Vjerojatnosne distribucije i t-test

Table of Contents

# Vjerojatnosne distribucije

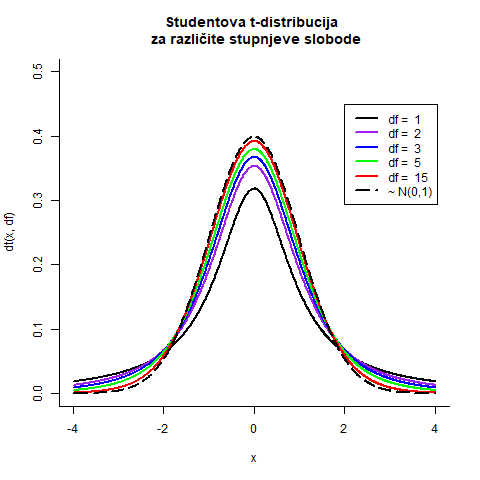
## Studentova t-distribucija ~t(df)

Studentova t-distribucija koristi se za procjenu populacijskih parametara kada je veličina uzorka mala (<30) i/ili kada je varijanca populacije nepoznata. U tom slučaju, računa se t vrijednost (engl. t score) po formuli:

gdje je srednja vrijednost uzorka, µ je srednja vrijednost populacije, s je standardna devijacija uzorka, a n je veličina uzorka. t je slučajna varijabla koja ima studentovu t-distribuciju s (n - 1) stupnjeva slobode.

Svojstva studentove t-distribucije: \* srednja vrijednost jednaka je nuli, \* distribucija je simetrična oko vrijednosti 0, \* varijanca je jednaka df/(df-2), gdje je df broj stupnjeva slobode (engl. degrees of freedom), \* varijnca je uvijek veća od 1. Kako veličina uzorka raste (veliki df), t-distribucija teži standardnoj normalnoj distribuciji. \* ima deblje repove od normalne distribucije, odnosno manje je osjetljiva na ekstremne vrijednosti

knitr::include\_graphics("../Slika/studentova\_t-distribucija.png")

 Na grafikonu je ucrtana i standardna normalna distribucija jer je ona ‘limes’ studentovoj t-distribuciji kako broj stupnjeva slobode raste.

## Uniformna distribucija ~U(a, b)

Slučajna varijabla kojoj svaki mogući ishod ima jednaku vjerojatnost uspjeha zove se uniformna distribucija. Može biti i kontinuirana (npr. nad skupom relanih brojeva [0,10]) i diskretna (npr. bacanje kocke od kojih svaki ishod 1, 2, …, 6 ima jednake šanse).

Funkcija gustoće vjerojatnosti je:

Srednja vrijednost je gdje su a i b početak i kraj intervala nad kojim distribucija ima vjerojatnost veću od 0.

Varijanca je .

## Bernoullijeva distribucija ~Bernoulli(p)

Diskretna slučajna varijabla koja može popirmiti samo 2 ishoda (kao npr. rezultat bacanja novčića (pismo ili glava), odgovor na da-ne pitanje, ishod 0 ili 1) od kojih se jedan događa s vjerojatnošću p, a drugi s vjerojatnošću (1 - p) ima Bernoulijevu distribuciju s parametrtom p.

Bernoulijeva distribucija s parametrom 0.5 je specijalan slučaj uniformne distribucije s dva ishoda.

sredinja vrijednost je p, varijanca je .

## Binomna distribucija ~B(n, p)

Ako zbrojimo n slučajnih varijabli od kojih svaka ima Bernoulijevu distribuciju s parametrom p, takav zbroj je slučajna varijabla s binomnom distribucijom s parametrima n i p.

Koristi se kod brojanja uspijeha u nizu pokusa od kojih svaki pokus ima 2 moguća ishoda: uspijeh i neuspjeh.

Primjeri: \* Novčić se baca 50 puta. Kolika je šansa da se dobije m pisama, gdje je m broj između 0 i 50? \* U zdjeli se nalazi 10 crnih i 10 bijelih loptica. Kolika je šansa da se nakon 5 izvlačenja s vraćanjem loptice dobije m crnih loptica, gdje je m broj između 0 i 5.

Srednja vrijednost je , a varijanca je jednaka .

## Hipergeometrijska distribucija ~Hipergeometrijska(N, K, n)

Ako zamislimo scenarij izvlačenja crnih i bijelih loptica kao što je gore opisano, ali ovaj put bez vraćanja loptice, vjerojatnost uspjeha (odabira crne loptice) mijenja se sa svakim izvlačenjem. Varijabla koja broji uspjehe u takvom eksperimentu ima hipergeometrijsku distribuciju.

Općenito, hipergeometriska distribucija s parametrima N, K, n je diskretna distribucija koja opisuje vjerojatnost da se dogodi k uspjeha u n izvlačenja iz skupa od N članova od kojih je K onih sa traženim svojstvom za uspjeh.

Srednja vrijednost ovakve varijable je , a varijanca je dana izrazom

## Geometrijska distribucija ~Geom(p)

Koliko puta je potrebno napraviti eksperiment da bi se dogodio prvi uspjeh? Ili koliko neuspjeha će se dogoditi prije prvog uspjeha? Varijabla koja mjeri ovakve događaje ima geometrijsku distribuciju.

Još jedna distribucija nastala iz Bernoullijeve, diskretna je i određena parametrom p - vjerojatnošću uspjeha.

Funkcija vjerojatnosti, odnosno vjerojatnost da se dogodi (k-1) neuspjeha prije prvog uspjeha, tj. da je potrebno k pokušaja do prvog uspjeha dana je izrazom:

Srednja vrijednost joj je , a varijanca .

## Negativna binomna distribucija ~ NB(r, p)

Negativna binomna je poopćenje geometrijske distribucije i broji koliko neuspijeha će se dogoditi prije r uspjeha. Pokušaji su nezavisni i identično distribuirani po Bernoulijevoj razdiobi s parametrom p.

Srednja vrijednost je , a varijanca .

## Poissonova distribucija ~Pois(λ)

Poissonovom distribucijom opisuje se vjerojatnost da se neki događaj dogodi n puta unutar određenog vremenskog perioda (ili prostora, volumena, udaljenosti), pri čemu je prosjek događanja po vremenskoj jedinici λ poznat i konstantan. λ je realan, pozitivan broj. Svi događaji su međusobno neovisni, tj. pojava jednog događaja ne ovisi o pojavi događaja prije njega.

Primjer: \* kolika je šansa za 5 poziva u jednoj minuti u službi za korisnike ako prosječno dobivaju 1000 poziva dnevno. \* kolika je vjerojatnost da će restoran imati 50 zahtjeva za večeru u danu, ako u prosjeku posluži 200 večera tjedno.

Poissonova distribucija je diskretna koja može poprimiti samo vrijendosti iz cijelih nenegativnih brojeva.

Parametar λ označava prosječan broj događanja u nekom intervalu i jedini je parametar koji utječe na distribuciju. e je Eulerova broj 2.71828…. S obzirom na to da je baza za prirodno logaritmiranje i potenciranje, u sustavu R dobije se kao exp(1). e^2 je exp(2), e^3 je exp(3), itd. k! je k faktorijela, tj. umnožak prvih k prirodnih brojeva. U sustavu R dobije se naredbom factorial(k).

Srednja vrijednost i varijanca jednaki su λ.

## Eksponencijalna distribucija ~Exp(λ)

Koliko dugo vremena prije nego se dogodi uspjeh? Ako je sekunda prevelika mjera i treba je raščlaniti na infinitezimalne dijeliće, odnosno ako geometrijsku distribuciju pustimo u krajnji limit, dobit ćemo eksponencijalnu distribuciju. Ovo je dakle kontinuirana vjerojatnosna distribucija.

Dok je Poissonova distribucija dala odgovor o broju događaja u vremenskom intervalu, eksponencijalna distribucija daje odgovar na vrijeme čekanje do prvog događaja.

Zapravo, za događaje koji se mogu opisati Poissonovom distribucijom vrijedi da se vrijeme između dva događaja može opisati eksponencijalnom distribucijom. Te dvije distribucije su povezane na takav način.

Ova distribucija je također određena parametrom λ koji opisuje prosječni broj događaja u nekom intervalu i pretpostavlja se da je konstantan za događaje koji se mjere. λ je realan, pozitivan broj (λ > 0).

Srednja vrijednost je ,a varijanca je .

## Ostale distribucije

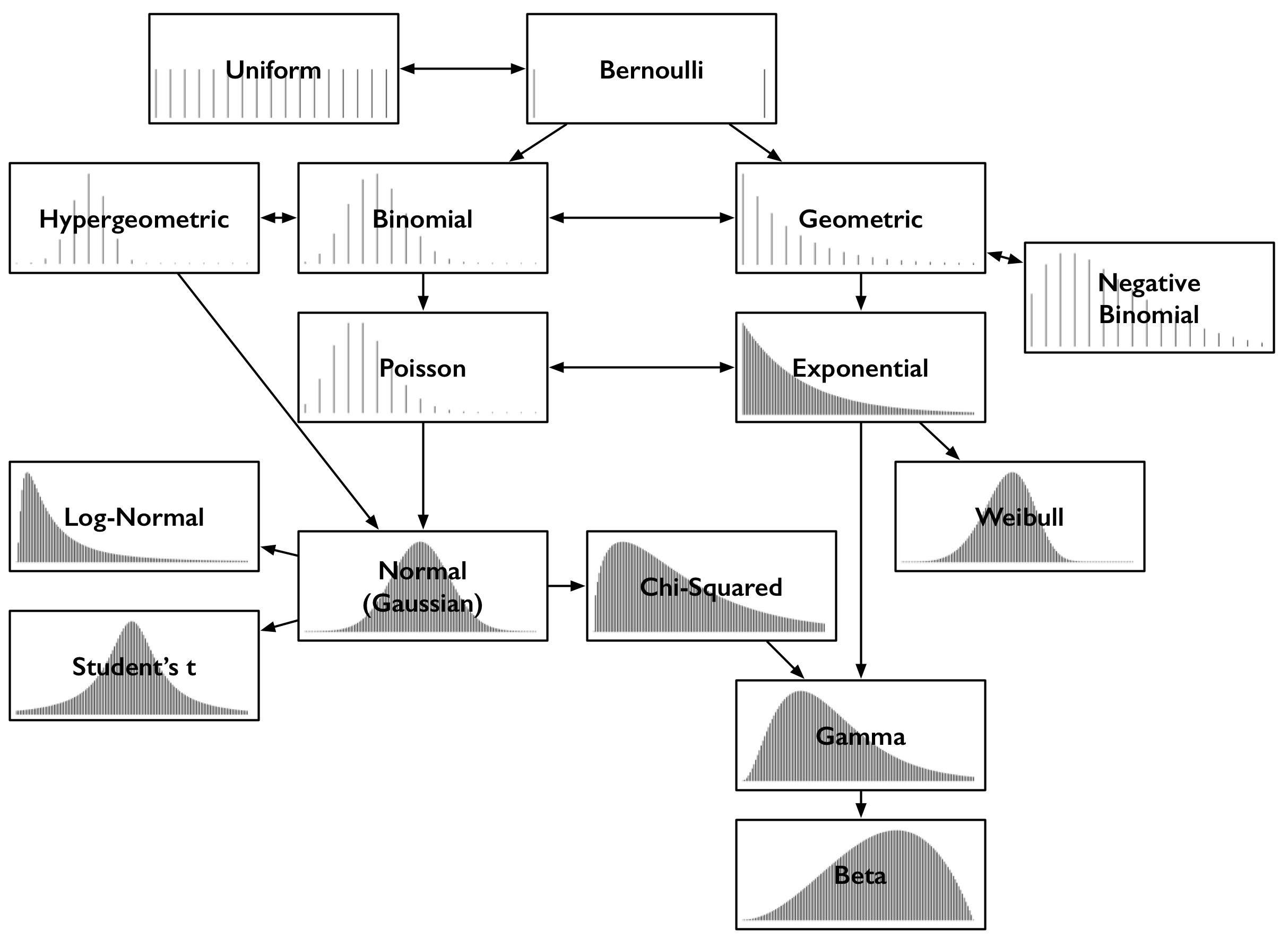
**Hi-kvadratna distribucija** s k stupnjeva slobode () je distribucija sume kvadrata k nezavisnih standardnih, normalnih distribucija te se koristi za testiranje hipoteza o standardnoj devijaciji, kada se želi testirati koliko dobro neki model pristaje opisu veze između varijabli (engl. goodness of fit testing), za konstrukciju intervala pouzdanosti populacijske standardne devijacije, itd.

Srednja vrijednost je k, a varijanca 2k.

**Gama distribucija** je sofisticirana verzija eksponencijalne, te opisuje vrijeme čekanja do n-te pojave nekog događaja (suma n eksponencijalnih distribucija kada je n cijeli broj), ili vrijme trajanja do isteka valjanosti/smrti.

Postoje još mnoge druge koje nećemo spominjati, ali ćemo dati dijagram koji slikovito prikazuje vezu između prethodno opisanih distribucija. Distribucije diskretnih slučajnih varijabli ispunjene su razmaknutim vodoravnim linijama, dok su distribucije kontinuiranih slučajnih varijabli prikazane gusto popunjenim linijama.

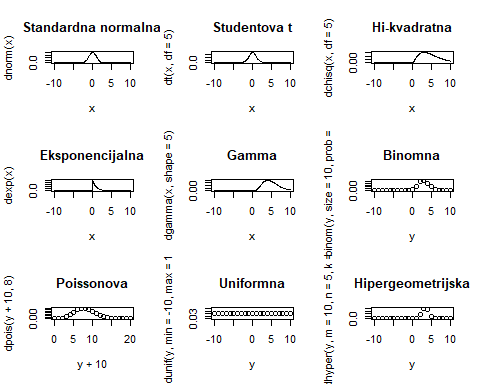
knitr::include\_graphics("../Slika/Distribucije.png")



## Primjeri:

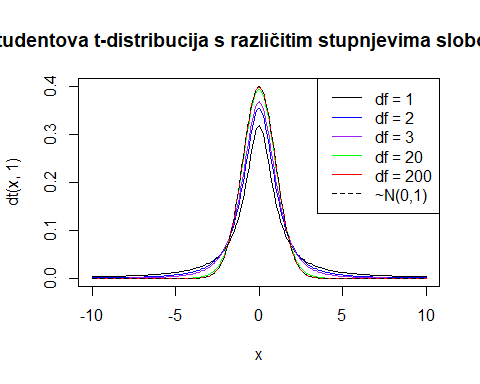
**Primjer 1**  
Grafički prikažite funkcije gustoće nekih od gore opisanih distribucija.

x <- seq(-10,10,0.1) # gusto raspoređen vektor za prikaz kontinuiranih distribucija  
y <- seq(-10,10,1) # cjelobrojni vektor za prikaz diskretnih distribucija  
  
par(mfrow=c(3,3))  
plot(x, dnorm(x), type="l", main="Standardna normalna")  
plot(x, dt(x, df = 5), type="l", main="Studentova t")  
plot(x, dchisq(x, df = 5), type="l", main="Hi-kvadratna")  
plot(x, dexp(x), type="l", main="Eksponencijalna")  
plot(x, dgamma(x, shape = 5), type="l", main="Gamma")  
plot(y, dbinom(y, size = 10, prob = 0.3), main="Binomna")  
plot(y+10, dpois(y+10, 8), main="Poissonova")  
plot(y, dunif(y, min=-10, max=10), main = "Uniformna")  
plot(y, dhyper(y, m=10, n=5, k=5), main = "Hipergeometrijska")



**Primjer 2**  
Prikažimo grafički kako se studentova t-distribucija mijenja ovisno o stupnjevima slobode:

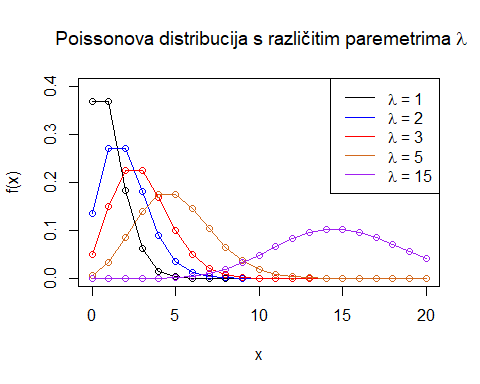
x <- seq(-10,10,0.1)  
  
plot(x, dt(x, 1), type="l", col="black", ylim=c(0, 0.4), main= "Studentova t-distribucija s različitim stupnjevima slobode df")  
lines(x, dt(x, 2), type="l", col="blue")  
lines(x, dt(x, 3), type="l", col="purple")  
lines(x, dt(x, 20), type="l", col="green")  
lines(x, dt(x, 200), type="l", col="red")  
  
lines(x, dnorm(x), col="black", lty=2) #standardna normalna distribucija  
  
legend("topright", legend=c("df = 1", "df = 2", "df = 3", "df = 20", "df = 200", "~N(0,1)"), col=c("black", "blue", "purple", "green", "red", "black"), lty = c(rep(1, 5), 2))



Dodavanjem standardne normalde distribucije, može se vidjeti kako studentova t-distribucija teži standardnoj normalnoj kako se povećava broj stupnjeva slobode, odnosno kako veličina uzorka iz populacije raste. Već za uzorke preko 20 primjeraka, distribucija je jako slična standardnoj normalnoj.

**Primjer 3**  
Prikažimo grafički kako se poissonova distribucija mijenja ovisno o parametru λ:

x <- seq(0,20,1)  
  
plot(x, dpois(x, 1), type="o", col="black", ylim=c(0, 0.4), main=" Poissonova distribucija s različitim paremetrima"~lambda, ylab = "f(x)")  
lines(x, dpois(x, 2), type="o", col="blue")  
lines(x, dpois(x, 3), type="o", col="red")  
lines(x, dpois(x, 5), type="o", col="chocolate")  
lines(x, dpois(x, 15), type="o", col="purple")  
legend("topright", legend=c(expression(paste(lambda, " = ", 1)), expression(paste(lambda, " = ", 2)), expression(paste(lambda, " = ", 3)), expression(paste(lambda, " = ", 5)), expression(paste(lambda, " = ", 15))), col=c("black", "blue", "red", "chocolate", "purple"), lty = 1)



**Primjer 4**  
Mladi astronom ide na promatranje metora na noć kada je najavljeno da će ih padati prosječno 15 po satu. 1. Kolika je vjerojatnost da će astronom vidjeti baš 15 meteora u sat vremena.

Ovdje se radi o Poissonovj distribuciji. Zadan je parametar λ = 15 koji označava prosjek meteora po satu. Vjerojatnost da će astronom vidjeti točno 10 meteora je vrijednost funkcije vjerojatnosti f(x) u točki 10, drugim riječima tražimo f(10). U sustavu R to odgovara funkciji d, točnije dpois():

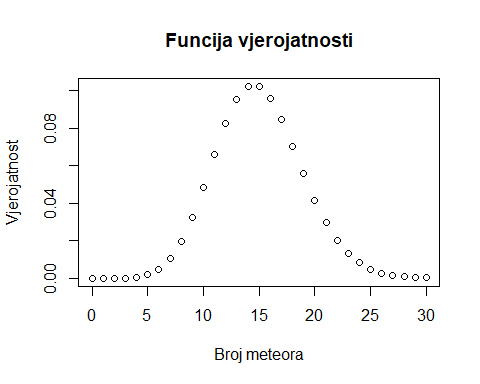
dpois(15, lambda = 15)

## [1] 0.1024359

Vjerojatnost da ćemo vidjeti točno 15 meteora je samo 10.2%. Iznenađujuće?

1. Grafički prikažite funckiju vjerojatnosti (f(x)) ovih slučajnih događaja. Raspon vrijednosti na x osi neka bude od 0 do 30.

plot(0:30, dpois(0:30, 15), ylab="Vjerojatnost", xlab="Broj meteora", main = "Funcija vjerojatnosti")



1. Kolika je vjerojatnost da ćemo vidjeti maximalno 10 meteora u sat vremena:

Tu se traži zbroj vjerojatnosti da ćemo vidjeti jedan meteor, ili dva meteora ili … ili deset meteora, odnosno traži se zbroj f(1)+f(2)+…f(10). Taj zbroj je zapravo kumulativna funkcija gustoće F(x) u točki 10. U sustavu R to je funkcija ppois()

ppois(10, 15)

## [1] 0.1184644

Šansa da ćemo vidjeti maksimalno 10 meteora u sat vremena je samo 11.8%.

1. Kolika je šansa da ćemo vidjeti više od 10 meteora u sat vremena?

To možemo izračunati na 2 načina:

# 1. način: 1 - P(X <= 10)  
1-ppois(10,15)

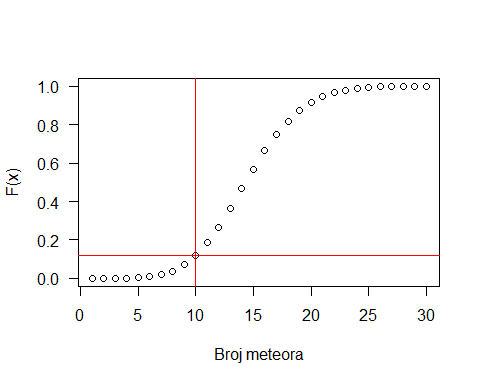
## [1] 0.8815356

# 2.način: P(X>10)   
ppois(10, 15, lower.tail=F)

## [1] 0.8815356

1. Funkcijom plot nacrtajte kumulativnu funkciju distribucije ove vjerojatnosti i prikažite vrijednost u točki 10.

plot(1:30, ppois(1:30, 15), xlab="Broj meteora", ylab="F(x)", las=1)  
abline(v=10, col="red")  
abline(h=ppois(10,15), col=2)



1. Neka se po satu očekuje 15 meteora, odnosno 1 meteor svake 4 minute. Neka je astronom došao na promatranje meteora u nasumičnom trenutku kada nebom nije proletio niti jedan meteor.
2. Izračunajte vjerojatnost da će astronom do prvog meteora čekati duže od 2 minute?

Rješenje:  
Prisjetimo se da je Vrijeme čekanja između dvaju događaja Poissnovog procesa distribuirano po eksponencijalnoj razdiobi, te primjenimo potrebnu funkciju uz odgovarajuće parametre.

Moramo voditi računa da je sve u istim jedinicama. Ako je vrijeme čekanja zadano u minutama, onda i lambda mora biti definirana kao mjera po minuti:

λ = 15 meteora po satu = 15 meteora u 60 minuta = 1/4 meteora po minuti (λ ne mora biti cijeli broj)

1- pexp(2, 1/4)

## [1] 0.6065307

#ili  
pexp(2, 1/4, lower.tail = FALSE)

## [1] 0.6065307

Inutitivno bi možda pomislili da je šansa 50%, no vjerojatnost eksponencijalno opada kada vrijeme čekanja raste.

1. Kolika je vjerojatnost da će astronom do prvog meteora čekati manje od 4 minute?

pexp(4, 1/4)

## [1] 0.6321206

#### Zadatak

Pizzerija primi u prosjeku 20 narudžbi po satu. a) Izračunaj vjerojatnost da će u 10 minuta primiti točno 2 narudžbe. b) Kolika je vjerojatnost da će u sat vremena primiti do 20 narudžbi? c) Kolika je vjerojatost da će u 30 min primiti više od 10 narudžbi?

# T-test za testiranje hipoteza

t-test se koristi za usporedbu dviju srednjih vrijednosti uzoraka za koje se pretpostavlja da su normalno distribuirani, ali se koristi kada nije poznata ni srednja vrijednost ni standardna devijacija populacije, te kada je veličina uzorka mala (<30). Nastala je u tvornici piva Guiness za potrebe analize ječma gdje se raspolagalo na uzorcima od samo 3 podatka u prosjeku.

Testna statstike ima studentovu t-distribuciju, a postoje 3 verzije testa:

* t-test za jedan uzorak - uspoređuje se sredina uzorka s nekom prepostavljenom ili poznatom sredinom,
* t-test za dva nezavisna uzorka - uspoređuju se srednje vrijednosti dvaju nezavisnih uzoraka,
* upareni t-test - za dva zavisna uzorka, npr. kada se uspoređuju srednje vrijednosti iste grupe u različitim trenutcima.

Uvjet za t-test \* uzorak/uzorci su normalno distribuirani, \* varijance uzoraka približno su jednake.

Formule za testne statistike za svaku od ove tri vrste testa mogu se pronaći npr. na poveznici [!link](https://en.wikipedia.org/wiki/Student%27s_t-test), a ovdje ćemo navesti kako se koriste u R-u.

**t-test za jedan uzorak:**

set.seed(4)  
X <- rnorm(10, 10, 3)  
  
# H0: mean(X) = 10  
# H1: mean(X) ≠ 10   
  
t.test(X, mu=10)

##   
## One Sample t-test  
##   
## data: X  
## t = 1.7105, df = 9, p-value = 0.1213  
## alternative hypothesis: true mean is not equal to 10  
## 95 percent confidence interval:  
## 9.451894 13.947281  
## sample estimates:  
## mean of x   
## 11.69959

Odluka o odbacivanju ili ne odbacivanju nul-hipoteze donosi se na temelju p-vrijednosti. Ako je veća od razine značajnosti α (po standardnim postavkama je 0.05), onda ne odbacujemo nul-hipotez, u protivnom je odbacujemo.

U ovom slučaju, p-vrijednost je veća od razine značajnosti α = 0.05, te nemamo dovoljno dokaza da odbacimo H0.

**t.test za dva nezavisna uzorka**

U ovom slučaju uspoređujemo sredine dvaju nezavisnih uzoraka. Funckji t.test() potrebno je proslijediti oba uzorka (redoslijed nije svejedan), te navesti smjer alternativne hipoteze.

set.seed(1)  
X <- rnorm(10, 10, 3)  
Y <- rnorm(10, 13, 3)  
  
#H0: mean(X) - mean(Y) = 0  
#H1: mean(X) - mean(Y) < 0  
  
t.test(X, Y, alternative = "less")

##   
## Welch Two Sample t-test  
##   
## data: X and Y  
## t = -2.6669, df = 16.469, p-value = 0.008289  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -1.160696  
## sample estimates:  
## mean of x mean of y   
## 10.39661 13.74653

U ovom primjeru, p-vrijednost je manja od nivoa značajnosti α = 0.05, te odbacujemo nul-hipotezu.

**upareni t.test**

Da bi dobili upareni t-test za 2 zavisna uzorka, funkciji t.test() potrebno je proslijediti 2 zavisna uzorka i navesti argumetnt “paired = TRUE”.

set.seed(4)  
X1 <- rnorm(10, 10, 3)  
X2 <- rnorm(10, 7, 3)  
  
#H0: mean(X1) - mean(X2) = 0  
#H1: mean(X1) - mean(X2) > 0  
  
t.test(X1, X2, alternative="greater", paired=TRUE)

##   
## Paired t-test  
##   
## data: X1 and X2  
## t = 3.2013, df = 9, p-value = 0.005404  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 1.769967 Inf  
## sample estimates:  
## mean of the differences   
## 4.141398

**Primjer:**  
Farmaceutska kompanija testira efikasnost novog lijeka. Grupa od 6 pacijenata primila je terapiju s novim lijekom dok je jednako velika kontrolna grupa primila tzv. placebo. Mjereno je vrijeme reakcije pacijenata. Postoji li razlika u reakcijama dviju grupa?

Placebo <- c(91, 87, 99, 77, 88, 91)  
Lijek <- c(101, 110, 103, 93, 99, 104)  
  
#H0: mean(Placebo) - mean(Lijek) = 0  
#H1: mean(Placebo) - mean(Lijek) ≠ 0  
  
var(Placebo)

## [1] 51.36667

var(Lijek)

## [1] 31.86667

t.test(Placebo, Lijek, var.equal = F)

##   
## Welch Two Sample t-test  
##   
## data: Placebo and Lijek  
## t = -3.4456, df = 9.4797, p-value = 0.006782  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -21.194292 -4.472375  
## sample estimates:  
## mean of x mean of y   
## 88.83333 101.66667

**Zadatak:**  
Provedeno je istraživanje o kontroliranju broja zdravstvenih pritužbi starijih pacijenata smještenih u staračkom domu. Nasumično je odabrano 30 pacijenata za sudjelovanje u istraživanju. Polovici pacijenata dana je biljka za koju su se sami morali brinuti, a drugoj polovici dana je biljka za koju se brinulo osoblje doma. Tjekom idućih sedam dana bilježio se broj zdravstvenih pritužbi svakog pacijenta koje su dane u vektorima Brinu\_sami i Ne\_brinu. Pomoću t-testa odredite da li briga o biljci utječe na broj zdravstvenih pritužbi pacijenata.

# Izvori:

* Owen, Sean. “Common Probability Distributions: The Data Scientist’s Crib”, 06.018, (Sheethttps://[medium.com/@srowen/common-probability-distributions-347e6b945ce4](mailto:medium.com/@srowen/common-probability-distributions-347e6b945ce4)) 23.10.2019.
* Koehrsen, Will. “The Poisson Distribution and Poisson Process Explained”, 01.2019, (<https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459>) 24.10.2019.