Testiranje hipoteza

Oznake:

- H_0 - **nulta hipoteza**
- H_1 - **alternativna hipoteza**
- T - **test statistika**
- $\hat{T}$ - **realizovana vrednost test statistike**
- W - **kritična oblast**
- $\alpha = P\{T \in W | H_0\}$ - **nivo značajnosti testa** (**verovatnoća greške prve vrste**)
- $\beta = P\{T \notin W | H_1\}$ - **verovatnoća greške druge vrste**
- $\gamma = 1 - \beta = P\{T \in W | H_1\}$ - **moć testa**
- **p-vrednost testa** - najmanji nivo značajnosti za koji prihvatamo H_0, tj. ako je $p < \alpha$ odbacujemo H_0

Testiranje u normalnom modelu

- **Test za parametar m**

1. H_0: $m = m_0$
2. H_1: $m > m_0$, $m < m_0$, $m \neq m_0$
3. T:

$$T = \sqrt{n} \cdot \frac{\overline{X_n} - m_0}{\sigma} \sim N(0,1),\ \sigma\ poznato$$

$$T = \sqrt{n} \cdot \frac{\overline{X_n} - m_0}{\tilde{S}_n} \sim t_{n-1},\ \sigma\ nepoznato$$

4. W: $W = \{T > c\}$, $W = \{T < c\}$, $W = \{T < -c\} \cup \{T > c\}$
5. p-vrednost: $p = P\{T > \hat{T} | H_0\}$, $p = P\{T < \hat{T} | H_0\}$, $p = 2\min\left(P\{T > \hat{T} | H_0\},\ P\{T < \hat{T} | H_0\}\right)$

```
z.test(x, sigma.x = ..., mu = ..., alternative = "greater/less/two.sided")
x -> uzorak
sigma.x -> standardno odstupanje
mu -> očekivanje pri H_0
```

- **Test za parametar σ**

$$T = \frac{(n-1)\tilde{S}_n^2}{\sigma_0^2} \sim X_{n-1}^2$$

```
t.test(x, mu = ..., alternative = "greater/less/two.sided")
```

Testiranje u binomnom modelu

$$T = \sqrt{n} \cdot \frac{\overline{X_n} - p_0}{\sqrt{p_0(1-p_0)}} \sim N(0,1)$$

Testiranje hipoteza u slučaju 2 uzorka

Testiranje u normalnom modelu

- **Test za parametar** m

0. Testiramo da li su disperzije jednake:
   1) H_0: $\sigma_1^2 = \sigma_2^2$
   2) H_1: $\sigma_1^2 \neq \sigma_2^2$
   3) T:

$$T = \frac{\tilde{S}_1^2}{\tilde{S}_2^2} \sim F_{n_1-1,\, n_2-1}$$

   4) W: $W = \{T < c_1\} \cup \{T > c_2\}$

```
var.test(A, B, ratio = 1, alternative = "two.sided")
```

1. H_0: $m_1 = m_2$
2. H_1: $m_1 > m_2$, $m_1 < m_2$, $m_1 \neq m_2$
3. T:

$$T = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},\ S^2 = \frac{(n_1-1)\tilde{S}_1^2 + (n_2-1)\tilde{S}_2^2}{n_1+n_2-2},\ \sigma_1 \approx \sigma_2$$

$$T = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{\tilde{S}_1^2}{n_1} + \frac{\tilde{S}_2^2}{n_2}}} \sim t_\nu,\ \nu\ dato,\ \sigma_1 \not\approx \sigma_2$$

```
t.test(A, B, mu = 0, var.equal = T/F, alternative = "greater/less/two.sided")
var.equal -> pretpostavka o jednakosti disperzija
```

- **Spareni test** - ako imamo dva zavisna uzorka istog obima spajamo ih u jedan (D = A - B) i radimo testiranje u slučaju jednog uzorka.

```
t.test(prvi, drugi, mu = 0, paired = TRUE, alternative = "greater/less/two.sided")
t.test(prvi - drugi, mu = 0, alternative = "greater/less/two.sided")
```

Testiranje u binomnom modelu

$$T = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{\overline{X_{n_1}}(1-\overline{X_{n_1}})}{n_1} + \frac{\overline{Y_{n_2}}(1-\overline{Y_{n_2}})}{n_2}}} \sim N(0,1)$$

Neparametarski testovi

- **Test znakova** (**medijane**) - kada nemamo pretpostavku o normalnoj raspodeli uzorka

1. H_0: $m_e = m_0$
2. H_1: $m_e > m_0$, $m_e < m_0$, $m_e \neq m_0$
3. T:

$$T = \sum_{i=1}^{n} I\{X_i > m_0\} \sim B(n, \frac{1}{2}) \sim N(\frac{n}{2}, \frac{n}{4}) \; ako \; n > 10$$

4. W: $W = \{T \geq c\}$, $W = \{T \leq c\}$, $W = \{T \leq c_1\} \cup \{T \geq c_2\}$

```
SIGN.test(x, md = ..., alternative = "greater/less/two.sided")
md -> medijana pod pretpostavkom H_0
```

- **Spareni test znakova** - ako imamo dva zavisna uzorka istog obima spajamo ih u jedan (D = A - B) i radimo testiranje u slučaju jednog uzorka.
- **Vilkoksonov test zbira rangova** - testiramo da li dva uzorka koji imaju istu raspodelu do na konstantu, zapravo imaju istu raspodelu. Odnosno, da li je $c = 0$ za $X = Y + c$.

1. H_0: $c = 0$
2. H_1: $c > 0$, $c < 0$, $c \neq 0$
3. T:

$$T = \sum_{i=1}^{r} r_i \sim N(\frac{n(n+m+1)}{2}, \frac{nm(n+m+1)}{12}) \; ako \; n, \; m \geq 10$$

   Vrednost r_i predstavlja rang elementa X_i u sortiranom uzorku $X_1, \ldots, X_n, Y_1, \ldots, Y_m$. Ako k elemenata ima istu vrednost sabiramo rangove koje bi imali i delimo sa brojem k i taj rang dodeljujemo svakom od njih.
4. W: $W = \{T \geq d\}$, $W = \{T \leq d\}$, $W = \{T \leq -d\} \cup \{T \geq d\}$

```
wilcox.test(x, y, alternative = "greater/less/two.sided")
W koje se dobija je T.hat - n(n + 1)/2
```

Testovi saglasnosti sa raspodelom

- **Kolmogorov-Smirnov test** - za neprekidne raspodele gde imamo mali uzorak.

1. H_0: $F = F_0$
2. H_1: $F \neq F_0$
3. D_n:

$$D_n = \sup_{x \in R} |(F_0(x) - F_n(x))|,\ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leq x\}$$

   Pri računanju supremuma računamo razlike samo u tačkama uzorka, kao i njihovim levim krajevima.
4. W: $W = \{D_n > c\}$, gde se c računa na osnovu tablice u zavisnosti od n i α

```
ks.test(x, "pexp", 0.2)
drugi argument je raspodela, a ostali argumenti parametri raspodele, ovde E(0.2)
```

- **Kolmogorov-Smirnov test za dva uzorka** - da li su raspodele dva uzorka iste.

```
ks.test(x, y, alternative = "greater/less/two.sided")
```

- X^2 **test saglasnosti** - i za neprekidne i za diskretne raspodele.

1. H_0: $F = F_0$
2. H_1: $F \neq F_0$
3. T:

$$T = \sum_{k=1}^{r} \frac{(M_k - np_k)^2}{np_k} \sim X^2_{r-1}$$

   Vrednost r predstavlja broj kategorija uzorka, M_k broj elemenata uzorka u k-toj kategoriji, a p_k verovatnoću da slučajno izabrani element uzorka upadne u k-tu kategoriju pri H_0.
4. W: $W = \{T > c\}$
5. Ako F_0 zavisi od nepoznatih parametara treba ih oceniti i tada važi $T \sim X^2_{r-q-1}$, gde je q broj ocenjenih parametara.
6. Ako kategorije ne pokrivaju sve vrednosti obeležja X dodati kategoriju koja obuhvata sve preostale vrednosti.
7. Za sve kategorije mora da važi $np_k \geq 5$, a ako ne važi spajati kategorije dok se to ne ispuni.

```
chisq.test(M, p, correct = FALSE)
M -> broj elemenata po kategorijama, tj. vektor koji sadrži sve M_k
p -> vrv da element uzorka upadne u svaku od kategorija, tj. vektor koji sadrži sve p_k
```

- X^2 **test nezavisnosti** - za diskretne raspodele X i Y koje imaju redom k i l kategorija.

1. H_0: X i Y su nezavisna obeležja
2. H_1: X i Y nisu nezavisna obeležja
3. T:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(M_{ij} - np_{i.}p_{.j})^2}{np_{i.}p_{.j}} \sim X^2_{(k-1)(l-1)}$$

   Vrednost M_{ij} predstavlja broj elemenata uzorka u i-toj kategoriji obeležja X i j-toj kategoriji obeležja Y. Vrednost $p_{i.}$ predstavlja verovatnoću da slučajno izabrani element uzorka upadne u i-tu kategoriju obeležja X, a $p_{.j}$ predstavlja verovatnoću da slučajno izabrani element uzorka upadne u j-tu kategoriju obeležja Y.
4. W: $W = \{T > c\}$

Linearna regresija

Prosta linearna regresija podrazumeva da kroz skup tačaka želimo provući pravu koja ih najbolje opisuje. Model:

$$y_i = \beta_0 + \beta_1 x_i + \mathcal{E}_i$$

```
model = lm(y ~ x)
summary(model)
```

Osobine **greške** $\mathcal{E}$:

- $E(\mathcal{E}_i) = 0$
- nekolinearnost: $E(\mathcal{E}_i \mathcal{E}_j) = 0,\ i \neq j$
- uniformno raspodeljena disperzija: $D(\mathcal{E}_i) = \sigma^2,\ \forall i$
- $\mathcal{E}_i$ je nezavisno od x_i
- $\mathcal{E}_i \sim N(0,\ \sigma^2)$

Crtanje tačaka i **regresione prave**:

```
plot(x, y) # plot(model)
abline(beta0, beta1, col = 'red') # abline(model, col = 'red')
```

Predviđanje vrednosti za date vrednosti x:

```
predict(model, newdata = data.frame(c(...)))
```

Interval poverenja za β_1 računa se pomoću test statistike:

$$\frac{\hat{\beta_1} - \beta_1}{sd(\hat{\beta_1})} \sim t_{n-2},\ P\{|T| \leq c\} = 0.95$$

Vrednost $sd(\hat{\beta_1})$ nalazi se u drugoj koloni. Analogno za parametar β_0. 95% interval poverenja za koeficijente modela:

```
ip <- confint(model)
abline(ip[1, 1], ip[2, 1], col = 'cyan') # donja granica
abline(ip[1, 2], ip[2, 2], col = 'green') # gornja granica
```

Test za značajnost parametra β_1:

- H_0: $\beta_1 = 0$, tj. β_1 nije značajan parametar
- H_1: $\beta_1 \neq 0$, tj. β_1 je značajan parametar
- T:

$$T = \frac{\hat{\beta_1}}{sd(\hat{\beta_1})} \sim t_{n-2}$$

- Za male **p-vrednosti testa** odbacujemo nultu hipotezu, tj. parametar je značajan. P-vrednosti se nalaze u četvrtoj koloni. Analogno za parametar β_0.

Reziduali modela:

```
reziduali <- model$residuals
hist(reziduali, probability = T, main = "", ylab = "")
qqnorm(reziduali)
```

Koeficijent determinacije R^2 ukazuje na tačnost modela. Važi:

$$SSE = \sum_{i=1}^{n} \mathcal{E}_i^2,\ SSR = \sum_{i=1}^{n} (\hat{y_i} - \overline{y_n})^2,\ SSTO = SSE + SSR = \sum_{i=1}^{n} (y_i - \overline{y_n})^2,\ R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Nalazi se u polju *Multiple R-squared*. Uzima vrednosti iz intervala [0, 1]. Želimo da bude što bliže jedinici, ali ako je previše blizu onda se model preprilagodio.

ANOVA test se koristi za poređenje 2 modela. Nulta hipoteza glasi: složeniji model ne doprinosi kvalitetu modela.

Linearni model sa više prediktora:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \mathcal{E}_i$$

- Test statistika za interval poverenja:

$$\frac{\hat{\beta_i} - \beta_i}{sd(\hat{\beta_i})} \sim t_{n-p-1},\ p\ je\ broj\ prediktora$$

- Test statistika za značajnost parametra:

$$T = \frac{\hat{\beta_i}}{sd(\hat{\beta_i})} \sim t_{n-p-1}$$

Kategorički predikatori su predikatori koji mogu da se svrstaju u kategorije. Umesto da delimo uzorak na dva dela prema toj promenljivoj, bolje je koristiti kategoričke predikatore jer onda nema gubitka informacija. Na primer, ako je kategorička promenljiva $smoke$ i model vrati $-200 \cdot smokeyes$, to znači da će slobodni član biti za 200 manji kod ljudi koji puše.

Baze

```
head(baza) # prvih 6 kolona baze
names(baza) # nazivi promenljivih u bazi
```

```r
# Linearni model nad promenljivama u bazi (p1 = beta0 + beta1 * p2)
model = lm(p1 ~ p2, data = baza)

# Model je linearan po koeficijentima, ne funkciji od predikatora
model = lm(p1 ~ log(p2), baza)

# Više predikatora
model = lm(p1 ~ p2+p3, baza)

# Prediktori su sve promenljive iz baze
model = lm(p1 ~ ., baza)

# Predikatori su sve promenljive sem p2 i p3
model = lm(p1 ~ .-p2-p3, baza)

# Izmena postojećeg modela
model1 = update(model, ~.-p4)

# Operator *
model = lm(p1 ~ p2*p3, baza) # uključuje predikatore p2, p3 i p2*p3
model = lm(p1 ~ I(p2*p3), baza) # uključuje predikator p2*p3
model = lm(p1 ~ p2+I(p2^2), baza) # uključuje predikatore p2 i p2^2
```

Logistička regresija

Kod **logističke regresije** promenljiva koju modeliramo je kategoričkog tipa. Verovatnoća da upadne u jednu kategoriju je p, a u drugu $1-p$. Ideja je da se ta verovatnoća transformiše u R što se može modelirati linearnom regresijom. **Model proste binarne logističke regresije**:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i$$

Formula sa leve strane naziva se **logit transformacija**. Odavde dobijamo model za p:

$$p = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_i)}}$$

Podatke klasifikujemo na osnovu **praga klasifikacije c**:

$$y_i = \begin{cases} 1, \ p_i > c \\ 0, \ inače \end{cases}$$

```r
model <- glm(y ~ x, family = binomial)
# Model se pravi na osnovu uzorka, umesto na osnovu cele baze
# Vrednosti možemo predviđati na preostalom delu baze
model <- glm(y ~ x, data = baza, family = binomial, subset = uzorak)
summary(model)
preds <- predict(model, newdata = data.frame(c(...)), type = "response") # vraća verovatnoće, ne odgovor 0/1
```

Matrica konfuzije:

y \ $\hat{y}$	0	1
0	TN	FP
1	FN	TP

Mere preciznosti:

- **tačnost** (**accuracy**): $\frac{TN+TP}{TN+FP+FN+TF}$, ova ocena nije dobra ako kategorije nisu izbalansirane po kardinalnosti.
- **senzitivnost/odziv** (**recall/true positive rate**): $\frac{TP}{FN+TP}$
- **specifičnost** (**specificity/true negative rate**): $\frac{TN}{FP+TN}$
- **preciznost**: $\frac{TP}{FP+TP}$

```
fbeta_score(y, preds)
recall(y, preds)
```

Test značajnosti parametra je isti kao u linearnoj regresiji, a koristi se test statistika:

$$T = \frac{\hat{\beta}_i}{\sqrt{\hat{D}(\hat{\beta}_i)}} \sim N(0,\ 1)$$

Koeficijent determinacije jednak je:

$$R^2 = 1 - \frac{D}{D_0} = 1 - \frac{\log L(y,\ \hat{\beta})}{\log L(y,\ \hat{\beta}_0)}$$

Vrednost D predstavlja **devijaciju** modela i nalazi se u polju *Residual deviance*, a D_0 devijaciju modela koji ima samo slobodan član i nalazi se u polju *Null deviance*.

AIC test se koristi za poređenje 2 modela. Bolji je model onaj koji ima manji AIC.

ROC kriva je kriva čije su x vrednosti specifičnost, a y-osa senzitivnost. **AUC** je površina ispod krive i želimo da ona bude što veća, odnosno da obe mere budu što bliže jedinici. Prag za klasifikaciju se bira tako da bude što bliži tački (1, 1).

Raspodele

Diskretne raspodele

1. **Bernulijeva raspodela** (**indikator**) X ~ $Ber(p)$

$$EX = p,\ DX = p(1-p),\ P\{X=k\} = p^k(1-p)^{k-1},\ k \in \{0,1\}$$

2. **Binomna raspodela** X ~ $B(n,p)$

$$EX = np,\ DX = np(1-p),\ P\{X=k\} = \binom{n}{k} p^k(1-p)^{n-k}$$

3. **Geometrijska raspodela** X ~ $G(p)$

$$EX = \frac{1}{p},\ DX = p^2,\ P\{X = k\} = (1-p)^{k-1}p$$

4. **Poasonova raspodela** X ~ $P(\lambda)$

$$EX = \lambda,\ DX = \lambda,\ P\{X = k\} = \frac{\lambda^k e^{-k}}{k!},\ \lambda > 0$$

Apsolutno-neprekidne raspodele

1. **Uniformna raspodela** X ~ $U[a,b]$

$$EX = \frac{a+b}{2},\ DX = \frac{(b-a)^2}{12},\ F(x) = \begin{cases} \frac{x-a}{b-a},\ x \in [a,b] \\ 0,\ x \notin [a,b] \end{cases},\ f(x) = \begin{cases} \frac{1}{b-a},\ x \in [a,b] \\ 0,\ x \notin [a,b] \end{cases}$$

2. **Normalna (Gausova) raspodela** X ~ $N(m, \sigma^2)$

$$EX = m,\ DX = \sigma^2,\ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-m)^2}{2\sigma^2}},\ x \in R,\ m \in R,\ \sigma^2 > 0$$

3. **Eksponencijalna raspodela** X ~ $\mathcal{E}(\lambda)$

$$EX = \frac{1}{\lambda},\ DX = \frac{1}{\lambda^2},\ F(x) = \begin{cases} 1 - e^{-\lambda x},\ x \geq 0 \\ 0,\ x < 0 \end{cases},\ f(x) = \begin{cases} \lambda e^{-\lambda x},\ x \geq 0 \\ 0,\ x < 0 \end{cases},\ \lambda > 0$$

4. **Gama raspodela** X ~ $\gamma(\alpha, \beta)$

$$EX = \frac{\alpha}{\beta},\ DX = \frac{\alpha}{\beta^2},\ f(x) = \frac{x^{\alpha-1}e^{-\beta x}\beta^\alpha}{\Gamma(\alpha)},\ \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1),\ x > 0$$