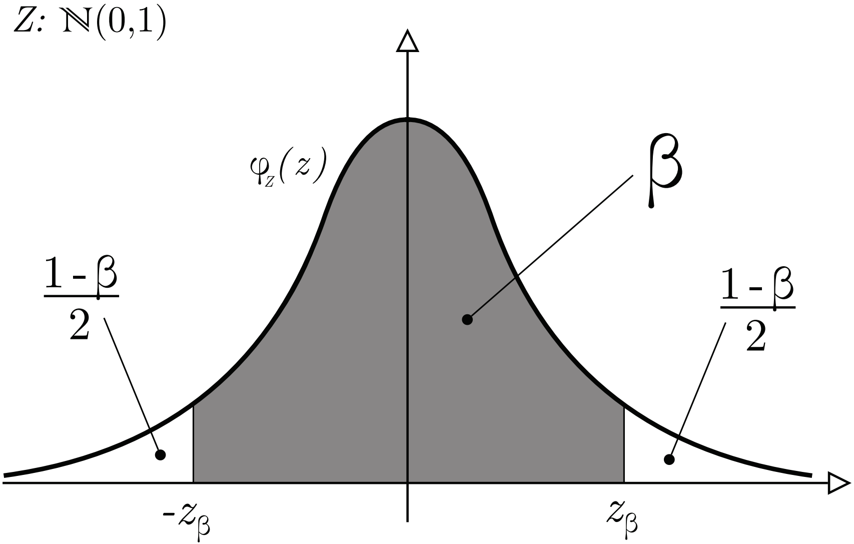
Vežbe 9 i 10

Intervali poverenja

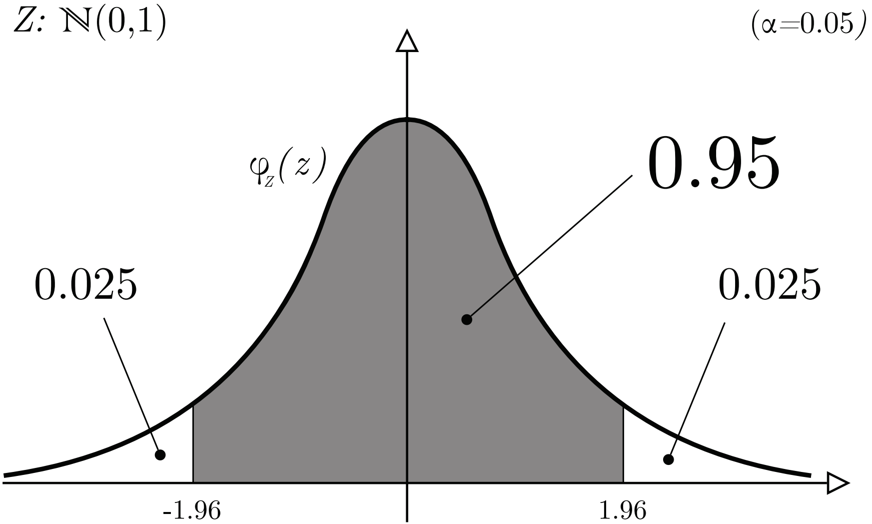
Intervali poverenja

1. Čak i najefikasnija centrirana tačkasta ocena parametra ne mora da tačno oceni traženi parametar populacije.
2. Nova ideja u vidu intervalne ocene (interval unutar koga je očekivano da se nalazi parametar):
$$\hat{\theta}_l \leq \theta \leq \hat{\theta}_r,$$
gde su $\hat{\theta}_l$ i $\hat{\theta}_r$ statistike koje zavise od $\hat{\theta}$, koje opet zavisi od uzorka.
3. Ako se nađu $\hat{\theta}_l$ i $\hat{\theta}_r$ koje zadovoljavaju jednakost:
$$P(\hat{\theta}_l \leq \theta \leq \hat{\theta}_r) = 1 - \alpha, \qquad \text{za } \alpha \in (0, 1),$$
onda je verovatnoća da se θ nalazi u intervalu izračunatom na osnovu slučajno odabranog uzorka jednaka $1 - \alpha$.
4. Interval (θ_l, θ_r), izračunat na osnovu datog uzorka, se zove $100(1 - \alpha)$-to postotni interval poverenja, a broj $\beta = 1 - \alpha$ nivo poverenja.

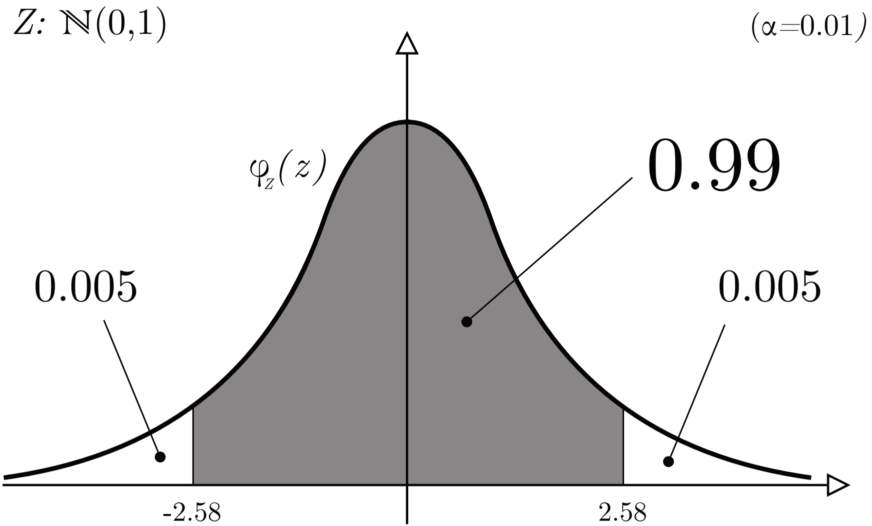
Interval poverenja - ideja



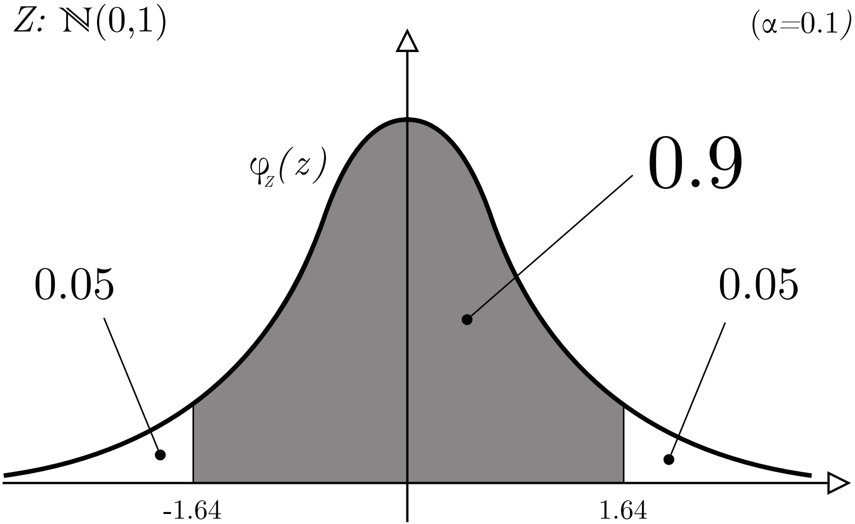
Interval poverenja - ideja



Interval poverenja - ideja



Interval poverenja - ideja



Intervali poverenja za m gde $X : N(m, \sigma)$, σ poznato

1. Prema Centralnoj graničnoj teoremi je $\bar{X}_n : N(m, \sigma/\sqrt{n})$, što znači da:

$$Z = \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \text{ ima aproksimativno } N(0,1) \text{ raspodelu.}$$

2. Zato ako važi $P(-z_\beta < Z < z_\beta) = \beta$, važi i:

$$P\left(-z_\beta < \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} < z_\beta\right) = \beta,$$

   odnosno, nakon sređivanja:

$$P\left(\bar{X}_n - z_\beta \frac{\sigma}{\sqrt{n}} \ < \ m \ < \ \bar{X}_n + z_\beta \frac{\sigma}{\sqrt{n}}\right) = \beta,$$

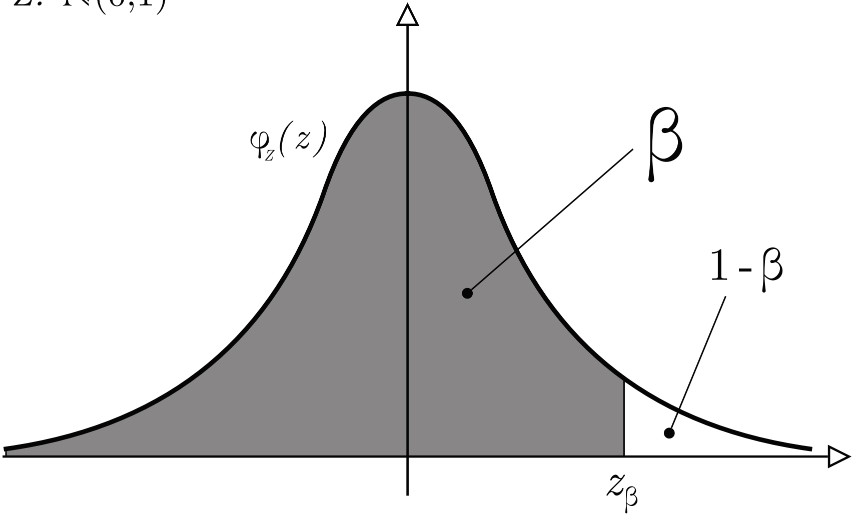
   gde je $z_\beta = \Phi^{-1}\left(\beta + \frac{1-\beta}{2}\right) = \Phi^{-1}\left(\frac{1+\beta}{2}\right) = -\Phi^{-1}\left(\frac{1-\beta}{2}\right)$ (`qnorm`).

Intervali poverenja za m gde $X : N(m, \sigma)$, σ nepoznato

1. Pokazano je na predavanjima da:
$$T = \frac{\bar{X}_n - m}{\bar{S}_n/\sqrt{n-1}} = \frac{\bar{X}_n - m}{\hat{S}_n/\sqrt{n}}$$
ima aproksimativno Studentovu t-raspodelu sa $(n-1)$ stepeni slobode.
2. Zato ako važi $P(-t_\beta < T < t_\beta) = \beta$, važi i:
$$P\left(-t_\beta < \frac{\bar{X}_n - m}{\hat{S}_n/\sqrt{n}} < t_\beta\right) = \beta,$$
odnosno, nakon sređivanja:
$$P\left(\bar{X}_n - t_\beta\frac{\hat{S}_n}{\sqrt{n}} \ < \ m \ < \ \bar{X}_n + t_\beta\frac{\hat{S}_n}{\sqrt{n}}\right) = \beta,$$
gde je t_β kvantil reda $(1+\beta)/2$ Studentove raspodele sa $(n-1)$ stepena slobode.

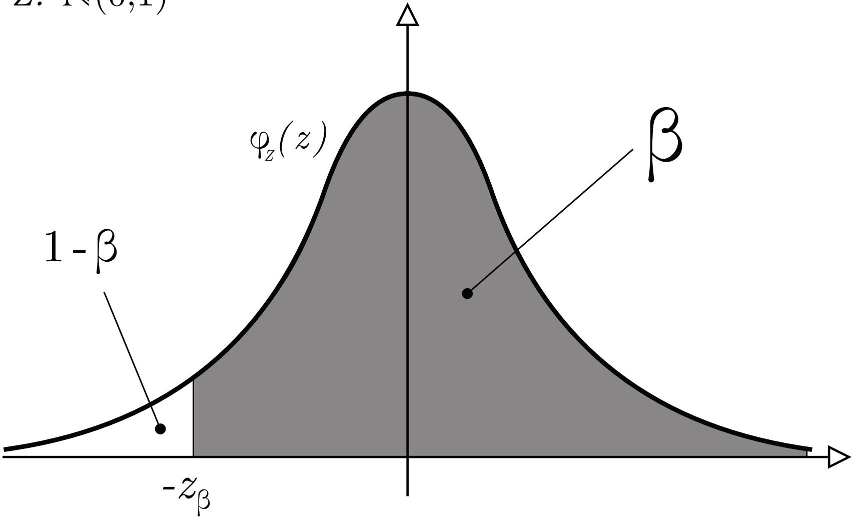
Interval poverenja - ideja

Z: $\mathbb{N}(0,1)$



Interval poverenja - ideja

Z: $\mathbb{N}(0,1)$



Jednostrani interval poverenja za m gde $X : N(m, \sigma)$, σ poznato

1. Slična ideja, ali ovoga puta je $P(Z < z_\beta) = \beta$, odnosno:

$$P\left(\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} < z_\beta\right) = \beta,$$

odakle direktno sledi i:

$$P\left(m \ > \ \bar{X}_n - z_\beta \frac{\sigma}{\sqrt{n}}\right) = \beta, \qquad z_\beta = \Phi^{-1}(\beta).$$

2. Za gornje ograničenje $P(Z > -z_\beta) = \beta$:

$$P\left(\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} > -z_\beta\right) = \beta,$$

odakle direktno sledi i:

$$P\left(m \ < \ \bar{X}_n + z_\beta \frac{\sigma}{\sqrt{n}}\right) = \beta, \qquad \text{gde je } -z_\beta = \Phi^{-1}(1-\beta), \text{ to jest } z_\beta = \Phi^{-1}(\beta).$$

Jednostrani interval poverenja za m gde $X : N(m, \sigma)$, σ nepoznato

1. Slična ideja, ali ovoga puta je $P(T < t_\beta) = \beta$, odnosno:
$$P\left(\frac{\bar{X}_n - m}{\hat{S}_n/\sqrt{n}} < t_\beta\right) = \beta,$$
odakle direktno sledi i:
$$P\left(m \ > \ \bar{X}_n - t_\beta \frac{\hat{S}_n}{\sqrt{n}}\right) = \beta, \qquad \mathtt{t}_\beta = \mathtt{qt}(\beta, \mathtt{df} = \mathtt{n} - 1).$$
2. Za gornje ograničenje $P(T > -t_\beta) = \beta$:
$$P\left(\frac{\bar{X}_n - m}{\hat{S}_n/\sqrt{n}} > -t_\beta\right) = \beta,$$
odakle direktno sledi i:
$$P\left(m \ < \ \bar{X}_n + t_\beta \frac{\hat{S}_n}{\sqrt{n}}\right) = \beta, \qquad \mathtt{t}_\beta = \mathtt{qt}(\beta, \mathtt{df} = \mathtt{n} - 1).$$

Intervali poverenja parametra σ^2, gde je $X : N(m, \sigma)$

1. Za σ^2, gde je $X : N(m, \sigma)$:
   - y_- je kvantil reda $(1-\beta)/2$ od χ^2_{n-1} raspodele (komanda `qchisq`),
   - y_+ je kvantil reda $(1+\beta)/2$ od χ^2_{n-1} raspodele (`qchisq`),

$$P\left(\frac{n\bar{S}_n^2}{y_+} < \sigma^2 < \frac{n\bar{S}_n^2}{y_-}\right) = P\left(\frac{(n-1)\hat{S}_n^2}{y_+} < \sigma^2 < \frac{(n-1)\hat{S}_n^2}{y_-}\right) = \beta.$$

2. Jednostrani interval poverenja (gornje ograničenje):

$$P\left(0 < \sigma^2 < \frac{n\bar{S}_n^2}{y_0}\right) = P\left(0 < \sigma^2 < \frac{(n-1)\hat{S}_n^2}{y_0}\right) = \beta,$$

gde je y_0 kvantil reda β od χ^2_{n-1} raspodele (`qchisq`).

Intervali poverenja za nepoznati parametar p, gde $X : B(n, p)$

1. Za nepoznati binomni parametar p, gde X ima binomnu $B(n, p)$ raspodelu:

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{X/n - p}{\sqrt{pq/n}} \text{ ima normalnu } N(0,1) \text{ raspodelu,}$$

i proporcija uspešne realizacije procenjuje se na osnovu:

$$P\left(p_u - z_\beta\sqrt{\frac{p_u q_u}{n-1}} \;<\; p \;<\; p_u + z_\beta\sqrt{\frac{p_u q_u}{n-1}}\right) = \beta,$$

pri čemu je:

- $z_\beta = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$ kvantil normalne $N(0,1)$ raspodele,
- K broj uspešnih realizacija,
- n obim uzorka,
- $p_u = K/n$, proporcija uspešno realizovanih,
- $q_u = 1 - p_u$.

Intervali poverenja - dva nezavisna uzorka, $m_1 - m_2$

1. Ako je $E(X_1) = m_1$, $E(x_2) = m_2$, $D(X_1) = \sigma_1$ i $D(X_2) = \sigma_2$, onda prema CGT:
$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \text{ ima normalnu } N(0,1) \text{ raspodelu,}$$
koja se koristi za ocenu razlike parametara $m_1 - m_2$.
2. Zato ako važi da je verovatnoća:
$$P\left(-z_\beta \; < \; \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \; < \; z_\beta\right) = \beta,$$
onda je i posledično:
$$P\left((\bar{x}_{n_1} - \bar{x}_{n_2}) - z_\beta\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \; < \; m_1 - m_2 \; < \; (\bar{x}_{n_1} - \bar{x}_{n_2}) + z_\beta\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = \beta,$$
gde je z_β kvantil reda $(1+\beta)/2$ normalne $N(0,1)$ raspodele.

Intervali poverenja - dva nezavisna uzorka, $p_1 - p_2$

1. Ako su $X_1 : B(n_1, p_1)$ i $X_2 : B(n_2, p_2)$ nezavisne, onda prema prethodnim slajdovima:

$$Z = \frac{(X_1/n_1 - X_2/n_2) - (p_1 - p_2)}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}} \text{ ima normalnu } N(0,1) \text{ raspodelu,}$$

koja se koristi za ocenu razlike parametara $p_1 - p_2$.

2. Zato ako važi jednakost:

$$P\left(-z_\beta \ < \ \frac{(X_1/n_1 - X_2/n_2) - (p_1 - p_2)}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}} \ < \ z_\beta \right) = \beta,$$

onda je sa verovatnoćom β i:

$$(p_{u_1} - p_{u_2}) - z_\beta \sqrt{\frac{p_{u_1} q_{u_1}}{n_1} + \frac{p_{u_1} q_{u_1}}{n_2}} \ < \ p_1 - p_2 \ < \ (p_{u_1} - p_{u_2}) + z_\beta \sqrt{\frac{p_{u_1} q_{u_1}}{n_1} + \frac{p_{u_1} q_{u_1}}{n_2}},$$

gde je z_β kvantil reda $(1+\beta)/2$ normalne $N(0,1)$ raspodele.

Intervali poverenja - dva nezavisna uzorka, količnik uz. disperzija

1. Ako su σ_1^2 i σ_2^2 disperzije normalnih populacija, intervalna procena za količnik σ_1^2/σ_2^2 može se dobiti koristeći statistiku:

$$F = \frac{\sigma_1^2 \hat{S}_1^2}{\sigma_2^2 \hat{S}_2^2},$$

za koju je poznato da ima Fišerovu raspodelu sa $n_1 - 1$ i $n_2 - 1$ stepeni slobode. Naime, ako važi $P(f_{(1-\beta)/2} < F < f_{(1+\beta)/2}) = \beta$, onda važi i:

$$P\left(\frac{\hat{S}_1^2}{\hat{S}_2^2} f_{(1-\beta)/2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\hat{S}_1^2}{\hat{S}_2^2} f_{(1+\beta)/2}\right) = \beta,$$

gde su $f_{(1\pm\beta)/2}$ kvantili reda $(1 \pm \beta)/2$ Fišerove raspodele sa $n_1 - 1$ i $n_2 - 1$ stepeni slobode.

p-vrednost, α^*

Kritična oblast

- Dat je nivo poverenja β, koji direktno služi za računanje kvantila, koji se koriste za proračun granica intervala poverenja. $\alpha = 1 - \beta$ zove prag značajnosti. Kritična oblast je (skupovni) komplement intervala poverenja, i on zavisi od α (samim tim i od β).

- Statistički testovi služe za utvrđivanje postojanja eventualno statistički značajnih razlika u raspodelama, svojstvima itd., na osnovu uzorka. Svaki statistički test ima svoju nultu (H_0) i alternativnu (H_1) hipotezu. Nulta hipoteza uvek tvrdi da ne postoji statistički značajan efekat, a alternativna suprotno.

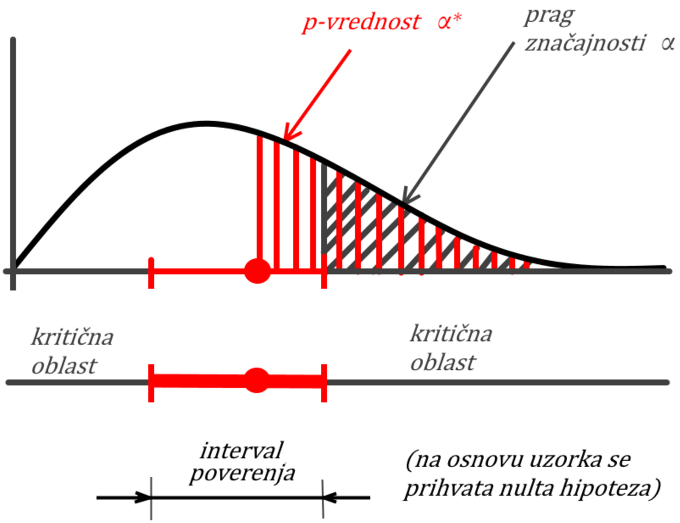
- Postoje dva načina testiranja:
  - (ne)pripadanje pretpostavljene vrednosti parametra kritičnoj oblasti;
  - upoređivanje p-vrednosti α^* sa pragom značajnosti $\alpha = 1 - \beta$.

p-vredność

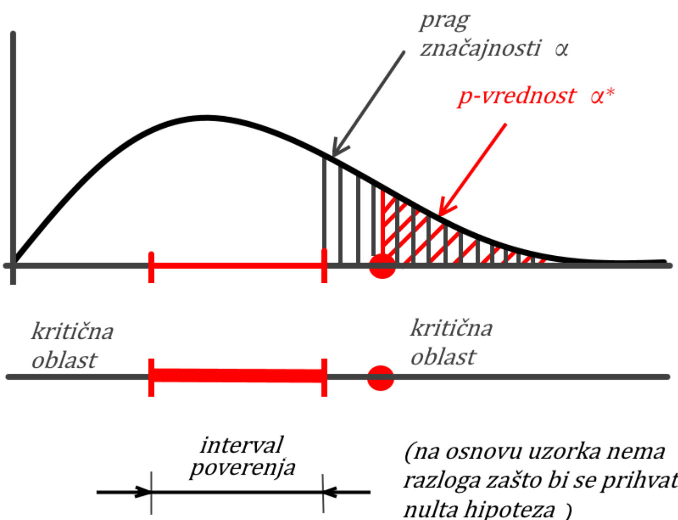
- p-vrednost α^* je verovatnoća da će se, pod pretpostavkom da je statistički model odgovarajući, test statistika uzimati vrednosti koje su "ekstremnije" od realizovane.

- Ako postoje statistički značajne razlike:
  - pretpostavljene vrednosti parametra ne pripadaju intervalu poverenja;
  - p-vrednost je manja od praga značajnosti ($\alpha^* < \alpha$);
  - kaže se da "na osnovu uzorka nema razloga zašto bi se prihvatila nulta hipoteza".
- Ako ne postoje statistički značajne razlike:
  - pretpostavljene vrednosti parametra pripadaju intervalu poverenja;
  - p-vrednost je veća od praga značajnosti ($\alpha^* > \alpha$);
  - kaže se da se "na osnovu uzorka prihvata nulta hipoteza".

p-vrednost - II



p-vrednost - III



(na osnovu uzorka nema razloga zašto bi se prihvatila nulta hipoteza)

Parametarski testovi

Test nepoznatog očekivanja m obeležja $X : N(m, \sigma)$, σ poznato

- Nulta hipoteza: $H_0(m = m_0)$;
- Alternativna hipoteza: $H_1(m \neq m_0)$;
- Ne postoji posebna komanda, ili se vrednost test statistike upoređuje sa `qnorm((1-β)/2)` ili se direktno izračuna p-vrednost:

```
xn ← mean(x)
n ← length(x)
beta ← 0.95
m0 ← 15                              #pretpostavljena vrednost za m
sigma ← 3.5                          #poznata standardna devijacija

z ← abs(xn-m0)*sqrt(n)/sigma         #vrednost test statistike
p ← 2*pnorm(-z)                      #p-vrednost, P(|Z|>z)
```

Test nepoznatog očekivanja m obeležja $X : N(m, \sigma)$, σ nepoznato

- Nulta hipoteza: $H_0(m = m_0)$;
- Alternativna hipoteza: $H_1(m \neq m_0)$;

- Zadatak 174 iz Zbirke, input:

```
xi ← c(1,2,3,4,5); fi ← c(5,9,4,5,5); x ← rep(xi,fi)
t.test(x,mu=3.5)
```

- Output:

```
t = -2.4182, df = 27, p-value = 0.02262
alternative hypothesis: true mean is not equal to 3.5
95 percent confidence interval: 2.311677 3.402608
sample estimates:
        mean of x
        2.857143
```

Test nepoznate proporcije p obeležja sa binomnom raspodelom

- Nulta hipoteza: $H_0(p = p_0)$;
- Alternativna hipoteza: $H_1(p \neq p_0)$;

- Zadatak 172 iz Zbirke, input:

```
prop.test(200,1000,p=1/6,conf.level=0.99,correct=FALSE)
```

  - prvi parametar je broj uspešnih realizacija;
  - drugi parametar je ukupan broj eksperimenata;
  - treći parametar je zadata vrednost p_0;
  - `conf.level` je nivo poverenja, β;
  - `correct=FALSE` ne umanjuje grešku zaokruživanja koja nastaje kao posledica aproksimacije diskretne slučajne promenljive neprekidnom (“Yates’ continuity correction”).

Test nepoznate proporcije p obeležja sa binomnom raspodelom - II

- Nulta hipoteza: $H_0(p = p_0)$;
- Alternativna hipoteza: $H_1(p \neq p_0)$;

- Zadatak 172 iz Zbirke, input:

```
prop.test(200,1000,p=1/6,conf.level=0.99,correct=FALSE)
```

- Output:

```
X-squared = 8, df = 1, p-value = 0.004678
alternative hypothesis: true p is not equal to 0.1666667
99 percent confidence interval: 0.1694428 0.2345119
sample estimates:
        p
        0.2
```

Test nepoznate proporcije p obeležja sa binomnom raspodelom - III

- Alternativno se koristi `binom.test`, on koristi tačne (neaproksimirane) vrednosti i vraća tačnu verovatnoću odbacivanja nulte hipoteze.

- Zadatak 172 iz Zbirke, input:

```
binom.test(200,1000,p=1/6,conf.level=0.99)
```

- Output:

```
successes = 200, trials = 1000, p-value = 0.00577
alternative hypothesis: true probability of success is
not equal to 0.1666667
99 percent confidence interval: 0.1684186 0.2344886
sample estimates:
        probability of success
        0.2
```

Test jednakosti varijanse, F-test

- Utvrđuje da li su dva uzorka dobijena uzorkovanjem normalne raspodele sa istom varijansom (nulta hipoteza je oblika $H_0(\sigma_1^2 = \sigma_2^2)$);
- Uzorci ne moraju biti istog obima, niti imati isto m.

- Input 1:

```
x1 ← c(...); x2 ← c(...)                    #vektori uzorka
var.test(x1,x2, conf.level = 0.99)
```

- Input 2:

```
Dfr ← data.frame(Brojevi=x, Faktori=y))
#može se i importovati dataframe Dfr
var.test(Brojevi~Faktori, data=Dfr)
```

Test jednakosti varijanse, F-test - II

```
        F test to compare two variances
F = 1.2, num df = 149, denom df = 299, p-value = 0.1893
alternative hypothesis: true ratio of variances not  1
99 percent confidence interval:
 0.8396564 1.7488966
ratio of variances:          1.200043
```

```
        F test to compare two variances
F = 0.85714, num df = 1, denom df = 2, p-value = 0.9046
alternative hypothesis: true ratio of variances is not 1
95 percent confidence interval:
   0.02225979 685.28571429
ratio of variances:         0.8571429
```

Neparametarski testovi

Pirsonov χ^2-test

- Nulta hipoteza: $H_0(F = F_0)$;
- Alternativna hipoteza: $H_1(F \neq F_0)$.

- Zadatak 186 iz Zbirke, input:

```
#prosleđuju se frekvencije i teorijske verovatnoće
chisq.test(c(35,35,18,12), p=c(.4375,.3125,.1875,.0625))
```

- Output:

```
Chi-squared test for given probabilities
data: c(35, 35, 18, 12)
X-squared = 7.52, df = 3, p-value = 0.05705
```

Tabele kontingencije

- Nulta hipoteza: obeležja su nezavisna;
- Alternativna hipoteza: obeležja nisu nezavisna.

- Zadatak 197 iz Zbirke, input:

```
#prosleđuje se matrica
x ← matrix(c(16, 39, 15, 30), ncol=2)   #bolje po kolonama
chisq.test(x, correct=FALSE)
```

- Output:

```
Pearson's Chi-squared test
data: x
X-squared = 0.20825, df = 1, p-value = 0.6481
```

λ-test Kolmogorov-Smirnov

- Nulta hipoteza: $H_0(F = F_0)$;
- Alternativna hipoteza: $H_1(F \neq F_0)$.

- Zadatak 185 iz Zbirke, input:

```
#prosleđuje se uzorak x
x ← c(0.0790, 0.2672, ...,  0.0108)
ks.test(x,"pexp",5)                  #navodnici su obavezni
```

- Output:

```
One-sample Kolmogorov-Smirnov test
data: c(0.079, 0.2672, ..., 0.0108)
D = 0.2172, p-value = 0.3818
alternative hypothesis: two-sided
```

Testovi normalnosti

- Nulta hipoteza: $H_0(F(x) = \Phi(\frac{x-m}{\sigma}))$;
- Alternativna hipoteza: $H_1(F(x) \neq \Phi(\frac{x-m}{\sigma}))$.

- Može Kolmogorov-Smirnov:

```
xr ← c(...)                                          #vektor uzorka
ks.test(xr, "pnorm", mean=1.5, sd=2.5)               #ne uvek 𝒩(0,1)
```

- Output:

```
        One-sample Kolmogorov-Smirnov test
data:   xr
D = 0.028608, p-value = 0.9668
alternative hypothesis: two-sided
```

Testovi normalnosti - II

- Može Anderson-Darling test (možda i najčešće korišćeni, package `nortest`);
- Može Cramér-von Mises test (ili modifikovani, package `nortest`);
- Može Jarque-Bera test (pogodno za uzorke većeg obima);

- Može Shapiro-Wilk test (obim može biti najviše 5000):

```
xr ← c(...)                              #vektor uzorka
shapiro.test(xr)
```

- Output:

```
        Shapiro-Wilk normality test
data:   xr
W = 0.99606, p-value = 0.6605
```

Domaći rad

(za radoznale i one koji očekuju iole visoku ocenu)

1. Izvesti izraz za interval poverenja nepoznate uzoračke disperzije dva uzorka, ako su nepoznate, ali jednake.
2. Izvesti izraz za interval poverenja nepoznate uzoračke disperzije dva uzorka, ako su nepoznate i nejednake.
3. Istražiti šta su to intervali predviđanja (prediction intervals).
4. Ispitati kako se intervali predviđanja mogu koristiti za detekciju outliera.
5. Istražiti šta su to intervali tolerancije (tolerance intervals).
6. Objasniti razliku između "Hypothesis Testing with Fixed Probability of Type I Error" i "Significance Testing (p-Value Approach)".
7. Istražiti šta su to testovi homogenosti.
8. Naći šta je to Jejtsova korekcija (Yates' correction for continuity) kod χ^2-testa.
9. Naći šta su vrednosti d_n i λ_α u λ-testu Kolmogorov-Smirnova.

10 Kolmogorov-Smirnov testom proveriti da li je uzorak:

0.3, 0.4, 0.3, 0.4, 0.7, 0.3, 0.1, 0.2, 0.4, 0.6, 0.7, 0.1, 0.2, 0.6, 1.0, 0.4, 0.6, 0.7, 0.3, 0.4, 0.8, 0.4, 0.7, 0.4, 0.6, 0.4, 0.3, 0.7, 0.5, 0.7, 0.2, 0.8, 0.9, 0.8, 0.5, 1.0, 1.0, 0.3

generisan slučajnom promenljivom koja ima geometrijsku $\mathcal{G}(0.3)$ raspodelu.

11 Generisati intervalni uzorak od uzorka u prethodnom zadatku koristeći komandu `table` i komandu `cut(x, breaks = seq(0, 1.1, 0.25))`.

12 Prethodni zadatak uraditi sa χ^2 testom (granice intervalnog uzorka su date sa `seq(0, 1.1, 0.25)`), ako u svakom ima najmanje pet elemenata.

13 Prethodni zadatak uraditi sa χ^2 testom ako su granice intervalnog uzorka date sa `seq(0, 1.1, 0.35)`. Da li je zaključak isti?

14 Anderson-Darling testom (package `nortest`) ispitati da li je uzorak iz zadatka 10 dobijen iz populacije sa normalnom raspodelom.

(budite
vredni!)