ACTA MATHEMATICA SPALATENSIA
Series didactica
Vol.6 (2023) 65–85
https://doi.org/10.32817/amssd.6.5

Zaprimljen: 14. ožujka 2023.
Prihvaćen: 12. lipnja 2023.

Statističko modeliranje ekstrema

Ana Marija Filipas, Ivana Slamić

Sažetak

U različitim područjima, od hidrologije, seizmologije, telekomunikacija i biomedicine pa do ekonomije i aktuarstva, postoji potreba za analizom ekstremnih događaja – događaja čija je vjerojatnost pojavljivanja vrlo mala, a čije posljedice, nasuprot tome, mogu biti vrlo velike. Teorija ekstremnih vrijednosti grana je statistike koja proučava upravo takve događaje. U članku su opisane osnove statističkog modeliranja ekstremnih događaja primjenom distribucija ekstremnih vrijednosti, vjerojatnosnih distribucija koje opisuju granično ponašanje maksimuma za niz nezavisnih i jednako distribuiranih slučajnih varijabli. Primjeri su riješeni upotrebom programskog jezika R.

Ključni pojmovi: vjerojatnost i statistika, teorija ekstremnih vrijednosti, distribucije ekstremnih vrijednosti

1. Uvod - što smatramo ekstremnim događajem?

S vremena na vrijeme priroda nas iznenadi svojim ekstremnim ponašanjem. Prisjetimo se tako samo nekih od nedavnih:

- 8.1.2021. Madrid i okolicu pogodilo snježno nevrijeme te je zabilježen snježni pokrivač od 50 cm (najveći od 1971. godine);
- 23.9.2022. Floridu i Južnu Karolinu pogodio uragan Ian koji je odnio najmanje 42 života, a bez struje je ostalo 2 milijuna ljudi;
- 28.9.2022. u Rijeci je palo 93 litre kiše u jednom satu, čime je oboren apsolutni rekord;
- 27.11.2022. najveći aktivni vulkan na svijetu, Mauna Loa na Havajima, eruptirao je prvi put nakon gotovo 40 godina.

Slika 1. Različite prirodne katastrofe ubrajaju se u ekstremne događaje – događaje koji se ne javljaju često, ali kada se dogode, sa sobom donose velike štete: (a) uragan Ian za vrijeme najjačeg intenziteta, 23. rujna 2022. (b) poplava u New Orleansu kao posljedica uragana Katrina, 28. kolovoza 2005. (c) erupcija vulkana Mauna Loa, studeni 2022.

Svaka od tih ekstremnih situacija često donosi sa sobom ogromne štete i mnoštvo pitanja. Možemo li takve pojave predvidjeti ili barem kontrolirati? Ako se u nekom trenutku pojavio takav ekstremni događaj te je najjači po intenzitetu u posljednjih 50 ili 100 godina, znači li to da se u sljedećih 50 ili 100 godina više neće javiti? Možemo li na temelju podataka iz prošlosti procijeniti vjerojatnost da se u razdoblju od 100 godina dogodi poplava prilikom koje će razina vode premašiti određenu vrijednost? Što ako podaci postoje samo za neki kraći period, primjerice 30 godina?

Događaji kojih smo se prethodno prisjetili ubrajaju se u klasu onih čiji ishod ne možemo sa sigurnošću predvidjeti, a upravo takvi događaji predmet su interesa teorije vjerojatnosti i matematičke statistike. Termin „100 year flood" opisuje poplavu sa svojstvom da je vjerojatnost da se dogodi takva ili veća u bilo kojoj godini jednaka 0.01. Iako bi doslovni prijevod mogao sugerirati da je to „poplava koja se dogodi jednom u 100 godina", to je daleko od ispravne interpretacije – u istom smislu kao kada bismo na temelju činjenice da je vjerojatnost pojave jedinice prilikom bacanja kocke 1/6 izveli zaključak da će se, ako kocku bacimo 12 puta, jedinica pojaviti točno dva puta. Kao što ni prilikom bacanja novčića ne možemo predvidjeti ishod sljedećeg bacanja (bez obzira je li se, primjerice, u prethodnih 5, 10 ili 50 bacanja pojavilo samo pismo), tako ne možemo očekivati ni da će informacija o vjerojatnosti takvog događaja dati precizan odgovor na pitanje kada će se on realizirati, no dobar statistički model pomoći će u kontroli i smanjenju šteta. Primjerice, na temelju statističkih modela za razinu mora formiranih za razdoblje od 100 godina (čak i ako su poznati podaci samo za proteklih 30 godina), možemo dobiti informaciju o visini valobrana koji treba izgraditi. Ponašanje ovih pojava, s druge strane, podložno je i promjenama uzrokovanim raznoraznim vanjskim utjecajima pa tako i vjerojat-

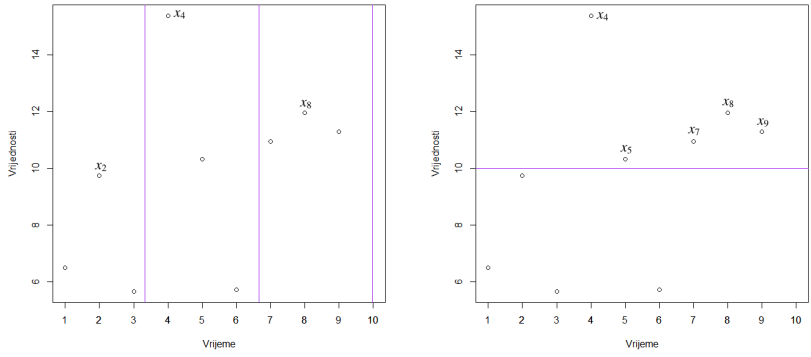
nosti stogodišnjih poplava posljednjih godina postaju sve veće, kao posljedica promjene klime, što dodatno otežava odabir modela.

Slika 2. Prirodne katastrofe ne možemo spriječiti, a najčešće ni predvidjeti, no dobrim modelima možemo umanjiti njihove posljedice. (a) Izgradnja valobrana na temelju podataka iz prošlosti spriječit će štetu koju mogu izazvati izrazito veliki oceanski valovi. (b) Prilikom poplave 1993. godine u Missouriju, SAD, razina vode premašila je nivo 100-godišnjih poplava.

No, modeliranje ekstremnih događaja nije važno samo u kontroli štete uzrokovane vremenskim nepogodama te svoju primjenu ne pronalazi samo u hidrologiji, već i u drugim područjima, kao što su ekonomija, financije te aktuarstvo, u kojima je također potrebno pokušati predvidjeti ili kontrolirati velike gubitke. Primjerice, zadaća sektora rizika unutar određene banke je zaštititi banku od takvih gubitaka. U tu svrhu, potrebno je modelirati financijske vremenske nizove te koristiti metode teorije ekstremnih vrijednosti kako bismo procijenili potencijalni rizik s kojim će se ta banka susresti u budućnosti. Slična pitanja, vezana uz velike štete, zanimaju i aktuare te se zbog toga teorija ekstremnih vrijednosti sve više koristi u teoriji rizika te osiguranju.

Pretpostavimo sada da imamo podatke koje želimo analizirati. Prirodno pitanje je što uopće smatrati ekstremnim događajem. U teoriji ekstremnih vrijednosti postoje dva načina na koje možemo izdvojiti ekstremne vrijednosti iz zadanog skupa podataka. Kod prvog načina vremensko razdoblje u kojem promatramo podatke dijelimo na određeni broj segmenata/blokova, a zatim uzmemo najveću vrijednost u svakom od tih blokova. Drugi način odabira ekstremnih vrijednosti je izdvajanjem vrijednosti koje su veće ili jednake od neke zadane vrijednosti, koju nazivamo *prag*. Sve vrijednosti koje su veće ili jednake od praga u tom slučaju ćemo smatrati ekstremnim vrijednostima. Drugi način najčešće se koristi u analizi financijskih podataka te podataka u osiguranju, dok se prvi način uvelike koristi za analizu podataka koje karakterizira sezonalost, kao što su primjerice klimatološki i hidrološki podaci.

Kod prvog pristupa, cilj je odrediti vjerojatnost da maksimum poprimi vrijednost veću od zadane, a odgovor na takvo pitanje možemo dati ako poz-



Slika 3. Izdvajanje ekstremnih vrijednosti dvjema metodama. (a) Ako vremensko razdoblje u kojem promatramo podatke podijelimo na tri jednaka podsegmenta, vrijednosti x_2, x_4 i x_8 bismo smatrali ekstremnim vrijednostima. (b) Ako bismo za prag uzeli npr. $u = 10$, vrijednosti x_4, x_5, x_7, x_8 i x_9 bismo smatrali ekstremnim vrijednostima.

najemo distribuciju maksimuma. U radu ćemo opisati osnovne aspekte statističkog modeliranja ekstremnih događaja primjenom distribucija ekstremnih vrijednosti – vjerojatnosnih distribucija koje se javljaju kao granične distribucije maksimuma za niz nezavisnih i jednako distribuiranih slučajnih varijabli. Članak je temeljen na diplomskom radu [3]. Kako bismo čitanje omogućili i onima koji nisu upoznati s osnovama vjerojatnosti, započinjemo s kratkim pregledom osnovnih pojmova, u kojem posebno ističemo pojam slučajne varijable (sve definicije koje nisu navedene, mogu se pronaći u [14]).

2. Distribucija maksimuma *n* slučajnih varijabli

Teorija vjerojatnosti je grana matematike koja se bavi onim pojavama čiji ishod nije unaprijed određen, odnosno ishodima *slučajnog pokusa*. Najjednostavniji primjer takvog pokusa je bacanje simetričnog novčića kod kojeg postoje dva moguća ishoda (*pismo* i *glava*) te bacanje simetrične kocke kod kojeg postoji šest mogućih ishoda $(1, 2, 3, 4, 5, 6)$. Pojedinom ishodu pokusa htjeli bismo pridružiti *vjerojatnost* njegovog pojavljivanja. Tako je kod primjera bacanja simetrične kocke razumno smatrati da svaki od spomenutih šest ishoda (zovemo ih *elementarnim događajima*) ima jednaku vjerojatnost pojavljivanja, tj. $1/6$. Matematički, skup svih elementarnih događaja u ovom primjeru prikazujemo na sljedeći način: $\Omega = \{1, 2, 3, 4, 5, 6\}$. Sada nas, vezano uz prethodne pokuse, mogu zanimati primjerice i sljedeća, nešto složenija, pitanja:

(i). Kolika je vjerojatnost da pri jednom bacanju dvije simetrične kocke zbroj brojeva na kockama bude 7?

(ii). Kolika je vjerojatnost da se prilikom 100 bacanja jedne simetrične kocke broj 6 pojavi 10 puta?

(iii). Kolika je vjerojatnost da prilikom 1000 bacanja simetrične kocke najdulji niz jedinica bude barem 10?

Kako bismo matematički opisali ova pitanja, potreban je pojam *slučajne varijable*, koja predstavlja neku funkciju $X : \Omega \to \mathbb{R}$. Tako u pitanju (i) imamo tri slučajne varijable: varijablu X koja predstavlja broj koji je pao na prvoj kocki, varijablu Y koja predstavlja broj koji je pao na drugoj kocki te varijablu $Z = X + Y$. Ako s $\mathbb{P}(A)$ označimo vjerojatnost događaja A, onda nas u ovom pitanju zanima vjerojatnost $\mathbb{P}(Z = 7)$. U pitanju (ii) bacamo samo jednu kocku pa je potrebna samo jedna varijabla, X koja govori o tome je li se prilikom bacanja kocke pojavila 6 - dakle, X poprima vrijednosti 0 ili 1. Taj pokus ponavljamo nezavisno 100 puta te ćemo promatrati 100 nezavisnih „kopija" takve varijable X, odnosno niz nezavisnih slučajnih varijabli

$$X_1, \ldots, X_{100}$$

koje imaju jednaku distribuciju kao X. Kako bismo odgovorili na pitanje, definiramo novu slučajnu varijablu kao njihov zbroj, odnosno $Y = X_1 + \ldots + X_{100}$, a zanima nas vjerojatnost $\mathbb{P}(Y = 10)$. Za pitanje (iii), opet krećemo od niza nezavisnih i jednako distribuiranih slučajnih varijabli $X_1, \ldots, X_{1000}$, ali ovaj put X_i govori o tome je li se u i-tom bacanju na kocki pojavila vrijednost 1. Nadalje, definiramo slučajne varijable Y_i na sljedeći način: stavimo $Y_1 := X_1$; za $i \geqslant 2$, ako je vrijednost od X_i jednaka 0, neka Y_i poprima vrijednost 0; ako je vrijednost od X_i jednaka 1 i ako je vrijednost od Y_{i-1} veća od 0, neka Y_i poprima zbroj vrijednosti od Y_{i-1} i X_i; ako je vrijednost od X_i jednaka 1 i ako je vrijednost od Y_{i-1} jednaka 0, neka Y_i poprima vrijednost 1. Uočimo da $Y_1, \ldots, Y_{1000}$ upravo mjeri niz uzastopnih jedinica u tih 1000 bacanja, odnosno ako bi realizacije od $X_1, \ldots, X_{12}$ bile redom $0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1$, onda bi realizacije od $Y_1, \ldots, Y_{12}$ bile redom $0, 0, 1, 2, 3, 4, 0, 1, 2, 0, 1, 2$. Kako bismo odgovorili na pitanje (iii), potrebno je definirati slučajnu varijablu

$$Z := \max\{Y_1, \ldots, Y_{1000}\},$$

koja predstavlja najdulji niz uzastopnih jedinica u tih 1000 bacanja.

Iako se pokus bacanja novčića čini suviše jednostavan da bi bio primjenjiv na primjere iz svakodnevnog života, mnoge se pojave mogu modelirati upravo korištenjem takvog tipa pokusa - sve one koje možemo shvatiti kao slučajne pokuse s dva moguća ishoda, pri čemu jedan zovemo *uspjeh*, a drugi *neuspjeh*. Takav pokus naziva se *Bernoullijev pokus*, a slučajnu varijablu koja bilježi je

li se prilikom nekog pokusa dogodio uspjeh ili neuspjeh nazivamo *Bernoullijevom slučajnom varijablom*. Vratimo se sada na pitanje (iii). U tom pitanju zanima nas najdulji niz uspjeha, odnosno koliko puta zaredom će na kocki pasti niz od barem 10 uzastopnih jedinica (u ovom slučaju uspjeh predstavlja događaj da je na kocki pala 1, dok neuspjeh predstavlja događaj da je na kocki pao bilo koji drugi broj osim 1). Slučajna varijabla Y iz pitanja (ii) ima distribuciju koja se naziva *binomna*, a njen zakon razdiobe dan je s

$$Y \sim \begin{pmatrix} 0 & 1 & \dots & n \\ p_0 & p_1 & \dots & p_n \end{pmatrix},$$

pri čemu su $p_k = \mathbb{P}(Y = k) = \binom{n}{k} p^k q^{n-k}$ vjerojatnosti da Y poprimi vrijednost $k \in \{0, \dots, n\}$, $p \in \langle 0, 1\rangle$ je vjerojatnost uspjeha i $q = 1 - p$ vjerojatnost neuspjeha. Ovako definirana slučajna varijabla može se interpretirati na sljedeći način: ukoliko Bernoullijev pokus ponavljamo nezavisno n puta i ako s p označimo vjerojatnost uspjeha u svakom pojedinom pokusu, onda slučajna varijabla Y predstavlja broj uspjeha u tih n pokusa.

Jedna od najpoznatijih i u primjenama najčešće korištenih distribucija je *normalna distribucija*. Za razliku od Bernoullijeve i binomne, za koje je skup vrijednosti koje mogu poprimiti konačan, te se ubrajaju u *diskretne slučajne varijable*, normalna slučajna varijabla može poprimiti bilo koju vrijednost iz $\mathbb{R}$ te se ubraja u *neprekidne slučajne varijable*, koje možemo zadati njihovim funkcijama gustoća. Slučajna varijabla X ima *normalnu distribuciju* s parametrima $\mu \in \mathbb{R}$ i $\sigma > 0$ (i pišemo $X \sim N(\mu, \sigma^2)$) ako joj je funkcija gustoće dana s

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Osim funkcijom gustoće, svaka slučajna varijabla određena je svojom funkcijom distribucije, odnosno funkcijom $F : \mathbb{R} \to \mathbb{R}$ definiranom s

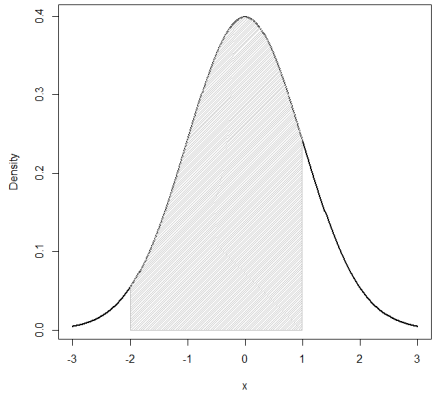
$$F(x) := \mathbb{P}(X \leqslant x), \quad x \in \mathbb{R}.$$

Ako je X neprekidna slučajna varijabla, onda, geometrijski, $\mathbb{P}(a < X < b)$ predstavlja površinu lika što ga graf funkcije gustoće određuje s pravcima $y = 0$, $x = a$ i $x = b$ (slika 4).

Razlog velike primjene normalne distribucije jedan je od najpoznatijih rezultata teorije vjerojatnosti, *centralni granični teorem*. Prema tom rezultatu, za niz nezavisnih, jednako distribuiranih slučajnih varijabli s konačnim očekivanjem $\mathbb{E}[X_i] = \mu$ i varijancom $\mathrm{Var}\,[X_i] = \sigma^2$ vrijedi

$$\frac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n} \xrightarrow{\mathrm{D}} N(0, 1), \text{ kada } n \to \infty,$$

gdje je $\overline{X_n} = \frac{X_1 + \dots + X_n}{n}$. Drugim riječima, ako je n dovoljno velik, aritmetička sredina $\overline{X_n}$ imat će približno normalnu distribuciju.



Slika 4. Graf funkcije gustoće $N(0,1)$ razdiobe. Površina osjenčanog područja ispod grafa funkcije gustoće jednaka je vjerojatnosti $\mathbb{P}(-2 < X < 1)$.

U primjeni X_i, primjerice, mogu predstavljati vrijednosti neke veličine mjerene u jednom danu (temperature zraka, brzine vjetra, koncentracije ozona, razine vode i slično). Centralni granični teorem (u slučaju da su ispunjene njegove pretpostavke) mogao bi opisivati distribuciju prosječnih vrijednosti tih veličina te bi histogram tih vrijednosti trebao nalikovati funkciji gustoće normalne distribucije. No, u problemu koji nas zanima, u kojem bismo se htjeli osigurati od velikih šteta, uzrokovanih primjerice velikim poplavama, ne bi nas zanimala prosječna dnevna razina vode, već vjerojatnost da maksimalna razina vode (unutar određenog razdoblja) premaši danu vrijednost. Drugim riječima, imajući na umu prvi pristup, promatrali bismo slučajne varijable $X_1, \ldots, X_n$ unutar nekog vremeneskog intervala (na primjer, godine), a zatim njihov maksimum

$$M_n = \max\{X_1, \ldots, X_n\}, \tag{1}$$

te bismo htjeli opisati distribuciju od M_n.

Uočimo, ako su $X_1, \ldots, X_n$ nezavisne i jednako distribuirane slučajne varijable, s funkcijom distribucije F, onda vrijedi:

$$\begin{aligned}\mathbb{P}(M_n \leqslant x) &= \mathbb{P}(\max\{X_1, \ldots, X_n\} \leqslant x) \\ &= \mathbb{P}(X_1 \leqslant x) \cdot \mathbb{P}(X_2 \leqslant x) \cdot \ldots \cdot \mathbb{P}(X_n \leqslant x) \\ &= [\mathbb{P}(X_1 \leqslant x)]^n = F^n(x).\end{aligned}$$

Ako je poznata distribucija od X_1, onda prethodni račun daje distribuciju od M_n. Međutim, u praksi obično nije poznata distribucija od X_i. Teorem koji na-

vodimo u sljedećoj cjelini daje odgovor na pitanje koje se vjerojatnosne distribucije javljaju kao granične distribucije maksimuma (centriranih i normiranih, kao kod centralnog graničnog teorema) za dani niz slučajnih varijabli, pri čemu pretpostavka o vrsti distribucije od X_i nije potrebna, već samo da one imaju jednaku distribuciju te da su nezavisne.

3. Distribucije ekstremnih vrijednosti

Prisjetimo se komentara iz uvodne cjeline: koliki utjecaj na ishod pokusa ima poznavanje činjenice da je vjerojatnost pojave jedinice prilikom jednom bacanja kocke jednaka $1/6$ ako kocku bacimo 12 puta? Ne preveliki. Međutim, ako pokus ponavljamo „dovoljno velik" broj puta, recimo N, broj 1 će se pojaviti približno $N/6$ puta. Danas je ocjenu o tome koji N je dovoljno velik moguće dobiti upotrebom računala, odnosno simulacija.

Simulacijama tako možemo pokušati dobiti informaciju o distribuciji slučajne varijable iz pitanja (iii), ali i distribuciji od M_n. Za provedbu simulacija koristimo R[1]. Za početak konstruirajmo niz duljine 1000 te nađimo najdulji niz uspjeha (prisjetimo se, pojavu broja jedan definirali smo kao uspjeh). Bacanje jedne kocke možemo simulirati naredbom:

```
x <- sample(1:6, size = 1, replace = TRUE)
```

odnosno 1000 kocaka naredbom:

```
N <- 1000
x <- sample(1:6, size = N, replace = TRUE)
```

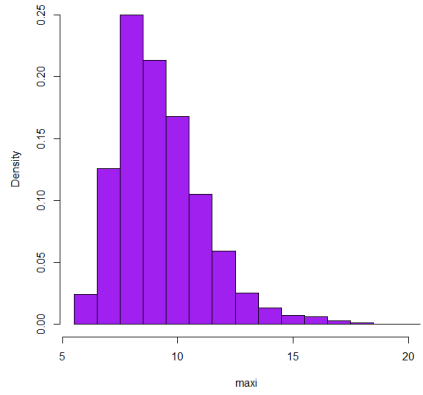
Najdulji niz uspjeha tražimo pomoću sljedećih naredbi:

```
y <- c(0, N)
for (i in 2:N){
if (x[i]==1) y[i] <- y[i-1] + x[i]
else y[i] <- 0}
maxi <- max(y)
```

Zatim, korištenjem još jedne `for` petlje, prethodni postupak ponovimo primjerice 1000 puta, a dobiveni histogram podataka prikazan je na slici 5.

Vratimo se sada na slučajne varijable (M_n) definirane u poglavlju 2. Želimo dobiti odgovor na sljedeće pitanje: kojim distribucijama možemo modelirati maksimume nezavisnih, jednako distribuiranih slučajnih varijabli? Provedimo simulacije za nekoliko različitih razdioba. Simulirajmo realizacije $x_1, \ldots, x_n$ (recimo, za $n = 1000$) slučajnih varijabli koje imaju određenu distribuciju te izračunajmo maksimum m. Postupak ponovimo $N = 100000$ puta te za podatke $m_1, \ldots, m_N$ nacrtajmo histogram. Provedimo najprije simulaciju geometrijske distribucije s parametrom $p = 0.5$ te nađimo maksimum. To možemo napraviti koristeći sljedeći kod u R-u:

[1] R je programski jezik i okruženje za statističke izračune i vizualizaciju.



Slika 5. Histogram podataka za najdulji niz uspjeha prilikom 1000 bacanja simetrične kocke

```
n <- 1000
M <- rgeom(n, 0.5)
maksimum <- max(M)
```

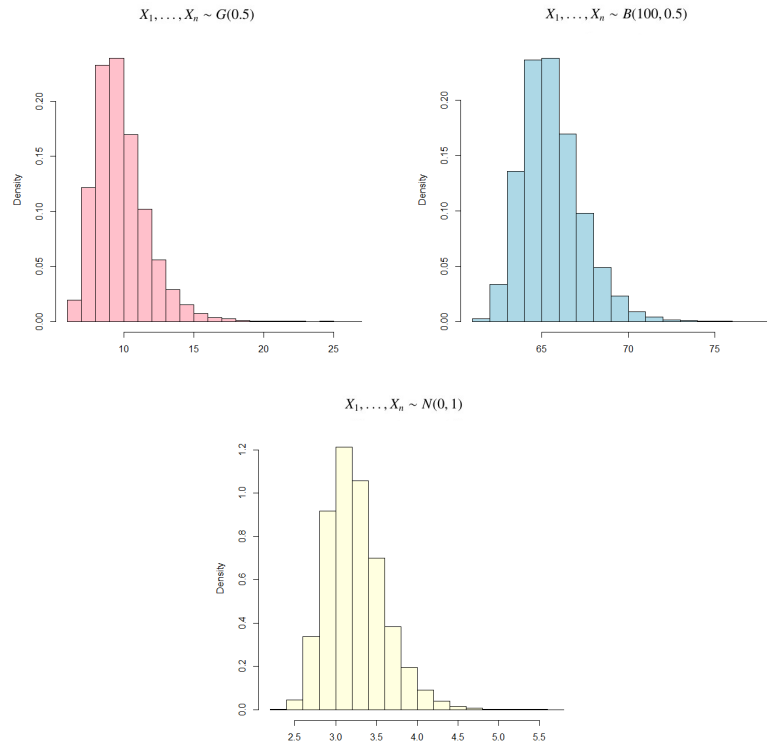
Prethodni postupak ponovimo 100000 puta te prikažimo histogram maksimuma koji smo pri tome dobili:

```
vec <- c()
for (i in 1:100000){
  M <- rgeom(n, 0.5)
   maks <- max(M)
   vec <- c(vec, maks)}
hist(vec, prob=T, col='pink')
```

Uzimanjem uzoraka iz različitih distribucija i modificiranjem prethodnog koda dobivamo histograme kako je prikazano na slici 6. Slika 6 (a) prikazuje histogram za geometrijsku distribuciju s parametrom $p = 0.5$, slika 6 (b) za binomnu distribuciju s parametrima $n = 100$ i $p = 0.5$, a slika 6 (c) za standardnu normalnu distribuciju. Uočimo da su ti histogrami vrlo slični. Sljedeći teorem pokazuje da to nije slučajno.

Teorem 1 (Fisher-Tippett-Gnedenko)**.** *Neka su $X_1, \ldots, X_n$ nezavisne, jednako distribuirane slučajne varijable te neka je $M_n = \max\{X_1, \ldots, X_n\}$. Ako postoje konstante $a_n \in \langle 0, +\infty \rangle$, $b_n \in \mathbb{R}$ takve da vrijedi:*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \longrightarrow G(x), \text{ kada } n \to \infty, \tag{2}$$



Slika 6. Usporedba histograma za maksimume geometrijske, binomne te normalne razdiobe

gdje je G neka nedegenerirana funkcija distribucije, tada G pripada jednoj od sljedeće tri familije:

$$\textit{Gumbel:} \quad G_0(x) = \exp\left[-e^{-\left(\frac{x-\mu}{\sigma}\right)}\right], \ -\infty < x < \infty;$$

$$\textit{Fréchet:} \quad G_1(x) = \exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}\right] 1_{[\mu, +\infty\rangle}(x), \ \alpha > 0, \ x \in \mathbb{R};$$

$$\textit{Weibull:} \quad G_2(x) = \exp\left\{-\left[-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}\right]\right\} 1_{\langle -\infty, \mu]}(x), \ \alpha < 0, \ x \in \mathbb{R},$$

gdje su $\mu \in \mathbb{R}$, $\sigma > 0$.

Drugim riječima, ovaj teorem govori o tome da se maksimumi nezavisnih i jednako distribuiranih slučajnih varijabli mogu modelirati jednom od sljedeće tri distribucije: Gumbelovom, Fréchetovom ili Weibullovom distribucijom. Funkcije gustoća ovih distribucija prikazane su na slici 7, a ove distribu-

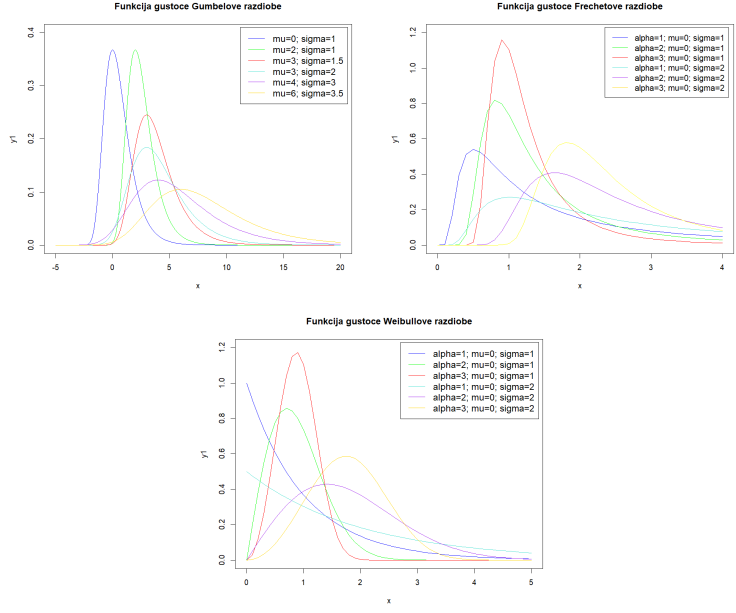
cije nazivamo *distribucijama ekstremnih vrijednosti* [2] te ih zajednički možemo opisati na sljedeći način:

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{3}$$

za x takve da je $1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0$, gdje su $\mu, \xi \in \mathbb{R}$, $\sigma > 0$. Model opisan izrazom (3) nazivamo *generalizirani model teorije ekstremnih vrijednosti* ili kraće *GEV*. Parametar μ zovemo *parametar lokacije*, σ *parametar skale*, a ξ *parametar oblika*. U slučaju $\xi = 0$ zapravo promatramo limes kada $\xi \to 0$, tj.

$$G(x) = \lim_{\xi\to 0}\left\{\exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}\right\} = \exp\left[-e^{-\left(\frac{x-\mu}{\sigma}\right)}\right],\ -\infty < x < \infty.$$

Uočimo da u tom slučaju dobivamo Gumbelovu distribuciju. Nadalje, slučaj $\xi > 0$ odgovara Fréchetovoj distribuciji, a slučaj $\xi < 0$ odgovara Weibullovoj distribuciji.



Slika 7. Grafovi Gumbelove, Fréchetove i Weibullove gustoće za različite vrijednosti parametara α, μ, σ

[2] U slučaju kada je $\mu = 0, \sigma = 1$ te $\alpha = 1$, govorimo o *standardnim funkcijama distribucijama ekstremnih vrijednosti.*

Uočimo da histogrami na slici 6 zaista nalikuju gustoćama na slici 7, kao i histogram na slici 5. Preciznije, nalikuju funkciji gustoće Gumbelove razdiobe. No, kao što je sugerirano teoremom 1, distribucija maksimuma ipak neće uvijek težiti Gumbelovoj. Točnije, odgovor ovisi o izboru distribucije od X_i. Uočimo, ako su X_i nezavisne varijable koje imaju eksponencijalnu distribuciju s parametrom $\lambda = 1$, odnosno ako je $F(x) = 1 - e^{-\lambda x}$, onda bismo za $a_n = 1$ i $b_n = \ln n$ dobili:

$$F^n\ (a_n x + b_n) = \left(1 - e^{-x - \ln n}\right)^n = \left(1 - \frac{e^{-x}}{n}\right)^n \longrightarrow \exp\left(-e^{-x}\right), \text{ kada } n \to \infty.$$

Drugim riječima, niz $(M_n - \ln n)$ konvergira po distribuciji prema nekoj slučajnoj varijabli Y koja ima standardnu Gumbelovu razdiobu. No, ako su X_i nezavisne varijable koje imaju Pareto distribuciju s parametrom $\alpha > 0$, odnosno ako je $F(x) = 1 - x^{-\alpha}$, $\alpha > 0$ onda bismo za $a_n = n^{1/\alpha}$ te $b_n = 0$ dobili:

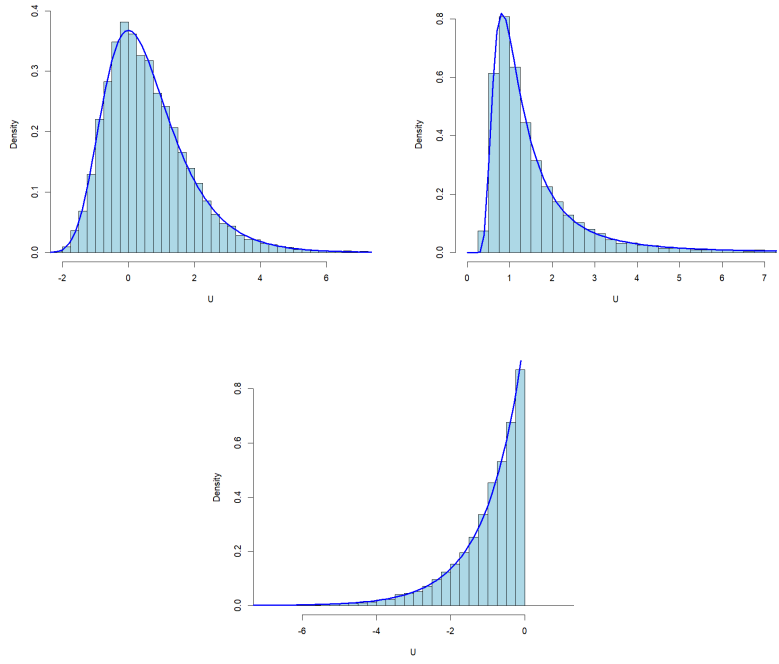
$$F^n\ (a_n x) = \left(1 - (a_n x)^{-\alpha}\right)^n = \left(1 - \frac{x^{-\alpha}}{n}\right) \longrightarrow \exp\left(-x^{-\alpha}\right), \text{ kada } n \to \infty.$$

Drugim riječima, niz $\left(\frac{M_n}{n^{1/\alpha}}\right)$ konvergira po distribuciji prema nekoj slučajnoj varijabli Z koja ima Fréchetovu razdiobu. Ako su pak X_i nezavisne varijable koje imaju uniformnu distribuciju na segmentu $[0, 1]$ za $a_n = 1/n$ i $b_n = 1$ vrijedilo bi:

$$F^n\ (a_n x + b_n) = F^n\left(n^{-1} x + 1\right) = \left(1 + \frac{x}{n}\right) \longrightarrow e^x, \text{ kada } n \to \infty.$$

Drugim riječima, niz $(n\, M_n - 1)$ u ovom slučaju konvergira po distribuciji prema nekoj slučajnoj varijabli W koja ima standardnu Weibullovu razdiobu.

Iako osnovni teorem teorije ekstremnih vrijednosti ima veliku primjenu, ima i poprilično jake pretpostavke. Naime, polazimo od pretpostavke da su slučajne varijable nezavisne i jednako distribuirane. Međutim, distribucije ekstremnih vrijednosti ponekad se mogu koristiti i u slučajevima kada pretpostavka nezavisnosti nije ispunjena. Nadalje, važno pitanje prilikom odabira ekstrema kod prvog pristupa je odrediti veličinu blokova s obzirom na koje promatramo maksimume. Promotrimo na primjer problem mjerenja temperature zraka. Ako bismo mjerili maksimalne dnevne temperature, logično je očekivati da će one varirati ovisno o godišnjim dobima pa neće biti jednako distribuirane. Kad bismo, primjerice, gledali maksimalne kvartalne temperature, pretpostavka je da ni onda dobiveni maksimumi ne bi bili jednako distribuirani jer bi ljetne maksimalne temperature koje bismo na taj način sakupili imale puno veće vrijednosti od zimskih. Zbog ovog razloga najčešće vremenska razdoblja u kojem promatramo temperature gledamo na godišnjoj razini.



Slika 8. Gumbelova, Fréchetova te Weibullova distribucija dobivene kao limesi niza maksimuma za niz nezavisnih eksponencijalnih, Pareto te uniformnih slučajnih varijabli

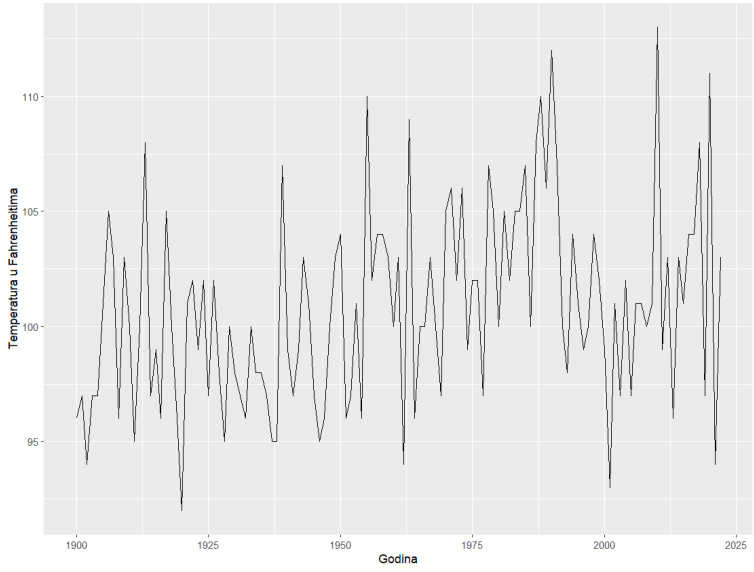
4. Primjeri

Pretpostavimo sada da smo mjerenjem prikupili podatke o maksimalnoj dnevnoj temperaturi, brzini vjetra ili slično. Drugim riječima, prikupili smo niz $x_1, \ldots, x_n$ realizacija slučajnih varijabli $X_1, \ldots, X_n$ promatranih u poglavljima 2 i 3. Osnovna zadaća statističke analize je na temelju danog uzorka donijeti zaključke o populaciji.

Na slici 9 prikazan je vremenski niz koji predstavlja godišnje maksimume temperatura zraka (u Fahrenheitima) izmjerene u gradskoj četvrti Downtown Los Angeles, od 1900. do 2022. godine[3]. Na temelju podataka želimo odrediti model koji opisuje distribuciju temperaturnih maksimuma.

Ako vrijednosti temperature zabilježene u uzorku shvatimo kao realizacije nekog niza slučajnih varijabli, na temelju teorema 1 mogli bismo očekivati da bi ovi podaci mogli imati jednu od sljedeće tri distribucije: Gumbelovu,

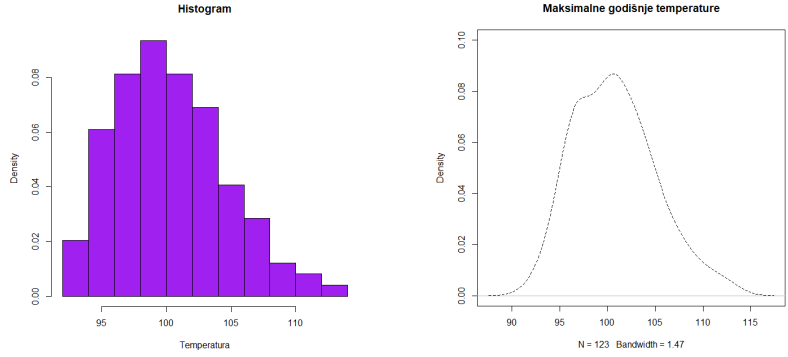
[3]Podaci preuzeti iz [9].



Slika 9. Maksimalne godišnje temperature u Fahrenheitima od 1900. do 2022. godine, Downtown Los Angeles

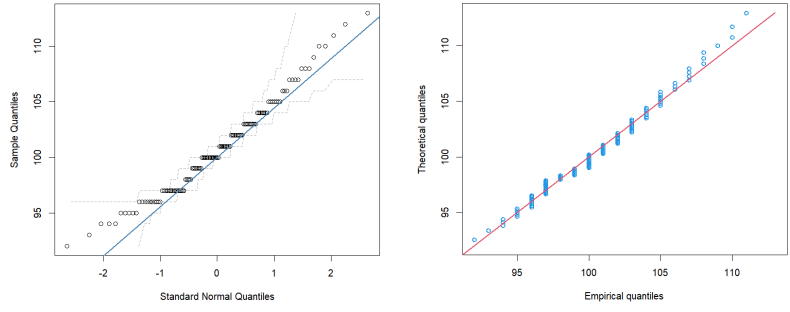
Fréchetovu ili Weibulovu distribuciju. No, općenito, proces odabira odgovarajućeg modela za stvarne podatke može biti poprilično zahtjevan. Za početak možemo vizualizirati podatke koristeći sljedeće metode: prikaz podataka pomoću histograma ili pomoću metode procjene gustoće jezgrom (*engl. kernel density method*), što za ovaj primjer možemo vidjeti na slici 10.

QQ-plot ili *graf kvantila* jedna je od najčešće korištenih metoda koja se koristi za određivanje tipa distribucije kojoj određeni uzorak pripada te pomoću *QQ*-plota možemo usporediti distribuciju uzorka s unaprijed određenom teorijskom distribucijom. Ako je distribucija populacije kojoj pripada uzorak jednaka teorijskoj, točke koje predstavljaju kvantile trebale bi se nalaziti približno na istom pravcu. Na taj način možemo ocijeniti je li pretpostavljeni model prikladan za dane podatke i kasnije to potkrijepiti statističkim testom. Primjerice, ako bismo htjeli provjeriti dolaze li podaci iz normalne distribucije, nacrtali bismo *QQ*-plot za normalnu razdiobu koji uspoređuje kvantile podataka s kvantilima normalne distribucije. Na slici 11 prikazan je normalni *QQ*-plot i *QQ*-plot Gumbelove razdiobe za naš primjer.



Slika 10. Maksimalne godišnje temperature u Los Angelesu: (a) Histogram podataka (b) Procjena gustoće jezgrom

Iz normalnog *QQ*-plota (slika 11 (a)) možemo uočiti da postoji značajno odstupanje točaka od pravca pa iz toga zaključujemo da ovi podaci neće biti normalno distribuirani. Nadalje, način na koji podaci odstupaju od pravca, može ukazati na klasu distribucija. U ovom slučaju, moglo bi se raditi o tzv. *distribucijama lakog repa*. Upravo je Gumbelova distribucija takvog tipa. Dakle, mogli bismo pretpostaviti da podatke možemo modelirati Gumbelovom distribucijom, što potvrđuje i Gumbelov *QQ*-plot na slici 11 (b).



Slika 11. (a) Normalni *QQ*-plot (b) *QQ*-plot za Gumbelovu razdiobu

Sljedeći korak je potkrijepiti ovu tvrdnju statističkim testovima. Za svaku od distribucija iz teorema 1, provest ćemo test o pripadnosti toj distribuciji odnosno testirat ćemo nul-hipotezu H_0 da podaci dolaze iz određene razdiobe (Gumbelove, Fréchetove odnosno Weibullove), nasuprot alternativnoj hi-

potezi H_1 da ne dolaze iz te razdiobe. Statističke testove obično provodimo na određenoj razini značajnosti α (α je neki zadani broj iz $\langle 0, 1\rangle$) i koja predstavlja vjerojatnost pogreške prve vrste (vjerojatnost odbacivanja istinite nul-hipoteze). S druge strane, umjesto toga, na temelju vrijednosti test-statistike može se računati i *p-vrijednost testa*, koja predstavlja najmanju razinu značajnosti uz koju bi nul-hipoteza bila odbačena u korist alternativne hipoteze. Pri testiranju smo koristili korelacijski test (*engl. correlation test*), čiju funkciju možemo pronaći u paketu `goft` u R-u te dobili p-vrijednosti redom $p = 0.1172$, $p = 0.0558$ odnosno $p = 0.0011$ za pripadnost podataka Gumbelovoj, Fréchetovoj odnosno Weibullovoj razdiobi. Primjerice, za testiranje pripadnosti danih podataka Gumbelovoj razdiobi, koristili smo sljedeći kod:

```
ev_test(temp, dist = 'gumbel', method = 'cor', N = 1000)
```

te na temelju informacije da je p-vrijednost jednaka $p = 0.1172$ zaključujemo da ne odbacujemo nul-hipotezu o pripadnosti Gumbelovoj razdiobi na razini značajnosti od 5%. Uočimo da je dobivena p-vrijednost za Gumbelovu razdiobu u skladu sa slutnjom temeljenom na izgledu QQ-plota (prisjetite se izgleda QQ-plota sa slike 11 (b)).

Alternativno, umjesto provođenja testa za svaku razdiobu posebno, može se testirati pripadnost generaliziranom modelu. Klasični testovi pomoću kojih možemo testirati pripadnost podataka generaliziranom modelu teorije ekstremnih vrijednosti su Anderson-Darlingov (AD) test, Cramer-vonMises (CVM) test, Kolmogorov-Smirnovljev (KS) test te Watsonov (W) test. Dakle, testirat ćemo nul-hipotezu H_0 da maksimalne godišnje temperature pripadaju *GEV* razdiobi, nasuprot alternativnoj hipotezi H_1 da ne pripadaju *GEV* razdiobi. U tu svrhu, koristit ćemo funkciju `gnfit` iz istoimenog paketa u R-u. Naime, ta funkcija izračunava vrijednosti test-statistika i p-vrijednosti Cramer-vonMisesovog i Anderson-Darlingovog testa. Međutim, prije korištenja te funkcije, potrebno je procijeniti vrijednosti nepoznatih parametara μ, σ, ξ za *GEV* razdiobu. Iz tog razloga najprije koristimo funkciju `gev.fit` iz paketa `ismev`. U toj funkciji implementirana je metoda maksimalne vjerodostojnosti pa kao ispis dobivamo procjene $\hat{\mu}, \hat{\sigma}, \hat{\xi}$. Dakle, koristeći sljedeći kod:

```
model <- gev.fit(temp)$mle
model
```

dobivamo:

```
model
[1] 99.0039464  3.7764979 -0.1228323
```

odnosno procjene redom iznose: $\hat{\mu} = 99.0039464, \hat{\sigma} = 3.7764979$ i $\hat{\xi} = -0.1228323$. Nadalje, korištenjem naredbe:

```
gnfit(temp, 'gev', df = NULL, pr = model, threshold = NULL)
```

dobivamo p-vrijednosti za Anderson-Darlingov i Cramer-vonMisesov test:

```
Test of Hypothesis for gev distribution
Cramer-von Misses Statistics:  0.0974    p-Value:  0.12216
Anderson-Darling Statistics:   0.5603    p-Value:  0.14763
```

Iz dobivenih p-vrijednosti ($p = 0.1476$ nakon provođenja AD testa odnosno $p = 0.1222$ nakon provođenja CVM testa), zaključujemo da ne odbacujemo nul-hipotezu o pripadnosti zadanih podataka GEV razdiobi na razini značajnosti od 5%.

Sada možemo postaviti sljedeće pitanje: je li uistinu Gumbelov model prikladniji za dane podatke od Fréchetova i Weibullova modela? Drugim riječima, unutar GEV modela, testiramo hipotezu H_0 da je parametar $\xi = 0$ nasuprot alternativnoj hipotezi H_1 da je $\xi \neq 0$. Uočimo da smo već procijenili vrijednost parametra ξ za naš primjer. Naime, dobili smo da je $\hat{\xi} = -0.1228323$. Kako je procijenjeni parametar oblika blizu nule, to bi moglo biti u skladu s našom pretpostavkom da je Gumbelov model prikladniji za modeliranje ovih podataka od Fréchetova i Weibullova modela. Da bismo potvrdili tu pretpostavku, koristit ćemo test omjera vjerodostojnosti (*engl. likelihood ratio test*). Detaljnije o testu omjera vjerodostojnosti moguće je pronaći u [12], poglavlje 4.2. Test ćemo provesti u R-u koristeći funkcije `fevd` i `lr.test` iz paketa `ExtRemes`. Naime, funkcija `fevd` prilagođava dane podatke određenoj distribuciji. Ukoliko ne specificiramo kojoj distribuciji želimo prilagoditi naše podatke, zadana opcija kojoj ta funkcija prilagođava podatke je GEV distribucija. Konkretno, mi ćemo prilagoditi naše podatke Gumbelovoj distribuciji i GEV distribuciji, a zatim ćemo iskoristiti funkciju `lr.test` koja će provesti test omjera vjerodostojnosti. Koristeći sljedeći kod:

```
fit0 <- fevd(temp, type='Gumbel')
fit1 <- fevd(temp)
lr.test(fit0, fit1)
```

dobivamo:

```
Likelihood-ratio Test

data:  temp
Likelihood-ratio = 2.8105, chi-square critical value = 3.8415,
alpha = 0.0500, Degrees of Freedom = 1.0000, p-value = 0.09365
alternative hypothesis: greater
```

Možemo očitati da je p-vrijednost jednaka $p = 0.0937$ pa ne odbacujemo nul-hipotezu na razini značajnosti od 5%. Dakle, na razini značajnosti od 5% ne odbacujemo hipotezu o prikladnosti Gumbelovog modela i ne možemo tvrditi da Gumbelov model nije prikladniji za modeliranje maksimalnih godišnjih temperatura od Fréchetova i Weibullova modela.

Promotrimo sada drugi primjer. U tablici 1 prikazani su godišnji maksimumi dnevnih mjerenja brzine vjetra u km/h u Vancouveru od 1947. do 1984. godine[4].

[4]Podaci preuzeti iz [12].

god.	brzina	god.	brzina	god.	brzina	god.	brzina	god.	brzina	god.	brzina	god.	brzina
1947.	79.5	1953.	64.8	1959.	64.8	1965.	61.0	1971.	70.3	1977.	48.1	1983.	51.8
1948.	68.4	1954.	59.2	1960.	88.8	1966.	51.8	1972.	68.4	1978.	53.6	1984.	48.1
1949.	74.0	1955.	79.5	1961.	88.8	1967.	62.9	1973.	55.5	1979.	55.5		
1950.	59.2	1956.	62.9	1962.	75.8	1968.	64.8	1974.	64.8	1980.	62.9		
1951.	74.0	1957.	59.2	1963.	68.4	1969.	61.0	1975.	77.7	1981.	61.0		
1952.	64.8	1958.	68.2	1964.	68.4	1970.	61.0	1976.	57.3	1982.	61.0		

Tablica 1. Godišnji maksimumi dnevnih mjerenja brzine vjetra u km/h u Vancouveru

Za svaku godinu promatrana je vrijednost maksimuma iz skupa podataka koji su u toj godini bili zabilježeni. Slika 12 (a) prikazuje zadane podatke po godinama.

Želimo odrediti razdiobu koja najbolje opisuje ove podatke. Za početak, testirajmo pripadnost zadanih maksimalnih godišnjih brzina vjetra *GEV* razdiobi. Dakle, testiramo nul-hipotezu H_0 da podaci pripadaju *GEV* razdiobi, nasuprot alternativnoj hipotezi H_1 da ne pripadaju *GEV* razdiobi. Ponovno, koristeći funkciju `gev.fit`:

```
model <- gev.fit(brzina)$mle
model
```

dobivamo procjene parametara *GEV* razdiobe: $\hat{\mu} = 60.8603081$, $\hat{\sigma} = 8.5096874$ i $\hat{\xi} = -0.1131792$. Nadalje, koristeći sljedeći kod:

```
gnfit(brzina, 'gev', df = NULL, pr = model, threshold = NULL)
```

dobivamo p-vrijednosti za Anderson-Darlingov i Cramer-vonMisesov test:

```
Test of Hypothesis for gev distribution
Cramer-von Misses Statistics:  0.0492   P-Value:  0.5216
Anderson-Darling Statistics:   0.2979   P-Value:  0.58862
```

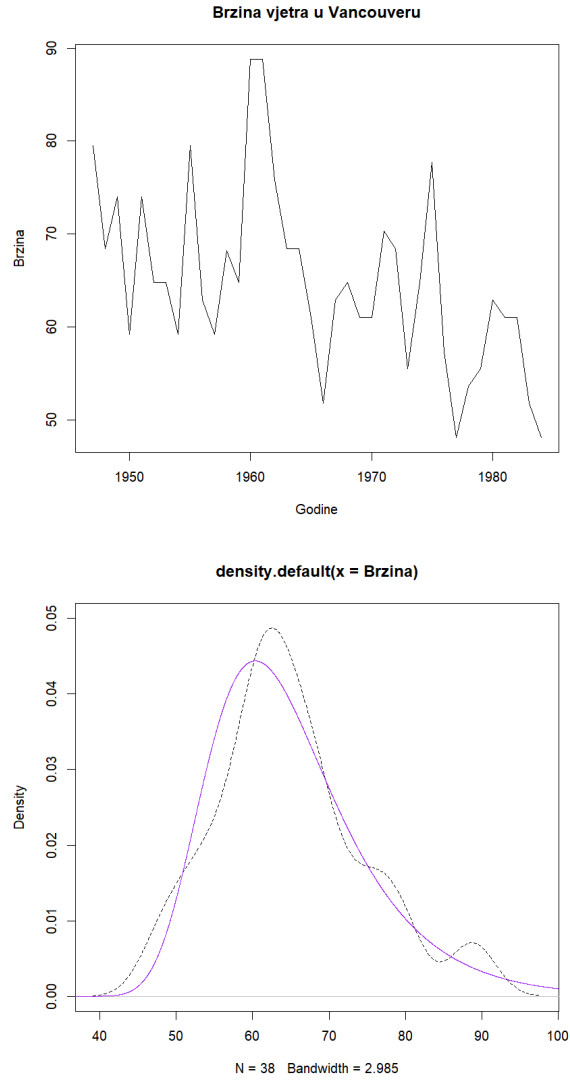
Iz dobivenih p-vrijednosti ($p = 0.5886$ nakon provođenja *AD* testa, odnosno $p = 0.5216$ nakon provođenja *CVM* testa), zaključujemo da ne odbacujemo nul-hipotezu o pripadnosti zadanih podataka *GEV* razdiobi na razini značajnosti od 5%. Nadalje, kako je opet vrijednost procijenjenog parametra oblika $\hat{\xi} = -0.1131792$ blizu nuli, možemo pretpostaviti da je Gumbelov model prikladniji za modeliranje ovih podataka od Fréchetova i Weibullova modela. Dakle, testiramo nul-hipotezu H_0 da je parametar $\xi = 0$ nasuprot alternativnoj hipotezi H_1 da je $\xi \neq 0$. Koristeći sljedeći kod u R-u:

```
fit0 <- fevd(brzina, type='Gumbel')
fit1 <- fevd(brzina)
lr.test(fit0, fit1)
```

dobivamo:

```
Likelihood-ratio Test

data:  brzina
```



Slika 12. (a) Prikaz podataka iz tablice 1 (b) Procjena gustoće jezgrom i funkcija gustoće Gumbelove razdiobe s procijenjenim parametrima

```
Likelihood-ratio = 0.84379, chi-square critical value = 3.8415,
alpha = 0.0500, Degrees of Freedom = 1.0000, p-value= 0.3583
alternative hypothesis: greater
```

Možemo očitati da je p-vrijednost jednaka $p = 0.3583$ pa ne odbacujemo nulhipotezu na razini značajnosti od 5%. Dakle, na razini značajnosti od 5% ne odbacujemo hipotezu o prikladnosti Gumbelovog modela i ne možemo tvrditi da Gumbelov model nije prikladniji za modeliranje maksimalnih godišnjih brzina vjetra od Fréchetova i Weibullova modela.

5. Zaključak

Spomenimo na kraju što na temelju dobivenog modela možemo zaključiti te istaknimo neke nedostatke i alternativni pristup izdvajanja ekstremnih vrijednosti iz skupa podataka. Jednom kada je model odabran, moguće je dati odgovore na neka od pitanja postavljena u uvodu. Konkretno, mogli bismo odgovoriti kolika je vjerojatnost $\mathbb{P}(M_n \geqslant L)$, za bilo koju vrijednost L koja nas zanima, odnosno mogli bismo primjerice odrediti koja vrijednost karakterizira 1000-godišnju poplavu. No, jesu li modeli promatrani u poglavlju 4 i najbolji mogući za te primjere? U članku smo već spomenuli neke od poteškoća koje se mogu javiti prilikom odabira modela. Prisjetimo ih se još jednom. Osnovni teorem koji smo koristili sadrži jake pretpostavke o nezavisnosti i jednakoj distribuiranosti, a ponašanje promatranih veličina može se mijenjati kroz vrijeme zbog raznoraznih vanjskih utjecaja. Tako, primjerice, na slici 9 možemo uočiti blagi linearni rast (posebno se to primjećuje za razdoblje do 1950. godine), što bi također trebalo uvažiti prilikom formiranja modela. Također, na slici 12 (a) možemo uočiti pad te bi bolji model svakako trebao uključiti uklanjanje trenda.

Postoji i drugi pristup izdvajanja ekstrema. Aktuare primjerice može zanimati vjerojatnost velikih šteta te se korisnijim od promatranja maksimalne vrijednosti nad nekim vremenskim intervalima čini promatranje vrijednosti iznad određenog praga. Na taj način u analizi koristimo sve podatke čije su vrijednosti veće od vrijednosti unaprijed zadanog praga, a ne samo jednu ekstremnu vrijednost po razdoblju u kojem provodimo naše promatranje. U tom slučaju za modeliranje ne koristimo distribucije ekstremnih vrijednosti već drugu klasu distribucija koje se nazivaju generalizirane Pareto distribucije.

Literatura

[1] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer Verlag, Berlin, 2001.

[2] Dokumentacija za rad s R-om, *RDocumentation*, `https://www.rdocumentation.org/`

[3] A. M. Filipas, *Statističko modeliranje ekstremnih vrijednosti*, diplomski rad, Rijeka, 2021.

[4] E. Gilleland, R. W. Katz, *extRemes 2.0: An Extreme Value Analysis Package in R*, Journal of Statistical Software, 2014.

[5] E. Gilleland, R. W. Katz, *Package extRemes*, `https://CRAN.R-project.org/package=extRemes`, 2022.

[6] E. Gonzalez-Estrada, J. A. Villasenor-Alva, *An R package for testing goodness of fit: goft*, Journal of Statistical Computation and Simulation, 88(1): 1–26, 2018.

[7] E. Gonzalez-Estrada, J. A. Villasenor-Alva, *Package goft*, https://CRAN.R-project.org/package=goft, 2020.

[8] J. E. Heffernan, *Package ismev*, `https://CRAN.R-project.org/package=ismev`, 2018.

[9] Los Angeles Almanac, `https://www.laalmanac.com/index.php`, 2023.

[10] Manual za rad s R-om, *The R Manuals*, `https://cran.r-project.org/manuals.html`

[11] S. P. Millard, *Package EnvStats*, `https://CRAN.R-project.org/package=EnvStats`, 2022.

[12] R. D. Reiss, M. Thomas, *Statistical Analysis of Extreme Values with Applications To Insurance, Finance, Hydrology and Other Fields*, Birkhäuser Basel, 2001.

[13] A. Saeb, *Package gnFit*, https://CRAN.R-project.org/package=gnFit, 2018.

[14] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

Ana Marija Filipas
Sveučilište u Rijeci, Ekonomski fakultet
E-mail adresa: `ana.marija.filipas@efri.hr`

Ivana Slamić
Sveučilište u Rijeci, Fakultet za matematiku
E-mail adresa: `islamic@math.uniri.hr`