Rešitve - uporaba testov

Nataša Kejžar

Naloga 1

```
n = 2207
pA = 0.44
pB = 0.47
nA = round(n*pA)
nB = round(n*pB)
n0 = n - nA - nB
##LRT
# ocena parametra pod H0
phat0 = (nA+nB)/(2*n)
# oceni parametrov pod HA (oceni dveh parametrov)
phatA = pA
phatB = pB

LRT = -2*((nA+nB)*log(phat0) + n0*log(1-2*phat0)-
            nA*log(phatA)-nB*log(phatB)-n0*log(1-phatA-phatB))
pchisq(LRT,df=1)
```

```
## [1] 0.8591765
```

```
## hi-kvadrat
testna = (nA-phat0*n)^2/(phat0*n)+(nB-phat0*n)^2/(phat0*n)+
  (n0 - (1-2*phat0)*n)/((1-2*phat0)*n)
pchisq(testna,df=1)
```

```
## [1] 0.8592118
```

Naloga 2 - Geissler

a. Lahko uporabimo test χ^2, ali test za posplošeno razmerje verjetij. V primeru testa χ^2 ima testna statistika porazdelitev $\chi^2_{df=12}$, saj je število vrstic 13, v primeru LRT pa $\chi^2_{df=11}$, saj pod alternativno domnevo ocenjujemo 12 parametrov (13. se sešteje v 6115), pod H_0 pa 1 (p_0).

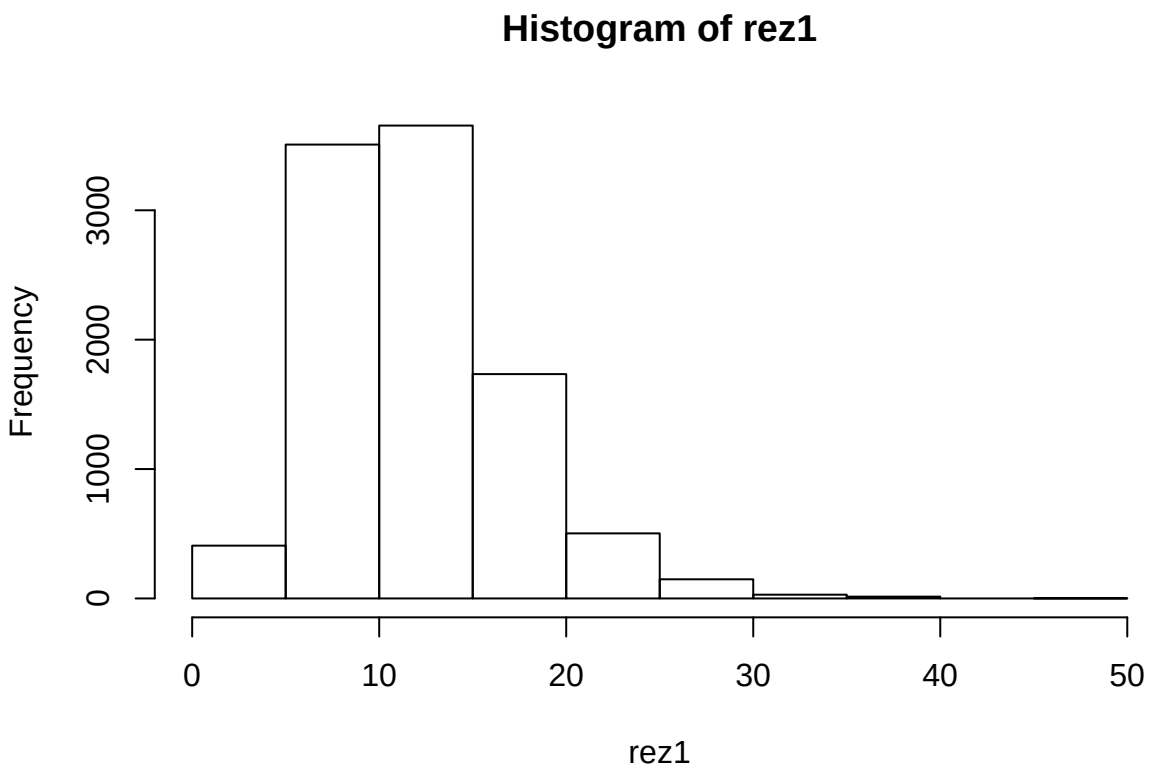
```
pod = read.csv("data/data_Geissler.csv")

n <- 6115
p0 <- 105/205
# analiza podatkov
Oi = pod$frekvenca
Ei = n*dbinom(0:12,12,p0)
T1 = sum((Oi-Ei)^2/Ei)
pchisq(T1,df=12,lower.tail=FALSE) # st.stolpcev*st.vrstic-1 =1*13-1 = 12
```

```
## [1] 1.058281e-95
```

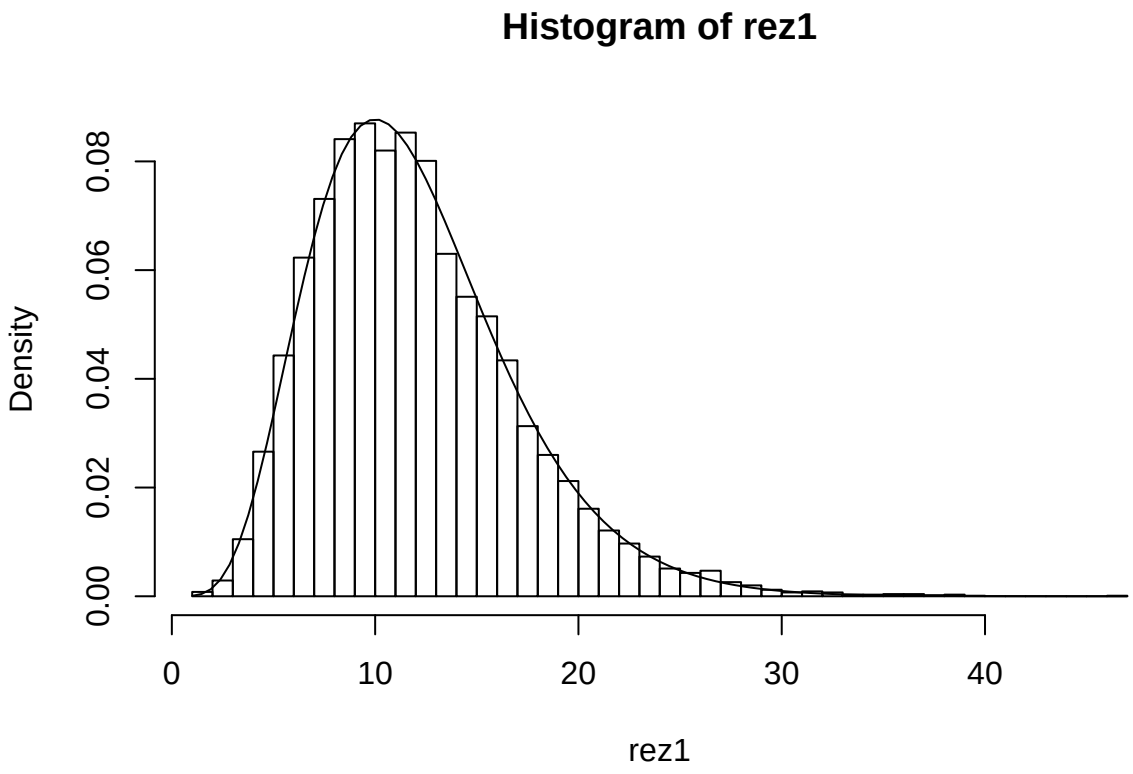
```r
#posploseno razmerje verjetij
T2 = 2*sum(Oi*log(Oi/Ei))
pchisq(T2,df=12,lower.tail=FALSE) # st.ocenjenih parametrov =12-0 = 12
```

```
## [1] 2.520612e-75
```

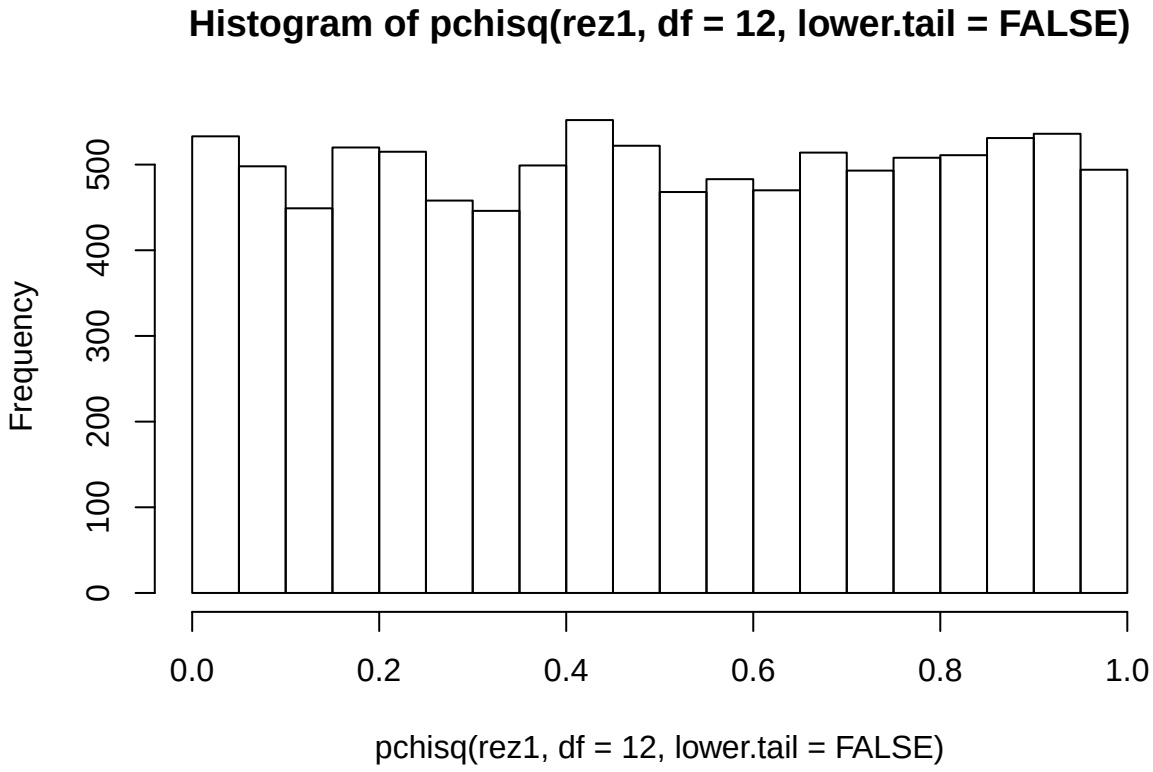
```r
meja = qchisq(0.95,df=12)

runs <- 10000
rez1 <- rez2  <- rep(NA,runs)
for(i in 1:runs){
  ##p = rnorm(n,p0,0.02)
  ##frekv <- rbinom(n,12,p) # st. fantkov
  frekv <- rbinom(n,12,p0) # st. fantkou
  Oi <- table(factor(frekv,levels=0:12))
  Ei <-  dbinom(0:12,12,p0)*n
  hi2 = sum((Oi-Ei)^2/Ei) #obicajni hi-kvadrat
  LR = 2*sum(Oi*log(Oi/Ei)) #formula iz razmerja verjetij
  rez1[i] <- hi2
  rez2[i] <- LR
}

hist(rez1)
```



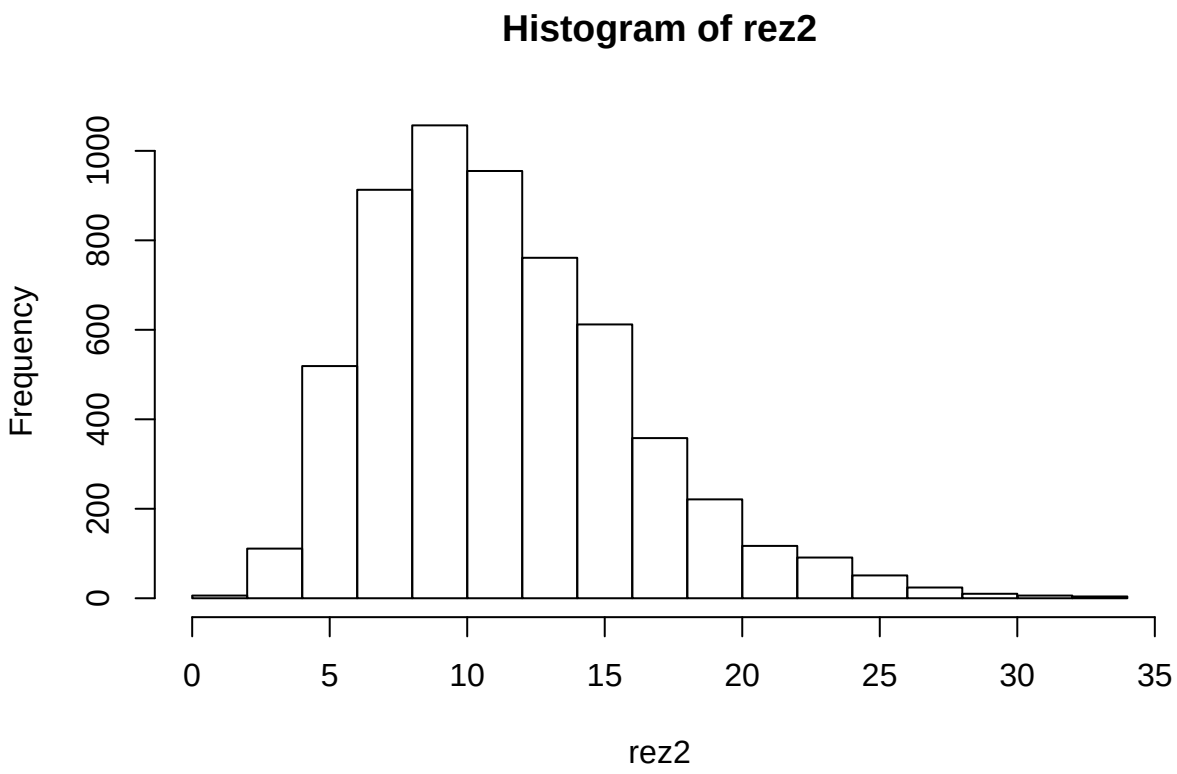

```r
# z vrisano porazdelitvijo pod H0
hist(rez1,breaks = 50,freq=FALSE)
curve(dchisq(x,df=12),add=TRUE)
```



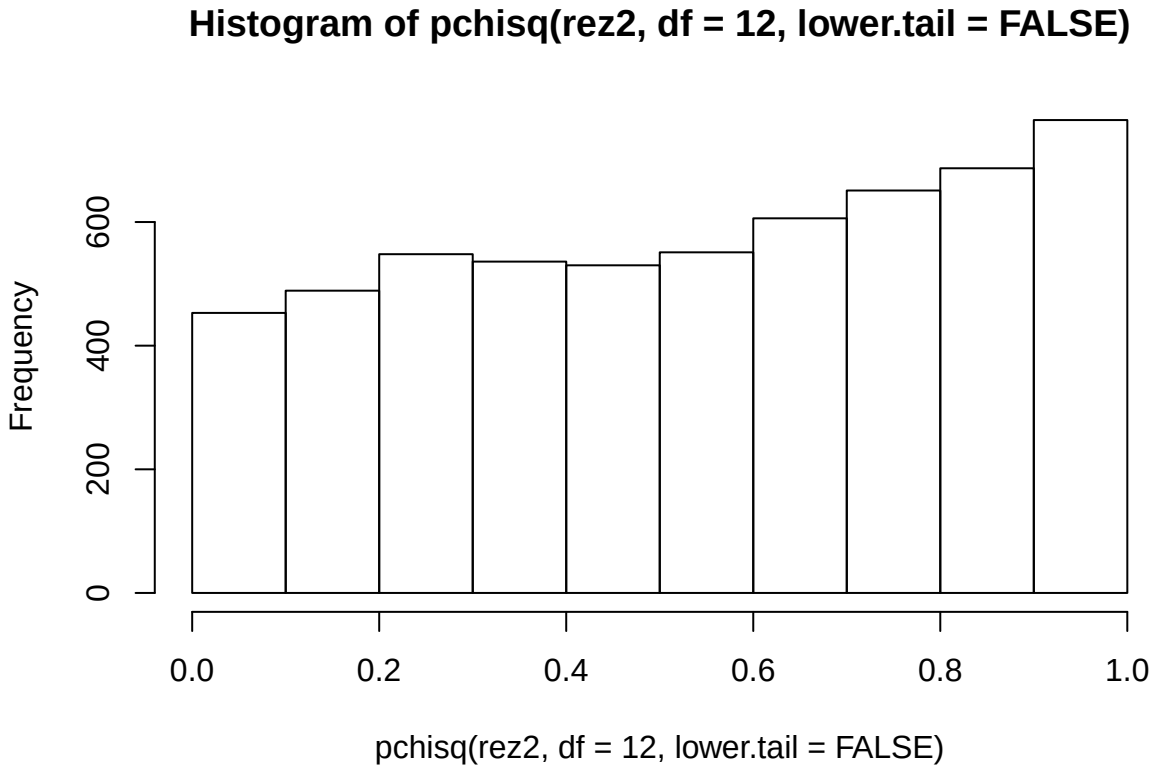

```
hist(pchisq(rez1,df=12,lower.tail=FALSE))
```




```
hist(rez2)
```




```
hist(pchisq(rez2,df=12,lower.tail=FALSE))
```

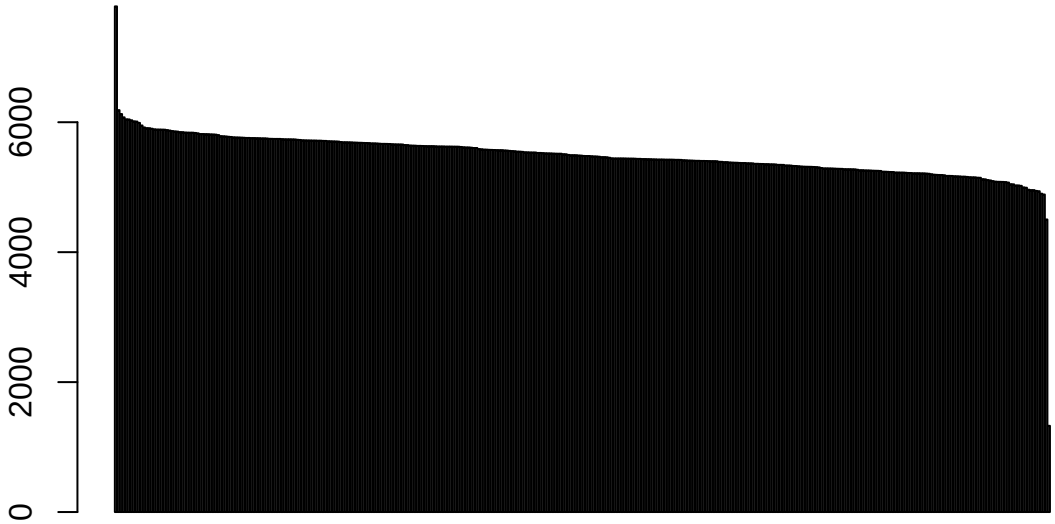


Naloga 3 - rojstni dnevi

```
rd = read.csv("data/data_RD.csv")
names(rd)
```

```
## [1] "Mesto"       "Rojstni.dan" "Stevilo"
```

```
barplot(rd[,3])
```




```
Oi = rd[,3]
Oi = Oi[1:365]
n = sum(Oi)
Ei = 1/365*n
testna = sum((Oi-Ei)^2/Ei)
pchisq(testna,df=364,lower.tail=FALSE)
```

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

```
testna2 = 2*sum(Oi*log(Oi/Ei))
# st.ocenjenih parametrov =364-0 =364
pchisq(testna2,df=364,lower.tail=FALSE)
```

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

Naloga 4 - potresi Poisson lambda=1

```
# testna statistika bo sum(X), X~Pois(lambda)
# vemo, da sum(X)~Pois(n lambda), ce X iid

lambda0 = 1
podatki = c(0,0,0,1,1,1,1,1,2,2,3,3,3,3,5)
T1 = sum(podatki)

# obmocje zavrnitve (enostransko, zavrnemo za velike vrednosti T)
meja = qpois(0.95,length(podatki)*lambda0)
ppois(meja,length(podatki)*lambda0,lower.tail = FALSE)
```

```
## [1] 0.03274424
```

```r
# zavrnemo, ce T >= meja

# nas primer
ppois(T1,length(podatki),lower.tail=FALSE)
```

```
## [1] 0.003311898
```

```r
# ce v resnici
lambda1 = 0.8
ppois(meja-1, length(podatki)*lambda1,lower.tail=FALSE)
```

```
## [1] 0.006065147
```

```r
## gre za napako 1.vrste
# zavrnemo pravilno nicelno domnevo (p = 0.006)

#ce
lambda2 = 1.4
ppois(meja, length(podatki)*lambda2,lower.tail=TRUE)
```

```
## [1] 0.6404561
```

```r
## gre za napako 2.vrste
# ne zavrnemo napacne nicelne domneve (p = 1-0.64)

# kako velik vzorec?
# vzorec je sestavljen iz let

moc = NULL
leta = 100:250
for(n in leta){
  meja = qpois(0.95,n*lambda0)
  moc = c(moc,ppois(meja-1, n*1.2,lower.tail=FALSE))
}

leta[min(which(moc > 0.9))] # 228 let
```

```
## [1] 228
```

Naloga 5 - potresi Poisson lambda=?

```r
### posploseni hi-kvadrat

lambdahat <- mean(podatki)
kat <- max(podatki)
podatki1 <- factor(podatki,levels=0:kat)
oi <- table(podatki1)
ei <- dpois(0:(kat-1),lambdahat)
ei <- c(ei, 1-sum(ei))*length(podatki)
T2 = sum((oi-ei)^2/ei)
1-pchisq(T2,kat-1) # st.nevezanih vrednosti - 1
```

```
## [1] 0.4204028
```

```r
#
ei <- dpois(0:(kat-1),1)
```

```r
ei <- c(ei, 1-sum(ei))*length(podatki)
1-pchisq(sum((oi-ei)^2/ei),kat) # st.kat-1 (st. nevezanih vrednosti)
```

```
## [1] 3.290808e-05
```

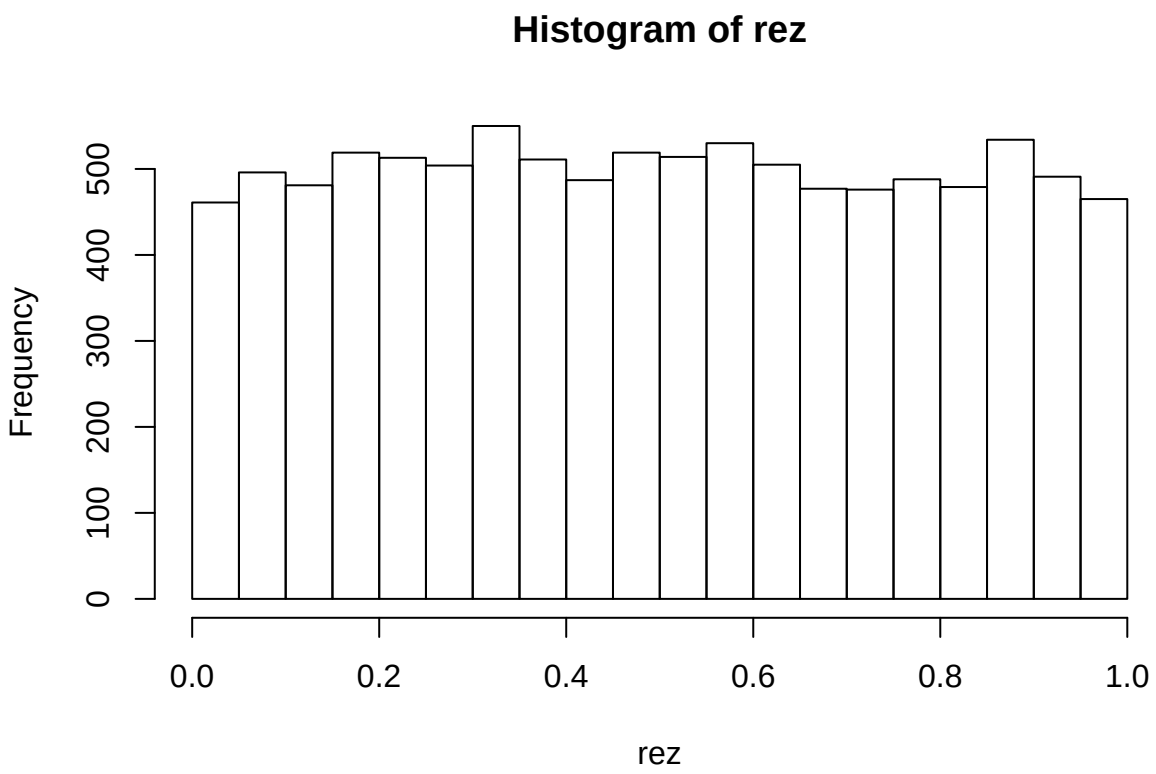
```r
#
# vpliv razlicnega stevila kategorij na statisticni test hi-kvadrat
# simulacije
runs <- 10000
rez <- rep(NA,runs)
size <- 120
lambda=1
for(it in 1:runs){
  vzorec <- (rpois(size,lambda))
  lambdahat <- mean(vzorec)
  nu <- max(vzorec) # določimo število kategorij za ta vzorec
  vzorec <- factor(vzorec,levels=0:nu)
  oi <- table(vzorec)
  ei <- dpois(0:(nu-1),lambdahat)
  ei <- c(ei,1-sum(ei))
  ei <- ei*size
  rez[it] <- 1-pchisq(sum((oi-ei)^2/ei),nu-1)
}
hist(rez)
```


```r
sum(rez<0.05)/runs
```

```
## [1] 0.1015
```

```r
runs <- 10000
rez <- rep(NA,runs)
```

```r
size <- 120
lambda=1
kat=4
for(it in 1:runs){
  vzorec <- (rpois(size,lambda))
  lambdahat <- mean(vzorec)
  vzorec[vzorec>kat] <- kat # št. kategorij fiksiramo
  vzorec <- factor(vzorec,levels=0:kat)
  oi <- table(vzorec)
  ei <- dpois(0:3,lambdahat)
  ei <- c(ei,1-sum(ei))
  ei <- ei*size
  rez[it] <- 1-pchisq(sum((oi-ei)^2/ei),kat-1)
}
hist(rez)
```




```r
sum(rez<0.05)/runs
```

```
## [1] 0.0461
```

Naloga 6 - potresi verjetnost

```r
### 7% verjetnost
# verjetnost za potres v mesecu je 7%

n = length(podatki)*12
p = 0.07
phat = sum(podatki)/n # st.potresov/st.mesecev
T3 = sum(podatki) # Binomska porazdelitev
# p-vrednosti iz eksaktne binomske porazdelitve
```

```r
1-pbinom(T3,prob=p,size=n)
```

```
## [1] 0.000151921
```

Naloga 7 - trgovina

Pri χ^2 testu so pričakovane frekvence v $4/7$ primerov < 5, zato tu predpostavke niso izpolnjene. To lahko popravimo z združevanjem kategorij. Združimo torej kategoriji z 0 in 1 dobrim dnevom in s 5 in 6 dobrimi dnevi. Potem dobimo malce drugačno testno statistiko, 2 stopnji prostosti manj.

Pri PRV lahko rečemo, da $n = 52$ ni dovolj velik vzorec.

Naloga 8 - ruleta

a. Testna statistika naj bo enaka
$$T = \sum_{i=0}^{36} \frac{(O_i - E_i)^2}{E_i}$$
kjer je O_i število pojavljanj številke i, E_i pa je za vse i enak $n/37$. Testna statistika je porazdeljena kot χ^2_{36}.

b. Testna statistika je asimptotska, saj je test χ^2 izpeljan le ko gre n proti neskončnosti.

c. Seveda. Verjetnost takega dogodka je enak 0.005.

d. Ne. Na podlagi vzorca ne moremo z gotovostjo trditi ničesar. Na našem cilindru smo dobili nekaj odstopanja od ničelne domneve, naš rezultat nam pove, da je verjetnost, da pride do opaženega (ali še večjega) odstopanja po naključju enaka $0,23$. Dejanskih razlik torej v našem primeru ne moremo ločiti od naključnih.

e. Naj bo spremenljivka X_i enaka 1, če cilinder natančneje preverijo in 0 sicer. Verjetnost $P(X_i = 1) = 0,01$. Zanima nas pričakovana vrednost vsote, spremenljivke so med seboj neodvisne (okvare cilindrov so neodvisne):
$$E\left(\sum_{k=1}^{100} X_k\right) = \sum_{k=1}^{100} E\left(X_k\right) = \sum_{k=1}^{100} P\left(X_k == 1\right) = 100 \cdot 0,01 = 1$$
V povprečju preverimo en cilinder na dan.

f. Tak cilinder bomo opazili večkrat, koliko večkrat je odvisno od tega, kako velika odstopanja povzroča okvara (moč testa).

Naloga 9 - žarnice

a. Ničelna hipoteza in prostor vrednosti za H_0 in H_A: % $H_0 : \lambda_0 = 1/1200$, $\Theta_0 = \{1/1200\}$ in $\Theta = \{\lambda; \lambda \in (0, \infty)\}$, $\Theta_A = \Theta \backslash \Theta_0$.

b. Uporabi lahko posplošeno razmerje verjetij. Testna statistika $(-2 \log \Lambda)$ za ta statistični test je porazdeljena v tem primeru po χ^2_1 porazdelitvi. Za testno statistiko (za alternativno domnevo) potrebujemo oceno za parameter λ. Tega ocenimo po metodi največjega verjetja.

$$
\begin{aligned}
L &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\
l &= \sum_{i=1}^{n} (\log \lambda - \lambda x_i) \\
&= n \log \lambda - \lambda \sum_{i=1}^{n} x_i \\
\frac{\partial}{\partial \lambda} l &= 0 = \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i \\
\hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} x_i}
\end{aligned}
$$

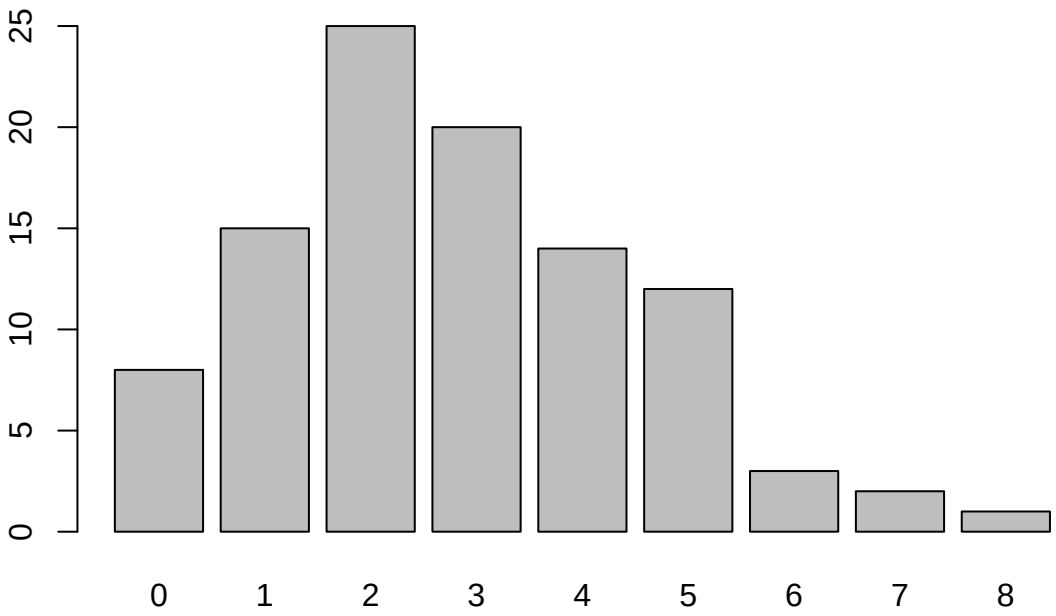
Zapišimo sedaj posplošeno razmerje verjetij:

$$
\begin{aligned}
\Lambda &= \frac{\prod_{i=1}^{n} \lambda_0 e^{-\lambda_0 x_i}}{\prod_{i=1}^{n} \hat{\lambda} e^{-\hat{\lambda} x_i}} \\
\log \Lambda &= n \log \lambda_0 - \lambda_0 \sum_{i=1}^{n} x_i - \left(n \log \hat{\lambda} - \hat{\lambda} \sum_{i=1}^{n} x_i \right) \\
\log \Lambda &= n \log \lambda_0 - \lambda_0 \sum_{i=1}^{n} x_i - n \log \left(\frac{n}{\sum_{i=1}^{n} x_i} \right) + \frac{n}{\sum_{i=1}^{n} x_i} \sum_{i=1}^{n} x_i \\
\log \Lambda &= n \log \lambda_0 - \lambda_0 \sum_{i=1}^{n} x_i - n \log n + n \log \sum_{i=1}^{n} x_i + n \\
-2 \log \Lambda &= -2n \left(\log \lambda_0 - \log n + \log \sum_{i=1}^{n} x_i + 1 \right) + 2\lambda_0 \sum_{i=1}^{n} x_i
\end{aligned}
$$

Naloga 10 - večkratno testiranje

```
set.seed(1)
n = 20
mu = 4
sd = 3

rez = data.frame(truepos=0,falsepos=0)
for(i in 1:100){
  # generiraj podatke - 60 spremenljivk
  data=data.frame(V1 = 1:n)
  for(j in 1:5)
    data[,j] = rnorm(n,mu,sd)
  for(j in 6:60)
    data[,j] = rnorm(n,0,sd)
    # preveri s t-testom
  testi = numeric(60)
  for(j in 1:60)
    testi[j] = t.test(data[,j])$p.value
  #testi = p.adjust(testi,method="bonferroni")
  rez[i,] = c(sum(testi[1:5] < 0.05),sum(testi[6:60] < 0.05))
```

```
}

barplot(table(rez$falsepos))
```




```
p.adjust(testi,method="bonferroni")
```

```
##  [1] 3.009070e-02 7.168640e-04 4.507115e-02 5.097977e-05 4.353178e-02
##  [6] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [11] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [16] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [21] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [26] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [31] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [36] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [41] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [46] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [51] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
## [56] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
```

```
#p.adjust(testi,method="holm")
```

Naloga 11 - psiholog

a. $H_0 : \mu_M = \mu_Z$

b. $H_0 : \mu_M \neq \mu_Z$, je sestavljena, dvostranska

c. $T = \bar{X}_M - \bar{X}_Z$, $X_i \sim N(\mu_0, \sigma)$, iid
$E(\bar{X}_M - \bar{X}_Z) = E(\bar{X}_M) - E(\bar{X}_Z) = \mu_0 - \mu_0 = 0$
$var(\bar{X}_M - \bar{X}_Z) = var(\bar{X}_M) + var(\bar{X}_Z) = \frac{\sigma^2}{n_M} + \frac{\sigma^2}{n_Z}$
v našem primeru je $n_M = n_Z = n$, torej
$var(\bar{X}_M - \bar{X}_Z) = 2\frac{\sigma^2}{n}$. Torej je naša testna statistika $T \sim N(0, \sqrt{2}\frac{\sigma}{\sqrt{n}})$

d. $(\infty, -0,020) \cup (0, 020, \infty)$

e.

```
pnorm(mean(x$M) -mean(x$F),0,sd=sqrt(2/nrow(x))*0.028)}= 0.0020
```

f. $[-0.049, -0.009]$
g. Razlika je statistično značilna. Razlika je med 0.05 in 0.009 sekund. Ker je zgornja meja intervala zaupanja precej blizu 0, obstaja dvom o strokovni pomembnosti. Verjetno bi bilo dobro nabrati večji vzorec.
h. Moški: $[0.224, 0.253]$, ženske: $[0.254, 0.282]$. Intervala se ne prekrivata, kar je pričakovano, saj smo dobili, da je obstaja populacijska razlika med povprečjema moških in žensk.

Naloga 12 - intervali zaupanja

a. Prva spremenljivka: $[-0.5, 0.5]$, druga spremenljivka: $[0.25, 1.25]$, se prekrivata
interval zaupanja za razliko: $[-1.46, -0.43]$. Kljub prekrivajočima samostojnima intervaloma zaupanja, statistični test trdi, da obstaja razlika v populacijskem povprečju med spremenljivkama. Do tega lahko pride, ker imamo majhen vzorec.
b. Prva spremenljivka: $[-0.58, 0.5]$, druga spremenljivka: $[0.17, 1.33]$, se prekrivata
interval zaupanja za razliko: $[-1.57, 0.07]$. V tem primeru do te anomalije ne pride. Pri velikih vzorcih je precej vseeno, ali računamo z s ali s σ.
c. To se zgodi, ker:

$$var(\bar{X}_1 - \bar{X}_2) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

torej je SE koren variance, kar pa ni enako vsoti posameznih SE:

$$\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \frac{\sigma}{\sqrt{n_1}} + \frac{\sigma}{\sqrt{n_2}}$$

Če gledamo vsako spremenljivko posebej, torej dobimo večjo variabilnost, zato je mogoče, da se posamezna intervala zaupanja prekrivata, skupni pa kaže, da je med njima statistično značilna razlika. Veljati mora, da je razlika med spremenljivkama relativno majhna glede na njuno posamično variabilnost. Če želimo preveriti, za kakšne vrednosti parametrov je to mogoče, je potrebno zapisati neenakosti, ki v tem primeru držijo. Npr. (naš primer) naj se spodnja meja druge spremenljivke prekriva z zgornjo mejo prve spremenljivke:

$$\bar{X}_1 + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} > \bar{X}_2 - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

torej $\bar{X}_1 - \bar{X}_2 > -2z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$. Velja pa tudi, da je zgornja meja skupnega intervala zaupanja pod 0:

$$\bar{X}_1 - \bar{X}_2 + z_{1-\alpha/2} \cdot \sqrt{2} \frac{\sigma}{\sqrt{n}} < 0$$

Ti dve neenačbi lahko zapišemo skupaj:

$$\sqrt{2} < (\bar{X}_1 - \bar{X}_2) \cdot c < 2; \quad c = \frac{\sqrt{n}}{z_{1-\alpha/2}\sigma}$$

Naloga 13 - teža

c. Tu vzamemo test-t za 2 odvisna vzorca. S tem zmanjšamo varianco in tako lahko naredimo bolj natančne zaključke. Ko uporabljamo test t za dva odvisna vzorca, predpostavljamo, da je v vzorcu $2n$ opazovanj (za obe skupini/spremenljivki po n. Varianca razlike med tema skupinama/spremenljivkama je

$$
\begin{aligned}
var(\bar{X}_1 - \bar{X}_2) =& var(\bar{X}_1) + var(\bar{X}_2) - 2cov(\bar{X}_1, \bar{X}_2) \\
=& \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - \frac{2}{n^2} cov(\sum X_{1i}, \sum X_{2j}) \\
=& \frac{1}{n} \left[\sigma_1^2 + \sigma_2^2 - 2cov(X_{1i}, X_{2i})\right] \\
\overset{\sigma_1 = \sigma_2}{=}& \frac{2\sigma^2}{n}(1 - \rho)
\end{aligned}
$$

Če odvisnosti ne predpostavljamo, korelacije ni, zato pridemo do variance $\frac{2\sigma^2}{n}$, ki je večja.

d. Manjša moč testa.