Primjena računara u biologiji

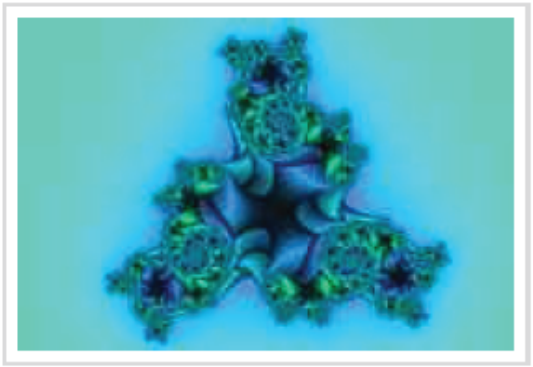
Vladimir Filipović

vladaf@matf.bg.ac.rs

Numeričke mere
i
raspodele verovatnoća

Numeričke mere

Numeričke mere



Opisuju se različite numeričke mere i dijagrami.
Biće objašnjeno (sa primerima) kako se mere i dijagrami određuju u sistemu R. Numeričke mere i dijagrami:

- Sredina
- Medijana
- Mod
- Kvartil i percentil
- Opseg i opseg među kvartilima
- Kutija dijagram (Box plot)
- Disperzija
- Standardna devijacija
- Kovarijansa
- Koeficijent korelacije
- Centralni moment
- Iskošenost i spljoštenost

Sredina

Sredina posmatrane promenljive je numerička mera centralne tendencije za vrednosti podataka.

Sredina se određuje tako što se suma vrednosti podataka podeli sa brojem podataka.

Dakle, za uzorak dimenzije n, ***sredina uzorka*** se određuje po sledećoj formuli:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Slično tome, za populaciju koja broji N jedinki, ***sredina populacije*** se određuje na sledeći način:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

U sistemu R, sredina se određuje korišćenjem funkcije **mean**.

Sredina (2)

Problem. Odrediti srednje trajanje erupcije gejzira u okviru sa podacima faithful.

Rešenje. Treba izvršiti isecanje okvira po koloni eruptions, pa nad izdvojenim podacima primeniti funkciju mean:

```
> duration = faithful$eruptions     # the eruption durations
> mean(duration)                    # apply the mean function
[1] 3.4878
```

Dakle, srednje trajanje erupcije gejzira je 3.4878 minuta.

Zadatak 1. Odrediti srednje trajanje čekanja između dve erupcije u okviru sa podacima faithful.

Medijana

Medijana posmatrane promenljive je ordinalna mera centralne tendencije za vrednosti podataka.

Medijana je središnji skor u distribuciji rezultata - broj podataka koji se nalaze ispod i iznad medijane je jednak. Drugim rečima, medijana deli na pola niz podataka ordinalno poređanih (od manjeg ka većem) - ispod i iznad medijane se nalazi 50% podataka .

Izračunavanje medijane se vrši na sledeći način:

1. Vrednosti tj. podaci se poređaju od najmanjeg do najvećeg.
2. Ukoliko uzorak sadrži neparan broj podataka, medijana je podatak tačno u sredini uzorka.
3. Ukoliko uzorak sadrži paran broj podataka, medijana je aritmetička sredina dva središnja elementa.

Medijana (2)

Primer. Pretpostavimo da smo merili visinu devojčica u 5_2 odeljenju osnovne škole "X" i da su dobijeni sledeći podaci izraženi u centimetrima:

140, 141, 138, 140, 122, 160, 154, 132, 148, 135, 140.

Prvi korak u izračunavanju medijane je da se sortiraju elementi:

122, 132, 135, 138, 140, 140, 140, 141, 148, 154 i 160.

Uzorak sadrži 11 elemenata i pozicija srednjeg je 6.

Medijana se nalazi na šestom mestu u sortiranom nizu elemnata, a to je 140.

Dakle, medijana visine devojčica u 5_2 odeljenju osnovne škole "X" iznosi 140.

U sistemu R, medijana se određuje korišćenjem funkcije **median**.

Medijana (3)

Problem. Odrediti medijanu za trajanje erupcije gejzira u okviru sa podacima faithful.

Rešenje. Treba izvršiti isecanje okvira po koloni eruptions, pa nad izdvojenim podacima primeniti funkciju median:

```
> duration = faithful$eruptions     # the eruption durations
> median(duration)                  # apply the median function
[1] 4
```

Dakle, medijana za trajanje erupcije gejzira je 4 minuta.

Zadatak 1. Odrediti medijanu za trajanje čekanja između dve erupcije u okviru sa podacima faithful.

Mod

Mod je najčešće opaženi skor u uzorku.

Mod je jedina mera centralne tendencije koja se može primeniti na nominalnim podacima, a može se koristiti i na ostalim podacima.

Na mod, kao i na medijanu, slabo utiču ekstremne vrednosti (outliners).

Ukoliko bi se iz neke populacije izdvajali uzorci i za svaki izračunala aritmetička sredina, medijana i mod, tada bi mod više varirao od uzorka do uzorka nego medijana i aritmetička sredina.

Jedna podela raspodela obesrviranih podataka:

- Ukoliko u nizu podataka postoji samo jedan sa najvećom frekvencijom onda se takva raspodela zove unimodalna raspodela.
- Moguće je da se u raspodeli pojavljuje više rezultata sa najvećom frekvencijom i onda se takva raspodela zove multimodalna raspodela.
- Posebna podvrsta multimodalne raspodele je bimodalna raspodela gde dva podatka imaju najveću frekvenciju.
- Ukoliko se u nizu podataka svi pojavljuju isti broj puta onda su svi rezultati modovi i takva raspodela se naziva uniformnom.

Mod (2)

Primer. Pretpostavimo da smo merili visinu devojčica u 5_2 odeljenju osnovne škole "X" i da su dobijeni sledeći podaci izraženi u centimetrima:

122, 132, 135, 138, 140, 140, 140, 141, 148, 154 i 160.

Kako su rezultati poređeni od najmanjeg ka najvećem, jasno se uočava da je jedini rezultat koji se pojavljuje više od jednom 140.

Ovaj rezultat se pojavljuje tri puta, a svi ostali rezultati se pojavljuju jednom.

Dakle, mod visine devojčica u 5_2 odeljenju osnovne škole "X" iznosi 140cm.

Kako ova raspodela ima samo jedan mod, radi se o unimodalnoj raspodeli.

U sistemu R, mod se može izračunati tako što se prvo, korišćenjem funkcije **table**, od podataka uzorka dobije tabela frekfencija – promenljiva tabela.frekfencija .

- Ako treba izdvojiti samo jedan mod, dovoljno je funkcijom **sort** sortirati tabelu frekfencija i iz tako sortirane tabele izdvojiti elemenat sa indeksom **nrow()**.
- Ako treba izdvojiti sve modove i nijihov broj pojava, koristi se funkcija **max** nad tabelom frekfencija, ona vraće koliko se puta pojavljuje mod u populaciji. Oformi se logički vektor izrazom tabela.frekfencija==max(tabela.frekfencija), a modovi se dobiju isecanjm tabele frekfencija po logičkom vektoru.

Mod (3)

Zadatak 1. Odrediti mod za trajanje čekanja između dve erupcije u okviru sa podacima faithful.

Kvartili i percentili

U statistici se razlikuje nekoliko ***kvartila***. ***Prvi kvartil***, ili ***donji kvartil***, je vrednost koja „odseca“ prvih 25% podataka kada su podaci sortirani u rastući poredak. ***Drugi kvartil***, ili ***medijana***, je vrednost koja „odseca“ prvih 50% podataka. ***Treći kvartil***, ili ***gronji kvartil***, je vrednost koja „odseca“ prvih 75% podataka.

Kvartili se u R-u određuju primenom funkcije **quantile**, pri čemu se prosleđuje samo jedan argument – uzorak sa podacima čiji se kvartili izračunavaju.

Problem. Odrediti kvartile za trajanje erupcije u okviru sa podacima faithful.
Rešenje. Iz okvira sa podacima se iseku trajanja erupcija i potom se na njih primeni funkcija quantile:

```
> duration = faithful$eruptions     # the eruption durations
> quantile(duration)                # apply the quantile function
    0%    25%    50%    75%   100%
1.6000 2.1627 4.0000 4.4543 5.1000
```

Dakle, prvi kvartil ima vrednost 2.1627, drugi ima vrednost 4, a treći 4.4543.

Zadatak. Odrediti kvartile za čekanje između erupcija u okviru sa podacima faithful.

Kvartili i percentili (2)

Percentili su uopštenja kvartila. Tako je ***n-ti percentil*** vrednost koja „odseca“ prvih n% podataka kada su podaci sortirani u rastući poredak.

Percentili se u R-u određuju primenom funkcije **quantile**, pri čemu se prosleđuju dva argumenta – uzorak sa podacima nad kojim se izračunavaju percentili i vektor koji sadrži vrednosti procenata za koje se izračunavaju percentili.

Problem. Odrediti 32-gi, 57-mi i 98-mi percentil za trajanje erupcije u okviru sa podacima faithful.
Rešenje. Iz okvira sa podacima se iseku trajanja erupcija i potom se na njih primeni funkcija quantile:

```
> duration = faithful$eruptions     # the eruption durations
> quantile(duration, c(.32, .57, .98))
   32%    57%    98%
2.3952 4.1330 4.9330
```

Dakle, 32-gi percentil ima vrednost 2.3952, 57-mi ima vrednost 4.1330, a 98-mi 4.9330.

Zadatak 1. Odrediti 17-ti, 43-ći, 67-mi i 85-ti percentil za čekanje između erupcija u okviru sa podacima faithful.

Kvartili i percentili (3)

Napomena. Kvartili i percentili se mogu izračunavati na više različitih načina. Više o tome se može naći u sistemu pomoći kod R-a, pozivom funkcije help.

```
> help(quantile).
```

Opseg i opseg među kvartilima

Opseg posmatrane promenljive je razlika između najveće i najmanje vrednosti u uzorku. Opseg predstavlja meru u kom intervalu bivaju raspršeni podaci.

Opseg se određuje po sledećoj formuli:

$$Opseg = Najveća\ vrednost\ - Najmanja\ vrednost$$

Opseg se u R-u može odrediti određuju primenom funkcija **min** i **max**.

Problem. Odrediti opseg za trajanje erupcije u okviru sa podacima faithful.
Rešenje. Iz okvira sa podacima se iseku trajanja erupcija i potom se opseg izračuna kao razlika između navećeg i najmanjeg:

Dakle, opseg za trajanje erupcije je 3.5.

Zadatak 1. Odrediti opseg za čekanje između erupcija u okviru sa podacima faithful.

Opseg i opseg među kvartilima (2)

Opseg među kvartilima posmatrane promenljive je razlika između gornjeg i donjeg kvartila uzorka. Opseg predstavlja meru u kom intervalu bivaju raspršeni podaci iz sredine uzorka.

Opseg među kvartilima se određuje po sledećoj formuli:

$$Opseg\ među\ kvartilima = Gornji\ kvartil\ - Donji\ kvartil$$

Opseg među kvartilima se u R-u može odrediti određuju primenom funkcije **IQR**.

Problem. Odrediti opseg među kvartilima za trajanje erupcije u okviru sa podacima faithful.
Rešenje. Jedan način je direktno primenom funkcije IQR:

```
> duration = faithful$eruptions     # the eruption durations
> IQR(duration)                     # apply the IQR function
[1] 2.2915
```

Dakle, opseg među kvartilima za trajanje erupcije je 2.2915 minuta.

Zadatak 1. Odrediti, prema gornjoj formuli, opseg među kvartilima za trajanje erupcije u okviru sa podacima faithful.
Zadatak 2. Odrediti opseg među kvartilima za čekanje između erupcija u okviru sa podacima faithful.

Kutija dijagram

Kutija dijagram (Box plot) posmatrane promenljive je grafička reprezentacija zasnovana na kvartilima, kao i na nakmanjoj i najvećoj vrednosti uzorka. Pomoću nje se pokušava dati vizuelni oblik za raspodelu podataka.

Kod ove vrsta dijagrame, krajnje crte predstavlaju najmanju i najveću vrednost uzorka, ivice kutije predstavlaju donji i gornji kvartil, a podebljana linija u kutiji predstavlja medijanu.

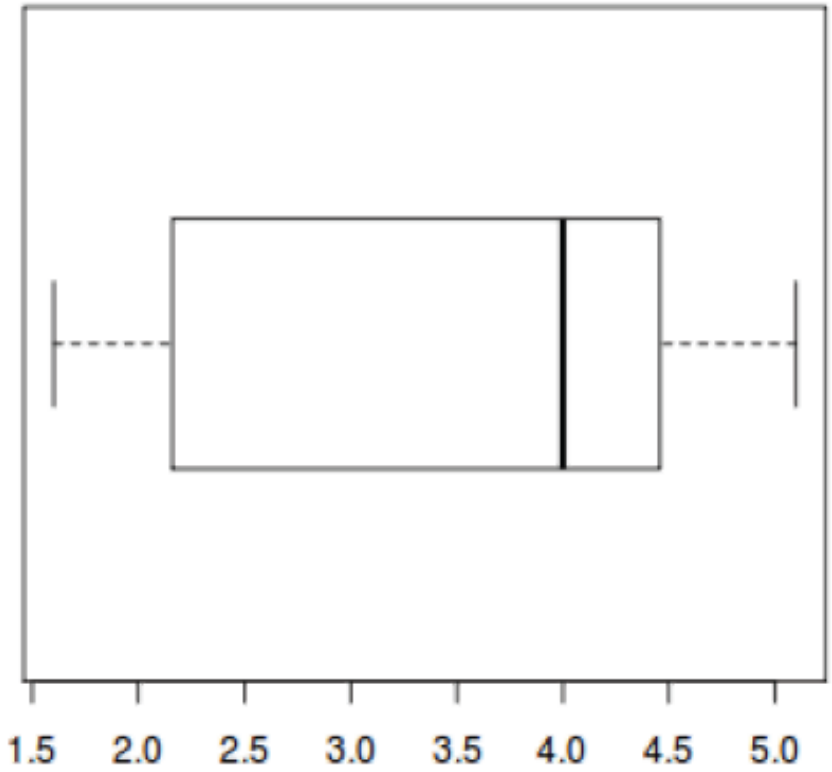
U sistemu R, kutija dijagram se dobija primenom **boxplot** funkcije.

Problem. Odrediti horizontalni kutija dijagram za trajanje erupcije u okviru sa podacima faithful.
Rešenje. Prvo se iseku trajanja iz okvira sa podacima, potom se primenoi funkcija boxplot:

```
> duration = faithful$eruptions       # the eruption durations
> boxplot(duration, horizontal=TRUE)  # horizontal box plot
```

Kutija dijagram (2)

Rešenje (nastavak). Kao rezultat, dobija se sledeći dijagram:



Zadatak 1. Odrediti vertikalni kutija dijagram za trajanje erupcije u okviru sa podacima faithful.

Zadatak 2. Odrediti kutija dijagram za čekanje između erupcija u okviru sa podacima faithful.

Disperzija

Disperzija posmatrane promenljive je numerička mera koja opisuje kako su podaci raspršeni oko sredine. Ona predstavlja prosečno kvadratno odstupanje od sredine.

Za uzorak dimenzije n sa sredinom $\bar{X}$, ***disperzija uzorka*** se određuje po sledećoj formuli:

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

Slično tome, za populaciju koja broji N jedinki i ima sredinu μ, ***disperzija populacije*** se određuje na sledeći način:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

U sistemu R, disperzija se određuje korišćenjem funkcije **var**.

Disperzija (2)

Problem. Odrediti disperziju za trajanje erupcije gejzira u okviru sa podacima faithful.

Rešenje. Treba izvršiti isecanje okvira po koloni eruptions, pa nad izdvojenim podacima primeniti funkciju var:

```
> duration = faithful$eruptions    # the eruption durations
> var(duration)                    # apply the var function
[1] 1.3027
```

Dakle, disperzija za trajanje erupcije gejzira je 1.3027 minuta.

Zadatak 1. Odrediti disperziju za trajanje čekanja između dve erupcije u okviru sa podacima faithful.

Standardna devijacija

Standardna devijacija posmatrane promenljive je kvadratni koren njene disperzije.

Za izračunavanje standardne devijacije u sistemu R koristi se funkcija **sd**.

Problem. Odrediti standardnu devijaciju za trajanje erupcije gejzira u okviru sa podacima faithful.

```
> duration = faithful$eruptions    # the eruption durations
> sd(duration)                     # apply the sd function
[1] 1.1414
```

Dakle, standardna devijacija za trajanje erupcije gejzira je 1.1414 minuta.

Zadatak 1. Odrediti standardnu devijaciju za trajanje čekanja između dve erupcije u okviru sa podacima faithful.

Kovarijansa

Kovarijansa za dve promenljive X i Y je numerička mera koja opisuje koliko su ove dve promenljive linearno korelirane.

Pozitivna kovarijansa ukazuje da postoji pozitivna linearna korelacija između ove dve promenljive, dok negativna kovarijansa ukazuje na postojanje negativne linearne korelacije (veze).

Za ***uzorak*** dimenzije n i promenljive X i Y sa sredinama $\bar{X}$ i $\bar{Y}$, ***kovarijansa*** se određuje po sledećoj formuli:

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

Slično tome, za ***populaciju*** od N jedinki i promenljive X i Y sa sredinama μ_X, μ_Y ***kovarijansa*** se određuje na sledeći način:

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

U sistemu R, kovarijansa se određuje korišćenjem funkcije **cov**, pri čemu se prosleđuju dva argumenta-vektora.

Kovarijansa (2)

Problem. Odrediti kovarijansu za trajanje erupcije gejzira i čekanje između dve erupcije u okviru sa podacima faithful.

Rešenje. Prvo se izvrši isecanje okvira po kolonama eruptions i waiting, pa se nad izdvojenim podacima primeniti funkcija cov:

```
> duration = faithful$eruptions   # the eruption durations
> waiting = faithful$waiting      # the waiting period
> cov(duration, waiting)          # apply the cov function
[1] 13.978
```

Dakle, kovarijansa za trajanje erupcije gejzira i čekanje između dve erupcije je 13.978.

Ovako visoka kovarijansa (visoka, imajući u vidu da je medijum za trajanje 4, a da je opseg trajanja 3.5) ukazuje da postoji pozitivna linearna korelacija između ove dve promenljive.

Koeficijent korelacije

Koeficijent korelacije za dve promenljive X i Y u uzorku podataka je normalizovana mera koja opisuje koliko su one linearno korelirane.

Koeficijent korelacije se izračunava tako što se kovarijansa ove dve promenljive podeli sa proizvodom njihovih standardnih devijacija.

Dakle, za ***uzorak*** se ***koeficijent korelacije*** za promenljive X i Y se određuje po sledećoj formuli:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Slično tome, za ***populaciju*** se ***koeficijent korelacije*** za promenljive X i Y se određuje po sledećoj formuli:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Koeficijent korelacije blizak broju 1 ukazuje da su promenljive pozitivno linearno korelisane. To znači da su tačke XY dijagrama grupisane oko linije koja ima pozitivan nagib.

Koeficijent korelacije blizak broju -1 ukazuje da su promenljive negativno linearno korelisane - tačke XY dijagrama su grupisane oko linije koja ima negativan nagib.

Koeficijent korelacije blizak 0 ukazuje na slabu linearnu korelaciju.

Koeficijent korelacije (2)

U sistemu R, koeficijent korelacije se određuje korišćenjem funkcije **cor**.

Problem. Odrediti koeficijent korelacije za trajanje erupcije gejzira i čekanje između dve erupcije u okviru sa podacima faithful.

Rešenje. Prvo se izvrši isecanje okvira po kolonama eruptions i waiting, pa se nad izdvojenim podacima primeniti funkcija cor:

```
> duration = faithful$eruptions   # the eruption durations
> waiting = faithful$waiting      # the waiting period
> cor(duration, waiting)          # apply the cor function
[1] 0.90081
```

Dakle, koeficijent korelacije za trajanje erupcije gejzira i čekanje između dve erupcije je 0.90081.

Koeficijent korelacije koji je ovako blizu 1 ukazuje da postoji pozitivna linearna korelacija između trajanje erupcije gejzira i čekanje između dve erupcije.

Centralni moment

Centralni moment k-tog reda za promenljivu X meri stepen udaljenosti elemenata od sredine.

Kada je dat ***uzorak***, tada se ***centralni moment k-tog reda*** za promenljivu X određuje po sledećoj formuli:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^k$$

Slično tome, za ***populaciju*** se ***centralni moment k-tog reda*** promenljive X određuje po sledećoj formuli:

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^k$$

Može se uočiti da je centralni moment uopštenje disperzije – centralni moment drugog reda se poklapa sa disperzijom.

U sistemu R, k-ti centralni moment se može računati funkcijom **moment**, koja se nalazi u biblioteci **e1071**. Pre korišćenja ove funkcije, potrebno je učitati biblioteku u kojoj se ta funkcija sadrži.

Centralni moment (2)

Problem. Odrediti centralni moment trećeg reda za trajanje erupcije gejzira u okviru sa podacima faithful.

Rešenje. Prvo se izvrši isecanje okvira po kolonama eruptions, pa se nad izdvojenim podacima primeni funkcija moment:

```
> library(e1071)                    # load e1071
> duration = faithful$eruptions     # eruption durations
> moment(duration, order=3, center=TRUE)
[1] -0.6149
```

Dakle, centralni moment trećeg reda za trajanje erupcije gejzira je -0.6149.

Zadatak 1. Odrediti centralni moment trećeg reda za trajanje čekanja između erupcija gejzira u okviru sa podacima faithful.

Iskošenost i spljoštenost

Iskošenost (skewness) u populaciji je definisana sledećom formulom:

$$\gamma_1 = \frac{\mu_3}{{\mu_2}^{3/2}}$$

Pri tome, μ_3 i μ_2 su centralni momenti reda 3 i reda 2 respektivno.

Intuitivno gledano, iskošenost je mera simetričnosti. Po pravilu, negativna iskošenost ukazuje da je sredina populacije manja od medijane, što znači da je populacija *iskošena ulevo*. Pozitivna iskošenost ukazuje da je sredina veća od medijane, pa je populacija *iskošena udesno*.

U sistemu R, iskošenost se može računati funkcijom **skewness**, koja se nalazi u biblioteci **e1071**. Pre korišćenja ove funkcije, potrebno je učitati biblioteku koja sadrži tu funkciju.

Iskošenost i spljoštenost (2)

Problem. Odrediti iskošenost za trajanje erupcije gejzira u okviru sa podacima faithful.

Rešenje. Prvo se izvrši isecanje okvira po kolonama eruptions, pa se nad izdvojenim podacima primeni funkcija skewness:

```
> library(e1071)                    # load e1071
> duration = faithful$eruptions     # eruption durations
> skewness(duration)                # apply the skewness function
[1] -0.41355
```

Dakle, iskošenost za trajanje erupcije gejzira je -0.41355. To ukazuje da je raspodela trajanja erupcija u okviru sa podacima faithful iskošena ulevo.

Zadatak 1. Odrediti iskošenost za trajanje čekanja između erupcija gejzira u okviru sa podacima faithful.

Iskošenost i spljoštenost (3)

Spljoštenost (kurtosis) je u univarijantnoj populaciji je definisana sledećom formulom:

$$\gamma_2 = \frac{\mu_4}{{\mu_2}^2} - 3$$

Pri tome, μ_4 i μ_2 su centralni momenti reda 4 i reda 2 respektivno.

Intuitivno gledano, spljoštenost je mera postojanja ispupčenja (tj. glatkosti). Po pravilu, negativna kurtoza ukazuje na glatku raspodelu. Pozitivna kurtoza ukazuje da raspodela sa većim brojem ispupčenja.

U sistemu R, spljoštenost se može računati funkcijom **kurtosis**, koja se nalazi u biblioteci **e1071**. Pre korišćenja ove funkcije, potrebno je učitati biblioteku koja sadrži tu funkciju.

Iskošenost i spljoštenost (4)

Problem. Odrediti spljoštenost za trajanje erupcije gejzira u okviru sa podacima faithful.

Rešenje. Prvo se izvrši isecanje okvira po kolonama eruptions, pa se nad izdvojenim podacima primeni funkcija kurtosis:

```
> library(e1071)                    # load e1071
> duration = faithful$eruptions     # eruption durations
> kurtosis(duration)                # apply the kurtosis function
[1] -1.5116
```

Dakle, spljoštenost za trajanje erupcije gejzira je -1.5116.

Zadatak 1. Odrediti spljoštenost za trajanje čekanja između erupcija gejzira u okviru sa podacima faithful.

Raspodela verovatnoća

Slučajne promenljive i raspodele

Ako se događaji (tj. ishodi eksperimenta) mogu predstaviti kao realni brojevi, onda se eksperiment može zamisliti kao izbor jedne promenljive. Promenljiva koja te brojne vrednosti uzima sa određenim verovatnoćama naziva se ***slučajna promenljiva***.

Primer. Novčić se baca jednom. Ako se brojem 0 označi ishod da padne grb, a sa 1 da padne pismo, tada se eksperiment bacanja novčića može posmatrati kao izbor 0 i 1, obe sa verovatnoćom 1/2.
Ovakav pristup omogućava jedan apstraktan model koji može da opiše i druge eksperimente sličnog oblika.

Definicija. Funkcija X koja svakom mogućem ishodu eksperimenta ω dodeljuje realni broj X (ω) zove se ***slučajna promenljiva***. Slučajne promenljive se obeležavaju velikim slovima X,Y,Z...

Primer. Novčić se baca dva puta. Neka je slučajna promenljiva X broj registrovanih pisama. Kako je slup mogućih ishoda {PP,PG,GP,GG} , onda je X(PP) = 2,X(GP)=1,X(PG)=1,X(GG) = 0. Dakle, slučajna promenljiva uzima tri moguće vrednosti 0,1,2.

$$X: \begin{pmatrix} PP & PG & GP & GG \\ 2 & 1 & 1 & 2 \end{pmatrix}$$

Slučajne promenljive i raspodele (2)

Razlikuju se dva osnovna tipa slučajnih promenljivih, **diskretne** i **neprekidne** slučajne promenljive. Podela se vrši u zavisnosti da li slučajna promenljiva uzima vrednosti u konačnom (odnosno prebrojivom) ili neprebrojivom (odnosno kontinum) skupu vrednosti.

Definicija. Neka slučajna promenljiva X može da uzme vrednosti $x_1, x_2,..., x_n$ sa verovatnoćama $p_1, p_2,..., p_n$ respektivno, pri čemu je $\sum_{i=1}^{n} p_i = 1$.
Zakon raspodele verovatnoća je skup uređenih parova $\{(x_i, p_i)\}$, $i = 1,2, \dots, n$.
Zakon raspodele verovatnoća za diskretnu slučajnu promenljivu X se obično zapisuje na sledeći način:

$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$$

Dakle, zakon raspodele verovatnoća slučajne promenljive je pravilo po kome se svakoj od vrednosti slučajne promenljive X pridružuje odgovarajuća verovatnoća.

Njime se daje odgovor na pitanje kolika je verovatnoća da slučajna promenljiva X ima tačno vrednost x_i, tj. verovatnoća događaja $\{X = x_i\}$.

Zakonom raspodele je ukupna verovatnoća (koja je jednaka 1) raspodeljena na pojedine vrednosti slučajne promenljive.

Slučajne promenljive i raspodele (3)

Primer. Novčić se baca dva puta. Neka je slučajna promenljiva X broj registrovanih pisama. Zakon raspodele za slučajnu promenljivu X je:

$$\begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Napomena. Zakon raspodele verovatnoća za slučajnu promenljivu postoji samo kod diskretnih slučajnih promenljivih.
Za neprekidne sličajne promenljive ne postoji zakon raspodele verovatnoća.

Kod neprekidne slučajne promenljive X odgovor na pitanje „kolika je verovatnoća da X ima vrednost tačno x_i, tj. kolika je verovatnoća događaja $\{X = x\ \}$“ je jedinstven i iznosi 0. Nulta verovatnoća ne znači da će da se u praksi događaj $\{X = x\ \}$ nikada neće ostvariti, već samo da je mala, veoma mala verovatnoća da će se on ostvariti. To je prihvatljivo, jer slučajna promenljiva X može da uzme bilo koju vrednost sa intervala, a njih je neprebrojivo (tj. kontinum) mnogo.

Međutim, iako za neprekidnu slučajnu promenljivu nema mnogo smisla postavljati pitanje o verovatnoći događaja $\{X = x_i\}$, veoma je smisleno postaviti pitanje o verovatnoći događaja $\{X < x_i\}$.

Slučajne promenljive i raspodele (4)

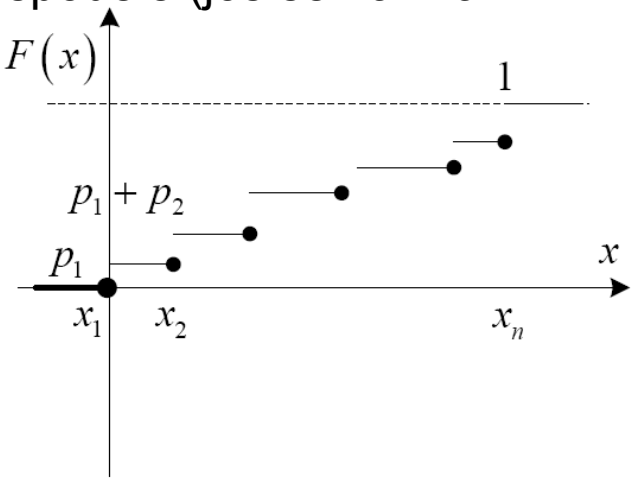
Do potpune karakterizacije za slučajnu promenljivu (i diskretnu i neprekidnu) se dolazi ako se posmatraju događaji $\{X < x\}$, a ne $\{X = x\}$.

Verovatnoća događaja $\{X < x\}$, zavisi od x, odnosno predstavlja funkciju od x, a ta funkcija F(x) se naziva ***funkcija raspodele verovatnoća*** ili ***funkcija raspodele***.

Drugim rečima, funkcija raspodele F(x) je statistička karakteristika slučajne promenljive X, koja omogućava da se izračuna verovatnoća događaja da slučajna promenljiva ima vrednost unutar datog intervala.

Za diskretnu slučajnu promenljivu X, njena funkcija raspodele (još se naziva i ***kumulativna funkcija***) je sledećeg oblika:

$$F(x) = \begin{cases} 0, & x \le x_1 \\ p_1, & x_1 < x \le x_2 \\ p_1 + p_2, & x_2 < x \le x_3 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ 1, & x > x_n \end{cases}$$



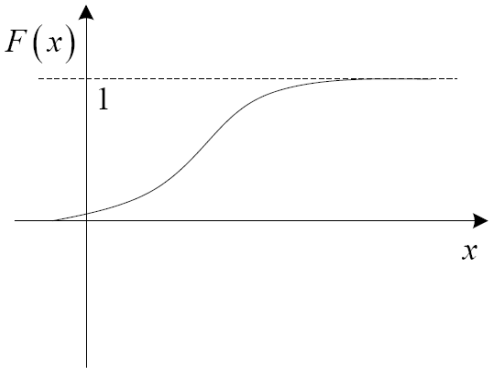
Slučajne promenljive i raspodele (5)

Primer. Novčić se baca dva puta. Neka je slučajna promenljiva X broj registrovanih pisama. Zakon raspodele i funkcija raspodele za slučajnu promenljivu X je:

$$\begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix} \qquad F(x) = \begin{cases} 0, & x < 0 \\ 1/4, & 0 \le x < 1 \\ 3/4, & 1 \le x < 2 \\ 1, & 2 \le x \end{cases}$$

Za neprekidnu slučajnu promenljivu X, funkcija raspodele F(X) je realna, neopadajuća i neprekidna funkcija F(x), takva da je F(x)=P(X<x) i koja ispunjava uslove $F(-\infty) = 0, F(+\infty) = 1$.

Primer. Funkcija prikazana sledećim grafikom može predstavliati funkciiu raspodele neprekidne slučajne promenljive:

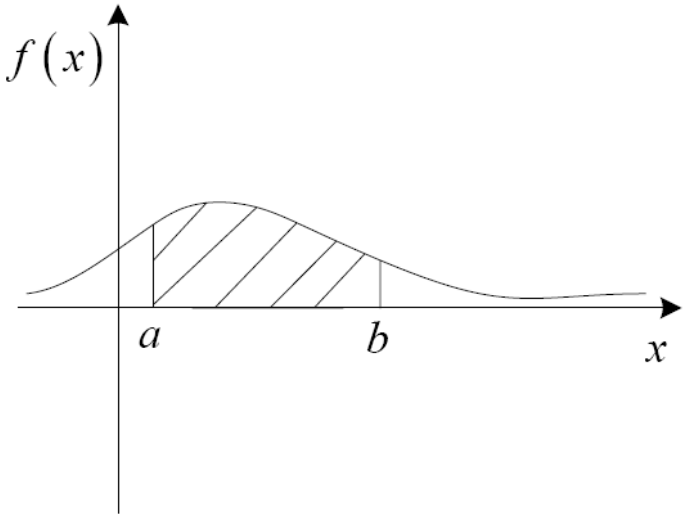


Slučajne promenljive i raspodele (6)

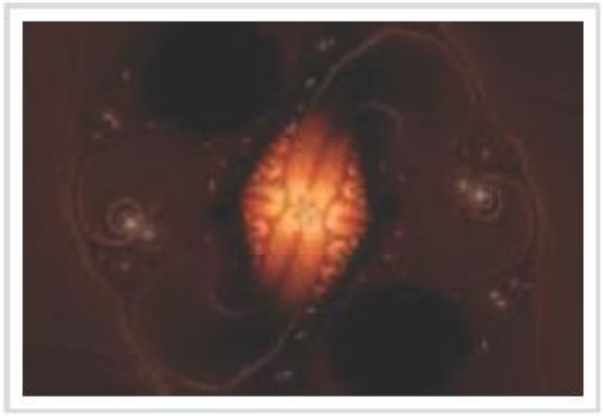
Funkcija raspodele određuje bitne osobine slučajne promenljive. Tako, verovatnoća da je vrednost slučajne promenljive X u intervalu (a,b) može da se odredi preko funkcije raspodele: $P\{a < X < b\} = F(b) - F(a)$

Definicija. Neka je X slučajna promenljiva i F njena funkcija raspodele. Kada postoji funkcija f, takva da je $f(x) = F'(x)$, funkcija f naziva se ***funkcijom gustine raspodele*** slučajne promenljive X.

Verovatnoća da je vrednost slučajne promenljive X u intervalu (a,b) može da se odredi preko funkcije gustine raspodele: $P\{a < X < b\} = \int_a^b f(x)dx$



Slučajne promenljive i raspodele (7)



U okviru ove prezentacije, biće upratko opisane neke od poznatijih raspodela slučajnih promenljivih, kao i funkcije u R-u koje se koriste za rad sa tim raspodelama.

Biće reči o sledećim raspodelama:

- Binomna raspodela
- Poasonova raspodela
- Uniformna raspodela
- Eksponencionalna raspodela
- Normalna raspodela
- Hi-kvadratna raspodela
- Studentova t raspodela
- Fišerova F raspodela

Binomna raspodela

Binomna raspodela je raspodela diskretne slučajne promenljive. Ona opisuje ishod n nezavisnih proba iste vrste u okviru eksperimenta. Svaka od proba može imati samo dva ishoda: uspeh ili neuspeh. Ako je verovatnoća uspeha probe p, tada je verovatnoća da će u n nezavisnih ponavljanja probe biti tačno x uspeha ($0 \leq x \leq n$) data sa sledećom formulom:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Zapis da slučajna promenljiva X ima binomnu raspodelu je: $X \sim \boldsymbol{B}(n, p)$.

U sistemu R, za rad sa binomnom raspodelom koriste se sledeće funkcije:

- **dbinom** - zakon raspodele,
- **pbinom** – funkcija raspodele,
- **qbinom** – određivanje kvantila,
- **rbinom** - generator pseudoslucajnih brojeva.

Kod ovih funkcija, imenovani argument **size** označava broj nezavisnih proba (to je n u gornjoj formuli), a imenovani argument **prob** označava verovatnoću uspešnog ishoda za jednu probu (u gornjoj formuli, ta veličina je označena sa p).

Binomna raspodela (2)

Problem. Odrediti verovatnocu da student "na srecu" tačno da 4 odgovora na 12 pitanja (za svako pitanje je ponuđeno 5 alternativa za odgovor).

Rešenje. Verovatnoća da student tačno odgovori na jedno pitanje je 1/5=0.2. Korišćenjem funkcije qbinom dobija se:

```
> tacna.4.od.12 <- dbinom(4, size=12, prob=0.2)
> tacna.4.od.12
[1] 0.1328756
```

Dakle, verovatnoća je da student „na sreću“ pogodi tačno 4 odgovora je približno 13,3%.

Problem. Odrediti verovatnocu da student „pogodi" ne više od 4 odgovora na 12 pitanja (za svako pitanje je ponuđeno 5 alternativa za odgovor).

Rešenje. Korišćenjem funkcije za zakon raspodele verovatnoća binomne promenljive qbinom se dobija:

```
> tacna.do.4.od.12 <- dbinom(0, size=12, prob=0.2
+                            )  + dbinom(1, size=12, prob=0.2
+                                        ) + dbinom(2, size=12, prob=0.2
+                    ) + dbinom(3, size=12, prob=0.2) + dbinom(4, size=12, prob=0.2)
> tacna.do.4.od.12
[1] 0.9274445
```

Verovatnoća je da student „na sreću“ pogodi ne više od 4 odgovora je oko 92,74%.

Binomna raspodela (3)

Problem. Odrediti verovatnocu da student „na sreću" tačno odgori na ne više od 4 odgovora na 12 pitanja (za svako pitanje je ponudjeno 5 alternativa za odgovor).

Rešenje. Isti problem se sada rešava na nešto drugačiji način. Korišćenjem funkcije za kumulativnu raspodelu verovatnoća binomne promenljive pbinom se dobija:

```
> tacna.do.4.od.12 <- pbinom(4, size=12, prob=0.2)
> tacna.do.4.od.12
[1] 0.9274445
```

Dakle, isto kao u prethodnom problemu (što je i očekivano) verovatnoća je da student „na sreću“ pogodi ne više od 4 odgovora je približno 92,74%.

Binomna raspodela (4)

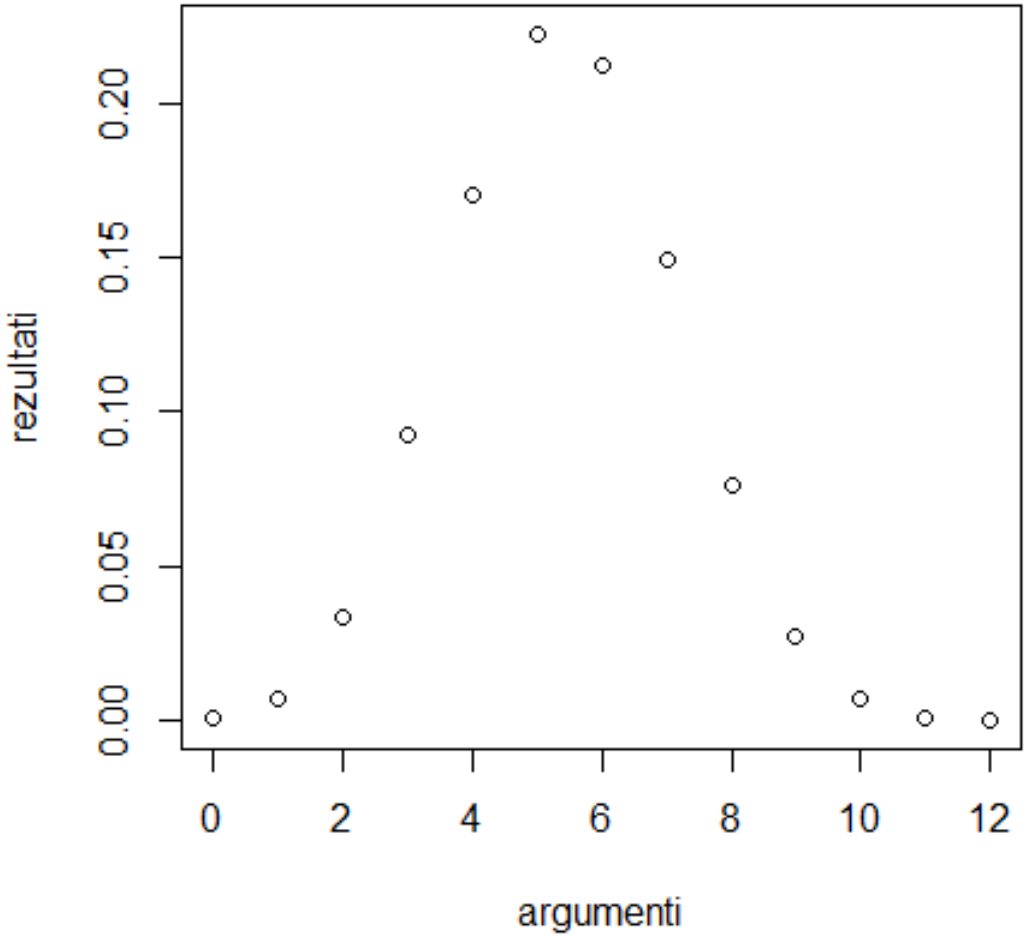
Problem. Odrediti raspodelu verovatnoca da student "na sreću" pogodi tačno x od ukupno n pitanja. Verovatnoca da pogodi jedno pitanje je p. Neka u ovom slučaju bude n=12, p=0.45.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije za zakon raspodele verovatnoća binomne promenljive dbinom dobije vektor rezultata i onda se funkcijom plot izvrši iscrtavanje:

```
> n <- 12
> p <- 0.45
> argumenti <- seq(0,n)
> argumenti
 [1]  0  1  2  3  4  5  6  7  8  9 10 11 12
> rezultati <- dbinom( argumenti, size=n, prob=p)
> rezultati
 [1] 7.662179e-04 7.522866e-03 3.385290e-02 9.232609e-02 1.699639e-01 2.224982e-01 2.123847e-01
 [9] 7.616511e-02 2.769640e-02 6.798208e-03 1.011304e-03 6.895252e-05
> plot(argumenti, rezultati)
```

Binomna raspodela (5)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Binomna raspodela (6)

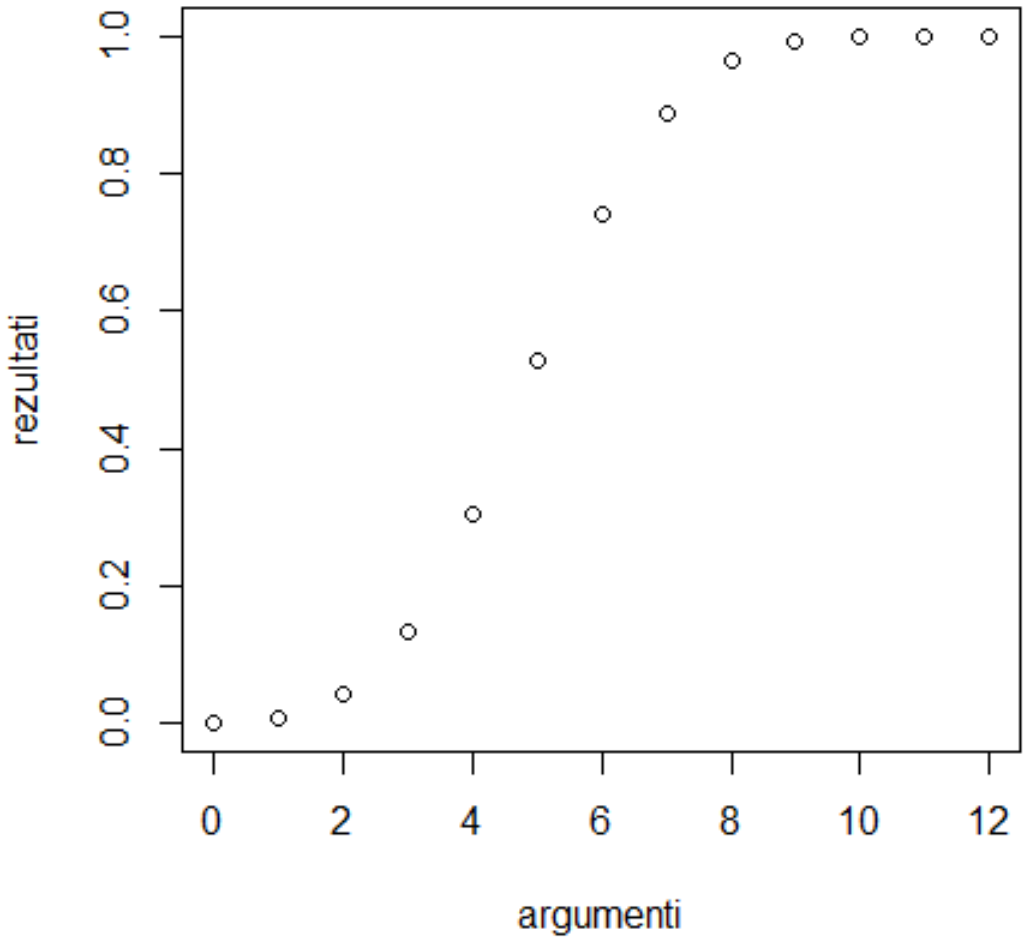
Problem. Odrediti raspodelu verovatnoca da student "na sreću" pogodi x ili manje od ukupno n pitanja. Verovatnoca da pogodi jedno pitanje je p. Neka u ovom slučaju bude n=12, p=0.45.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije za kumulativnu raspodelu verovatnoća binomne promenljive pbinom dobije vektor rezultata i onda se funkcijom plot izvrši iscrtavanje:

```
> n <- 12
> p <- 0.45
> argumenti <- seq(0,n)
> argumenti
 [1]  0  1  2  3  4  5  6  7  8  9 10 11 12
> rezultati <- pbinom( argumenti, size=n, prob=p)
> rezultati
 [1] 0.0007662179 0.0082890842 0.0421419826 0.1344680692 0.3044320013 0.5269302398 0.7393149219
 [9] 0.9644251325 0.9921215357 0.9989197438 0.9999310475 1.0000000000
> plot(argumenti, rezultati)
```

Binomna raspodela (7)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Poasonova raspodela

Poasonova raspodela je raspodela diskretne slučajne promenljive. Ona predstavlja raspodelu verovatnoće za broj nezavisnih događaja u datom intervalu.

Ako je λ srednja vrednost za broj događanja događaja u datom intervalu, tada verovatnoća da će datom intervalu biti tačno x događaja data sa:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Zapis da slučajna promenljiva X ima Poasonovu raspodelu je: $X \sim \boldsymbol{P}(\lambda)$.

Kada je kod binomne raspodele $n \geq 30$ i $np < 10$, razlika između binomne i Puasonove raspodele je vrlo mala, pa se u tim slučajevima binomna raspodela može aproksimirati Poasonovom, gde je $\lambda = np$.

U sistemu R, za rad sa Posaonovom raspodelom koriste se sledeće funkcije:

- **dpois** - zakon raspodele,
- **ppois** – funkcija raspodele,
- **qpois** – određivanje kvantila,
- **rpois** - generator pseudoslucajnih brojeva.

Imenovani argument **lambda** označava srednju vrednost broja događaja u datom intervalu (tj. λ u gornjoj formuli).

Poasonova raspodela (2)

Problem. Neka 12 automobila u proseku predje preko mosta za 1 minut. Odrediti verovatnocu da u datom minutu preko mosta pređe tačno 16 automobila.

Rešenje. Korišćenjem funkcije za zakon raspodele verovatnoća qpois dobija se:

```
> verovatnoca.16 <- dpois(16, lambda=12)
> verovatnoca.16
[1] 0.05429334
```

Dakle, verovatnoća da preko mosta pređe tokom jednog minuta pređe tačno 16 automobila je približno 5,43%.

Problem. Neka 12 automobila u proseku predje preko mosta za 1 minut. Odrediti verovatnocu da u datom minutu preko mosta pređe 17 ili više automobila.

Rešenje. Korišćenjem funkcije za kumulativnu raspodelu verovatnoća ppois se prvo odredi verovatnoća da je preko mosta za minut prošlo 16 ili manje automobila, a konačan rezultat se dobije kad se od jedinice oduzme ta razlika :

```
> verovatnoca.16.ili.manje <- ppois(16, lambda=12) # donji kraj
> verovatnoca.17.ili.vise <- 1 - verovatnoca.16.ili.manje
> verovatnoca.17.ili.vise
[1] 0.101291
```

Verovatnoća da preko mosta pređe tokom jednog minuta pređe 17 ili više automobila je približno 1,01%.

Poasonova raspodela (3)

Problem. Neka 12 automobila u proseku predje preko mosta za 1 minut. Odrediti verovatnocu da u datom minutu preko mosta pređe 17 ili više automobila.

Rešenje. Problem iz prethodnog primera se sada rešava na nešto drugačiji način. Korišćenjem funkcije za kumulativnu raspodelu verovatnoća binomne promenljive pbinom, pri čemu se imenovani argument lower postavi na FALSE (što ukazuje da se računa gornja a ne donja vrednost), direktno se dobija traženi rezultat:

```
> verovatnoca.17.ili.vise <- ppois(16,
+                                  lambda=12,
+                                  lower=FALSE # gornji kraj
+                                  )
> verovatnoca.17.ili.vise
[1] 0.101291
```

Dakle, isto kao u prethodnom problemu (što je i očekivano) da preko mosta pređe tokom jednog minuta pređe 17 ili više automobila je približno 1,01%.

Poasonova raspodela (4)

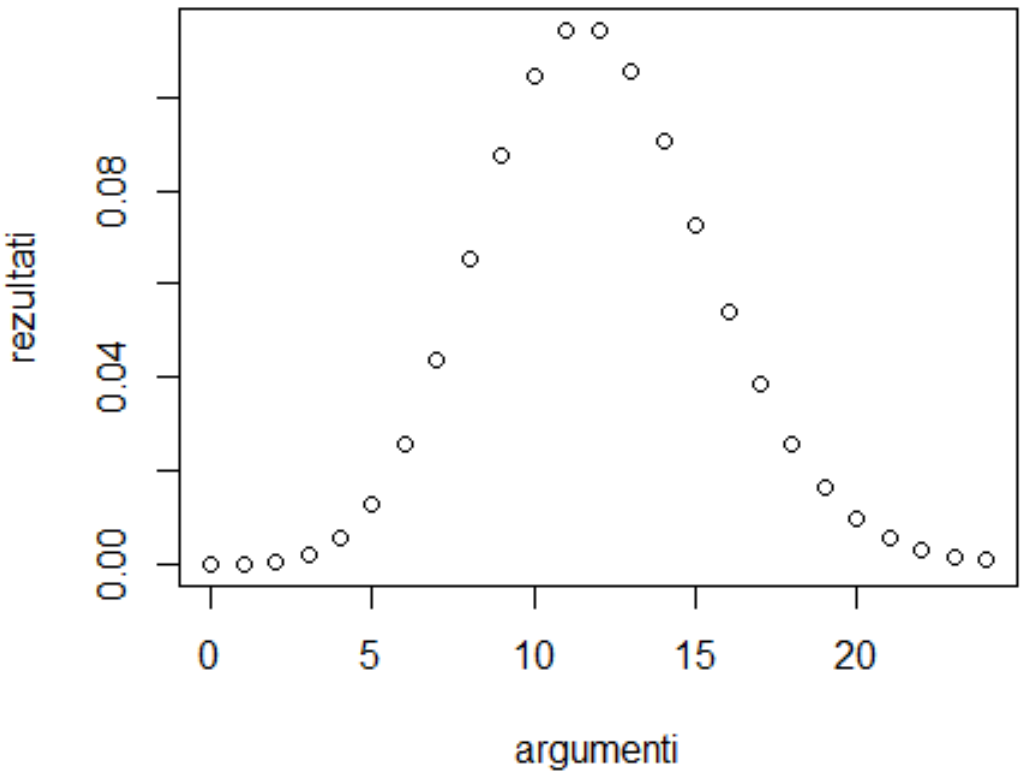
Problem. Prikazati zakon funkcije raspodele za Poasonovu raspodelu. Neka u ovom slučaju bude λ=12.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije za zakon raspodele verovatnoća Poasonove promenljive dpois dobije vektor rezultata i na kraju se funkcijom plot izvrši iscrtavanje:

```
> lambda.parametar <- 12
> opseg <- 2 * lambda.parametar
> argumenti <- seq(0,opseg)
> rezultati <- dpois(argumenti, lambda=lambda.parametar)
> plot(argumenti, rezultati)
```

Poasonova raspodela (5)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Poasonova raspodela (6)

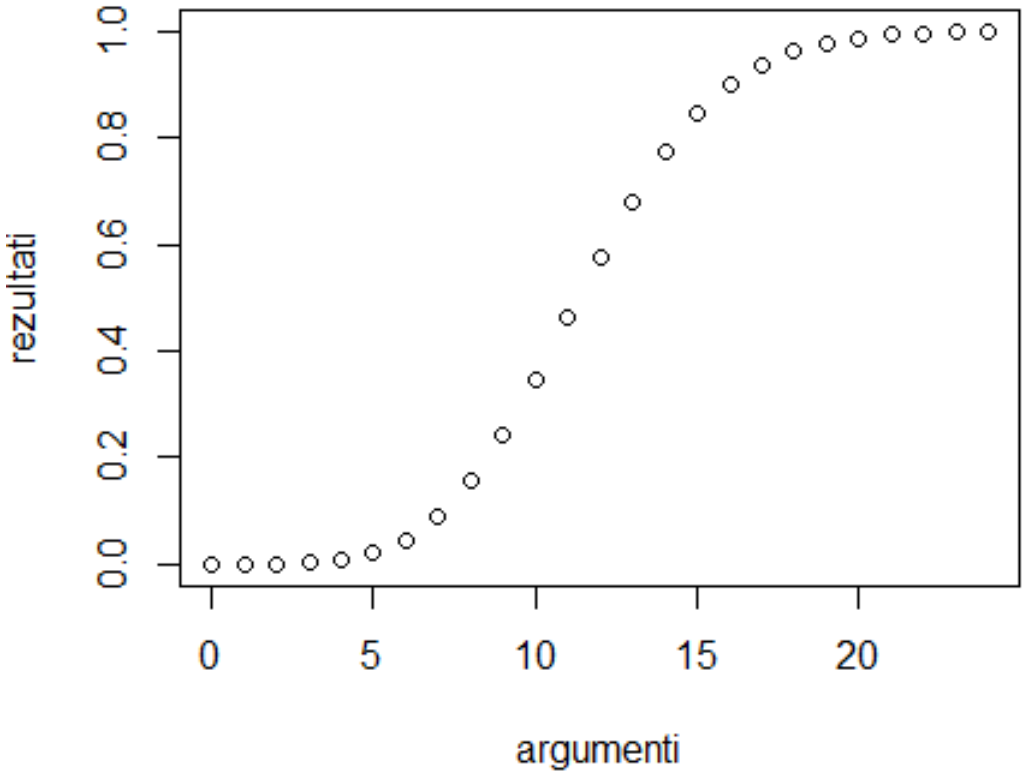
Problem. Prebrojavaju se automobili koji prelaze preko mosta. Neka slučajna promenljiva X predstavlja broj automobila ima Posaonovu raspodelu. Neka lambda, tj. parametar koji predstavlja broj automobila koji u proseku predje preko mosta za 1 minut ima vrednost 12. Odrediti kumulativnu raspodelu za X, tj. verovatnocu da u datom minutu broj automobila koji prelaze preko mosta bude manji ili jednak od x.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije za kumulativnu raspodelu verovatnoća Poasonove promenljive ppois dobije vektor rezultata, a na kraju se funkcijom plot izvrši iscrtavanje:

```
> lambda.parametar <- 12
> opseg <- 2 * lambda.parametar
> argumenti <- seq(0,opseg)
> rezultati <- ppois( argumenti, lambda=lambda.parametar)
> plot(argumenti, rezultati)
```

Poasonova raspodela (7)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Uniformna raspodela

Uniformna (ravnomerna) raspodela je raspodela neprekidne slučajne promenljive. Ona predstavlja raspodelu promenljive čija se vrednost dobija slučajnim izborom iz intervlala (a, b). Funkcija gustine raspodele ove slučajne promenljive data je sledećom formulom:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \leq x \leq b \\ 0 & inače \end{cases}$$

Zapis da slučajna promenljiva X ima Uniformnu raspodelu je: $X \sim \boldsymbol{U}(a, b)$.

U sistemu R, za rad sa uniformnom raspodelom koriste se sledeće funkcije:

- **dunif** – funkcija gustine raspodele,
- **punif** – funkcija raspodele,
- **qunif** – određivanje kvantila,
- **runif** - generator pseudoslucajnih brojeva.

Kod ovih funkcija, imenovani argument **min** označava levi kraj intervala (to je a u gornjoj formuli), a imenovani argument **max** označava desni kraj intervala (u gornjoj formuli, ta veličina je označena sa b).

Uniformna raspodela (2)

Problem. Oformiti osam pseudoslucajnih brojeva izmedju 1 i 3.

Rešenje. Ovi pseudoslucajni brojevi ima uniformnu raspodelu – predstavljaju moguće vrednosti slučajne promenljive sa uniformnom raspodelom i krajevima intervala 1 i 3. Korišćenjem funkcije runif, dobija se traženi rezultat:

```
> slucajni <- runif(8, min=1, max=3)
> slucajni
[1] 1.615499 2.999713 1.120795 1.328624 1.205883 1.106537 1.835271 1.865801
```

Uniformna raspodela (3)

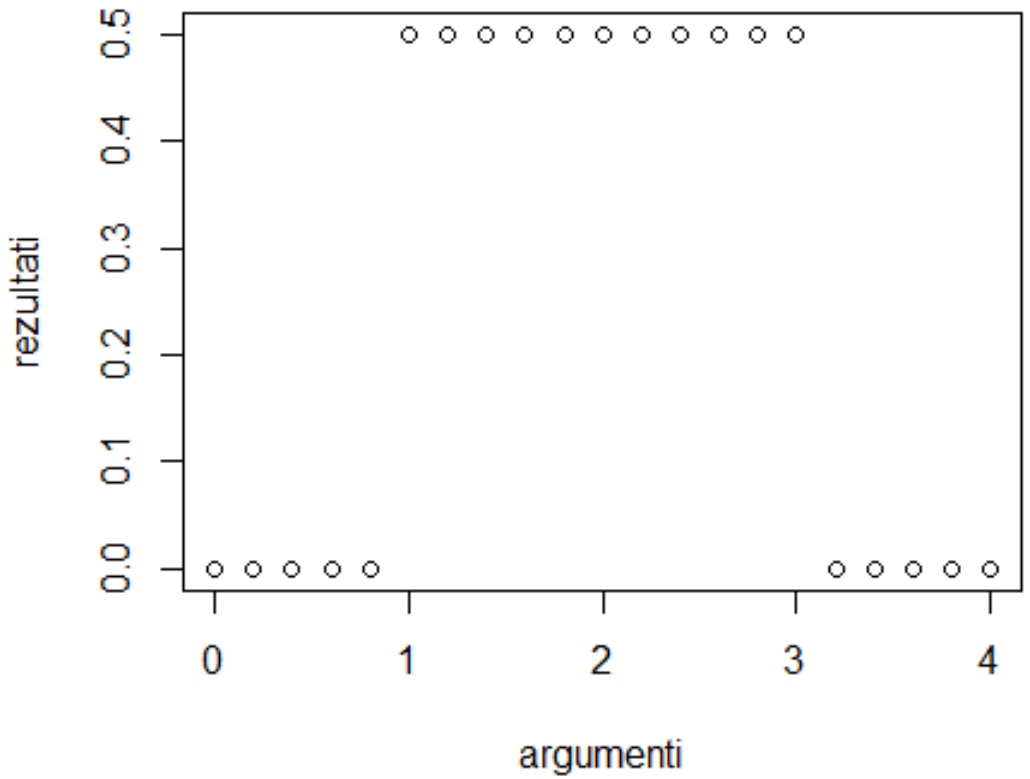
Problem. Prikazati gustinu raspodele za uniformnu raspodelu, pri čemu je ovom slučaju $a = 1 \;\; b = 3$.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije dunif dobije vektor rezultata i na kraju se funkcijom plot izvrši iscrtavanje:

```
> donja.granica <- 1
> gornja.granica <- 3
> opseg <- gornja.granica - donja.granica
> podela <- 10
> argumenti <- seq(donja.granica-1, gornja.granica+1, by=opseg/podela)
> rezultati <- dunif( argumenti, min=donja.granica, max=gornja.granica )
> plot(argumenti, rezultati)
```

Uniformna raspodela (4)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Uniformna raspodela (5)

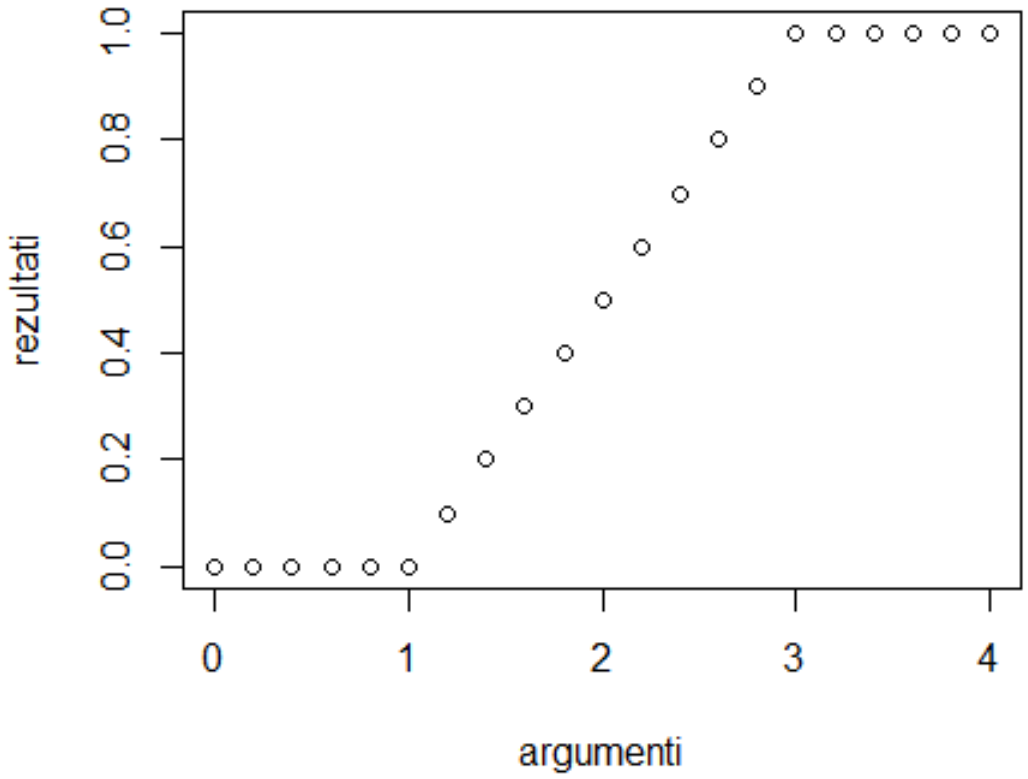
Problem. Prikazati funkciju raspodele za slučajnu promenljivu koja ima uniformnu raspodelu, gde je $a = 1$ $b = 3$. Drugim rečima, prikazati raspodelu verovatnoca da slucajno izabran broj iz intervala $(1,3)$ bude manji ili jednak od x.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, potom se funkcijom punif dobije vektor rezultata i na kraju se funkcijom plot izvrši iscrtavanje:

```
> donja.granica <- 1
> gornja.granica <- 3
> opseg <- gornja.granica - donja.granica
> podela <- 10
> argumenti <- seq(donja.granica-1, gornja.granica+1, by=opseg/podela)
> rezultati <- punif( argumenti, min=donja.granica, max=gornja.granica )
> plot(argumenti, rezultati)
```

Uniformna raspodela (6)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Eksponencionalna raspodela

Eksponencionalna raspodela je raspodela neprekidne slučajne promenljive. Ona predstavlja vreme dešavanja za sekvencu nezavisnih događaja koji se ponavljaju s vremena na vrema u sličajnim intervalima. Ako je sa μ označeno srednje vreme čekanja do sledećeg događaja, tada je funkcija gustine raspodele ovakve slučajne promenljive data sledećom formulom:

$$f(x) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}} & x \geq 0 \\ 0 & inače \end{cases}$$

Zapis da slučajna promenljiva X ima Eksponencionalnu raspodelu je: $X \sim \boldsymbol{E}(\mu)$.

U sistemu R, za rad sa eksponencionalnom raspodelom se koriste:

- **dexp** – funkcija gustine raspodele,
- **pexp** – funkcija raspodele,
- **qexp** – određivanje kvantila,
- **rexp** - generator pseudoslucajnih brojeva.

Kod ovih funkcija, imenovani argument **rate** označava srednje vreme čekanja do sledećeg događaja (u gornjoj formuli, ta veličina je označena sa μ).

Eksponencionalna raspodela (2)

Problem. Neka je srednje vreme za obradu na kasi 3 minuta. Odrediti verovatnoću da je obrada završena za manje od 2 minuta.

Rešenje. Slučajna promenljiva koja predstavlja vreme čekanja na kasi ima eksponecionalnu raspodelu, pri čemu je brzina obrade je 1/3 musterije po minuti. Korišćenjem funkcije qexp, dobija se traženi rezultat – verovatnoća da će obrada biti završena za manje od dva minuta:

```
> verovatnoca <- pexp(2, rate=1/3)
> verovatnoca
[1] 0.4865829
```

Dakle, verovatnoća da je obrada zavrsena za manje od 2 minuta je približno 48,66%.

Eksponencionalna raspodela (3)

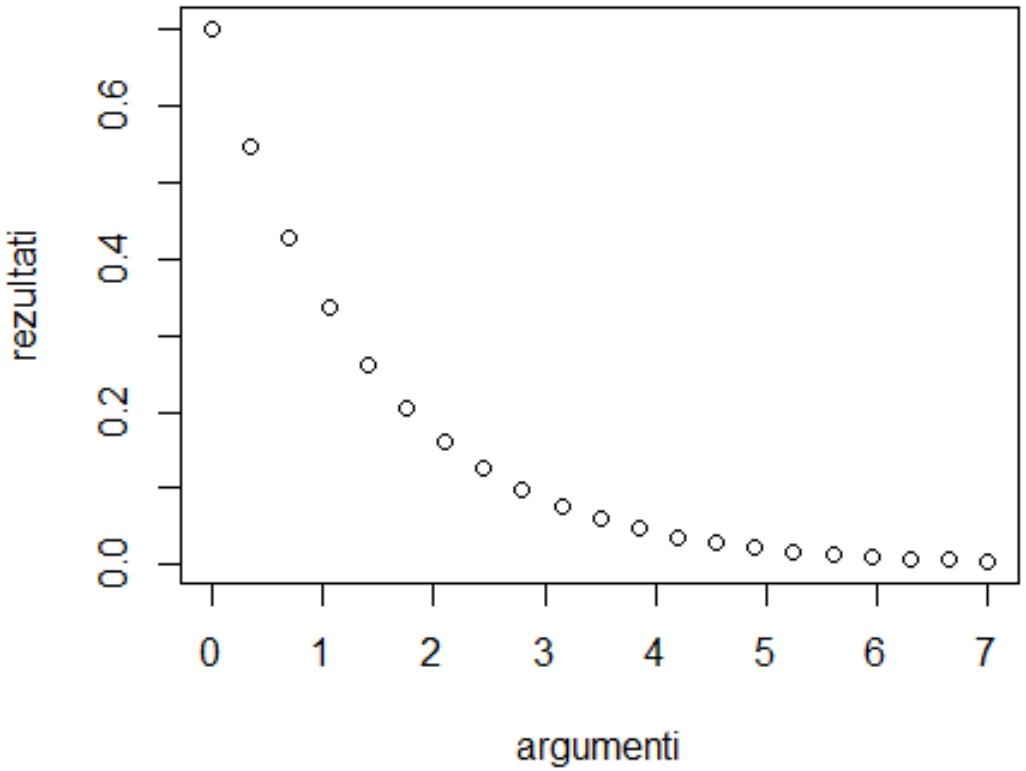
Problem. Prikazati gustinu raspodele za eksponencionalnu raspodelu, pri čemu je ovom slučaju $\mu = 0.7$.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije dexp dobije vektor rezultata i na kraju se funkcijom plot izvrši iscrtavanje:

```
> mi.parametar <- 0.7
> donja.granica <- 0
> gornja.granica <- 7
> opseg <- gornja.granica - donja.granica
> broj.tacaka <- 20
> argumenti <- seq(donja.granica, gornja.granica, by=opseg/broj.tacaka)
> rezultati <- dexp( argumenti, rate = mi.parametar )
> plot(argumenti, rezultati)
```

Eksponencionalna raspodela (4)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Eksponencionalna raspodela (5)

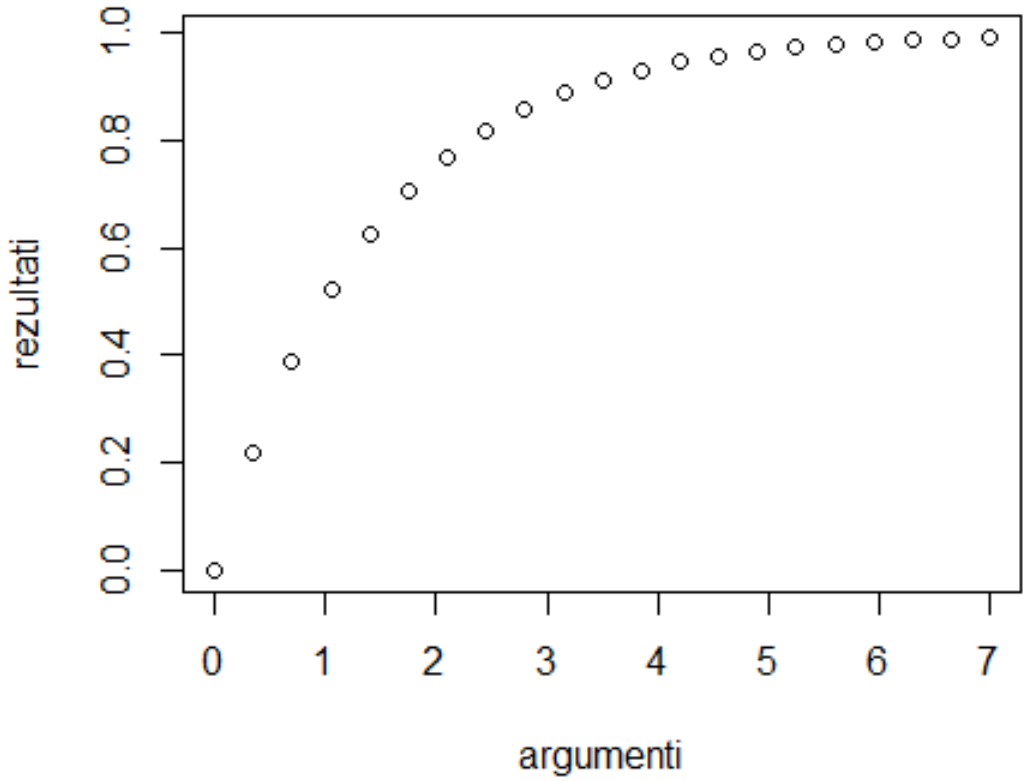
Problem. Prikazati funkciju raspodele za slučajnu promenljivu koja ima eksponencionalnu raspodelu sa $\mu = 0.7$. Drugim rečima, prikazati raspodelu verovatnoca do je obrada zavrsena za manje od x minuta, kada je srednje vreme obrade μ.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, potom se funkcijom pexp dobije vektor rezultata. Na kraju, funkcijom plot se izvrši iscrtavanje:

```
> mi.parametar <- 0.7
> donja.granica <- 0
> gornja.granica <- 7
> opseg <- gornja.granica - donja.granica
> broj.tacaka <- 20
> argumenti <- seq(donja.granica, gornja.granica, by=opseg/broj.tacaka)
> rezultati <- pexp( argumenti, rate = mi.parametar )
> plot(argumenti, rezultati)
```

Eksponencionalna raspodela (6)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Normalna raspodela

Normalna raspodela je raspodela neprekidne slučajne promenljive. Ovo je jedna od najznačajnijih raspodela verovatnoća. Ako je sa μ označena sredina populacije, a sa σ^2 disperzija populacije, tada je funkcija gustine raspodele slučajne promenljive data sledećom formulom:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Zapis da slučajna promenljiva X ima normalnu raspodelu je: $X \sim \boldsymbol{N}(\mu, \sigma^2)$.

Kada je $\mu = 0$ i $\sigma = 1$, tada slučajna promenljiva X ima normiralnu normalnu raspodelu, tj. $X \sim \boldsymbol{N}(0,1)$.

Ako slučajna promenljiva X ima normalnu raspodelu (tj. $X \sim \boldsymbol{N}(\mu, \sigma^2)$), tada slučajna promenljiva $Y = \frac{X-\mu}{\sigma}$ ima normiranu normalnu raspodelu, tj. $Y \sim \boldsymbol{N}(0,1)$.
Kaže se da je slučajna promenljiva Y kreirana centriranjem i normiranjem slučajne promenljive X.

Dovoljno je znati kako da se izračunaju verovatnoće za normiranu normalnu raspodelu i onda se može izračunati sve što je potrebno za normalnu raspodelu.

Normalna raspodela (2)

Normalna raspodela ima veliki značaj zbog ***Centalne granične teoreme***, koja tvrdi da je (normirana i centrirana) suma velikog broja nezavisnih i identično raspoređenih slučajnih promenljivih teži normalnoj raspodeli verovatnoće.

Teorema. Neka su $X_1, X_2, \ldots, X_n$ nezavisne slučajne promenljive koje imaju istu, ne nužno normalnu raspodelu sa sredinom μ i standardnom devijacijom σ. Tada slučajna promenljiva $S_n = X_1 + X_2 + \cdots + X_n$ ima sredinu $n\mu$ i disperziju $n\sigma^2$. Uočimo slučajnu promenljivu $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. Kada $n \to \infty$, tada raspodela verovatnoća za Z_n teži ka normiranoj normalnoj raspodeli, tj. ka $\boldsymbol{N}(0,1)$.

U sistemu R, za rad sa normalnom raspodelom koriste se sledeće funkcije:

- **dnorm** – funkcija gustine raspodele,
- **pnorm** – funkcija raspodele,
- **qnorm** – određivanje kvantila,
- **rnorm** - generator pseudoslucajnih brojeva.

Kod ovih funkcija, imenovani argument **mean** označava sredinu (tj. μ), a imenovani argument **sd** označava standardnu devijaciju (tj. σ).

Normalna raspodela (3)

Problem. Neka se rezultati testova uklapaju u normalnu raspodelu sa sredinom 72 i sa stanadrdnom devijacijom 15.2. Odrediti procenat studenata koji su osvojili 84 ili vise bodova.

Rešenje. Korišćenjem funkcije pnorm, dobija se traženi rezultat. S obzirom da je normalna raspodela simetrična u odnosu na sredinu, može se birati da li se računa i gornja i donja strana, ili samo gornja ili samo donja. U ovom slučaju, potrebna je samo gornja strana, pa se imenovani parametar **lower.tail** postavlja na FALSE:

```
> verovatnoca <- pnorm(84, # vrednost
+                      mean = 72, # sredina
+                      sd = 15.2, # standardnd devijacija
+                      lower.tail=FALSE # interesuje nas samo gornja strana
+   )
> verovatnoca
[1] 0.2149176
> procenat <- 100 * verovatnoca
> procenat
[1] 21.49176
> staro <- options(digits=4)
> procenat
[1] 21.49
> options(staro)
```

Dakle, procenat studenata sa 84 ili više bodova je približno 21,49%.

Normalna raspodela (4)

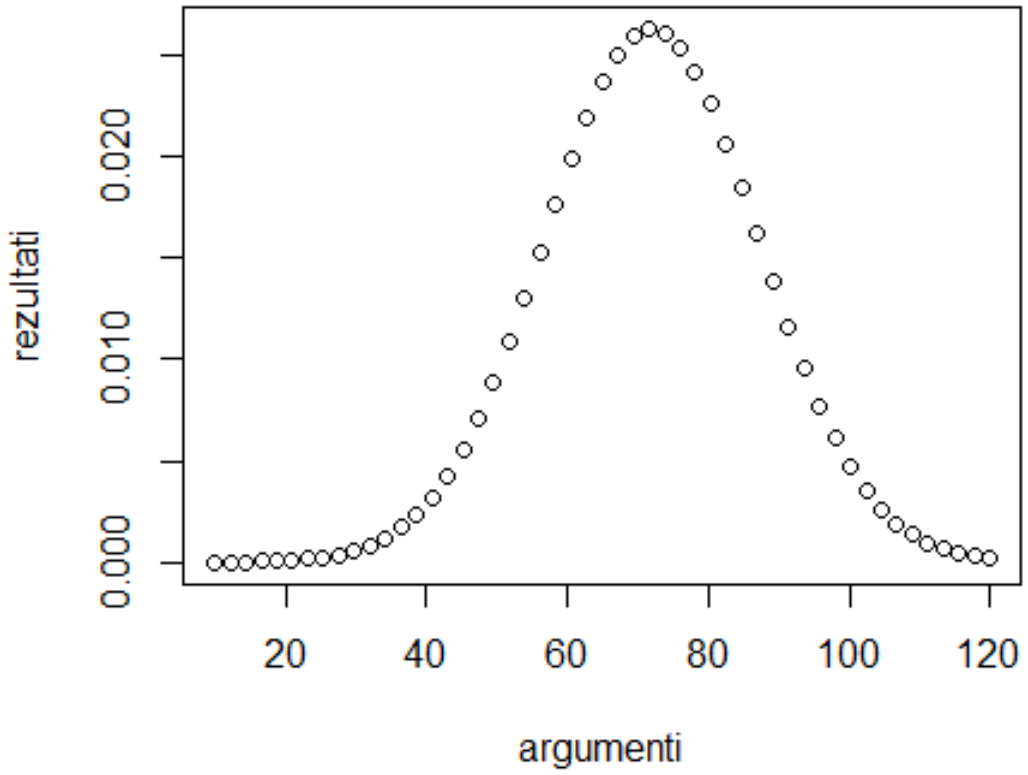
Problem. Prikazati gustinu raspodele za normalnu raspodelu, gde je $\mu = 72$ i $\sigma = 15.2$.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije dnorm dobije vektor rezultata i na kraju se funkcijom plot izvrši iscrtavanje:

```
> mean.parametar <- 72
> sd.parametar <- 15.2
> donja.granica <- 10
> gornja.granica <- 120
> opseg <- gornja.granica - donja.granica
> broj.tacaka <- 50
> argumenti <- seq(donja.granica, gornja.granica, by=opseg/broj.tacaka)
> rezultati <- dnorm( argumenti, mean = mean.parametar, sd = sd.parametar )
> plot(argumenti, rezultati)
```

Normalna raspodela (5)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Normalna raspodela (6)

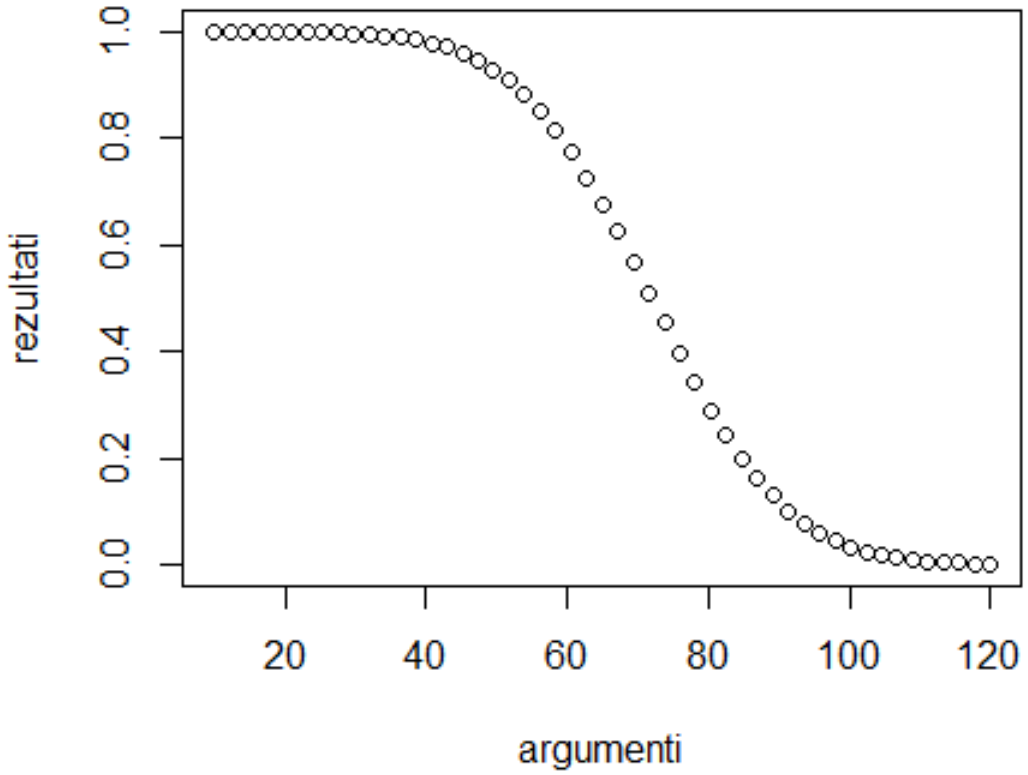
Problem. Prikazati funkciju raspodele za slučajnu promenljivu koja ima normalnu raspodelu sa $\mu = 72$ i $\sigma = 15.2$. Drugim rečima, ako se rezultati testova uklapaju u normalnu raspodelu sa sredinom 72 i sa standardnom devijacijom 15.2, grafički prikazati raspodelu verovatnoća da je student osvojio x ili više bodova.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, potom se funkcijom pnorm dobije vektor rezultata. Na kraju, funkcijom plot se izvrši iscrtavanje:

```
> mean.parametar <- 72
> sd.parametar <- 15.2
> donja.granica <- 10
> gornja.granica <- 120
> opseg <- gornja.granica - donja.granica
> broj.tacaka <- 50
> argumenti <- seq(donja.granica, gornja.granica, by=opseg/broj.tacaka)
> rezultati <- pnorm( argumenti, mean = mean.parametar, sd = sd.parametar, lower.tail=FALSE )
> plot(argumenti, rezultati)
```

Normalna raspodela (7)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Hi-kvadratna raspodela

Hi-kvadratna raspodela je raspodela neprekidne slučajne promenljive.
Ako su $X_1, X_2, \ldots, X_m$ nezavisne slučajne promenljive sa normiranom normalnom raspodelom, tada će slučajna promenljiva $V = {X_1}^2 + {X_2}^2 + \cdots + {X_m}^2$ imati hi-kvadratnu raspodelu sa m stepena slobode.

Zapis da slučajna promenljiva X ima hi-kvadratnu raspodelu sa m stepena slobode je: $\mathrm{V} \sim \boldsymbol{X}^2(m)$.

U sistemu R, za rad sa hi-kvadratnom raspodelom se koriste:

- **dchisq** – funkcija gustine raspodele,
- **pchisq** – funkcija raspodele,
- **qchisq** – određivanje kvantila,
- **rchisq** - generator pseudoslucajnih brojeva.

Kod ovih funkcija, imenovani argument **df** označava broj stepena slobode (u gornjoj formuli, ta veličina je označena sa m).

Hi-kvadratna raspodela (2)

Problem. Odrediti 95-ti percentil hi-kvadratne raspodela sa 7 stepena slobode.

Rešenje. Korišćenjem funkcije za računanje kvantila qchisq, dobija se traženi rezultat:

```
> percentil.95 <- qchisq(
+   0.95, # koji se percentil racuna
+   df = 7 # broj stepena slobode hi-kvadratne raspodele
+   )
> percentil.95
[1] 14.06714
```

Dakle, 95-ti percentil hi-kvadratne raspodela sa 7 stepena slobode je 14.06714.

Hi-kvadratna raspodela (3)

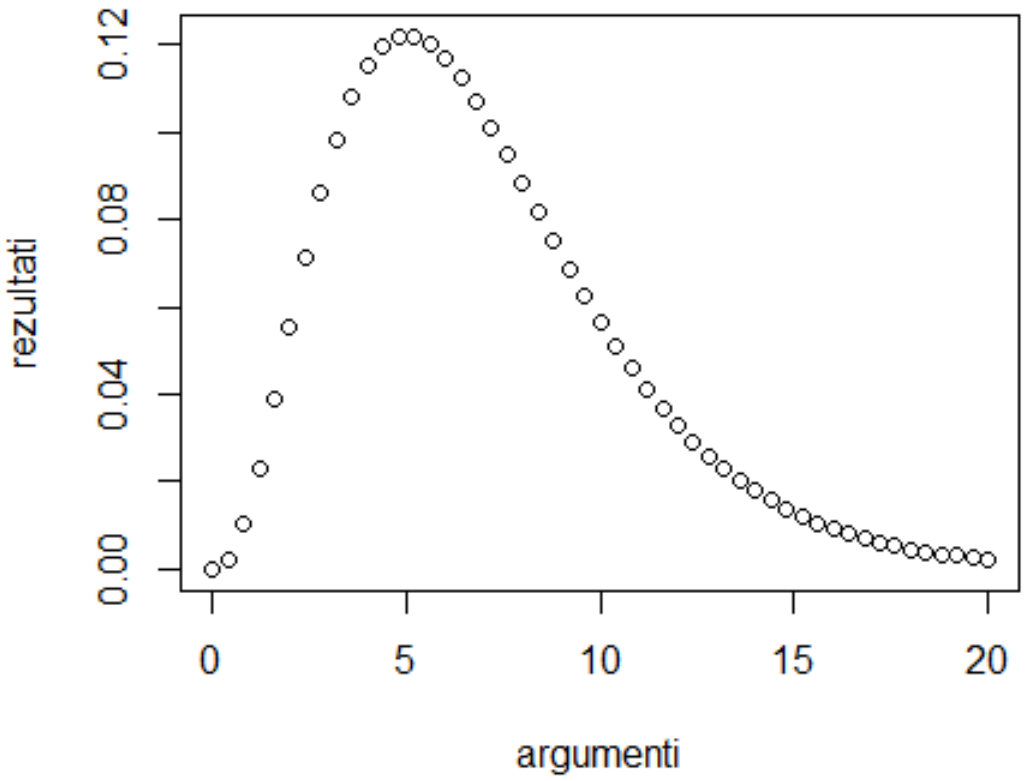
Problem. Prikazati gustinu raspodele za hi-kvadratnu raspodelu sa brojem stepena slobode $m = 7$.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije dchisq dobije vektor rezultata i na kraju se funkcijom plot izvrši iscrtavanje:

```
> broj.stepena.slobode <- 7
> donja.granica <- 0
> gornja.granica <- 20
> opseg <- gornja.granica - donja.granica
> broj.tacaka <- 50
> argumenti <- seq(donja.granica, gornja.granica, by=opseg/broj.tacaka)
> rezultati <- dchisq( argumenti, df = broj.stepena.slobode )
> plot(argumenti, rezultati)
```

Hi-kvadratna raspodela (4)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Studentova raspodela

Studentova raspodela je raspodela neprekidne slučajne promenljive.
Ako su Z i V nezavisne slučajne promenljive takve da Z ima normiranu normalnu raspodelu (tj. $Z \sim \boldsymbol{N}(0,1)$) i da V ima hi-kvadratnu raspodelu sa m stepena slobode, tada će slučajna promenljiva $T = \frac{Z}{\sqrt{V/m}}$ imati Studentovu raspodelu sa m stepena slobode.

Zapis da slučajna promenljiva T ima Studentovu raspodelu sa m stepena slobode je: $\mathrm{T} \sim \boldsymbol{t}(m)$.

U sistemu R, za rad sa Studentovom raspodelom se koriste:

- **dt** – funkcija gustine raspodele,
- **pt** – funkcija raspodele,
- **qt** – određivanje kvantila,
- **rt** - generator pseudoslucajnih brojeva.

Kod ovih funkcija, imenovani argument **df** označava broj stepena slobode (u gornjoj formuli, ta veličina je označena sa m).

Studentova raspodela (2)

Problem. Odrediti 2.5-ti i 97.5-ti percentil Studentove raspodele sa 5 stepena slobode.

Rešenje. Korišćenjem funkcije za računanje kvantila Studentove raspodele qt, dobija se traženi rezultat:

```
> percentili <- qt(c(0.025, 0.975), df = 5)
> percentili
[1] -2.570582  2.570582
```

Dakle, 2.5-ti percentil Studentove raspodele sa 5 stepena slobode je -2.570582, a 97.5-ti percentil je 2.570582.

Studentova raspodela (3)

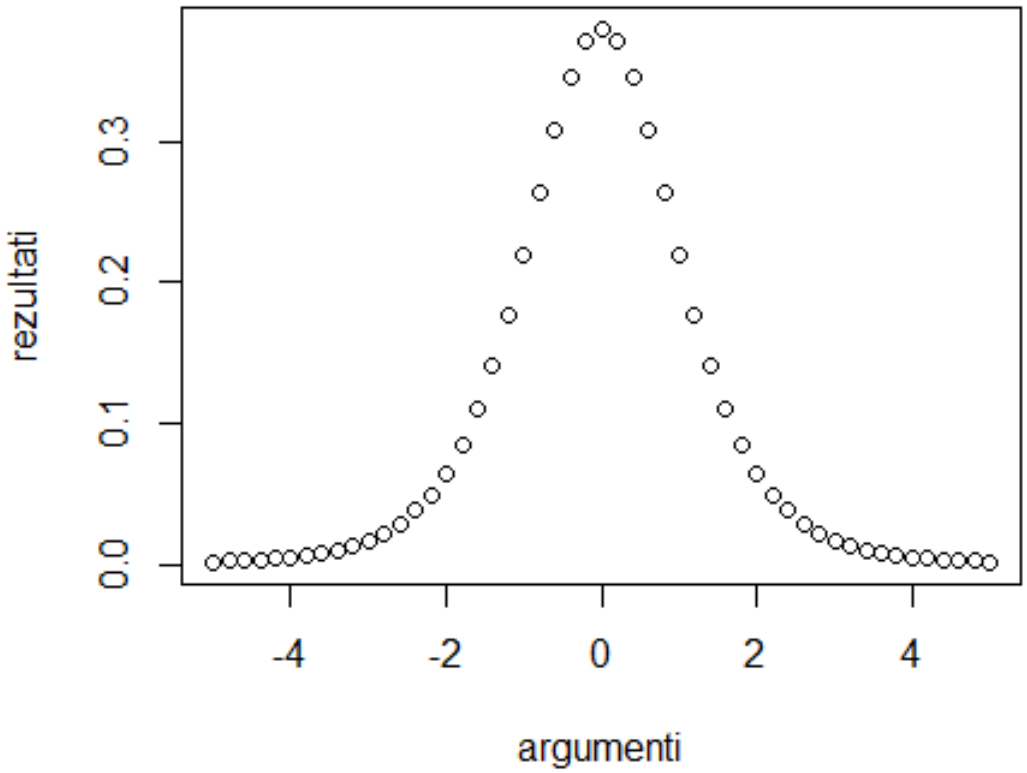
Problem. Prikazati gustinu raspodele za Studentovu raspodelu sa brojem stepena slobode $m = 5$.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije dt dobije vektor rezultata i na kraju se funkcijom plot izvrši iscrtavanje:

```
> broj.stepena.slobode <- 5
> donja.granica <- -5
> gornja.granica <- 5
> opseg <- gornja.granica - donja.granica
> broj.tacaka <- 50
> argumenti <- seq(donja.granica, gornja.granica, by=opseg/broj.tacaka)
> rezultati <- dt( argumenti, df = broj.stepena.slobode )
> plot(argumenti, rezultati)
```

Studentova raspodela (4)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Fišerova raspodela

Fišerova raspodela je raspodela neprekidne slučajne promenljive.
Ako su V_1 i V_2 nezavisne slučajne promenljive takve da V_1 ima hi-kvadratnu raspodelu sa m_1 stepena slobode i V_2 ima hi-kvadratnu raspodelu sa m_2 stepena slobode, tada će slučajna promenljiva $F = \frac{V_1/m_1}{V_2/m_2}$ imati Fišeovu raspodelu sa (m_1, m_2) stepena slobode.

Zapis da slučajna promenljiva F ima Fišerovu raspodelu sa(m_1, m_2)stepena slobode je: $F \sim \boldsymbol{F}(m_1, m_2)$.

U sistemu R, za rad sa Fišerovom raspodelom se koriste:

- **df** – funkcija gustine raspodele,
- **pf** – funkcija raspodele,
- **qf** – određivanje kvantila,
- **rf** - generator pseudoslucajnih brojeva.

Kod ovih funkcija, imenovani argumenti **df1** i **df2** označavaju brojeve stepena slobode (u gornjoj formuli, te veličine su označene sa m_1 i m_2).

Fišerova raspodela (2)

Problem. Odrediti 95-ti percentil Fišerove raspodele sa (5,2) stepena slobode.

Rešenje. Korišćenjem funkcije za računanje kvantila Fišerove raspodele qf, dobija se traženi rezultat:

```
> percentil.95 <- qf(
+   0.95, # koji se percentil racuna
+   df1 = 5, df2 = 2 # stepeni slobode F raspodele
+ )
> percentil.95
[1] 19.29641
```

Dakle, 95-ti percentil Fišerove raspodele sa (5,2) stepena slobode je 19.29641.

Fišerova raspodela (3)

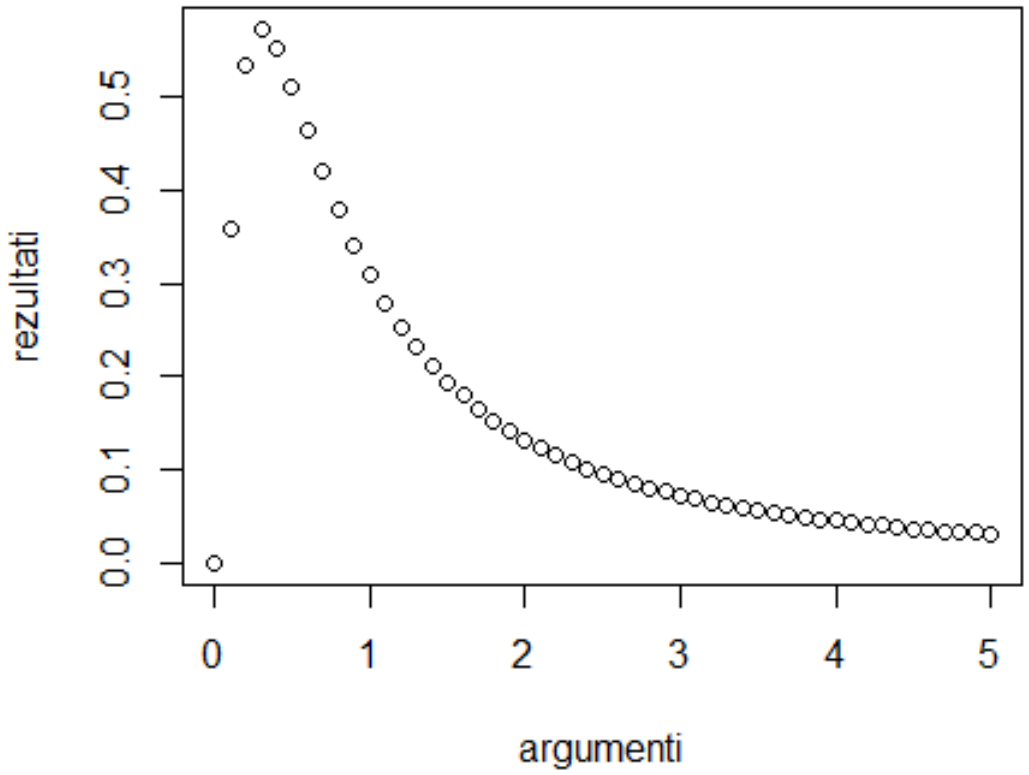
Problem. Prikazati gustinu raspodele za Studentovu raspodelu sa brojem stepena slobode $m = 5$.

Rešenje. Prvo se oformi vektor koji predstavlja argumente, pa se korišćenjem funkcije dt dobije vektor rezultata i na kraju se funkcijom plot izvrši iscrtavanje:

```
> broj.stepena.slobode.1 <- 5
> broj.stepena.slobode.2 <- 2
> donja.granica <- 0
> gornja.granica <- 5
> opseg <- gornja.granica - donja.granica
> broj.tacaka <- 50
> argumenti <- seq(donja.granica, gornja.granica, by=opseg/broj.tacaka)
> rezultati <- df( argumenti, df1 = broj.stepena.slobode.1, df2 = broj.stepena.slobode.2 )
> plot(argumenti, rezultati)
```

Fišerova raspodela (4)

Rešenje (nastavak). Kao rezultat se dobija sledeći dijagram:



Korišćeni izvori

Deo materijala ove prezentacije je preuzet sa sajta

http://www.r-tutor.com/

Deo materijala je preuzet sa sajta

http://www.e-statistika.rs