Seminar iz statstičkog praktikuma

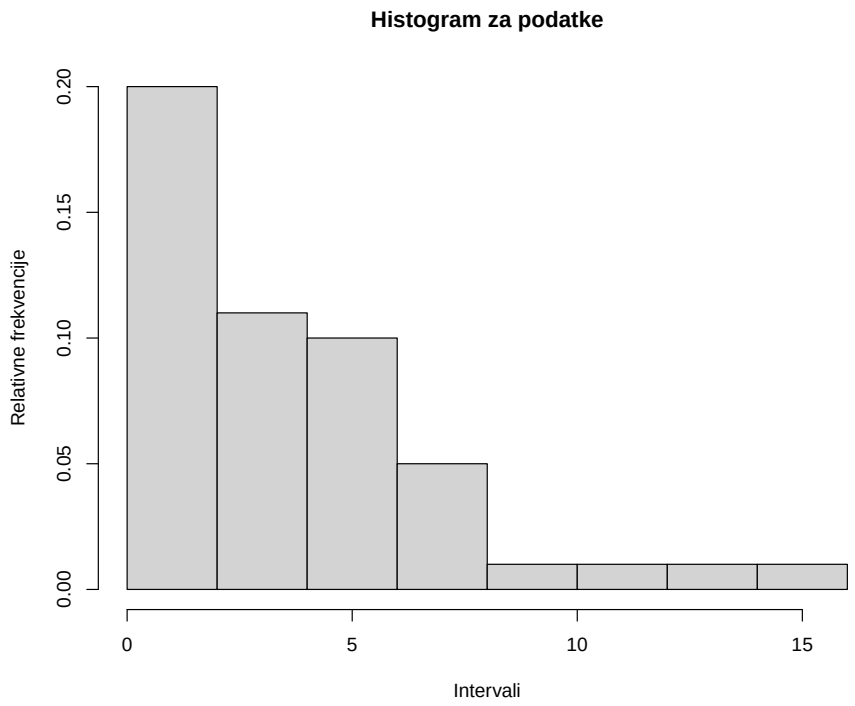
Krunoslav Ivanović, **zadatak 38.**

13. siječnja 2023.

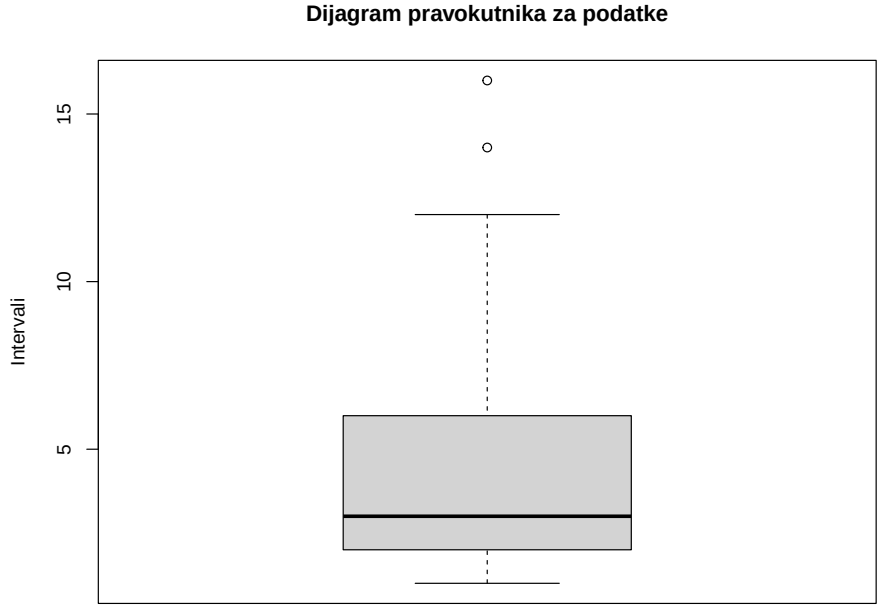
U ovom seminaru, bavimo se proučavanjem distribucije vremenskih intervala međudolazaka automobila na jednom određenom križanju u Australiji. Ukupno imamo 50 mjerenja. Ispitujemo pripadanje tih podataka trokutastoj i eksponencijalnoj distribuciji, pronalazimo pouzdani interval za očekivanje vremena između međudolazaka automobila i testiramo hipotezu o tome je li to očekivanje jednako 4.

Podaci su nam zadani u tablici iz koje ih iščitamo i upišemo u R. Osnovna svojstva podataka su dana s

```
> summary(podaci)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.00    3.00    4.24    5.75   16.00
```



Već prvi pogled na histogram nam daje naslutiti da je riječ o eksponencijalno distribuiranim podacima, no, naravno, to ćemo kasnije detaljno provjeriti. Također, nacrtati ćemo i dijagram pravokutnika.



Gustoća trokutaste razdiobe je dana s

$$f(x) = \frac{2}{3m}\left(1 - \frac{x}{3m}\right)\mathbb{1}_{\langle 0,3m\rangle}(x).$$

Na standardan način računamo očekivanje i varijancu.

$$\mathbb{E}X = \int_0^\infty x f(x)\mathrm{d}x = \int_0^{3m} \frac{2x}{3m}\left(1 - \frac{x}{3m}\right)\mathrm{d}x = \ldots = m.$$
$$\mathbb{E}X^2 = \int_0^\infty x^2 f(x)\mathrm{d}x = \int_0^{3m} \frac{2x^2}{3m}\left(1 - \frac{x}{3m}\right)\mathrm{d}x = \ldots = \frac{3}{2}m^2.$$

Naposljetku, naravno, vrijedi

$$\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{3}{2}m^2 - m^2 = \frac{m^2}{2}.$$

Kako će nam kasnije trebati, računamo i funkciju distribucije za trokutastu razdiobu. Vrijedi

$$F(x) = \int_{-\infty}^x f(t)\mathrm{d}t = \int_0^x \frac{2}{3m}\left(1 - \frac{t}{3m}\right)\mathbb{1}_{\langle 0,3m\rangle}(t)\mathrm{d}t$$

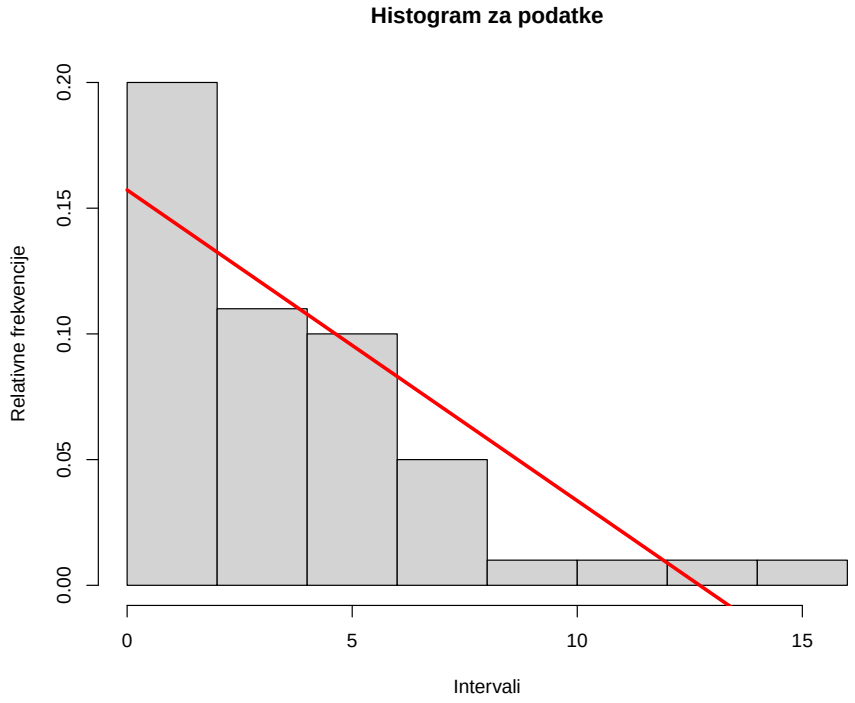
i odavde se lagano vidi

$$F(x) = \begin{cases} 0 & x \leqslant 0 \\ \frac{-x(x-6m)}{9m^2} & 0 < x < 3m \\ 1 & x \geqslant 3m \end{cases}$$

Usporedit ćemo histogram s gustoćom prilagođene trokutaste razdiobe s parametrom, kojeg procjenjujemo metodom momenata i MLE. Prvo, kako je $\mathbb{E}X = m$, znamo da je procjenjitelj dobiven metodom momenata upravo

$$\hat{m} = \overline{X} \approx 4.24.$$

Na sljedećoj slici je prikazana navedena usporedba.



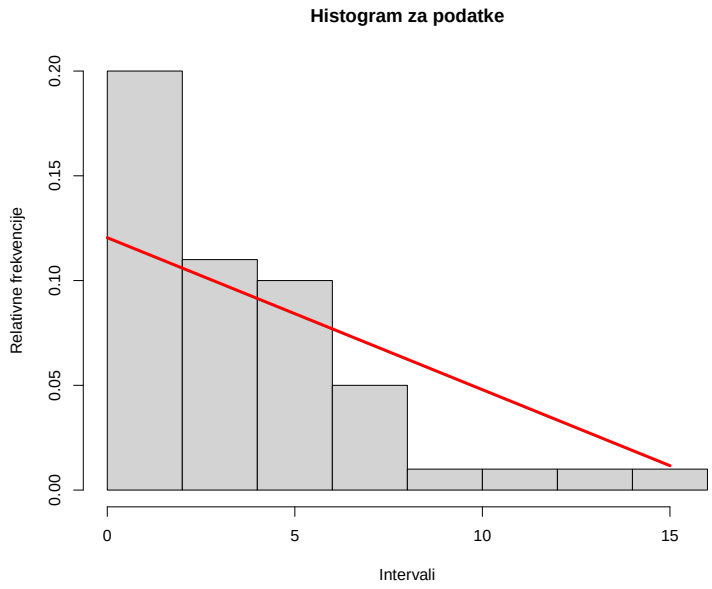
MLE procjenjitelj dobivamo standardnim načinom, odnosno, definiramo

$$L(x;m) = \prod_{i=1}^{50} f(x_i;m)$$

gdje je $f(x;m)$ gustoća trokutaste razdiobe s parametrom m, a $x_1, \ldots, x_{50}$ ostvarenje našeg uzorka. MLE procjenjitelj dobijemo tako što nalazimo m koji maksimizira L. Nakon računanja, ispadne

$$\hat{m}_{\text{MLE}} = 5.535.$$

Opet, nacrtamo histogram i odgovarajuću funkciju gustoće.

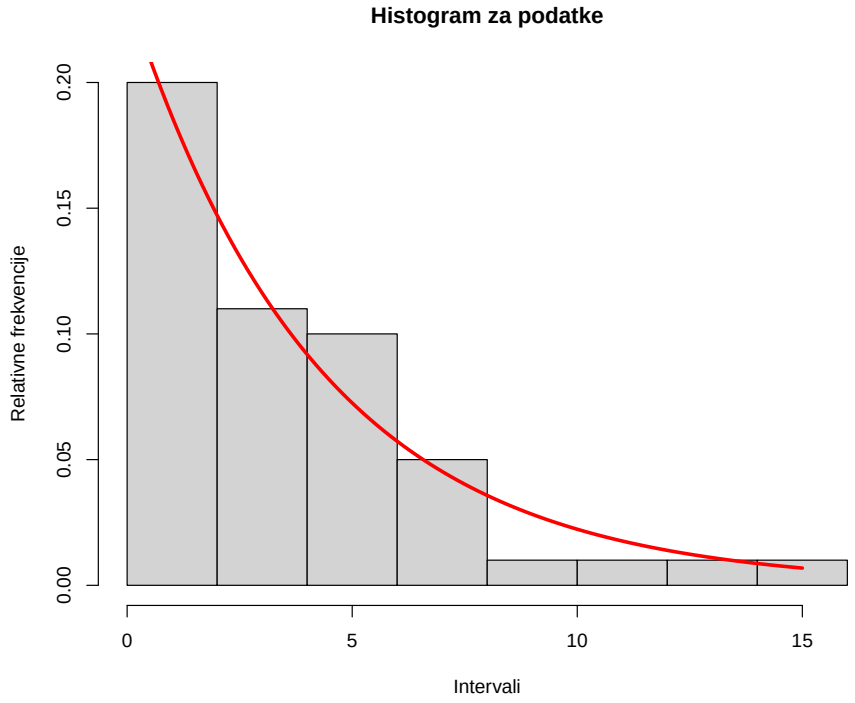


Histogram također uspoređujemo s grafom funkcije gustoće prilagođene eksponencijalne razdiobe s odgovarajućim parametrom. Znamo da je za eksponencijalnu razdiobu, parametar dobiven MLE procjenom ili metodom momenata isti pa je svejedno što ćemo raditi, a kako je lakše, koristimo metodu momenata.

Znamo da je za $Y \sim \text{Exp}(\lambda)$, $\mathbb{E}Y = \frac{1}{\lambda}$ pa je

$$\hat{\lambda} = \frac{1}{\overline{X}} = \frac{1}{4.24} \approx 0.236.$$

Dakle, crtamo zajedno histogram za podatke te gustoću eksponencijalne razdiobe s parametrom $\lambda = 0.236$, što vidimo na sljedećoj slici (crvenom bojom je označena funkcija gustoće za opisanu eksponencijalnu razdiobu).



Vidimo da ta funkcija gustoće daleko bolje „pristaje" histogramu nego u slučaju trokutaste razdiobe.

Sada ćemo provesti χ^2-test pripadnosti, prvo za trokutastu, a onda za eksponencijalnu razdiobu, pri čemu parametar određujemo „minimum χ^2-metodom". Prvo, frekvencije naših podataka su dane u sljedećoj tablici:

Duljina	1	2	3	4	5	6	7	8	10	12	14	16
Frekvencija	9	11	7	4	6	4	2	3	1	1	1	1

Vidimo da su dobivene frekvencije u nekim skupinama relativno male pa grupiramo na sljedeći način:

Duljina	1	2	3	{4,5}	{6,7}	$\geqslant 8$
Frekvencija	9	11	7	10	6	7

Ideja iza minimum χ^2-metode jest naći parametar λ za koji će testna statistika χ^2-testa biti najmanja moguća. To dobivamo minimiziranjem funkcije

$$f(\theta) = \sum_{j=1}^{6} \frac{\big(N_j - n_j(\theta)\big)^2}{n_j(\theta)},$$

gdje su N_j i n_j redom dobivene, odnosno, očekivane frekvencije u j-tom razredu. Brojeve $n_j(\theta)$ dobivamo iz formule

$$n_j(\theta) = n \cdot P(X_\theta \in A_j),$$

gdje je n veličina uzorka (u našem slučaju, $n = 50$), a X_θ odgovarajuća slučajna varijabla s parametrom θ. Zbog strukture razreda A_j, vjerojatnost na desnoj strani možemo lagano računati kao razliku funkcije distribucije u gornjoj i donjoj granici j-tog razreda.

Dobivamo da je optimalni m za trokutasti model jednak $m \approx 3.61$, a optimalni λ za eksponencijalni model jednak $\lambda \approx 0.268$.

Sada provodimo χ^2-test da provjerimo pripadnost uzorka modelima s danim parametrima. To radimo na standardan način. Testiramo hipotezu

$$H_0 \colon X \sim F_0$$
$$H_1 \colon \text{ne } H_0,$$

gdje je F_0 neka konkretna distribucija. Testna statistika je

$$H = \sum_{i=1}^{k} \frac{(N_i - n_i)^2}{n_i}\bigg|_{H_0} \sim \chi^2(k - d - 1),$$

gdje je k broj razreda, N_i broj opaženih uzoraka u i-tom razredu, a n_i očekivani broj uzoraka u i-tom razredu. Parametar k je jednak broju razreda i iznosi 6, a kako smo procijenili samo jedan parametar, $d = 1$. Naravno, pazimo da svaki od n_i bude veći ili jednak 5.

Nakon računanja, dobivamo:

Tip modela	Parametar	χ^2-statistika	p-vrijednost
eksponencijalni	$\lambda = 0.268$	1.268	0.867
trokutasti	$m = 3.61$	2.100	0.717

Vidimo da ni u jednom modelu ne možemo odbaciti H_0 ni na jednoj smislenoj razini značajnosti, no, isto tako, vidimo da je eksponencijali model bolji (p-vrijednost je veća).

Ostaje naći 95%-CI za $\mu = \mathbb{E}X$, pri čemu je X bolji model iz prethodnog dijela, odnsono, eksponencijalni. Znamo da je zbroj eksponencijalnih slučajnih varijabli gamma slučajna varijabla, odnosno, imamo

$$\sum X_n \sim \Gamma(n, \lambda),$$

gdje je λ parametar eksponencijalne distribucije. Znamo da je $\mu := \mathbb{E}X = \frac{1}{\lambda}$.

Iz svojstava gamma distribucije, znamo da vrijedi

$$\frac{\sum X_n}{\mu} \sim \Gamma(50, \lambda\mu) = \Gamma(50, 1).$$

Sada jednostavno možemo odrediti 95%-CI. Naime, vrijedi

$$\mathbb{P}\left(\alpha_1 \leqslant \frac{\sum X_n}{\mu} \leqslant \alpha_2\right) = 1 - \alpha,$$

gdje su α_1 i α_2 redom $\frac{\alpha}{2}$ i $1-\frac{\alpha}{2}$ kvantili za $\Gamma(50,1)$ distribuciju. Sada laganom algebrom vidimo da se interval pouzdanosti dobije kao

$$\frac{\sum X_n}{\alpha_1} \geqslant \mu \geqslant \frac{\sum X_n}{\alpha_2}.$$

Uvrštavanjem, dobijemo da je interval poudzanosti

$$[3.2726, 5.7126].$$

Alternativno, mogli smo dobiti interval pouzdanosti iz asimptotske normalnosti (dobije se sličan rezultat jer je uzorak dovoljno velik, rezultat dolje iz t-testa nije „najbolji" jer ne iskorištavamo restrikciju da je $\operatorname{Var} X = \mathbb{E}X^2$).

Naposljetku, testiramo hipotezu

$$H_0 \colon \mu = 4$$
$$H_1 \colon \mu \neq 4.$$

To radimo t-testom koji je implementiran u R-u i dobivamo:

```
> t.test(podaci, alternative = "two.sided", mu = 4)

  One Sample t-test

data:  podaci
t = 0.50149, df = 49, p-value = 0.6183
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 3.278279 5.201721
sample estimates:
mean of x
     4.24
```

Vidimo da ne možemo odbaciti H_0 ni na jednoj standardnoj razini značajnosti.