Statistika
(za biologe, 2.dio)

Bojan Basrak

rujan 2006

Osnovni pojmovi vjerojatnosti

Neizvjesnost je karakteristika gotovo svih procesa u našim životima.

Neizvjesni su npr.:

- rezultati fizikalnih mjerenja
- poslovni rezultati
- ishodi medicinskih tretmana
- dobit u igrama na sreću itd.

Iako nam nije lako definirati što su to neizvjesni dogadaji i kolika je njihova vjerojatnost nekakvu intuitivnu pretpostavku o ovim pojmovima imamo svi.

Matematički modeli u znanosti često pokušavaju modelirati neizvjesnost - npr. modelirajući mutacije DNK u evoluciji. Modeli zato uključuju tzv. vjerojatnosne ili stohastičke komponente.

Takav način modeliranja je dopušten čak i ako prihvatimo riječi A. Einstena "God doesn't play dice" izrečene protiv kvantne teorije.

Iako su vjerojatnosni modeli najčešće samo približno točni pravo je pitanje mogu li nam biti od koristi. Primjetite da su i Newtonovi zakoni samo približni iako dovoljno precizni da odvedu čovjeka na mjesec.

Suočeni s rezultatima nekog pokusa u statistici mi moramo moći kvantificirati neizvjesnost.

Vratimo se primjeru sa 100 bacanja novčića. Neka je 80 puta palo pismo. Prije nego pozovemo policiju jer je novčić neispravan, možemo uočiti da je takvo što moguće čak i ako je novčić savršeno simetričan, pa su ishod pismo i glava jednako vjerojatni. Samo ovakav ishod nije baš vjerojatan.

Zato je korisno odrediti vjerojatnost da se ovako ekstremno odstupanje od 50tak pisama koje očekujemo uopće dogodi kod ispravnog novčića. Kad postavimo matematički model ovog pokusa pokazat ćemo da je ova vjerojatnost zapravo

$$0.0000000011\ldots$$

Dakle izuzetno mala.

Već smo rekli da pojam vjerojatnost nije lako definirati. No intuitivno je lako razumjeti da ako ponavljamo isti slučajan pokus veliki broj puta n i bilježimo koliko puta se pojavio izvjestan rezultat A recimo n_A, tada očekujemo da će omjer (ili relativna frekvencija)

$$\frac{n_A}{n}$$

težiti k nekom broju, recimo p_A, taj broj bismo mogli zvati vjerojatnost dogadjaja A.

Iako neprecizna ovakva intuicija motivira matematičku definiciju vjerojatnosti.

U slučaju bacanja novčića rezultat $A = \{\text{palo je pismo}\}$ je ishod slučajnog pokusa. A ukoliko je novčić simetričan očekivali bismo da se relativna frekvencija njegovog pojavljivanja nakon puno bacanja približava $1/2$. Tada bismo pisali $p_A = 1/2$.

Dakle vjerojatnost vežemo uz procese koji imaju neizvjestan ishod. Takav proces ćemo zvati **slučajni pokus**.

Primjeri slučajnih pokusa

- nakon bacanja igraće koce zabilježimo rezultat
- nakon 100 ponovljenih bacanja novčića prebrojimo pisma
- nakon ispitivanja 1000 odabranih birača na tzv. exit polls, prebrojimo za koje su glasali liste.
- nakon sadnje 100 sadnica iste vrste biljaka u laboratorijskim uvjetima, bilježimo parametre njihova rasta tokom sljedeće 2 godine
- istu vrstu bakterija napadnemo in vitro s dva različita antibiotika, sa svakim u 20 odvojenih posuda i bilježimo rezultate nakon 48 sati.
- odabiremo na slučajan način bebu rodjenu u RH tokom protekle godine i bilježimo njene karakteristike.

Elementarni dogadjaj je svaki mogući (i nedjeljivi) ishod slučajnog pokusa.

Skup svih elementarnih dogadjaja zovemo **prostor elementarnih dogadjaja**, tipična matematička oznaka je Ω.

Slučajan dogadjaj je bilo koji podskup od Ω.

Posebno svaki slučajni dogadjaj skup je elementarnih dogadjaja. Slučajne dogadjaje obilježavamo velikim slovima abecede tipično. Uočite: sl. dogadjaj se može sastojati i od samo jednog elementarnih dogadjaja.

Primjer

i) Kod bacanja jedne kocke elementarni ishodi su 1,2,3,4,5 i 6, pa je i

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

a slučajni dogadjaj bi mogao biti npr.

$$A = \{ \text{ pao je paran broj} \} = \{2, 4, 6\}.$$

ii) Kod bacanja jednog novčića ishodi su pismo ili glava, (alternativno 0 ili 1), pa pišemo

$$\Omega = \{\text{P,G}\} \text{ ili } \{0, 1\}.$$

a slučajni dogadjaj bi mogao biti npr. $A = \{\text{P}\}$.

iii) Kod slučajnog odabira bebe rodjene u RH prošle godine

$$\Omega = \{\omega_1, \ldots, \omega_N\},$$

gdje je N broj takvih beba, a ω_i su njihove jednoznačne oznake npr. JMBG. Slučajni dogadjaj bi mogao biti npr.

$$A = \{ \text{ odabrana beba je ženskog spola } \}$$

ili u drugom zapisu

$$A = \{\omega_i : \omega_i \text{ je ž. spola}\}.$$

iv) Kod bacanja para kocaka možemo staviti

$$\Omega = \{(i, j) : i, j = 1, \ldots, 6\}.$$

$\diamond$ Ω može biti i beskonačno neprebrojiv skup npr. kada odabiremo slučajnu točku na nekom intervalu ili u kvadratu. No tu je matematička teorija puno zahtjevnija.

$\diamond$ Uvedimo i pojam **suprotnog dogadaja**, kao A^c

$\diamond$ Sl. dogadjaji su skupovi dakle elementarnih dogadjaja pa na njima ima smisla provjeravati različite skupovne relacije kao npr.

$A \subseteq B$ dogadjaj A povlači dogadjaj B

$A = B^c$ dogadjaji A i B su suprotni

$A \cap B = \emptyset$ dogadjaji A i B su disjunktni ili se isključuju

i uvoditi skupovne operacije npr.

$A \cap B$ dogodili su se i dogadjaj A i dogadjaj B

$A \cup B$ dogodio se bar jedan on dogadjaja A i B

$\diamond$ Relacije i operacije ćemo najlakše predstaviti tzv. Vennovim dijagramima.

Vjerojatnost sl. dogadjaja

Vjerojatnost sl. dogadjaja A označit ćemo brojem $P(A)$. Od vjerojatnosti ćemo tražiti da zadovoljava izvjesna svojstva:

- Za svaki sl. dogadjaj A

$$0 \leq P(A) \leq 1.$$

- Vjerojatnost da će se išta dogoditi je 1, tj.

$$P(\Omega) = 1.$$

- Ako se dogadjaji $A_1, A_2, \ldots$ medjusobno isključuju tada vrijedi

$$P(\cup_i A_i) = \sum_i P(A_i).$$

Unatoč ovim formalnim pravilima nije jasno kako doći do broja $P(A)$ za zadani sl. dogadjaj A.

Pokušajmo ipak s nekim primjerima

Primjer C. de Mere je u 17 st. pitao niz matematičara – što je vjerojatnije da će u 4 bacanja kocke pasti bar jedna šestica (pokus 1) ili u 24 bacanja kocke pasti bar jedna dvostruka šestica (pokus 2).

Primjetite očekivani broj uspješnih pokusa u prvom slučaju je (intuitivno, za sada)

$$\frac{4}{6},$$

a u drugom

$$\frac{24}{36},$$

dakle isto. No vjerojatnosti nisu jednake naslućivao je de Mere. Odgovor je nastao u korespodenciji B. Pascala i P. de Fermata i smatra se početkom teorije vjerojatnosti.

Primjer Kolika je vjerojatnost da će nakon bacanja igraće kocke pasti paran broj? Već smo odredili sve moguće ishode tj.

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

A tražimo

$$P(A)$$

za $A = \{$ pao je paran broj$\} = \{2, 4, 6\}$. Ako je kocka simetrična, svi elementarni dogadjaji su jednako vjerojatni, tj.

$$P(1) = P(2) = \cdots = P(6).$$

Kako se elementarni dogadjaji medjusobno isključuju definicija vjerojatnosti nam kaže

$$P(\Omega) = P(\{1\} \cup \ldots \cup \{6\}) = \sum_{i=1}^{6} P(i) = 6P(1).$$

No $P(\Omega) = 1$ pa dakle vrijedi

$$P(i) = \frac{1}{6}$$

za sve i.

Sad je

$$P(A) = P(\{2\} \cup \{4\} \cup \{6\}) = P(2) + P(4) + P(6) = \frac{3}{6}.$$

Ovaj primjer nam daje vrlo važno općenito pravilo.

$\diamond$ Ako možemo pretpostaviti da su elementarni dogadjaji jednako vjerojatni i ako ih je konačno, npr. $n \in \mathbb{N}$, tad je vjerojatnost svakog od njih

$$\frac{1}{n}.$$

Nadalje, vjerojatnost bilo kojeg sl. dogadjaja A u tom slučaju je

$$P(A) = \frac{\text{broj elementarnih dogadjaja koji su povoljni za } A}{\text{broj svih elementarnih dogadjaja}}$$

Dakle jednako vjerojatni elementarni dogadjaji daju vrlo jednostavan način za računanje vjerojatnosti.

Prema definiciji vjerojatnosti, vjerojatnost bilo kojeg dogadjaja A možemo naći kao

$$P(A) = \sum_{i:\omega_i \in A} P(\omega_i)\,,$$

čak i ako nisu svi ω_i jednako vjerojatni. No općenito je ovu vjerojatnosti teže izračunati kada elem. dogadjaji nisu jednako vjerojatni.

OPREZ U praksi je česta greška pretpostaviti da su dogadjaji jednako vjerojatni iako oni to nisu.

Npr. ako izlaznu anketu na izborima radimo samo u urbanim sredinama, nije razumno pretpostaviti da svi birači imaju jednaku vjerojatnost biti članovima uzorka.

Primjer Pretpostavimo da se sl. pokus sastoji u promatranju spola prvo dvoje djece koja će se roditi sutra u Zagrebu i Splitu. Odredimo vjerojatnost da će se prve dvije bebe biti različitog spola u različitim gradovima. Pretpostavimo (aproksimativno) da su kod beba oba spola jednako vjerojatna. Prije odredjivanja vjerojatnosti trebamo odabrati dobar matematički model.

Prva ideja: postavimo

$$\Omega = \{bb, bg, gg\} \text{ i } P(bb) = P(bg) = P(gg) = 1/3.$$

Druga ideja

$$\Omega = \{(b,b),(b,g),(g,b),(g,g)\} \text{ i } P(b,b) = P(b,g) = P(g,b) = P(g,g) = 1/4.$$

Iako su u oba modela elem. dogadjaji jednako vjerojatni, samo je jedan model razuman, koji?

Pravila za računanje vjerojatnosti

Direktno iz matematičkog opisa vjerojatnosti slijedi

$$P(A^c) = 1 - P(A).$$

$$A \subseteq B \text{ povlači } P(A) \leq P(B)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

te

$$P(\emptyset) = 0.$$

Posebno ako su A i B disjunktni

$$P(A \cup B) = P(A) + P(B).$$

Lako je provjeriti da sve ovo vrijedi i u posebnom slučaju kad su elem. dogadjaji jednako vjerojatni. No tada mi znamo i više

$$P(A) = \frac{\text{broj elementarnih dogadjaja koji su povoljni za } A}{\text{broj elementarnih dogadjaja u } \Omega}.$$

Gornja formula sugerira da je jako važno u modelu s jednako vjerojatnim elem. dogadjajima znati prebrojati elemente proizvoljnog skupa $A \subseteq \Omega$. Ponovimo neka općenita pravila o broju elem. ishoda pojedinih pokusa.

$\diamond$ (Pravilo množenja) Ako imamo dva sl. dogadjaja A_1 i A_2 vezana uz dva odvojena pokusa, tada možemo uvesti novi dogadjaj $A = \{$u prvom pokusu dogodio se A_1 u drugom $A_2\}$. Broj povoljnih elem. dogadjaja za A je sada

$$n_1 \cdot n_2$$

gdje je n_i broj elementarnih dogadjaja koji su povoljni za A_i.

$\diamond$ Slično ako imamo više sl. dogadjaja $A_1, A_2, \ldots A_k$ vezanih uz k odvojena pokusa, tada možemo uvesti novi dogadjaj $A = \{$u i-tom pokusu dogodio se $A_i\}$. Broj povoljnih elem. dogadjaja za A je sada

$$n_1 n_2 \cdots n_k$$

gdje je n_i broj elementarnih dogadjaja koji su povoljni za A_i.

$\diamond$ (Broj permutacija) Pretpostavimo da iz skupa od N elemenata izabiremo uzorak od njih r pazeći pri tom na poredak. To možemo učiniti na

$$N(N-1)\cdots(N-r+1)$$

načina. Taj broj je broj **permutacija** duljine r iz skupa od N elemenata.

$\diamond$ Ako se pitamo samo na koliko načina možemo poredati N elemenata, to je specijalni slučaj permutacija za $r = N$, pa je taj broj

$$N(N-1)\cdots 2\cdot 1 = N!$$

Po definiciji je $0! = 1$.

Primjer
Na koliko načina možete posložiti 3 od 4 slova A,C,G,T kako biste kreirali sve kodone u kojima nema istih nukleotida?

Odgovor je dakako

$$4 \cdot 3 \cdot 2 = 24.$$

Usput, koliko je kodona sveukupno?

$\diamond$ (Broj kombinacija) Pretpostavimo da iz skupa od N elemenata izabiremo uzorak od njih r **ne** pazeći pri tom na poredak. To možemo učiniti na

$$\frac{N(N-1)\cdots(N-r+1)}{r!} = \frac{N!}{r!(N-r)!} = \binom{N}{r}$$

načina. Taj broj zovemo broj **kombinacija** duljine r iz skupa od N elemenata.

Primjer
Na koliko načina možete izabrati 7 brojeva od 39?

Odgovor je dakako

$$\binom{39}{7} = 15380937\,.$$

Pa je i vjerojatnost da ćete uplatom jedne kombinacije u igri lota 7 od 39 osvojiti glavni dobitak izuzetno mala i iznosi otprilike

$$6.5 \cdot 10^{-8}.$$

Uvjetna vjerojatnost

Ponekad imamo dijelomičnu informaciju o ishodu sl. pokusa. Naprimjer kod bacanja igraće kocke iako nam je nepoznat točan ishod mi bismo mogli znati da je rezultat broj veći od 3. Vjerojatnost elem. dogadjaja 4 uz ovaj uvjet je

$$\frac{1}{3}.$$

Naime, sad znamo da su dogadjaji 4,5 i 6 jedini mogući, ali i jednako vjerojatni. Tako da je vjerojatnost zapravo omjer broja povoljnih elem. dogadjaja i broja dogadjaja koji se uopće mogu dogoditi uz informaciju koju mi imamo.

Općenito ako imamo pokus sa jednako vjerojatnim elem. ishodima, i znamo da se dogodio neki od ishoda u skupu B te se pitamo kolika je vjerojatnost da se dogodio sl. dogadjaj A uz ove uvjete, odgovor je

$$P(A|B) = \frac{\text{broj elem. dogadjaja u } A \cap B}{\text{broj elem. dogadjaja u } B} = \frac{P(A \cap B)}{P(B)}.$$

Motivirani ovim primjerom možemo definirati **uvjetnu vjerojatnost** dogadjaja A uz uvjet da se dogodio dogadjaj B

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

naravno uz pretpostavku da je $P(B) > 0$.

Uvjetna vjerojatnost pokazuje kako naše prethodno znanje o ishodu pokusa mijenja vjerojatnosti svih ostalih dogadjaja.

Npr. vjerojatnost da će sl. odabrani student u RH biti student poljoprivrednog fakulteta se mijenja ako nam je poznato da student dolazi iz Rijeke npr.

S druge, strane vjerojatnost da će ako taj student baci kocku pasti 6, ostaje $1/6$ čak i ako znamo ovu činjenicu.

Rekli bismo da su dogadjaji $\{$ sl. odabrani student bacio je 6 na kocki$\}$ i $\{$ sl. odabrani student dolazi iz Rijeke$\}$ nezavisni.

Matematički precizno, sl. dogadjaji A i B su **nezavisni** ako vrijedi

$$P(A \cap B) = P(A)P(B)\,.$$

Posebno je tada

$$P(A|B) = P(A)\,.$$

Dakle vjerojatnost dogadjaja A se ne mijenja čak i ako znamo da se dogodio dogadjaj B.

Primjer Kod bacanja dvije kocke dogadjaji da je na prvoj odn. drugoj kocki pala šestica su nezavisni.

Primjer
Sad možemo pokušati riješiti de Mereov problem. Pretpostavimo da pokus bacanja dvije igraće kocke opisuje sljedeći model: $\Omega = \{(i,j) : i, j+1, \ldots, 6\}$, te da su svi elem. dogadaji jednako vjerojatni.

Prema tome za svaki ishod (i,j) imamo

$$P(i,j) = \frac{1}{36}\,.$$

Vjerojatnost dogadjaja da će u 24 bacanja dvije kocke pasti barem jedan par 6-ica je

$$P(A) = 1 - P(\text{niti u jednom od 24 bacanja nisu pale dvije 6})$$

ili

$$1 - P(\cap_{i=1}^{24}\{\text{u bacanju } i \text{ palo je nešto drugo od dvije 6}\})$$

$$P(\cap_{i=1}^{24}\{\text{u bacanju } i \text{ palo je nešto drugo od dvije 6}\})$$

je zbog nezavisnosti izmedju bacanja zapravo umnožak 24 vjerojatnosti oblika

$$P(\text{u bacanju } i \text{ palo je nešto drugo od dvije 6})\,.$$

No za sve i ova vjerojatnost je ista kako su svi ishodi jednako vjerojatni i iznosi

$$\frac{35}{36}.$$

Pa je tražena vjerojatnost

$$P(A) = 1 - \left(\frac{35}{36}\right)^{24} = 0.4914039.$$

Slično se može pokazati (pokažite) da je vjerojatnost bar jedne 6-ice u 4 bacanja jedne kocke

$$1 - \left(\frac{5}{6}\right)^4 = 0.5177469.$$

Dajle ovakva oklada je povoljnija. TIme semo našli odgovor na de Mereov problem.

Bayesova formula

Direktno iz definicije uvjetne vjerojatnosti može se dobiti sljedeća tzv. **Bayesova formula**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}\,.$$

Primjer (prisoner's dilemma)
Pretpostavimo da zatvorski čuvar napravi sljedeće: ako je B onaj koji neće biti pušten on otkriva da će pušten biti C i obrnuto. Jedino u slučaju kad A neće biti pušten, stražar odabire na slučajan način jednog od B i C i to govori zatvoreniku A.

Dogadjaji $\{$B je otkriven$\}$ i $\{$A je zadržan$\}$ su uz ove uvjete nezavisni. Da bismo to pokazali uočite $\{$B je otkriven$\} = \{$C je otkriven$\}^c$. Pa je zbog simetričnosti naših uvjeta

$$P(\text{B je otkriven}) = P(\text{C je otkriven}) = \frac{1}{2}\,.$$

Nadalje, prema pretpostavkama je $P(\text{B je otkriven} \mid \text{A je zadržan}) = 1/2$. Pa Bayesova formula daje

$$P(\text{A je zadržan} \mid \text{B je otkriven}) = \frac{1/2 \cdot 1/3}{1/2}$$

Dakle

$$P(\text{A je zadržan}|\text{B je otkriven}) = P(\text{A je zadržan}) = \frac{1}{3}.$$

Ako se skup Ω može napisati kao unija disjunktnih podskupova $H_i,\ i = 1, 2, \ldots$, tada je

$$P(B) = \sum_i P(B|H_i)P(H_i)\,.$$

Zato se za $A = H_1$ Bayesova formula može napisati i u ovom obliku

$$P(H_1|B) = \frac{P(B|H_1)P(H_1)}{\sum_i P(B|H_i)P(H_i)}\,.$$

Primjer

Postoji test koji ispituje prisutnost odredjene rijetki bolesti na osnovu uzorka krvi. Test nije savim precizan, pa je poznato da vrijedi

$$P(\text{test je pozitivan}|\text{pacijent je bolestan}) = 0.99$$

ali i

$$P(\text{test je pozitivan}|\text{pacijent je zdrav}) = 0.02$$

Pretpostavite da je sl. odabrana osoba pozitivna na test, ako se bolest javlja u 1 od 10000 osoba u populaciji, kolika je vjerojatnost da je ta osoba bolesna?

Iskoristimo Bayesovu formulu, stavimo $B = \{\text{test je pozitivan}\}$, te

$$H_1 = \text{pacijent je bolestan}, \ H_2 = \text{pacijent je zdrav}\,.$$

Slijedi

$$P(H_1|B) = \frac{0.99 \cdot 1/10000}{0.99 \cdot 1/10000 + 0.02 \cdot 9999/10000} = 0.0049$$

Modeli razdioba i slučajne varijable

Kako smo već istaknuli u primjerima, na osnovi sl. pokusa, u statistici često prikupljamo podatke o nekom numeričkom obilježju jedinki u populaciji. Npr. visina, težina ili uspjeh na ispitu za slučajno odabranog studenta su takva obilježja. Matematički model za numeričke rezultate sl. pokusa je **slučajna varijabla**.

Slučajna varijabla, npr. X je funkcija koja svakom elem. ishodu pokusa ω pridružuje broj $X(\omega)$.

Primjer Pretpostavimo da nakon bacanja novčića ako padne pismo dobijamo 1 EUR, a ako padne glava gubimo 2 EUR. Pokus opisuje npr. $\Omega = \{\mathsf{P},\mathsf{G}\}$, a naša zarada je sl. varijabla definirana da

$$X(\mathsf{P}) = 1,\ X(\mathsf{G}) = -2.$$

Diskretne slučajne varijable

U gornjem primjeru sl. varijabla X može primiti samo dvije vrijednosti 1 i -2, pa bismo rekli da je diskretna.

Općenito medjusobno različite vrijednosti koje poprima **diskretna sl. varijabla** možemo napisati kao niz: $a_1, a_2, a_3, \ldots$. Definirajmo niz

$$p_i = P(X = a_i) = P(\omega \in \Omega : X(\omega) = a_i)\,.$$

Kažemo da ova dva niza odn. tablica

$$\begin{pmatrix} a_1 & a_2 & a_3 & \ldots \\ p_1 & p_2 & p_3 & \ldots \end{pmatrix},$$

predstavljaju **razdiobu ili distribuciju sl. varijable** X.

Pišemo

$$X \sim \begin{pmatrix} a_1 & a_2 & a_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix}.$$

Alternativno koristimo i tablicu sljedećeg oblika

x	a_1	a_2	a_3	$\dots$
$P(X = x)$	p_1	p_2	p_3	$\dots$

Primjetite da su brojevi p_i uvijek nenegativni, te da vrijedi

$$\sum_i p_i = 1.$$

No primjetimo i da svaka tablica oblika koji smo vidjeli na prethodnom slajdu zadaje razdiobu neke sl. varijable ako i samo ako su svi p_i nenegativni i vrijedi gornji uvjet.

Primjer

Ako bacamo dva novčića, pokus opisuje npr. $\Omega = \{PP, PG, GP, GG\}$. Ako nas zanima broj pisama koji je pao, to je sl. varijabla zadana sa $X(PP) = 2,\ X(PG) = X(GP) = 1$, i $X(GG) = 0$. Ako je novčić nepristran, svi su elem. dogadjaji jednako vjerojatni. Tako da X ima sljedeći zakon razdiobe

x	0	1	2
$P(X = x)$	$1/4$	$1/2$	$1/4$

Zakon razdiobe možemo takodjer prikazati grafički stupčastim dijagramom ili tzv. vjerojatnosnim histogramom na jednostavan način. Napravite to za broj pisama u prethodnom primjeru, ali i za broj pisama nakon bacanja 3 novčića.

Matematičko očekivanje

Ako je X sl. varijabla tako da

$$X \sim \begin{pmatrix} a_1 & a_2 & a_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix} .$$

matematičko očekivanje definiramo kao broj

$$E\,X = E(X) = \sum_i a_i p_i \,.$$

Primjer
Nadjite matematičko očekivanje za broj pisama nakon bacanje dva novčića.

$$E(X) = 0\frac{1}{4} + 1\frac{1}{2} + 2\frac{1}{4} = 1.$$

Napravite to za 3 i 4 novčića.

▷ Ako su X i Y dvije sl. varijable a α proizvoljan realan broj, matematičko očekivanje zadovoljava

$$E(\alpha X) = \alpha E(X)$$

$$E(X + Y) = E(X) + E(Y)$$

▷ Ako je $g : \mathbb{R} \to \mathbb{R}$ proizvoljna funkcija tada vrijedi

$$E[g(X)] = \sum_i g(a_i) p_i \,. \tag{1}$$

Varijanca i standardna devijacija

Ako stavimo $g(u) = (u - E(X))^2$, tada definiramo **varijancu sl. varijable** X kao broj

$$\mathrm{var}X = E[g(X)] = E[(X - EX)^2].$$

Prema (1) je

$$\mathrm{var}X = \sum_i (a_i - E(X))^2 p_i \,.$$

Može se pokazati da za varijancu vrijedi

$$\mathrm{var}(X) = E(X^2) - (E(X))^2,$$

a za sve $a, b \in \mathbb{R}$

$$\mathrm{var}(aX + b) = a^2 \mathrm{var} X.$$

Standardnu devijaciju sl. varijable X definiramo kao broj

$$\sigma_X = +\sqrt{\mathrm{var} X}.$$

Primjer

Nadjite varijancu odn. standardnu devijaciju za broj pisama nakon bacanja 3 novčića. Pokažite

$$\mathrm{var}X = \frac{3}{4}$$

Primjetite mat. očekivanje, varijanca, i st. devijacija imale su svoj ekvivalent medju deskriptivnim statistikama.

Varijanca je intuitivno mjera raspršenosti razdiobe sl. varijable oko njenog očekivanja. Ponekad se koristi i kao mjera rizika, npr. u financijskoj industriji ako dvije investicije imaju slučajan dobitak, no istog očekivanja, manje rizičnom se smatra ona koja ima manju varijancu tog dobitka.

Za razdiobe možemo definirati i medijan, kvartile odn. percentile kao što to radili i za sl. uzorak. Pokušajte pogoditi kako bismo definirali medijan npr.

Zajednička razdioba dvije sl. varijable

Dvije sl. varijable X i Y istom elem dogadjaju pridružuju dva realna broja.

Primjer Za sl. pokus bacanja igraće kocke neka je $X(\omega) = 1$ za parne ω, a 0 za neparne. A $Y(\omega) = 1$ za $\omega = 5$ ili 6, a 0 za sve ostale. Tada zajednički zakon razdiobe od X i Y možemo zadati tablicom

	0	1
0	$P(X = 0, Y = 0) = 2/6$	$P(X = 1, Y = 0) = 2/6$
1	$P(X = 0, Y = 1) = 1/6$	$P(X = 1, Y = 1) = 1/6$

Općenito zajednički zakon razdiobe za diskretne sl. varijable X i Y možemo zadati tablicom

	a_1	a_2	a_3	$\dots$
b_1	$P(X=a_1, Y=b_1)$	$P(X=a_2, Y=b_1)$	$P(X=a_3, Y=b_1)$	$\dots$
b_2	$P(X=a_1, Y=b_2)$	$P(X=a_2, Y=b_2)$	$P(X=a_3, Y=b_2)$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Primjetite da je suma vjerojatnosti u tablici uvijek 1.

Za diskretne sl. varijable X iY kažemo da su **nezavisne** ako za sve a_i i b_j iz gornje tablice vrijedi

$$P(X = a_i,\, Y = b_j) = P(X = a_i)P(Y = b_j)\,.$$

Za sl. varijable iz prethodnog primjera možemo vidjeti da su nezavisne jer npr.

$$P(X = 0, Y = 0) = 1/3 = P(X = 0)P(Y = 0)\,.$$

Ako definiramo na istom sl. pokusu sl. varijablu Z koja je 1 samo za $\omega = 1$ a 0 inače. Tada

$$P(X = 1, Z = 1) = 0 \neq P(X = 1)P(Z = 1)\,,$$

pa X i Z nisu nezavisne.

Za nezavisne sl. varijable vrijedi

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

i

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$$

Za sl. varijable X i Y definiramo **kovarijancu** od X i Y kao

$$\operatorname{cov}(X, Y) = E[(X - EX)(Y - EY)].$$

Lako se vidi da vrijedi

$$\operatorname{cov}(X, Y) = E(XY) - EXEY.$$

Posebno je za nezavisne sl. varijable X i Y $\operatorname{cov}(X, Y) = 0$.

Iz zajedničke razdiobe kovarijancu računamo po formuli

$$\mathrm{cov}(X,Y) = \sum_i \sum_j (a_i - EX)(b_j - EY) P(X = a_i, Y = b_j)\,.$$

Nadalje vrijedi

$$\mathrm{var}(X+Y) = \mathrm{var}X + \mathrm{var}Y + 2\mathrm{cov}(X,Y).$$

Za sl. varijable X i Y definiramo **koeficijent korelacije** kao

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Ako $\operatorname{cov}(X, Y) = 0$, slijedi $\operatorname{corr}(X, Y) = 0$, u tom slučaju kažemo da su X i Y **nekorelirane**. Npr. nezavisne sl. varijable su uvijek nekorelirane, no obrnuto ne vrijedi.

Broj $\operatorname{corr}(X, Y)$ je uvijek izmedju -1 i 1. U slučaju $\operatorname{corr}(X, Y) = \pm 1$ sl. varijable su X i Y potpuno linearno zavisne, preciznije, za neke $a, b \in \mathbb{R}$

$$Y = aX + b\,.$$

Binomna razdioba

Zakonom razdiobe odredjujemo **vjerojatnosni model** za izvjesno numeričko obilježje.

Najvažniji primjer pokusa je onaj kod kojeg imamo samo dva moguća ishoda: "uspjeh" i "neuspjeh". Ako je vjerojatnost uspjeha $p \in [0, 1]$ zakon razdiobe sl. varijable X koja iznosi 1 ako se dogodio "uspjeh", a 0 ako se dogodio "neuspjeh", izgleda ovako

$$X \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix},$$

gdje je $q = 1 - p$. Sl. varijablu X s ovakom razdiobom zovemo **Bernoullijeva sl. varijabla s parametrom** p.

Ako sada gore opisan pokus ponavljamo n puta *nezavisno*, tada nam je interesantno vidjeti kako se ponaša ukupan broj uspjeha, recimo ponovo X. X je svakako izmedju 0 i n, no kolika je vjerojatnost dogadjaja $\{X = k\}$?

Ispostavlja se za $k = 0, 1, \ldots, n$

$$P(X = k) = \binom{n}{k} p^k q^{n-k}.$$

Imate li argument za ovu formulu? Sl. varijablu X za koju ona vrijedi zovemo **binomna sl. varijabla s parametrima** n i p. Pišemo katkad $X \sim B(n, p)$. Za $n = 4$, $p = 1/4$ napišite tablicu razdiobe za $X \sim B(4, 1/4)$.

Primjetite, Bernoullijeva sl. varijabla je specijalno i binomna, no tada je broj ponavljanja, odn. parametar $n = 1$.

Provjerite da je binomna razdioba dobro definirana.

Primjer
Pretpostavimo da iz posijanog sjemena razvije biljka u 80% slučajeva. Kolika je vjerojatnost da će iz 5 sjemenki niknuti barem dvije? A manje od 2?

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = \cdots = 0.99328.$$

Mat. očekivanje za X binomnu sl. varijablu s parametrima n i p je

$$EX = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = np.$$

A varijanca je

$$\mathrm{var} X = npq\,.$$

Pa je st. devijacija

$$\sigma_X = \sqrt{npq}\,.$$

Hipergeometrijska razdioba

Pretpostavimo da u populaciji od m jedinki od kojih je $r \leq m$ na neki način obilježeno biramo sl. uzorak od njih $n \leq m$. Ako sa X označimo sl. varijablu koja prebroji obilježene jedinke u našem uzorku, kažemo da je X **hipergeometrijska sl. varijabla s parametrima** m, r i n. Očito je X broj izmedju 0 i r, zakon razdiobe od X formulom zapisujemo kao

$$P(X = k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}},$$

$k = 0, 1, \ldots, r$.

Primjer U kutiji su 15 bijelih i 10 crvenih kuglica, ako na sl. način izaberemo 8 kuglica iz kutije, kolika je vjerojatnost da je medju njima bar 2 crvene?

Postavimo $m = 25,\ r = 10,\ n = 8$, a tražimo

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = \cdots$$

Hipergeometrijska sl. varijabla X ima očekivanje

$$EX = \frac{rn}{m}$$

te varijancu koju ne moramo pamtiti.

Poissonova razdioba

Pretpostavimo da je u vrlo velikoj populaciji relativno malo označenih jedinki. Ako uzmemo veliki uzorak iz populacije ispostavlja se da razdioba broja obilježenih jedinki, recimo X ima sljedeći oblik

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

za $k = 0, 1, 2, \ldots$ i neku konstantu $\lambda > 0$. Za ovakvu sl. varijablu X kažemo da je **Poissonova sl. varijabla a parametrom** λ.

Ako je X Poissonova sl. varijabla a parametrom λ tada je

$$EX = \mathrm{var}X = \lambda.$$

Primjer
Pretpostavljamo da želimo prebrojati broj izvjesnih rijetkih planktona u uzorku od 1l vode iz Jadranskog mora. Označimo taj broj sa X i pretpostavimo da znamo da je očekivani broj plantkona u 1l vode 3. Tada će X imati Poissonovu razdiobu s parametrom 3.

Razmislite što je ovdje populacija odn. jedinka.

Neprekidne sl. varijable

Diskretne sl. varijable su nam služile kao model kod prebrojavanja raznih elemenata u uzorku. Primale su vrijednosti tipično u skupu $0, 1, 2, \ldots$. Neka numerička obilježja, npr. visina u m, padaline u l/m^2, duljina životnog vijeka u godinama i sl. mogu teoretski poprimiti sve vrijednosti u nekom intervalu. Kako u praksi sva mjerenja zaokružujemo na odredjeni broj decimalnih mjesta i njih bismo mogli modelirati diskretnim sl. varijablama. No pokazuje se da postoji matematički jednostavniji model, posebno kad je različitih vrijednosti koje poprimaju naša mjerenja zaista mnogo.

Sjetimo se histograma koji smo koristili kod neprekidnih numeričkih obilježja. Podijelili smo skup mogućih vrijednosti u intervale oblika $I_j = [a_j, a_{j+1})$, izračunali bismo relativnu frekvenciju podataka u j-tom razredu i podijelili ga s njegovom duljinom.

Histogram je bio stepenasti dijagram kojemu je ukupna površina koju omedjuje iznosila 1. Relativna frekvencija svakog razreda bila je jednaka površini histograma iznad tog intervala.

Intuitivno očekujemo da relativna frekvencija podataka u j-tom intervalu odgovara vjerojatnosti da će mjerenje za sl. odabranu jedinku iz populacije pasti u taj interval.

IDEJA Ako poželimo svakom intervalu dodijeliti njegovu vjerojatnost, mogli bismo to učiniti preko neke općenitije funkcije f, tako da vjerojatnost upadanja obilježja (odn. sl. varijable) u taj interval bude jednaka površini omedjenoj rubovima intervala i grafom funkcije f.

Po analogiji s histogramom, prirodno je tada da je ukupna površina ispoda grafa od f jednaka 1, te da je f nenegativna funkcija. Takvu fukciju $f : \mathbb{R} \to \mathbb{R}$ zvat ćemo **funkcija gustoće**, dakle za nju vrijedi

$$\int_{-\infty}^{\infty} f(t)\mathrm{d}t = 1\,, \;\; f(t) \geq 0.$$

Za sl. varijablu X kažemo da je **neprekidna** ako postoji funkcija gustoće $f = f_X$ takva da za sve $a < b$ vrijedi

$$P(a \leq X \leq b) = \int_a^b f_X(t)\mathrm{d}t\,.$$

Posebno za neprekidnu sl. varijablu X i bilo koji realan broj a vrijedi

$$P(X = a) = 0\,!?!$$

Nadalje

$$P(X \leq b) = \int_{-\infty}^{b} f_X(t)\mathrm{d}t\,.$$

Funkcija $F_X(b) = P(X \leq b)$ zove se funkcija distribucije od X. I očito je rastuća (zašto?)

Očekivanje i varijanca neprekidnih razdioba

Matematičko očekivanje neprekidne sl. varijable X definiramo kao

$$EX = \int_{-\infty}^{\infty} t f_X(t) \mathrm{d}t.$$

Ako sl. varijabla X prima vrijednosti samo u konačnom intervalu $[a, b]$, ovaj integral je lakše izračunati jer tada

$$EX = \int_{a}^{b} t f_X(t) \mathrm{d}t.$$

Ako je $g : \mathbb{R} \to \mathbb{R}$ neprekidna funkcija tada vrijedi

$$E[g(X)] = \int_{-\infty}^{\infty} g(t) f_X(t) \mathrm{d}t.$$

Tako da je **varijanca** od X

$$\mathrm{var} X = E[(X - EX)^2] = \int_{-\infty}^{\infty} (t - EX)^2 f_X(t) \mathrm{d}t.$$

Normalna razdioba

Neprekidna sl. varijabla X ima **normalnu razdiobu s parametrima** $\mu \in \mathbb{R}$ i $\sigma^2 > 0$ ako joj je funkcija gustoće

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

Skraćena oznaka za ovu tvrdnju je $X \sim N(\mu, \sigma^2)$.

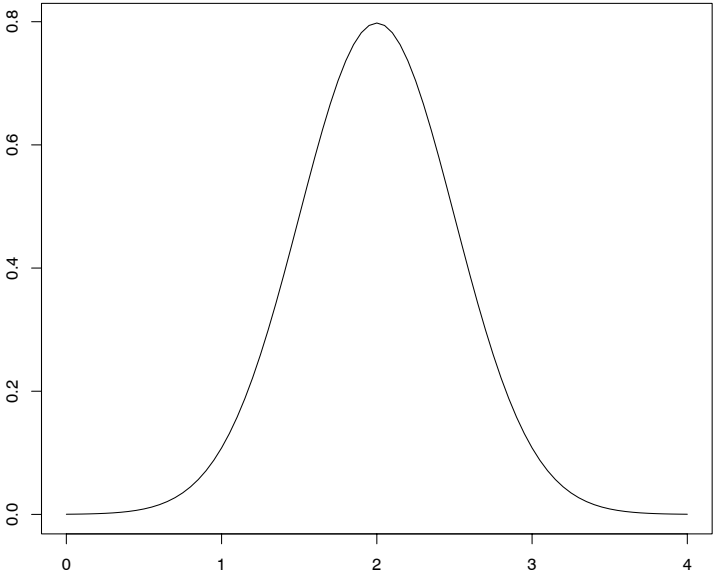
Primjetimo $f_X(t) > 0$ za sve $t \in \mathbb{R}$, tako da X prima vrijednosti u cijelom skupu realnih brojeva.

Matematičko očekivanje sl. varijable $X \sim N(\mu, \sigma^2)$ je

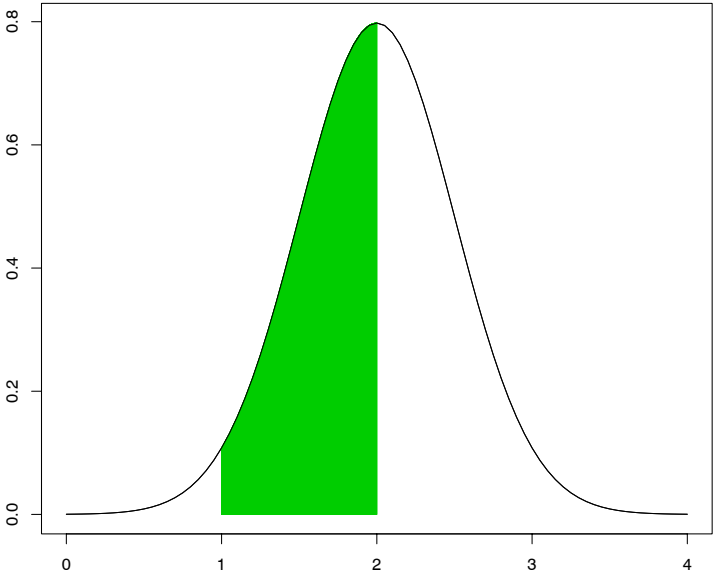
$$EX = \int_{-\infty}^{\infty} t \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}t = \mu.$$

Dok varijanca iznosi upravo

$$\mathrm{var}X = \sigma^2.$$



Normalna gustoća s parametrima 2 i $1/\sqrt{2}$. Graf funkcije gustoće simetričan je oko očekivanja $\mu = 2$ i vrlo brzo pada k nuli.



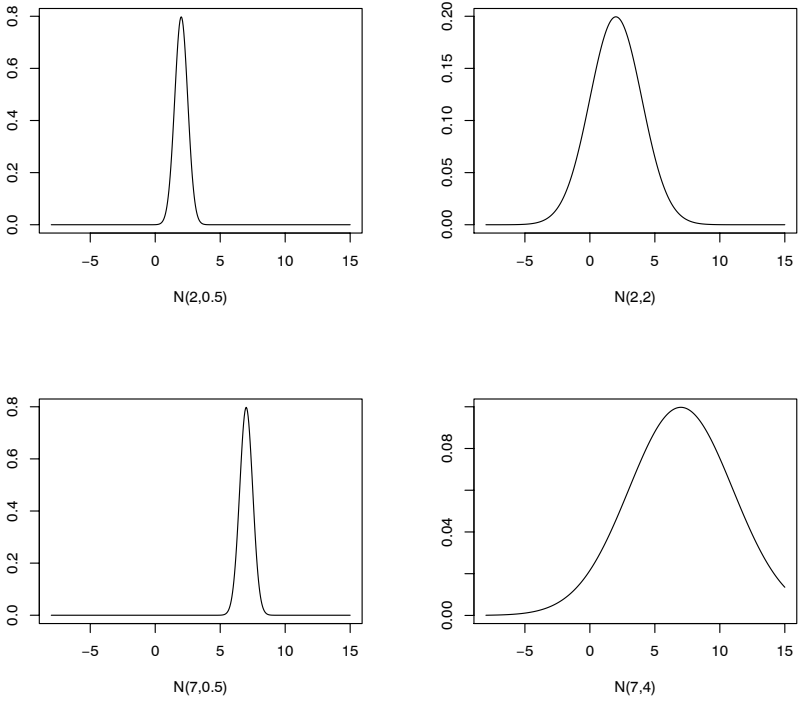
Normalna gustoća s parametrima 2 i $1/\sqrt{2}$. Zeleno osjenčana površina daje vjerojatnost da je X izmedju 1 i 2.

Ako je $X \sim N(\mu, \sigma^2)$, za zadane $a, b \in \mathbb{R}$, $a \neq 0$ vrijedi

$$Y = aX + b \sim N(a\mu + b, a^2\sigma^2).$$

Npr. ako je $a > 0$, za proizvoljne $c < d \in \mathbb{R}$, slijedi

$$\begin{aligned}
& P(c \leq Y \leq d) \\
&= P(c \leq aX + b \leq d) \\
&= P(\frac{c-b}{a} \leq X \leq \frac{d-b}{a}) \\
&= \int_{\frac{c-b}{a}}^{\frac{c-b}{a}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \mathsf{d}t \\
&= \text{ zamijena } u = at + b \\
&= \int_c^d \frac{1}{a\sigma\sqrt{2\pi}} e^{-\frac{(u-a\mu-b)^2}{2a^2\sigma^2}} \mathsf{d}u
\end{aligned}$$



Normalne gustoće s raznim parametrima.

Posebno ako je $X \sim N(\mu, \sigma^2)$, tada

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Ovakva transformacija sl. varijable X se zove **standardizacija**, a sl. varijabla $Z \sim N(0, 1)$ ima tzv. **standardnu (ili jediničnu) normalnu razdiobu**. Za nju definiramo i funkciju

$$\Phi(x) = P(Z \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t.$$

Ovu funkciju nije lako izračunati za različite x, no zato njene vrijednosti tabeliramo.

Vrijednosti funkcije Φ su nam od velike koristi za sve normalne sl. varijable. Naime za sve $X \sim N(\mu, \sigma^2)$ i $a < b$ nalazimo

$$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}),$$

a za standardiziranu sl. varijablu $Z = (X-\mu)/\sigma$, to je jednako

$$P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}).$$

Iz osnovnih pravila za računanje vjerojatnosti (ali i integrala) za $a < b$ slijedi

$$\begin{aligned} & P(a \leq Z \leq b) \\ &= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t \\ &= P(Z \leq b) - P(Z \leq a) \end{aligned}$$

Tako da za $X \sim N(\mu, \sigma^2)$ vrijedi

$$P(a \leq X \leq b) = P(Z \leq \frac{b-\mu}{\sigma}) - P(Z \leq \frac{a-\mu}{\sigma}) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}).$$

Posebno uočite

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P(-2 \leq Z \leq 2) = \Phi(2) - \Phi(-2) \approx 0.9545,$$

a

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = P(-3 \leq Z \leq 3) = \Phi(3) - \Phi(-3) \approx 0.9973,$$

te

$$P(X > \mu + 1.65\sigma) = P(Z > 1.65) = 1 - \Phi(1.65) \approx 0.05.$$

No pitanje ostaje – zašto smo kao prvu neprekidnu razdiobu uveli baš normalnu? Za razliku od binomne, hipergeometrijske ili Poissonove, do nje nismo dospjeli preko nekakvog pokusa. Ipak normalna razdioba je daleko najvažnija razdioba u statistici. Postoje dva bitna razloga:

i) Naime mnoge pojave u prirodi su normalno distribuirane: npr. fizičke karakteristike u raznim populacijama biljaka i životinja.

ii) Postoji više matematičkih rezultata tzv. *centralnih graničnih teorema* koji pokazuju da se razne razdiobe mogu aproksimirati normalnom.

Zapravo se ii) može protumačiti kao razlog i za i).

Normalna aproksimacija

Kao prvu diskretnu razdiobu uveli smo binomnu razdiobu, tj. $X \sim B(n,p)$. Kao što znamo ova sl. varijabla ima očekivanje np i varijancu npq, pa nam sljedeća transformacija

$$\frac{X-np}{\sqrt{npq}}$$

daje sl. varijablu koja ima očekivanje 0 i varijancu 1. No to i dalje ostaje diskretna razdioba. Zbog toga je možda iznenadjujuće da za sve $0<p<1$

$$P\left(\frac{X-np}{\sqrt{npq}} \leq x\right) \approx P(Z \leq x),$$

gdje je $Z \sim N(0,1)$ za dovoljno velike n.

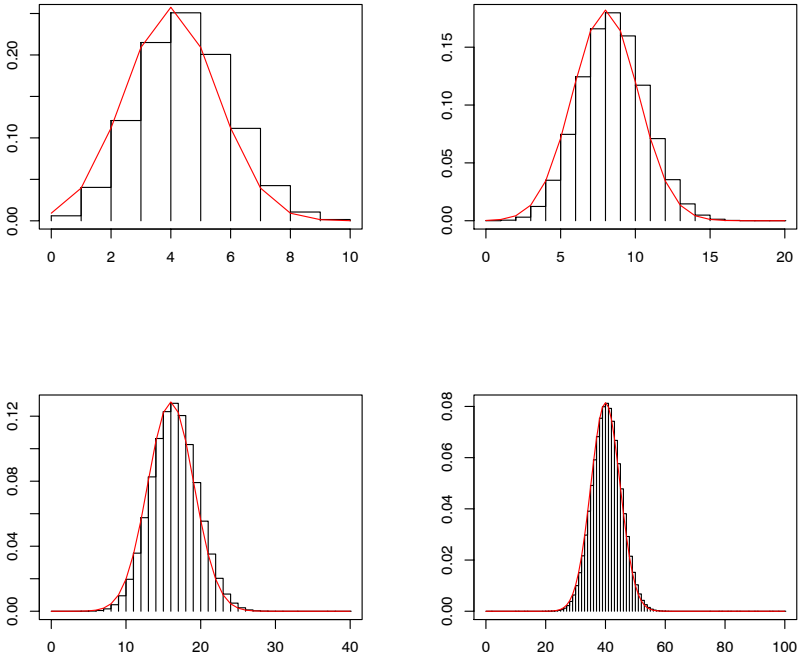
Zbog toga vrijedi

$$P(a \leq X \leq b) \approx P\left(\frac{a-np}{\sqrt{npq}} \leq Z \leq \frac{b-np}{\sqrt{npq}}\right).$$

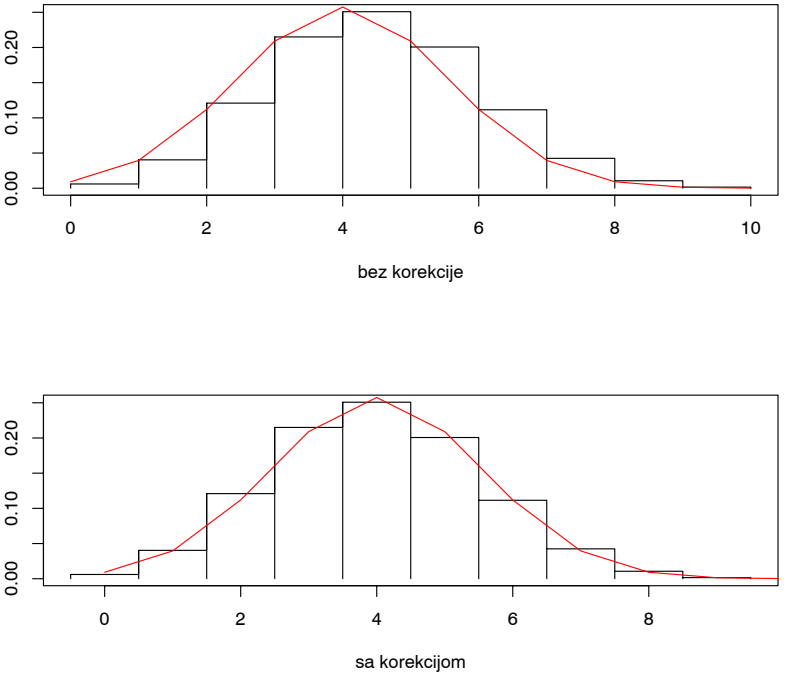
Ova aproksimacija je korisna i upotrebljava se kada $np \geq 5$ i $nq \geq 5$.

Kako je binomna razdioba diskretna, a standardna normalna neprekidna u praksi se često koristi i **korekcija po neprekidnosti** koja aproksimaciju čini još boljom

$$P(a \leq X \leq b) \approx P\left(\frac{a-1/2-np}{\sqrt{npq}} \leq Z \leq \frac{b+1/2-np}{\sqrt{npq}}\right).$$



Aproksimacija binomne razdiobe normalnom.



Aproksimacija binomne razdiobe normalnom sa i bez korekcije po neprekidnosti.

Prvi su ovu korisnu aproksimaciju uočili de Moivre i Laplace. No nakon toga je pokazano da ona vrijedi i puno općenitije. Naime binomna sl. varijabla je zapravo zbroj od n nezavisnih sl. varijabli koje sve imaju vrijednosti 0 ili 1, ovisno o ishodima pojedinačnih pokusa. Njena razdioba za velike n poprima isti oblik kao i normalna uz dobro odabranu varijancu i očekivanje. To vrijedi u vrlo generalnoj situaciji.

Centralni granični teorem

Sva numerička obilježja koja su rezultat puno malih i nepovezanih slučajnih utjecaja imat' će približno normalnu razdiobu.

χ^2 razdioba

Sl. varijabla X ima χ^2 **razdiobu s ν stupnjeva slobode** ako prima samo nenegativne vrijedosti i ima gustoću

$$f_X(t) = \frac{1}{2^{\nu/2}\Gamma\nu/2} t^{\nu/2-1} e^{-\frac{t}{2}}, \ t > 0.$$

Može se pokazati da vrijedi: $EX = \nu$, $\mathrm{var}X = 2\nu$.

Ako je $X \sim N(0,1)$, onda X^2 ima χ^2 razdiobu s 1 stupnjem slobode.

Uniformna razdioba

Sl. varijabla X ima **uniformnu razdiobu na intervalu** $[a, b]$, $a < b$, ako prima vrijedosti u intervalu $[a, b]$ i ima gustoću

$$f_X(t) = \frac{1}{b-a}\,,\ t \in [a, b].$$

Može se pokazati da vrijedi:

$$EX = \frac{a+b}{2}, \quad \mathrm{var}X = \frac{(b-a)^2}{12}.$$