Goodness of fit $\mathcal{X}^2$ тест. Това са тестове за добро състояние и проверяват дали данните идват от някаква конкретна популация. Тези тестове се основават на асимптотични теореми, следователно те работят по-добре с големи извадки.

Пример. 1 Джон е хвърлил зарче 150 пъти и е открил, че има следното разпределение:

Лице:	1	2	3	4	5	6
Брой хвърляния:	22	21	22	27	22	36

Помогнете му да провери дали зарчето е честно (правилно).

Реш:

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

f=c(22, 21, 22, 27, 22, 36)
e=sum(f) * 1/6
xemp=sum((f-e)^2/e); xemp
# [1] 6.72
df <- length(f) - 1
alpha <- 0.05
pchisq(xemp, df, lower.tail = F)
```

Кратко решение:

```
# H0: the die is fair
# HA: the die is not fair

freq <- c(22, 21, 22, 27, 22, 36)
prob <- rep(1/6, 6)
res=chisq.test(freq, p = prob)
if(res$p.value>0.05) print("no evidence to reject H0") else print("reject H0 in favour of HA")
# [1] "no evidence to reject H0"
```

Пример. 2 Постановката е същата като от предходната задача, но този път нека допуснем, че имаме оригиналните данни. По-точно, нека симулираме хвърляне на зарче 150 пъти и да запишем наблюденията. След това да направим същия тест.

Реш:

```
set.seed(1)
X=sample(x=1:6, size=150, replace=T)
freq=table(X); freq

prob <- rep(1/6, 6)
res=chisq.test(freq, p = prob)
if(res$p.value>0.05) print("no evidence to reject H0") else print("reject H0 in favour of HA")
```

Пример. 3 Когато извадката не е достатъчно голяма.

```
X=sample(1:6, 150, replace = T, prob = c(0.75, 0.75, 1, 1, 1.25, 1.25)/6)
freq <- table(X); freq
prob <- rep(1/6, 6)
chisq.test(freq, p = prob)
# p-value = 0.1653
```

Пример. 4 Когато извадката нарастне значително.

```
X=sample(1:6, 1500, replace = T, prob = c(0.75, 0.75, 1, 1, 1.25, 1.25)/6)
freq <- table(X); freq
prob <- rep(1/6, 6)
chisq.test(freq, p = prob)
# p-value = 6.358e-16
```

Пример. 5 Компания принтира бейзболни картички. Тя твърди, че
30% са rookies; 60% са ветерани но не All-stars; 10% са ветерани All-stars.
В случайна извадка от 100 карти се окачва, че:
50% са rookies; 45% са ветерани но не All-stars; 5% са ветерани All-stars.
Проверете дали разпределението на случайната извадка се различава статистически значимо от разпределението заявено от компанията? Използвайте ниво на значимост 0,05.

Реш:

Формулираме нулевата хипотеза по следния начин:
H0: Наблюденията съответстват на:

	rookies	vet. not All-Stars	vet. All-Stars	
Теорирични вер.:	0.3	0.6	0.1	общо: 1
Очаквани набл.:	30	60	10	общо: 100

HA: Съществува статистически значима разлика между теоритичното и тестванаото разпр.

```
f=c(50,45,5)
prob=c(30,60,10)/100
res=chisq.test(f, p = prob)
if(res$p.value>0.05) print("no evidence to reject H0") else print("reject H0 in favour of HA")
# [1] "reject H0 in favour of HA"
```

Имаме основание да подозираме, че компанията послъгва.

Пример. 6 6-те най-популярни букви в англииската азбука са E, T, A, N, R, O и техните честоти са:
E - 12%; T - 9%; A - 8%; N - 7%; R - 7%; O - 7%; Another - 50%
Анализираме даден текст и получаваме, че срещанията на тези букви са следните:
E - 100; T - 110; A - 90; N - 80; R - 55; O - 14; Another - 600
На англиийски език ли е тествания текст?

Реш:
Формулираме нулевата хипотеза по следния начин:
H0: Наблюденията съответстват на вероятностите probs=c(12, 9, 8, 7, 7, 7, 50)/100
HA: Поне една категория не съответства на нейната теоритична вероятност от H0

```
f=c(100, 110, 90 ,80 ,55 ,14, 600)
prob=c(12, 9, 8, 7, 7, 7, 50)/100
res=chisq.test(f, p = prob)
if(res$p.value>0.05) print("no evidence to reject H0") else print("reject H0 in favour of HA")
# [1] "reject H0 in favour of HA"
```

Пример. 7 Постановката е същата като предходната задача, но този път разглеждаме текст от учебника на Verzani "Simple R", стр. 8, текста от секция 3 преди заглавието "Categorical data".

```
f=c(215, 165, 149, 103 ,100, 113, 826)
prob=c(12, 9, 8, 7, 7, 7, 50)/100
res=chisq.test(f, p = prob)
if(res$p.value>0.05) print("no evidence to reject H0") else print("reject H0 in favour of HA")
[1] "no evidence to reject H0"
```

Пример. 8 В дадена индустриална зона има шевни машини. Всеки ден в 18:00 часа машините се проверяват и се определя броя на машините, които трябва да бъдат ремонтирани. Производителят е наблюдавал 450 дни. Резултатите от наблюдението са дадени в следващата таблица:

Брой на машини за ремонт	Брой дни
0	149
1	141
2	74
3	37
4	32
повече	15

Използвайте $\mathcal{X}^2$ тест, за да проверите дали наблюдаваната случайна променлива е разпределена **поасоново**, както се очаква.

Разпределението на Поасон има един параметър. Това е неговата средна стойност. Следователно ще изчислим параметъра λ чрез средната стойност на повредените шевни машини в рамките на един ден.

За да изчислим средната стойност, заместваме повече с „5“.

Реш:

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

Средно има 1.34 повредени мапини дневно в тази индустриална зона.

H0: Броя на повредените машини в рамките на един ден е поасоново разпределен с параметър lambda=1.34
HA: Машините не се експлоатират внимателно/правилно/според указанията и т.н. (чупят се повече отколкото трябва или производителя не е отговорил на изискванията за здравина)

Взимаме теоритичните вероятности за поасоновото разпр. с този параметър:

```
probs <- c(dpois(0:4, lambda), ppois(4, lambda, lower.tail = FALSE))
res=chisq.test(f, p = probs)
if(res$p.value>0.05) print("no evidence to reject H0") else print("reject H0 in favour of HA")
# [1] "reject H0 in favour of HA"
```

$\mathcal{X}^2$ тест за независимост.

Тестовете за $\mathcal{X}^2$ за независимост обикновено се използват, когато поне една от наблюдаваните променливи е категориина. Тези тестове са по смислени когато се изпозлват за много на брой наблюдения, тъй като се основават на асимптотични заключения за разпределенията.

Пример. 9 Изследователи искат да проверят хипотезата дали коланите влияят на нивото на нараняване при сблъсък. Направена е извадка от n=86789 независими наблюдения и са получени следните резултати:

Сблъсък\Ниво на нараняване	Никакво	Минимално	Малко	Голямо
с колан	12813	647	359	42
без колан	65963	4000	2642	303

Независими ли са двете променливи? Прави ли предпазния колан статистически значима разлика в нивата на нараняване?

Реш:

```
yesbelt=c(12813, 647, 359, 42)
nobelt=c(65963, 4000, 2642, 303)
res=chisq.test(data.frame(yesbelt, nobelt))
if(res$p.value>0.05) print("no evidence to reject H0") else print("reject H0 in favour of HA")
# [1] "reject H0 in favour of HA"
```

Пример. 10 Правителство иска да провери хипотезата дали полът влияе върху предпочитанията за гласуване. Направени са n=1000 независими наблюдения и са получени следните резултати:

Пол\Предпочитания	Републиканци	Демократи	Независими	Общо:
Мъж	200	150	50	400
Жена	250	300	50	600
Общо:	450	450	100	1000

Разликите в гласуването на мъжете значително ли се различава от предпочитанията на жените?

Реш:

H0: Предпочитанията за гласуване и пол са независими
HA: Предпочитанията за гласуване и пол не са независими

```
male=c(200, 150, 50)
female=c(250, 300, 50)
res=chisq.test(data.frame(male, female))
if(res$p.value>0.05) print("no evidence to reject H0") else print("reject H0 in favour of HA")
# [1] "reject H0 in favour of HA"
```

$\mathcal{X}^2$ тест за хомогенност.

Често трябва да проверяваме дали разпределението на две популации съвпада на практика.

Пример. 11 Хвърляме два зара:
- правилен зар 200 пъти
- неправилен зар 100 пъти

```
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

res=chisq.test(freq)
if(res$p.value>0.05) print("no evidence to reject H0") else print("reject H0 in favour of HA")
# [1] "reject H0 in favour of HA"
```

Зад. 12.1

Problem 12.1

In an effort to increase student retention, many colleges have tried block programs. Suppose 100 students are broken into two groups of 50 at random. One half are in a block program, the other half not. The number of years in attendance is then measured. We wish to test if the block program makes a difference in retention. The data is:

Program	1 yr	2 yr	3 yr	4 yr	5+ yr
Non-Block	18	15	5	8	4
Block	10	5	7	18	10

Do a test of hypothesis to decide if there is a difference between the two types of programs in terms of retention.

```
> nonBlock <- c(18, 15, 5, 8 ,4)
> block <- c(10, 5, 7, 18, 10)
> chisq.test(rbind(nonBlock, block))

    Pearson's Chi-squared test

data:  rbind(nonBlock, block)
X-squared = 14.037, df = 4, p-value = 0.007179
```

The $p - value = 0.007179 < 0.05 = \alpha$, so we reject H_0. The block programs makes a difference in a retention.

Зад. 12.2

Problem 12.2

A survey of drivers was taken to see if they had been in an accident during the previous year, and if so was it a minor or major accident. The results are tabulated by age group:

Age \ Accident Type	None	Minor	Major
under 18	67	10	5
18 - 25	42	6	5
26 - 40	75	8	4
41 - 65	56	4	6
over 65	57	15	1

Do a chi-squared hypothesis test of homogeneity to see if there is difference in distributions based on age.

```
> under18 <- c(67, 10, 5)
> between18and25 <- c(42, 6, 5)
> between26and40 <- c(75, 8, 4)
> between40and65 <- c(56, 4, 6)
> over65 <- c(57, 15, 1)
> chisq.test(rbind(under18, between18and25, between26and40, between40and65, over65))
Warning in chisq.test(rbind(under18, between18and25, between26and40,
between40and65, : Chi-squared approximation may be incorrect

    Pearson's Chi-squared test

data:  rbind(under18, between18and25, between26and40, between40and65,     over65)
X-squared = 12.586, df = 8, p-value = 0.1269
```

The $p - value = 0.1269 > 0.05 = \alpha$, so we have no evidence to reject H_0. The age does not influence the accident type.

Зад. 12.3

Problem 12.3

A fish survey is done to see if the proportion of fish types is consistent with previous years. Suppose, the 3 types of fish recorded: parrotfish, grouper, tang are historically in a $5:3:4$ proportion and in a survey the following counts are found

Parrotfish	Grouper	Tang
53	22	49

Do a test of hypothesis to see if this survey of fish has the same proportions as historically.

We perform goodness of fit test

```
> freq <- c(53, 22, 49)
> prob <- c(5, 3, 4) / 12
> chisq.test(freq, p = prob)

    Chi-squared test for given probabilities

data:  freq
X-squared = 4.0694, df = 2, p-value = 0.1307
```

The $p - value = 0.1307 > 0.05 = \alpha$, so we have no evidence to reject H_0. This survey of fish have the same proportion as historically observed.

Зад. 12.4

Problem 12.4

The R data set `UCBAdmissions` contains data on admission to UC Berkeley by gender. We wish to investigate if the distribution of males admitted is similar to that of females. To do so, we need to first do some spade work as the data set is presented in a complex contingency table. The `ftable` (flatten table) command is needed. To use it try

```
> 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 = ftable(UCBAdmissions)
> x
                Dept   A   B   C   D   E   F
Admit    Gender
Admitted Male        512 353 120 138  53  22
         Female       89  17 202 131  94  24
Rejected Male        313 207 205 279 138 351
         Female       19   8 391 244 299 317
```

We want to compare rows 1 and 2. Treating x as a matrix, we can access these with x[1:2,]. Do a test for homogeneity between the two rows. What do you conclude? Repeat for the rejected group.

```
> chisq.test(x[1:2,])

    Pearson's Chi-squared test

data:  x[1:2, ]
X-squared = 463.09, df = 5, p-value < 2.2e-16
```

The $p - value < 2.2e - 16 < 0.05 = \alpha$, so we reject H_0. The difference in the admitted men and women is statistically significant.

```
> chisq.test(x[3:4,])

    Pearson's Chi-squared test

data:  x[3:4, ]
X-squared = 552.62, df = 5, p-value < 2.2e-16
```

The $p - value < 2.2e - 16 < 0.05 = \alpha$, so we reject H_0. The difference in the rejected men and women is statistically significant.