Primjena diskretne slučajne varijable u teoriji kodiranja

Sanja Tipurić-Spužević[1], Marina Jolić[2] i Ante Maleš[3]

Sažetak: U ovom je radu diskretna slučajna varijabla primijenjena na rješavanje problema u informatici. Pokazano je kako se u takvim problemima za izračun vjerojatnosti mogu koristiti programski paketi *GeoGebra* i *Excel*.

Uvod

Slučajna varijabla numerički je ishod slučajnog pokusa. Veliki broj pojava u prirodi, kao i mnogi proizvodni procesi mogu se matematički opisati pojmom slučajne varijable. Teorijska znanja iz ovog područja omogućuju primjene u modeliranju širokog spektra realnih problema, a u ovom će članku biti naglasak na primjenama iz područja kodiranja. Slučajne varijable mogu biti diskretne, kao što je broj ispisanih pogrešaka u knjizi ili broj prometnih nesreća u Republici Hrvatskoj, ili kontinuirane kao što su visina osobe, koncentracija i slično.

U ovom radu želimo pokazati primjenu diskretne slučajne varijable u teoriji kodiranja i pokazati kako se za izračune mogu koristiti programski paketi *Excel* i *GeoGebra*, što olakšava primjenu računa vjerojatnosti svima onima koji nisu primarno matematičari.

Pojam slučajne varijable i njezine razdiobe

Slučajna varijabla numerički je ishod slučajnog pokusa. Neka je S skup svih elementarnih događaja ω_k (ishoda) u nekom pokusu. Elementarni događaj je pojedini ishod nekog pokusa. Svakom događaju ω_k na odgovarajući način pridružimo realan broj x_k.

Definicija 1. Slučajna varijabla X je funkcija $X : S \to \mathbb{R}$, gdje je S skup ishoda u nekom pokusu.

Skup vrijednosti slučajne varijable X označavat ćemo s $R(X)$.

[1]Sanja Tipurić-Spužević, Kemijsko-tehnološki fakultet, Sveučilište u Splitu
[2]Marina Jolić, studentica Fakulteta prirodoslovno-matematičkih i odgojnih znanosti, Sveučilište u Mostaru
[3]Ante Maleš, student Fakulteta prirodoslovno-matematičkih i odgojnih znanosti, Sveučilište u Mostaru

Diskretna slučajna varijabla

Definicija 2. *Diskretna slučajna varijabla je slučajna varijabla kojoj je skup vrijednosti konačan ili beskonačan prebrojiv.*

Drugim riječima, slučajna je varijabla diskretna ako je $R(X)=\left\{x_1,x_2,...,x_n\right\}$ ili $R(X)=\left\{x_1,x_2,...\right\}$.

Događaji $\left\{X=x_k\right\}$ čine potpun skup događaja. To znači da se oni međusobno isključuju i da je suma njihovih vjerojatnosti jednaka 1:

$$\sum P(X=x_k)=1 \,.$$

Razdioba vjerojatnosti slučajne diskretne varijable X može se zapisati kao

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

ako je skup vrijednosti konačan, odnosno kao

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$$

ako je skup vrijednosti beskonačan prebrojiv, pri čemu se s p_k označuje $P(X = x_k)$.

Vrlo korisni pojmovi u području informacijskih sustava pojmovi su koje vežemo uz slučajnu varijablu kao što su razdioba slučajne varijable, očekivanje slučajne varijable i varijanca.

Definicija 3. *Funkcija razdiobe slučajne varijable X funkcija je koja daje vjerojatnost da će slučajna varijabla X poprimiti vrijednost jednaku ili manju od nekog realnog broja x_k:*

$$F(x_k)=P(X\le x_k)=\sum_{i=1}^{k} p_i \,.$$

Definicija 4. *Očekivana vrijednost slučajne varijable zbroj je umnožaka vrijednosti varijable X i njoj pripadajućih odgovarajućih vjerojatnosti p_k:*

$$E(X)=\sum_{k=1}^{\infty} x_k p_k=\mu \,.$$

Definicija 5. *Varijanca slučajne varijable očekivanje je kvadratnog odstupanja vrijednosti varijable X od njezinog očekivanja:*

$$V(X)=E\left[(X-\mu)^2\right]=E(X^2)-\left[E(X)^2\right]=\sum_{k=1}^{\infty} x_k^2 p_k-\mu^2.$$

Binomna razdioba

Ako slučajni pokus koji se ponavlja n puta ima dva moguća ishoda, *uspjeh i neuspjeh,* i ako je vjerojatnost nastupa ishoda *uspjeh* poznata, unaprijed utvrđena i konstantna tijekom cijelog istraživanja i iznosi p, kaže se da se diskretna slučajna varijabla X ravna prema binomnoj razdiobi.

Vjerojatnost da se neki slučajni događaj X u n pokusa realizira k puta prema zakonu razdiobe binomne distribucije jednaka je:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}=\binom{n}{k}p^kq^{n-k},$$

gdje je n broj ponavljanja pokusa, p vjerojatnost realizacije povoljnog ishoda, q vjerojatnost da se slučajni događaj ne realizira, te k broj povoljnih ishoda n ponavljanja slučajnog pokusa.

Očekivanje slučajne varijable X kod binomne razdiobe jednako je $E(X) = np$. Varijanca slučajne varijable X kod binomne razdiobe jednaka je $\sigma^2=npq$. Općenito se kaže da se slučajna varijabla X ravna po binomnoj razdiobi koja je određena parametrima n i p *i* zapisuje se $X\sim B(n,p)$.

Napomena. Primjer koji slijedi sastavili su studenti preddiplomskog studija informatike u okviru svojih završnih radova.

Primjer 1.

Koristeći EBCDI kod od 8 bitova, mogućnost pogreške na 1 bitu na izabranih 40 kodova je 2 koda.

Kolika je vjerojatnost da se u 100 kodova

a) ne pojavi nijedan pogrešan kod,

b) pojave najviše 2 pogrešna koda,

c) pojavi barem 5 pogrešnih kodova?

Koliki je očekivani broj pogrešnih kodova na 200 kodova?

Napomena. U današnjim tehnologijama kodiranje prevodi niz znakova, slova, brojeva, simbola itd. u digitalni oblik (format) koji se koristi za prijenos ili pohranu podataka. Kodiranje podrazumijeva upotrebu kodova pri promjeni originalnih podataka u neki drugi format. Ako kodiranjem želimo prikazati sve znamenke dekadskog brojevnog sustava, potrebno je koristiti kombinacije od 4 bita koje se sastoje od 0 ili 1. S 4 bita možemo napraviti $2^4=16$ mogućih kombinacija.

Alfanumerički kodovi su kodovi koji omogućuju kodiranje slova i znakova. Najpoznatiji je i najviše u upotrebi ASCII kod od 7 bitova, što daje 128 kombinacija, i EBCDIC kod koji sadrži 8 bitova i 256 mogućih kodnih kombinacija. Pod pogrešnim kodom podrazumijevamo kod kod kojeg je došlo do pogreške na samo jednom bitu.

Rješenje. Zadatak ćemo riješiti na tri načina – prvo numerički, a potom pomoću programa *GeoGebra* i *Excel.*

Numeričko rješavanje. Vjerojatnost pogreške na jednom kodu iznosi $p=\frac{2}{40}=0.05$ Vjerojatnost da je kod točan je $q=1-0.05=0.95$. Sada redom imamo:

a) $P(X=0)=\binom{100}{0}0.05^0\cdot 0.95^{100}=0.0059$.

b) $P(X\le 2)=P(X=0)+P(X=1)+P(X=2)$

$$=\binom{100}{0}0.05^0\cdot 0.95^{100}+\binom{100}{1}0.05^1\cdot 0.95^{99}+\binom{100}{2}0.05^2\cdot 0.95^{98}$$

$$=0.0059+0.0312+0.0812=0.1183.$$

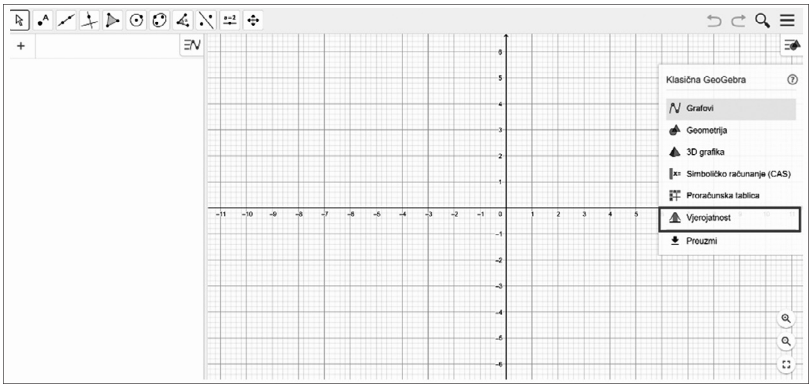
c) $P(X\ge 5)=1-P(X<5)=1-(P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4))$

$$=1-(0.1183+\binom{100}{3}0.05^3\cdot 0.95^{97}+\binom{100}{4}0.05^4\cdot 0.95^{96})$$

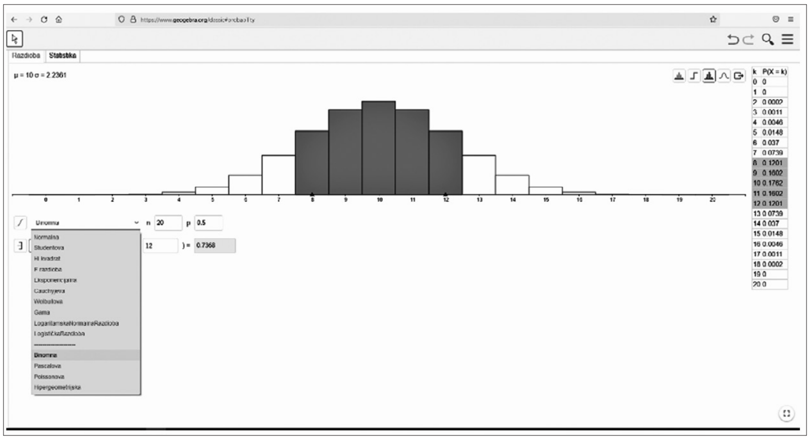
$$=1-(0.1183+0.1396+0.1781)=0.564.$$

d) $E(X)=200\cdot 0.05=10$.

Primjena programskog paketa Geogebra. Primjer ćemo sada riješiti pomoću matematičkog alata *GeoGebra*. *GeoGebra* je program koji je moguće koristiti bez prethodne instalacije, direktno iz preglednika. Odabirom opcije *Klasična GeoGebra* otvara se sučelje koje je prikazano na Slici 1. U izborniku treba s desne strane kliknuti



Slika 1.

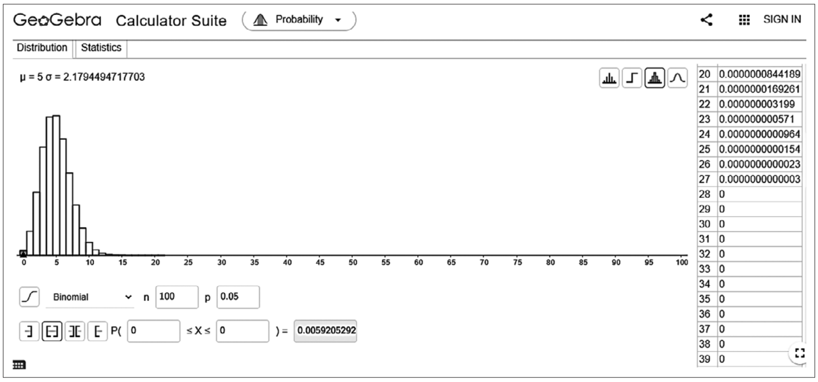


Slika 2.

na ***Vjerojatnost,*** a nakon otvaranja novog prozora u padajućem je izborniku potrebno odabrati opciju ***Binomna***, što je vidljivo na Slici 2.

Na Slici 3. vidimo da najprije treba unijeti vrijednost za broj uzoraka n i vrijednost za p koji predstavlja vjerojatnost da će se neki događaj odviti. U našem je primjeru $n = 100$ i $p = 0.05$.

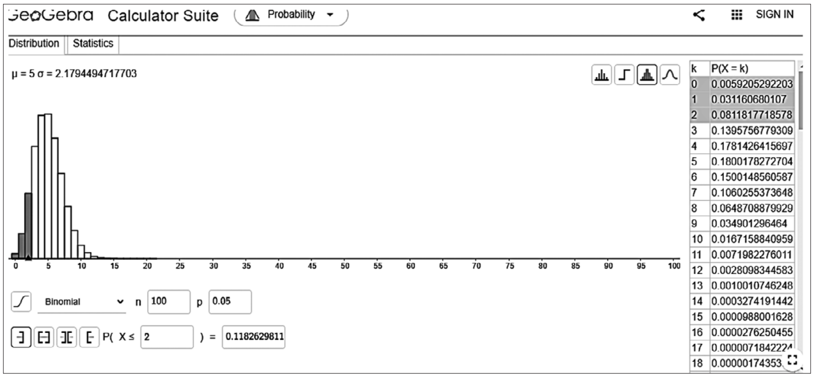
S obzirom na to da je u zadatku pod a) pitanje kolika je vjerojatnost da se u 100 kodova ne pojavi niti jedan pogrešan kod, prvi će red ostati nepromijenjen, a drugom je redu potrebno izabrati ***Interval*** (Slika 3.), a potom unijeti vrijednost 0 slijeva i zdesna. Dobivena vjerojatnost iznosi 0.0059.



Slika 3.

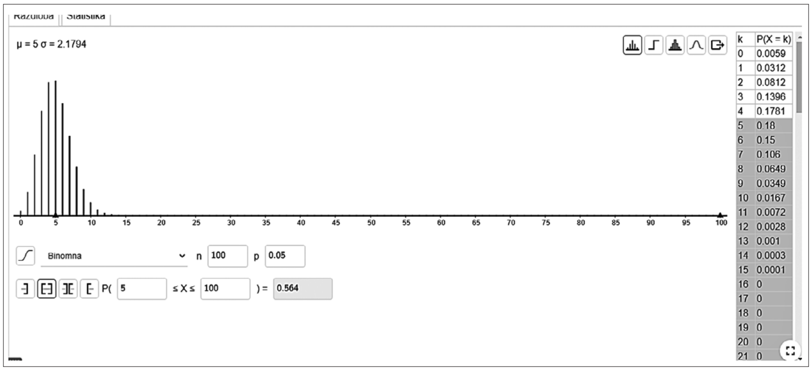
U zadatku pod b) pitanje je kolika je vjerojatnost da se u 100 kodova pojave najviše 2 pogrešna koda, stoga je potrebno izabrati ***Jednostrani lijevi interval***, a potom unijeti vrijednost 2 (Slika 4.).

Dobivena vjerojatnost iznosi 0.1183 i jednaka je zbroju vjerojatnosti da se neće pojaviti nijedan pogrešan kod, da će se pojaviti 1 i da će se pojaviti 2 pogrešna koda. Pojedinačne vjerojatnosti za svaki točno definirani broj uzoraka prikazane su u tablici s desne strane (Slika 4.).



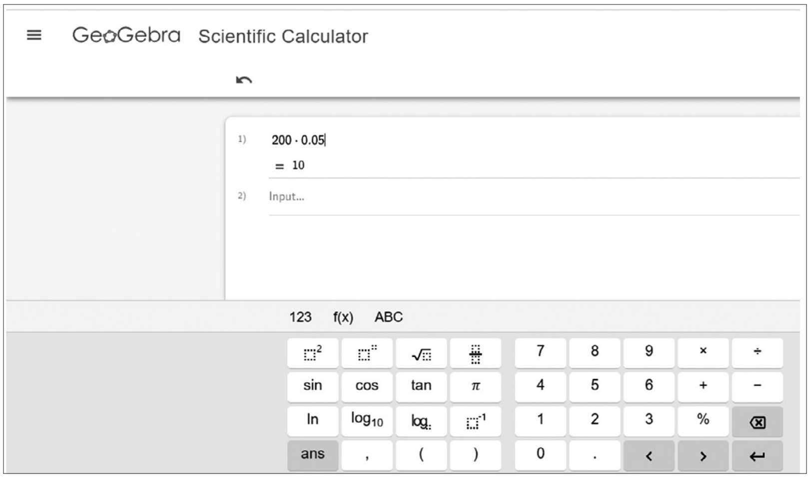
Slika 4.

Budući da je u zadatku pod c) pitanje kolika je vjerojatnost da se u 100 kodova pojavi barem 5 pogrešnih kodova prvi će red ostati nepromijenjen, u drugom je redu potrebno izabrati ***Interval*** (Slika 4.), a potom unijeti vrijednost 5 slijeva i vrijednost 100 zdesna. Dobivena vrijednosti iznosi 0.564. Pojedinačne vjerojatnosti za svaki točno definirani broj uzoraka prikazane su u tablici s desne strane (Slika 5.)



Slika 5.

U zadatku pod d), gdje treba izračunati očekivani broj pogrešnih kodova na 200 kodova, radi se samo o množenju pa se možemo poslužiti ponuđenim kalkulatorom u *GeoGebri* (Slika 6.).



Slika 6.

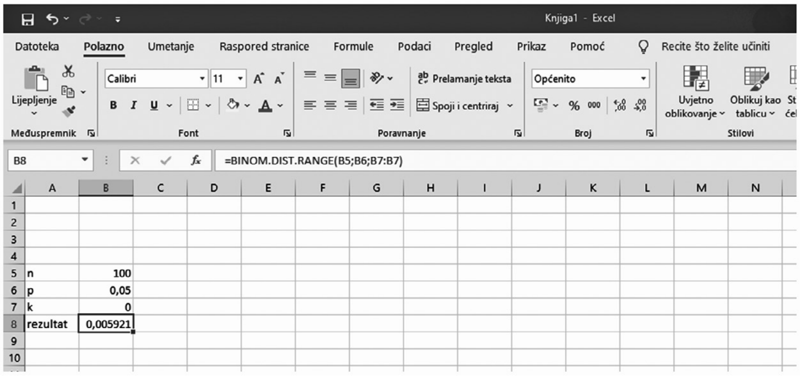
Dobivena je vrijednost 10.

Primjena programskog paketa Excel. Za rješavanje u *Excelu* potrebno je koristiti sljedeće parametre i funkcije:

- n – broj ponavljanja pokusa
- p – vjerojatnost realizacije povoljnog ishoda
- k – broj nezavisnih pokušaja
- funkciju BINOM.DIST (pomoću koje izračunavamo vjerojatnost najviše k uspjeha ili točno k uspjeha) ili BINOM.DIST.RANGE (pomoću koje izračunavamo vjerojatnost da ćemo imati najmanje k uspjeha)

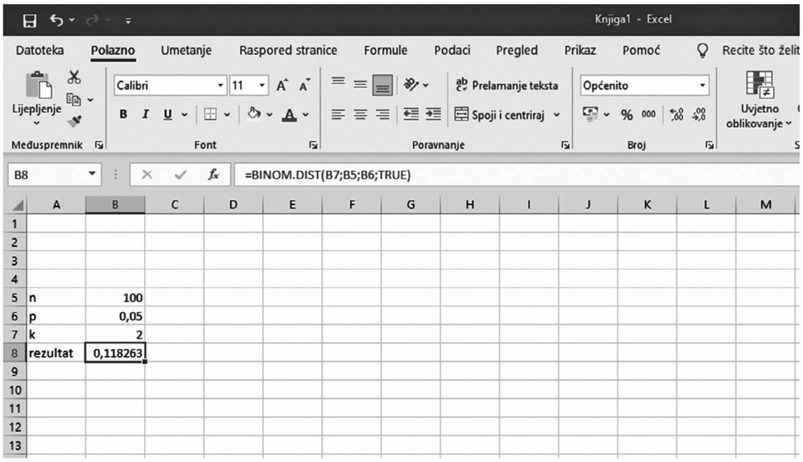
- ako biramo funkciju BINOM.DIST, moramo izabrati i logičku vrijednost koja određuje oblik funkcije: opciju BINOM.DIST: TRUE biramo ako želimo izračunati vjerojatnost da imamo najviše *k* uspjeha, a opciju BINOM.DIST: FALSE ako želimo izračunati vjerojatnost da imamo točno *k* uspjeha.

Pri rješavanju zadatka pod a) potrebno je unijeti funkciju BINOM.DIST.RANGE(0;100;0,05;0;0), (Slika 7.). Dobivena je vjerojatnost 0.0059.



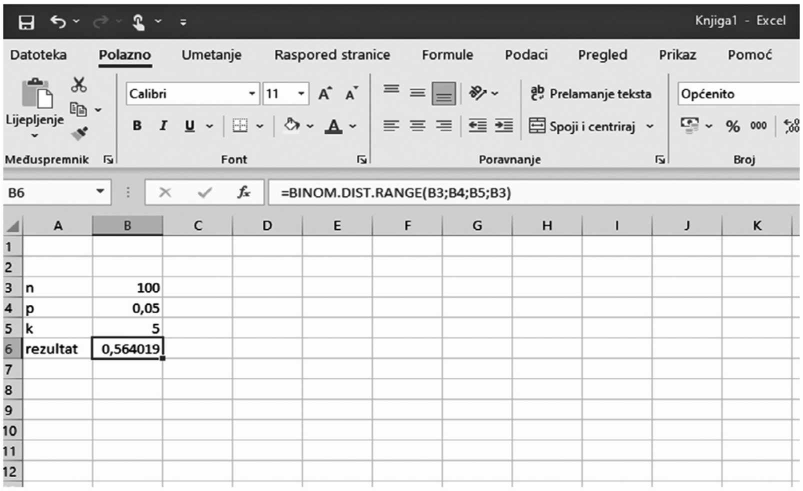
Slika 7.

Pod b) je potrebno koristiti funkciju BINOM.DIST(2;20;0,05;TRUE) pomoću koje računamo vjerojatnost da se u 100 kodova pojave najviše 2 pogrešna koda (Slika 8.). Dobivena je vjerojatnost 0.1183.



Slika 8.

Pri rješavanju zadatka pod c) potrebno je koristiti funkciju BINOM.DIST.RANGE (100;0,05;5;100) pomoću koje računamo vjerojatnost da se u 100 kodova pojavi barem 5 pogrešnih kodova (Slika 9.). Dobivena je vjerojatnost 0.564.



Slika 9.

Poissonova razdioba

Ako je vjerojatnost da se dogodi neki događaj poznata, unaprijed utvrđena, konstantna i jako mala tijekom cijelog istraživanja te broj pokusa teži u beskonačnost, umjesto binomne razdiobe, koristi se Poissonova razdioba. Uglavnom se upotrebljeva za rijetke događaje, odnosno one događaje koji imaju veliki uzorak i malu vjerojatnost. Važna primjena Poissonove razdiobe je u slučajevima kada ne znamo broj pokusa ni vjerojatnost realizacije slučajnog događaja, ali znamo njihov umnožak odnosno očekivanje slučajne varijable (vidi nastavak). Poissonova distribucija koristi se ako vrijedi $n \geq 50$ i $p \leq 0.1$.

Prema zakonu razdiobe Poissonove distribucije vrijedi da je

$$P(X=k)=\frac{\mu \cdot e^{-\mu}}{k!}, \; k=0,1,2,\ldots,$$

gdje je n broj pokusa, p vjerojatnost realizacije slučajnog događaja, $\mu = np$ i k broj povoljnih ishoda u n pokusa.

Očekivanje slučajne varijable X kod Poissonove razdiobe iznosi $E(X)=np=\mu$, a varijanca slučajne varijable X kod Poissonove razdiobe iznosi $\sigma^2 = np = \mu$.

Općenito se kaže da se slučajna varijabla X ravna po Poissonovoj razdiobi koja je određena parametrom μ i zapisuje se $X \sim Poi(\mu)$.

Napomena. Primjer koji slijedi sastavili su studenti preddiplomskog studija informatike u okviru svojih završnih radova.

Primjer 2.

Neka je prosječna duljina kodne riječi 5.8 na 100 kodova. Kolika je vjerojatnost da među 100 proizvoljno odabranih kodova

a) prosječna duljina kodne riječi neće biti veća od 4,

b) prosječna će duljina kodne riječi biti veća od 7?

Napomena. Entropijsko kodiranje bazirano je na činjenici da svaki signal ima jedinstvenu informaciju, a prosječna duljina koda vezana je za entropiju izvora informacije, što je poznato kao Shannonov prvi teorem. Kodna riječ za svaki simbol ne mora biti stalne duljine, nego duljina može biti varijabilna, ovisno o količini entropije. Osnovna je ideja entropijskog kodiranja predstaviti vjerojatnije kvantizacijske indekse kraćim kodnim riječima, a one manje vjerojatne duljim kodnim riječima, da bi se postigla manja prosječna duljina kodne riječi. Na ovaj način skup kvantizacijskih indeksa može biti prikazan manjim brojem bitova.

Napomena. U ovom primjeru ne znamo broj pokusa ni vjerojatnost realizacije slučajnog događaja, ali znamo njihov umnožak odnosno očekivanje slučajne varijable stoga ovaj primjer možemo riješiti samo pomoću Poissonove razdiobe.

Rješenje. Ponovno ćemo ilustrirati tri moguća načina rješavanja zadatka – direktno, numeričkim putem, pomoću *GeoGebre* te pomoću *Excela*.

Numeričko rješavanje. Označimo s X= „prosječna duljina kodne riječi na 100 kodova". Znamo da je $X \sim Poi(5.8)$.

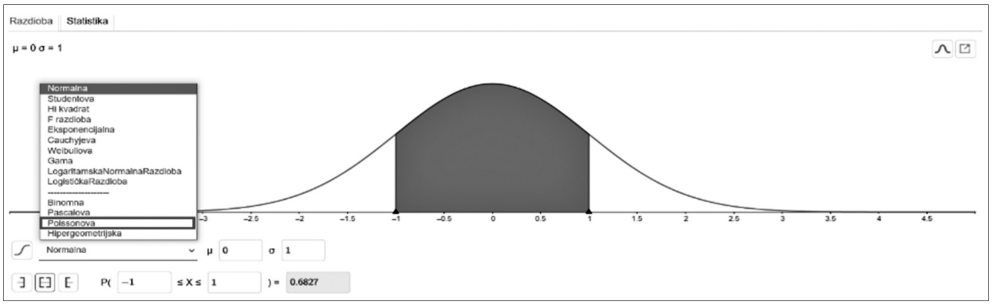
a) $P(X \leq 4) = P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)$

$$= \frac{5.8^0}{0!}\cdot e^{-5.8} + \frac{5.8^1}{1!}\cdot e^{-5.8} + \frac{5.8^2}{2!}\cdot e^{-5.8} + \frac{5.8^3}{3!}\cdot e^{-5.8} + \frac{5.8^4}{4!}\cdot e^{-5.8} = 0.3127 .$$

b) $P(X > 7) = 1 - P(X \leq 7) = \; = 1-(P(X=0)+\cdots+P(X=7))$

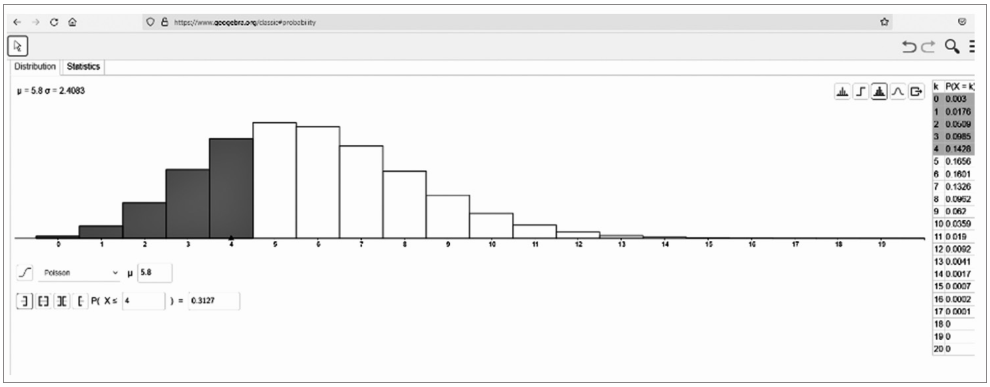
$$= 1-(\frac{5.8^0}{0!}\cdot e^{-5.8} + \cdots + \frac{5.8^7}{7!}\cdot e^{-5.8}) = 0.228974$$

Primjena programskog paketa GeoGebra. Za rješavanje primjera u *GeoGebr*i postupamo na isti način kao u prethodnom primjeru, s tim što nakon opcije **Vjerojatnost** u padajućem izborniku odaberemo opciju **Poissonova** (Slika 10.), a potom unosimo vrijednost μ, koja predstavlja očekivanje slučajne varijable, pa je u ovom primjeru potrebno unijeti vrijednost 5.8.



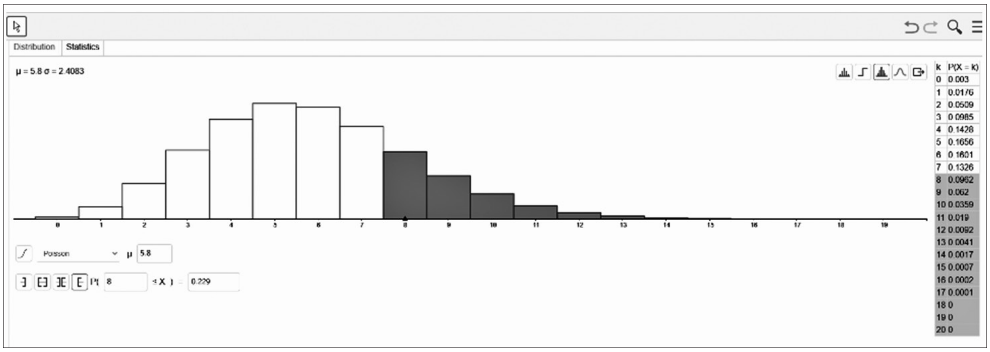
Slika 10.

U zadatku pod a) pitanje je kolika je vjerojatnost da među bilo kojih 100 kodova prosječna duljina kodne riječi neće biti veća od 4, stoga je potrebno izabrati ***Jednostrani lijevi interval***, a potom unijeti vrijednost 4 (Slika 11.). Dobivena vjerojatnost iznosi 0.3127 i jednaka je zbroju vjerojatnosti da će duljina kodne riječi biti jednaka 0, 1, 2, 3 ili 4. Pojedinačne vjerojatnosti za svaki točno definirani broj uzoraka prikazane su u tablici s desne strane (Slika 11.).



Slika 11.

U zadatku pod b) pitanje je kolika je vjerojatnost da će među bilo kojih 100 kodova prosječna duljina kodne riječi biti veća od 7, stoga je potrebno izabrati ***Jednostrani desni interval***, a potom unijeti vrijednost 8 (Slika 12.). Dobivena vjerojatnost iznosi 0.229.



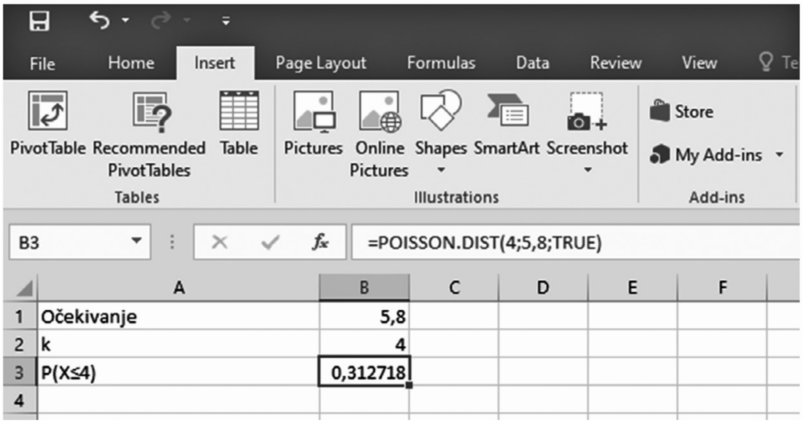
Slika 12.

Primjena programskog paketa Excel. Za rješavanje primjera u *Excelu* potrebno je koristiti sljedeće parametre i funkcije

- k – broj događaja
- μ – očekivani ishod
- funkciju za izračunavanje POISSON.DIST

- logičku vrijednost koja određuje oblik funkcije POISSON.DIST: TRUE biramo ako želimo izračunati vjerojatnost da će broj slučajnih događaja biti između nula i *k*, a FALSE ako želimo izračunati vjerojatnost da će broj događaja biti točno *k*.

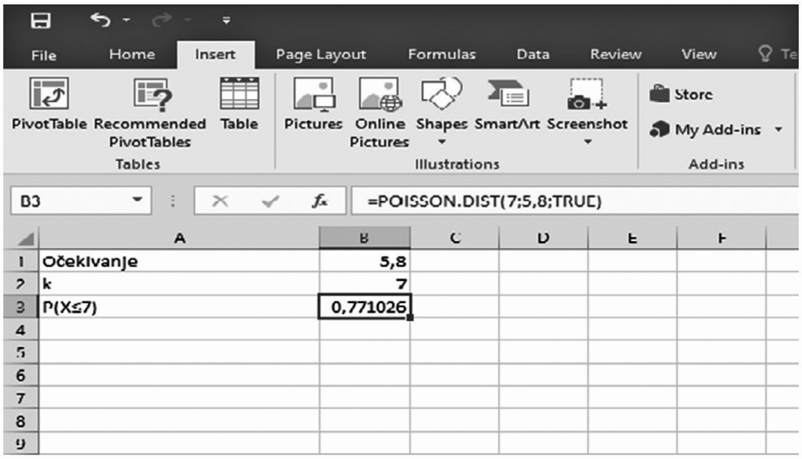
U primjeru pod a) potrebno je unijeti funkciju POISSON.DIST (4;5,8;TRUE) pomoću koje računamo vjerojatnost da među bilo kojih 100 kodova prosječna duljina kodne riječi neće biti veća od 4 (Slika 13.). Dobivena je vjerojatnost 0.3127.



Slika 13.

U primjeru pod b) potrebno je unijeti funkciju POISSON.DIST (7;5,8;TRUE) pomoću koje računamo vjerojatnost da među bilo kojih 100 kodova prosječna duljina kodne riječi neće biti veća od 7 (Slika 14.). Dobivena je vjerojatnost 0.771026 (Slika 14.). Potom računamo suprotnu vjerojatnost, odnosno vjerojatnost da će među bilo kojih 100 kodova prosječna duljina kodne riječi biti veća od 7:

$$P(X>7)=1-P(X\leq 7)=1-0.771026=0.228974\ .$$



Slika 14.

Zaključak

Iz navedenih je primjera očito da uporaba gotovih programskih paketa korisnicima približava matematičke koncepte i olakšava njihovo razumijevanje. Numeričko je rješavanje ovakvih problema složeno, ne samo u samom izračunu, nego i u postavci zadataka, dok gotovi programskih paketi pomažu pri provjeravanju pretpostavki, obradi i razmjeni podataka i informacija te za rješavanje problema i modeliranje. Na ovaj način svaki inženjer *nematematičar* može sam računati vjerojatnost ishoda pojedinih problema i pretpostaviti rezultate.

Literatura:

1. W. Volk, *Applied Statistics for Engineers,* McGraww-Hill Book Company, Inc., United States of America, 1958.
2. S. Pivac, A. Rozga, *Statistika za sociološka istraživanja,* Sveučilište u Splitu, Split, 2006.
3. II. SLUČAJNE VARIJABLE, http://matematika.fkit.hr/novo/statistika_i_vjerojatnost/predavanja/3-Osnove_teorije_vjerojatnosti_za_inzenjere_2.dio.pdf (3. 5. 2022.)
4. N. Koceić Bilan, *Primijenjena statistika*, Prirodoslovno-matematički fakultet u Splitu, Split, 2011.
5. https://personal.oss.unist.hr/~mnizetic/Laboratorijska vježba 1.pdf (3. 5. 2022.)
6. H.R. Wu, K.R. Rao, *Digital Video Image Quality and Perceptual Coding*, Boca Raton: CRC Press, 2006.