Statistika
(za biologe, 3.dio)

Bojan Basrak

rujan 2006

Procjena parametara modela

Nakon što smo prikupili numeričke podatke i odredili primjeren vjerojatnosni model, tipično želimo procjeniti parametre na osnovu podataka.

Primjer
Ako označimo sa X broj ženskih potomaka u 20 slučajeva prvorodjene djece kraljevskih obitelji ili broj adenina u uzorku od 1000 nukleotida odabranih slučajno sa genoma neke vrste, mogli bismo modelirati X kao binomnu sl. varijablu, s parametrom $n = 20$ odn. 1000, no s nepoznatim parametrom p. Dakako, jedan prirodan procjenitelj za p je jednostavno X/n.

Primjer
Ako označimo s X raspon krila slučajno odabrane mušice u nekoj koloniji i prepostavimo da je X normalno distribuirana sl. varijabla, postavlja se pitanje možemo li odrediti parametre ove razdiobe na osnovu prikupljenog uzorka. Kako su parametri normalne razdiobe upravo njeno očekivanje i varijanca, nije teško pretpostaviti da očekivanje i varijanca uzorka mogu poslužiti u procjeni.

Slučajni uzorak je niz sl. varijabli $X_1, \ldots, X_n$, koji zadovoljava

- sl. varijable X_i su nezavisne,
- sl. varijable X_i imaju istu razdiobu.

Sl. uzorak predstavlja numeričke podatke koje namjeravamo prikupiti tokom istraži vanja.

Procjenitelji parametara su uvijek samo funkcije slučajnog uzorka. Kako se proizvoljna funkcija uzorka zove **statistika**, možemo reći da parametre procjenjujemo preko statistika. Uočite: svaka je statistika i sama slučajna varijabla.

Primjer
Aritmetička sredina slučajnog uzorka $X_1, X_2, \ldots, X_n$ je

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Primjetite, sada opažanja modeliramo sl. varijablama, pa smo u ovoj definiciji koristili velika slova X.

Procjena parametra p binomne razdiobe

Pretpostavimo da se naš sl. uzorak sastoji od n sl. varijabli $X_1, \ldots, X_n$ koje sve imaju vrijednosti 0 ili 1, dakle X_i je 1 ako se u i-tom pokusu dogodio "uspjeh" inače je $X_i = 0$. Uz oznaku

$$S = X_1 + \cdots + X_n,$$

mi znamo da S zbog nezavisnosti i jednake distribucije sl. varijabli X_i ima binomnu razdiobu s parametrima n i p. Primjetite da je za Bernoullijeve sl. varijable X_i, parametar p ujedno njihovo očekivanje. Pitanje je kako iz uzorka procjeniti parametar p.

Primjer

i) na ispitivanju uzorka od 1000 glasača utvrdjeno je da izvjesnog kandidata za gradonačelnika podržava njih 513. Procjenite vjerojatnost da taj kandidat uživa podršku sl. izabranog glasača.

ii) u sl. uzorku od 800 nukleotida sa genoma *C. elegans* utvrdjeno je da 312 njih predstavlja baze C i G. Koliki je postotak ovih nukelotida u genomu ove vrste?

iii) na financijskim tržištima uzorak od 200 uzastopnih dana pokazuje da su dionice kompanije ZXY porasle u 143 dana. Kolika je vjerojatnost da vrijednost ove dionice poraste tokom sl. odabranog dana?

iv) U 600 bacanja novčića od 1 eura pismo je palo u 289 slučajeva, kolika je vjerojatnost da padne pismo nakon pojedinog bacanja?

Napomena Pretpostavka da su X_i nezavisne i jednako distribuirane u ovim slučajevima bi trebalo u praksi i opravdati.

U svim gornjim primjerima, jedan prirodan procjenitelj za vjerojatnost uspjeha p je relativna frekvencija

$$\hat{p} = \frac{S}{n}.$$

Uočite da je $\hat{p}$ slučajna varijabla (prije nego nam je poznat uzorak). Možemo izračunati njeno očekivanje, ono jasno iznosi

$$E\hat{p} = \frac{1}{n}E(S) = \frac{1}{n}np = p.$$

Dakle očekivanje procjenitelja $\hat{p}$ je jednako vrijednosti koju želimo procjeniti, takve procjenitelje zovemo **nepristranim**.

Izračunajmo i varijancu od $\hat{p}$ ona je

$$\mathrm{var}\hat{p} = \frac{1}{n^2}\mathrm{var}S = \frac{1}{n^2}n\mathrm{var}X_1 = \frac{pq}{n},$$

gdje je $q = 1 - p$ kao i do sada. **Standardnu grešku** procjenitelja definiramo kao

$$\text{s.e.}(\hat{p}) = \sqrt{\mathrm{var}\hat{p}} = \sqrt{\frac{pq}{n}}.$$

Dakle s.e. $(\hat{p})$ je zapravo standardna devijacija sl. varijable $\hat{p}$. Primjetite da ona teži k 0, za $n \to \infty$, jer $\mathrm{var}\hat{p} \to 0$ za $n \to \infty$. Za nepristrane procjenitelje koji zadoljavaju ovo svojstvo kažemo da su **konzistentni**.

Kako je naš uzorak slučajan, nije jako vjerojatno da vrijedi $p = \hat{p}$, no očekujemo da vrijedi $p \approx \hat{p}$ za velike n. Dakle, procjenitelj je "blizu" prave vrijednosti no koliko blizu i što to točno znači?

Prisjetimo se de Moivre-Laplaceovog teorema koji kaže da za $0 < p < 1$ i "velike" n slučajna varijabla

$$\frac{S - np}{\sqrt{npq}} = \sqrt{\frac{n}{pq}}(\hat{p} - p)$$

ima približno N(0,1) razdiobu.

Dakle za $a < b$ vrijedi

$$P\left(a \leq \sqrt{\frac{n}{pq}}(\hat{p} - p) \leq b\right) \approx P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Ili malo drugačije zapisano

$$P\left(p + a\sqrt{\frac{pq}{n}} \leq \hat{p} \leq p + b\sqrt{\frac{pq}{n}}\right) \approx \Phi(b) - \Phi(a).$$

Koristeći da je s.e.$(\hat{p}) = \sqrt{pq/n}$ te da vrijedi $\Phi(1.96) - \Phi(-1.96) = 0.95$ (provjerite u tablicama) možemo pisati i

$$P\left(p - 1.96\text{s.e.}(\hat{p}) \leq \hat{p} \leq p + 1.96\text{s.e.}(\hat{p})\right) \approx 0.95,$$

ili uobičajenije

$$P\left(\hat{p} - 1.96\text{s.e.}(\hat{p}) \leq p \leq \hat{p} + 1.96\text{s.e.}(\hat{p})\right) \approx 0.95.$$

Dakle, vjerojatnost da p leži u intervalu

$$(\hat{p} - 1.96\text{s.e.}(\hat{p}), \hat{p} + 1.96\text{s.e.}(\hat{p}))$$

je 95%. No ovdje je ipak potreban oprez u interpretaciji.

Općenijite, definirajmo za $\alpha \in (0,1)$, broj z_α kao $1-\alpha$-kvantil razdiobe od $Z \sim N(0,1)$, tj. kao broj koji zadovoljava

$$1 - \Phi(z_\alpha) = P(Z > z_\alpha) = \alpha.$$

Gornji argumenti pokazuju da vrijedi

$$P\left(\hat{p} - z_{\alpha/2}\text{s.e.}(\hat{p}) \leq p \leq \hat{p} + z_{\alpha/2}\text{s.e.}(\hat{p})\right) \approx 1 - \alpha,$$

za "velike" n. Primjetite, na prethodnim slajdovima smo koristili: $z_{0.025} = 1.96$.

Zbog svega navedenog interval

$$\left(\hat{p} - z_{\alpha/2}\text{s.e.}(\hat{p}), \hat{p} + z_{\alpha/2}\text{s.e.}(\hat{p})\right)$$

zovemo $(1-\alpha)$%-tni **interval pouzdanosti** za parametar p.
Ipak, postoji problem u praksi, da bismo odredili gornji interval morali bismo znati standardnu pogrešku s.e., no ona sadrži nama nepoznati parametar p, zbog toga je zamijenjujemo nama dostupnom procjenom

$$\text{s.e.}(\hat{p}) \approx \sqrt{\hat{p}(1-\hat{p})/n}.$$

Tako dobijemo interval

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

koji koristimo u praksi i takodjer zovemo $(1-\alpha)$%-tni **interval pouzdanosti** za parametar p.

Ovaj interval se sužava kako raste veličina našeg uzorka n ili kada se $\hat{p}$ približava 0 ili 1. No što ako je $\hat{p} = 0$.

Primjetite da smo očekivanje sl. varijabli X_i procjenili preko aritmetičke sredine uzorka. Zbog toga se katkad govori da populacijsko očekivanje u ovom slučaju možemo procjeniti očekivanjem uzorka. To i nije neko iznenadjenje.

Nadalje i intervale pouzdanosti uobičajeno zovemo procjeniteljima za traženi parametar, no oni su takozvani intervalni procjenitelji. Intervalni procjenitelji su općenito korisniji od točkovnih procjenitelja, jer nam daju i ocjenu pogreške u našoj procjeni.

Procjena parametra μ normalne razdiobe uz poznatu varijancu

Primjer

Prepostavimo da nam je poznato da odredjeni instrument uspjeva izmjeriti udio šećera u 100g nekog proizvoda. Pri tom prilikom svakog mjerenja pravi i grešku koja je normalno distribuirana s očekivanjem 0 i standardnom devijacijom od 1.5 g. Nakon 5 mjerenja količine šećer u jednoj vrsti gumenih bombona dobiveni su sljedeći rezultati: 77.7,78.2,78.9,76.9 i 76.7 g u 100g proizvoda. Možete li procjeniti stvarni sadržaj šećera u ovom proizvodu?

Uočite, ako je greška $\varepsilon \sim N(0, \sigma^2)$ distribuirana, tada uz prepostavku da je stvarna koncentracija μ, dobiveni podaci imaju prikaz $\mu + \varepsilon$, te imaju $N(\mu, \sigma^2)$ razdiobu.

Iako smo nezavisnost definirali samo za diskretne sl. varijable, to možemo učiniti i za neprekidne sl. varijable. Tako su npr. dvije sl. varijable X i Y nezavisne, ako za sve intervale (a, b) i (c, d) u $\mathbb{R}$ vrijedi

$$P(X \in (a, b), Y \in (c, d)) = P(X \in (a, b))P(Y \in (c, d)).$$

Suma nezavisnih normalnih sl. varijabli ponovo je normalna sl. varijabla. Dakle, ako je $X \sim N(\mu_1, \sigma_1^2)$ i $Y \sim N(\mu_2, \sigma_2^2)$ tada je uz prepostavku njihove nezavisnosti

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Nadalje, ako $X_1, \ldots, X_n$ čine sl. uzorak iz normalne razdiobe s parametrima μ i σ^2 tada je njihova suma $S = X_1 + \cdots + X_n$ ponovo normalno distribuirana s očekivanjem $n\mu$ i varijancom $n\sigma^2$.

Kako je parametar μ ujedno očekivanje razdiobe od X_i i njega možemo pokušati procjeniti preko očekivanja uzorka, odn. njegove aritmetičke sredine

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

Kako je $\bar{X} = S/n$ i ovo je normalna sl. varijabla samo ima očekivanje μ i varijancu σ^2/n (pokažite). Dakle $\bar{X}$ je nepristran i konzistentan (jer mu varijanca konvergira k 0) procjenitelj parametra μ.

No i ovdje bismo željeli imati intervalni procjenitelj. A njega lako dobijemo ako nam je poznata varijanca σ^2. Naime, standardizacijom sl. varijable $\bar{X}$ slijedi

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1).$$

Posebno za sve $a < b$ vrijedi

$$P\left(a \leq \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \leq b\right) = P(a \leq Z \leq b) = \Phi(b) - \Phi(a),$$

za sl. varijablu $Z \sim N(0, 1)$. Ili

$$P\left(\mu + a\sigma/\sqrt{n} \leq \bar{X} \leq \mu + b\sigma/\sqrt{n}\right) = P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Što možemo pisati i kao

$$P\left(\bar{X} + a\sigma/\sqrt{n} \leq \mu \leq \bar{X} + b\sigma/\sqrt{n}\right) = P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Tako da je sada 95%-tni **interval pouzdanosti** za parametar μ jednostavno

$$\left(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}\right).$$

Općenitije je

$$\left(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right), \tag{1}$$

$(1-\alpha)$100%-tni **interval pouzdanosti** za parametar μ.

I ovdje se greška u procjeni smanjuje s rastom veličine uzorka, ali i s opadanjem populacijske varijance σ^2. No pretpostavka o tome da nam je varijanca poznata tipično je nerealistična u praksi, zato i nju moramo procjeniti. Kako?
Naravno, razumno je pokušati **varijancom uzorka**

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Varijanca uzorka ovako definirana nepristran je i konzistentan procjenitelj za populacijsku varijancu. Primjetite da smo ponovo podatke označili velikim slovima jer su prije pokusa i oni za nas sl. varijable.

Procjena parametra μ normalne razdiobe uz nepoznatu varijancu

Neka $X_1, \ldots, X_n$, $n \geq 1$, čine sl. uzorak iz normalne razdiobe s parametrima μ i σ^2, te neka su nam oba parametra nepoznata. Kako smo vidjeli tada njihova aritmetička sredina ima $N(\mu, \sigma^2/n)$ razdiobu. Da bismo konstruirali interval povjerenja za μ, σ^2 možemo procijeniti s $\hat{s}^2$. No, iako

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

ima standardnu normalnu razdiobu to ne vrijedi za sl. varijablu

$$\sqrt{n}\frac{\bar{X} - \mu}{\hat{s}}.$$

Za ovu sl. varijablu prvi je razdiobu izračunao William Gosset, i objavio je 1908. u časopisu "Biometrika" pod pseudonimom Student. Gosset je inače radio za pivovaru "Guiness", koja je od njega zahtjevala da koristi pseudonim.
Danas ovu razdiobu zovemo **Studentova t razdioba**. Ona pripada klasi važnih neprekidnih razdioba, tako da i za nju tabeliramo funkciju distribucije odn. njene kvantile.

Posebno, sl. varijabla

$$\sqrt{n}\frac{\bar{X}-\mu}{\hat{s}}.$$

ima Studentovu t razdiobu s $n-1$ **stupnjem slobode**.

Broj stupnjeva slobode $n-1$ uvijek je cijeli i pozitivan, on predstavlja parametar t razdiobe. Ova razdioba ima gustoću simetričnu oko 0, pa joj je i očekivanje 0.

Napomenimo da je za velike n (npr. $n > 100$) Studentova t razdioba gotovo istovjetna sa standardnom normalnom razdiobom.

U tablicama nalazimo brojeve t_α, koji nam daju $(1-\alpha)$-kvantil razdiobe sl. varijable T koja ima Studentovu t razdiobu s d stupnjeva slobode. Dakle t_α zadovoljava

$$P(T > t_\alpha) = \alpha.$$

Pri tom, tipično različiti reci tablice odgovaraju različitim stupnjevima slobode d.

Sada možemo reći da je

$$\left(\bar{X} - t_{\alpha/2}\hat{s}/\sqrt{n}, \bar{X} + t_{\alpha/2}\hat{s}/\sqrt{n}\right), \tag{2}$$

$(1-\alpha)100\%$-tni **interval pouzdanosti** za parametar μ u ovom slučaju.

Ovaj interval možemo pisati i u obliku

$$\left(\bar{X} - t_{\alpha/2}\text{s.e.}(\bar{X}), \bar{X} + t_{\alpha/2}\text{s.e.}(\bar{X})\right),$$

gdje je s.e.$(\bar{X})$ standardna greška procjenitelja $\bar{X}$ tj. naša procjena za nju

$$\text{s.e.}(\bar{X}) = \hat{s}/\sqrt{n}$$

Važna napomena

Iako su intervalni procjenitelji parametra μ izvedeni pod pretpostavkom da sami podaci dolaze iz normalne razdiobe, oni predstavljaju razuman izbor i koriste se u praksi čak i u onim slučajevima kada podaci slijede neku drugu razdiobu. Tada su ovi intervali zapravo intervalni procjenitelji očekivanja razdiobe iz koje dolaze podaci.

Ako je uzorak dovoljno velik (npr. $n \geq 30$ ili bolje $n \geq 100$), tada centralni granični teorem omogućuje da procjenimo interval pouzdanosti za μ preko normalne razdiobe dakle formulom (1). Dakako, u toj formuli poznatu standardnu devijaciju σ zamjenjujemo procjenom tj. sa $\hat{s}$.

Za manje uzorke taj teorem ne možemo koristiti, no ako je gustoća podataka u uzorku približno istog "zvonastog" oblika kao i normalna, u praksi možemo i dalje koristiti intervale dobivene iz formule (2), uz ogradu da su ovako dobiveni intervali zapravo aproksimativni.

Usporedba podataka s normalnom razdiobom

U praksi je ipak važno znati usporediti razdiobu prikupljenih podataka s normalnom, kako bismo znali da li su korištene statističke metode primjenjive na naše podatke. Usporediti ih možemo npr. koristeći deskriptivne statističke metode koje smo već spominjali.

Jedna je mogućnost da nacrtamo histogram podataka s funkcijom gustoće normalne razdiobe s procjenjenim vrijednostima očekivanja i varijance.

Druga metoda, koja je ujedno ilustrativnija u praksi, je metoda koja usporedjuje kvantile u slučajnom uzorku s kvantilima standardne normalne radzdiobe.

Ako uredimo sl. uzorak $X_1, X_2, \ldots, X_n$ po veličini dobit ćemo uzlazni niz

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

Za bilo koji $0 < \alpha < 1$ možemo definirati α-**kvantil** sl. uzorka $X_1, X_2, \ldots, X_n$, kao

$$q_\alpha = X_{(\alpha(n+1))}\,.$$

gdje pretpostavljamo da je $s = \alpha(n+1)$ realan broj izmedju 1 i n.

Pritom kao i prije, broj s izmedju 1 i n, prikažemo u obliku

$$s = k + r$$

tako da je $k = 1, 2, \ldots, n$ njegov cijeli, a $0 \leq r < 1$ njegov razlomljeni dio, te definiramo

$$X_{(s)} = (1 - r)X_{(k)} + rX_{(k+1)}\,.$$

Kvantili neprekidne sl. varijable X ili njene razdiobe se definiraju preko funkcije distribucije F od X, tako da za $\alpha \in (0,1)$, α-kvantil od X bude broj x_α takav da vrijedi

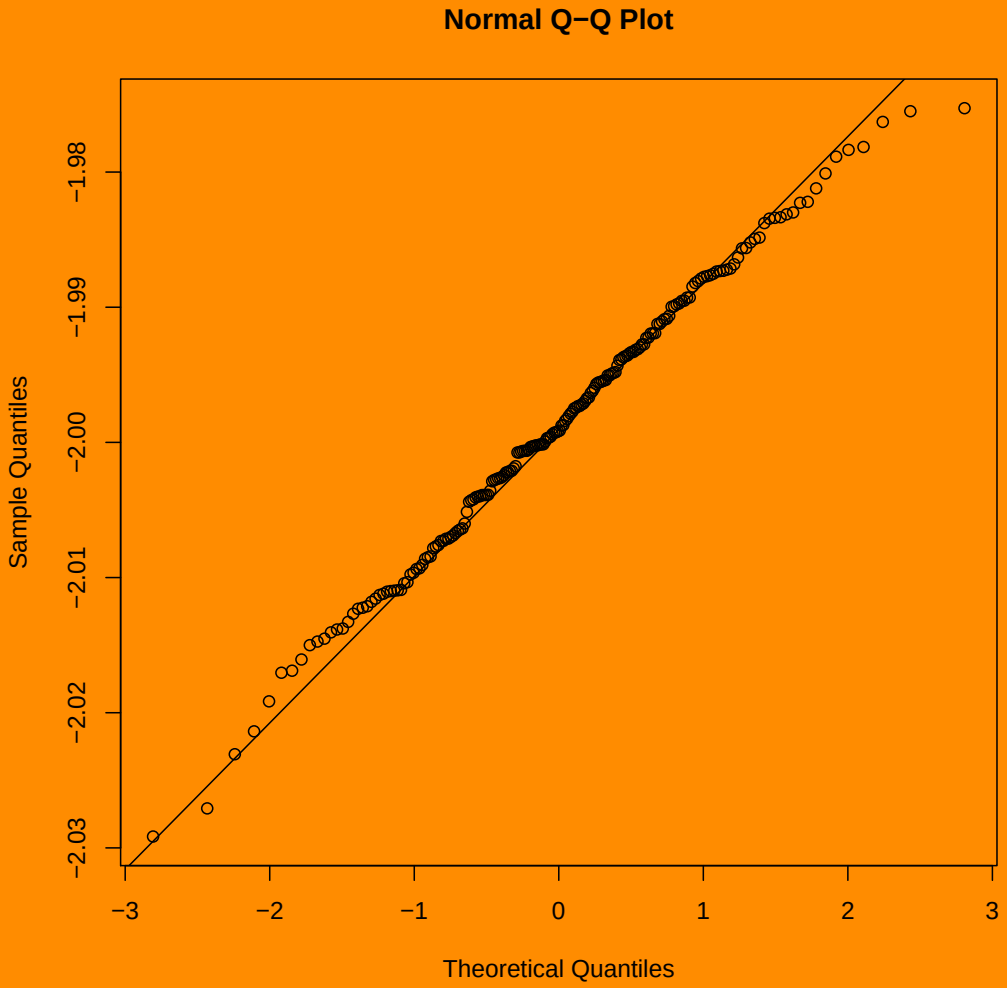
$$P(X \leq x_\alpha) = F(x_\alpha) = \alpha.$$

Ako dvije neprekidne sl. varijable imaju istu funkciju gustoće (pa i istu funkciju distribucije) tada su im i kvantili jednaki. Ova činjenica sugerira da bismo mogli usporediti kvantile sl. uzorka s kvantilima normalne razdiobe kako bi napravili još jednu usporedbu.

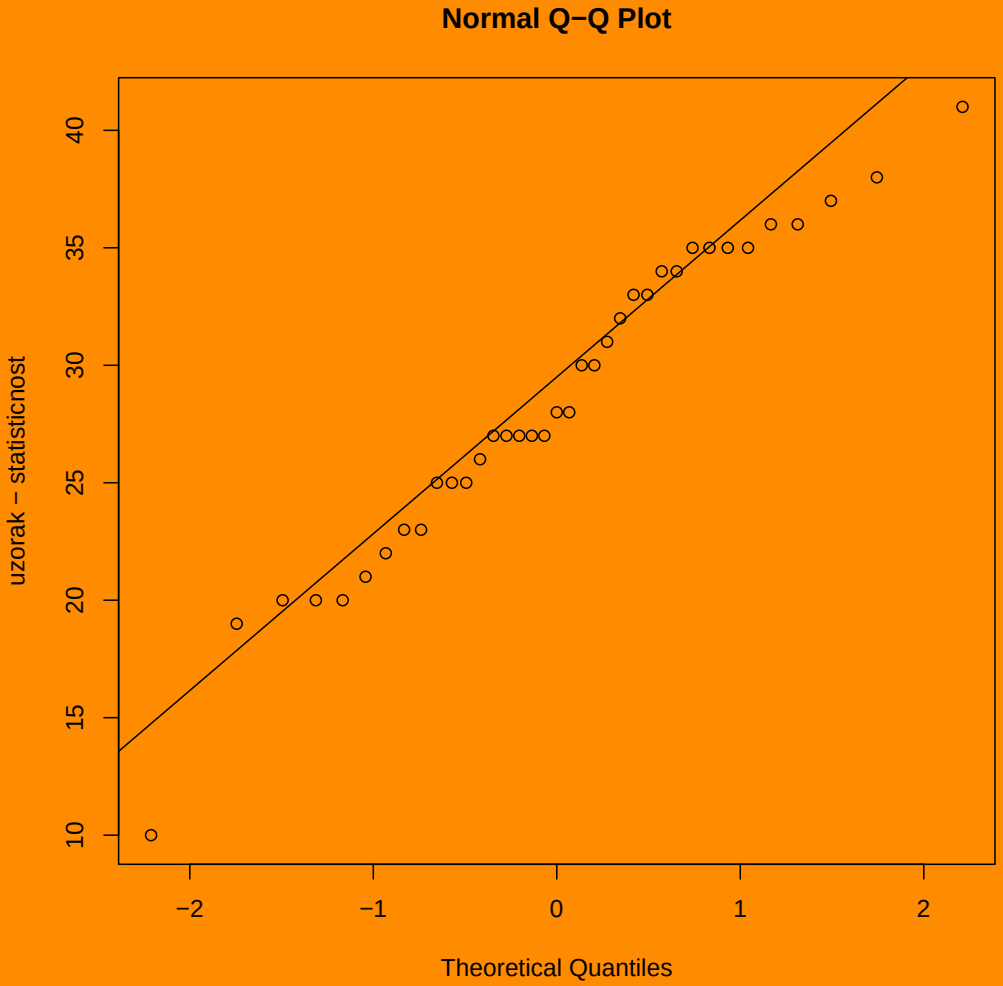
Uzmimo da je $\alpha_i = i/(n+1)$, $i = 1, \ldots, n$. Za ovakve α_i je α_i-kvantil našeg uzorka upravo $X_{(i)}$, pa ako je x_{α_i} α_i-kvantil normalne razdiobe iz koje dolazi uzorka očekivali bismo

$$X_{(i)} \approx x_{\alpha_i}.$$

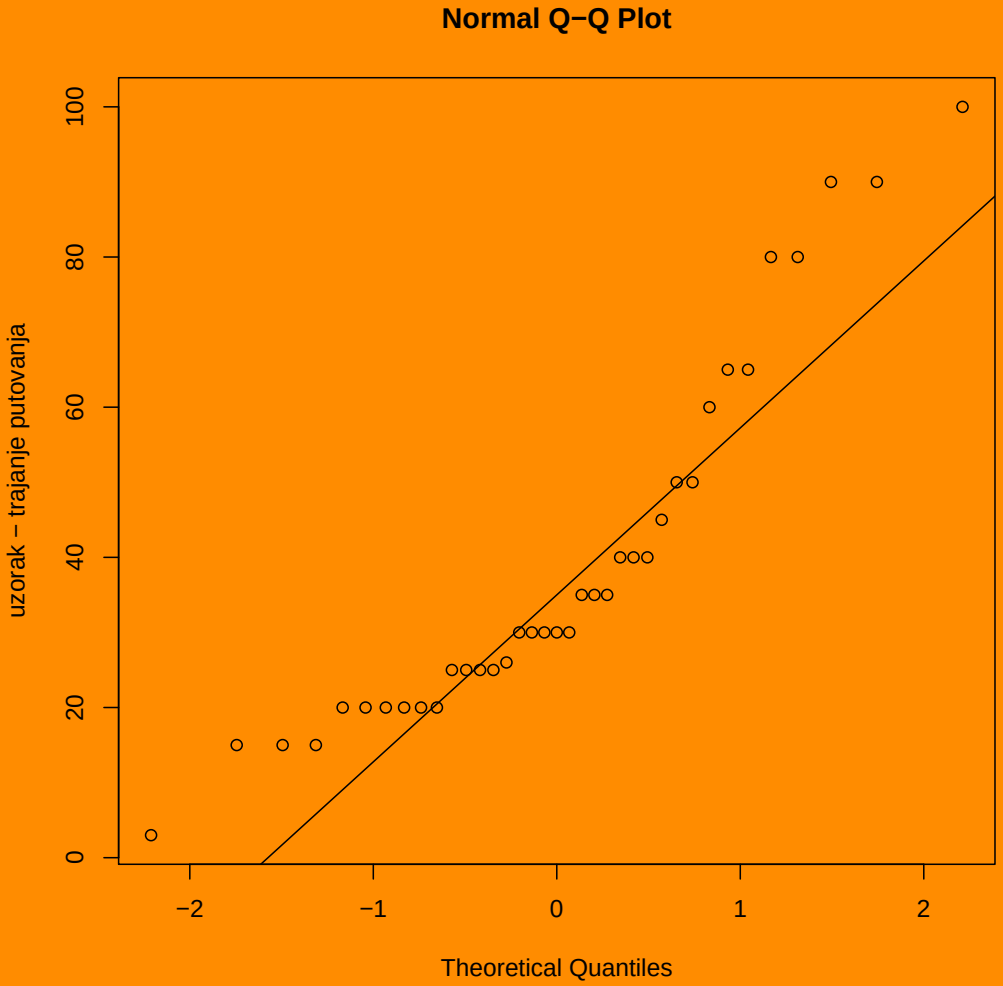
Koliko su blizu kvantili uzorka i teorijske normalne razdiobe najbolje prikazuje *graf kvantila* ili qq-*plot* u odn. na normalnu razdiobu.



Graf kvantila za sl. uzorak **simuliran** iz $N(-2, 0.0001)$ razdiobe.



Graf kvantila za uzorak varijable "statisticnost".



Graf kvantila za uzorak varijable "trajanje putovanja".

Rezultati 1. kolokvija, deskriptivne statistike

Lako se nadju:

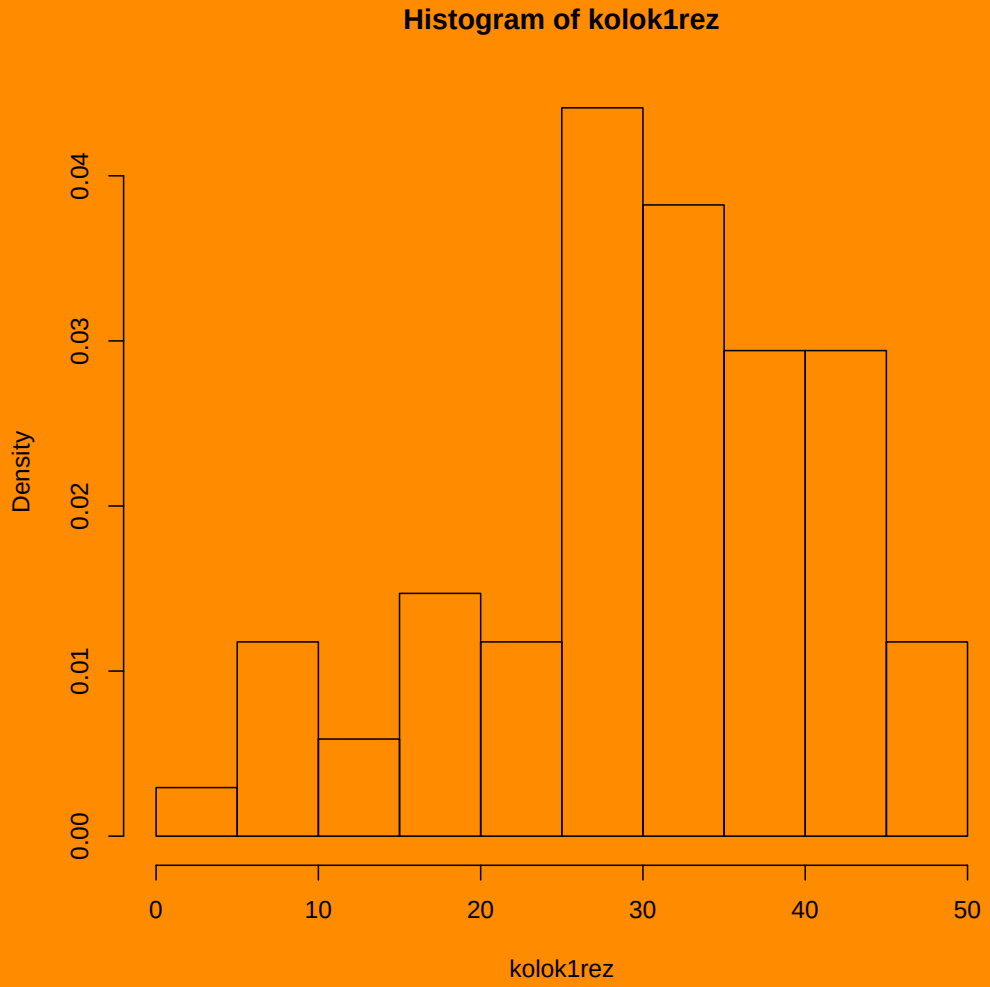
medijan: $m = 32.50$,

aritm. sredina: $\bar{x} = 30.75$,

donji kvartil: $Q_1 = 26.00$,

gornji kvartil: $Q_3 = 39.00$,

stand. devijacija: $s = 10.81$.



Histogram za rezultate kolokvija.

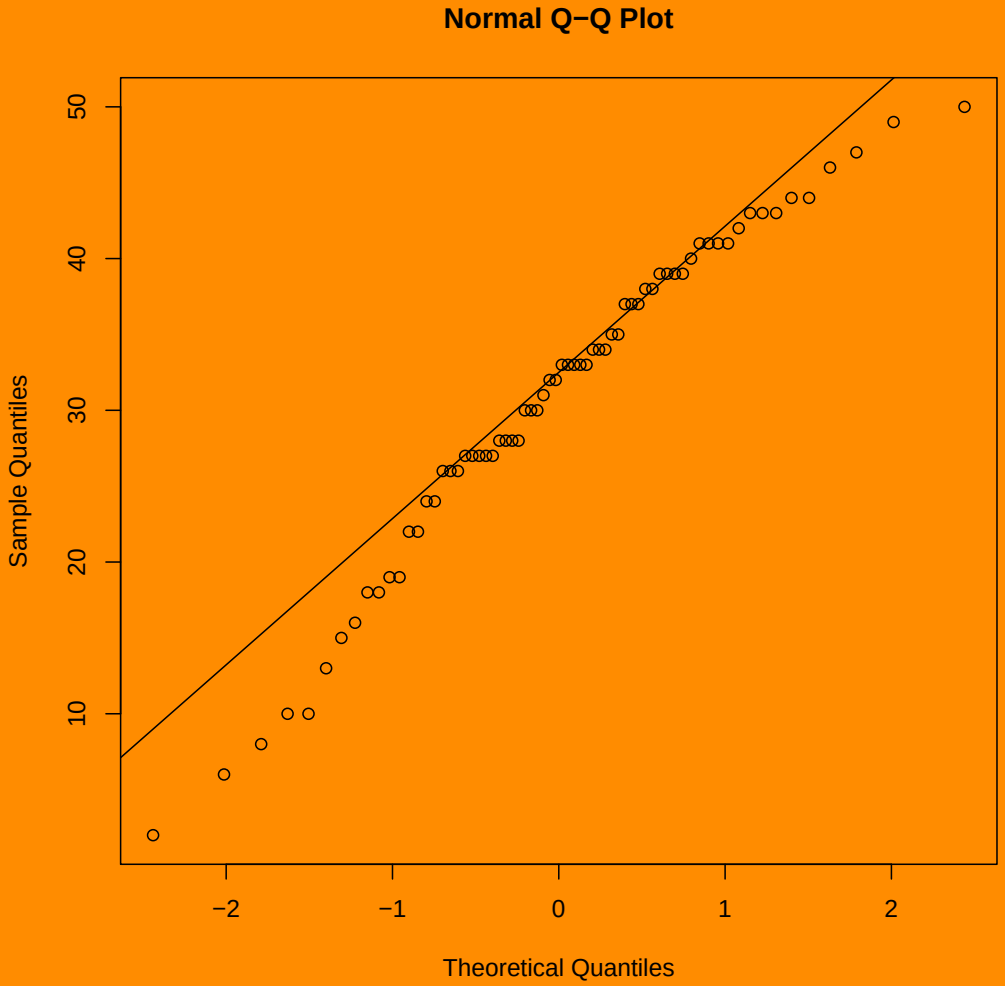
Rezultati 1. kolokvija, interval pouzdanosti za očekivanje

Pitanja:

- da li je razdioba rezultata normalno distribuirana?

- da li je $n = 68$ dovoljno velik za normalnu aproksimaciju?

- ima li smisla ovo smatrati uzorkom, ako smo ispitali cijelu populaciju zapravo?



Graf kvantila za rezultate kolokvija.

Uz prepostavku da ovo možemo smatrati uzorkom, izračunamo 95%-tni interval pouzdanosti kako smo istaknuli u važnoj napomeni, tj. kao

$$\left(\bar{x} - 1.96s/\sqrt{n}, \bar{x} + 1.96s/\sqrt{n}\right).$$

Tako dobijemo:

$$(28.13185,\ 33.36815)$$