$H_A : X \overset{d}{\neq} Y$. The differences in the samples are statistically significant.

We chose α.

$$W_\alpha = \left\{ \sum_{i=1}^{k} \frac{(\hat{p}_{X_i} - \hat{p}_{Y_i})^2}{\frac{m\hat{p}_{X_i} + s\hat{p}_{Y_i}}{ms}} \geq x_{1-\alpha, \mathcal{X}^2(k-1)} \right\}$$

Example 11

Let us suppose we roll two dice.

- Fair die 200 times
- Biased die 100 times

Simulate the results and apply the chi-square test for homogeneity.

```
> set.seed(1)
> die.fair <- sample(1:6, 200, prob = c(1, 1, 1, 1, 1, 
1)/6, replace = TRUE)
> set.seed(1)
> die.bias <- sample(1:6, 100, prob = c(0.5, 0.5, 1, 1, 
1, 2)/6, replace = TRUE)
> fair.freq <- table(die.fair)
> bias.freq <- table(die.bias)
> freq <- rbind(fair.freq, bias.freq)
> freq
           1  2  3  4  5  6
fair.freq 32 22 34 42 36 34
bias.freq  4 10 25 13 21 27
> chisq.test(freq)

    Pearson's Chi-squared test

data:  freq
X-squared = 16.154, df = 5, p-value = 0.006419
```

The $p-value = 0.006419 < 0.05 = \alpha$, so we reject H_0. Both sets of data come from different distributions.