Cvičení 9 – Testování hypotéz

Příklad 1

Příklad 1a

Podíváme se na kvantilovou funkci (tj. inverzní distribuční funkci) normálního rozdělení:

```
qnorm(1-0.05/2)
```

```
## [1] 1.959964
```

```
qnorm(0.05/2)
```

```
## [1] -1.959964
```

Také můžeme přímo použít kvantilovou funkci N(5,1)

```
qnorm(1-0.05/2,5,1)
```

```
## [1] 6.959964
```

```
qnorm(0.05/2,5,1)
```

```
## [1] 3.040036
```

Ještě to ověříme pomocí samplování:

```
n=10^5
data = rnorm(n,mean=5,sd=1)
sum(data >= 5-1.96 & data <= 5+1.96)/n
```

```
## [1] 0.95127
```

Příklad 1b

Podíváme se na inverzní funkci normálního rozdělení. Rozptyl je roven 1/n, takže směrodatná odchylka 1/sqrt(n).

```
n=10
qnorm(1-0.05/2,sd=1/sqrt(n))
```

```
## [1] 0.619795
```

```
qnorm(0.05/2,sd=1/sqrt(n))
```

```
## [1] -0.619795
```

```
qnorm(1-0.05/2,5,1/sqrt(n))
```

```
## [1] 5.619795
```

```
qnorm(0.05/2,5,1/sqrt(n))
```

```
## [1] 4.380205
```

Příklad 1c

Zajímá nás distribuční funkce v bodech

```
-1.96+sqrt(10)
```

```
## [1] 1.202278
```

```
1.96+sqrt(10)
```

```
## [1] 5.122278
```

Nezamítneme s pravděpodobností

```
pnorm(1.96+sqrt(10))-pnorm(-1.96+sqrt(10))
```

```
## [1] 0.1146278
```

```
pnorm(5+1.96/sqrt(10),mean=4,sd=1/sqrt(10))-pnorm(5-1.96/sqrt(10),mean=4,sd=1/sqrt(10))
```

```
## [1] 0.1146278
```

Spočetli jsme tedy chybu 2. druhu cca 11 %. Pravděpodobnost, že tuto chybu neuděláme se nazývá síla test, v tomto případě tedy 88 %.

Pro kontrolu ještě zkusíme nasamplovat. (Seriózní kód v R by nepoužíval for-cyklus, ale takhle je to asi názornější.)

```
n=10
opak=10^4
chyb = 0   # počet chyb druhého druhu
for(i in 1:opak){
  data = rnorm(n,4,1)
  m = mean(data)
  if (m > 5-1.96/sqrt(n) & m<5+1.96/sqrt(n)){ chyb = chyb+1 }
}

chyb/opak
```

```
## [1] 0.1165
```

Příklad 1d

Podle CLV funguje stejný postup, ale "platí jen v limitě". Tady zkusíme jenom samplovat, pro rozdělení, které má střední hodnotu 5 a rozptyl 1.

```
n=71
opak=10^4
chyb = 0
for(i in 1:opak){
  data = rbinom(n,4,1/2)+3
  m = mean(data)
  if (m > 5-1.96/sqrt(n) & m<5+1.96/sqrt(n)){ chyb = chyb+1 }
}

chyb/opak
```

```
## [1] 0.9511
```

Příklad 2

Příklad 2a

Stačí použít funkce mean() a var(), ale pro zkoušku (a připomenutí vzorců) to spočítáme i přímo:

```
data = c(8.47,10.91,10.87,9.46,10.40)
n = length(data); n
```

```
## [1] 5
```

```
mu=mean(data); mu
```

```
## [1] 10.022
```

```
sum(data)/n
```

```
## [1] 10.022
```

```
v = var(data); v
```

```
## [1] 1.09377
```

```
sum((data-mu)^2)/(n-1)
```

```
## [1] 1.09377
```

Příklad 2b

Postupujeme stejně jako minule. Považujeme průměr za n.v. s rozdělením N(9,v/n). Opět z cvičných důvodů vyřešíme dvěma způsoby: buď převedením na kvantilovou funkci standardního normálního rozdělení, nebo přímo.

```
9+qnorm(1-0.05/2)*sqrt(v/n)
```

```
## [1] 9.916698
```

```
9+qnorm(0.05/2)*sqrt(v/n)
```

```
## [1] 8.083302
```

```
qnorm(1-0.05/2,mean=9,sd=sqrt(v/n))
```

```
## [1] 9.916698
```

```
qnorm(0.05/2,mean=9,sd=sqrt(v/n))
```

```
## [1] 8.083302
```

Příklad 2c

```
9+qt(0.975,4)*sqrt(v/n)
```

```
## [1] 10.29858
```

```
9+qt(0.025,4)*sqrt(v/n)
```

```
## [1] 7.701425
```

```
28/600
```

```
## [1] 0.04666667
```

```r
qbinom(0.95,600,0.03)
```

```
## [1] 25
```

```r
p = 0.03
n=600
pnorm((28-600*p)/sqrt(p*(1-p)*n))
```

```
## [1] 0.9916488
```

```r
qnorm(0.95)
```

```
## [1] 1.644854
```

```r
qbinom(0.95,1000,.5)
```

```
## [1] 526
```

Příklad 4

Příklad 4a

Postup přes CLV – aproximace normálním rozdělením. Napřed spočteme parametry.

```r
emaily = c(34,35,29,31,30)
n = length(emaily); n
```

```
## [1] 5
```

```r
mu = mean(emaily); mu
```

```
## [1] 31.8
```

```r
var(emaily)
```

```
## [1] 6.7
```

Poisson má stejný rozptyl a střední hodnotu, tak aproximujeme průměr pomocí N(35,35/n). Naměřený výběrový rozptyl je o hodně menší než 35, ale to teď pomiňme.

```r
abs(mu-35) < (qnorm(0.975)*sqrt(35/n))
```

```
## [1] TRUE
```

```r
35+qnorm(0.975)*sqrt(35/n)
```

```
## [1] 40.18558
```

```r
35+qnorm(0.025)*sqrt(35/n)
```

```
## [1] 29.81442
```

```r
#var(emaily)
```

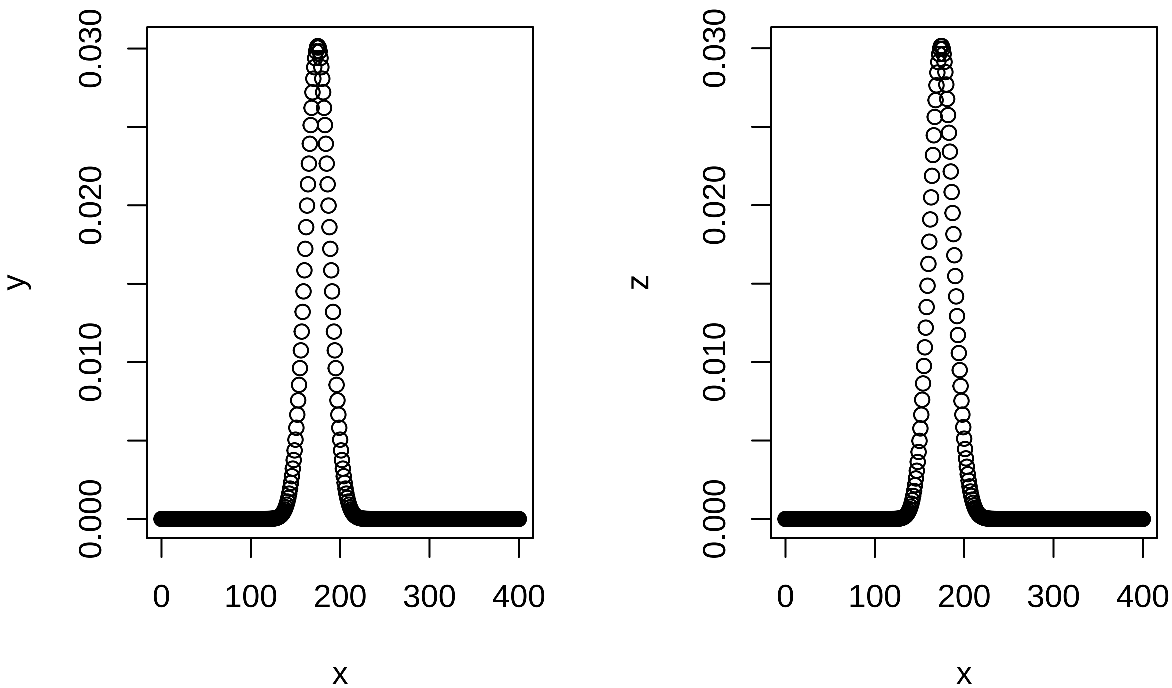
Ilustrace přesnosti (kreslíme dvakrát, abychom si ilustrovali i ukládání do souboru).

```r
x = 0:400
y = dnorm(x,35*5,sqrt(35*5))
z = dpois(x,35*5)
jpeg(file="pois_CLV.jpg")
par(mfrow=c(2,1))
```

```r
plot(x,y)
plot(x,z)
dev.off()
```

```
## pdf
##   2
```

```r
par(mfrow=c(1,2))
plot(x,y)
plot(x,z)
```



##
Příklad 4b Přístup přes součet Poissonů: X_1 + … + X_5 má rozdělení (přesně) Pois(35*5).

```r
cat("dolní okraj intervalu:", qpois(0.025,35*n), "\n")
```

```
## dolní okraj intervalu: 150
```

```r
cat("horní okraj intervalu:", qpois(0.975,35*n), "\n")
```

```
## horní okraj intervalu: 201
```

```r
cat("naměřený součet:", sum(emaily))
```

```
## naměřený součet: 159
```

Nad rámec zadání: pravděpodobnost, že počet emailů za týden bude nejvýše 159 je cca 12 %, tj. nic moc divného.

```r
ppois(159,35*n)
```

```
## [1] 0.1196265
```

```r
d = rpois(5,35)
mean(d)
```

```
## [1] 32
```

```r
var(d)
```

```
## [1] 43.5
```

```r
var(rpois(5,35))
```

```
## [1] 17.7
```

A ještě jednou nad rámec: podíváme se, jak často je výběrový rozptyl tak malý, jak nám vyšel. Žádná teorie, jenom jednoduché samplování. Vychází cca 6 %, tj. ne zas tolik podezřelé, jak to vypadá.

```r
opak=10^4

divne = 0
for(i in 1:opak){
  if (var(rpois(5,35)) <= 6.7){
    divne = divne+1
  }
}

divne/opak
```

```
## [1] 0.0634
```

Totéž elegantněji (ale trochu pomaleji).

```r
opak=10^4
sum(replicate(opak, var(rpois(5,35)))<=6.7)/opak
```

```
## [1] 0.0601
```