Prilagodbe modela podacima (2)

Prirodoslovno-matematički Fakultet, Sveučilište u Zagrebu

6. studenoga 2026.

Zašto provjeravamo normalnost?

Simulacijom provjerimo mogu li odstupanja od normalnosti promijeniti učestalost pogrešnih odluka.
Generiramo dva nezavisna uzorka iz *iste* distribucije. U svakoj iteraciji na isti par primijenimo Welchov t-test i F-test. Udio odbacivanja procjenjuje stvarnu vjerojatnost pogreške prve vrste.

Zašto provjeravamo normalnost?

Simulacijom provjerimo mogu li odstupanja od normalnosti promijeniti učestalost pogrešnih odluka.
Generiramo dva nezavisna uzorka iz *iste* distribucije. U svakoj iteraciji na isti par primijenimo Welchov t-test i F-test. Udio odbacivanja procjenjuje stvarnu vjerojatnost pogreške prve vrste.

Distribucija i veličina svake skupine	Welchov t-test	F-test
Normalna, $n = 25$	0.052	0.050
Studentova s 5 s.s., $n = 25$	0.050	0.164
Uniformna, $n = 25$	0.049	0.005
Eksponencijalna, $n = 25$	0.047	0.285
Eksponencijalna, $n = 100$	0.050	0.310

Q–Q graf

Q–Q je kratica za *quantile–quantile*. U normalnom Q–Q grafu uspoređujemo uređene podatke s kvantilima standardne normalne distribucije.
Ako su dvije distribucije jednake - njihovi kvantili će se poklapati.
Pretpostavimo da želimo provjeriti dolaze li podaci $x_1, ..., x_n$ iz $N(0, 1)$ distribucije. Prva ideja – nacrtati $(z_k, x_{(k)})$ uz $z_k = \Phi^{-1}(k/n)$. Problem?

Q–Q graf

Q–Q je kratica za *quantile–quantile*. U normalnom Q–Q grafu uspoređujemo uređene podatke s kvantilima standardne normalne distribucije.
Ako su dvije distribucije jednake - njihovi kvantili će se poklapati.
Pretpostavimo da želimo provjeriti dolaze li podaci $x_1, ..., x_n$ iz $N(0,1)$ distribucije. Prva ideja – nacrtati $(z_k, x_{(k)})$ uz $z_k = \Phi^{-1}(k/n)$. Problem?
Umjesto toga uzimamo

$$p_k = \frac{k - \frac{1}{2}}{n}, \quad z_k = \Phi^{-1}(p_k).$$

Q–Q graf

Q–Q je kratica za *quantile–quantile*. U normalnom Q–Q grafu uspoređujemo uređene podatke s kvantilima standardne normalne distribucije.
Ako su dvije distribucije jednake - njihovi kvantili će se poklapati.
Pretpostavimo da želimo provjeriti dolaze li podaci $x_1, ..., x_n$ iz $N(0,1)$ distribucije. Prva ideja – nacrtati $(z_k, x_{(k)})$ uz $z_k = \Phi^{-1}(k/n)$. Problem?
Umjesto toga uzimamo

$$p_k = \frac{k - \frac{1}{2}}{n}, \quad z_k = \Phi^{-1}(p_k).$$

Što ako želimo provjeriti dolaze li podaci iz $N(\mu, \sigma^2)$ distribucije?

Q–Q graf

Q–Q je kratica za *quantile–quantile*. U normalnom Q–Q grafu uspoređujemo uređene podatke s kvantilima standardne normalne distribucije.
Ako su dvije distribucije jednake - njihovi kvantili će se poklapati.
Pretpostavimo da želimo provjeriti dolaze li podaci $x_1, ..., x_n$ iz $N(0,1)$ distribucije. Prva ideja – nacrtati $(z_k, x_{(k)})$ uz $z_k = \Phi^{-1}(k/n)$. Problem?
Umjesto toga uzimamo

$$p_k = \frac{k - \frac{1}{2}}{n}, \quad z_k = \Phi^{-1}(p_k).$$

Što ako želimo provjeriti dolaze li podaci iz $N(\mu, \sigma^2)$ distribucije?
Gledamo je li $x_{(k)} \approx \mu + \sigma z_k$

Normalni Q–Q graf

Zadatak

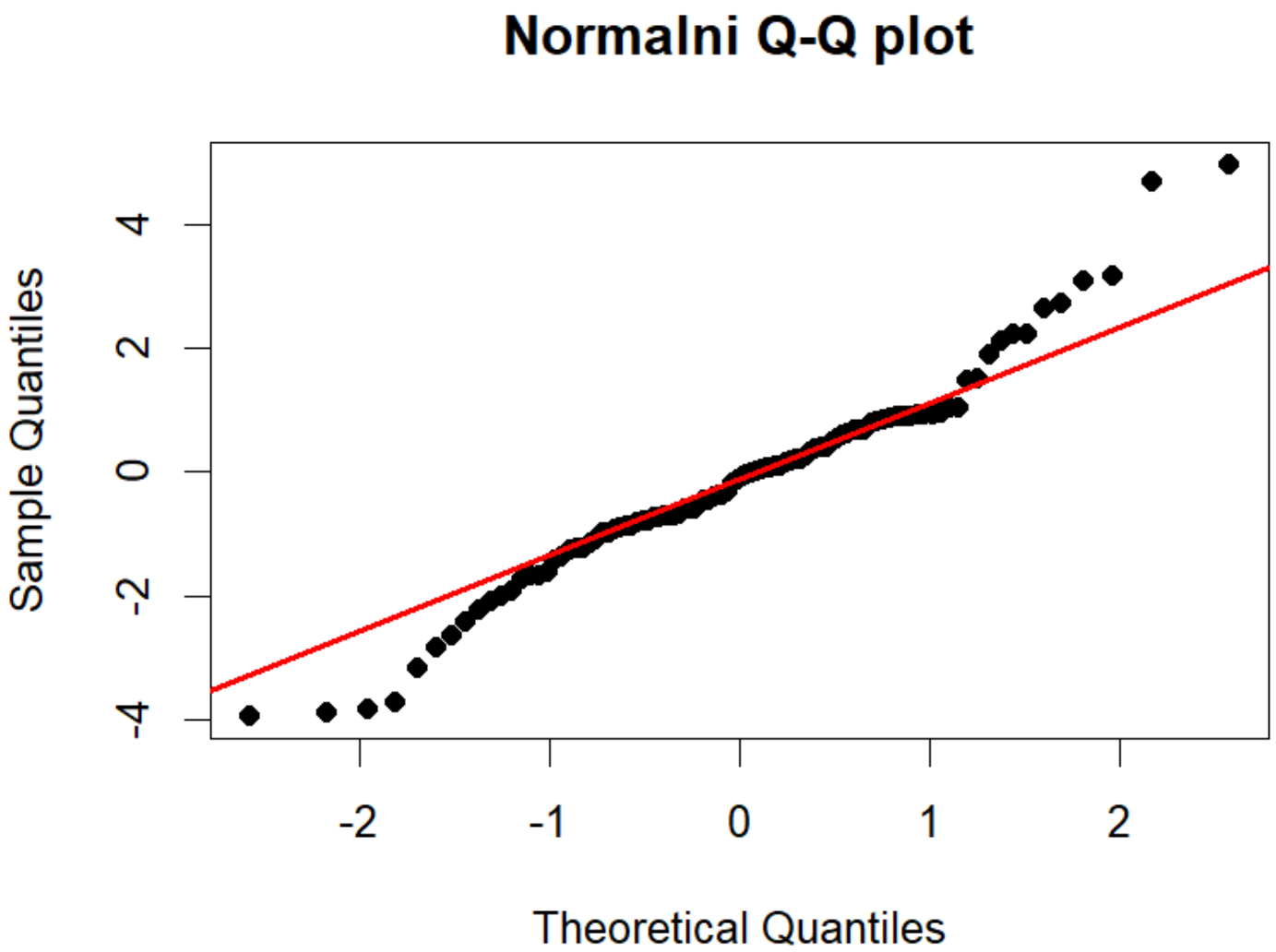
Generirajte uzorak duljine 200 iz $N(5, 2^2)$ distribucije. Nacrtajte Q–Q graf.

Rješenje.

```
set.seed(601)
x <- rnorm(200, mean = 5, sd = 2)
qqnorm(x, main = "Normalni Q-Q graf", pch = 16)
qqline(x, col = "red", lwd = 2)
```

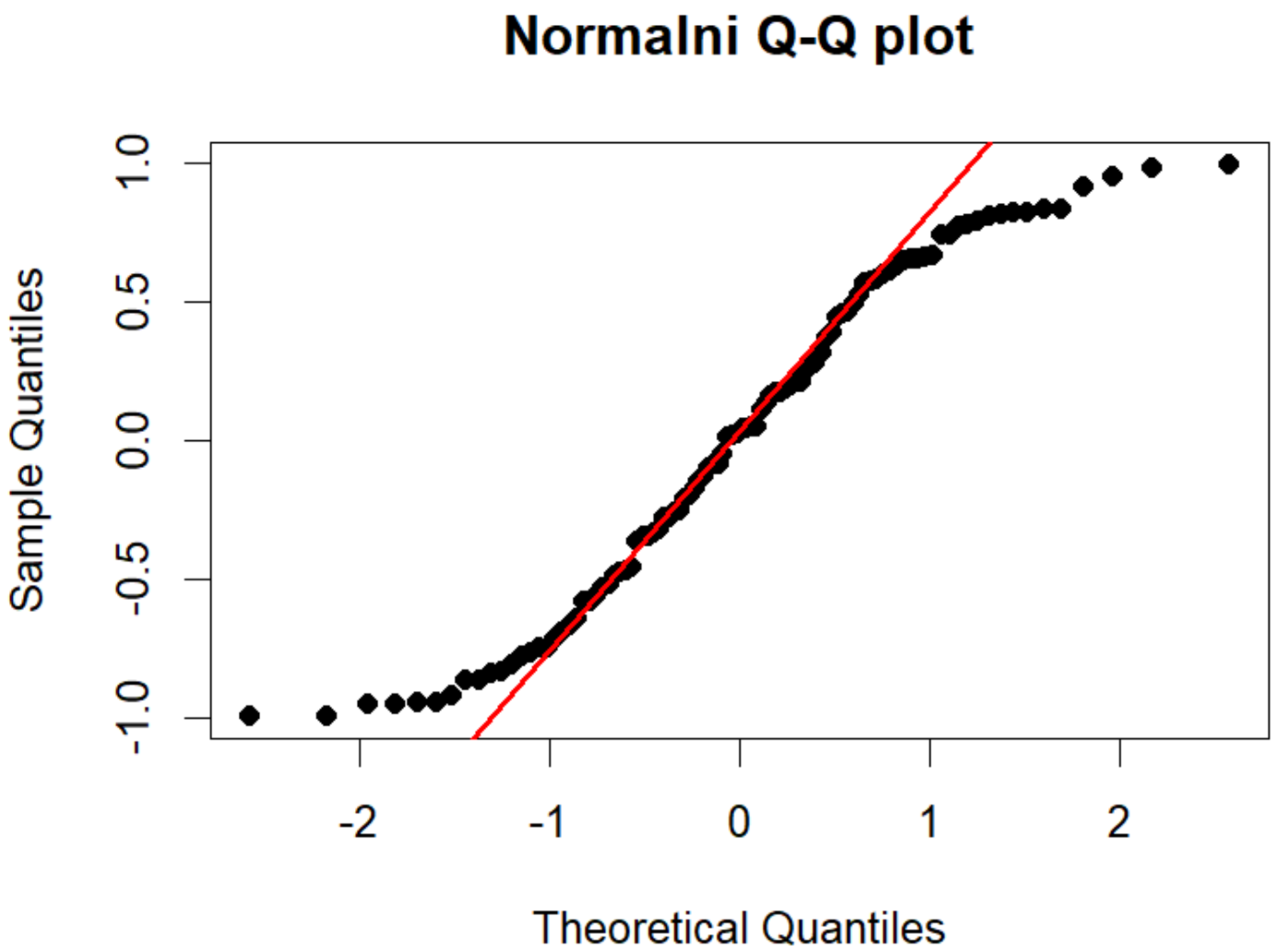


Primjer: Teški repovi



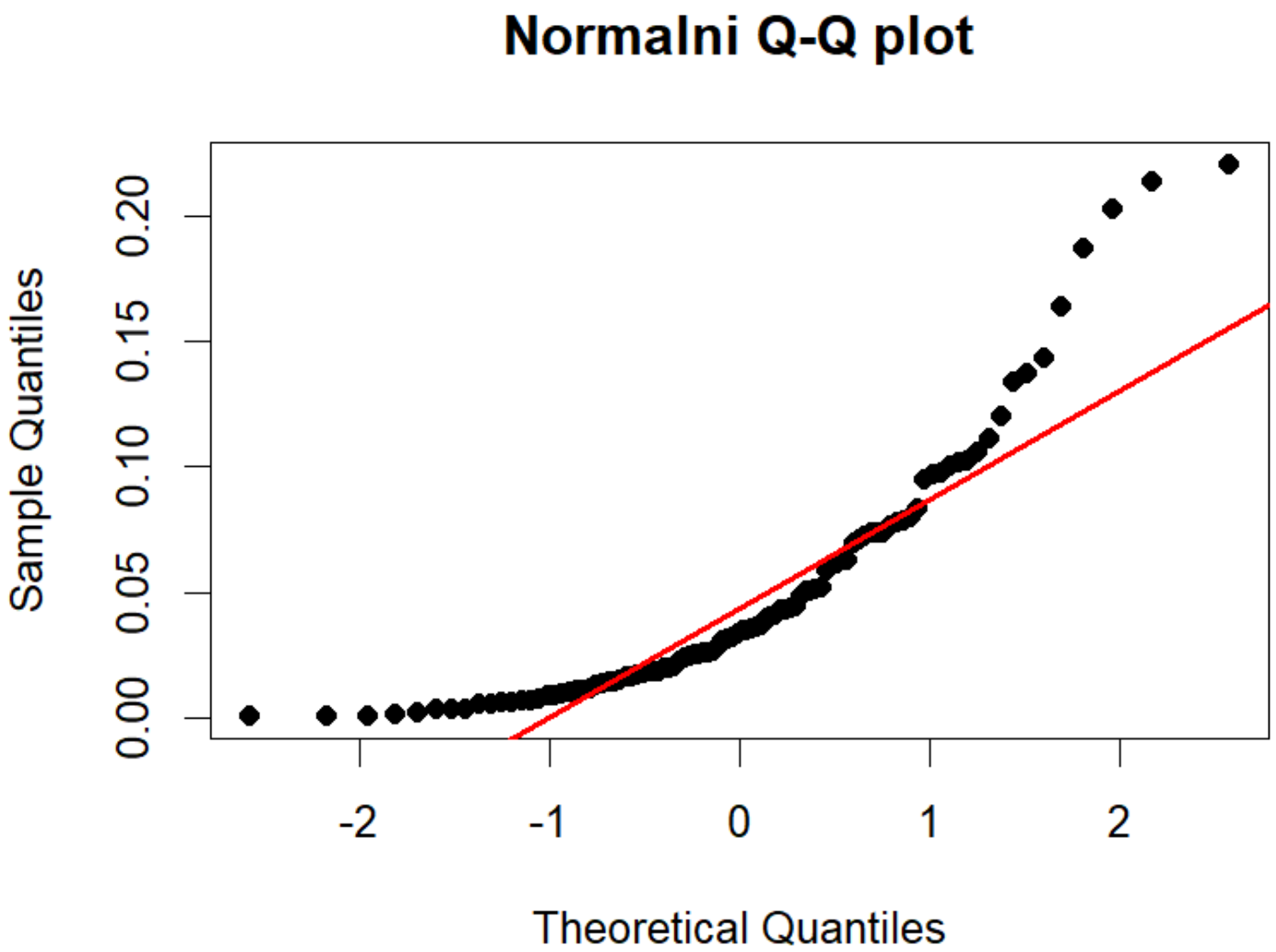
Slika: Na primjer, studentova distribucija s malim brojem stupnjeva slobode.

Primjer: Laki repovi



Slika: Na primjer, uniformna distribucija

Primjer: Desna asimetrija



Slika: Na primjer, eksponencijalna distribucija

Zadatak

Zadatak

Generirajte podatke iz $N(5, 2^2), t(3), U(-1, 1)$ i $Exp(20)$ distribucije duljine 20 i duljine 2000. Usporedite histograme i Q–Q plotove za svaku od distribucija i svaku duljinu uzoraka.

Q–Q graf za druge distribucije

Zadatak

Generirajte podatke na sljedeći način

```
set.seed(123)
exp_data <- rexp(100, rate = 2)
```

Nacrtajte Q–Q graf kojim ćete vizualno provjeriti dolaze li podaci iz $\text{Exp}(2)$ distribucije.

Q–Q graf za druge distribucije

Zadatak

Generirajte podatke na sljedeći način

```
set.seed(123)
exp_data <- rexp(100, rate = 2)
```

Nacrtajte Q–Q graf kojim ćete vizualno provjeriti dolaze li podaci iz $\text{Exp}(2)$ distribucije.

Rješenje.

```
n <- length(sample_data)
kvantili <- qexp((1:n-1/2)/n, rate = 2)
qqplot(kvantili, exp_data,
       main = "Exp Q-Q plot",
       xlab = "Theoretical Quantiles", ylab = "Sample Quantile
abline(0, 1, col = "red")
```

Empirijska funkcija distribucije

Empirijska funkcija distribucije u točki t daje udio opažanja koja nisu veća od t:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}. \tag{1}$$

Funkcija F_n stepenasta je i u svakom opažanju skače za $1/n$ ako nema ponovljenih vrijednosti. U R-u je dobivamo naredbom `ecdf`.

Empirijska funkcija distribucije

Empirijska funkcija distribucije u točki t daje udio opažanja koja nisu veća od t:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}. \tag{1}$$

Funkcija F_n stepenasta je i u svakom opažanju skače za $1/n$ ako nema ponovljenih vrijednosti. U R-u je dobivamo naredbom `ecdf`.

Primjer

```
x <- c(1.2, 1.8, 2.4, 3.1, 4.5)
Fn <- ecdf(x)
plot(Fn, main = "Empirijska funkcija distribucije")
```

Empirijska funkcija distribucije

Empirijska funkcija distribucije u točki t daje udio opažanja koja nisu veća od t:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}. \tag{1}$$

Funkcija F_n stepenasta je i u svakom opažanju skače za $1/n$ ako nema ponovljenih vrijednosti. U R-u je dobivamo naredbom `ecdf`.

Primjer

```
x <- c(1.2, 1.8, 2.4, 3.1, 4.5)
Fn <- ecdf(x)
plot(Fn, main = "Empirijska funkcija distribucije")
```

S empirijskom fjom distribucije uspoređujemo (teorijsku) fju distribucije.

Zadatak

Generirajte uzorak duljine 20 iz neke normalne distribucije te nacrtajte empirijsku funkciju distribucije. Potom generirajte to ponovite za uzorak duljine 2000.

Kolmogorov–Smirnovljev test

Za $X \sim F$ i konkretnu neprekidnu distribuciju F_0 testiramo

$$H_0 : F = F_0, \qquad H_1 : F \neq F_0. \tag{2}$$

KS-test uspoređuje najveće okomito odstupanje empirijske funkcije od zadane funkcije distribucije:

$$D_n = \sup_{t \in \mathbb{R}} |F_n(t) - F_0(t)|. \tag{3}$$

Kolmogorov–Smirnovljev test

Za $X \sim F$ i konkretnu neprekidnu distribuciju F_0 testiramo

$$H_0 : F = F_0, \qquad H_1 : F \neq F_0. \tag{2}$$

KS-test uspoređuje najveće okomito odstupanje empirijske funkcije od zadane funkcije distribucije:

$$D_n = \sup_{t \in \mathbb{R}} |F_n(t) - F_0(t)|. \tag{3}$$

Primjer

Proizvođač tvrdi da mjerenja imaju distribuciju $N(5, 2^2)$. Provjeravamo vizualno uspoređujući funkcije distribucije, a potom korištenjem naredbe `ks.test()` za provedbu KS-testa.

```
set.seed(602); xx <- rnorm(80, mean = 5, sd = 2)
plot(ecdf(xx), main = "Uzorak i unaprijed zadani model")
curve(pnorm(x, mean = 5, sd = 2), add = TRUE,
      col = "red", lwd = 2)
ks.test(xx, "pnorm", mean = 5, sd = 2)
```

KS test

Zadatak

Za uzorak iz prethodnog primjera testirajte dolazi li iz $N(0,1)$ distribucije. Prije testiranja provjerite grafički uz pomoć Q–Q plota.

KS test

Zadatak

Za uzorak iz prethodnog primjera testirajte dolazi li iz $N(0,1)$ distribucije. Prije testiranja provjerite grafički uz pomoć Q–Q plota.

Što ako želimo testirati dolaze li podaci iz **neke** normalne distribucije?

Lillieforsov test normalnosti

Lillieforsov test, u R-u `lillie.test`, prilagođava KS-pristup situaciji u kojoj očekivanje i varijancu normalnog modela procjenjujemo iz uzorka.

$$H_0 : F(t) = \Phi\left(\frac{t-\mu}{\sigma}\right) \text{ za sve } t \text{ i neke } \mu \in \mathbb{R},\ \sigma > 0,$$

$$H_1 : F \text{ nije normalna.}$$

Statistika je najveća razlika između F_n i prilagođene normalne funkcije:

$$D_{\mathrm{L}} = \sup_{t\in\mathbb{R}} \left| F_n(t) - \Phi\left(\frac{t-\overline{X}}{S}\right) \right|. \tag{4}$$

Lillieforsov test normalnosti

Lillieforsov test, u R-u `lillie.test`, prilagođava KS-pristup situaciji u kojoj očekivanje i varijancu normalnog modela procjenjujemo iz uzorka.

$$H_0 : F(t) = \Phi\left(\frac{t-\mu}{\sigma}\right) \text{ za sve } t \text{ i neke } \mu \in \mathbb{R},\ \sigma > 0,$$

$$H_1 : F \text{ nije normalna.}$$

Statistika je najveća razlika između F_n i prilagođene normalne funkcije:

$$D_{\mathrm{L}} = \sup_{t\in\mathbb{R}} \left| F_n(t) - \Phi\left(\frac{t-\overline{X}}{S}\right)\right|. \tag{4}$$

U R-u ga pozivamo uz pomoć funkcije `lillie.test()` koja je dio paketa `nortest`.

Lillieforsov test normalnosti

Primjer

Za podatke iz normalne razdiobe provjeravamo dolaze li iz neke normalne distribucije.

```
# Instalacija je potrebna samo jednom, uz internetsku vezu:
# install.packages("nortest")
set.seed(603)
x <- rexp(80, rate = 1)
qqnorm(x)
qqline(x, col = "red")
nortest::lillie.test(x)
```

Lillieforsov test normalnosti

Primjer

Za podatke iz normalne razdiobe provjeravamo dolaze li iz neke normalne distribucije.

```
# Instalacija je potrebna samo jednom, uz internetsku vezu:
# install.packages("nortest")
set.seed(603)
x <- rexp(80, rate = 1)
qqnorm(x)
qqline(x, col = "red")
nortest::lillie.test(x)
```

Kod malih uzoraka test može imati malu jakost i propustiti znatna odstupanja. Kod jako velikih uzoraka, zbog velike jakosti, može detektirati i mala, trivijalna odstupanja kao odstupanja od normalne distribucije. P-vrijednost zato nije mjera veličine odstupanja od normalnosti. Uz rezultat uvijek pogledamo graf, veličinu uzorka i cilj analize.