Rešitve - statistični testi

Nataša Kejžar

Naloga 1 - semafor

a. $H_0 : \lambda_0 = 3$, $H_A : \lambda_A = 5$, obe hipotezi sta enostavni
b. Razmerje verjetij je:

$$T = \prod_{i=1}^{n} \frac{f_0(x_i)}{f_A(x_i)}$$

c. Razmerje verjetij je:

$$T = \prod_{i=1}^{n} \frac{\lambda_0^{x_i} e^{-\lambda_0}}{x_i!} \cdot \frac{x_i!}{\lambda_A^{x_i} e^{-\lambda_A}}$$

$$T = \prod_{i=1}^{n} \left(\lambda_0^{x_i} \lambda_A^{-x_i} e^{-\lambda_0 + \lambda_A}\right)$$

$$\log T = \sum_{i=1}^{n} x_i \log \lambda_0 - \sum_{i=1}^{n} x_i \log \lambda_A + \sum_{i=1}^{n} (\lambda_A - \lambda_0)$$

Zadnji člen je pozitiven, konstanten, prvi in drugi člen sta vsota opazovanj, pomnožena z neko konstanto, tako da lahko zapišemo testno statistiko:

$$\log T = \sum_{i=1}^{n} x_i$$

d. $\sum X_i \sim Pois(\sum \lambda_i)$, če $X_i \sim Pois(\lambda_i)$, torej za našo testno statistiko velja

$$\log T = \sum_{i=1}^{n} x_i \sim Pois(n\lambda_0)$$

Meji zavrnitve sta v splošnem $P(\log T < t_{min}) = \alpha/2$ in $P(\log T > t_{max}) = \alpha/2$. Ker pa imamo enosmerni test (le preverjanje λ_A, ki so večje od λ_0), potem imamo le zgornjo mejo $P(\log T > t_{max}) = \alpha$.

```
qpois(0.975,n*lambda0)
```

območje zavrnitve je $[329, \infty)$. f. Testna statistika je: 326, vrednost p je

```
ppois(sum(x)-1,n*lambda0,lower.tail=FALSE)
```

g. Test ni statistično značilen, zato ne moremo zavreči hipoteze, da so prihodi avtomobilov na križišče porazdeljeni po Poissonovi porazdelitvi s hitrostjo 3.
h. Pokazati je torej potrebno, da ima testna statistika največjo moč za katerokoli enostavno H_A, kjer je $\lambda_A > \lambda_0$. Zapišimo moč testa (t_{max} je meja zavrnitve, odvisna je od λ_0 in n):

$$P_A(T > t_{max})$$

t_{max} torej ni odvisna od vrednosti H_A, velja pa, da večjo vrednost λ_A kot imamo, (enakomerno) večja bo moč tega testa. Torej je ta testna statistika res enakomerno najmočnejša za testiranje H_0.

```
pod = read.csv("data/data_semafor.csv")
n=length(pod$x)
lambda0=3
qpois(0.95,n*lambda0) # tmax- meja zavrnitve
```

```
## [1] 329
```

```
ppois(sum(pod$x)-1,n*lambda0,lower.tail=FALSE)
```

```
## [1] 0.07192727
```

Naloga 5 – študenti

```
studenti = read.csv("data/data_studenti.csv")

modelC=lm(teza~studij,data=studenti)
model0=lm(teza~1,data=studenti)
testna = as.numeric(-2*(logLik(model0)-logLik(modelC)))
df = model0$df.residual - modelC$df.residual
pchisq(testna,df=df,lower.tail=FALSE)
```

```
## [1] 0.2495559
```

```
# vkljucimo se visino
modelC1=lm(teza~visina+studij,data=studenti)
model1=lm(teza~visina,data=studenti)
testna = as.numeric(-2*(logLik(model1)-logLik(modelC1)))
df = model1$df.residual - modelC1$df.residual
pchisq(testna,df=df,lower.tail=FALSE)
```

```
## [1] 0.6266136
```

```
# vkljucimo se spol
modelC2=lm(teza~visina+spol+studij,data=studenti)
model2=lm(teza~visina+spol,data=studenti)
testna = as.numeric(-2*(logLik(model2)-logLik(modelC2)))
df = model2$df.residual - modelC2$df.residual
pchisq(testna,df=df,lower.tail=FALSE)
```

```
## [1] 0.7002433
```

Naloga 8 – hemoglobin

a. Ničelna domneva:
$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$
Alternativna domneva: $$H_1 : \sigma_i^2 \text{ niso vse enake}$$
b. Funkcija verjetja je enaka

$$L(x, \mu, \sigma) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(x_j - \mu)^2}{2\sigma^2},$$

del njenega logaritma v katerem nastopata parametra, ki ju želimo oceniti, pa je enak

$$\log L(x, \mu, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu)^2.$$

Poiščimo maksimum po μ:

$$\begin{aligned}
\frac{\partial \log L(x, \mu, \sigma)}{\partial \mu} &= 0 \\
-\frac{1}{2\widehat{\sigma}^2} \sum_{j=1}^{n} (x_j - \widehat{\mu})(-2) &= 0 \\
\sum_{j=1}^{n} (x_j - \widehat{\mu}) &= 0 \\
\widehat{\mu} &= \frac{1}{n} \sum_{j=1}^{n} x_j
\end{aligned}$$

Pa še za varianco:

$$\begin{aligned}
\frac{\partial \log L(x, \mu, \sigma)}{\partial \sigma} &= 0 \\
-\frac{n}{\widehat{\sigma}} - \frac{1}{2} \sum_{j=1}^{n} (x_j - \widehat{\mu})^2 \frac{-2}{\widehat{\sigma}^3} &= 0 \\
-\widehat{\sigma}^2 n + \sum_{j=1}^{n} (x_j - \widehat{\mu})^2 &= 0 \\
\widehat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^{n} (x_j - \widehat{\mu})^2
\end{aligned}$$

c. Funkcija verjetja pod alternativno domnevo je enaka

$$L(x, \mu, \sigma) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi}\sigma_i} \exp -\frac{(x_{ij} - \mu_i)^2}{2\sigma_i^2}$$

Vsak člen vsote, ki jo dobimo po logaritmiranju gornje funkcije, je sestavljen le iz parametrov enega posameznika, ko odvajamo po tistem parametru torej ostanejo le členi, ki so vezani na tistega posameznika. Za ocenjevanje parametrov za nekega posameznika i torej potrebujemo izključno njegove vrednosti, parametre posameznikov torej ocenimo povsem neodvisno drug od drugega.

d. Pod ničelno domnevo je σ_i enak za vse i, zato ga v logaritmu funkcije verjetja lahko izpostavimo in ne vpliva na našo oceno posameznih povprečij. Ocena posameznih povprečij je zato enaka kot pod alternativno domnevo.

e. Del logaritma funkcije verjetja, ki nas zanima, je enak

$$\log L(x, \mu, \sigma_0) = -\sum_{i=1}^{k} n_i \log \sigma_0 - \frac{1}{2\sigma_0^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2.$$

Odvod po σ izenačimo z 0 in dobimo

$$\hat{\sigma}_0^2 = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu}_i)^2$$

f. Zapišemo Wilksov Λ (zgoraj je funkcija verjetja pod alternativno domnevo, spodaj pod ničelno): Pod alternativno optimiziramo po več parametrih - vedno je večja vrednost, zato je razmerje večje od 1 in log večji od 0

$$\begin{aligned}
\Lambda &= \frac{\prod_{i=1}^{k} \prod_{j=1}^{n_i} \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}_i}} \exp\{-\frac{(x_{ij}-\widehat{\mu}_i)^2}{2\widehat{\sigma}_i^2}\} \right)}{\prod_{i=1}^{k} \prod_{j=1}^{n_i} \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}_0}} \exp\{-\frac{(x_{ij}-\widehat{\mu}_{0i})^2}{2\widehat{\sigma}_0^2}\} \right)} \\
&= \frac{\left(\prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\widehat{\sigma}_i}} \right) \prod_{i=1}^{k} \exp\{-\frac{\sum_{j=1}^{n_i} (x_{ij}-\widehat{\mu}_i)^2}{2\widehat{\sigma}_i^2}\}}{\left(\prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\widehat{\sigma}_0}} \right) \exp\{-\sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(x_{ij}-\widehat{\mu}_{0i})^2}{2\widehat{\sigma}_0^2}\}}
\end{aligned}$$

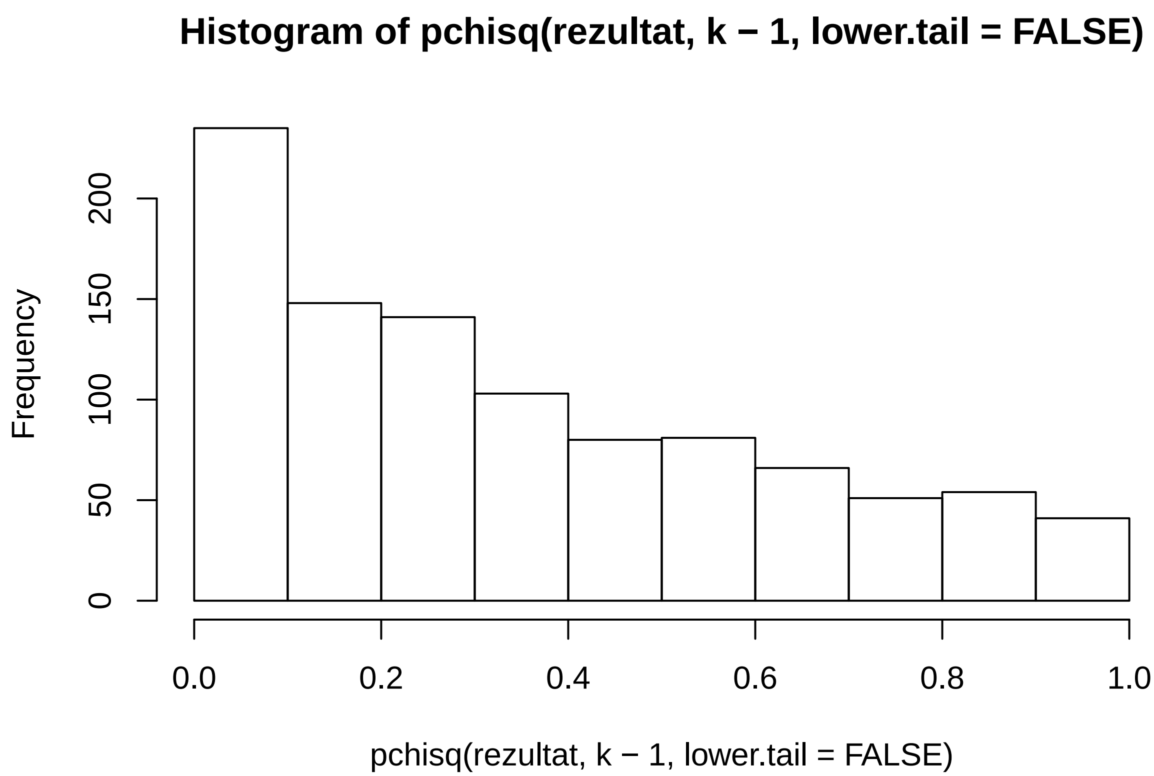
Vstavimo ocene za variance v eksponent in tako v števcu kot tudi v imenovalcu dobimo $\exp\{-\frac{1}{2} \sum_{i=1}^{k} n_i\}$, ki se zato pokrajša. Logaritem Λ je enak

$$\begin{aligned}
log\Lambda &= -\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} \log(\widehat{\sigma}_i) \right) + \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} \log(\widehat{\sigma}_0) \right) \\
&= \left(\sum_{i=1}^{k} n_i \log(\widehat{\sigma}_0) \right) - \left(\sum_{i=1}^{k} n_i \log(\widehat{\sigma}_i) \right) \\
&= \sum_{i=1}^{k} n_i \left[\log(\widehat{\sigma}_0) - \log(\widehat{\sigma}_i) \right]
\end{aligned}$$

Dvakratna vrednost logaritma verjetij je porazdeljena kot χ^2_{k-1}, saj smo pod alternativno domnevo ocenili $k-1$ parametrov več kot pod ničelno.

```
set.seed(1)

k = 30 #st. sportnikov
ni = 10 #st. meritev za vsakega sportnika
si = rep(5,k) # varianca vsakega sportnika
mi = rnorm(k,148,7) # povprecje vsakega sportnika
sportnik = data.frame(mi=mi,si=si)

meritve = as.data.frame(apply(sportnik,1,
                        FUN=function(x){rnorm(ni,x[1],x[2])}))
# vsak stolpec prikazuje meritve za enega sportnika

#ocenimo parametre
miHat = apply(meritve,2,mean)# za vsak stolpec
siHat = apply(meritve,2,function(x){sqrt(1/ni*sum((x-mean(x))^2))})
```

```
# MLE - pristranska cenilka

#ocena s_0 pod H0
#vsi si so enaki
xMinusMi = apply(meritve,2,function(x){sum((x-mean(x))^2)})
si0Hat = sqrt(sum(xMinusMi)/(ni*k))

# -2 log Wilksov Lambda

logL = 2*ni*sum(log(si0Hat) - log(siHat))

ponovi = 1000
rezultat = numeric(ponovi)
for(it in 1:ponovi){
  mi = rnorm(k,148,7) # povprecje vsakega sportnika
  sportnik = data.frame(mi=mi,si=si)
  meritve = as.data.frame(apply(sportnik,1,FUN=function(x){rnorm(ni,x[1],x[2])}))
  miHat = apply(meritve,2,mean)# za vsak stolpec
  siHat = apply(meritve,2,function(x){sqrt(1/ni*sum((x-mean(x))^2))})
  xMinusMi = apply(meritve,2,function(x){sum((x-mean(x))^2)})
  si0Hat = sqrt(sum(xMinusMi)/(ni*k))
  logL = 2*ni*sum(log(si0Hat) - log(siHat))
  rezultat[it] = logL
}

hist(pchisq(rezultat,k-1,lower.tail=FALSE))
```



```r
hist(rezultat)
```

Komentar: porazdelitev vrednosti p ni čisto taka, kot bi jo pričakovali. Zakaj? Razmislite o velikosti vzorca in o številu parametrov, ki jih preverjate (ter o *le* asimptotskih lastnostih testne statistike). Kako bi bilo smiselno spremeniti število parametrov in velikost vzorca, da bi dobili bolj pričakovano porazdelitev vrednosti p.