cv5.R

Marek

2023-03-16

```r
library('latexpdf')
library('latex2exp')
```

Jednovyberove testy hypotez pre parametre normalneho rozdelenia.

Cize musi platit predpoklad, ze data su normalne rozdelene (Shapiro Wilk test) Hypoteza je tvrdenie, ktoreho pravdivost je treba overit statistickymi metodami. Vysledok setrenia je zatazeny dvomi chybami (pravdepodobnostou tychto chyb). Prva chyba, chyba prveho druhu (pravdepodobnost chyby) α je hladina vyznamnosti. Zamietam nulovu hypotezu H_0 (tvrdenie) a ona plati. Chyba druheho druhu je β, nezamietam H_0 a H_0 neplati. Jednu chybu fixneme a druhu minimalizujeme tak, ze spravne vykoname test. K H_0 sa pridava alternativna hypoteza H_1. Ak hypotezu H_0 zamietnem, tak prijmem aternativu H_1

Test pre strednu hodnotu μ ak σ je zname

$$H_0 \quad \mu = \mu_0 \qquad H_1 \mu \neq \mu_0$$

Zvolime vhodnu statistiku, vypocitame ju, rozdelime R na oblast, kde zamietam H_0 (to sa vola kriticka oblast testu) a kde nezamietam H_0 (dokopy musia dat obor hodnot) Priklad 1 Vyrobca uvadza, ze v kopirke je nutne menit toner v priemere po 2500 skopirovanych stranach so smerodajnou odchylkou $\sigma = 30$. Na hladine vyznamnosti $\alpha = 0.05$ testujte, ci tvrdenie vyrobcu je v sulade so skutocnostou. Najprv vypoctom statistiky, potom prikazom v R. Kazdy test sa zacina stanovenim nulovej a alternativnej hypotezy.

$$H_0 \quad \mu = 2500 \qquad H_1 \quad \mu \neq 2500$$

```r
x <- c(2445, 2450, 2453, 2462, 2463, 2463, 2466, 2471, 2474, 2475, 2475,
       2484, 2485, 2486, 2487, 2490, 2491, 2493, 2499, 2501, 2501, 2503,
       2504, 2505, 2505, 2506, 2506, 2507, 2509, 2511, 2511, 2513, 2514,
       2515, 2518, 2523, 2523, 2524, 2525, 2527, 2529, 2530, 2530, 2533,
       2535, 2536, 2537, 2539, 2560, 2571)
alfa <- 0.05
mean <- mean(x)
sigma <- 30
n <- length(x)
mi0 <- 2500
t <- (mean - mi0) * sqrt(n) / sigma
t
```

```
## [1] 0.7683894
```

```r
k1 <- qnorm(alfa / 2, 0, 1)
k2 <- qnorm(1 - alfa / 2, 0, 1)
c(t, k1, k2)
```

```
## [1]  0.7683894 -1.9599640  1.9599640
```

Statistika padla do intervalu (-1.96, 1.96), nezamietame H_0, tvrdenie vyrobcu je pravdive alebo v sulade so skutocnostou. pomocou kniznice DescTools a prikazu v R.

```r
library('DescTools')
ZTest(x, mu = 2500, sd_pop = 30) #chyba je defaultne nastavena na 0.05
```

```
##
##  One Sample z-test
##
## data:  x
## z = 0.76839, Std. Dev. Population = 30, p-value = 0.4423
## alternative hypothesis: true mean is not equal to 2500
## 95 percent confidence interval:
##  2494.945 2511.575
## sample estimates:
## mean of x
##   2503.26
```

rozhodnutie bud pozriem, ci testovana hodnota je z prislusneho intervalu spolahlivosti, alebo rozhodneme podla P-Value, pravdepodobnostnej hodnoty P-Hodnoty. Ak je p-hodnota mensia ako alfa (0.05), tak zamietam nulovu hypotezu H_0.

```r
ZTest(x, mu = 2500, sd_pop = 30)$p.value
```

```
## [1] 0.4422559
```

P-Hodnota je 0.44 > 0.05, teda nezamietam H_0

```r
pvalue <- pnorm(-abs(t)) * 2
pvalue
```

```
## [1] 0.4422559
```

Toto bolo o vyrobcovi. Mna ako pouzivatela skor zaujima overenie

tvrdenia, ze nakopirujem aspon 2500 stran

$$H_0 \quad \mu = 2500 (\geq 2500) \quad H_1 \quad \mu < 2500$$

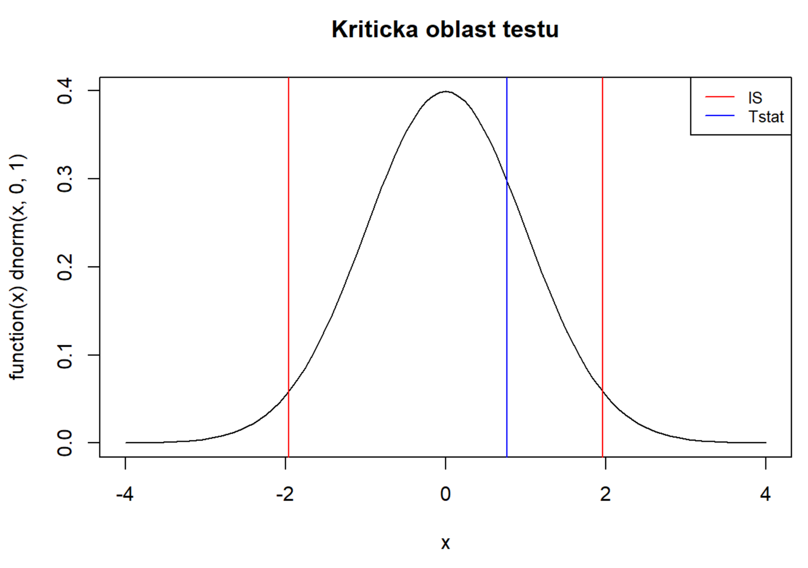
alternativa je less

```r
ZTest(x, mu=2500, sd_pop=30, alternative = 'l')
```

```
##
##  One Sample z-test
##
## data:  x
## z = 0.76839, Std. Dev. Population = 30, p-value = 0.7789
## alternative hypothesis: true mean is less than 2500
## 95 percent confidence interval:
##      -Inf 2510.239
## sample estimates:
## mean of x
##   2503.26
```

P-Hodnota je 0.77>0.05, nezamietam H_0 nakopirujem tolko alebbo viac, ale urcite nie menej Graficky vystup

```r
plot(function(x)dnorm(x, 0, 1),
     xlim=c(-4, 4),
     main = 'Kriticka oblast testu')
abline(v = k1, col = 'red')
abline(v = k2, col = 'red')
abline(v = t, col = 'blue')
legend('topright', c('IS', 'Tstat'), col = c('red', 'blue'), lt = 1, cex = 0.8)
```



Test pre strednu hodnotu, ak σ nie je zname

Rieste rovnaku ulohu ako hore, ale za predpokladu, ze σ nepozname. Najskor vypoctom, potom prikazom.

```r
sd <- sd(x)
t <- (mean - mi0) * sqrt(n) / sd
k1 <- qt(alfa / 2, n - 1)
k2 <- qt(1 - alfa / 2, n - 1)
c(t, k1, k2)
```

```
## [1]  0.8214202 -2.0095752  2.0095752
```

Hodnota vypocitanej statistiky padne do Intervalu Spolahlivosti, nezamietam H_0.

```r
#abline(v = k1, col = 'green')
#abline(v = k2, col = 'green')
#abline(v = t, col = 'purple')
```

prikazom v R, rozhodneme podla P hodnoty

```r
t.test(x, mu = 2500)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 0.82142, df = 49, p-value = 0.4154
## alternative hypothesis: true mean is not equal to 2500
## 95 percent confidence interval:
##  2495.285 2511.235
## sample estimates:
## mean of x
##   2503.26
```

```r
t.test(x, mu = 2500)$p.value
```

```
## [1] 0.4153858
```

P-Hodnota = 0.41 > 0.05, nezamietame H_0. Tvrdenie o pocte skopirovanych stran je pravdive Zmena hladiny vyznamnosti prikladom Maly rodinny podnik predava 100% bio jablkovu stavu v baleni 0.5l. Po oprave pniacej linky sme namerali tieto hodnoty objemu balenia(ml) na hladine vyznamnosti $\alpha = 0.1$ testujte hypotezu, ze plniaca linka je dobre nastavena

$$H_0 \quad \mu = 500 \quad H_1 \mu \neq 500$$

```r
js <- c(495.2,496.8,502.1,498.5,501,503,500.7,
        501.5,501.8,499.1,500.9,502.2,501.7, 500.4,
        500.2,501.1,499.9,500.2,501.1,500.8,499.3)
t.test(js, mu = 500, conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  js
## t = 0.89828, df = 20, p-value = 0.3797
## alternative hypothesis: true mean is not equal to 500
## 90 percent confidence interval:
##  499.6714 501.0429
## sample estimates:
## mean of x
##  500.3571
```

P-Hodnota je 0.37, co je viac ako 0.1, takze H_0 nezamietame, plniaca linka je dobre nastavena Dalsi priklad Firma ABC, ktora vyraba baterie do notebookov tvrdi, ze jednu vyrobi v priemere do 13 minut (max 13 minut). Na hladine vyznamnosti $\alpha = 0.05$ overte toto tvrdenie, ak mate k dispozicii dataset dlzok vyroby, greater

$$H_0 \quad \mu = 13 (\leq 13) \quad H_1 \quad \mu > 13$$

```r
bat <- c(12,19,16,11,7,12,14,18,15,19,17,19,13,9,
         11,20,12,19,8,13)
t.test(bat, mu = 13, alternative = 'g')
```

```
##
##  One Sample t-test
##
## data:  bat
## t = 1.3346, df = 19, p-value = 0.09889
## alternative hypothesis: true mean is greater than 13
## 95 percent confidence interval:
##  12.6453     Inf
## sample estimates:
## mean of x
##      14.2
```

```r
t.test(bat, mu = 13, alternative = 'g')$p.value
```

```
## [1] 0.09888518
```

P hodnota = 0.098 < 0.05, nezamietam H_0, tvrdenie firmy je v sulade so skutocnostou. Ak by sa zmenila hladina vyznamnosti na 0.1, tak 0.098 < 0.1 na tejto hladine vyznamnosti by sme zamietli hypotezu Dalsi priklad Firma XYZ nakupuje baterky do elektronickych hraciek. Vyrobca garantuje, ze vydrzia minimalne 19 hodin nepretrzitej prevadzky (19 a viac). Kontrolor nahodne vyberie 12 bateriek a zisti udaje o zivotnosti. Testujte na hladine vyznamnosti $\alpha = 0.05$.

$$H_0 \quad \mu = 19 (\geq 19) \qquad H_1 \quad \mu < 19$$

alternativa bude less

```r
bat1 <- c(20.2,19.6,18.6,19.4,17,18.5,18,18.4,19,
18,17.9,18.1)
t.test(bat1, mu = 19, alternative = 'l')$p.value
```

```
## [1] 0.05397316
```

P hodnota 0.053 > 0.05, nezamietam H_0, vydrzia 19 a viac, nie menej

Test pre disperziu overte tvrdenie vyrobcu (prvy priklad) aj co sa tyka disperzie, najprv vypoctom a potom prikazom v R

```r
alfa <- 0.05
t <- (n - 1) * var(x) / 30^2
k1 <- qchisq(alfa / 2, n - 1)
k2 <- qchisq(1 - alfa / 2, n - 1)
c(t, k1, k2)
```

```
## [1] 42.87736 31.55492 70.22241
```

Pomocou kniznice EnvStats

```r
library('EnvStats')
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

```r
VarTest(x, sigma.squared = 30^2)
```

```
##
##  One Sample Chi-Square test on variance
##
## data:  x
## X-squared = 42.877, df = 49, p-value = 0.5632
## alternative hypothesis: true variance is not equal to 900
## 95 percent confidence interval:
##   549.5342 1222.9353
## sample estimates:
## variance of x
##      787.5433
```

Statistika padla do oblasti, kde nezamietam H_0, tvrdenie vyrobcu o disperzii je pravdive # dvojvyberove testy o parametroch normalneho rozdelenia parove (zavisle merania) neparove (nezavisle merania) Parove testy su urcene pre dvojice dat, ktore su spojene jednym objektom, prah pocutelnosti, prave lave ucho, ojazdenost pneu predna/zadna data by sme mali dostat ako dvojice, najcastejsie sa pouzivaju v pripadoch pred liecbou, po liecbe, pred zakrokom, po zakroku, pred skolenim, po skoleni, zaujima nas zmena (X, Y), Z = X - Y, a sa nic nemzenilo, tak vlastne testujem jednovyberovym testom, ze stredna hodnota Z je nula. Su dane casy v sekundach, pocas ktorych vyriesili kontrolne ulohy pred a po specialnych cviceniach z pamatoveho pocitania. Zlepsili cvicenia schopnost ziakov rychlejsie riesit ulohy? (zlepsili sa, ak to zvladli rychlejsie), pred - po >= 0, alternativa je menej ako 0

```r
pred <- c(87,61,98,90,93,74,83,72,81,75,83)
po <- c(50,45,79,90,88,65,52,79,84,61,52)
z <- pred - po
t.test(z, mu = 0, alternative = 'l')
```

```
##
##  One Sample t-test
##
## data:  z
## t = 3.1128, df = 10, p-value = 0.9945
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##      -Inf 21.86391
## sample estimates:
## mean of x
##  13.81818
```

P-Hodnota = 0.99 nezamietam H_0, nezmenili alebo zlepsili, nie zhorsili

```r
t.test(pred, po, mu = 0, paired = T, alternative = 'l')
```

```
##
##  Paired t-test
##
## data:  pred and po
## t = 3.1128, df = 10, p-value = 0.9945
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##      -Inf 21.86391
## sample estimates:
## mean difference
##        13.81818
```