

# Metoda bootstrapa u statističkom zaključivanju

---

Vajdić, Ivan

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: <https://um.nsk.hr/um:nbn:hr:217:744511>

Rights / Prava: [In copyright](#)/[Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2024-11-05**



Repository / Repozitorij:

[Repository of the Faculty of Science - University of Zagreb](#)



# Metoda bootstrapa u statističkom zaključivanju

---

Vajdić, Ivan

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: <https://um.nsk.hr/um:nbn:hr:217:744511>

Rights / Prava: [In copyright](#)/[Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2024-06-18**



Repository / Repozitorij:

[Repository of the Faculty of Science - University of Zagreb](#)



**SVEUČILIŠTE U ZAGREBU**  
**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**  
**MATEMATIČKI ODSJEK**

Ivan Vajdić

**METODA *BOOTSTRAPA* U**  
**STATISTIČKOM ZAKLJUČIVANJU**

Diplomski rad

Voditelj rada:  
prof. dr. sc. Miljenko Huzak

Zagreb, srpanj 2020.

Ovaj diplomski rad obranjen je dana \_\_\_\_\_ pred ispitnim povjerenstvom u sastavu:

1. \_\_\_\_\_, predsjednik
2. \_\_\_\_\_, član
3. \_\_\_\_\_, član

Povjerenstvo je rad ocijenilo ocjenom \_\_\_\_\_.

Potpisi članova povjerenstva:

1. \_\_\_\_\_
2. \_\_\_\_\_
3. \_\_\_\_\_

*Djedu Josipu*

# Sadržaj

<b>Sadržaj</b>	<b>iv</b>
<b>Uvod</b>	<b>2</b>
<b>1 Metoda <i>bootstrapa</i></b>	<b>3</b>
1.1 Osnovno o metodi . . . . .	3
1.2 Neparametarska procjena varijance procjenitelja . . . . .	5
<b>2 Testiranje statističkih hipoteza</b>	<b>9</b>
2.1 Uvod . . . . .	9
2.2 Parametarski testovi . . . . .	13
2.3 Permutacijski testovi . . . . .	24
2.4 Neparametarski testovi . . . . .	28
2.5 Metoda korekcije $p$ -vrijednosti . . . . .	38
<b>3 Pouzdani intervali</b>	<b>41</b>
3.1 Uvod . . . . .	41
3.2 Osnovni pouzdani intervali . . . . .	42
3.3 Percentilni pouzdani intervali . . . . .	52
3.4 Pouzdani intervali uz dvostruki <i>bootstrap</i> . . . . .	61
3.5 Pouzdana područja . . . . .	64
<b>Bibliografija</b>	<b>69</b>

# Uvod

Područje statistike ugrubo se može podijeliti na opisnu i inferencijalnu statistiku. Opisna statistika nastoji opisati sakupljene podatke koristeći neke numeričke vrijednosti ili grafičke prikaze. Često se ti rezultati koriste da bi se analizirao *uzorak*, to jest da bi se odredile neke njegove zanimljive karakteristike. S druge strane, u inferencijalnoj statistici koristimo informaciju danu u uzorku da bismo donijeli neke zaključke o cijeloj *populaciji* iz koje je taj uzorak uzet. Ti se zaključci mogu odnositi na procjene parametara, donošenje odluka, predviđanja budućih vrijednosti ili bilo kakve generalizacije o populaciji. Najčešće korištene metode u ovoj grani statistike su testiranja statističkih hipoteza i konstrukcija pouzdanih intervala. Za obje metode razvijeni su mnogi postupci i teorijski rezultati. Međutim, često je u primjeni tih metoda potrebno biti oprezan budući da mnoge pretpostavke pod kojima su te metode razvijene ne moraju u praksi biti ispunjene. Na primjer, poznati Studentov  $t$ -test razvijen je za slučaj da uzorak dolazi iz normalno distribuirane populacije. Ta pretpostavka u praksi često nije istinita ili ju je teško provjeriti. Cilj je ovog diplomskog rada prezentirati metodu koja se može koristiti u situacijama gdje klasične statističke metode nisu upotrebljive.

U prvome poglavlju prezentirat ćemo motivacijski primjer na kojem ćemo pokazati osnovnu ideju metode ponovljenog uzorkovanja kako bismo procijenili uzoračku razdiobu jedne statistike od interesa čija je egzaktna razdioba nepoznata. U ostatku poglavlja dajemo metodu neparametarske procjene varijance procjenitelja pomoću takozvanih funkcija utjecaja. Standardna pogreška procjenitelja igrat će važnu ulogu u konstrukciji pouzdanih intervala, a ova metoda daje brz, direktan i precizan način kako se ona može procijeniti.

U drugome poglavlju pokazujemo kako koristeći metodu *bootstrapa* možemo procijeniti  $p$ -vrijednosti raznih statističkih testova. Krećemo s Monte Carlo testovima koji koriste simulirane uzorke kako bi se procijenila nul-razdioba testne statistike, a samim time i  $p$ -vrijednost testa. Zatim pokazujemo kako se postupak može generalizirati na razne parametarske modele odnosno na testiranje hipoteza kod kojih uz nultu hipotezu imamo neki parametarski model za opažanja opisan konačnim brojem parametara. Često je bez nekih općenitih pretpostavki nul-razdioba testne statistike nepoznata, ali može se procijeniti metodom ponovljenog uzorkovanja. Pokazujemo također kako se mogu procijeniti pouzdane pruge koje koristimo za testiranje pripadnosti distribucije pomoću vjerojatnos-

nog grafa. Nakon parametarskih testova prelazimo na neparametarske kod kojih uz nultu hipotezu nemamo poznati parametarski model koji opisuje naša opažanja. U tom slučaju najbolju procjenu modela dobivamo iz empirijske funkcije distribucije. Iz te razdiobe simuliramo mnogo uzoraka na temelju kojih procjenjujemo nul-razdiobu testne statistike i  $p$ -vrijednost testa. Posebnu pažnju stavljamo na permutacijske testove koji su vrsta uvjetnih testova. Kod takvih testova uvjetujemo nekom poznatom statistikom ne bismo li dobili točniju procjenu  $p$ -vrijednosti. Kako i sam naziv kaže, permutacijski testovi koriste permutirane vrijednosti uzorka za procjenu nul-razdiobe testne statistike. Na samom kraju dajemo metodu koja koristi dvostruki *bootstrap* i samim time daje nepristranu i precizniju procjenu  $p$ -vrijednosti.

Konstrukcija pouzdanih intervala tema je trećeg poglavlja. U praksi su rijetki slučajevi u kojima se može odrediti egzaktni pouzdani interval za neki parametar te je stoga nužno imati neku aproksimativnu metodu pri ruci. Osnovnu ideju ilustriraju takozvani osnovni i studentizirani *bootstrap* pouzdani intervali. Posebno razmatramo parametarske i neparametarske slučajeve. Međutim, pokazujemo da oni ne daju uvijek zadovoljavajuće rješenje. Predlažemo korekciju tih metoda koja daje puno bolje rezultate, ali može biti teška za provesti u praksi. Nakon toga prezentiramo percentilnu metodu koja je u praksi izrazito jednostavna, ali ne rezultira naročito dobrim procjenama. Predlažemo korekciju te metode, to jest metodu  $BC_a$  pouzdanog intervala koja se pokazala izrazito uspješnom u raznim situacijama. U parametarskim slučajevima metoda daje gotovo egzaktna rješenja. U neparametarskim modelima rezultati su se pokazali nešto lošijima, ali ipak preciznijim nego u svim ostalim metodama. Metodu je lako provesti budući da se zasniva na algoritamskom pristupu. Također, dajemo metodu korekcije pouzdanih intervala koja koristi dvostruki *bootstrap* i ilustriramo ju na osnovnim pouzdanim intervalima. Na kraju pokazujemo metodu konstrukcije pouzdanih područja za proizvoljan broj nepoznatih parametara.

Svaka je metoda praćena primjerima na kojima ilustriramo opisanu metodu. U gotovo svim se primjerima koriste stvarni podaci preuzeti iz raznih statističkih udžbenika.



# Poglavlje 1

## Metoda *bootstrapa*

### 1.1 Osnovno o metodi

Započnimo ovaj odjeljak motivacijskim primjerom koji će ilustrirati korištenje metode *bootstrapa* u jednom konkretnom problemu.

**Primjer 1.1.1.** *Pretpostavimo da nam je dano sljedećih  $n = 12$  opažanja (realizacija) neke slučajne varijable:*

$i$	1	2	3	4	5	6	7	8	9	10	11	12
$x_i$	14.16	12.9	13.9	9.1	13.54	16.04	11.05	8.3	7.13	10.93	13.42	14.09

*Želimo procijeniti koeficijent varijacije, njegovu standardnu pogrešku i procjenu razdiobe uzoračkog koeficijenta varijacije.*

*Koeficijent varijacije procjenjujemo uzoračkim koeficijentom varijacije definiranim kao  $C = \frac{s}{\bar{x}}$ , gdje su  $\bar{X}$  i  $S$  redom uzoračka aritmetička sredina i uzoračka standardna devijacija. Procjenu koeficijenta varijacije lako je izvršiti:*

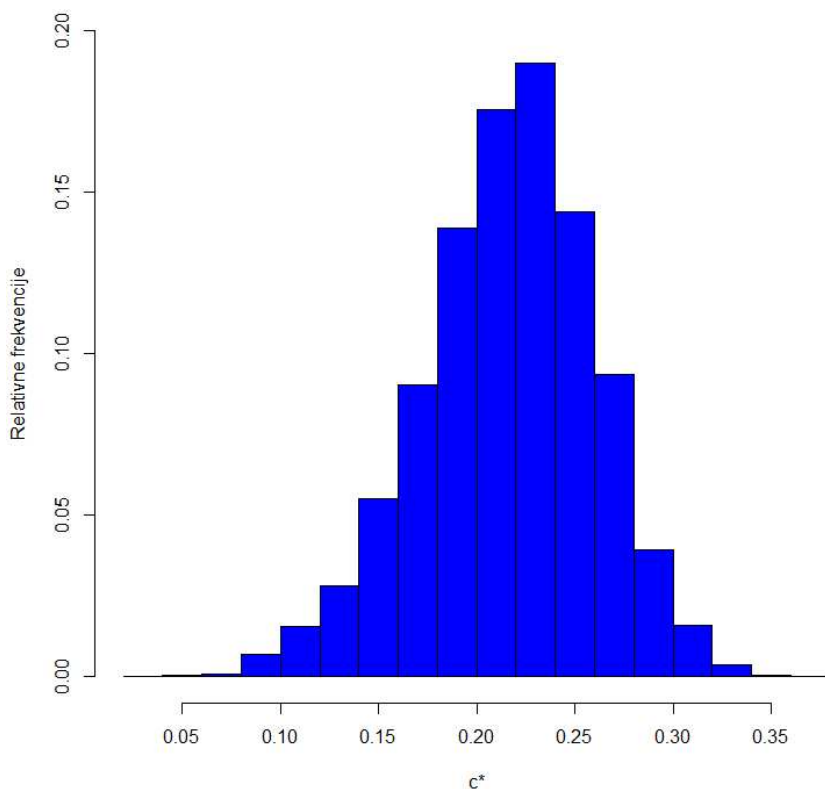
$$\bar{x} = 12.047, s = 2.73 \implies c = \frac{2.73}{12.047} = 0.227.$$

*Međutim, za procjenu standardne greške procjenitelja  $C$  trebali bismo znati izraz za njegovu varijancu (ili, ekvivalentno, za njegovu standardnu devijaciju). Iako postoje neke aproksimacije te standardne pogreške, one vrijede tek uz neke određene uvjete, a njihova točnost može biti upitna, pogotovo za ovako mali uzorak.*

*Zbog tih razloga koristimo metode ponovljenog uzorkovanja kako bismo procijenili razdiobu procjenitelja  $C$  i njegovu standardnu pogrešku. Neka je  $F$  funkcija distribucije slučajne varijable čije smo realizacije opazili. Budući da nam je  $F$  nepoznata, procjenjujemo ju s  $\hat{F}$  koja u ovom slučaju može biti empirijska funkcija distribucije, budući da nemamo neki konkretan (parametarski) model koji bi opisivao naše podatke. Ideja je ponovljenog uzorkovanja simulirati mnogo (nezavisnih) uzoraka iste duljine  $n = 12$  iz razdiobe*

čija je funkcija distribucije procjena  $\hat{F}$  i za svaki simulirani uzorak izračunati uzorački koeficijent varijacije. Te nam simulirane vrijednosti daju ideju (procjenu) kako izgleda prava razdioba procjenitelja  $C$ . Koristeći simulirane vrijednosti možemo i procijeniti standardnu grešku od  $C$  kao standardnu devijaciju svih simuliranih vrijednosti.

Neka je  $R = 9999$  broj simulacija, to jest broj uzoraka koje ćemo simulirati (zajedno s početnim uzorkom to daje ukupno 10000 uzoraka, to jest simulacija vrijednosti procjenitelja  $C$ ). Budući da je ovdje  $\hat{F}$  empirijska funkcija distribucije, simuliranje iz te razdiobe znači uzimanje slučajnog uzorka duljine  $n = 12$  s ponavljanjem iz skupa opaženih vrijednosti. Neka je  $x_{1,r}^*, \dots, x_{12,r}^*$   $r$ -ta simulacija. Iz tih vrijednosti izračunamo vrijednost uzoračkog koeficijenta varijacije  $c_r^*$  i to učinimo za sve  $r \in \{1, 2, \dots, 9999\}$  (odnosno, to ponovimo 9999 puta). Na Slici 1.1 prikazan je histogram simuliranih vrijednosti  $c_r^*$ ,  $r \in \{1, 2, \dots, 9999\}$  i početne vrijednosti  $c = c_0^*$ . Taj nam grafički prikaz daje ideju kako izgleda prava (egzaktna) razdioba procjenitelja  $C$ . Osim toga, standardna pogreška pro-



Slika 1.1: Histogram simuliranih vrijednosti uzoračkog koeficijenta varijacije u Primjeru 1.1.1

cjenitelja  $C$  može se procijeniti pomoću

$$s_{C,boot} = \sqrt{\frac{1}{R} \cdot \sum_{r=0}^R (c_r^* - \bar{c}^*)^2}, \quad (1.1)$$

gdje je  $\bar{c}^*$  prosjek svih simuliranih (i početne) vrijedosti. Budući da je  $\bar{c}^* = 0.216$  iz (1.1) slijedi da je

$$s_{C,boot} = 0.044$$

procjena standardne greške procjenitelja  $C$ .

Napomenimo još da smo u prethodnom primjeru ilustrirali upotrebu *neparametarskog bootstrapa* budući da nismo imali neki parametarski model koji je modelirao naša opažanja.

Prethodni je primjer pokazao tek jednu od mnogobrojnih upotreba metode *bootstrapa* i ponovljenog uzorkovanja. U iduća dva poglavlja pokazat ćemo u koje se još svrhe može koristiti ova metoda.

## 1.2 Neparametarska procjena varijance procjenitelja

Neka je  $\theta$  nepoznati parametar koji nam je od interesa, a  $X_1, \dots, X_n$  slučajni uzorak duljine  $n$  iz razdiobe s funkcijom distribucije  $F$  opisane parametrom  $\theta$ . Neka je  $x_1, \dots, x_n$  realizacija tog slučajnog uzorka (opažene vrijednosti). Neka je  $T$  statistika koja procjenjuje parametar  $\theta$ . Često će nam (većinom u zadnjem poglavlju kod konstrukcije pouzdanih intervala) trebati (procijenjena) varijanca procjenitelja  $T$ . U prethodnom primjeru vidjeli smo kako možemo koristiti metodu ponovljenog uzorkovanja da procijenimo tu varijancu. Međutim, ta će nam varijanca trebati kod svakog simuliranog *bootstrap* uzorka, tako da bi ovaj postupak pronalaženja varijance od  $T$  ipak bio komputacijski prezahtjevan. Želimo ipak nešto *direktniju* procjenu, stoga ćemo u ovom potpoglavlju opisati takozvanu *neparametarsku delta metodu* za procjenu varijance procjenitelja  $T$ .

Ta se procjena vrši pomoću funkcija utjecaja (*engl. influence functions*) koje su se originalno koristile radi analize robusnosti nekih procjenitelja. Međutim, svoju su ulogu pronašle i u procjeni varijance procjenitelja kao što ćemo sada vidjeti.

Neka je  $\hat{F}$  empirijska funkcija distribucije i pretpostavimo da procjenitelj  $T$  možemo zapisati kao  $T = T(\hat{F})$ . Tada je, jasno,  $\theta = T(F)$ . Neka je  $\varepsilon \in \langle 0, 1 \rangle$  odabran po volji i  $x \in \mathbb{R}$  neki realan broj. Definirajmo

$$F_{x,\varepsilon}(t) = (1 - \varepsilon)F(t) + \varepsilon\Delta_x(t),$$

gdje je

$$\Delta_x(t) = \begin{cases} 0, & t < x \\ 1, & t \geq x \end{cases}.$$

Definiramo funkciju utjecaja procjenitelja  $T$  u točki  $x$  kao

$$\text{IF}(x, T) = \lim_{\varepsilon \rightarrow 0} \frac{T(F_{x,\varepsilon}) - T(F)}{\varepsilon} \quad (1.2)$$

ukoliko ovaj limes postoji. Iz definicije je jasno da funkciju utjecaja možemo računati i kao

$$\text{IF}(x, T) = \left. \frac{\partial T(F_{x,\varepsilon})}{\partial \varepsilon} \right|_{\varepsilon=0}. \quad (1.3)$$

Empirijske procjene funkcije utjecaja označit ćemo s  $l(x) = \text{IF}(x, T; \hat{F})$ . Dodatak  $\hat{F}$  znači da umjesto (nepoznate)  $F$  koristimo procjenu  $\hat{F}$ . Pokažimo na jednom jednostavnom primjeru kako možemo izračunati funkciju utjecaja.

**Primjer 1.2.1.** *Pretpostavimo da je  $T = \bar{X}$  uzoračka aritmetička sredina. Tada je u našim oznakama  $T(F) = \mu$  (populacijsko očekivanje), a  $T(\hat{F}) = \bar{X}$ . Kako je  $\mu = \int_{\mathbb{R}} y dF(y) = T(F)$  imamo da je, za proizvoljni  $x \in \mathbb{R}$  i  $\varepsilon \in \langle 0, 1 \rangle$ ,*

$$T(F_{x,\varepsilon}) = T((1 - \varepsilon)F + \varepsilon\Delta_x) = \int_{\mathbb{R}} (1 - \varepsilon)y dF(y) + \int_{\mathbb{R}} \varepsilon y d\Delta_x = (1 - \varepsilon)\mu + \varepsilon x.$$

Deriviramo li gornji izraz po  $\varepsilon$  dobivamo

$$\frac{\partial T(F_{x,\varepsilon})}{\partial \varepsilon} = -\mu + x.$$

Konačno, evaluiramo gornji izraz u  $\varepsilon = 0$  i dobivamo funkciju utjecaja:

$$\text{IF}(x, T) = x - \mu.$$

Empirijska procjena ove funkcije utjecaja dana je s

$$l(x) = x - \bar{x},$$

gdje je  $\bar{x}$  opažena vrijednost uzoračke aritmetičke sredine.

Promotrimo sada  $\text{IF}(X)$  kao slučajnu varijablu (za neki procjenitelj  $T$  kojeg nećemo pisati). Može se pokazati da vrijedi  $\mathbb{E}(\text{IF}(X)) = 0$  pa je stoga  $\text{Var}(\text{IF}(X)) = \mathbb{E}(\text{IF}^2(X))$  ukoliko taj drugi moment postoji. Budući da su  $\text{IF}(X_1), \dots, \text{IF}(X_n)$  nezavisne i jednako distribuirane te budući da vrijedi

$$(T - \theta) \cdot \sqrt{n} = \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^n \text{IF}(X_i) + o_p(1),$$

to uz pretpostavku da  $\text{IF}(X_i)$  imaju konačne varijance po Levyjevom centralnom graničnom teoremu vrijedi

$$(T - \theta) \cdot \sqrt{n} \xrightarrow{D} N(0, \mathbb{E}(\text{IF}^2(X_1))).$$

Označimo  $v_{n,T}(F) = \frac{1}{n} \cdot \mathbb{E}(\text{IF}^2(X_1))$  pa imamo da je asimptotska varijanca procjenitelja  $T$  upravo  $v_{n,T}(F)$ . Ako umjesto  $F$  stavimo procjenu  $\hat{F}$  dobivamo procjenu varijance od  $T$  neparametarskom delta metodom danu s

$$v_{n,T}(\hat{F}) = \frac{1}{n^2} \cdot \sum_{i=1}^n l^2(x_i). \quad (1.4)$$

Često se parametar  $\theta$  može napisati preko jednadžbe oblika  $\int_{\mathbb{R}} h(x, \theta) dF(x) = 0$ . Očekivanje  $\theta = \mu$  je ovog oblika. Tada se procjenitelj  $T$  za  $\theta$  može napisati preko jednadžbe  $\sum_{i=1}^n h(x_i, t) \cdot \frac{1}{n} = 0$ , gdje je  $t$  opažena vrijednost procjenitelja  $T$ . Funkcija utjecaja se u tom slučaju može izraziti kao

$$\text{IF}(x, T) = \frac{h(x, \theta)}{\mathbb{E}\left(-\frac{\partial h}{\partial \theta}(X, \theta)\right)}.$$

Empirijska procjena funkcije utjecaja je

$$l(x_i) = \frac{-nh(x_i, t)}{\sum_{i=1}^n \frac{\partial h}{\partial t}(x_i, t)}$$

pa je procjena varijance od  $T$  neparametarskom delta metodom dana s

$$v_{n,T}(\hat{F}) = \frac{\sum_{i=1}^n (h(x_i, t))^2}{\left(\sum_{i=1}^n \frac{\partial h}{\partial t}(x_i, t)\right)^2}. \quad (1.5)$$

### **Jackknife procjene**

Ako je teško ili nemoguće pronaći funkciju utjecaja, pa samim time i empirijske vrijednosti funkcije utjecaja, možemo ih aproksimirati *jackknifeom*. Neka je kao i ranije  $T$  procjenitelj nepoznatog parametra  $\theta$ , a  $t$  njegova opažena vrijednost. Neka je  $t_{-i}$  opažena vrijednost procjenitelja  $T$  dobivena iz početnog uzorka, ali bez elementa  $x_i$ . Empirijska procjena  $l(x_i)$  tada se procjenjuje s

$$l_{\text{jack}}(x_i) = (n - 1) \cdot (t - t_{-i}) \quad (1.6)$$

i tu procjenu zovemo *jackknife* procjena funkcije utjecaja. Procjena pristranosti procjenitelja  $T$  je

$$b_{T,\text{jack}} = -\frac{1}{n} \cdot \sum_{i=1}^n l_{\text{jack}}(x_i)$$

pa je procjena varijance od  $T$  dana s

$$v_{T,\text{jack}} = \frac{1}{n(n-1)} \cdot \left( \sum_{i=1}^n l_{\text{jack}}^2(x_i) - nb_{T,\text{jack}}^2 \right). \quad (1.7)$$

U praksi možemo koristiti (1.7) kao procjenu varijance ili uvrstiti procjene (1.6) u (1.4) te tako dobiti procjenu za  $v_{n,T}(\hat{F})$ .

**Primjer 1.2.2.** U Primjeru 1.2.1 izračunali smo empirijsku procjenu funkcije utjecaja  $l(x_i) = x_i - \bar{x}$ . Stoga je procjena varijance procjenitelja  $\bar{X}$  dobivena neparametarskom delta metodom (1.4)

$$v_{n,\bar{X}}(\hat{F}) = \frac{1}{n^2} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n^2} \cdot s^2,$$

gdje je  $s^2$  opažena vrijednost uzoračke varijance.

Kako se očekivanje  $\theta$  može napisati preko jednadžbe  $\int_{\mathbb{R}} (x - \theta) dF(x) = 0$  imamo da je  $h(x, \theta) = x - \theta$  pa je  $h(x_i, \bar{x}) = x_i - \bar{x}$ . Stoga je  $\frac{\partial h}{\partial \bar{x}}(x_i, \bar{x}) = -1$  pa je u (1.5)

$$v_{n,\bar{X}}(\hat{F}) = \frac{1}{n^2} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n^2} \cdot s^2,$$

jednako kao i ranije.

Ukoliko izbrišemo  $i$ -tu opservaciju  $x_i$ , opažena vrijednost uzoračke aritmetičke sredine je  $\bar{x}_{-i} = \frac{n\bar{x} - x_i}{n-1}$  pa je  $l_{\text{jack}}(x_i) = (n-1) \cdot \left( \bar{x} - \frac{n\bar{x} - x_i}{n-1} \right) = x_i - \bar{x}$ . Dakle je  $b_{\bar{X},\text{jack}} = 0$  pa je

$$v_{\bar{X},\text{jack}} = \frac{1}{n(n-1)} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \cdot s^2.$$

Ukoliko uvrstimo procjenu  $l_{\text{jack}}(x_i)$  u (1.4) za procjenu dobivamo ponovno  $v_{\bar{X}}(\hat{F})$ .

Sjetimo se još da je uobičajena procjena varijance procjenitelja  $\bar{X}$

$$s_{\bar{X}}^2 = \frac{1}{n} \cdot s^2,$$

a sve gornje procjene su vrlo blizu ove (jackknife procjena je čak i jednaka ovoj procjeni).

## Poglavlje 2

# Testiranje statističkih hipoteza

### 2.1 Uvod

Često je u statističkim analizama potrebno provesti testove značajnosti nekih hipoteza. U velikom broju slučajeva dešava se da klasične statističke metode nisu primjenjive na naš problem. Neke neparametarske metode mogu dati rješenje, međutim njihova preciznost zna biti upitna. U tu svrhu razvijene su metode ponovljenog uzorkovanja kao što su, na primjer, randomizacijski i permutacijski testovi. S pojavom računala razvijaju se i Monte Carlo testovi koji koriste simulirane podatke. Naš je cilj u ovom poglavlju opisati metode ponovljenog uzorkovanja i načine na koji se one mogu iskoristiti za testiranje statističkih hipoteza.

Statističke hipoteze mogu se podijeliti na *jednostavne* i *složene*. Kod jednostavnih hipoteza svi su mogući parametri u potpunosti određeni nultom hipotezom. Na primjer, želimo testirati dolazi li uzorak  $x_1, \dots, x_n$  iz standardne normalne populacije pa nulta hipoteza glasi  $H_0 : X \sim N(0, 1)$  gdje je  $X$  opažanje (slučajna varijabla) uzeto iz iste populacije kao i uzorak. S druge strane, kod složenih hipoteza neki parametri ostaju neodređeni. Tako je primjer složene hipoteze  $H_0 : X \sim N(0, \sigma^2)$ , gdje je parametar  $\sigma^2$  nepoznat.

Testiranje bilo kojih hipoteza uglavnom se provodi slijedeći istu proceduru. Definiramo testnu statistiku  $T$  koja na neki način mjeri koliko naš uzorak odstupa od nulte hipoteze. Ukoliko je to odstupanje *veliko*, uzorak je dao dovoljno dokaza da možemo odbaciti nultu hipotezu i zaključiti u korist alternativne hipoteze. Koliko je to odstupanje zapravo *veliko* mjeri se uz pomoć veličine koju nazivamo  $p$ -vrijednost testa. Pretpostavimo za početak da je nulta hipoteza jednostavna i označimo s  $t$  opaženu vrijednost testne statistike. Definiramo opaženu  $p$ -vrijednost kao

$$p = \mathbb{P}(T \geq t \mid H_0). \quad (2.1)$$

Dakle,  $p$ -vrijednost testa je vjerojatnost da opazimo vrijednost testne statistike još *eks-*

*tremniju* od one koju smo prvi put opazili, uz uvjet da je nulta hipoteza točna. Ukoliko test provodimo na nekoj zadanoj razini značajnosti  $\alpha$  tada ćemo nultu hipotezu odbaciti ako je  $p$ -vrijednost testa manja ili jednaka  $\alpha$ . U suprotnom nećemo odbaciti nultu hipotezu. Napomenimo još da ćemo razdiobu testne statistike  $T$  uz pretpostavku da je nulta hipoteza točna nazivati nul-razdiobom od  $T$ .

Primijetimo da je  $p$ -vrijednost testa definirana s (2.1) slučajna varijabla i označimo ju s  $P$ . Ako pretpostavimo da je testna statistika  $T$  apsolutno neprekidna slučajna varijabla (to najčešće i jest slučaj pa ovakva pretpostavka ne predstavlja veliko ograničenje) onda se može pokazati da je  $P$  uniformna slučajna varijabla na segmentu  $[0, 1]$ . Dakle, za  $p$ -vrijednost testa vrijedi

$$\mathbb{P}(P \leq p \mid H_0) = p. \quad (2.2)$$

Zaključujemo da ako je nulta hipoteza zaista istinita, u  $\alpha \cdot 100\%$  slučajeva test će pogrešno odbaciti nultu hipotezu, gdje je  $\alpha$  razina značajnosti zadana prije provođenja testa. Svaki *dobar* test mora imati svojstvo (2.2) te će nam to svojstvo biti jedan od kriterija za konstrukciju testova.

## Odabir testne statistike

Ukoliko je razumno pretpostaviti da naša opažanja slijede neku (poznatu) razdiobu opisanu (konačnim) brojem parametara tada imamo parametarski model. Pretpostavimo, na primjer, da je model opisan nepoznatim parametrom  $\theta$  i da želimo testirati hipoteze

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1.$$

Ako s

$$L(\theta) = f_{X_1, \dots, X_n}(x_1, \dots, x_n \mid \theta)$$

označimo funkciju vjerodostojnosti, onda je *najbolja* testna statistika za testiranje gornjih hipoteza omjer vjerodostojnosti:

$$T = \frac{L(\theta_0)}{L(\theta_1)}. \quad (2.3)$$

Testna statistika može se poopćiti za slučajeve složenih hipoteza.

Ako nemamo neki razuman parametarski model za naše podatke, nalazimo se u neparametarskom okruženju te nemamo općeniti izraz za *najbolju* testnu statistiku. Međutim, tada za  $T$  uzimamo onu statistiku koja se čini *najprirodnija* za problem u kojem se nalazimo. Drugim riječima, za  $T$  uzimamo neku statistiku koja *dobro* opisuje (procjenjuje) nepoznati parametar  $\theta$  kojeg testiramo i koja svojom opaženom vrijednošću, na temelju uzorka, može ukazati na to da nulta hipoteza možda nije istinita. Na primjer, recimo da želimo testirati jesu li neka dva obilježja (slučajne varijable)  $X$  i  $Y$  nezavisna. Populacijski



koeficijent korelacije  $\rho = \text{Cor}(X, Y)$  je mjera linearne zavisnosti slučajnih varijabli  $X$  i  $Y$  te je  $\rho = 0$  uz nultu hipotezu. Dakle, za testnu statistiku uzeli bismo  $T = \hat{\rho}$ , odnosno uzorački koeficijent korelacije. Ako je alternativna hipoteza dvostrana, možemo uzeti  $T = \hat{\rho}^2$  budući da nam nije bitno je li korelacija pozitivna ili negativna, već samo koliko iznosi po apsolutnoj vrijednosti.

## Uvjetni testovi

Do sad smo opisivali slučajeve u kojima je nulta hipoteza bila jednostavna. Kod složenih hipoteza nisu svi parametri određeni nultom hipotezom, to jest vjerojatnosna funkcija distribucije  $F$  elemenata uzorka nije u potpunosti određena. Zbog toga  $p$ -vrijednost u (2.1) nije dobro definirana jer vjerojatnost  $\mathbb{P}(T \geq t \mid F)$  ovisi o tome koja funkcija distribucije  $F$  zadovoljava nultu hipotezu. Tom se problemu može pristupiti tako da za testnu statistiku  $T$  odaberemo onu čija je razdioba ista bez obzira na  $F$ . Na primjer, ako je uzorak normalno distribuiran s nepoznatim očekivanjem  $\mu$  i nepoznatom varijancom  $\sigma^2$ , onda

$$T = \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n},$$

uz nultu hipotezu, ima Studentovu  $t$ -razdiobu s  $n - 1$  stupnjem slobode koja ne ovisi o nepoznatom parametru  $\sigma^2$ . Nažalost, u velikom broju slučajeva ovakve je testne statistike teško pronaći, pogotovo ako se radi o neparametarskim modelima, to jest ako nemamo neku konkretnu razdiobu koja aproksimira razdiobu našeg uzorka.

Drugi pristup ovom problemu jest da se na testnu statistiku  $T$  uvjetuje nekom dovoljnom statistikom  $S$  za parametre koji su ostali neodređeni nultom hipotezom. Dovoljna statistika  $S$  ima svojstvo da uvjetna razdioba bilo koje statistike  $V$  uz dano  $S = s$  ne ovisi o parametrima za koje je statistika  $S$  dovoljna. Uvjetna  $p$ -vrijednost testa je onda dobro definirana s

$$p = \mathbb{P}(T \geq t \mid S = s, H_0). \quad (2.4)$$

Studentov  $t$ -test opisan ranije primjer je ovakvog testa budući da je uz pretpostavku normalnosti statistika  $S$  (uzoračka standardna devijacija) dovoljna za  $\sigma$ .

U slučaju da nije moguće pronaći dovoljnu statistiku kojom bismo uvjetovali testnu statistiku  $T$ , možemo procijeniti funkciju distribucije  $F$  s  $\hat{F}_0$  koja zadovoljava nultu hipotezu. Procjena  $p$ -vrijednosti testa je tada dana s

$$p = \mathbb{P}(T \geq t \mid \hat{F}_0). \quad (2.5)$$

Kažemo da je ovo *procjena*  $p$ -vrijednosti budući da ona ne zadovoljava (2.2) egzaktno, ali je izrazito dobra aproksimacija prave  $p$ -vrijednosti.

## Pivotni testovi

Za testiranje hipoteza moguće je koristiti ekvivalenciju između testova značajnosti i pouzdanih intervala: ako vrijednost  $\theta_0$  pretpostavljena u nultoj hipotezi leži izvan  $(1 - \alpha) \cdot 100\%$  pouzdanog intervala za nepoznati parametar  $\theta$  onda na razini značajnosti  $\alpha$  možemo odbaciti nultu hipotezu u korist alternativne hipoteze. Naravno, alternativna hipoteza određuje oblik pouzdanog intervala (jednostrani ili dvostrani) i obratno. Na primjer, ako alternativna hipoteza glasi  $H_1 : \theta < \theta_0$  onda je potrebno konstruirati donji (jednostrani) pouzdani interval za  $\theta$ , koji će biti oblika  $(-\infty, t_U]$ , gdje je  $t_U$  gornja granica tog pouzdanog intervala.

Testovi ovakvog oblika zovu se pivotni testovi. Pretpostavimo da želimo testirati hipoteze

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta > \theta_0$$

gdje je  $\theta \in \mathbb{R}$  nepoznati parametar, a  $\theta_0 \in \mathbb{R}$  pretpostavljena vrijednost. Neka je  $T$  procjenitelj za  $\theta$  s procijenjenom varijancom  $V$ . Neka je studentizirana statistika  $Z = \frac{T - \theta}{\sqrt{V}}$  pivotna, to jest razdioba joj je ista za svaki  $F$ , odnosno za svaku vrijednost  $\theta$ . Neka je još  $z_0 = \frac{t - \theta_0}{\sqrt{v}}$  opažena vrijednost studentizirane testne statistike. Tada je  $p$ -vrijednost ovog testa dana s

$$p = \mathbb{P}\left(\frac{T - \theta_0}{\sqrt{V}} \geq \frac{t - \theta_0}{\sqrt{v}} \mid H_0\right).$$

Međutim,  $Z$  je pivotna pa je

$$\mathbb{P}\left(Z \geq \frac{t - \theta_0}{\sqrt{v}} \mid H_0\right) = \mathbb{P}\left(Z \geq \frac{t - \theta_0}{\sqrt{v}} \mid F\right).$$

Dakle, imamo da je

$$p = \mathbb{P}(Z \geq z_0 \mid F). \quad (2.6)$$

U parametarskom modelu postoji poseban slučaj pivotnog testa. Neka je  $\psi$  parametar (ili vektor parametara) kojeg testiramo i neka je  $\lambda$  vektor preostalih (moguće nepoznatih) parametara. Definiramo generalizirani omjer vjerodostojnosti kao

$$\mathcal{L} = \frac{\max_{H_0} L(\psi, \lambda)}{\max L(\psi, \lambda)}.$$

Pretpostavimo da želimo testirati sljedeće hipoteze:

$$H_0 : \psi = \psi_0$$

$$H_1 : \psi \neq \psi_0.$$

Gornji omjer vjerodostojnosti se tada može napisati kao

$$\mathcal{L} = \frac{L(\psi_0, \hat{\lambda}_0)}{L(\hat{\psi}, \hat{\lambda})}, \quad (2.7)$$

gdje su  $\hat{\psi}$  i  $\hat{\lambda}$  procjenitelji metodom maksimalne vjerodostojnosti, a  $\hat{\lambda}_0$  također procjena metodom maksimalne vjerodostojnosti, ali uz nultu hipotezu. Uz neke određene pretpostavke može se pokazati da  $T = -2 \ln(\mathcal{L})$  ima asimptotski  $\chi^2$ -razdiobu s  $m$  stupnjeva slobode, gdje je  $m$  broj parametara koji se testiraju (dimenzija od  $\psi$ ). Dakle,  $p$ -vrijednost ovog testa je

$$p \approx \mathbb{P}(T \geq t) \quad (2.8)$$

i ona ne ovisi o  $\lambda$  pa zaključujemo da je  $T$ , odnosno  $\mathcal{L}$ , aproksimativno pivotna testna statistika.

## 2.2 Parametarski testovi

Ponekad je teško ili gotovo nemoguće opisati egzaktna teorijska svojstva testne statistike  $T$  koju koristimo za testiranje nekih hipoteza (na primjer, njenu nul-razdiobu), čak i ako imamo razuman parametarski model za naše podatke. U ovom odjeljku opisat ćemo korištenje parametarskog ponovljenog uzorkovanja uz nultu hipotezu kao moguću metodu za testiranje hipoteza.

### Monte Carlo testovi

Pretpostavimo da imamo parametarski model i testnu statistiku  $T$ . Pretpostavimo također je nul-razdioba od  $T$  u potpunosti određena nultom hipotezom. Tada je  $p$ -vrijednost testa dana s (2.1), a u slučaju uvjetovanja dovoljnom statistikom s (2.4). Cilj Monte Carlo simulacija jest dati aproksimaciju te  $p$ -vrijednosti budući da je u velikom broju slučajeva vrlo teško pronaći egzaktnu  $p$ -vrijednost testa.

Ideja je simulirati velik broj opažanja testne statistike  $T$  uz nultu hipotezu i tako dobiti aproksimativnu nul-razdiobu od  $T$ . Preciznije, simuliramo  $R$  nezavisnih uzoraka iste dužine, simuliranih iz modela kojeg određuje nulta hipoteza. Iz svakog simuliranog uzorka izračunamo opažene vrijednosti testne statistike, označene s  $t_1^*, \dots, t_R^*$ . Da bismo dobili (aproksimativnu)  $p$ -vrijednost testa, uspoređujemo opaženu vrijednost testne statistike  $t$  sa simuliranim vrijednostima. Uz nultu hipotezu, svaka od  $R + 1$  vrijednosti  $t, t_1^*, \dots, t_R^*$  jednako je vjerojatna pa, uz pretpostavku da je  $T$  apsolutno neprekidna, imamo da je

$$\mathbb{P}(T < T_{(r)}^* | H_0) = \frac{r}{R + 1}, \quad (2.9)$$

gdje je  $s_{(r)}^*$  označena  $r$ -ta po redu vrijednost u nizu uređenih simuliranih testnih statistika (poredanih uzlazno). Dakle, ako je točno  $k$  simuliranih vrijednosti strogo veće od opažene vrijednosti  $t$  onda je

$$p = \mathbb{P}(T \geq t \mid H_0) \approx \frac{k+1}{R+1} = p_{\text{mc}}. \quad (2.10)$$

Razlomak na desnoj strani iza znaka aproksimacije zvat ćemo Monte Carlo  $p$ -vrijednost. Pokažimo da je ova definicija dobra. Ako je  $T$  apsolutno neprekidna, onda iz (2.9) i (2.10) slijedi da je razdioba slučajne varijable  $P_{\text{mc}}$  uniformna na skupu  $\left\{\frac{1}{R+1}, \dots, \frac{R}{R+1}, 1\right\}$ . Dakle, ovo je diskretni analogon od (2.2) pa zaključujemo da je Monte Carlo test egzaktan (u smislu svoje definicije). Naravno, test će biti egzaktan u smislu definicije (2.2) tek ako uzmemo broj simulacija  $R = +\infty$ , što naravno nije izvedivo, ali ni potrebno budući da je (2.10) dobra aproksimacija  $p$ -vrijednosti već i za manje vrijednosti  $R$ . U slučaju da je  $T$  diskretna mogu se pojaviti ponavljajuće vrijednosti od  $t_r^*$  za neki  $r$  pa stoga za općenitu definiciju Monte Carlo  $p$ -vrijednosti uzimamo

$$p_{\text{mc}} = \frac{1 + k(\{r; t_r^* \geq t\})}{R+1}, \quad (2.11)$$

gdje  $k(A)$  označava kardinalnost (broj elemenata) skupa  $A$ .

**Primjer 2.2.1.** *Pretpostavimo da je trajanje neke komponente eksponencijalno distribuirano s nepoznatim parametrom  $\lambda > 0$ . Na slučajan način odabrano je deset takvih komponenti i mjereno im je vrijeme života. Trajanje tih komponenti dano je u sljedećoj tablici:*

$i$	1	2	3	4	5	6	7	8	9	10
Trajanje	85	21	32	235	77	26	13	81	54	113

Proizvođač tih komponenti tvrdi da im je očekivano trajanje 80 jedinica vremena. Želimo testirati njegovu tvrdnju, to jest, želimo testirati:

$$H_0 : \mu = 80$$

$$H_1 : \mu < 80,$$

gdje je  $\mu = \mathbb{E}(X)$  očekivano trajanje slučajno odabrane komponente. Budući da je  $\mu = \frac{1}{\lambda}$ , gdje je  $\lambda$  parametar eksponencijalne razdiobe, imamo da su gornje hipoteze ekvivalentne s

$$H_0 : \lambda = \frac{1}{80}$$

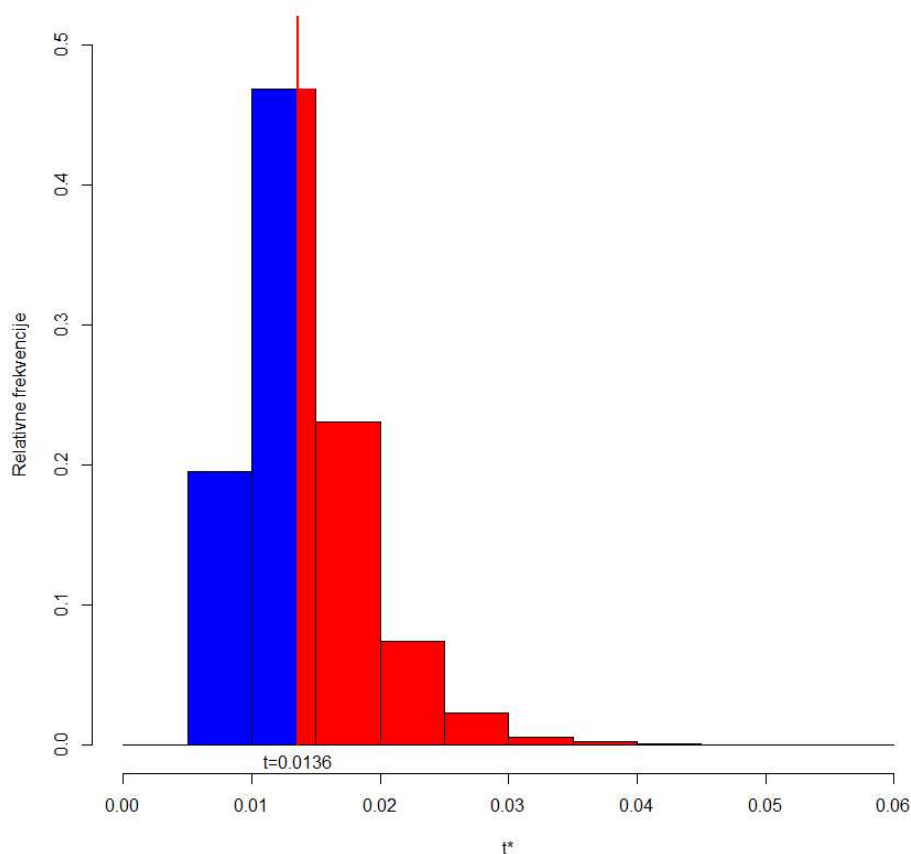
$$H_1 : \lambda > \frac{1}{80}.$$

Dakle, nulta hipoteza u potpunosti određuje sve parametre i uz nultu hipotezu je  $X \sim \text{Exp}\left(\frac{1}{80}\right)$  razdioba trajanja komponenti.

Za testnu statistiku  $T$  možemo uzeti  $T = \frac{1}{\bar{X}}$  budući da je  $\bar{X}$  (jako) konzistentan procjenitelj za  $\mu$ . Opažena vrijednost testne statistike je  $t = 0.0136$ .

Neka je  $R = 9999$  broj simulacija. Simuliramo uzorak duljine  $n = 10$  iz  $\text{Exp}\left(\frac{1}{80}\right)$  razdiobe i izračunamo vrijednost testne statistike  $t^*$ . Ponovimo  $R$  puta da dobijemo simulirane vrijednosti  $t_1^*, \dots, t_{9999}^*$ . Među simuliranim vrijednostima, njih 4472 veće je ili jednako  $t = 0.0136$  iz čega dobivamo da Monte Carlo  $p$ -vrijednost testa (2.11) iznosi  $p = 0.4473$  te ne odbacujemo hipotezu da je očekivano trajanje života tih komponenti 80 jedinica vremena.

Na Slici 2.1 prikazan je histogram simuliranih vrijednosti  $t^*$ . Vertikalna crvena linija označava opaženu vrijednost testne statistike, a dio označen crvenom bojom površinu histograma koja daje Monte Carlo  $p$ -vrijednost.



Slika 2.1: Histogram simuliranih vrijednosti testne statistike u Primjeru 2.2.1

Egzaktna  $p$ -vrijednost ovog testa može se dobiti koristeći činjenicu da statistika

$$U = 2\lambda_0 \cdot \sum_{i=1}^n X_i \stackrel{H_0}{\sim} \chi^2(2n)$$

uz nultu hipotezu ima hi-kvadrat razdiobu s  $2n = 20$  stupnjeva slobode. Vrijednost te testne statistike je  $u = 18.425$ , a test odbacuje nultu hipotezu za male vrijednosti  $U$ . Dakle je  $p$ -vrijednost ovog testa  $F_U(u) = 0.4406$ , gdje je  $F_U$  funkcija distribucije hi-kvadrat slučajne varijable s  $df = 20$  stupnjeva slobode. Uspoređujući s Monte Carlo  $p$ -vrijednosti, vidimo da su egzaktna i aproksimativna  $p$ -vrijednost testa gotovo identične.

Jasno je da točnost u aproksimaciji Monte Carlo  $p$ -vrijednosti raste s porastom broja simulacija  $R$ . Kasnije ćemo komentirati koliki bi  $R$  trebao biti, a za sad preporučamo koristiti  $R = 9999$ , što za današnja računala nije komputacijski naporno.

Kod Monte Carlo testova, sve što nam je potrebno da bismo ih proveli jest da znamo simulirati vrijednosti iz razdioba određenih nultom hipotezom, što je najčešće vrlo jednostavno. Zbog tog su razloga Monte Carlo testovi vrlo laki za provesti, a s druge strane izuzetno efikasni.

## Parametarski *bootstrap* testovi

Kod Monte Carlo testova pretpostavljali smo da nulta hipoteza u potpunosti određuje sve parametre razdiobe testne statistike  $T$ . Međutim, često se dešava da to nije slučaj te da neki parametri uz nultu hipotezu ostanu neodređeni. U nekim je slučajevima teško ili nemoguće pronaći pivotnu testnu statistiku ili pak pronaći dovoljnu statistiku kojom bismo uvjetovali testnu statistiku i time izgubili preostale (nepoznate) neodređene parametre. Tada ne bismo mogli direktno provesti Monte Carlo test.

Mogući pristup ovom problemu je da procijenimo funkciju distribucije  $F$  s  $\hat{F}_0$  koja zadovoljava nultu hipotezu. Procjena  $p$ -vrijednosti testa je tada dana s (2.5) te se takav test naziva *bootstrap* test. Na primjer, pretpostavimo da želimo testirati  $H_0 : \psi = \psi_0$  nasuprot nekoj alternativni. Pretpostavimo da je  $\lambda$  vektor nepoznatih parametara koji nisu određeni nultom hipotezom. Prvi je korak procijeniti  $\lambda$  metodom maksimalne vjerodostojnosti i to uz nultu hipotezu (to jest, tako da stavimo  $\psi = \psi_0$ ). Ako je  $f(x | \psi, \lambda)$  pretpostavljena funkcija gustoće koja modelira naša opažanja, tada je  $f(x | \psi_0, \hat{\lambda}_0)$  procjena te gustoće uz nultu hipotezu. Konačno, za  $\hat{F}_0$  uzimamo funkciju distribucije od  $f(\cdot | \psi_0, \hat{\lambda}_0)$ .

$P$ -vrijednost testa dana s (2.5) često se ne može egzaktno izračunati (ili ne postoji neka zadovoljavajuća aproksimacija) pa podliježemo simulacijama. Slično kao kod Monte Carlo testova, simuliramo  $R$  nezavisnih uzoraka iz razdiobe određene s  $\hat{F}_0$  i računamo opaženu vrijednost testne statistike, tako da dobijemo  $R$  vrijednosti  $t_1^*, \dots, t_R^*$  testne statistike  $T$ . Slično kao u (2.11), ako je  $t$  opažena vrijednost testne statistike, onda  $p$ -vrijednost danu s

(2.5) procjenjujemo s

$$p_{\text{boot}} = \frac{1 + k(\{r; t_r^* \geq t\})}{R + 1}. \quad (2.12)$$

**Primjer 2.2.2.** Na slučajnom uzorku od  $n = 25$  slučajno odabranih muških studenata mjerena je visina ( $Y_1$ ), duljina desne ruke ( $Y_2$ ), duljina desnog stopala ( $Y_3$ ), duljina desnog dlana ( $Y_4$ ), opseg glave ( $Y_5$ ) i duljina nosa ( $Y_6$ ). Podaci su preuzeti iz [7]. Želimo testirati je li razdioba slučajnog vektora  $\mathbf{X} = (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)$  6-dimenzionalna normalna, tj. želimo testirati sljedeće hipoteze:

$$H_0 : \mathbf{X} \sim \mathbf{N}_6(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$H_1 : H_0 \text{ nije istinita,}$$

gdje su  $\boldsymbol{\mu}$  i  $\boldsymbol{\Sigma}$  oboje nepoznati parametri. Za testnu statistiku definiramo tzv. energetska udaljenost po uzoru na [10]:

$$\mathcal{E} = n \left( \frac{2}{n} \cdot \sum_{j=1}^n E|y_j - Z|_d - 2 \cdot \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} - \frac{1}{n^2} \cdot \sum_{j=1}^n \sum_{k=1}^n |y_j - y_k|_d \right),$$

gdje je još

$$E|a - Z|_d = \frac{\sqrt{2} \cdot \Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} + \sqrt{\frac{2}{\pi}} \cdot \sum_{k=0}^{\infty} \left( \frac{(-1)^k}{k! \cdot 2^k} \cdot \frac{|a|_d^{2k+2}}{(2k+1)(2k+2)} \cdot \frac{\Gamma(\frac{d+1}{2}) \cdot \Gamma(k + \frac{3}{2})}{\Gamma(k + \frac{d}{2} + 1)} \right).$$

Ovdje je  $d = 6$  dimenzija vektora  $\mathbf{X}$  a  $y_j$  su standardizirane opservacije. Opažena vrijednost testne statistike je  $e = 1.424$ .

Prvo procjenjujemo parametre  $\boldsymbol{\mu}$  i  $\boldsymbol{\Sigma}$ . Vektor očekivanja  $\boldsymbol{\mu}$  procjenjujemo vektorom uzoračkih aritmetičkih sredina  $\hat{\boldsymbol{\mu}} = (\bar{Y}_1, \dots, \bar{Y}_6)$ , a kovarijacijsku matricu  $\boldsymbol{\Sigma}$  matricom uzoračkih kovarijanci

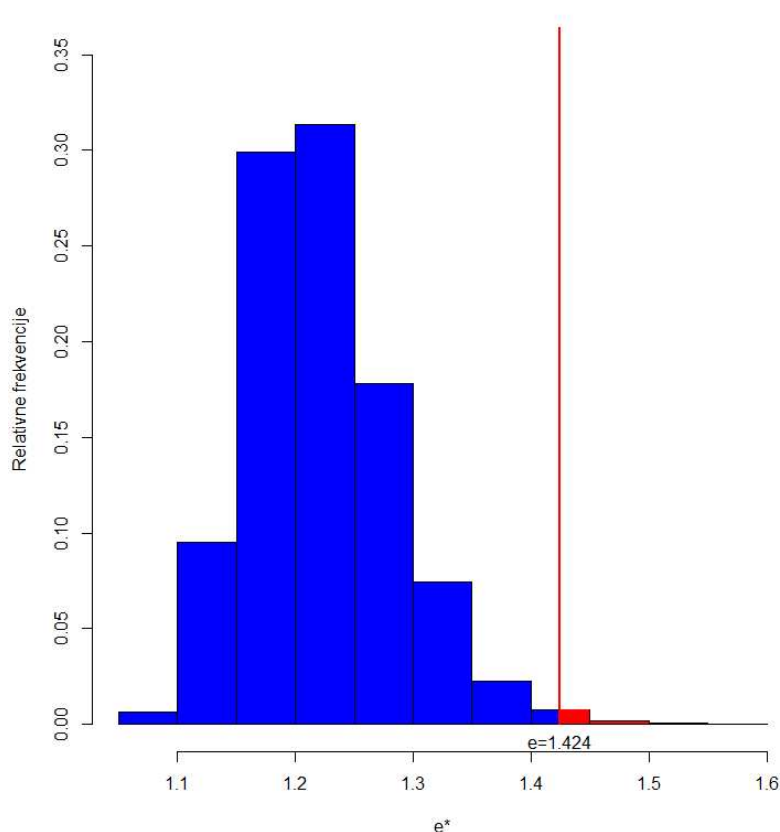
$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{Cov}(Y_1, Y_1) & \hat{Cov}(Y_1, Y_2) & \cdots & \hat{Cov}(Y_1, Y_6) \\ \hat{Cov}(Y_2, Y_1) & \hat{Cov}(Y_2, Y_2) & \cdots & \hat{Cov}(Y_2, Y_6) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Cov}(Y_6, Y_1) & \hat{Cov}(Y_6, Y_2) & \cdots & \hat{Cov}(Y_6, Y_6) \end{bmatrix}.$$

Sada je  $\hat{F}_0$  funkcija distribucije slučajnog vektora koji ima  $\mathbf{N}_6(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$  razdiobu.

Neka je  $R = 9999$  broj simulacija. Simuliramo uzorak duljine  $n = 25$  iz  $\mathbf{N}_6(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$  razdiobe i izračunamo vrijednost testne statistike  $e^*$ . Ponovimo  $R$  puta da dobijemo simulirane vrijednosti  $e_1^*, \dots, e_{9999}^*$ . Među simuliranim vrijednostima, njih 57 veće je ili jednako  $e = 1.424$  iz čega dobivamo da bootstrap  $p$ -vrijednost testa (2.12) iznosi  $p = 0.0058$

te odbacujemo hipotezu da je zajednička razdioba 6 promatranih obilježja multivarijatna normalna.

Na Slici 2.2 prikazan je histogram simuliranih vrijednosti  $e^*$ . Vertikalna crvena linija označava opaženu vrijednost testne statistike, a dio označen crvenom bojom površinu histograma koja daje bootstrap  $p$ -vrijednost.



Slika 2.2: Histogram simuliranih vrijednosti testne statistike u Primjeru 2.2.2

## Grafički testovi

Često se nakon prilagodbe nekog modela podacima provjerava kvaliteta te prilagodbe nekim grafičkim prikazom. Na primjer, prilagodba normalnog modela podacima može se testirati normalnim vjerojatnosnim grafom. U većini slučajeva graf bi trebao pratiti neki pravac, ukoliko je prilagođeni model zaista dobar. Naravno, zbog slučajne pogreške gotovo



uvijek postoje odstupanja od savršene linearnosti i često je dosta teško vizualno odrediti je li neko odstupanje *dovoljno veliko* da možemo sigurno zaključiti da prilagođeni model nije dobar. Zbog toga je poželjno svakom takvom grafičkom prikazu dodati neku pouzdanu prugu, to jest odrediti dio ravnine u kojem dozvoljavamo da se naše točke grafa pojavljuju. Ukoliko puno točaka leži izvan tog dijela ravnine, možemo zaključiti da prilagođeni model nije dobar. Te pouzdane pruge konstruirat ćemo koristeći Monte carlo ili parametarske *bootstrap* metode. Napomenimo da grafički testovi nemaju alternativnu hipotezu, to jest u slučaju da odbacujemo nultu hipotezu (odbacujemo tvrdnju da je prilagođeni model dobar) nemamo konkretnu alternativu kojoj bismo se okrenuli. Ipak, grafički prikazi mogu nam dati ideju o kojoj bi se alternativni radilo, to jest koja bi alternativa bila bolja u slučaju da prvi model kojeg testiramo nije prihvatljiv.

Neka je kao i ranije  $T$  testna statistika i  $\mathcal{A} \subseteq \mathbb{R}$  ograničen skup. Neka je  $\{t(a); a \in \mathcal{A}\}$  opaženi grafički prikaz. Naš je cilj usporediti opaženu vrijednost  $t(a)$  s razdiobom  $T(a)$  ukoliko vrijedi nulta hipoteza, za sve  $a \in \mathcal{A}$ . Na primjer, želimo provjeriti dolaze li opažanja uzorka  $x_1, \dots, x_n$  iz normalno distribuirane populacije. U tom je slučaju  $\mathcal{A} = \left\{ \Phi^{-1}\left(\frac{i}{n+1}\right); i \in \{1, 2, \dots, n\} \right\}$  skup kvantila standardne normalne razdiobe (to jest,  $a_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$ ,  $i \in \{1, 2, \dots, n\}$ ) te  $t(a_i) = x_{(i)}$   $i$ -ta po redu vrijednost u (neopadajuće) uređenom nizu opservacija. U Kartezijevom koordinatnom sustavu crtamo uređene parove  $(a_i, x_{(i)})$ , za sve  $i \in \{1, 2, \dots, n\}$ .

Za početak pretpostavimo da razdioba slučajnog vektora  $\{T(a); a \in \mathcal{A}\}$  ne ovisi o nikakvim nepoznatim parametrima. Na primjer, to je istina ako u prethodnom primjeru studentiziramo opažanja, to jest ako definiramo  $z_i = \frac{x_i - \bar{x}}{s}$  i stavimo  $t(a_i) = z_{(i)}$ . Tada možemo testirati svaki  $a \in \mathcal{A}$  Monte Carlo testom. Naime, za  $a \in \mathcal{A}$  simuliramo  $R$  nezavisnih uzoraka iz razdiobe određene nultom hipotezom i izračunamo  $t_r^*(a)$ , za sve  $r \in \{1, 2, \dots, R\}$ . Uz nultu hipotezu, slučajne varijable  $T(a), T_1^*(a), \dots, T_R^*(a)$  nezavisne su i jednako distribuirane pa vrijedi (2.9), to jest za sve  $a \in \mathcal{A}$  je

$$\mathbb{P}(T(a) < T_{(j)}^*(a) \mid H_0) = \frac{j}{R+1}. \quad (2.13)$$

Dakle, mogli bismo koristiti (2.11) za traženje  $p$ -vrijednosti (jednostranog) testa ukoliko je testna statistika  $T$  takva da test odbacuje nultu hipotezu za velike vrijednosti  $t(a)$ . Za dvostrani test (ili za test koji odbacuje za male vrijednosti  $t(a)$ ) potrebno je modificirati testnu statistiku  $T$ .

Gornji postupak testiranja potrebno je provesti za sve  $a \in \mathcal{A}$ . Međutim, umjesto testiranja svake pojedine vrijednosti  $a \in \mathcal{A}$  možemo naći interval vrijednosti u koji dozvoljavamo da opažanje  $t(a)$  upadne ukoliko je nulta hipoteza istinita. Dakle, za svaki  $a \in \mathcal{A}$  tražimo donju i gornju kritičnu vrijednost. Te su kritične vrijednosti takve da interval između njih ima *pokrivenost*  $(1 - 2\alpha) \cdot 100\%$ , gdje je  $\alpha \in \langle 0, 1 \rangle$  unaprijed zadana razina značajnosti (odnosno, vrijedi  $\mathbb{P}(T(a) < t_L^*) = \mathbb{P}(T(a) > t_U^*) = \alpha$ ). U koordinatnom sustavu crtamo uređene

parove točaka  $a_i$  i njihovih kritičnih vrijednosti i promatramo nalazi li se koja točka izvan svojih kritičnih vrijednosti – u tom slučaju imamo dokaz protiv nulte hipoteze.

Da bismo pronašli kritične vrijednosti prvo biramo  $k \in \mathbb{N}$  i  $R \in \mathbb{N}$  takve da je  $\frac{k}{R+1} = \alpha$  i kritične vrijednosti definiramo kao

$$t_{(k)}^*(a) \quad \text{i} \quad t_{(R+1-k)}^*(a),$$

gdje su  $t_{(1)}^*(a) \leq \dots \leq t_{(R)}^*(a)$  uređene simulirane vrijednosti, kao i ranije. Iz konstrukcije imamo da ukoliko je  $t(a)$  veće od svoje gornje kritične vrijednosti ili manje od svoje donje kritične vrijednosti onda je  $p$ -vrijednost (jednostranog) testa opisanog ranije manja od  $\alpha$ .

Konačno, definiramo pouzdanu prugu kao skup svih intervala između donjih i gornjih kritičnih vrijednosti svake točke  $a \in \mathcal{A}$ , to jest

$$\mathcal{P}^{1-2\alpha} = \{[t_{(k)}^*(a), t_{(R+1-k)}^*(a)]; a \in \mathcal{A}\}. \quad (2.14)$$

Točke koje leže izvan  $\mathcal{P}^{1-2\alpha}$  upućuju na to da nulta hipoteza nije točna.

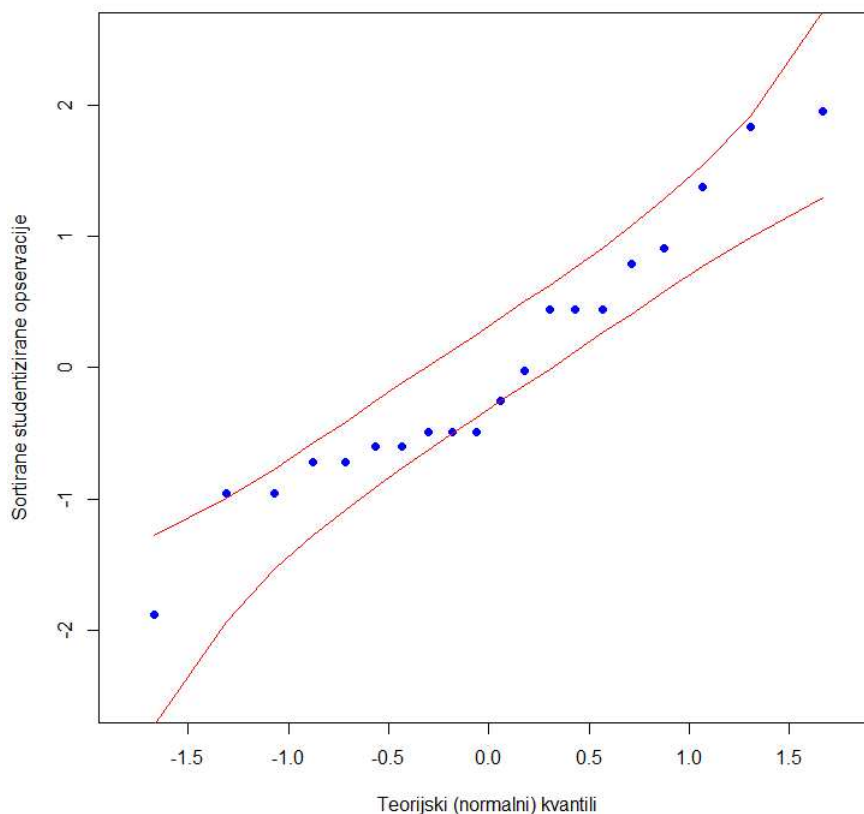
**Primjer 2.2.3.** Na slučajnom uzorku  $n = 20$  vozača mjerena je brzina (u mph) kojom su vozili nekom dionicom ceste. Podaci su preuzeti iz [7]. Želimo provjeriti jesu li brzine vožnje na toj dionici ceste normalno distribuirane, koristeći grafički test.

Pretpostavimo da je  $\alpha = 0.025$  i uzмимо  $R = 9999$ . Tada je  $k = 0.025 \cdot 10000 = 250$ . Prvi je korak studentizirati opservacije ( $z_i = \frac{x_i - \bar{x}}{s}$ , gdje su  $x_1, \dots, x_n$  opažene brzine) tako da se riješimo nepoznatih parametara. Nakon toga možemo simulirati iz standardne normalne razdiobe. Za svaku točku  $a_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$  simuliramo uzorak duljine  $n = 20$  iz  $N(0, 1)$  razdiobe, standardiziramo ga i nađemo vrijednost  $t_r^*(a_i) = z_{(i)}^*$  koja je  $i$ -ta po redu uređena vrijednost u standardiziranom uzorku. To ponovimo  $R$  puta i izračunamo kritične vrijednosti  $t_{(250)}^*(a_i)$  i  $t_{(9750)}^*(a_i)$ . Kritične vrijednosti na ovaj način nađemo za svaku točku  $a_i$ ,  $i \in \{1, 2, \dots, n\}$ . U Kartezijevom koordinatnom sustavu crtamo parove točaka  $(a_i, z_{(i)})$  i njihove intervale pouzdanosti. Područje između dviju krivulja definira pouzdanu prugu.

Na Slici 2.3 prikazan je normalni vjerojatnosni graf podataka zajedno s 95%-pouzdanom prugom. Primjećujemo da dvije točke leže izvan pouzdane pruge čime normalnost podataka iz uzorka postaje upitna.

Iako smo dosad promatrali slučaj u kojem nulta hipoteza određuje sve parametre razdiobe, postupak se može poopćiti na slučaj kad neki parametri ostanu neodređeni. Slično kao i kod *bootstrap* testova, procijenili bismo funkciju distribucije  $F$  s  $\hat{F}_0$  koja zadovoljava nultu hipotezu i simulirali iz te procijenjene razdiobe. U tom slučaju je (2.13) aproksimativno, a procjena  $p$ -vrijednosti testa je dana s (2.12).

Konstrukcija pouzdane pruge koju smo opisali ranije ima jednu manu: (jednostrana) razina značajnosti je  $\alpha$  za svaku točku zasebno, dakle ukupna razina značajnosti svih točaka je puno veća od  $\alpha$ . Drugim riječima, vjerojatnost da ovako konstruirana pouzdana pruga



Slika 2.3: Normalni vjerojatnosni graf i pouzdana pruga za Primjer 2.2.3

sadrži sve točke je puno manja od  $1 - 2\alpha$ . Dakle, to što dvije točke ne leže u pouzdanoj prugi u Primjeru 2.2.3 ne mora nužno značiti da uzorak ne dolazi iz normalno distribuirane populacije.

Puno je korisnije znati kolika je ukupna razina značajnosti svih točaka, to jest kolika je vjerojatnost da neka točka ne leži unutar pouzdane pruge. To se može odrediti empirijski. Pretpostavimo da smo izračunali  $R$  nizova točaka (kao gore) iz kojih određujemo pouzdanu prugu. Neka je  $\mathcal{P}_{-r}^{1-2\alpha}$  pouzdana pruga dobivena micanjem  $r$ -tog niza, to jest skupa  $\{t_r^*(a); a \in \mathcal{A}\}$ . Ideja je usporediti upravo taj niz  $\{t_r^*(a); a \in \mathcal{A}\}$  s pouzdanom prugom  $\mathcal{P}_{-r}^{1-2\alpha}$ , za sve  $r \in \{1, 2, \dots, R\}$ . Zanima nas koliko tih nizova u potpunosti izlazi iz pouzdane pruge dobivene micanjem tog niza, a ukupna razina značajnosti će tada biti proporcija takvih nizova, to jest

$$\alpha_{\text{uk}} = \frac{k(\{r; \{t_r^*(a); a \in \mathcal{A}\} \not\subseteq \mathcal{P}_{-r}^{1-2\alpha}\})}{R}. \quad (2.15)$$

Primijetimo da skup  $\{t_r^*(a); a \in \mathcal{A}\}$  neće u potpunosti biti sadržan u  $\mathcal{P}_{-r}^{1-2\alpha}$  ako i samo ako

$$\exists a \in \mathcal{A} \text{ tako da } R(t_r^*(a)) \leq k \quad \text{ili} \quad \exists a \in \mathcal{A} \text{ tako da } R(t_r^*(a)) \geq R + 1 - k,$$

gdje  $R(t_r^*(a))$  označava rang vrijednosti  $t_r^*(a)$  među vrijednostima  $t_1^*(a), \dots, t_R^*(a)$ . Vrijednost  $k$  je, kao i ranije, takva da  $\alpha = \frac{k}{R+1}$ .

**Primjer 2.2.4.** (Nastavak Primjera 2.2.3) Koristeći (2.15) možemo izračunati da je ukupna razina značajnosti svih točaka u Primjeru 2.2.3 jednaka  $\alpha_{uk} = 0.637$ , što je daleko veće od specificirane razine značajnosti  $2\alpha = 0.05$ .

Direktnom provjerom možemo se uvjeriti da je potrebno staviti  $\alpha = 0.0013$  da bi ukupna razina značajnosti svih točaka bila  $\alpha_{uk} = 0.05$ . Dakle, ako stavimo  $k = 13$  i  $R = 9999$  te ponovimo konstrukciju iz Primjera 2.2.3 dobivamo pouzdanu prugu koja je takva da je vjerojatnost da neka točka ispadne iz nje upravo  $\alpha_{uk} = 0.05$ . Na Slici 2.4 prikazan je normalni vjerojatnosni graf podataka zajedno s pouzdanom prugom kojoj je ukupna razina značajnosti  $\alpha_{uk} = 0.05$ . Primjećujemo da u ovom slučaju niti jedna točka ne leži izvan pouzdane pruge pa na temelju grafičkog prikaza nemamo dovoljno dokaza da odbacimo tvrdnju da opažanja dolaze iz normalno distribuirane populacije.

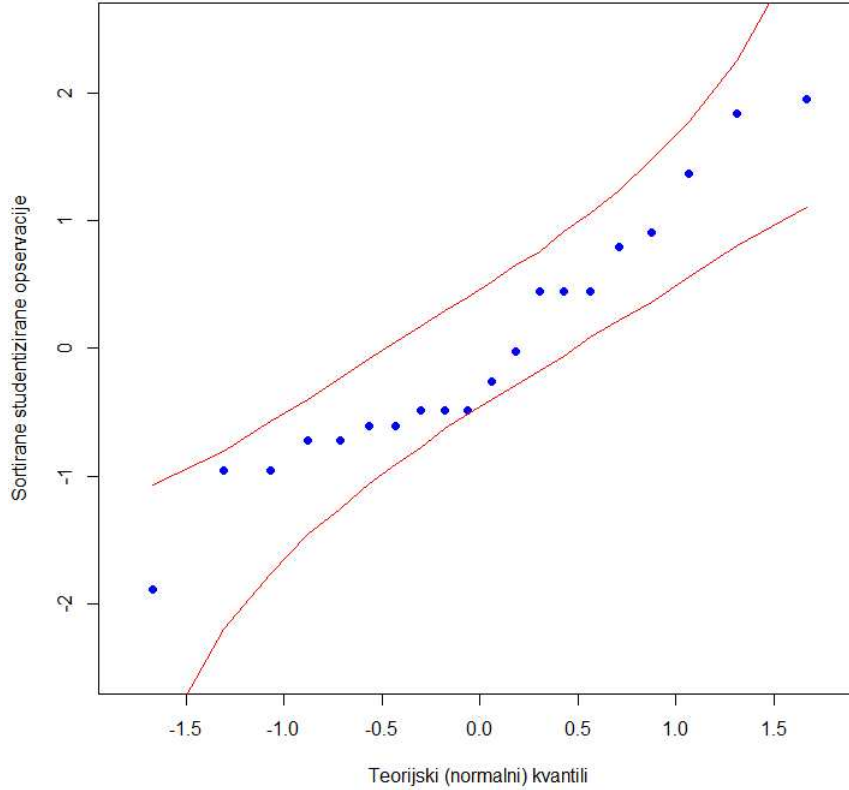
## Odabir broja simulacija $R$

Dosad smo u svakom primjeru uzimali broj simulacija  $R = 9999$  bez nekog posebnog razloga. U ovom pododjeljku pokazat ćemo da je u nekim slučajevima to i više nego dovoljan broj simulacija. Naravno, intuitivno je jasno da s porastom broja simulacija dobivamo sve bolje i bolje aproksimacije (na primjer,  $p$ -vrijednosti testova), ali koliki bi trebao biti taj broj simulacija da bi aproksimacije bile zadovoljavajuće? Kako uopće mjeriti zadovoljivost aproksimacija? Jedan od mogućih kriterija je jakost testa, što se definira kao vjerojatnost odbacivanja nulte hipoteze ukoliko ona stvarno nije istinita.

Pretpostavimo da želimo testirati nultu hipotezu  $H_0$  naspram alternativni  $H_1$  na razini značajnosti  $\alpha \in \langle 0, 1 \rangle$ . Dakle, takav test odbacuje nultu hipotezu ako i samo ako je  $p$ -vrijednost tog testa  $p \leq \alpha$ . Neka je  $k \in \mathbb{N}$  takav da je  $\alpha = \frac{k}{R+1}$  i neka su  $t_1^*, \dots, t_R^*$  simulirane vrijednosti testne statistike, a  $t$  opažena vrijednost testne statistike. Nultu hipotezu odbacujemo ako i samo ako je  $t_{(R+1-k)}^* < t$  pa je jakost ovog testa dana s

$$\gamma_R(\alpha) = \mathbb{P}(\text{Odbacujemo } H_0 \mid H_1) = \mathbb{P}(T_{R+1-k}^* < T \mid H_1).$$

Pretpostavimo još da je  $T$  apsolutno neprekidna slučajna varijabla. Neka su  $g_0$  i  $G_0$  redom funkcija gustoće i funkcija distribucije od  $T$  uz nultu hipotezu, a  $g_1$  i  $G_1$  redom funkcija gustoće i funkcija distribucije od  $T$  uz alternativnu hipotezu. Budući da je  $T_{(R+1-k)}^*$  uređajna statistika uzorka  $T_1^*, \dots, T_R^*$  (zaista, to je uzorak jer su sve te varijable nezavisne i jednako



Slika 2.4: Normalni vjerojatnosni graf i korigirana pouzdana pruga za Primjer 2.2.3

distribuirane), njena funkcija gustoće uz nultu hipotezu je

$$f_{T_{(R+1-k)}^*}(t) = \frac{R!}{(R-k)! \cdot (k-1)!} \cdot [G_0(t)]^{R-k} [1 - G_0(t)]^{k-1} g_0(t) \cdot \mathbf{1}_{\mathcal{S}},$$

gdje je  $\mathcal{S}$  nosač slučajne varijable  $T_{(R+1-k)}^*$ . Sada se funkcija jakosti testa može napisati kao

$$\gamma_R(\alpha) = \int_{-\infty}^{+\infty} \int_{-\infty}^t R \cdot \binom{R-1}{k-1} \cdot [G_0(x)]^{R-k} [1 - G_0(x)]^{k-1} g_0(x) g_1(t) dx dt.$$

Zamjenom varijabli i sređivanjem dobivamo

$$\gamma_R(\alpha) = \int_0^1 \gamma_{\infty}(u) h(u; (R+1)\alpha, (R+1)(1-\alpha)) du, \quad (2.16)$$

gdje je  $\gamma_\infty$  jakost egzaktnog testa (kad je  $R = +\infty$ ), a  $h$  je funkcija gustoće beta slučajne varijable s parametrima  $(R + 1)\alpha$  i  $(R + 1)(1 - \alpha)$ .

Kako je  $\gamma_R$  konkavna funkcija od  $\alpha$ , to je  $\gamma_R(u) \geq \frac{u}{\alpha} \cdot \gamma_R(\alpha) = \tilde{\gamma}_R(u)$  za  $u \leq \alpha$  te je  $\gamma_R(u) \geq \gamma_R(\alpha)$  za  $u > \alpha$ . Iz (2.16) onda slijedi

$$\begin{aligned} \gamma_\infty(\alpha) - \gamma_R(\alpha) &\leq \frac{\gamma_\infty(\alpha)}{2\alpha} \cdot \int_0^1 |u - \alpha| h(u; (R + 1)\alpha, (R + 1)(1 - \alpha)) du \\ &= \frac{\gamma_\infty(\alpha) \alpha^{(R+1)\alpha} (1 - \alpha)^{(R+1)(1-\alpha)} \Gamma(R + 1)}{(R + 1)\alpha \Gamma((R + 1)\alpha) \Gamma((R + 1)(1 - \alpha))}. \end{aligned}$$

Sada koristimo Stirlingovu aproksimaciju koja vrijedi za velike  $x$  :

$$\Gamma(x) \approx \sqrt{2\pi} \cdot x^{x-\frac{1}{2}} \cdot e^{-x}$$

i dobivamo da je

$$\frac{\gamma_R(\alpha)}{\gamma_\infty(\alpha)} \leq 1 - \sqrt{\frac{1 - \alpha}{2\pi(R + 1)\alpha}}. \quad (2.17)$$

U donjoj tablici nalaze se gornje ograde na omjere jakosti testova za neke veličine  $R$  i neke klasične razine značajnosti  $\alpha$ :

$R$	99	999	4999	9999	19999	49999	79999	99999
Omjer jakosti ( $\alpha = 0.10$ )	0.880	0.962	0.983	0.988	0.992	0.995	0.996	0.996
Omjer jakosti ( $\alpha = 0.05$ )	0.826	0.945	0.975	0.983	0.988	0.992	0.994	0.995
Omjer jakosti ( $\alpha = 0.025$ )	0.751	0.921	0.965	0.975	0.982	0.989	0.991	0.992
Omjer jakosti ( $\alpha = 0.01$ )	0.603	0.874	0.944	0.960	0.972	0.982	0.986	0.987

Za  $R = 9999$  su omjeri jakosti veći od 0.95 za sve uobičajene razine značajnosti. Budući da za današnja računala  $R = 9999$  simulacija nije komputacijski naporno, taj nam se odabir čini *najbolji* za daljnje praktične primjene.

## 2.3 Permutacijski testovi

Permutacijski testovi poseban su slučaj neparametarskih testova koje ćemo opisati u idućem poglavlju. Neparametarske metode koristan su alat u slučajevima kad nemamo neki razuman parametarski model koji opisuje naša opažanja u uzorku ili ako samo želimo potvrditi neki rezultat dobiven korištenjem parametarskih metoda.

Kod neparametarskih testova testna statistika uz nultu hipotezu ne ovisi ni o kakvim parametrima. Primjeri neparametarskih statistika su statistike bazirane na rangovima, kao

što su, na primjer, Wilcoxonova statistika, Mann-Whitneyjeva statistika, Kruskal-Wallisova statistika i tako dalje. Odabir testne statistike u neparametarskom slučaju nije očit – za testnu statistiku uzimamo ono što nam se čini *najpogodnije* za naš problem ili ono što bi bila testna statistika kad bismo se nalazili u parametarskom slučaju. Na primjer, ako želimo testirati jesu li očekivanja u dvije različite populacije jednaka, prirodan odabir za testnu statistiku bila bi razlika uzoračkih aritmetičkih sredina (ili uzoračkih medijana ukoliko želimo nešto robusniju procjenu).

Permutacijski testovi na neki način uspoređuju funkcije distribucije. To nas motivira da  $p$ -vrijednost permutacijskog testa tražimo preko izraza (2.4), budući da znamo dovoljnu statistiku za funkciju distribucije. Neka je  $X_1, \dots, X_n$  slučajni uzorak duljine  $n$ ,  $F$  njihova funkcija distribucije i  $Y_1 < Y_2 < \dots < Y_n$  njegove uređajne statistike. Razdioba slučajnog vektora  $(X_1, \dots, X_n)$  uvjetno na  $(Y_1, \dots, Y_n) = (y_1, \dots, y_n)$  je diskretna razdioba koja stavlja vjerojatnost  $\frac{1}{n!}$  na svaku od  $n!$  mogućih permutacija vektora  $(y_1, \dots, y_n)$ . Ovdje pretpostavljamo da su  $X_1, \dots, X_n$  apsolutno neprekidne slučajne varijable. Dakle, ova uvjetna razdioba ne ovisi o  $F$  pa je stoga vektor uređajnih statistika  $(Y_1, \dots, Y_n)$  dovoljna statistika za  $F$ . Uvjet  $S = s$  u (2.4) će predstavljati uvjet  $(Y_1, \dots, Y_n) = (y_1, \dots, y_n)$ .

$P$ -vrijednost danu  $s$  (2.4) aproksimirat ćemo koristeći Monte Carlo test i  $p$ -vrijednost danu  $s$  (2.11). Primijetimo da možemo koristiti Monte Carlo test jer uvjetovanjem  $S = s$  na testnu statistiku gubimo sve nepoznate parametre. Neka je  $T$  testna statistika, a  $R$  broj simulacija. Jer smo uvjetovali vektorom uređajnih statistika uzorka, simuliranje pod nultom hipotezom znači jednostavno permutiranje početnog uzorka. Dakle, kreiramo  $R$  slučajnih permutacija uzorka i za svaku permutaciju izračunamo opaženu vrijednost testne statistike  $t_r^*$ ,  $r \in \{1, 2, \dots, R\}$ .  $P$ -vrijednost permutacijskog testa tada aproksimiramo s (2.11), to jest

$$p \approx p_{mc} = \frac{1 + k(\{r; t_r^* \geq t\})}{R + 1},$$

gdje  $t$  označava opaženu vrijednost testne statistike  $T$ .

**Primjer 2.3.1.** Tijekom  $n = 30$  ljetnih dana na mjernim postajama u Los Angelesu mjeren je jutarnji prosjek insolacije (mjera količine sunčeve svjetlosti) ( $X$ ) i maksimalna količina nekog fotokemijskog zagađivača ( $Y$ ). Podaci su preuzeti iz [9]. Želimo testirati jesu li ta dva obilježja nekorelirana naspram alternativni da postoji pozitivna korelacija između njih, to jest želimo testirati sljedeće hipoteze:

$$H_0 : \rho(X, Y) = 0$$

$$H_1 : \rho(X, Y) > 0,$$

gdje  $\rho(X, Y)$  označava (populacijski) koeficijent korelacije između obilježja (slučajnih varijabli)  $X$  i  $Y$ .

Za testnu statistiku možemo uzeti uzorački koeficijent korelacije

$$T = \frac{1}{n-1} \cdot \sum_{i=1}^n \frac{(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{S_X \cdot S_Y},$$

budući da je to (uzoračka) mjera linearne povezanosti između dvaju obilježja te je to i procjena koeficijenta korelacije. Opažena vrijednost testne statistike je  $t = 0.505$ .

Primijetimo da je ovdje  $F$  bivarijatna funkcija distribucije, to jest funkcija distribucije slučajnog vektora  $(X, Y)$ . Dakle, dovoljna statistika za  $F$  sastoji se od uređajnih statistika uzorka  $X_1, \dots, X_n$  i  $Y_1, \dots, Y_n$ , to jest  $S = (X_{(1)}, \dots, X_{(n)}, Y_{(1)}, \dots, Y_{(n)})$ . Odatle zaključujemo da uz uvjet  $S = s = (x_{(1)}, \dots, x_{(n)}, y_{(1)}, \dots, y_{(n)})$  vrijedi da je  $(X_1, \dots, X_n) = (x_{(1)}, \dots, x_{(n)})$ , a  $(Y_1, \dots, Y_n) = (y_{p_1}, \dots, y_{p_n})$ , gdje je  $(p_1, \dots, p_n)$  neka permutacija vektora indeksa  $(1, \dots, n)$ . Nadalje, uz nultu hipotezu sve su takve permutacije jednako vjerojatne. Preciznije, vrijedi

$$\mathbb{P} \left( T = \frac{1}{n-1} \cdot \sum_{i=1}^n \frac{(x_{(i)} - \bar{x}) \cdot (y_{p_i} - \bar{y})}{S_X \cdot S_Y} \mid H_0, S = s \right) = \frac{1}{n!}.$$

Budući da alternativna hipoteza pretpostavlja pozitivnu korelaciju, test će odbaciti nultu hipotezu ako je vrijednost testne statistike  $T$  velika. Dakle je (2.11) aproksimativna  $p$ -vrijednost ovog testa.

Naravno, egzaktna  $p$ -vrijednost ovog permutacijskog testa dobije se kao proporcija svih mogućih permutacija  $(y_{p_1}, \dots, y_{p_n})$  za koje vrijedi  $t^* \geq t$ , gdje smo s  $t^*$  označili simuliranu vrijednost testne statistike, to jest

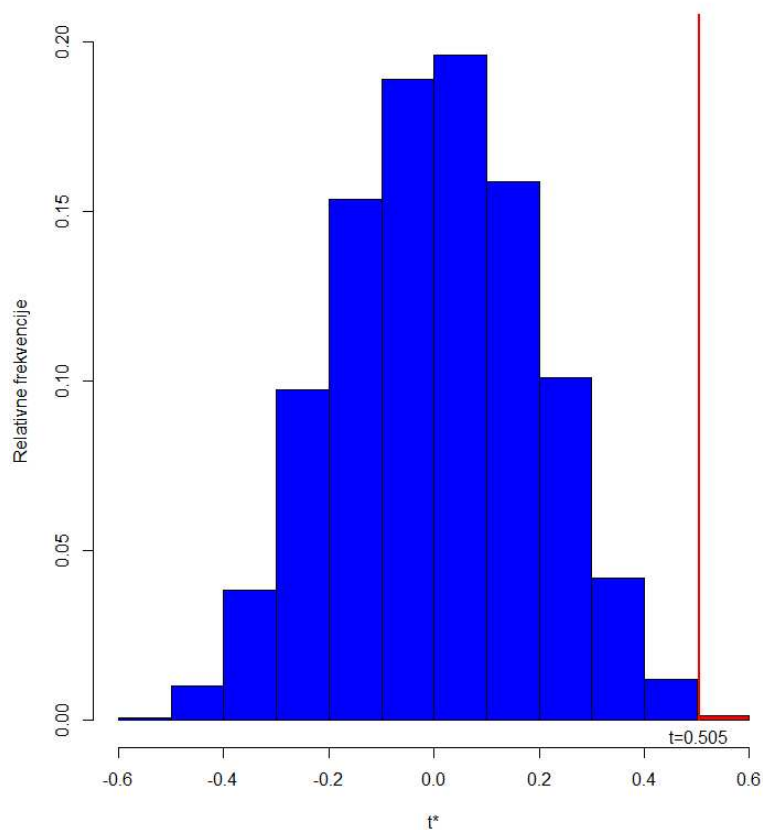
$$p = \frac{\text{broj permutacija takvih da je } t^* \geq t}{n!}.$$

Budući da za  $n = 30$  postoji  $30! \approx 2.7 \cdot 10^{32}$  permutacija vektora  $(y_{p_1}, \dots, y_{p_n})$ , egzaktnu je  $p$ -vrijednost gotovo nemoguće naći pa se umjesto nje koristimo aproksimacijom (2.11). Simuliramo  $R = 9999$  slučajnih permutacija vektora  $(y_1, \dots, y_n)$  i izračunamo koeficijent korelacije  $t_r^*$  između vektora  $(x_{(1)}, \dots, x_{(n)})$  i simulirane slučajne permutacije, za sve  $r \in \{1, 2, \dots, 9999\}$ .

Na Slici 2.5 prikazan je histogram simuliranih vrijednosti  $t^*$ . Vertikalna crvena linija označava opaženu vrijednost testne statistike, a dio označen crvenom bojom površinu histograma koja daje  $p$ -vrijednost. Budući da je 10 simuliranih vrijednosti testne statistike veće ili jednako  $t = 0.505$  imamo da je  $p$ -vrijednost testa  $p \approx p_{mc} = 0.0011$ . Dakle, odbacujemo nultu hipotezu u korist alternative i zaključujemo da postoji pozitivna korelacija između dva promatrana obilježja.

Naravno, u prethodnom smo primjeru za testnu statistiku mogli uzeti Spearmanov koeficijent korelacije ili pak Kendallov  $\tau$  koeficijent jer su obje statistike *dobre* mjere povezanosti. Ovo je jedna od prednosti permutacijskih testova – svejedno je što koristimo kao





Slika 2.5: Histogram simuliranih vrijednosti testne statistike u Primjeru 2.3.1

testnu statistiku dokle god njenu vrijednost možemo izračunati za sve simulirane permutacije.

**Primjer 2.3.2.** U svrhu istraživanja utjecaja vrste čokolade na kardiovaskularno zdravlje čovjeka sudjelovalo je  $n = 12$  slučajno odabranih osoba. Svaka je osoba prvi dan pojela ili 100 grama tamne čokolade ili 200 grama mliječne čokolade (odabrano na slučajan način) te im je izmjeren kapacitet antioksidansa u krvnoj plazmi. Sljedeći dan su osobe koje su prvi dan pojele tamnu čokoladu pojele mliječnu čokoladu i obratno te im je ponovno izmjeren kapacitet antioksidansa u krvnoj plazmi. Podaci su preuzeti iz [8]. Hipoteza je da konzumiranje tamne čokolade rezultira većem kapacitetu antioksidansa u krvnoj plazmi te želimo testirati tu tvrdnju.

Neka je  $X_1, \dots, X_n$  slučajni uzorak iz prve populacije (mjerenja osoba koje su pojele

tamnu čokoladu), a  $Y_1, \dots, Y_n$  slučajni uzorak iz druge populacije (mjerenja osoba koje su pojele mliječnu čokoladu). Neka su  $\mu_1$  i  $\mu_2$  (populacijska) očekivanja kapaciteta antioksidansa u krvnoj plazmi osoba koje su pojele tamnu čokoladu i mliječnu čokoladu, redom. Želimo testirati

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2.$$

Pretpostavimo još da su varijance u dvama populacijama jednake. Neka su  $F_1$  i  $F_2$  funkcije distribucije opažanja (slučajnih varijabli) iz prvog i drugog uzorka, redom. Tada je, zbog jednakog raspršenja,  $F_1(z) = G(z - \mu_1)$  i  $F_2(z) = G(z - \mu_2)$  za neku funkciju distribucije  $G$ . Stoga, ako vrijedi nulta hipoteza, onda su funkcije distribucije  $F_1$  i  $F_2$  jednake  $F_1 \equiv F_2 \equiv F$  i zapravo imamo jedan (združeni) uzorak iz iste populacije, čiji elementi imaju funkciju distribucije  $F$ . Zato je dovoljna statistika za  $F$  vektor uređajnih statistika združenog uzorka, to jest  $S = (Z_{(1)}, \dots, Z_{(2n)})$ , gdje je  $Z_1 = X_1, \dots, Z_n = X_n, Z_{n+1} = Y_1, \dots, Z_{2n} = Y_n$  združeni uzorak.

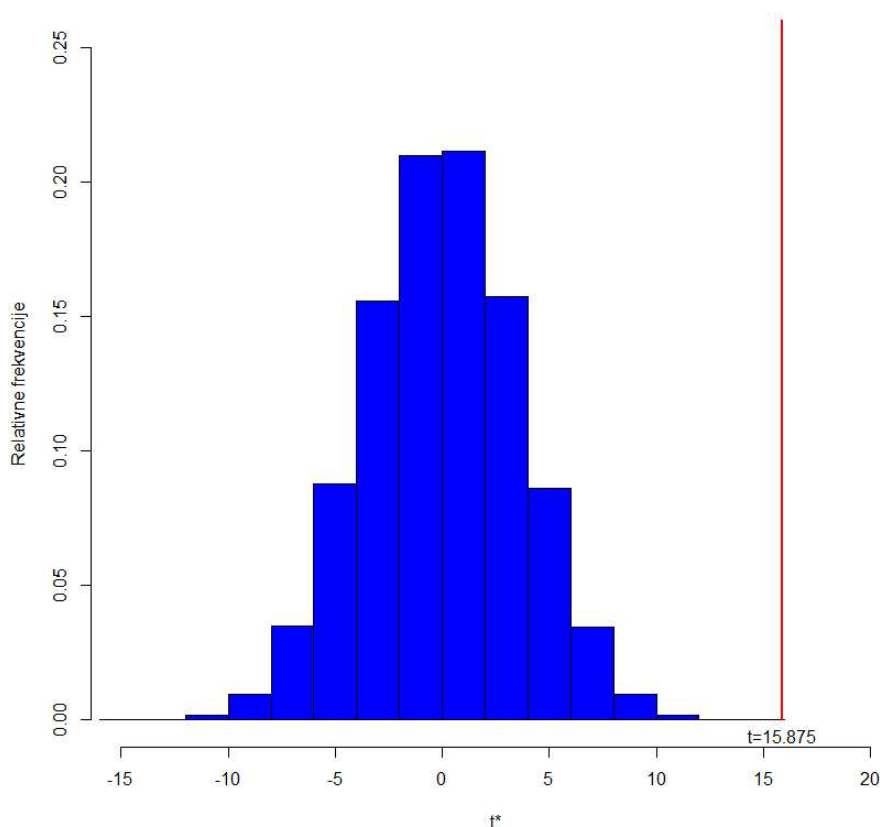
Za testnu statistiku možemo uzeti razliku uzoračkih aritmetičkih sredina, to jest  $T = \bar{X} - \bar{Y}$  jer će tada, zbog forme alternativne hipoteze, velike vrijednosti  $T$  ići u korist alternativnoj hipotezi i aproksimativna  $p$ -vrijednost bit će dana s (2.11). Opažena vrijednost testne statistike je  $t = 15.875$ .

Uvjet  $S = s = (z_1, \dots, z_{2n})$  znači da je združeni uzorak  $(Z_1, \dots, Z_{2n})$  neka permutacija od  $S$ . Prvih  $n$  elemenata te permutacije pripada prvoj populaciji, a zadnjih  $n$  pripada drugoj populaciji. Simuliramo  $R = 9999$  slučajnih permutacija vektora  $(z_1, \dots, z_{2n})$  i izračunamo vrijednost testne statistike  $t_r^*$ , za sve  $r \in \{1, 2, \dots, 9999\}$ .

Na Slici 2.6 prikazan je histogram simuliranih vrijednosti  $t^*$ . Vertikalna crvena linija označava opaženu vrijednost testne statistike. Ta se linija nalazi daleko od desnog ruba histograma jer ne postoji niti jedna simulirana vrijednost testne statistike veća ili jednaka  $t = 15.875$ . Dakle,  $p$ -vrijednost testa je  $p \approx p_{mc} = 0.0001$  pa odbacujemo nultu hipotezu u korist alternative i zaključujemo da je kapacitet antioksidansa u krvnoj plazmi veći nakon konzumiranja tamne čokolade nego nakon konzumiranja mliječne čokolade.

## 2.4 Neparametarski testovi

Permutacijski testovi koje smo obradili u prethodnom poglavlju poseban su slučaj malo šire klase neparametarskih testova. Glavna je razlika ta što je kod permutacijskih testova ponovljeno uzorkovanje bez ponavljanja dok će kod neparametarskih testova to ponovljeno uzorkovanje biti s ponavljanjem. Međutim, permutacijski testovi predstavljali su tip uvjetnih testova, to jest za njihovu  $p$ -vrijednost uzimali smo (2.4). Općenito, neparametarski testovi bit će raznih oblika a njihov oblik određivat će tip problema i sama nulta hipoteza.



Slika 2.6: Histogram simuliranih vrijednosti testne statistike u Primjeru 2.3.2

U prvome poglavlju na samom početku opisali smo primjer korištenja neparametarskog *bootstrapa*, odnosno neparametarskog ponovljenog uzorkovanja. Tamo smo uzorkovali iz empirijske funkcije distribucije  $\hat{F}$  što se pokazalo kao dobra ideja u slučaju neparametarskog modela. Međutim, problem s ovim pristupom kod testiranja hipoteza je taj da pretpostavljamo da je nulta hipoteza točna, odnosno moramo uzorkovati iz  $\hat{F}$  koja zadovoljava nultu hipotezu. To smo već razmatrali kod parametarskih *bootstrap* testova gdje smo uzorkovali iz (parametarske) procjene  $\hat{F}_0$ . Označimo kao i ranije s  $T$  testnu statistiku. Dakle, jednom kad smo prilagodili procjenu  $\hat{F}$  tako da ona zadovoljava nultu hipotezu (nazovimo tu procjenu s  $\hat{F}_0$ ),  $p$ -vrijednost testa možemo računati uz pomoć (2.5), što procjenjujemo s (2.12), gdje su, ponovno,  $t_1^*, \dots, t_R^*$  simulirane vrijednosti testne statistike, a  $R$  je broj simulacija.

**Primjer 2.4.1.** Na slučajnom uzorku  $n = 20$  osoba mjeren je srčani puls u mirovanju. Osobe su kategorizirane u dvije skupine prema tome bave li se nekim sportom ( $X$ ) ili ne ( $Y$ ). U sljedećoj tablici dana su mjerenja, preuzeta iz [7]:

$i$	1	2	3	4	5	6	7	8	9	10	11	12
Trenira ( $x_i$ )	62	72	60	63	75	64	60	52	64	80	68	64
Ne trenira ( $y_i$ )	72	84	66	72	62	84	76	60				

Smatra se da osobe koje se svakodnevno bave nekom fizičkom aktivnošću imaju u prosjeku niži puls nego one koje se ne bave. Želimo testirati tu tvrdnju, to jest želimo testirati

$$H_0 : \mu_2 = \mu_1$$

$$H_1 : \mu_2 > \mu_1$$

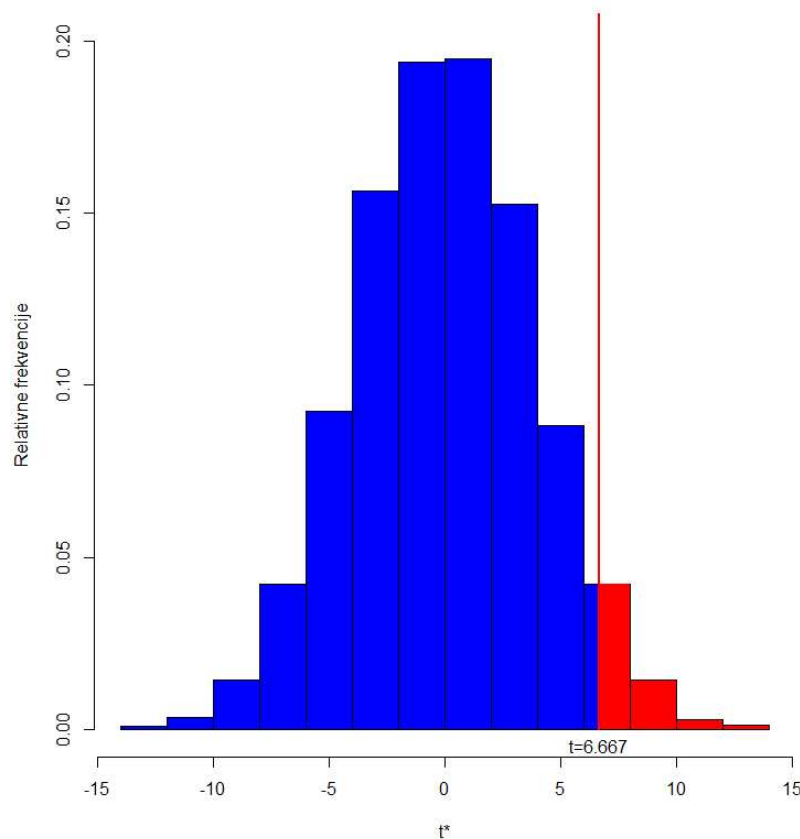
gdje su  $\mu_1$  i  $\mu_2$  očekivani srčani pulsevi osoba koje se, redom, bave i ne bave nekom fizičkom aktivnošću.

Pretpostavimo još da su raspršenja u obje populacije jednaka, to jest da su distribucije obje populacije istog oblika. Uz nultu hipotezu onda imamo jedan (združeni) uzorak iz iste populacije čiji elementi imaju funkciju distribucije  $F$ . Neka je  $Z_1, \dots, Z_n$  združeni uzorak te  $n_1 = 12$  i  $n_2 = 8$  veličine uzoraka iz svake grupe. Dakle, za  $\hat{F}_0$  možemo uzeti empirijsku funkciju distribucije združenog uzorka  $Z_1, \dots, Z_n$ , koja zadovoljava nultu hipotezu. Simuliranje iz te razdiobe svodi se na uzimanje slučajnog uzorka duljine  $n = 20$  s ponavljanjem iz skupa opažanja  $z_1, \dots, z_n$ . Prvih  $n_1 = 12$  elemenata tada pripada prvom uzorku, a zadnjih  $n_2 = 8$  elemenata drugom uzorku.

Neka je  $T = \bar{Y} - \bar{X}$  testna statistika. Velike vrijednosti  $T$  tada idu u korist alternativnoj hipotezi pa je (2.5) zaista  $p$ -vrijednost ovog testa, koju procjenjujemo s (2.12). Opažena vrijednost testne statistike je  $t = 6.667$ . Neka je  $R = 9999$  broj simulacija. Simuliramo (nezavisne) uzorke iste duljine kako smo ranije opisali i izračunamo vrijednosti testne statistike  $t_r^*$ , za sve  $r \in \{1, 2, \dots, 9999\}$ .

Na Slici 2.7 prikazan je histogram simuliranih vrijednosti  $t^*$ . Vertikalna crvena linija označava opaženu vrijednost testne statistike, a dio označen crvenom bojom površinu histograma koja daje  $p$ -vrijednost. Budući da je 423 simuliranih vrijednosti testne statistike veće ili jednako  $t = 6.667$ ,  $p$ -vrijednost testa je  $p \approx p_{boot} = 0.0424$ . Ovisno o razini značajnosti  $\alpha$  donosimo zaključak testa. Na primjer, na razini značajnosti  $\alpha = 0.05$  odbacujemo nultu hipotezu, ali na razini značajnosti  $\alpha = 0.01$  ju ne odbacujemo.

U zadnja dva primjera pretpostavljali smo da su oblici dvaju razdioba jednaki, to jest da su im raspršenja ista. Pokažimo da nije nužno restringirati se na tu pretpostavku, to jest pokažimo kako provesti neparametarski test bez te dodatne pretpostavke.



Slika 2.7: Histogram simuliranih vrijednosti testne statistike u Primjeru 2.4.1

**Primjer 2.4.2.** Mjerena je razina određene supstance (klorfeniramin maleat) u nekim tabletama proizvedenim u 7 različitih laboratorija. Na slučajan je način iz svakog laboratorija odabrano  $n_i = 10$  tableta i izmjerena im je razina te supstance, za sve  $i \in \{1, 2, \dots, 7\}$  tako da je ukupan broj mjerenja  $n = 70$ . Podaci su preuzeti iz [9]. Želimo testirati je li očekivana razina te supstance u tabletama ista za sve laboratorije, to jest želimo testirati

$$H_0 : \mu_1 = \dots = \mu_7$$

$$H_1 : \text{postoje } k \text{ i } l \text{ takvi da je } \mu_k \neq \mu_l,$$

gdje su  $\mu_i$  očekivane razine supstance tableta iz  $i$ -tog laboratorija,  $i \in \{1, 2, \dots, 7\}$ .

Ovdje ne pretpostavljamo da su raspršenja u svih 7 populacija (laboratorija) jednaka

te stoga modeliramo opažanja pomoću

$$X_{i,j} = \mu_i + \sigma_i \varepsilon_{i,j}, \quad i \in \{1, 2, \dots, 7\}, j \in \{1, 2, \dots, n_i\},$$

gdje su  $\varepsilon_{i,j}$  slučajne pogreške (slučajne varijable) s istom funkcijom distribucije  $G$ .

Za testnu statistiku možemo uzeti

$$T = \sum_{i=1}^7 n_i \cdot \left( \frac{\bar{X}_i - \bar{X}}{S_i} \right)^2$$

jer je to težinska suma takozvanih kvadriranih  $z$  – scoreova (odražava ukupnu udaljenost procjena  $\bar{X}_i$  od nulte vrijednosti  $\bar{X}$  u odnosu na raspršenje podataka). Ova će testna statistika odbacivati nultu hipotezu za velike vrijednosti  $T$  pa ponovno  $p$ -vrijednost testa možemo aproksimirati koristeći (2.12). U gornjoj su formuli  $\bar{X}_i$  i  $S_i$  uzoračka aritmetička sredina i uzoračka standardna devijacija, redom, u  $i$ -toj grupi, a  $\bar{X}$  je aritmetička sredina svih (združenih)  $n = 70$  podataka. Opažena vrijednost testne statistike je  $t = 105.03$ .

Varijance  $\sigma_i^2$  procjenjujemo s

$$\hat{\sigma}_i^2 = \frac{n_i - 1}{n_i} \cdot S_i^2 + (\bar{X}_i - \bar{X})^2.$$

Sada su opažene slučajne pogreške  $e_{i,j}$  uz nultu hipotezu dane s

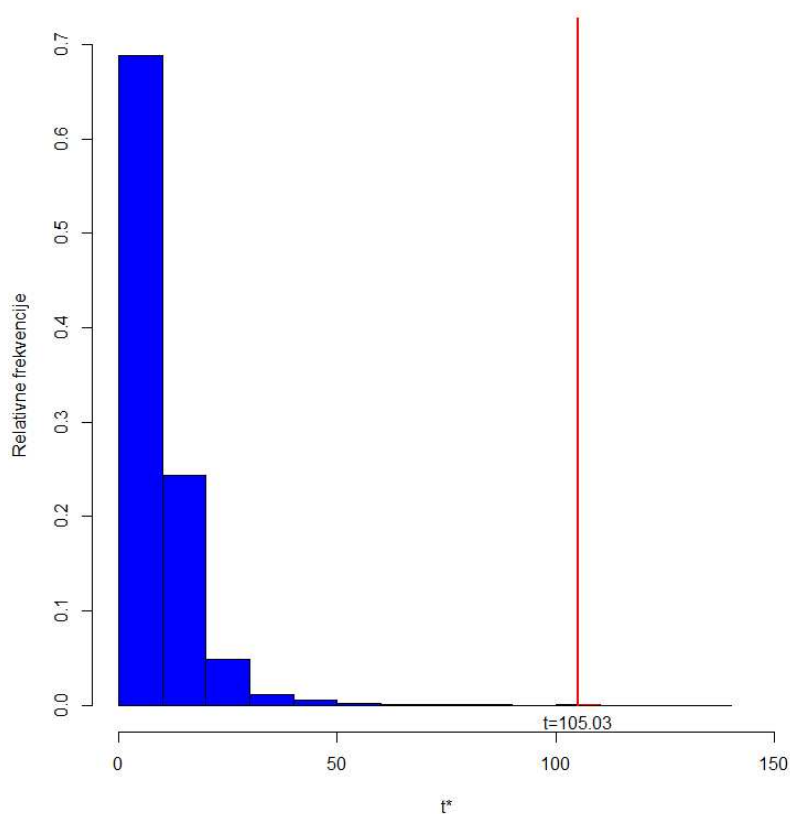
$$e_{i,j} = \frac{x_{i,j} - \bar{x}}{\hat{\sigma}_i}.$$

Stoga, za svaki  $i \in \{1, 2, \dots, 7\}$  simuliramo s ponavljanjem vrijednosti  $n_i$  slučajnih pogreški iz skupa združenih opažanja slučajnih pogrešaka  $e_{i,j}$  i označimo simulacije s  $\varepsilon_{i,j}^*$ . Stavimo

$$x_{i,j}^* = \bar{x} + \hat{\sigma}_i \varepsilon_{i,j}^*.$$

Iz tih vrijednosti izračunamo vrijednost testne statistike  $t^*$  i to ponovimo  $R = 9999$  puta tako da dobijemo simulirane vrijednosti testne statistike  $t_1^*, \dots, t_{9999}^*$ . Budući da je samo 3 simuliranih vrijednosti testne statistike veće ili jednako  $t = 105.03$ ,  $p$ -vrijednost testa je  $p \approx p_{boot} = 0.0004$  te stoga odbacujemo nultu hipotezu u korist alternativnoj hipotezi. Na Slici 2.8 prikazan je histogram simuliranih vrijednosti  $t^*$ . Vertikalna crvena linija označava opaženu vrijednost testne statistike, a dio označen crvenom bojom površinu histograma koja daje  $p$ -vrijednost.

U nekim je situacijama teško ili nemoguće prilagoditi empirijsku funkciju distribucije  $\hat{F}$  tako da ona zadovoljava nultu hipotezu. Pokazat ćemo jedan način na koji možemo u općenitim situacijama pronaći funkciju distribucije  $\hat{F}_0$  koja zadovoljava nultu hipotezu.



Slika 2.8: Histogram simuliranih vrijednosti testne statistike u Primjeru 2.4.2

Pretpostavimo da imamo  $k$  (nepoznatih) funkcija distribucija  $F_1, \dots, F_k$  i da je nulta hipoteza oblika  $H_0 : h(F_1, \dots, F_k) = 0$ . Neka su procjene tih funkcija distribucija restringirane na skup opaženih vrijednosti uzorka  $x_{i,1}, \dots, x_{i,n_i}$ , za sve  $i \in \{1, 2, \dots, k\}$ , gdje je  $n_i$  veličina  $i$ -tog uzorka. Procjena za  $F_i$  tada inducira vjerojatnosti  $p_i = (p_{i,1}, \dots, p_{i,n_i})$  opažanja vrijednosti  $x_{i,1}, \dots, x_{i,n_i}$ . S druge strane, empirijska funkcija distribucije  $\hat{F}_i$  inducira vjerojatnosti  $\hat{p}_i = (\frac{1}{n_i}, \dots, \frac{1}{n_i})$ . Sada na neki način želimo *izmjeriti* koliko se  $p_i$  i  $\hat{p}_i$  razlikuju. Neka je  $d(p_i, \hat{p}_i)$  ta mjera udaljenosti dana s

$$d(p_i, \hat{p}_i) = \sum_{j=1}^{n_i} p_{i,j} \cdot \ln \left( \frac{p_{i,j}}{\hat{p}_{i,j}} \right). \quad (2.18)$$

Ta je udaljenost minimalna moguća kad nemamo nikakvih restrikcija (specificiranih nul-tom hipotezom) pa je tada  $p_i = \hat{p}_i$  za sve  $i$ . Međutim, uz nultu hipotezu  $p_i$  mora biti takav

da vrijedi  $h(p_1, \dots, p_k) = 0$  te, jasno, suma elemenata vektora  $p_i$  mora biti 1, za sve  $i$ . Dakle, uz te dodatne uvjete minimiziramo Lagrangeovu funkciju

$$\sum_{i=1}^k d(p_i, \hat{p}_i) - \lambda h(p_1, \dots, p_k) - \sum_{i=1}^k \mu_i \cdot \left( \sum_{j=1}^{n_i} p_{i,j} - 1 \right). \quad (2.19)$$

Rješenja ovog problema minimizacije označit ćemo s  $\hat{p}_{i0}$ ,  $i \in \{1, 2, \dots, k\}$ . Rješenje  $\hat{p}_{i0,j}$  zapravo će predstavljati vjerojatnost s kojom će element  $x_{i,j}$  iz uzorka biti odabran kod simulacija. Pokažimo na primjeru koristeći podatke iz Primjera 2.4.1 kako u slučaju  $k = 2$  i testiranja jednakosti očekivanja možemo provesti ovakav test.

**Primjer 2.4.3.** Za podatke iz Primjera 2.4.1, koristeći iste oznake kao i tamo, želimo testirati

$$\begin{aligned} H_0 : \mathcal{M}_2 &= \mathcal{M}_1 \\ H_1 : \mathcal{M}_2 &> \mathcal{M}_1. \end{aligned}$$

Klasične oznake za očekivanja zamijenili smo s  $\mathcal{M}_1$  i  $\mathcal{M}_2$  da se oznake ne bi podudarale s oznakama parametara  $\mu_1$  i  $\mu_2$  u Lagrangeovoj jednadžbi (2.19). Za testnu statistiku uzimamo kao u Primjeru 2.4.1, to jest  $T = \bar{Y} - \bar{X}$ . Opažena vrijednost testne statistike je  $t = 6.667$ . Budući da su očekivanja dana s

$$\mathcal{M}_1 = \sum_{j=1}^{n_1} x_j p_{1,j} \quad i \quad \mathcal{M}_2 = \sum_{j=1}^{n_2} y_j p_{2,j}$$

imamo da je (2.19) jednako

$$\sum_{i=1}^2 \sum_{j=1}^{n_i} p_{i,j} \cdot \ln \left( \frac{p_{i,j}}{\hat{p}_{i,j}} \right) - \lambda \cdot \left( \sum_{j=1}^{n_1} x_j p_{1,j} - \sum_{j=1}^{n_2} y_j p_{2,j} \right) - \sum_{i=1}^2 \mu_i \cdot \left( \sum_{j=1}^{n_i} p_{i,j} - 1 \right).$$

Parcijalno derivirajući ovaj izraz po  $p_{i,j}$ , za sve  $i \in \{1, 2\}$  i  $j \in \{1, 2, \dots, n_i\}$  dobivamo da mora vrijediti

$$1 + \ln(p_{1,j}) - \mu_1 - \lambda x_j = 0 \quad i \quad 1 + \ln(p_{2,j}) - \mu_2 + \lambda y_j = 0$$

odakle korištenjem početnih restrikcija dobivamo da su rješenja

$$\hat{p}_{10,j} = \frac{e^{\lambda x_j}}{\sum_{k=1}^{n_1} e^{\lambda x_k}} \quad i \quad \hat{p}_{20,j} = \frac{e^{-\lambda y_j}}{\sum_{k=1}^{n_2} e^{-\lambda y_k}}. \quad (2.20)$$

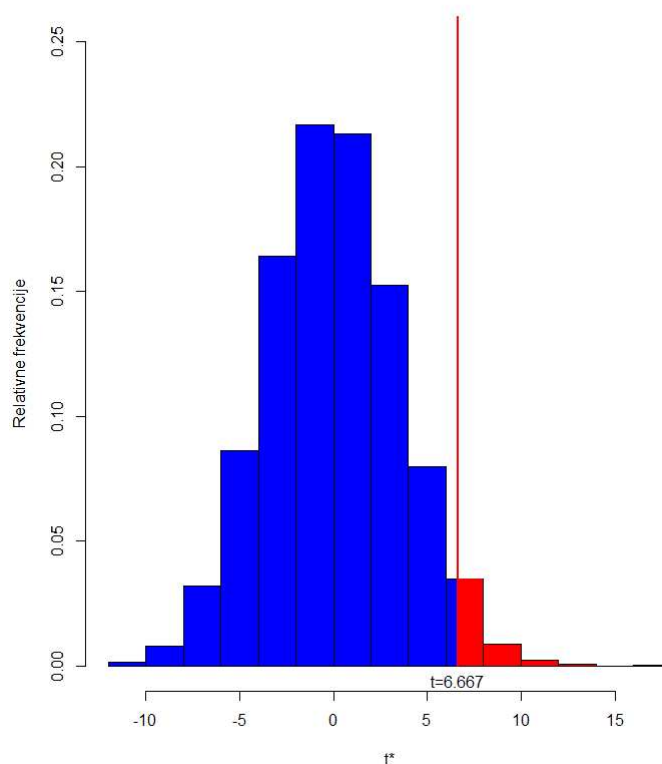
Vrijednost  $\lambda$  također dobivamo koristeći početne restrikcije koje se mogu izraziti kao

$$\frac{\sum_{k=1}^{n_1} x_k \cdot e^{\lambda x_k}}{\sum_{k=1}^{n_1} e^{\lambda x_k}} = \frac{\sum_{k=1}^{n_2} y_k \cdot e^{-\lambda y_k}}{\sum_{k=1}^{n_2} e^{-\lambda y_k}}.$$



Ova se jednadžba rješava numerički te u našem slučaju rješenje iznosi  $\lambda = 0.054319$ . Supstituiranjem te vrijednosti u (2.20) dobivamo rješenja jednadžbe (2.19).

Neka je  $R = 9999$  broj simulacija. U  $r$ -tom koraku simuliramo uzorak  $x_1^*, \dots, x_{n_1}^*$  tako da iz opažanja  $x_1, \dots, x_{n_1}$  na slučajan način odaberemo  $n_1$  elemenata s ponavljanjem gdje je element  $x_i$  odabran s vjerojatnosti  $\hat{p}_{10,i}$ . Nadalje, simuliramo uzorak  $y_1^*, \dots, y_{n_2}^*$  tako da iz opažanja  $y_1, \dots, y_{n_2}$  na slučajan način odaberemo  $n_2$  elemenata s ponavljanjem gdje je element  $y_i$  odabran s vjerojatnosti  $\hat{p}_{20,i}$ . Iz simuliranih uzoraka izračunamo vrijednost testne statistike  $t_r^*$ . Nakon što smo to ponovili  $R = 9999$  puta,  $p$ -vrijednost testa računamo koristeći (2.12). Budući da je 320 simuliranih vrijednosti testne statistike veće ili jednako  $t = 6.667$ ,  $p$ -vrijednost testa je  $p \approx p_{boot} = 0.0321$ . Usporedimo s Primjerom 2.4.1 gdje je  $p$ -vrijednost testa bila  $p = 0.0424$ . Na Slici 2.9 prikazan je histogram simuliranih vrijednosti  $t^*$ . Vertikalna crvena linija označava opaženu vrijednost testne statistike, a dio označen crvenom bojom površinu histograma koja daje  $p$ -vrijednost.



Slika 2.9: Histogram simuliranih vrijednosti testne statistike u Primjeru 2.4.3

Primijetimo da u gornjem primjeru nismo pretpostavljali da su dvije distribucije istog oblika, odnosno da imaju istu raspršenost. Dakle, ova metoda može se upotrebljavati u vrlo općenitim slučajevima i bez strogih pretpostavki.

### Metode studentiziranog *bootstrapa*

Ponekad je korisno studentizirati opažanja u uzorku i provesti test koji koristi studentizirana opažanja. Naime, u praksi je čest slučaj da studentizirani uzorak više ne ovisi ni o kakvim nepoznatim parametrima pa u tom slučaju možemo primijeniti pivotni test kojeg smo opisali u uvodnom dijelu ovog poglavlja. Neka je  $\theta \in \mathbb{R}$  nepoznati parametar i pretpostavimo da želimo testirati hipoteze

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta > \theta_0$$

gdje je  $\theta_0 \in \mathbb{R}$  pretpostavljena vrijednost. Neka je  $T$  procjenitelj za  $\theta$  s procijenjenom varijancom  $V$ . Neka je studentizirana statistika  $Z = \frac{T - \theta}{\sqrt{V}}$  (barem aproksimativno) pivotna, to jest razdioba joj je ista za svaki  $F$ , odnosno za svaku vrijednost  $\theta$ . Neka je još  $z_0 = \frac{t - \theta_0}{\sqrt{v}}$  opažena vrijednost studentizirane testne statistike.  $P$ -vrijednost tog testa dana je s (2.6) bez obzira na  $F$ , a aproksimira se pomoću (2.12) uz očite zamjene oznaka. Za procjenu varijance procjenitelja  $V$  možemo uzeti, na primjer, procjenu pomoću neparametarske delta-metode, opisanu u prethodnom poglavlju. Naravno, studentizirana testna statistika može se koristiti iako ona možda nije pivotna. U tom je slučaju potrebno naći procjenu funkcije distribucije  $\hat{F}_0$  koja zadovoljava nultu hipotezu te simulirati iz te procjene.

**Primjer 2.4.4.** *Za podatke iz Primjera 2.4.1, koristeći iste oznake kao i tamo, želimo testirati*

$$H_0 : \mu_2 = \mu_1$$

$$H_1 : \mu_2 > \mu_1.$$

*Za testnu statistiku uzimimo*

$$Z = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_Y^2}{n_2} + \frac{S_X^2}{n_1}}}$$

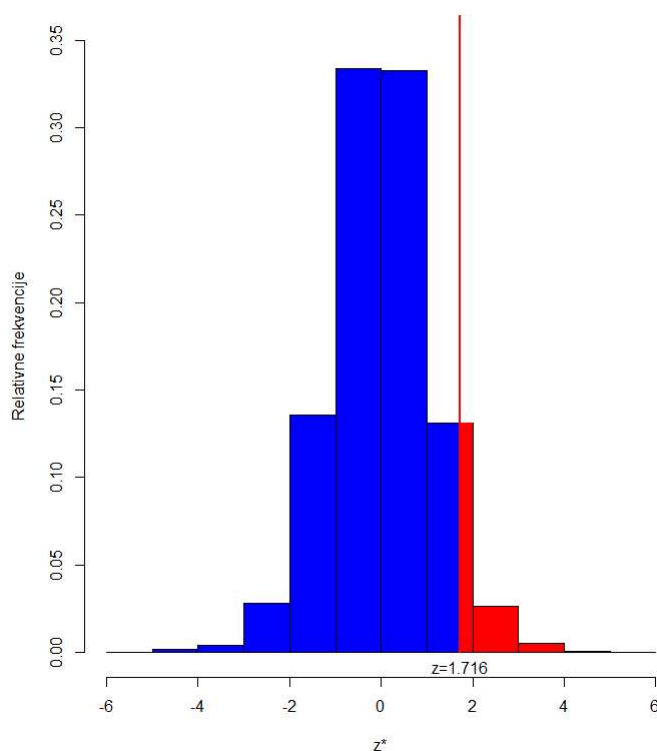
*budući da ima smisla pretpostaviti da je ona aproksimativno pivotna (naime, u slučaju normalnosti dvaju uzoraka ova statistika ima Studentovu  $t$ -razdiobu koja jest pivotna). Opažena vrijednost testne statistike je  $z = 1.716$ .*

Neka je  $R = 9999$  broj simulacija. U  $r$ -tom koraku simuliramo uzorke  $x_1^*, \dots, x_{n_1}^*$  i  $y_1^*, \dots, y_{n_2}^*$  tako da na slučajan način odaberemo elemente iz početnih uzoraka s ponavljanjem i izračunamo simuliranu vrijednost testne statistike

$$z_r^* = \frac{\bar{y}^* - \bar{x}^* - (\bar{y} - \bar{x})}{\sqrt{\frac{s_y^{*2}}{n_2} + \frac{s_x^{*2}}{n_1}}},$$

za sve  $r \in \{1, 2, \dots, 9999\}$ .

Naše simulacije dale su 524 vrijednosti testne statistike veće ili jednake  $z = 1.716$  pa je  $p$ -vrijednost ovog testa  $p \approx p_{boot} = 0.0525$ . Usporedimo s  $p$ -vrijednostima iz Primjera 2.4.1 ( $p \approx 0.0424$ ) i Primjera 2.4.3 ( $p \approx 0.0321$ ). Na Slici 2.10 prikazan je histogram simuliranih vrijednosti  $z^*$ . Vertikalna crvena linija označava opaženu vrijednost testne statistike, a dio označen crvenom bojom površinu histograma koja daje  $p$ -vrijednost.



Slika 2.10: Histogram simuliranih vrijednosti testne statistike u Primjeru 2.4.4

U slučaju da je  $\theta \in \mathbb{R}^p$  vektor parametara,  $p \in \mathbb{N}$ , tada za studentiziranu testnu statistiku uzimamo

$$Z = (T - \theta)^t V^{-1}(T - \theta),$$

gdje je  $T$  procjenitelj od  $\theta$ , a  $V$  njegova kovarijacijska matrica. Međutim, ova se testna statistika može koristiti u slučaju da je alternativna hipoteza oblika  $H_1 : \theta \neq \theta_0$  (dakle, za dvostrani test).

## 2.5 Metoda korekcije $p$ -vrijednosti

U testovima koje smo dosad obrađivali  $p$ -vrijednost nam je gotovo uvijek bila dana s  $p = \mathbb{P}^*(T^* \geq t \mid \hat{F}_0)$ , gdje je  $\hat{F}_0$  procjena funkcije distribucije koja zadovoljava nultu hipotezu,  $T^*$  simulacija testne statistike,  $t$  opažena vrijednost testne statistike, a  $\mathbb{P}^*$  vjerojatnost u odnosu na razdiobu dobivenu simuliranjem. Ukoliko je nulta hipoteza točna,  $p$ -vrijednost bi trebala biti uniformno distribuirana na segmentu  $[0, 1]$ , to jest trebalo bi vrijediti (2.2). Ta je tvrdnja istinita za permutacijske testove, no za ostale testove ne mora biti istinita. U ovome potpoglavlju dat ćemo metodu za korekciju  $p$ -vrijednosti koja se može koristiti u bilo kojem testu kojeg smo ranije opisali.

Ideja je dobivenu  $p$ -vrijednost testa shvatiti kao opaženu vrijednost (realizaciju) slučajne varijable  $P$  ( $p$ -vrijednosti). Razdiobu slučajne varijable  $P$ , uz nultu hipotezu, procjenjujemo simulacijama to jest ponovljenim uzorkovanjem. Definiramo korigiranu  $p$ -vrijednost kao

$$p_{\text{adj}} = \mathbb{P}(P^* \leq p \mid \hat{F}_0), \quad (2.21)$$

gdje je  $p$  opažena  $p$ -vrijednost, a  $P^*$  simulacija  $p$ -vrijednosti.

Ovakav pristup ustvari korigira moguću pristranost u inicijalnoj opaženoj  $p$ -vrijednosti. Označimo  $p$ -vrijednost s  $p = p_0(\hat{F})$ , gdje je  $p_0$  neka funkcija. Pretpostavimo da je nulta hipoteza točna i da je  $F_0$  funkcija distribucije koja zadovoljava nultu hipotezu. Nul-razdioba slučajne varijable  $P$  je tada

$$G(s, F_0) = \mathbb{P}(p_0(\hat{F}) \leq s \mid F_0). \quad (2.22)$$

Naravno, za  $s = \alpha$  ovo ima interpretaciju razine značajnosti testa. Iz (2.22) slijedi

$$\mathbb{P}(G(p_0(\hat{F}), F_0) \leq \alpha \mid F_0) = \alpha$$

pa  $G(p_0(\hat{F}), F_0)$  predstavlja korigiranu  $p$ -vrijednost. Procjenu funkcije distribucije  $G$  možemo dobiti tako da u (2.22) stavimo  $\hat{F}_0$  na mjesto  $F_0$ . Time dobijemo da je procjena jednaka

$$\mathbb{P}^*(p_0(\hat{F}^*) \leq s \mid \hat{F}_0).$$

Ponovno, stavimo  $s = p_0(\hat{F})$  pa dobivamo da je

$$G(p_0(\hat{F}), F_0) \approx \mathbb{P}^*(p_0(\hat{F}^*) \leq p_0(\hat{F}) \mid \hat{F}_0).$$

Dakle, ako je  $p_0(\hat{F}) = p$  dobivamo korigiranu  $p$ -vrijednost

$$p_{\text{adj}} = \mathbb{P}^*(p_0(\hat{F}^*) \leq p \mid \hat{F}_0), \quad (2.23)$$

što je ustvari isto kao i (2.21).

U (2.23) uvjet  $\hat{F}_0$  znači da uzorkujemo iz razdiobe određene funkcijom distribucije (to jest, nekom procjenom)  $\hat{F}_0$  koja još i zadovoljava nultu hipotezu. Oznaka  $\hat{F}^*$  predstavlja empirijsku funkciju distribucije uzorka simuliranog iz  $\hat{F}_0$ . Dakle, koristit ćemo metodu ponovljenog uzorkovanja ne bismo li procijenili korigiranu  $p$ -vrijednost. Neka je, kao i ranije,  $R$  broj simulacija (to jest, broj nezavisnih uzoraka koje smo simulirali) pod nultom hipotezom, to jest iz  $\hat{F}_0$ . Neka su  $t_1^*, \dots, t_R^*$  dobivene simulirane vrijednosti testne statistike. Kao što smo rekli,  $r$ -ti uzorak ima empirijsku funkciju distribucije  $\hat{F}_r^*$  koju prilagođavamo tako da zadovoljava nultu hipotezu. Označimo tu prilagođenu empirijsku funkciju distribucije s  $\hat{F}_{r0}^*$ . Neka je sada  $M$  broj simulacija iz te razdiobe (određene s  $\hat{F}_{r0}^*$ ). Simuliramo  $M$  puta iz te razdiobe i izračunamo vrijednosti testne statistike  $t_{r,m}^*$ , za sve  $m \in \{1, 2, \dots, M\}$ . Definiramo

$$p_r^* = \frac{1 + k(\{r; t_{r,m}^* \geq t_r^*\})}{M + 1}$$

pa je Monte Carlo aproksimacija korigirane  $p$ -vrijednosti dana s

$$p_{\text{adj, mc}} = \frac{1 + k(\{r; p_r^* \leq p\})}{R + 1}. \quad (2.24)$$

Zbog možda malo kompliciranog postupka poželjno je sažeti ga u jedan kraći algoritam:

**Algoritam 2.5.1.** Za  $r \in \{1, 2, \dots, R\}$  radi:

1. Generiraj  $x_1^*, \dots, x_n^*$  iz  $\hat{F}_0$  i iz njih izračunaj  $t_r^*$ .
2. Nađi procjenu funkcije distribucije  $\hat{F}_{r0}^*$  gornjeg uzorka koja zadovoljava  $H_0$ .
3. Za  $m \in \{1, 2, \dots, M\}$  radi :
  - 3.1. Generiraj  $x_1^{**}, \dots, x_n^{**}$  iz  $\hat{F}_{r0}^*$ .
  - 3.2. Izračunaj vrijednost testne statistike  $t_{r,m}^*$ .
4. Izračunaj  $p_r^*$ .

Izračunaj  $p_{\text{adj, mc}}$ .

Nećemo diskutirati koliki bi trebao biti broj simulacija  $M$ . Razmatranja u [2] pokazuju da je  $M = 249$  sasvim dovoljan broj simulacija pa ćemo mi za naše potrebe uzimati taj  $M$ . Međutim, smanjit ćemo broj simulacija  $R$  na  $R = 999$ .

**Primjer 2.5.2.** Za podatke i hipoteze iz Primjera 2.4.1 koristili smo gornji algoritam kako bismo izračunali korigirane  $p$ -vrijednosti za testove iz Primjera 2.4.1, Primjera 2.4.3 i Primjera 2.4.4. Za traženje simuliranih vrijednosti testnih statistika kao i  $p$ -vrijednosti koristili smo iste metode opisane u tim primjerima.

Za Primjer 2.4.1 s jednostavnim neparametarskim bootstrap testom u koraku 3.1. simulirali smo  $z_1^{**}, \dots, z_n^{**}$  s ponavljanjem iz simuliranog uzorka  $z_1^*, \dots, z_n^*$  (to je bila simulacija iz združenog uzorka). U koraku 3.2. računali smo vrijednost testne statistike kao  $t_{r,m}^* = \overline{y^{**}} - \overline{x^{**}}$ . Dobivena korigirana  $p$ -vrijednost je  $p_{adj, mc} = 0.033$  naspram originalne  $p$ -vrijednosti koja je iznosila  $p_{boot} = 0.038$  (u ponovljenom testu).

Za Primjer 2.4.3 s neparametarskim bootstrap testom s korigiranim vjerojatnostima slučajnog odabiranja elemenata uzorka u svakom ( $r$ -tom) koraku trebamo izračunati vrijednost  $\lambda$  iz simuliranih uzoraka  $x_1^*, \dots, x_{n_1}^*$  i  $y_1^*, \dots, y_{n_2}^*$  te vjerojatnosti iz (2.20). S tim vjerojatnostima onda simuliramo  $x_1^{**}, \dots, x_{n_1}^{**}$  i  $y_1^{**}, \dots, y_{n_2}^{**}$  u koraku 3.1. te za vrijednost testne statistike stavljamo  $t_{r,m}^* = \overline{y^{**}} - \overline{x^{**}}$ . Dobivena korigirana  $p$ -vrijednost je  $p_{adj, mc} = 0.04$  naspram originalne  $p$ -vrijednosti koja je iznosila  $p_{boot} = 0.035$  (u ponovljenom testu).

Za Primjer 2.4.4 sa studentiziranom testnom statistikom u koraku 3.1. gornjeg algoritma simulirali smo  $x_1^{**}, \dots, x_{n_1}^{**}$  s ponavljanjem iz simuliranog uzorka  $x_1^*, \dots, x_{n_1}^*$  te  $y_1^{**}, \dots, y_{n_2}^{**}$  s ponavljanjem iz simuliranog uzorka  $y_1^*, \dots, y_{n_2}^*$ . U koraku 3.2. računali smo vrijednost testne statistike kao

$$z_{r,m}^* = \frac{\overline{y^{**}} - \overline{x^{**}} - (\overline{y^*} - \overline{x^*})}{\sqrt{\frac{s_y^{**2}}{8} + \frac{s_x^{**2}}{12}}}.$$

Dobivena korigirana  $p$ -vrijednost je  $p_{adj, mc} = 0.037$  naspram originalne  $p$ -vrijednosti koja je iznosila  $p_{boot} = 0.049$  (u ponovljenom testu).

# Poglavlje 3

## Pouzdana intervali

### 3.1 Uvod

Ponekad umjesto testiranja samo jedne moguće vrijednosti nepoznatog parametra  $\theta$  želimo znati koje su *najvjerojatnije* vrijednosti tog parametra. Drugim riječima, željeli bismo znati skup svih vrijednosti za koji smo *vrlo sigurni* da se u njemu nalazi prava vrijednost parametra  $\theta$ . Samo procijeniti parametar  $\theta$  nekom statistikom  $T$  nije dovoljno. Čak i ako je taj procjenitelj  $T$  *dobar* (na primjer, UMVUE, to jest nepristrani procjenitelj uniformno minimalne varijance), uvijek postoji neka greška u procjeni. Pouzdani intervali nastoje ograničiti tu pogrešku što je više moguće, do na neku zadanu razinu sigurnosti. Ova nam interpretacija pomaže da za neki  $\varepsilon \in [0, 1]$  formalno definiramo  $\varepsilon \cdot 100\%$  pouzdani interval za  $\theta$  kao (slučajni) skup  $C_\varepsilon(X_1, \dots, X_n)$  takav da je

$$\mathbb{P}(\theta \in C_\varepsilon(X_1, \dots, X_n)) = \varepsilon, \quad (3.1)$$

gdje je  $X_1, \dots, X_n$  slučajni uzorak na temelju kojeg formiramo pouzdani interval. Naravno, pretpostavljamo da ne postoje nikakvi drugi nepoznati parametri, odnosno da skup  $C_\varepsilon(X_1, \dots, X_n)$  ovisi samo o tom slučajnom uzorku. Definicija se poklapa s našom interpretacijom s početka – ukoliko formiramo velik broj pouzdanih intervala (baziranih na nezavisnim slučajnim uzorcima iste duljine  $n$ ), otprilike njih  $\varepsilon \cdot 100\%$  sadržavat će pravu (nepoznatu) vrijednost parametra  $\theta$ .

Važna karakteristika pouzdanog intervala je da ga zaista definiramo kao *interval*. Primijetimo da skup  $C_\varepsilon$  zadan kao u definiciji (3.1) ne mora nužno biti interval, to jest može biti raznih *oblika*. Međutim, mi ćemo htjeti da to zaista bude interval i tako ćemo ga konstruirati. Dakle, pouzdani interval bit će (slučajni) interval oblika  $[\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha}]$  gdje su njegove granice određene jednakostima

$$\mathbb{P}(\theta < \hat{\theta}_\alpha) = \mathbb{P}(\theta > \hat{\theta}_{1-\alpha}) = \alpha,$$

za neki  $\alpha \in [0, \frac{1}{2}]$ . Iz te jednakosti slijedi

$$\mathbb{P}(\theta \in [\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha}]) = 1 - 2\alpha$$

pa kažemo da je *pokrivenost* tog pouzdanog intervala  $1 - 2\alpha$  ili da je to  $(1 - 2\alpha) \cdot 100\%$  pouzdani interval za  $\theta$ .

Najčešće korištene vrijednosti  $\alpha$  kod konstrukcije pouzdanih intervala su  $\alpha = 0.10$ ,  $\alpha = 0.05$ ,  $\alpha = 0.025$  i  $\alpha = 0.005$ , to jest najčešće korišteni pouzdani intervali su 80%, 90%, 95% i 99% pouzdani intervali. U nekom konkretnom problemu poželjno je konstruirati ih sve (a ne samo jedan) jer nam to daje bolju ideju o procjeni i pogrešci procjene za  $\theta$ .

Diskusija koju smo dosad proveli vrijedi ukoliko, osim  $\theta$ , nema drugih nepoznatih parametara. Na primjer, (3.1) će egzaktno vrijediti ako je procjenitelj  $T$  (ili njegova standardizirana statistika) pivotna statistika. Međutim, u praksi često nemamo takvu situaciju. Cilj je u ovome poglavlju pokazati kako iskoristiti metode ponovljenog uzorkovanja i za konstrukciju (aproksimativnih) *bootstrap* pouzdanih intervala u slučajevima kad egzaktne metode ne mogu dati rješenje. Dva osnovna tipa *bootstrap* pouzdanih intervala su osnovni i percentilni pouzdani intervali koje obrađujemo u sljedeća dva potpoglavlja.

## 3.2 Osnovni pouzdani intervali

Spomenuli smo u uvodnom dijelu da se pouzdani intervali konstruiraju zbog nesigurnosti (pogreške) u procjeni parametra  $\theta$  nekom statistikom  $T$ . Zato je dobra ideja taj pouzdani interval konstruirati tako da bude *okolina* procjenitelja  $T$ . Postupak konstrukcije razlikuje se za parametarske i neparametarske modele pa u ovom potpoglavlju razmatramo ta dva slučaja posebno.

### Parametarski modeli

Pretpostavimo da je  $\theta$  nepoznati parametar, a  $T$  statistika koja ga procjenjuje (i pretpostavimo još da je  $T$  apsolutno neprekidna slučajna varijabla). Na primjer, u parametarskom slučaju  $T$  može biti procjenitelj metodom maksimalne vjerodostojnosti (MLE). Neka je  $1 - 2\alpha$  željena pokrivenost pouzdanog intervala. Označimo s  $a_p$   $p$ -ti kvantil razdiobe od  $T - \theta$ . Tada je iz definicije kvantila

$$\mathbb{P}(T - \theta \leq a_\alpha) = \mathbb{P}(T - \theta \geq a_{1-\alpha}) = \alpha. \quad (3.2)$$

Događaj  $\{T - \theta \leq a_\alpha\}$  ekvivalentan je s  $\{\theta \geq T - a_\alpha\}$  dok je događaj  $\{T - \theta \geq a_{1-\alpha}\}$  ekvivalentan s  $\{\theta \leq T - a_{1-\alpha}\}$  pa imamo da je u (3.2)

$$\mathbb{P}(\theta \leq T - a_{1-\alpha}) = \mathbb{P}(\theta \geq T - a_\alpha) = \alpha$$



odakle slijedi da su granice  $(1 - 2\alpha) \cdot 100\%$  pouzdanog intervala za  $\theta$  dane s

$$\hat{\theta}_\alpha = t - a_{1-\alpha} \quad \text{i} \quad \hat{\theta}_{1-\alpha} = t - a_\alpha, \quad (3.3)$$

gdje je  $t$  opažena vrijednost procjenitelja  $T$ . Od interesa nam je procijeniti razdiobu od  $T - \theta$  budući da nam je ona nepoznata, ali je ključna kod formiranja pouzdanih intervala. Postoji nekoliko načina procjene te razdiobe koje ćemo sada analizirati.

### Pouzdana intervali uz normalnu aproksimaciju

Pretpostavimo da je  $T - \theta \sim N(0, v)$  barem aproksimativno, gdje je  $v$  varijanca procjenitelja  $T$ . Ovo je dosta jaka pretpostavka budući da osim normalnosti pretpostavljamo i da je procjenitelj  $T$  nepristran za  $\theta$ , odnosno da mu je očekivanje upravo  $\theta$ . Pokazat ćemo kasnije kako u praksi možemo provjeriti pretpostavku normalnosti. Ukoliko su gornje pretpostavke istinite, tada je u (3.3)  $a_\alpha = \sqrt{v} \cdot z_\alpha$  i  $a_{1-\alpha} = \sqrt{v} \cdot z_{1-\alpha}$ , gdje su  $z_\alpha = \Phi^{-1}(\alpha)$  i  $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$  kvantili standardne normalne razdiobe. Dakle su granice pouzdanog intervala dane s

$$\hat{\theta}_\alpha = t - \sqrt{v} \cdot z_{1-\alpha} \quad \text{i} \quad \hat{\theta}_{1-\alpha} = t - \sqrt{v} \cdot z_\alpha. \quad (3.4)$$

Ukoliko ne znamo izraz za egzaktnu varijancu procjenitelja  $T$ , za  $v$  možemo uzeti procjenu varijance dobivenu neparametarskom delta metodom danu s (1.4). Ukoliko je potrebno, empirijske funkcije utjecaja u toj formuli možemo procijeniti *jackknifeom*. Ako je  $T$  procjenitelj metodom maksimalne vjerodostojnosti, onda za  $v$  možemo uzeti njegovu procijenjenu asimptotsku varijancu danu s  $v_a = \frac{1}{nI(t)} = -\frac{1}{I''(t)}$ , gdje je  $I(\theta)$  Fisherova informacija,  $I$  funkcija log-vjerodostojnosti, a  $t$  procjena za  $\theta$ .

Ako se pak nijedna od ovih metoda procjena varijanci ne čini pouzdanom, varijancu možemo procijeniti koristeći parametarski *bootstrap*, slično kao što smo to učinili u Primjeru 1.1.1, samo što bismo ovdje simulirali iz nekog parametarskog modela s procijenjenim parametrima umjesto pravih (nepoznatih) vrijednosti. Dakle, simulirali bismo vrijednosti  $t_1^*, \dots, t_R^*$  procjenitelja  $T$  i za procjenu varijance uzeli varijancu simuliranih (zajedno s opaženom) vrijednosti. Korištenje *bootstrapa* za procjenu varijance ima dvije prednosti. Prvo, možemo iskoristiti simulirane vrijednosti kako bismo provjerili je li procjenitelj  $T$  aproksimativno normalno distribuiran (na primjer, koristeći normalni vjerojatnosni graf). Drugo, možemo procijeniti pristranost od  $T$  kao  $b = \bar{t}^* - t$ , gdje je  $\bar{t}^*$  prosjek simuliranih (i opažene) vrijednosti. Tada bismo u (3.4)  $t$  zamijenili s  $t - b$  pa bi granice pouzdanog intervala korištenjem *bootstrapa* (u slučaju aproksimativne normalnosti procjenitelja  $T$ ) bile

$$\hat{\theta}_\alpha = (t - b) - \sqrt{v_{\text{boot}}} \cdot z_{1-\alpha} \quad \text{i} \quad \hat{\theta}_{1-\alpha} = (t - b) - \sqrt{v_{\text{boot}}} \cdot z_\alpha.$$

### Osnovni i studentizirani pouzdani intervali

Osnovna je ideja ovdje koristiti simulacije da bismo procijenili razdiobu od  $T - \theta$ . Naime, neka je  $T^*$  simulirani procjenitelj, a  $t_1^*, \dots, t_R^*$  simulirane vrijednosti procjenitelja  $T$ . Tada

je smisleno razdiobu od  $T - \theta$  procijeniti pomoću  $T^* - t$ . Ako su  $t_{(1)}^* \leq \dots \leq t_{(R)}^*$  uređene simulirane vrijednosti procjenitelja  $T$ , onda  $p$ -ti kvantil  $a_p$  razdiobe od  $T - \theta$  procjenjujemo kao  $t_{((R+1)p)}^* - t$ , pri čemu moramo odabrati broj simulacija  $R$  tako da je  $(R + 1)p$  prirodan broj (kao i obično,  $R = 9999$  će u gotovo svim slučajevima biti dobar odabir). Stoga ćemo za kvantile  $a_\alpha$  i  $a_{1-\alpha}$  u (3.2) imati

$$\mathbb{P}(T - \theta \leq t_{((R+1)\alpha)}^* - t) \approx \mathbb{P}(T - \theta \geq t_{((R+1)(1-\alpha))}^* - t) \approx \alpha,$$

pa će, kao u (3.3) granice osnovnog pouzdanog intervala biti

$$\hat{\theta}_\alpha = 2t - t_{((R+1)(1-\alpha))}^* \quad \text{i} \quad \hat{\theta}_{1-\alpha} = 2t - t_{((R+1)\alpha)}^*. \quad (3.5)$$

Druga je ideja koristiti granice pouzdanog intervala kao u (3.4), ali bez pretpostavke da procjenitelj  $T$  ima normalnu razdiobu. Tada je potrebno prilagoditi kvantile  $z_\alpha$  i  $z_{1-\alpha}$  i to se može učiniti simuliranjem. Dakle, ako je  $T$  procjenitelj za  $\theta$  s varijancom  $V$  onda želimo procijeniti kvantile razdiobe  $Z = \frac{T-\theta}{\sqrt{V}}$ , što je zapravo studentizirani procjenitelj  $T$ . Simuliramo  $R$  nezavisnih slučajnih uzoraka iste duljine i iz svakog izračunamo vrijednost procjenitelja  $t_r^*$  te njegovu varijancu  $v_r^*$  (na primjer, neparametarskom delta metodom). Koristeći njih izračunamo  $z_r^* = \frac{t_r^* - t}{\sqrt{v_r^*}}$ . Ako su  $z_{(1)}^* \leq \dots \leq z_{(R)}^*$  uređene vrijednosti simuliranog studentiziranog procjenitelja  $T$ , onda je procjena  $p$ -tog kvantila razdiobe od  $Z$  dana kao  $z_{((R+1)p)}^*$ . Nakon što smo procijenili te kvantile, granice pouzdanog intervala su dane slično kao u (3.4):

$$\hat{\theta}_\alpha = t - \sqrt{v} \cdot z_{((R+1)(1-\alpha))}^* \quad \text{i} \quad \hat{\theta}_{1-\alpha} = t - \sqrt{v} \cdot z_{((R+1)\alpha)}^*, \quad (3.6)$$

gdje je  $v$  procjena varijance procjenitelja  $T$  dobivena iz početnog uzorka. Ovakav pouzdani interval nazivamo studentizirani pouzdani interval. Empirijski se pokazalo da studentizirani pouzdani interval u praksi daje bolje procjene nego osnovni pouzdani interval.

Napomenimo još jednom da je nužno uzeti broj simulacija  $R$  takav da su  $(R + 1)\alpha$  i  $(R + 1)(1 - \alpha)$  prirodni brojevi. Iako će to gotovo uvijek biti moguće učiniti, ako iz nekog razloga ne možemo pronaći takav  $R$  onda koristimo interpolaciju. Naime, ako je  $[(R + 1)\alpha] = d$  cijeli dio od  $(R + 1)\alpha$  onda definiramo

$$t_{((R+1)\alpha)}^* = t_{(d)}^* + \frac{\Phi^{-1}(\alpha) - \Phi^{-1}\left(\frac{d}{R+1}\right)}{\Phi^{-1}\left(\frac{d+1}{R+1}\right) - \Phi^{-1}\left(\frac{d}{R+1}\right)} \cdot (t_{(d+1)}^* - t_{(d)}^*), \quad (3.7)$$

gdje je  $\Phi$  funkcija distribucije standardne normalne slučajne varijable. Naravno, cijeli dio  $d$  ne smije biti jednak 0,  $R$  niti  $R + 1$ , ali to je uvijek moguće izbjeći.

### Transformacije parametra

Pouzdana intervala uz normalnu aproksimaciju mogu ponekad dati loše rezultate ako je razdioba procjenitelja  $T$  jako nagnuta (odnosno, asimetrična). Asimetrija se često može

prepoznati promatrajući varijancu procjenitelja  $T$ . Naime, ukoliko se  $\text{Var}(T)$  mijenja s promjenom  $\theta$  možemo zaključiti da je razdioba od  $T$  nagnuta i korištenje normalne aproksimacije za konstrukciju pouzdanog intervala nije dobra ideja. Zbog toga je poželjno koristiti neku transformaciju koja će kontrolirati varijancu procjenitelja  $T$ .

Neka je  $h$  rastuće i bijektivno (ili barem strogo rastuće) preslikavanje i  $\eta = h(\theta)$ . Neka je također  $U = h(T)$  i  $u = h(t)$ . Ideja je primijeniti bilo koju od prethodnih metoda konstrukcije pouzdanih intervala na transformirane vrijednosti  $\eta$  i  $u$ . Primjenom preslikavanja  $h^{-1}$  na granice pouzdanog intervala za  $\eta$  dobivamo pouzdani interval za  $\theta$ . Pokažimo kako to možemo realizirati za pouzdane intervale uz normalnu aproksimaciju. Neka je  $v_T$  varijanca procjenitelja  $T$ . Koristeći Taylorov razvoj može se pokazati da je

$$\text{Var}(U) = \text{Var}(h(T)) \approx (h'(t))^2 \cdot v_T = v_U$$

pa imamo da su granice pouzdanog intervala (3.4) za  $\eta$  dane kao

$$\hat{\eta}_\alpha = h(t) - \sqrt{v_U} \cdot z_{1-\alpha} \quad \text{i} \quad \hat{\eta}_{1-\alpha} = h(t) - \sqrt{v_U} \cdot z_\alpha.$$

Dakle, granice pouzdanog intervala za  $\theta$  (zbog strogog rasta funkcije  $h$ ) su

$$\hat{\theta}_\alpha = h^{-1}(h(t) - \sqrt{v_U} \cdot z_{1-\alpha}) \quad \text{i} \quad \hat{\theta}_{1-\alpha} = h^{-1}(h(t) - \sqrt{v_U} \cdot z_\alpha). \quad (3.8)$$

Na isti način dobivamo da je osnovni pouzdani interval za  $\theta$  oblika

$$\hat{\theta}_\alpha = h^{-1}(2h(t) - h(t_{((R+1)(1-\alpha))}^*)) \quad \text{i} \quad \hat{\theta}_{1-\alpha} = h^{-1}(2h(t) - h(t_{((R+1)\alpha}^*))) \quad (3.9)$$

te da je studentizirani pouzdani interval za  $\theta$  oblika

$$\hat{\theta}_\alpha = h^{-1}(h(t) - \sqrt{v_U} \cdot z_{((R+1)(1-\alpha))}^*) \quad \text{i} \quad \hat{\theta}_{1-\alpha} = h^{-1}(h(t) - \sqrt{v_U} \cdot z_{((R+1)\alpha}^*)), \quad (3.10)$$

gdje su sada  $z_r^* = \frac{h(t_r^*) - h(t)}{|h'(t_r^*)| \cdot \sqrt{v_r^*}}$  simulirani kvantili.

Preostaje odrediti tu transformaciju  $h$ . Ukoliko je nastupio slučaj s početka ovog odjeljka, dakle, ako je  $\text{Var}(T) = v(\theta)$  neka funkcija od  $\theta$ , onda definiramo

$$h(t) = \left( \int \frac{du}{\sqrt{v(u)}} \right) \Big|_{u=t} \quad (3.11)$$

i tu funkciju zovemo funkcijom stabilizirajuće varijance.

Sada se postavlja problem određivanja funkcije  $\theta \mapsto v(\theta)$ . Ako je  $T$  procjenitelj metodom maksimalne vjerodostojnosti, onda je asimptotska varijanca procjenitelja  $T$  dana izrazom  $v(\theta) = \frac{1}{nI(\theta)} = -\frac{1}{l''(\theta)}$ , gdje je  $I(\theta)$  Fisherova informacija, a  $l$  funkcija log-vjerodostojnosti. Ako ne možemo egzaktno odrediti preslikavanje  $\theta \mapsto v(\theta)$ , onda simuliramo uzorke iste dužine za razne vrijednosti  $\theta$  (po jedan uzorak za svaku vrijednost  $\theta$ ) i iz svakog uzorka procijenimo varijancu od  $T$  (na primjer, neparametarskom delta metodom). U Kartezijevom koordinatnom sustavu crtamo parove tih točaka i na temelju grafa pokušavamo procijeniti funkciju  $\theta \mapsto v(\theta)$ .

**Primjer 3.2.1.** U svrhu ispitivanja učinkovitosti metode sijanja oblaka kojom bi se trebala povećati količina oborina, na slučajnan je način odabrano  $n = 16$  oblaka koji su tada bili posijani. Mjerena je količina kiše koja je pala iz svakog oblaka. Podaci su dani u sljedećoj tablici, a preuzeti su iz [9]:

$i$	$x_i$	$i$	$x_i$
1	200.7	9	612.7
2	978.0	10	129.6
3	198.6	11	489.1
4	753.4	12	430.0
5	17.5	13	7.7
6	255.0	14	31.4
7	115.3	15	302.8
8	242.5	16	512.8

Pretpostavimo da podaci prate eksponencijalnu razdiobu (grafički prikazi i testovi pripadnosti distribuciji ne odbacuju tu tvrdnju) s nepoznatim parametrom  $\lambda > 0$ . Želimo procijeniti očekivanu količinu kiše  $\mu$  koja padne iz posijanog oblaka i naći 95% pouzdani interval za tu očekivanu vrijednost.

Budući da je očekivanje eksponencijalne slučajne varijable dano s  $\mu = \frac{1}{\lambda}$  te je procjenitelj metodom maksimalne vjerodostojnosti za parametar  $\theta = \lambda$  dan s  $\hat{\lambda}_{MLE} = \frac{1}{\bar{X}}$ , to koristeći invarijantnost MLE na funkcijske transformacije imamo da je MLE za očekivanu vrijednost  $\mu$  dan s  $\hat{\mu}_{MLE} = T = \bar{X}$ . Egzaktna varijanca ovog procjenitelja je  $\text{Var}(T) = \frac{\sigma^2}{n}$ , gdje je  $\sigma^2$  populacijska varijanca opažanja, pa za procjenu varijance od  $T$  uzimamo  $v = \frac{s^2}{n}$ . Opažena vrijednost procjenitelja  $T$  je  $t = \bar{x} = 329.82$ , a uzoračka varijanca je  $s^2 = 78185.08$  pa procijenjena varijanca od  $T$  iznosi  $v = \frac{78185.08}{16} = 4886.57 = 69.9^2$ .

Ako razdiobu od  $T$  aproksimiramo normalnom razdiobom i iskoristimo (3.4), onda imamo da su (uz  $z_{0.025} = -1.96$  i  $z_{0.975} = 1.96$ ) procjene granica 95% pouzdanog intervala za  $\mu$

$$\hat{\theta}_\alpha = 329.82 - 1.96 \cdot 69.9 = 192.82 \quad i \quad \hat{\theta}_{1-\alpha} = 329.82 + 1.96 \cdot 69.9 = 466.82$$

pa je 95% pouzdani interval za  $\mu$  dan s  $[192.82, 466.82]$ .

Pokušajmo pronaći funkciju stabilizirajuće varijance  $h$ . Budući da je  $v(\mu) = \frac{\mu^2}{n}$ , ignorirajući faktor  $n^{-1}$  imamo da je funkcija  $h$  dana preko (3.11) i možemo izračunati da je

$$h(t) = \left( \int \frac{du}{u} \right) \Big|_{u=t} = \ln(u) \Big|_{u=t} = \ln(t).$$

Ako stavimo da je  $U = h(T) = \ln(\bar{X})$  imamo da je  $\text{Var}(U) = v_U \approx (h'(t))^2 \cdot v = \frac{v}{\bar{x}^2} = 0.0449$  pa su po (3.8) procjene granica pouzdanog intervala

$$\hat{\theta}_\alpha = \exp\left(\ln(329.82) - 1.96 \cdot \sqrt{0.0449}\right) = 217.72$$

i

$$\hat{\theta}_{1-\alpha} = \exp\left(\ln(329.82) + 1.96 \cdot \sqrt{0.0449}\right) = 499.63,$$

to jest procjena 95% pouzdanog intervala za  $\mu$  je [217.72, 499.63].

Za konstrukciju osnovnog pouzdanog intervala simuliramo  $R = 9999$  nezavisnih slučajnih uzoraka duljine  $n = 16$  iz eksponencijalne razdiobe s parametrom  $\frac{1}{t} = \frac{1}{x} = \frac{1}{329.82}$ . Iz svakog simuliranog uzorka izračunamo vrijednost simuliranog procjenitelja  $t_r^*$ . Simulirane vrijednosti poredamo uzlazno i nađemo 250. i 9750. vrijednost u sortiranom nizu (to su traženi simulirani kvantili). U našim simulacijama je  $t_{(250)}^* = 189.50$  i  $t_{(9750)}^* = 505.91$  pa su po (3.5) procjene granica pouzdanog intervala dane s

$$\hat{\theta}_\alpha = 2 \cdot 329.82 - 505.91 = 153.73 \quad i \quad \hat{\theta}_{1-\alpha} = 2 \cdot 329.82 - 189.50 = 470.14,$$

to jest procjena 95% pouzdanog intervala za  $\mu$  je [153.73, 470.14].

Ako iskoristimo funkciju stabilizirajuće varijance  $h(\mu) = \ln(\mu)$ , iz (3.9) slijedi da su procjene granica pouzdanog intervala

$$\hat{\theta}_\alpha = \exp(2 \cdot \ln(329.82) - \ln(505.91)) = 215.02$$

i

$$\hat{\theta}_{1-\alpha} = \exp(2 \cdot \ln(329.82) - \ln(189.50)) = 574.04$$

pa je procjena 95% pouzdanog intervala za  $\mu$  interval [215.02, 574.04].

Konačno, za konstrukciju studentiziranog pouzdanog intervala simuliramo  $R = 9999$  nezavisnih slučajnih uzoraka duljine  $n = 16$  iz eksponencijalne razdiobe s parametrom  $\frac{1}{t} = \frac{1}{x} = \frac{1}{329.82}$ . Iz svakog simuliranog uzorka izračunamo vrijednost  $z_r^* = \frac{t_r^* - t}{\sqrt{v_r^*}}$ , gdje je  $v_r^* = \frac{s_r^{*2}}{n}$  procjena varijance simuliranog procjenitelja  $T_r^*$ . Simulirane vrijednosti  $z_r^*$  poredamo uzlazno i nađemo 250. i 9750. vrijednost u sortiranom nizu (to su traženi simulirani kvantili). U našim simulacijama je  $z_{(250)}^* = -3.28$  i  $z_{(9750)}^* = 1.66$  pa su po (3.6) procjene granica pouzdanog intervala dane s

$$\hat{\theta}_\alpha = 329.82 - 1.66 \cdot 69.9 = 213.79 \quad i \quad \hat{\theta}_{1-\alpha} = 329.82 + 3.28 \cdot 69.9 = 559.09,$$

to jest procjena 95% pouzdanog intervala za  $\mu$  je [213.79, 559.09].

Usporedimo ove procjene s egzaktnim 95% pouzdanim intervalom za  $\mu$ . Egzaktni se pouzdani interval dobiva koristeći činjenicu da statistika

$$W = \frac{2n\bar{X}}{\mu} \sim \chi^2(2n)$$

ima hi-kvadrat razdiobu s  $2n = 32$  stupnja slobode. Dakle je

$$\mathbb{P}\left(\chi^2_{\alpha} \leq \frac{2n\bar{X}}{\mu} \leq \chi^2_{1-\alpha}\right) = 1 - 2\alpha \iff \mathbb{P}\left(\frac{2n\bar{X}}{\chi^2_{1-\alpha}} \leq \mu \leq \frac{2n\bar{X}}{\chi^2_{\alpha}}\right) = 1 - 2\alpha$$

pa kako su ovi kvantili  $\chi^2_{0,025} = 18.29$  i  $\chi^2_{0,975} = 49.48$  imamo da je

$$\hat{\theta}_{\alpha} = \frac{2 \cdot 16 \cdot 329.82}{49.48} = 213.30 \quad i \quad \hat{\theta}_{\alpha} = \frac{2 \cdot 16 \cdot 329.82}{18.29} = 577.05,$$

to jest egzaktni 95% pouzdani interval za  $\mu$  je  $[213.30, 577.05]$ .

Radi usporedbe, u tablici ispod prikazane su gornje i donje granice svih aproksimativnih pouzdanih intervala zajedno s egzaktnim vrijednostima. Primjećujemo da je osnovni pouzdani interval uz korekciju s funkcijom stabilizirajuće varijance najbliži egzaktnom pouzdanom intervalu.

Metoda	$\hat{\theta}_{\alpha}$	$\hat{\theta}_{1-\alpha}$
Egzaktni	213.30	577.05
Normalni	192.82	466.82
Normalni sa stabilizacijom	217.72	499.63
Osnovni	153.73	470.14
Osnovni sa stabilizacijom	215.02	574.04
Studentizirani	213.79	559.09

## Neparametarski modeli

U neparametarskom slučaju nemamo nikakvu pretpostavku na razdiobu iz koje dolaze opažanja. Međutim, većina zaključaka koje smo donijeli za parametarske modele vrlo se lako prenose na slučaj neparametarskih modela. Za procjenitelj  $T$  nepoznatog parametra  $\theta$  uzimamo ono što nam se čini *najsmislenije* za naš problem ili ono što bi  $T$  bio kad bismo se nalazili u parametarskom slučaju.

### Pouzdana intervali uz normalnu aproksimaciju

Slično kao i ranije, pretpostavimo da je  $T - \theta \sim N(0, v)$  barem aproksimativno, gdje je  $v$  varijanca procjenitelja  $T$ . Ovdje se  $v$  procjenjuje s  $v_L$  koja je dobivena neparametarskom

delta metodom kao u (1.4). Ukoliko je potrebno, empirijske funkcije utjecaja u toj formuli možemo procijeniti *jackknifeom*. Ukoliko je pretpostavka normalnosti istinita, tada su granice pouzdanog intervala dane s

$$\hat{\theta}_\alpha = t - \sqrt{v_L} \cdot z_{1-\alpha} \quad \text{i} \quad \hat{\theta}_{1-\alpha} = t - \sqrt{v_L} \cdot z_\alpha, \quad (3.12)$$

gdje su ponovno  $z_\alpha$  i  $z_{1-\alpha}$  kvantili standardne normalne razdiobe.

Ako se ta procjena varijance ne čini pouzdanom, varijancu možemo procijeniti koristeći neparametarski *bootstrap* kao što smo to učinili u Primjeru 1.1.1. Dakle, simulirali bismo vrijednosti  $t_1^*, \dots, t_R^*$  procjenitelja  $T$  i za procjenu varijance uzeli varijancu simuliranih (zajedno s opaženom) vrijednosti. Također, možemo procijeniti pristranost od  $T$  kao  $b = \bar{t}^* - t$ , gdje je  $\bar{t}^*$  prosjek simuliranih (i opažene) vrijednosti. Tada u (3.12)  $t$  zamjenjujemo s  $t - b$  pa su granice pouzdanog intervala korištenjem *bootstrapa* (u slučaju aproksimativne normalnosti procjenitelja  $T$ ) dane s

$$\hat{\theta}_\alpha = (t - b) - \sqrt{v_{\text{boot}}} \cdot z_{1-\alpha} \quad \text{i} \quad \hat{\theta}_{1-\alpha} = (t - b) - \sqrt{v_{\text{boot}}} \cdot z_\alpha.$$

### Osnovni i studentizirani pouzdani intervali

Postupak konstrukcije osnovnog pouzdanog intervala identičan je kao u parametarskom slučaju, jedino što se ovdje uzorak simulira iz razdiobe određene empirijskom funkcijom distribucije  $\hat{F}$ . Otprije već znamo da je simuliranje iz te razdiobe ekvivalentno uzimanju slučajnog uzorka duljine  $n$  s ponavljanjem iz danog skupa opservacija. Dakle, simuliramo  $R$  nezavisnih uzoraka iste duljine na taj način i iz svakog simuliranog uzorka procijenimo  $t_r^*$ . Ako su  $t_{(1)}^* \leq \dots \leq t_{(R)}^*$  uređene simulirane vrijednosti, onda su procjene granica osnovnog pouzdanog intervala dane s (3.5), gdje je  $t$  ponovno opažena vrijednost procjenitelja  $T$ .

Studentizirani pouzdani interval konstruira se jednako kao i u parametarskom slučaju. Glavna je razlika ta što se ovdje varijanca procjenitelja  $T$  procjenjuje s  $v_L$  koja je dobivena neparametarskom delta metodom. Granice studentiziranog pouzdanog intervala su tada dane s

$$\hat{\theta}_\alpha = t - \sqrt{v_L} \cdot z_{((R+1)(1-\alpha))}^* \quad \text{i} \quad \hat{\theta}_{1-\alpha} = t - \sqrt{v_L} \cdot z_{((R+1)\alpha)}^*, \quad (3.13)$$

gdje su  $z_r^* = \frac{t_r^* - t}{\sqrt{v_{L,r}^*}}$  studentizirane simulirane vrijednosti procjenitelja  $T$ . Naravno, procjena  $v_{L,r}^*$  dobivena neparametarskom delta metodom mora se procijeniti u svakom od  $R$  uzoraka posebno.

### Transformacije parametra

Jednako kao i u parametarskom slučaju željeli bismo naći neku strogo rastuću transformaciju  $h$  koja će na neki način kontrolirati varijancu procjenitelja  $T$  i time dati bolje procjene pouzdanih intervala. Ako je  $\text{Var}(T) = v(\theta)$  neka funkcija od  $\theta$ , onda transformaciju  $h$

definiramo pomoću (3.11). Međutim, budući da je ustvari  $v(\theta) = \text{Var}(T | F)$ , gdje je  $F$  (egzaktna) funkcija distribucije naših opažanja, funkciju  $v$  općenito nije moguće naći jer u neparametarskom slučaju ne znamo  $F$ . Stoga funkciju  $\theta \mapsto v(\theta)$  procjenjujemo na sljedeći način, kojeg smo već spomenuli u parametarskom slučaju. Simuliramo uzorke iste duljine iz empirijske funkcije distribucije  $\hat{F}$  (dakle, uzimamo slučajne uzorke s ponavljanjem iz početnog skupa podataka) i iz svakog uzorka izračunamo procjenu  $t$  i procijenimo varijancu od  $T$  neparametarskom delta metodom. U Kartezijevom koordinatnom sustavu crtamo parove tih točaka  $(t_i, v_{L,i})$  i na temelju grafa pokušavamo procijeniti funkciju  $\theta \mapsto v(\theta)$ . Konačno, primjenjujemo (3.11) da bismo pronašli procjenu  $\hat{h}$  funkcije  $h$ . U tom slučaju granice pouzdanih intervala dane su s (3.8), (3.9) i (3.10) uz zamjenu oznake  $h$  oznakom  $\hat{h}$ .

**Primjer 3.2.2.** *Koristimo podatke iz Primjera 3.2.1 da bismo pronašli 95% pouzdani interval za  $\mu$ , što predstavlja očekivanu količinu oborina palu iz posijanih oblaka. Međutim, ovdje ne pretpostavljamo eksponencijalni model kao u prethodnom primjeru već se nalazimo u neparametarskom modelu.*

*Za očekivanje ponovno koristimo procjenu  $T = \bar{X}$  čija je opažena vrijednost  $t = \bar{x} = 329.82$ . Varijancu ovog procjenitelja procjenjujemo neparametarskom delta metodom. Izraz za procjenu izveli smo u Primjeru 1.2.2:*

$$v_L = \frac{1}{n^2} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = 4581.16 = 67.7^2.$$

*Ako razdiobu od  $T$  aproksimiramo normalnom i iskoristimo (3.12) dobivamo da su procjene granica pouzdanog intervala*

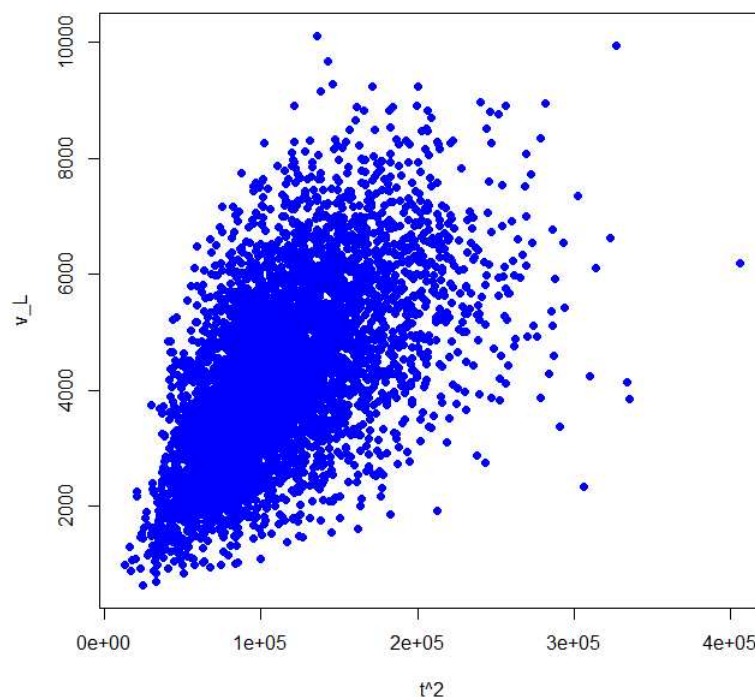
$$\hat{\theta}_\alpha = 329.82 - 1.96 \cdot 67.7 = 197.13 \quad i \quad \hat{\theta}_{1-\alpha} = 329.82 + 1.96 \cdot 67.7 = 462.51,$$

*to jest procjena 95% pouzdanog intervala za  $\mu$  je  $[197.13, 462.51]$ .*

*Pokušajmo sada pronaći aproksimaciju funkcije stabilizirajuće varijance  $\hat{h}$ . Kako ne znamo vezu između  $\mu$  i varijance procjenitelja  $T$  posežemo za simulacijama. Simuliramo  $N = 5000$  uzoraka iste duljine  $n = 16$  s ponavljanjem iz početnog uzorka i na temelju simuliranih uzoraka računamo procjene  $t_i$  i procjenjujemo varijancu  $v_{L,i}$  neparametarskom delta metodom. Točke  $(t_i, v_{L,i})$  crtati ćemo u koordinatnom sustavu ne bismo li uočili neku vezu. Uočavamo otprilike kvadratnu vezu (to jest,  $t_i^2 \propto v_{L,i}$ ) što se još bolje vidi sa Slike 3.1 gdje su prikazani parovi simuliranih točaka  $(t_i^2, v_{L,i})$ . Vidi se jasna linearna veza i veća raspršenost za velike vrijednosti  $t_i$ . Dakle, zaključujemo da je  $\hat{v}(\mu) \propto \mu^2$  pa (ignorirajući neke moguće, nepoznate faktore) dobivamo da je  $\hat{h}(t) = \ln(t)$  istim postupkom kao i u Primjeru 3.2.1. Ako stavimo da je  $U = \hat{h}(T) = \ln(\bar{X})$  imamo da je  $\text{Var}(U) = v_U \approx (\hat{h}'(t))^2 \cdot v_L = \frac{v_L}{\bar{x}^2} = 0.0421$  pa su po (3.8) procjene granica pouzdanog intervala*

$$\hat{\theta}_\alpha = 220.61 \quad i \quad \hat{\theta}_{1-\alpha} = 493.10,$$





Slika 3.1: Veza između vrijednosti procjenitelja  $T$  i njegove varijance u Primjeru 3.2.2

to jest procjena 95% pouzdanog intervala za  $\mu$  je  $[220.61, 493.10]$ .

Za osnovni pouzdani interval simuliramo  $R = 9999$  nezavisnih slučajnih uzoraka iz empirijske funkcije distribucije  $\hat{F}$  i iz svakog uzorka izračunamo procjenu  $t_r^*$ . Simulirane vrijednosti poredamo uzlazno i nađemo 250. i 9750. vrijednost u sortiranom nizu. U našim simulacijama je  $t_{(250)}^* = 204.68$  i  $t_{(9750)}^* = 468.33$  pa su po (3.5) granice pouzdanog intervala dane s

$$\hat{\theta}_\alpha = 2 \cdot 329.82 - 468.33 = 191.31 \quad i \quad \hat{\theta}_{1-\alpha} = 2 \cdot 329.82 - 204.68 = 454.96,$$

to jest procjena 95% pouzdanog intervala za  $\mu$  je  $[191.31, 454.96]$ .

Ukoliko primijenimo transformaciju  $\hat{h}$ , iz (3.9) slijedi da su granice pouzdanog intervala

$$\hat{\theta}_\alpha = 232.27 \quad i \quad \hat{\theta}_{1-\alpha} = 531.47,$$

pa je 95% pouzdani interval za  $\mu$  interval  $[232.27, 531.47]$ .

Za studentizirani pouzdani interval simuliramo  $R = 9999$  nezavisnih slučajnih uzoraka iz  $\hat{F}$  i iz svakog uzorka izračunamo procjenu  $t_r^*$  i procjenu varijance  $v_{L,r}^*$  od  $T$  neparametarskom delta metodom te nađemo vrijednost  $z_r^* = \frac{t_r^* - t}{\sqrt{v_{L,r}^*}}$ . Sortiramo uzlazno vrijednosti  $z_r^*$  i nađemo 250. i 9750. vrijednost u simuliranom nizu. U našim simulacijama je  $z_{(250)}^* = -2.77$  i  $z_{(9750)}^* = 1.92$  pa su po (3.13) procjene granica pouzdanog intervala

$$\hat{\theta}_\alpha = 329.82 - 1.92 \cdot 67.7 = 199.84 \quad i \quad \hat{\theta}_{1-\alpha} = 329.82 + 2.77 \cdot 67.7 = 517.35,$$

to jest procjena 95% pouzdanog intervala za  $\mu$  je [199.84, 517.35].

Radi usporedbe, u tablici ispod prikazane su gornje i donje granice svih aproksimativnih pouzdanih intervala. Osnovni pouzdani interval uz korekciju s funkcijom stabilizirajuće varijance najbliži je egzaktnom pouzdanom intervalu nađenom u Primjeru 3.2.1.

Metoda	$\hat{\theta}_\alpha$	$\hat{\theta}_{1-\alpha}$
Normalni	197.13	462.51
Normalni sa stabilizacijom	220.61	493.10
Osnovni	191.31	454.96
Osnovni sa stabilizacijom	232.27	531.47
Studentizirani	199.84	517.35

### 3.3 Percentilni pouzdani intervali

U prethodnom potpoglavlju vidjeli smo da pouzdani intervali mogu postati puno točniji ako na njih primijenimo funkciju stabilizirajuće varijance  $h$  koja će kontrolirati varijancu našeg procjenitelja. Međutim, ponekad je jako teško pronaći takvu funkciju. Percentilni pouzdani intervali oslanjaju se na simulirane kvantile razdiobe tog procjenitelja i samim time kontroliraju varijancu procjenitelja bez potrebe pronalaženja funkcije  $h$ .

#### Osnovni percentilni pouzdani intervali

Neka je  $T$  procjenitelj nepoznatog parametra  $\theta$ . Pretpostavimo da postoji neko bijektivno i rastuće preslikavanje  $h$ , možda nama nepoznato, takvo da  $U = h(T)$  ima simetričnu razdiobu. Ako je s  $a_p$  označen  $p$ -ti kvantil razdiobe od  $U$  onda zbog simetrije imamo da je  $a_\alpha = -a_{1-\alpha}$ . Neka je  $u = h(t)$  opažena vrijednost statistike  $U$ . Kao i ranije, procjenjujemo razdiobu od  $U - \eta$  pomoću  $U^* - u$ . Ako su  $u_{(1)}^* \leq \dots \leq u_{(R)}^*$  uređene simulirane vrijednosti procjenitelja  $U$ , onda  $p$ -ti kvantil razdiobe od  $U - \eta$  procjenjujemo kao  $u_{((R+1)p)}^* - u$ , ali budući da je, zbog simetrije,  $p$ -ti kvantil te razdiobe jednak negativnom  $(1 - p)$ -tom kvantilu te razdiobe, imamo da je procjena  $p$ -tog kvantila upravo  $u - u_{((R+1)(1-p))}^*$ . Dakle je

$$\mathbb{P}(U - \eta \leq u - u_{((R+1)(1-\alpha))}^*) \approx \mathbb{P}(U - \eta \geq u - u_{((R+1)\alpha)}^*) \approx \alpha$$

pa imamo da su granice osnovnog  $(1 - 2\alpha) \cdot 100\%$  pouzdanog intervala za  $\eta$

$$\hat{\eta}_\alpha = u_{((R+1)\alpha)}^* \quad \text{i} \quad \hat{\eta}_{1-\alpha} = u_{((R+1)(1-\alpha))}^*$$

Dakle, primijenivši inverznu transformaciju  $h^{-1}$  imamo da su granice osnovnog  $(1 - 2\alpha) \cdot 100\%$  pouzdanog intervala za  $\theta$

$$\hat{\theta}_\alpha = t_{((R+1)\alpha)}^* \quad \text{i} \quad \hat{\theta}_{1-\alpha} = t_{((R+1)(1-\alpha))}^* \quad (3.14)$$

i one ne ovise o (možda nepoznatoj) funkciji  $h$ . Dakle, ovaj se pouzdani interval može vrlo jednostavno konstruirati bez da znamo funkciju stabilizirajuće varijance  $h$ , a ima isti efekt kao da smo ju primijenili. Ovaj pouzdani interval nazivamo percentilni *bootstrap* interval.

Naravno, metoda se može primijeniti u parametarskom i u neparametarskom slučaju. U parametarskom slučaju simulacije  $t_r^*$  dobivamo simuliranjem iz pretpostavljene razdiobe s procijenjenim parametrom  $\hat{\theta} = t$ , a u neparametarskom slučaju simuliranjem iz empirijske funkcije distribucije  $\hat{F}$  (dakle, uzimanjem slučajnih uzoraka s ponavljanjem iz skupa opaženih vrijednosti).

Ipak, metoda ne daje uvijek dobre rezultate u neparametarskom slučaju. Zbog toga je razvijena korigirana metoda percentilnog *bootstrap* pouzdanog intervala koju opisujemo u ostatku ovog potpoglavlja. Prije te metode slijedi primjer.

**Primjer 3.3.1.** *Ponovno koristimo podatke iz Primjera 3.2.1 da bismo pronašli 95% pouzdani interval za  $\mu$ , što predstavlja očekivanu količinu oborina palu iz posijanih oblaka.*

*Prvo pretpostavljamo da opažanja dolaze iz eksponencijalne razdiobe čiji je parametar procijenjen kao  $\hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{329.82}$ . Simuliramo  $R = 9999$  nezavisnih slučajnih uzoraka iz eksponencijalne razdiobe s parametrom  $\hat{\lambda} = \frac{1}{329.82}$  i iz svakog uzorka izračunamo vrijednost  $t_r^*$ . Uzlazno sortiramo te vrijednosti i nađemo 250. i 9750. vrijednost u uređenom nizu. U našim simulacijama je  $t_{(250)}^* = 190.48$  i  $t_{(9750)}^* = 514.21$  pa je procjena 95% pouzdanog intervala za  $\mu$  interval  $[190.48, 514.21]$ .*

*Sada ne pretpostavljmo nikakvu razdiobu pa simuliramo  $R = 9999$  nezavisnih slučajnih uzoraka iz empirijske funkcije distribucije  $\hat{F}$  i iz svakog uzorka izračunamo vrijednost  $t_r^*$ . Isto kao i u prvom slučaju, uzlazno sortiramo te vrijednosti i nađemo 250. i 9750. vrijednost u uređenom nizu. U našim simulacijama je  $t_{(250)}^* = 202.06$  i  $t_{(9750)}^* = 471.4$  pa je procjena 95% pouzdanog intervala za  $\mu$  interval  $[202.06, 471.4]$ .*

### Korigirani percentilni pouzdani intervali

Glavni je problem percentilne metode to što procjenitelj  $U = h(T)$  možda nije nepristran za  $\eta = h(\theta)$  i to je ono što ova korigirana metoda pokušava popraviti. Također, *oblik* uzoračke

razdiobe procjenitelja  $T$  može se drastično promijeniti kad pravu funkciju distribucije  $F$  zamijenimo (to jest, aproksimiramo) s procjenom  $\hat{F}$ . Važno je napomenuti i da simetričnost razdiobe od  $U$  nema sasvim isti efekt kao i funkcija stabilizirajuće varijance  $h$ . Korigirana metoda koju ćemo sada predstaviti popravlja te nedostatke i u velikoj većini slučajeva daje zadovoljavajuće dobre rezultate.

### Parametarski modeli s jednim parametrom

Za početak pretpostavimo da naša opažanja dolaze iz iste funkcije distribucije  $F$  koja je opisana samo jednim nepoznatim parametrom  $\theta$ . Neka je  $T$  procjenitelj parametra  $\theta$  metodom maksimalne vjerodostojnosti, a  $t = \hat{\theta}$  njegova opažena vrijednost. Pretpostavimo da postoji bijektivna i rastuća funkcija  $h$ , faktor korekcije pristranosti  $w$  i faktor korekcije asimetrije  $a$  (svi troje nepoznati) takvi da je  $h(\theta) = \varphi$  i da je razdioba od  $U = h(T)$  normalna:

$$U \sim N(\varphi - w\sigma(\varphi), \sigma^2(\varphi)), \quad \text{gdje je } \sigma(\varphi) = 1 + a\varphi. \quad (3.15)$$

Ideja je metodom *bootstrapa* pronaći pouzdani interval za  $\varphi$  i inverznom transformacijom  $h^{-1}$  dobiti pouzdani interval za  $\theta$ . Neka je  $Z \sim N(0, 1)$  standardna normalna slučajna varijabla. Tada se  $U$  može napisati kao

$$U = \varphi + (1 + a\varphi)(Z - w).$$

Neka su, kao i obično, sa  $z_\alpha$  označeni kvantili standardne normalne razdiobe (to jest, kvantili od  $Z$ ). Množenjem gornje jednakosti s  $a$ , dodavanjem 1 i uzimanjem prirodnog logaritma slijedi

$$\ln(1 + aU) = \ln(1 + a\varphi) + \ln(1 + a(Z - w)).$$

Stavimo li kvantil  $z_\alpha$  umjesto  $Z$  i opaženu vrijednost  $u$  umjesto  $U$  te riješimo gornju jednadžbu po  $\varphi$  dobivamo da je donja granica pouzdanog intervala za  $\varphi$

$$\hat{\varphi}_\alpha = u + \sigma(u) \cdot \frac{w + z_\alpha}{1 - a(w + z_\alpha)}.$$

Primijenivši inverznu transformaciju  $h^{-1}$  dobivamo da je donja granica pouzdanog intervala za  $\theta$  upravo  $\hat{\theta}_\alpha = h^{-1}(\hat{\varphi}_\alpha)$ . Međutim, funkcija  $h$  je nepoznata.

Neka je  $T^*$  simulirani procjenitelj, a  $\hat{G}$  njegova funkcija distribucije. Neka je još i  $\mathbb{P}^*$  vjerojatnost u odnosu na razdiobu dobivenu simuliranjem. Tada je

$$\hat{G}(\hat{\theta}_\alpha) = \mathbb{P}^*(T^* < \hat{\theta}_\alpha \mid t) = \mathbb{P}^*(U^* < \hat{\varphi}_\alpha \mid u) = \Phi\left(\frac{\varphi_\alpha - u}{\sigma(u)} + w\right) = \Phi\left(w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\right).$$

Primjenom inverzne transformacije  $\hat{G}^{-1}$  dobivamo da je

$$\hat{\theta}_\alpha = \hat{G}^{-1}\left(\Phi\left(w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\right)\right). \quad (3.16)$$

Ali  $\hat{G}$  je funkcija distribucije od  $T^*$  pa ako su  $t_{(1)}^* \leq \dots \leq t_{(R)}^*$  uređene simulirane vrijednosti procjenitelja  $T$  onda (3.16) možemo izraziti kao

$$\hat{\theta}_\alpha = t_{((R+1)\tilde{\alpha})}, \quad \text{gdje je} \quad \tilde{\alpha} = \Phi\left(w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\right). \quad (3.17)$$

Primjećujemo da funkcija  $\hat{G}$  ustvari zamjenjuje nepoznatu funkciju  $h$ . Ostalo nam je još procijeniti vrijednosti  $a$  i  $w$  koje su obje nepoznate. Iz (3.15) slijedi

$$\mathbb{P}^*(T^* < t \mid t) = \mathbb{P}^*(U^* < u \mid u) = \mathbb{P}(U < \varphi \mid \varphi) = \Phi(w)$$

pa imamo da je

$$w = \Phi^{-1}(\hat{G}(t)).$$

Ponovno, kako je  $\hat{G}$  funkcija distribucije od  $T^*$  imamo da se  $w$  može aproksimirati simulacijama kao

$$\hat{w} = \Phi^{-1}\left(\frac{k(\{r; t_r^* \leq t\})}{R+1}\right). \quad (3.18)$$

Broj  $w$  (odnosno njegovu procjenu  $\hat{w}$ ) nazivamo *korekcijom pristranosti* (engl. *bias-correction*) budući da on, grubo rečeno, mjeri kolika je razlika između medijalne vrijednosti simulacija  $t_r^*$  i procjene  $t$  izražena u *normalnim kvantilima*.

Neka je  $l(\varphi)$  funkcija log-vjerodostojnosti modela danog s (3.15). Može se pokazati da, ignorirajući faktore reda  $n^{-1}$ , vrijedi

$$\frac{\mathbb{E}(l'(\varphi)^3)}{\text{Var}(l'(\varphi))^{3/2}} = 6a,$$

gdje  $l'(\varphi)$  shvaćamo kao slučajnu varijablu  $l'(\varphi) = l'(\varphi; U)$ , odnosno kao funkciju od  $U$ . Ovaj je omjer invarijantan na funkcijske transformacije pa ako iskoristimo preslikavanje  $h^{-1}$  da  $\varphi$  preslikamo *natrag* u  $\theta$  onda, ponovno ignorirajući faktore reda  $n^{-1}$ , imamo

$$\frac{\mathbb{E}(l'(\theta)^3)}{\text{Var}(l'(\theta))^{3/2}} = 6a.$$

Ukoliko je potrebno,  $a$  aproksimiramo simulacijama kao

$$\frac{1}{6} \cdot \frac{\mathbb{E}^*(l'^*(t)^3)}{\text{Var}^*(l'^*(t))^{3/2}} \approx \frac{(R-1)^{3/2}}{6R} \cdot \frac{\sum_{r=1}^R l_r'^*(t)^3}{\left[\sum_{r=1}^R (l_r'^*(t) - \overline{l'^*(t)})^2\right]^{3/2}} = \hat{a}, \quad (3.19)$$

gdje je  $l'^*$  funkcija log-vjerodostojnosti uzorka simuliranog iz pretpostavljene razdiobe s parametrom  $t$ . Broj  $a$  (odnosno njegova procjena  $\hat{a}$ ) zove se *ubrzanje* (engl. *acceleration*) jer on reflektira brzinu promjene standardne greške procjenitelja  $T$  u odnosu na promjenu vrijednosti parametra  $\theta$ .

Pouzdan interval dobiven ovom metodom naziva se  $BC_a$  (engl. *bias-corrected and accelerated*) pouzdani interval.

**Primjer 3.3.2.** Pronađimo  $BC_a$  95% pouzdani interval za očekivanu količinu oborina  $\mu$  iz Primjera 3.2.1.

Budući da smo pretpostavili eksponencijalni model s (nepoznatim) parametrom  $\theta = \mu = \frac{1}{\lambda}$  za naša opažanja imamo da je funkcija vjerodostojnosti uzorka  $X_1, \dots, X_n$  duljine  $n$  u ovisnosti o  $\mu = \frac{1}{\lambda}$  dana s

$$L(\mu) = \prod_{i=1}^n \left( \frac{1}{\mu} \cdot e^{-\frac{x_i}{\mu}} \right) = \frac{1}{\mu^n} \cdot e^{-\frac{x_1 + \dots + x_n}{\mu}}$$

pa je log-vjerodostojnost

$$l(\mu) = \ln(L(\mu)) = -n \ln(\mu) - \frac{1}{\mu} \cdot \sum_{i=1}^n x_i.$$

Stoga je derivacija log-vjerodostojnosti po  $\mu$  jednaka

$$l'(\mu) = -\frac{n}{\mu} + \frac{1}{\mu^2} \cdot \sum_{i=1}^n x_i.$$

Ako sada zamijenimo  $x_i$  sa slučajnim varijablama  $X_i$  možemo naći da je  $\mathbb{E}(l'(\mu)^3) = \frac{2n}{\mu^3}$  te da je  $\text{Var}(l'(\mu)) = \frac{n}{\mu^2}$  pa imamo da je ubrzanje jednako

$$a = \frac{1}{6} \cdot \frac{\frac{2n}{\mu^3}}{\frac{n^{3/2}}{\mu^3}} = \frac{n^{-1/2}}{3} = \frac{1}{3\sqrt{n}} = \frac{1}{3\sqrt{16}} = 0.0833.$$

Da nismo računali treći moment i varijancu, već da smo odlučili posegnuti za simulacijama (simuliramo puno uzoraka iz eksponencijalne razdiobe s parametrom  $\frac{1}{t}$  i iz svakog uzorka izračunamo vrijednost  $l_r^*(t)$  te primijenimo (3.19)) dobili bismo procjenu  $\hat{a} = 0.0841$ , što je vrlo blizu prave vrijednosti.

Za procjenu korekcije pristranosti simuliramo  $R = 9999$  nezavisnih slučajnih uzoraka iste duljine iz eksponencijalne razdiobe s parametrom  $\frac{1}{t} = \frac{1}{329.82}$  i iz svakog simuliranog uzorka izračunamo procjenu  $t_r^*$ . U našim je simulacijama 5321 simulirana procjena  $t_r^*$  manja ili jednaka  $t = 329.82$  pa je

$$\hat{w} = \Phi^{-1}\left(\frac{5321}{10000}\right) = 0.0805.$$

Iz vrijednosti  $a$  i  $\hat{w}$  dobivamo

$$\alpha_1 = \Phi\left(0.0805 + \frac{0.0805 + \overbrace{z_{0.025}}^{-1.96}}{1 - 0.0833 \cdot (0.0805 + z_{0.025})}\right) = \Phi(-1.544) = 0.0612$$

te

$$\alpha_2 = \Phi \left( 0.0805 + \frac{0.0805 + \overbrace{z_{0.975}}^{=1.96}}{1 - 0.0833 \cdot (0.0805 + z_{0.975})} \right) = \Phi(2.539) = 0.9944.$$

Dakle je donja granica  $BC_a$  95% pouzdanog intervala  $t_{((R+1)\alpha_1)}^* = t_{(612)}^* = 214.36$ , a gornja je granica  $t_{((R+1)\alpha_2)}^* = t_{(9944)}^* = 574.93$ . Stoga je  $BC_a$  95% pouzdani interval za  $\mu$  dan s [214.36, 574.93]. Primijetimo da je ovaj pouzdani interval gotovo identičan egzaktnom pouzdanom intervalu dobivenom u Primjeru 3.2.1.

### Parametarski modeli s više parametara

Korigirana percentilna  $BC_a$  metoda ne može se upotrijebiti ako u parametarskom modelu imamo više nepoznatih parametara. Problem se javlja kod računanja ubrzanja  $a$  budući da tada ne znamo izračunati funkciju log-vjerodostojnosti. Samo procijeniti ostale nepoznate parametre i staviti ih u (3.19) umjesto pravih vrijednosti parametara nije rješenje.

Pretpostavimo da je  $\psi = (\theta, \lambda)$  vektor svih parametara modela i da, kao i obično, želimo pouzdani interval za  $\theta$ . Dakle,  $\lambda$  predstavlja vektor preostalih nepoznatih parametara. Neka je

$$I(\psi) = \mathbb{E}(-Hl(\psi)) = \mathbb{E} \left( \left[ -\frac{\partial^2 l}{\partial \psi_i \partial \psi_j}(\psi) \right]_{ij} \right)$$

Fisherova informacijska matrica, gdje smo radi lakše notacije označili  $\psi = (\psi_1, \dots, \psi_d)$ . Neka je  $\hat{\psi}$  vektor procjena vektora parametara  $\psi$  (na primjer, dobiven metodom maksimalne vjerodostojnosti). Stavimo  $\hat{\delta} = I^{-1}(\hat{\psi})(1, 0, \dots, 0)^\tau$ . Definiramo *najnepovoljniju familiju*

*razdioba* kao jednoparametarsku familiju razdioba (s parametrom kojeg označavamo s  $\zeta$ ) dobivene iz početnog (pretpostavljenog) modela, ali s vektorom parametara  $\psi$  restriniranim na  $\hat{\psi} + \zeta \hat{\delta}$ . Funkcija log-vjerodostojnosti iz ove najnepovoljnije familije dana je s

$$l_{LF}(\zeta) = l(\hat{\psi} + \zeta \hat{\delta}),$$

gdje je kao i ranije  $l$  funkcija log-vjerodostojnosti početnog modela.

Konstrukciju  $BC_a$  pouzdanog intervala sada provodimo kao i ranije, ali uz najnepovoljniju familiju. Korekcija pristranosti  $w$  računa se na isti način, dakle pomoću (3.18), gdje su sada  $t_r^*$  simulirane vrijednosti procjenitelja  $T$  dobivene pomoću uzoraka simuliranih iz razdiobe određene najnepovoljnijom familijom. Za računanje procjene  $\hat{a}$  ubrzanja  $a$  koristimo

$$\frac{1}{6} \cdot \frac{\mathbb{E}^*(l'_{LF}^*(0)^3)}{\text{Var}^*(l'_{LF}^*(0))^3/2} \approx \frac{(R-1)^{3/2}}{6R} \cdot \frac{\sum_{r=1}^R l'_{LF,r}^*(0)^3}{\left[ \sum_{r=1}^R (l'_{LF,r}^*(0) - \overline{l'_{LF}^*(0)})^2 \right]^{3/2}} = \hat{a}, \quad (3.20)$$

budući da je procjena metodom maksimalne vjerodostojnosti parametra  $\zeta$  upravo  $\hat{\zeta} = 0$ . U (3.20) je  $l_{LF}^*$  funkcija log-vjerodostojnosti uzorka simuliranog iz razdiobe određene najnepovoljnijom familijom s parametrom  $\zeta = 0$ . Međutim, kako je

$$l'_{LF}(\zeta) = \hat{\delta}^\tau D l(\hat{\psi} + \zeta \hat{\delta})$$

imamo da je

$$l'_{LF}(0) = \sum_{i=1}^d \hat{\delta}_i \cdot \frac{\partial l}{\partial \psi_i}(\hat{\psi}) \quad (3.21)$$

pa možemo ustvari simulirati iz početne razdiobe s procijenjenim parametrima  $\hat{\psi}$  i koristeći jednakost (3.21) izračunati  $l'_{LF,r}^*(0)$  za sve  $r$ .

**Primjer 3.3.3.** Neka su dani podaci o količini oborina iz posijanih oblaka kao u Primjeru 3.2.1, ali pretpostavimo da ta opažanja dolaze iz gama razdiobe s (nepoznatim) parametrima  $\alpha > 0$  i  $\beta > 0$ . Ako je  $X \sim \Gamma(\alpha, \beta)$  onda je  $\mathbb{E}(X) = \frac{\alpha}{\beta}$  pa definirajmo  $\mu = \frac{\alpha}{\beta}$  te promatrajmo funkciju gustoće od  $X$  kao funkciju parametara  $\alpha$  i  $\mu$ , to jest  $f_X(x) = f_X(x; \alpha, \mu)$ . Tu smo reparametrizaciju učinili jer želimo pouzdani interval za očekivanu količinu oborina, pa sada ustvari tražimo  $BC_a$  95% pouzdani interval za  $\mu$ . Označimo još  $\psi = (\mu, \alpha)$ .

Uz ovu reparametrizaciju imamo da je funkcija log-vjerodostojnosti dana s

$$l(\mu, \alpha) = n\alpha \ln\left(\frac{\alpha}{\mu}\right) + (\alpha - 1) \cdot \sum_{i=1}^n \ln(x_i) - \frac{\alpha}{\mu} \cdot \sum_{i=1}^n x_i - n \ln(\Gamma(\alpha)).$$

Parcijalnim deriviranjem te funkcije lako dobijemo Hesseovu matricu od  $l$ :

$$Hl(\psi) = Hl(\mu, \alpha) = \begin{bmatrix} -\frac{n\alpha}{\mu^2} & 0 \\ 0 & \frac{n}{\alpha} + \frac{d^2}{d\alpha^2}\Gamma(\alpha) \end{bmatrix}.$$

Numerički možemo dobiti da su procjene metodom maksimalne vjerodostojnosti parametara  $\mu$  i  $\alpha$  dane s  $t = \hat{\mu} = 329.82$  i  $\hat{\alpha} = 1.018$ . Sada imamo da je

$$\hat{\delta} = \begin{bmatrix} \left(\frac{16 \cdot 1.018}{329.82^2}\right)^{-1} & 0 \\ 0 & \left(-\frac{16}{1.018} - \frac{d^2}{d\alpha^2}\Gamma(\alpha)\Big|_{\alpha=1.018}\right)^{-1} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = (6677.57, 0)^\tau.$$

Procjenu ubrzanja  $\hat{\alpha}$  sada možemo dobiti tako da simuliramo puno (na primjer,  $R = 100000$ ) nezavisnih slučajnih uzoraka iste duljine  $n = 16$  iz gama razdiobe s parametrima  $\hat{\alpha}$  i  $\hat{\mu}$  te iz svakog uzorka izračunamo  $l'_{LF,r}^*(0)$  pomoću (3.21) gdje je u ovom slučaju

$$l'_{LF,r}^*(0; x_{r,1}^*, \dots, x_{r,n}^*) = \hat{\delta}_1 \cdot \left( \frac{\hat{\alpha}}{\hat{\mu}^2} \cdot \sum_{i=1}^n x_{r,i}^* - \frac{n\hat{\alpha}}{\hat{\mu}} \right) = \frac{n\hat{\alpha}\hat{\delta}_1}{\hat{\mu}} \cdot \left( \frac{x_r^*}{\hat{\mu}} - 1 \right).$$



Na temelju simuliranih vrijednosti primijenimo (3.20) i s našim simulacijama dobivamo  $\hat{a} = 0.0845$ .

Za procjenu korekcije pristranosti simuliramo  $R = 9999$  nezavisnih slučajnih uzoraka iste duljine iz gama razdiobe s parametrima  $\hat{a}$  i  $\hat{\mu}$  i iz svakog simuliranog uzorka izračunamo procjenu  $t_r^* = \bar{x}_r^*$  (opažena uzoračka aritmetička sredina). U našim je simulacijama 5363 simuliranih procjena  $t_r^*$  manje ili jednako  $t = \hat{\mu} = 329.82$  pa je

$$\hat{w} = \Phi^{-1}\left(\frac{5363}{10000}\right) = 0.0911.$$

Iz vrijednosti  $\hat{a}$  i  $\hat{w}$  koristeći (3.17) dobivamo

$$\alpha_1 = \Phi\left(0.0911 + \frac{0.0911 + \overbrace{z_{0.025}}^{-1.96}}{1 - 0.0845 \cdot (0.0911 + z_{0.025})}\right) = \Phi(-1.517) = 0.0647$$

te

$$\alpha_2 = \Phi\left(0.0911 + \frac{0.0911 + \overbrace{z_{0.975}}^{1.96}}{1 - 0.0845 \cdot (0.0911 + z_{0.975})}\right) = \Phi(2.587) = 0.9952.$$

Dakle je donja granica  $BC_a$  95% pouzdanog intervala  $t_{((R+1)\alpha_1)}^* = t_{(647)}^* = 215.03$ , a gornja je granica  $t_{((R+1)\alpha_2)}^* = t_{(9952)}^* = 585.53$ . Stoga je  $BC_a$  95% pouzdani interval za  $\mu$  dan s [215.03, 585.53].

### Neparametarski modeli

$BC_a$  pouzdani intervali u neparametarskom slučaju dobivaju se tako da konstruiramo najnepovoljniju familiju razdioba iz multinomijalne razdiobe inducirane uzoračkim frekvencijama uzorka. Na dobivenu najnepovoljniju familiju razdioba primijenimo  $BC_a$  metodu za parametarski slučaj s jednim parametrom.

Neka je  $\theta$  nepoznati parametar za kojeg konstruiramo pouzdani interval i neka je  $T$  procjenitelj za  $\theta$  s opaženom vrijednosti  $t$ . Neka  $l(x_i)$  označava empirijsku procjenu funkcije utjecaja procjenitelja  $T$  u točki  $x_i$ . Definiramo najnepovoljniju familiju razdioba preko

$$p_i = \mathbb{P}(X^* = x_i) = \frac{e^{\eta l(x_i)}}{\sum_{k=1}^n e^{\eta l(x_k)}}$$

gdje je  $X^*$  slučajna varijabla iz iste razdiobe kao i svi  $X_i$  (elementi slučajnog uzorka) s ne nužno istim parametrima, a  $\eta$  nepoznati parametar i bijektivna funkcija  $g(\theta)$  od  $\theta$ . Procjena parametra  $\eta$  metodom maksimalne vjerodostojnosti je  $\hat{\eta} = 0$ .

Korekcija pristranosti  $\hat{w}$  računa se kao i ranije, dakle koristeći (3.18), gdje su sada  $t_r^*$  simulirane vrijednosti procjenitelja  $T$  dobivene iz slučajnih uzoraka simuliranih iz  $\hat{F}$  (empirijska funkcija distribucije početnog uzorka). To je posljedica činjenice da je  $\hat{\eta}_{MLE} = 0$  pa su  $p_i = \frac{1}{n}$ , za sve  $i \in \{1, 2, \dots, n\}$ .

Može se pokazati da je

$$l'(\theta) = g'(\theta) \cdot \left( \sum_{i=1}^n l(x_i^*) - n\mathbb{E}^*(\mathbf{IF}^*) \right).$$

Međutim, ukoliko gornji izraz evaluiramo u  $\theta = t$ , odnosno u  $\eta = \hat{\eta} = 0$  imamo da je  $\mathbb{E}^*(\mathbf{IF}^*) = 0$ . Ako primijenimo (3.19) onda se član  $g'(t)$  skрати i dobivamo da je

$$a = \frac{1}{6} \cdot \frac{\sum_{i=1}^n l^3(x_i)}{(\sum_{i=1}^n l^2(x_i))^{3/2}}. \quad (3.22)$$

Ukoliko je funkciju utjecaja teško ili nemoguće izračunati možemo koristiti *jackknife* procjenu danu s (1.6) pa dobivamo da je

$$\hat{a} = \frac{1}{6} \cdot \frac{\sum_{i=1}^n (t - t_{-i})^3}{(\sum_{i=1}^n (t - t_{-i})^2)^{3/2}}. \quad (3.23)$$

**Primjer 3.3.4.** Pronađimo  $BC_a$  95% pouzdani interval za očekivanu količinu oborina  $\mu$  iz Primjera 3.2.1 bez ikakvih pretpostavki na razdiobu iz koje opažanja dolaze.

Budući da je procjenitelj  $T = \bar{X}$  uzoračka aritmetička sredina, znamo iz Primjera 1.2.1 da je empirijska procjena funkcije utjecaja  $l(x_i) = x_i - \bar{x}$ . Stoga iz (3.22) slijedi da je ubrzanje jednako  $a = 0.0352$ . Korištenje formule (3.23) daje identičan rezultat.

Korekciju pristranosti  $\hat{w}$  dobivamo tako da simuliramo  $R = 9999$  nezavisnih slučajnih uzoraka iste duljine s ponavljanjem iz početnog skupa opažanja (dakle, iz empirijske funkcije distribucije  $\hat{F}$ ) i iz svakog simuliranog uzorka izračunamo procjenu  $t_r^* = \bar{x}_r^*$  (opažena uzoračka aritmetička sredina). U našim je simulacijama 5143 simuliranih procjena  $t_r^*$  manje ili jednako  $t = 329.82$  pa je

$$\hat{w} = \Phi^{-1}\left(\frac{5143}{10000}\right) = 0.0359.$$

Iz vrijednosti  $a$  i  $\hat{w}$  koristeći (3.17) dobivamo

$$\alpha_1 = \Phi\left(0.0359 + \frac{0.0359 + \overbrace{z_{0.025}}^{-1.96}}{1 - 0.0352 \cdot (0.0359 + z_{0.025})}\right) = \Phi(-1.766) = 0.0387$$

te

$$\alpha_2 = \Phi \left( 0.0359 + \frac{0.0359 + \overbrace{z_{0.975}}^{=1.96}}{1 - 0.0352 \cdot (0.0359 + z_{0.975})} \right) = \Phi(2.1825) = 0.9855.$$

Dakle je donja granica  $BC_a$  95% pouzdanog intervala  $t_{((R+1)\alpha_1)}^* = t_{(387)}^* = 214.51$ , a gornja je granica  $t_{((R+1)\alpha_2)}^* = t_{(9855)}^* = 482.25$ . Stoga je neparametarski  $BC_a$  95% pouzdani interval za  $\mu$  dan s  $[214.51, 482.25]$ .

### 3.4 Pouzdani intervali uz dvostruki *bootstrap*

Sve metode konstrukcije pouzdanih intervala temeljene na simulacijama koje smo dosad opisali su aproksimativne, to jest pokrivenost dobivenih pouzdanih intervala može značajno odudarati od željene pokrivenosti  $1 - \alpha$  čak i ako iskoristimo proizvoljno velik broj simulacija  $R$ . Ta razlika između dobivene i željene pokrivenosti posljedica je pristranosti koju svaka od metoda posjeduje. Zato je poželjno nekako korigirati tu pristranost kako bismo dobili točnije procjene. Jedna od mogućih metoda je korištenje dvostrukog *bootstrapa* koju smo već koristili kod korekcije  $p$ -vrijednosti testova u prethodnom poglavlju. Metodu ćemo primijeniti na osnovni pouzdani interval iako će iz postupka biti jasno kako se ona može primijeniti i na ostale pouzdane intervale (na primjer, studentizirane ili  $BC_a$  pouzdane intervale).

Neka je  $F$  funkcija distribucije naših opažanja,  $\theta$  nepoznati parametar te razdiobe,  $T$  njegov procjenitelj i  $\hat{F}$  procjena funkcije distribucije. Tada možemo pisati  $T = t(\hat{F})$  i  $\theta = t(F)$ . Ako je, kao i ranije,  $a_\alpha$  kvantil razdiobe od  $T - \theta$  onda za njega vrijedi

$$\mathbb{P}(T - \theta \leq a_\alpha \mid F) = \mathbb{P}(t(\hat{F}) - t(F) \leq a_\alpha(F) \mid F) = \alpha,$$

to jest  $t(\hat{F}) - a_\alpha(F)$  je (egzaktna) gornja granica pouzdanog intervala. Međutim, mi ne znamo egzaktno kvantil  $a_\alpha(F)$ , nego ga procjenjujemo s  $a_\alpha(\hat{F})$  tako da je naša procjena gornje granice pouzdanog intervala  $t(\hat{F}) - a_\alpha(\hat{F})$ . Zbog moguće pogreške u procjeni i pristranosti same metode općenito je  $a_\alpha(\hat{F}) \neq a_\alpha(F)$  tako da je

$$\mathbb{P}(t(F) \leq t(\hat{F}) - a_\alpha(\hat{F}) \mid F) \neq 1 - \alpha,$$

odnosno procjena pouzdanog intervala nema željenu pokrivenost  $1 - \alpha$ .

Jedan mogući pristup ovom problemu je da se pokuša korigirati vrijednost  $a_\alpha(\hat{F})$ . Međutim, pokazuje se da je bolje korigirati broj  $\alpha$  u toj procjeni  $a_\alpha(\hat{F})$ , odnosno, umjesto  $\alpha$  stavljamo  $q(\alpha)$  i procjenjujemo tu korekciju. Dakle, željeli bismo da je  $q(\alpha)$  takav da

$$\mathbb{P}(t(F) \leq t(\hat{F}) - a_{q(\alpha)}(\hat{F}) \mid F) = 1 - \alpha. \quad (3.24)$$

Naravno, ta će vrijednost ovisiti o (nepoznatoj)  $F$ , to jest  $q(\alpha) = q(\alpha; F)$  pa kako je  $F$  nepoznata ustvari procjenjujemo  $q(\alpha)$  s  $\hat{q}(\alpha) = q(\alpha; \hat{F})$ . Neka je  $\mathbb{P}^*$  vjerojatnost u odnosu na razdiobu dobivenu simuliranjem, a  $\mathbb{P}^{**}$  vjerojatnost u odnosu na razdiobu dobivenu simuliranjem iz simuliranog uzorka. U odnosu na vjerojatnost  $\mathbb{P}^*$  procjena  $\hat{q}(\alpha)$  zadovoljava

$$\mathbb{P}^*(t(\hat{F}) \leq t(\hat{F}^*) - a_{\hat{q}(\alpha)}(\hat{F}^*) \mid \hat{F}) = 1 - \alpha \quad (3.25)$$

što se može riješiti uz pomoć *bootstrapa*. Iz definicije kvantila  $a_\alpha(F)$  slijedi da se jednadžba (3.25) može zapisati kao

$$\mathbb{P}^*(\mathbb{P}^{**}(T^{**} \leq 2T^* - t \mid \hat{F}^*) \geq \hat{q}(\alpha) \mid \hat{F}) = 1 - \alpha, \quad (3.26)$$

gdje je sada  $T^*$  procjenitelj iz simuliranog uzorka,  $T^{**}$  procjenitelj dobiven simuliranjem iz već simuliranog uzorka, a  $t$  opažena vrijednost procjenitelja  $T$ . Naravno,  $\hat{F}^*$  predstavlja procjenu funkcije distribucije  $F$  dobivenu iz simuliranog uzorka.

Dakle, simuliramo  $R$  nezavisnih slučajnih uzoraka iz funkcije distribucije  $\hat{F}$  i iz svakog simuliranog uzorka ponovno procijenimo  $F$  s  $\hat{F}_r^*$ . Stavimo

$$u_r^* = \mathbb{P}(T^{**} \leq 2t_r^* - t \mid \hat{F}_r^*)$$

što procjenjujemo tako da simuliramo  $M$  nezavisnih slučajnih uzoraka iz funkcije distribucije  $\hat{F}_r^*$  i iz svakog od tih  $M$  uzoraka izračunamo vrijednost procjenitelja  $t_{r,m}^{**}$ . Procjena za  $u_r^*$  je tada dana s

$$u_{M,r}^* = \frac{1}{M} \cdot \sum_{m=1}^M \mathbf{1}_{\{t_{r,m}^{**} \leq 2t_r^* - t\}}$$

što je ustvari proporcija simuliranih procjenitelja  $t_{r,m}^{**}$  koji su manji ili jednaki  $2t_r^* - t$ . Lijeva strana u (3.26) se tada procjenjuje kao

$$\frac{1}{R} \cdot \sum_{r=1}^R \mathbf{1}_{\{u_{M,r}^* \geq \hat{q}(\alpha)\}}$$

i želimo da je ta procjena jednaka danoj pokrivenosti  $1 - \alpha$ . Dakle,  $\hat{q}(\alpha)$  dobiva se tako da uredimo simulirane vrijednosti  $u_{M,r}^*$  u neopadajući niz  $u_{M,(1)}^* \leq \dots \leq u_{M,(R)}^*$  i stavimo  $\hat{q}(\alpha) = u_{M,((R+1)\alpha)}^*$ .

Primijetimo da smo ovim postupkom dobili korekciju samo za gornju granicu pouzdanog intervala. Za korekciju donje granice pouzdanog intervala stavljamo  $\hat{q}(1 - \alpha) = u_{M,((R+1)(1-\alpha))}^*$ . Napomenimo još da je dovoljan broj simulacija  $M = 249$ , ali ipak moramo smanjiti  $R$  na  $R = 999$ . Ovaj je postupak sažet u sljedećem algoritmu:

**Algoritam 3.4.1.** Za  $r \in \{1, 2, \dots, R\}$  radi:

1. Generiraj  $x_1^*, \dots, x_n^*$  iz  $\hat{F}$  i iz njih izračunaj  $t_r^*$ .
2. Nađi procjenu funkcije distribucije  $\hat{F}_r^*$  gornjeg uzorka.
3. Za  $m \in \{1, 2, \dots, M\}$  radi :
  - 3.1. Generiraj  $x_1^{**}, \dots, x_n^{**}$  iz  $\hat{F}_r^*$ .
  - 3.2. Izračunaj vrijednost procjenitelja  $t_{r,m}^{**}$ .
4. Izračunaj  $u_{M,r}^*$ .

Poredaj  $u_{M,r}^*$  u neopadajući niz  $u_{M,(1)}^* \leq \dots \leq u_{M,(R)}^*$  i izračunaj  $\hat{q}(\alpha) = u_{M,((R+1)\alpha)}^*$  te  $\hat{q}(1-\alpha) = u_{M,((R+1)(1-\alpha))}^*$ .

Izračunaj granice pouzdanog intervala:  $\hat{\theta}_\alpha = 2t - t_{((R+1)\hat{q}(1-\alpha))}^*$  i  $\hat{\theta}_{1-\alpha} = 2t - t_{((R+1)\hat{q}(\alpha))}^*$ .

Razmatranja u [2] pokazuju da ova metoda korigira pristranost osnovnog pouzdanog intervala za jedan red veličine, to jest, ako dobiveni osnovni pouzdani interval ima pokrivenost  $1 - \alpha + O(n^{-1})$  onda je uz ovu korekciju pokrivenost intervala  $1 - \alpha + O(n^{-2})$ .

**Primjer 3.4.2.** Za podatke iz Primjera 3.2.1 i Primjera 3.2.2 pronađimo korigirani osnovni 95% pouzdani interval za očekivanu količinu oborina  $\mu$ .

Ukoliko smo u situaciji iz Primjera 3.2.1 onda pretpostavljamo eksponencijalni model za naša opažanja s procijenjenim parametrom  $\hat{\lambda} = \frac{1}{t} = \frac{1}{\bar{x}} = \frac{1}{329.82}$  pa simuliramo iz te procijenjene razdiobe. Prateći Algoritam 3.4.1 dobivamo  $\hat{q}(0.025) = u_{M,(25)}^* = 0$  i  $\hat{q}(0.975) = u_{M,(975)}^* = 0.920$ . Budući da je  $t_{((R+1)\hat{q}(\alpha))}^* = t_{(0)}^*$  nepostojeća vrijednost uzet ćemo  $t_{(1)}^* = 120.20$  umjesto te vrijednosti i dobiti da je gornja granica korigiranog osnovnog 95% pouzdanog intervala za  $\mu$  jednaka

$$\hat{\theta}_{1-\alpha} = 2 \cdot 329.82 - 120.20 = 539.44.$$

Slično, kako je  $t_{((R+1)\hat{q}(1-\alpha))}^* = t_{(920)}^* = 455.02$  imamo da je donja granica korigiranog osnovnog 95% pouzdanog intervala za  $\mu$  jednaka

$$\hat{\theta}_\alpha = 2 \cdot 329.82 - 455.02 = 204.62.$$

Dakle je procjena 95% pouzdanog intervala [204.62, 539.44]. Uspoređujući s rezultatima iz Primjera 3.2.1 vidimo da je ovaj interval puno bliži egzaktnom pouzdanom intervalu nego obični osnovni pouzdani interval.

Ako smo u situaciji Primjera 3.2.2 onda smo u neparametarskom modelu pa simuliramo iz empirijske funkcije distribucije  $\hat{F}$  što znači uzimanje slučajnog uzorka duljine  $n$  s

ponavljanjem iz početnog skupa opažanja. Prateći Algoritam 3.4.1 dobivamo  $\hat{q}(0.025) = u_{M,(25)}^* = 0$  i  $\hat{q}(0.975) = u_{M,(975)}^* = 0.972$ . Budući da je ponovno  $t_{((R+1)\hat{q}(\alpha))}^* = t_{(0)}^*$  nepostojeća vrijednost uzet ćemo  $t_{(1)}^* = 146.28$  umjesto te vrijednosti i dobiti da je gornja granica korigiranog osnovnog 95% pouzdanog intervala za  $\mu$  jednaka

$$\hat{\theta}_{1-\alpha} = 2 \cdot 329.82 - 146.28 = 513.37.$$

Slično, kako je  $t_{((R+1)\hat{q}(1-\alpha))}^* = t_{(972)}^* = 457.57$  imamo da je donja granica korigiranog osnovnog 95% pouzdanog intervala za  $\mu$  jednaka

$$\hat{\theta}_{\alpha} = 2 \cdot 329.82 - 457.57 = 202.07.$$

Dakle je procjena 95% pouzdanog intervala  $[202.07, 513.37]$ . Uspoređujući s rezultatima iz Primjera 3.2.2 vidimo da je i ovaj interval također puno bliži egzaktnom pouzdanom intervalu nego obični osnovni pouzdani interval te da su dva intervala dobivena u ovom primjeru vrlo slična.

### 3.5 Pouzdana područja

Dosad smo se fokusirali isključivo na slučaj jednog nepoznatog parametra  $\theta \in \mathbb{R}$ . Međutim, ponekad želimo konstruirati pouzdana područja koja su višedimenzionalni analogon pouzdanih intervala za nekoliko parametara. Neka je  $\theta = (\theta_1, \dots, \theta_d)^\tau \in \mathbb{R}^d$   $d$ -dimenzionalni nepoznati parametar. Jedna moguća ideja je konstruirati pouzdane intervale  $C_{\alpha,i}$  za svaki parametar  $\theta_i$  posebno i definirati  $C_{\alpha} = C_{\alpha,1} \times \dots \times C_{\alpha,d}$  kao pouzdano područje za parametar  $\theta$  (to je tada  $d$ -dimenzionalni kvadar). Međutim, ako želimo da je pokrivenost pouzdanog područja  $1 - \alpha$  tada jednodimenzionalni intervali  $C_{\alpha,i}$  ne mogu uvijek imati pokrivenost također  $1 - \alpha$ . Naime,

$$\mathbb{P}(\theta \notin C_{\alpha}) = \mathbb{P}\left(\bigcup_{i=1}^d \{\theta_i \notin C_{\alpha,i}\}\right) \leq \sum_{i=1}^d \underbrace{\mathbb{P}(\theta_i \notin C_{\alpha,i})}_{=\alpha} = \sum_{i=1}^d \alpha = d\alpha.$$

Međutim, ako stavimo  $\alpha_i = \frac{\alpha}{d}$  za sve  $i \in \{1, 2, \dots, d\}$  (takozvana *Bonferronijeva procedura*) i konstruiramo  $(1 - \alpha_i) \cdot 100\%$  pouzdane intervale za parametre  $\theta_i$ , onda će pokrivenost pouzdanog područja definiranog kao gore biti *najmanje*  $1 - \alpha$  kao što smo i htjeli. Ta je pokrivenost jako često puno veća od  $1 - \alpha$  jer je gornja ocjena dosta konzervativna. Osim toga, procjene za parametre  $\theta_i$  mogu biti međusobno korelirane pa pouzdano područje u obliku kvadra nema smisla. Naime, oblik kvadra je rezultat toga da su pouzdani intervali za  $\theta_i$  konstruirani nezavisno jedan od drugog, ali među procjenama možda postoji korelacija pa je neki drugi oblik puno prikladniji (na primjer, često se u praksi pojavljuje elipsoid).

Samim time pouzdano područje u obliku kvadra ima značajno veću pokrivenost nego što želimo. Zbog ovih razloga treba nam neki drugi pristup konstrukciji pouzdanih područja.

Neka je  $T$   $d$ -dimenzionalni procjenitelj za parametar  $\theta \in \mathbb{R}^d$ , a  $V$  njegova kovarijacijska matrica. Definiramo statistiku

$$Q = (T - \theta)^\tau V^{-1} (T - \theta). \quad (3.27)$$

Primijetimo da je statistika  $Q$  višedimenzionalni analogon kvadrata studentizirane statistike  $T$ . Kad bismo znali egzaktne kvantile  $a_p$  razdiobe od  $Q$  onda bi  $(1 - \alpha) \cdot 100\%$  pouzdano područje za  $\theta$  bilo

$$C_\alpha = \left\{ \theta \in \mathbb{R}^d; Q \leq a_{1-\alpha} \right\}.$$

Ukoliko  $T$  aproksimativno prati višedimenzionalnu normalnu razdiobu tada će razdioba od  $Q$  biti aproksimativno hi-kvadrat razdioba s  $d$  stupnjeva slobode. Međutim, često je upitno ili teško provjeriti je li pretpostavka multivarijatne normalnosti od  $T$  ispunjena pa je poželjno aproksimirati razdiobu od  $Q$  metodom *bootstrapa*. Dakle, simuliramo  $R$  nezavisnih slučajnih uzoraka iste duljine i iz svakog uzorka izračunamo vrijednost procjenitelja  $t_r^*$  te procijenimo kovarijacijsku matricu  $v_r^*$  procjenitelja  $T$ . Izračunamo  $q_r^* = (t_r^* - t)^\tau v_r^{*-1} (t_r^* - t)$  i te vrijednosti poredamo u neopadajući niz  $q_{(1)}^* \leq \dots \leq q_{(R)}^*$ . Tada je procjena kvantila  $a_{1-\alpha}$  dana s  $(R + 1)(1 - \alpha)$ -tom vrijednosti u uređenom nizu simuliranih vrijednosti  $q_r^*$ , odnosno s  $q_{((R+1)(1-\alpha))}^*$  te je *bootstrap*  $(1 - \alpha) \cdot 100\%$  pouzdano područje za  $\theta$  dano s

$$\left\{ \theta \in \mathbb{R}^d; (t - \theta)^\tau v^{-1} (t - \theta) \leq q_{((R+1)(1-\alpha))}^* \right\}. \quad (3.28)$$

Simulacije se u parametarskom slučaju izvode iz pretpostavljene razdiobe opažanja, a u neparametarskom slučaju iz empirijske funkcije distribucije  $\hat{F}$ . U parametarskom slučaju kovarijacijsku matricu  $V$  procjenitelja  $T$  možemo naći pomoću delta metode. U slučaju da je  $T$  procjenitelj metodom maksimalne vjerodostojnosti možemo naći asimptotsku kovarijacijsku matricu procjenitelja  $T$  pomoću Fisherove informacijske matrice. Mi ćemo se u našem primjeru koji slijedi poslužiti *bootstrapom* kako bismo procijenili kovarijacijsku matricu procjenitelja  $T$ . Naravno, ta metoda može poslužiti i u neparametarskom slučaju.

Kao i kod pouzdanih intervala, poželjno je koristiti (ako je moguće) neku transformaciju  $h = (h_1, \dots, h_d) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ . Na primjer, ako je  $D$  dijagonalna matrica s elementima  $\frac{\partial h_i}{\partial \theta_i}$  na dijagonali onda računamo  $q = (h(t) - h(\theta))^\tau (D^\tau v D)^{-1} (h(t) - h(\theta))$  koje procjenjujemo *bootstrapom* i simulacijama. Naime, simuliramo velik broj uzoraka i iz svakog uzorka izračunamo  $q_r^*$  te definiramo *bootstrap*  $(1 - \alpha) \cdot 100\%$  pouzdano područje za  $\theta$  kao

$$\left\{ \theta \in \mathbb{R}^d; (h(t) - h(\theta))^\tau (D^\tau v D)^{-1} (h(t) - h(\theta)) \leq q_{((R+1)(1-\alpha))}^* \right\} \quad (3.29)$$

Funkciju  $h$ , odnosno njene komponente  $h_i$  biramo istim metodama kao i kod osnovnih pouzdanih intervala.

**Primjer 3.5.1.** U svrhu istraživanja prostorne organizacije ljudskog kromosoma u staničnoj jezgri, parovi DNA nizova na slučajno odabranim mjestima u ljudskom kromosomu označeni su fluorescentnim bojama i mjerene su udaljenosti između baznih parova (nukleobaza povezanih vodikovim vezama). U sljedećoj je tablici dano tih  $n = 20$  mjerenja, preuzetih iz [9].

$i$	1	2	3	4	5	6	7	8	9	10
$x_i$	4.41	1.64	1.98	4.51	3.71	1.27	5.20	4.35	3.13	0.94
$i$	11	12	13	14	15	16	17	18	19	20
$x_i$	0.25	0.98	2.24	1.30	2.41	3.23	2.51	3.32	0.78	1.60

Pretpostavljamo da dani podaci prate gama razdiobu  $s$  (nepoznatim) parametrima  $\alpha > 0$  i  $\beta > 0$  te želimo naći 80%, 95% i 99% pouzdana područja za vektor parametara  $\theta = (\alpha, \beta)$ .

Neka je  $T = (\hat{\alpha}, \hat{\beta})$  procjenitelj metodom maksimalne vjerodostojnosti parametra  $\theta$ . Numerički možemo naći da su opažene vrijednosti tih procjenitelja  $\hat{\alpha} = 2.503$  i  $\hat{\beta} = 1.006$ , dakle,  $t = (2.503, 1.006)^\tau$ .

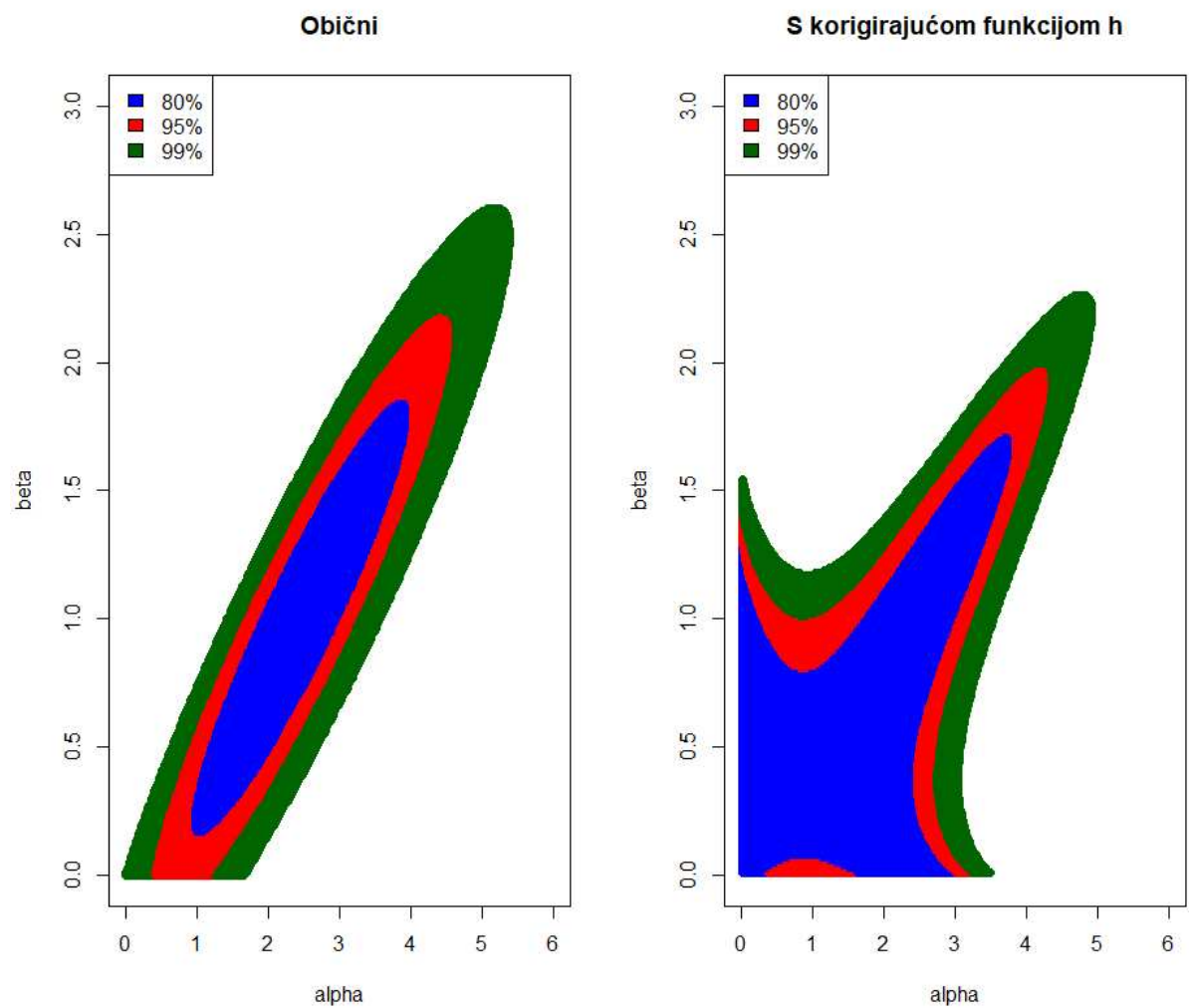
Procjenu u kovarijacijske matrice  $V$  procjenitelja  $T$  tražimo metodom bootstrapa. Simuliramo  $M = 499$  nezavisnih slučajnih uzoraka iste duljine  $n = 20$  iz gama razdiobe s parametrom  $t$  (gornja procjena) i iz svakog uzorka izračunamo opažene procjene parametara  $\alpha_m^*$  i  $\beta_m^*$ . Kovarijacijsku matricu procjenjujemo uz pomoć tih simulacija – na glavnoj dijagonali će se nalaziti (uzoračke) varijance simuliranih vrijednosti (zajedno s početnom procjenom), a na sporednoj dijagonali kovarijanca između vrijednosti  $\alpha_m^*$  i  $\beta_m^*$ . U našim simulacijama je

$$v = \begin{bmatrix} 0.946 & 0.407 \\ 0.407 & 0.204 \end{bmatrix}.$$

Napomenimo da smo uzeli  $M = 499$  budući da će ovu procjenu i simulacije biti potrebno napraviti za svaki od  $R = 2999$  simuliranih uzoraka (zato smo i skratili  $R$  na 2999). Dakle, simuliramo  $R = 2999$  nezavisnih slučajnih uzoraka iste duljine  $n = 20$  iz gama razdiobe s parametrom  $t$  i iz svakog uzorka izračunamo procjenu  $t_r^*$  i procjenu kovarijacijske matrice  $v_r^*$  (ponovno metodom bootstrapa kao i za početni uzorak). Iz tih vrijednosti izračunamo  $q_r^* = (t_r^* - t)^\tau v_r^{*-1} (t_r^* - t)$  i sve te vrijednosti poredamo u neopadajući niz. U našim simulacijama je  $q_{(2400)}^* = 1.747$ ,  $q_{(2850)}^* = 3.365$  i  $q_{(2970)}^* = 6.143$ . Dakle su pouzdana područja određena s (3.28) i prikazana na Slici 3.2.

Po uzoru na Primjer (3.2.1) i Primjer (3.2.2) definirali smo funkciju  $h(t_1, t_2) = (\ln(t_1), \ln(t_2))$  i ponovili gornji postupak, ali računajući  $q$  kao u (3.29). U našim simulacijama je  $q_{(2400)}^* = 2.068$ ,  $q_{(2850)}^* = 4.382$  i  $q_{(2970)}^* = 8.126$ . Pouzdana su područja određena s (3.29) i također prikazana na Slici 3.2.





Slika 3.2: Pouzdana područja za Primjer 3.5.1. Lijevo: obična; desno: uz korigirajuću funkciju  $h$



# Bibliografija

- [1] A. T. Craig, R. V. Hogg i J. W. McKean, *Introduction to Mathematical Statistics*, Pearson, 2018.
- [2] A. C. Davison i D. V. Hinkley, *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, 1997.
- [3] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, 1982.
- [4] B. Efron i B. Narasimhan, *The automatic construction of bootstrap confidence intervals*, (2018), 4–8, <http://statweb.stanford.edu/~ckirby/brad/papers/2018Automatic-Construction-BCIs.pdf>.
- [5] B. Efron i R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
- [6] P. I. Good, *Permutation, Parametric and Bootstrap Tests of Hypotheses*, Springer, 2005.
- [7] R. F. Heckard i J. M. Utts, *Mind on Statistics*, Cengage Learning, 2014.
- [8] D. C. Montgomery i G. C. Runger, *Applied Statistics and Probability for Engineers*, Wiley, 2013.
- [9] J. A. Rice, *Mathematical Statistics and Data Analysis*, Belmont, CA: Thomson/Brooks/Cole, 2007.
- [10] G. J. Szekely i M. L. Rizzo, *Energy Statistics: A Class of Statistics Based on Distances*, (2003), 9–10, [https://eric-bunch.github.io/static/Szekely\\_estats.pdf](https://eric-bunch.github.io/static/Szekely_estats.pdf).



# Sažetak

U ovome radu ilustrirali smo korištenje metode *bootstrapa* u svrhu testiranja statističkih hipoteza i konstrukcije pouzdanih intervala za parametre raznih modela. Iako su za obje ove metode statističkog zaključivanja izvedeni mnogobrojni teorijski rezultati, u praksi često pretpostavke pod kojima ti rezultati vrijede nisu ispunjeni. Tada je potrebno posegnuti za aproksimativnim metodama. Glavna je ideja ove metode procijeniti određene parametre iz uzorka i koristiti te procjene kao da su pravi parametri. Intuitivno, sva informacija o populaciji sadržana je u uzorku pa ovim postupkom tretiramo uzorak kao populaciju. Simuliranjem iz procijenjene razdiobe dobivamo ideju kako izgleda uzoračka razdioba neke statistike od interesa. Iako je metoda aproksimativna, kroz mnogobrojne primjere pokazali smo da daje zadovoljavajuće dobre rezultate te brzu i direktnu procjenu. Pokazali smo i kako se svaka od tih metoda može dodatno korigirati ne bi li rezultati bili još precizniji. Razvili smo metode koje iznenađujuće dobro funkcioniraju i u slučajevima kad nemamo nikakvih pretpostavki na razdiobu iz koje dolaze naša opažanja. Naravno, ako imamo razuman pretpostavljeni model, rezultati su još i bolji. Svaku predloženu metodu ilustrirali smo primjerima gdje smo koristili simulacije na računalu koje znatno olakšavaju izvrednjavanje kompliciranih izraza (na primjer, raznih vjerojatnosti) na koje se te metode oslanjaju.



# Summary

In this thesis we've illustrated the use of the *bootstrap* method for testing statistical hypotheses and constructing confidence intervals for parameters of various models. Even though there is a great number of theoretical results derived for both of these methods of statistical inference, often the assumptions under which those results work are not met in practice. This is the reason why we need to resort to approximate methods. The main point of this method is to use the available sample to estimate certain parameters and then use those estimates as if they were the real parameter values. Intuitively, all the information about the population of interest is contained in the sample, so basically we treat the sample as the population. By simulating samples from the fitted distribution we get an insight on the shape of sampling distribution of statistic of interest. Although this method is only approximate, we've shown through numerous examples how efficient it can be and that it results in satisfying and quick estimates. We've also demonstrated how each of the methods we presented can be improved so that they give adjusted and even more precise results. We've developed various methods that work surprisingly well even when no assumptions are being made on the distribution of the sample values. Certainly the results are even more accurate when we have a parametric model available. Every method that we've presented was accompanied by an example illustrating the use of the method. In the examples we used computer simulation to help us compute complicated expressions (for instance, some complicated probabilities) which are used in the methods.





# Životopis

Rođen sam 20. ožujka 1997. godine u Zagrebu. Završio sam Osnovnu školu Brezovica 2011. godine. Iste godine upisujem V. gimnaziju u Zagrebu koju završavam 2015. godine. Nakon završetka srednje škole upisujem preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu. Godine 2018. upisujem diplomski studij Matematička statistika. Na posljednjoj godini studija osvajam nagradu fakulteta za uspjeh na studiju.