

Анализа података у биологији

Марија Цупарић

Четврти час



Ако као резултат експеримента нека променљива може да узима различите вредности, које унапред нису познате, онда њу зовемо **случајна променљива** или **случајна величина**.

Примери

- (1) X - збир бројева на коцкицама приликом бацања две коцкице; скуп вредности 2,3,4,5,6,7,8,9,10,11,12 (коначно много вредности)
- (2) Y - број бацања коцкице док се не добије шестика; скуп вредности 1, 2, 3, 4, ... (пребројиво много вредности)
- (3) T - време првог позива између 7:30 и 8 часова стигао у полицијску станицу; скуп вредности интервал (7:30,8)
- (4) L - дужина извршавања компјутерског програма; скуп вредности је интервал $(0, t)$, где је t време за које ће се програм сигурно извршити

Којег су типа наведене случајне величине?



Када се ради са случајним величинама није довољно знати само које вредности може да узме, већ и са којим вероватноћама узима те вредности.

Дефиниција: Нека је X случајна величина. Функција F одређена са

$$F(x) = P\{X \leq x\} = P\{X \in (-\infty, x]\}$$

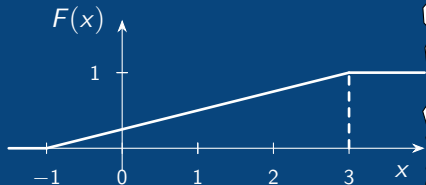
зове се функција расподеле случајне величине X .

Особине функције расподеле:

- 1) $0 \leq F(x) \leq 1$ (јер је дефинисана као вероватноћа);
- 2) $F(x)$ је неоппадајућа функција, тј. за $x_1 < x_2 \implies F(x_1) \leq F(x_2)$;
- 3) $F(x)$ је непрекидна сдесна и има граничну вредност слева у свакој тачки;
- 4) Ако се све могуће вредности које узима случајна величина X налазе у интервалу (a, b) , онда је $F(x) = 0$ за $x \leq a$, а $F(x) = 1$ за $x \geq b$.

Пример Случајна величина X задата је својом функцијом расподеле

$$F(x) = \begin{cases} 0, & x \leq -1, \\ \frac{x}{4} + \frac{1}{4}, & -1 < x \leq 3, \\ 1, & x > 3. \end{cases}$$



Одредити вероватноћу да ће као резултат експеримента X узети вредности из интервала $(0, 2)$.

$$\begin{aligned} \Rightarrow P\{0 < X < 2\} &= P\{X < 2\} - P\{X \leq 0\} \\ &= P\{X \leq 2\} - P\{X \leq 0\} \\ &= F(2) - F(0) \\ &= \frac{2}{4} + \frac{1}{4} - \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

- ★ Вредности дискретне случајне величине су изоловане тачке $x_1, x_2, \dots$ међусобно раздвојене интервалима
- ★ За сваку вредност x_i догађај да X узме вредност x_i има одређену вероватноћу која се обележава са: $p_i = p(x_i) = P\{X = x_i\}$.

Пар $\{(x_i, p_i), i = 1, 2, \dots\}$ називамо **закон расподеле** (вероватноћа) дискретне случајне величине X .

- ★ други запис: $X : \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$ или $X : \begin{pmatrix} x_1 & x_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix}$
- ★ $p_i \geq 0, i = 1, 2, \dots$ и $\sum_{i=1}^{\infty} p_i = 1$
- ★ Ако је позната расподела неке случајне величине X , онда за сваки скуп B може се одредити вероватноћа $P\{X \in B\}$.



Пример. Бацају се три новчића. Одредити вероватноћу да буде мање од 2 писма.

X - број писама који су пали

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

$$\Rightarrow P\{X < 2\} = P\{X = 0\} + P\{X = 1\} = \frac{1}{8} + \frac{3}{8} = \frac{4}{8}$$

Није исто што и $P\{X \leq 2\}$!



Пример. Следећа табела показује расподелу број људи који траже не-
 потребан преглед у хитној помоћи на дневном нивоу у некој малој бол-
 ници:

x	0	1	2	3	4	≥ 5
$P\{X = x\}$	0.01	0.1	0.3	0.4	0.1	?

Одредити вероватноћу да ће неког дана се појавити више од 3 особе
 са таквим захтевом?

Прво треба одредити недостајућу вероватноћу. Укупна сума свих веро-
 ватноћа треба да буде 1. Према томе

$$P\{X \geq 5\} = 1 - 0.01 - 0.1 - 0.3 - 0.4 - 0.1 = 0.09.$$

Тражена вероватноћа је

$$P\{X > 3\} = P\{X = 4\} + P\{X \geq 5\} = 0.1 + 0.09 = 0.19.$$

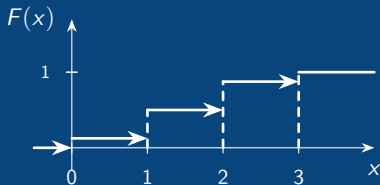


Дефиниција: Функција расподеле дискретне случајне величине X је функција F_X одређена са

$$F_X(x) = P\{X \leq x\} = \sum_{x_i \leq x} P\{X = x_i\}.$$

Пример Одредити функцију расподеле у примеру са три новчица.

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix} \Rightarrow F_X(x) = P\{X \leq x\} = \begin{cases} 0, & x < 0, \\ \frac{1}{8}, & 0 \leq x < 1, \\ \frac{4}{8}, & 1 \leq x < 2, \\ \frac{7}{8}, & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$



Дефиниција: Случајне величине X и Y су независне уколико је сваки догађај везан за X независан од догађаја везаног за Y , односно

$$P\{X = x_i | Y = y_j\} = P\{X = x_i\}$$

за свако x и y .



Поред расподеле случајних величина, битне су и њихове нумеричке карактеристике.

Нумеричке карактеристике случајне величине су бројеви којима се описују нека својства расподеле вероватноће.

Најзначајније су математичко очекивање и дисперзија.

Да ли смо се већ упознали са њима?



Математичко очекивање дискретне случајне величине

Математичко очекивање или очекивана вредност EX случајне величине X је дугорочни теоријски просек вредности, односно средња вредност, за X .

Пример Баца се коцкица n пута и X је број добијен на њој.
 Нека су добијене вредности: 1, 3, 2, 5, 2, 1, 1, 6, 5, 4, 2, 3, 6, 4...
 Ако после сваког бацања рачунамо дотадаши просек добијамо низ просека:
 1, 2, 2, 2.75, 2.6, 2.33, 2.14, 2.63, 2.89, 3.0, 2.91, 2.92, 3.15, 3.21...

Дефиниција: Нека дискретна случајна величина X има закон расподеле $X : \begin{pmatrix} x_1 & x_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix}$. Тада је

$$E(X) = \sum_{i=1}^{\infty} x_i p_i.$$

Пример Теоријски, очекивање у претходном примеру је:

$$EX = \sum_{k=1}^6 k \frac{1}{6} = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = 3.5.$$

Ако бацамо коцкицу (велики број) n пута, приближно у једној шестини од n бацања добићемо 1, исто важи и за остале бројеве.

$$\frac{\frac{n}{6} \cdot 1 + \frac{n}{6} \cdot 2 + \frac{n}{6} \cdot 3 + \frac{n}{6} \cdot 4 + \frac{n}{6} \cdot 5 + \frac{n}{6} \cdot 6}{n}$$

- ★ Математичко очекивање случајне променљиве X је **број**
- ★ Ако X узима коначно много вредности онда очекивање увек постоји (каначан је број), ако узима пребројиво много вредности сума на десној страни може бити и ∞ .



Пример Осигуравајућа компаније нуди осигурање за случај пожара, торнада, поплаве и крађе. Након истраживања, закључили су да су вероватноће да се ове незгоде десе 0.1, 0.01, 0.05 и 0.14 редом. Компаније исплаћује 100, 5000, 1000 и 75 евра, редом, осигураницима којима се догоде ове незгоде. Премија осигурања је 150 евра годишње. Да ли осигуравајућа компанија може да очекује профит од овакве полисе?

X - профит компаније по полиси, $X \in \{50, -4850, -850, 75, 150\}$

$$X : \begin{pmatrix} -4850 & -850 & 50 & 75 & 150 \\ 0.01 & 0.05 & 0.1 & 0.14 & 0.7 \end{pmatrix}.$$

$$\implies EX = -4850 \cdot 0.01 - 850 \cdot 0.05 + 50 \cdot 0.1 + 75 \cdot 0.14 + 150 \cdot 0.7 = 29.50.$$

С обзиром да је очекивање позитивно, значи да је очекивана добит од полисе позитивна. Односно оваква полиса се исплати компанији.



Често је потребно одредити очекивану вредност неке функције случајне величине. На пример, одредити очекивање од X^2 или $X + 2$.

Нека је $g(X)$ нека функција од X . Тада је $g(X)$ такође случајна величина која у општем случају има расподелу

$$g(X) : \begin{pmatrix} g(x_1) & g(x_2) & \dots & g(x_n) \\ p_1 & p_2 & \dots & p_n \end{pmatrix}.$$

Њено математичко очекивање је одређено са

$$E(g(X)) = g(x_1)p_1 + g(x_2)p_2 + \dots + g(x_n)p_n = \sum_{i=1}^n g(x_i)p_i$$

Особине математичког очекивања:

- 1) $E(c) = c$, где је c константа;
- 2) $E(aX + b) = aE(X) + b$, где су a и b константе (линеарност);
- 3) $E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$;
- 4) ако је $X \geq c$, онда је $EX \geq c$.



Пример Нека је

$$X : \begin{pmatrix} -0.01 & 0.01 \\ 0.5 & 0.5 \end{pmatrix}, \quad Y : \begin{pmatrix} -100 & 100 \\ 0.5 & 0.5 \end{pmatrix} \implies EX = 0, \quad EY = 0$$

Дакле, на основу очекивања ништа не можемо закључити о величинама.

Дефиниција: Нека дискретна случајна величина X има закон расподеле

$X : \begin{pmatrix} x_1 & x_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix}$. Дисперзија случајне величине X је број

$$D(X) = E(X - E(X))^2 = \sum_{i=1}^{\infty} (x_i - E(X))^2 p_i.$$

Пример За случајне величине из претходног примера важи

$$DX = (-0.01 - 0)^2 \cdot 0.5 + (0.01 - 0)^2 \cdot 0.5 = 0.0001,$$

$$DY = (-100 - 0)^2 \cdot 0.5 + (100 - 0)^2 \cdot 0.5 = 10000.$$

Особине дисперзије:

$$1) DX = E(X)^2 - (E(X))^2 \text{ јер}$$

$$\begin{aligned} E(X - E(X))^2 &= E(X^2 - 2XE(X) + (E(X))^2) \\ &= E(X^2) - 2E(XE(X)) + E((E(X))^2) \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

$$2) D(X) \geq 0;$$

$$3) D(c) = 0, \text{ где је } c \text{ константа};$$

$$4) D(aX + b) = a^2 D(X), \text{ где су } a \text{ и } b \text{ константе.}$$

Величина $\sigma(X) = \sqrt{D(X)}$ се назива стандардно одступање (стандардна девијација).

Уз стандардно одступање могу да иду мерне јединице (нпр. валута), док уз дисперзију не знамо како да их интерпретирамо.

Ако су случајне величине X и Y независне онда

$$\star E(XY) = EX \cdot EY$$

$$\star D(X + Y) = DX + DY$$

$$\star D(X - Y) = DX + DY$$



Пример

- ★ Индикатор да јединка припада угроженој врсти:

$$I = \begin{cases} 1, & \text{врста је угрожена} \\ 0, & \text{иначе} \end{cases}$$

- ★ Индикатор да је биодиверзитет низак (нпр. број врста мањи од 10):

$$I = \begin{cases} 1, & S < 10 \\ 0, & \text{иначе} \end{cases}$$

Нека је A неки догађај који се реализије са вероватноћом $p, p \in (0, 1)$.

Случајна величина са законом расподеле

$$I_A : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix} \text{ или } p(x; p) = P\{X = x\} = (1-p)^{1-x} p^x, x \in \{0, 1\}.$$

се назива индикатор догађаја A .

$$\star E I_A = p, D I_A = p(1-p)$$

- ★ R: `rbinom(sample size, 1, probaility)`



Пример

- ★ У узорку из једног језера, вероватноћа да случајно изабрана риба садржи повишен ниво загађујућих супстанци износи 0.1. Ако се изабере 8 риба, X је број таквих јединки.
- ★ У истраживању птица, вероватноћа да једна јединка преживи зиму износи 0.8. Посматра се 6 јединки независно, X је број јединки које су преживеле зиму.

Имамо n експеримената и у сваком посматрамо да ли се догодио одређени догађај који називамо успехом. Експерименти су међусобно независни и вероватноћа успеха у сваком од њих је p . За случајну променљиву која представља број успеха у n оваквих експеримената кажемо да има биномну расподелу с параметрима n и p . Закон расподеле је

$$p(x; n, p) = P\{X = x\} = \binom{n}{x} p^x (1 - p)^{n-x}, x \in \{0, 1, \dots, n\}.$$

- ★ $EX = np, DX = np(1 - p)$
- ★ **R:** `rbinom(sample size, n, probability)` - даје случајан узорак из $\mathcal{B}(n, p)$ расподеле



Пример У једном језеру вероватноћа да случајно изабрана риба има повишен ниво загађујућих супстанци износи 0.1. Насумично је изабрано 8 риба.

- а) Колика је вероватноћа да тачно 3 рибе имају повишен ниво загађујућих супстанци?
- б) Колика је вероватноћа да највише 3 рибе имају повишен ниво загађујућих супстанци?
- в) Колика је вероватноћа да бар 4 рибе имају повишен ниво загађујућих супстанци?
- г) Одредити очекивану вредност и дисперзију броја таквих риба.

X — број риба са повишеним нивоом загађујућих супстанци

а) $P\{X = 3\} = \binom{8}{3} 0.1^3 \cdot 0.9^5 = 0.033$

б) $P\{X \leq 3\} = \sum_{k=0}^3 \binom{8}{k} 0.1^k \cdot 0.9^{8-k} = 0.995$

в) $P\{X \geq 4\} = 1 - P\{X < 4\} = 1 - P\{X \leq 3\} = 1 - 0.995 = 0.005$

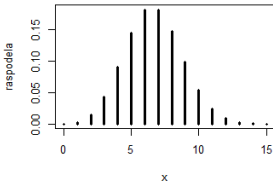
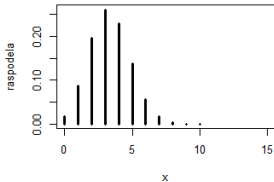
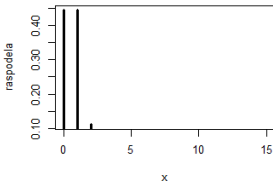
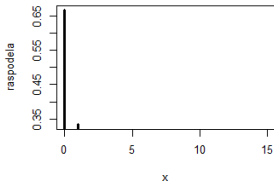
г) $EX = 8 \cdot 0.1 = 0.8, DX = 8 \cdot 0.1 \cdot 0.9 = 0.72$



```
> dbinom(3,8,0.1) #resenje dela a)
> pbinom(3,8,0.1) #resenje dela b)
```

R

```
> plot(0:20, dbinom(0:20, size=20, prob=1/3),xlim=c(0,15),
      type='h', axes=TRUE,xlab="x", ylab="raspodela", lwd=3)
```



Геометријска случајна величина: $X \in \mathcal{G}(p)$

Пример

- ★ Предатор у природи (нпр. птица) хвата плен (нпр. инсекте). Сваки покушај хватања је независан, а вероватноћа успешног хватања износи 0.8. Посматра се број успешних хватања пре него што предатор први пут не успе да ухвати плен.
- ★ Број гађања у мету пре него што се погоди

У свакој етапи имамо два исхода: „успех” и „неуспех”. Етапе су међусобно независне и вероватноћа успеха p је иста у свакој етапи. Бројимо експерименте док се не постигне први успех. Закон расподеле је

$$p(x; p) = P\{X = x\} = (1 - p)^{x-1}p, x = 1, 2, 3, \dots$$

- ★ $EX = \frac{1}{p}, DX = \frac{1-p}{p^2}$
- ★ може дефинисати и као број неуспеха док се не појави први успех и тада је закон расподеле

$$p(x; p) = P\{Y = x\} = P\{X = x + 1\} = (1 - p)^x p, x = 0, 1, 2, \dots$$

- ★ **R:** `rgeom(sample size, probability)` генерисање величине Y

Пример Птица предатор лови инсекте. Сваки покушај хватања је независан, а вероватноћа успешног хватања износи 0.8.

- Колика је вероватноћа да први инсект који птица није успела да ухвати буде четврти по реду?
- Колика је вероватноћа да је птица успешно ухватила мање од 3 инсекта пре првог неуспеха?
- Одредити очекивану вредност и дисперзију броја успешно ухваћених инсеката пре првог неуспеха.

Y — број успешно ухваћених инсеката пре првог промашаја, $Y \in \mathcal{G}(0.2)$

- $P\{Y = 3\} = 0.8^3 \cdot 0.2 = 0.1024$
- $P\{Y < 3\} = P\{Y = 0\} + P\{Y = 1\} + P\{Y = 2\}$

$$= 0.2 + 0.2 \cdot 0.8 + 0.2 \cdot 0.8^2 = 0.488$$
- $EY = E(X - 1) = EX - 1 = \frac{1}{0.2} - 1 = 4,$
 $DY = D(X - 1) = DX = \frac{1 - 0.2}{0.2^2} = 20$



```
> dgeom(3,0.2) #resenje dela a)
> pgeom(2,0.2) #resenje dela b)
```

Интерпретација резултата у претходном примеру:

- ★ Очекивана вредност је 4, што значи да очекујемо да ће птица ухватити 4 инсекта пре него што први пут промаши
- ★ Дисперзија је 20, односно стандардно одступање $\sqrt{20} \approx 4.47$, што је велика вредност.

Иако просечно птица ухвати 4 инсекта, број успеха може доста да варира. Велика дисперзија проистиче из мале вероватноће успеха.

То значи да у пракси, у једном низу хватања, птица може ухватити много мање или много више инсеката од 4 пре него што први пут промаши.



R

```
> plot(0:15, dgeom(0:15, prob = 1/3), xlim = c(0, 15),  
      type = 'h', axes = TRUE, xlab = "x",  
      ylab="raspodela", lwd=3)
```

