

Statystyka i modelowanie w naukach o środowisku

Czym jest statystyka?

Statystyka (niem. *Statistik*, „badanie faktów i osób publicznych”, z łac. *statisticus*, „polityczny, dot. polityki”, od *status*, „państwo, stan”) – nauka, której przedmiotem zainteresowania są metody pozyskiwania i prezentacji, a przede wszystkim analizy danych opisujących zjawiska, w tym masowe.

zebranie dużej ilości danych
prezentacja
analiza
interpretacja

Wykład 1

Elementy statystyki opisowej

Pozyskiwanie danych

Badanie statystyczne, w ramach którego dokonuje się obserwacji statystycznej (pomiarów lub zliczania)

Prezentacja danych

Przedstawienie danych w postaci ułatwiającej ich analizę i ocenę:

tabele

wykresy

Analiza i interpretacja danych

Analiza za pomocą metod opisu statystycznego

Statystyka matematyczna i metody rachunku prawdopodobieństwa

Badanie statystyczne

Dotyczy zawsze pełnych zbiorowości, których elementami są obiekty materialne lub zjawiska.

Zbiorowość statystyczna – populacja generalna lub zbiorowość generalna

Populacja generalna skończona – zbiorowość mieszkańców Polski, zbiorowość indywidualnych gospodarstw rolnych w pewnym województwie

Zbiór elementów nieskończony – zbiorowość rzutów monetą, zbiorowość możliwych wyników pomiarów wytrzymałości materiału

Populacja generalna

Elementy populacji posiadają cechy statystyczne - właściwości, które podlegają obserwacji statystycznej (populacja Polski – płeć, wiek, wzrost, wykształcenie)

Cechy statystyczne – mierzalne (wzrost, waga) i niemierzalne (płeć, kolor oczu)

Elementy populacji generalnej zwykle posiadają różne wartości cechy statystycznej – rozkład cechy (cech)

Rodzaje badań statystycznych

Pełne (wszystkie elementy zbiorowości generalnej)

Częściowe (część elementów zbiorowości generalnej - próbę)

Badania mogą być obciążone błędami podczas organizacji badania, pomiaru cech i przetwarzania wyników.

Badania częściowe są narażone na błędy wyboru próby.

Wnioskowanie statystyczne

Statystyka matematyczna - badania częściowe

Uogólnienie wyników na całą populację elementów i oszacowanie wielkości popełnionych błędów – wnioskowanie statystyczne:

Estymacja (szacowanie) nieznanych wartości parametrów rozkładu cechy w populacji

Sprawdzanie słuszności hipotez dotyczących wartości parametrów rozkładów cechy w populacji lub postaci rozkładu

Losowy dobór próby

Znalezienie się poszczególnych elementów populacji w próbie jest losowe:

- Każda jednostka populacji generalnej ma dodatnie znane prawdopodobieństwo znalezienia się w próbie

- Istnieje możliwość ustalenia prawdopodobieństwa znalezienia się w próbie dla każdego zespołu elementów populacji

Próba losowa – próba otrzymana w wyniku doboru losowego

Losowanie proste

Wszystkie elementy populacji mają jednakowe prawdopodobieństwo dostania się do próby i prawdopodobieństwo to nie zmienia się w czasie (losowanie ze zwracaniem, losowanie niezależne)

Populacja nieskończona – przeprowadzenie serii n niezależnych eksperymentów z zachowaniem identycznych warunków

Metody opisu statystycznego

Metody służące do przedstawienia charakterystyki zbioru badanej cechy, czyli wyznaczenie liczbowych parametrów określających badany zbiór danych

Punkt wyjścia do wnioskowania, gdy zbiór jest próbą losową

Zamknięte badanie, gdy analiza dotyczy pełnej zbiorowości generalnej

Empiryczny rozkład cechy

Przyporządkowanie uszeregowanym rosnąco wartościom, przyjmowanym przez cechę, odpowiednio zdefiniowanych częstości ich występowania

Empiryczny rozkład cechy - przykład

Sprawdzono 20 stron maszynopisu znajdując na nich następujące liczby błędów 0, 3, 1, 1, 2, 2, 0, 0, 3, 5, 0, 1, 2, 2, 1, 1, 0, 1, 1, 1

Zbiorowość – 20 stron maszynopisu

Cecha – liczba błędów na stronie

Cecha skokowa

Empiryczny rozkład cechy - przykład

liczba błędów x_i	liczba stron n_i	częstość stron w_i
0	5	0,25
1	8	0,4
2	4	0,2
3	2	0,1
4	0	0
5	1	0,05
suma	20	1

Wartości rzeczywiste cechy

Cechy ciągłe

Przedziały klasowe

Rozpiętość przedziału $h_i = x_{1i} - x_{0i}$

Środek przedziału $x_i = (x_{0i} + x_{1i})/2$

Liczba przedziałów, zwykle $5 \leq k \leq 20$, $k \leq 5 \log n$, n – liczebność zbiorowości

Szereg rozdzielczy, histogram

$x_1, \dots, x_n$

rozstęp cechy $x_{\max} - x_{\min}$

liczba pomiarów n	liczba klas k
30-60	6-8
60-100	7-10
100-200	9-12
200-500	11-17
500-1500	16-25

Wartości rzeczywiste cechy - przykład

Zmierzono czas obsługi przy kasie sklepowej 25 losowo
wybranych klientów, uzyskując dane (czas w s): 15, 37, 34, 9, 61,
24, 56, 52, 6, 35, 21, 46, 86, 40, 74, 39, 48, 55, 73, 92, 43, 78, 67, 30,
29.

Wartości rzeczywiste cechy - przykład

czas obsługi klienta (s) y_0i-y_1i	liczba klientów ni	częstość klientów
0-20	3	0,12
20-40	9	0,36
40-60	6	0,24
60-80	5	0,2
80-100	2	0,08
suma	25	1

Dystrybuanta empiryczna

- Przedstawienie empirycznego rozkładu cechy za pomocą skumulowanych częstości względnych
- Dystrybuanta empiryczna jest funkcją niemalejącą, ograniczoną w przedziale $<0,1>$

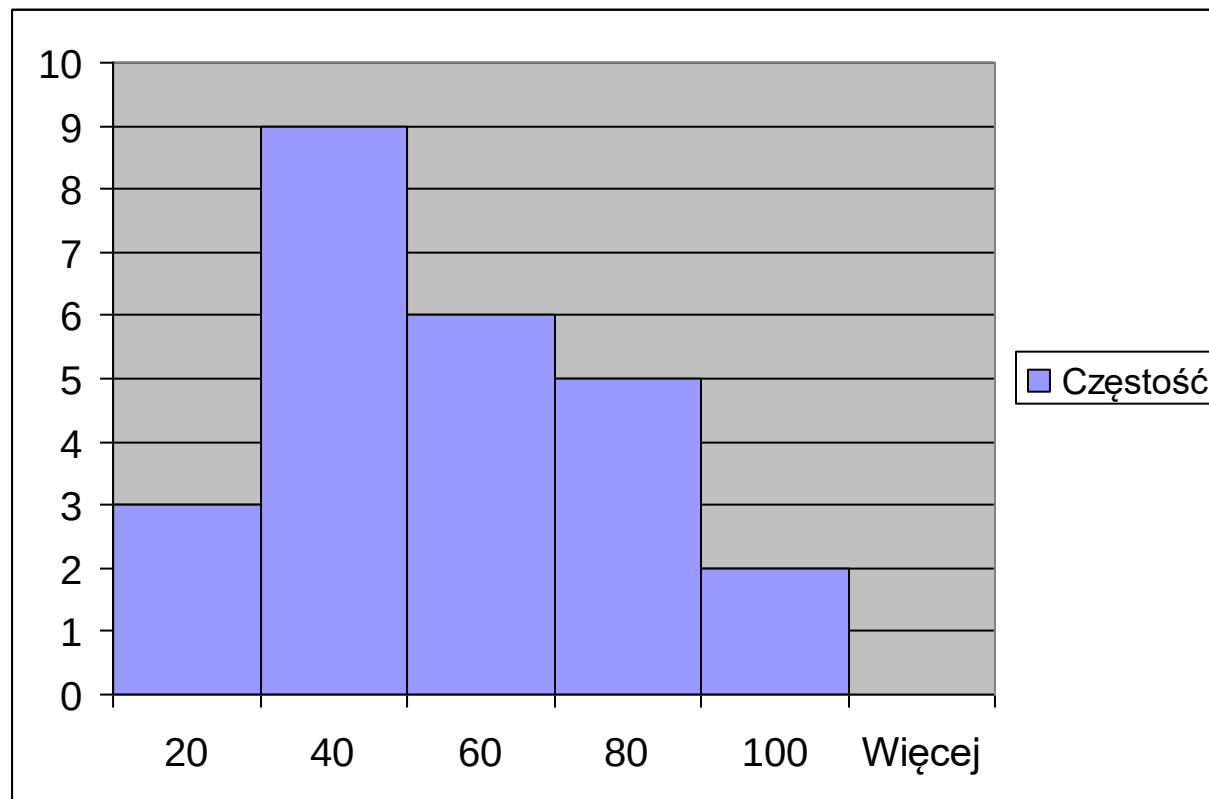
$$F(x) = \left\{ \begin{array}{ll} 0, & x < x_1 \\ \sum_{s=1}^i w_s, & x_i \leq x < x_{i+1} \\ 1, & x \geq x_k \end{array} \right\}$$

Dystrybuanta empiryczna - przykład

liczba błędów x_i	Skumulowana liczba stron n_i	dystrybuanta empiryczna $F_n(x_i)$
0	5	0,25
1	13	0,65
2	17	0,85
3	19	0,95
4	19	0,95
5	20	1

$n(2) = 17$, na 17-tu stronach było nie więcej niż po 2 błędy, $F(2)=0,85$, stanowiły one 85% wszystkich stron

Histogram



Miary położenia rozkładu

Średnia arytmetyczna

Mediana rozkładu empirycznego

Wartość modalna (moda, dominanta)

Miary położenia rozkładu

Średnia arytmetyczna

dla liczb $x_1, \dots, x_n$

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

Miary położenia rozkładu

Mediana rozkładu empirycznego – wartość cechy, dla której co najmniej połowa jednostek zbiorowości ma wartość cechy nie większą od niej i równocześnie co najmniej połowa jednostek ma wartość cechy nie mniejszą od tej wartości

$$m_e = \begin{cases} x_{\frac{n+1}{2}}, & \text{gdy } n \text{ nieparzyste} \\ \frac{1}{2} \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right), & \text{gdy } n \text{ parzyste} \end{cases}$$

Miary położenia rozkładu

Mediana rozkładu empirycznego – przykład

Z populacji generalnej pobrano $n=50$ -elementową próbkę i przebadano ze względu na cechę X . Wyznacz medianę.

3,6	5,6	4,6	5,9	4,7
5,0	3,5	5,1	4,2	6,4
4,0	5,4	4,7	6,4	5,1
4,7	5,2	3,0	5,3	3,4
5,2	4,1	5,5	4,5	5,2
5,9	5,0	6,1	4,9	6,2
4,5	3,1	3,8	4,0	4,4
5,3	5,8	4,9	5,2	4,3
5,5	4,8	5,6	3,3	5,8
3,9	4,4	6,1	5,4	3,7

Miary położenia rozkładu

Mediana rozkładu empirycznego – przykład

Porządkujemy próbkę niemalejąco.

3,0	3,1	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0	4,0	4,1	4,2	4,3	4,4	4,4	4,5	4,5	4,6	4,7	4,7	4,7	4,8	4,9	4,9
5,0	5,0	5,1	5,1	5,2	5,2	5,2	5,2	5,3	5,3	5,4	5,4	5,5	5,5	5,6	5,6	5,8	5,8	5,9	5,9	6,1	6,1	6,2	6,4	6,4

$$x_{25}=4,9$$

$$x_{26}=5,0$$

$$m_e = \frac{4,9 + 5}{2} = 4,95$$

Miary położenia rozkładu

Wartość modalna (moda, dominanta)

dla liczb $x_1, \dots, x_n$ o powtarzających się wartościach, to najczęściej powtarzająca się wartość

Miary położenia rozkładu

Wartość modalna (moda, dominanta) – przykład

próbka I: 16, 13, 15, 17, 16, 16, 15, 14, 12, 17, 16, 18, 14, 15, 17, 16

próbka II: 27, 24, 28, 24, 25, 23, 29, 26, 29, 25

12	1
13	1
14	2
15	2
16	5
17	2
18	1

23	1
24	2
25	2
26	1
27	2
28	1
29	2

wartość modalna: próbka I - 16, próbka II - brak

Miary zróżnicowania cechy

- Wariancja

$$s^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

- Współczynnik zmienności

$$V = \frac{s}{\bar{x}}$$

Miary zróżnicowania rozkładu

Kwantyl rzędu p ($0 < p < 1$) w rozkładzie empirycznym – wartość cechy k_p , dla której – jako pierwszej – dystrybuanta empiryczna: $F_n(k_p) \geq p$

Kwantyl rzędu $1/2$ to inaczej mediana.

Kwantyle rzędu $1/4$, $2/4$, $3/4$ są inaczej nazywane kwartylami.

Kwantyle rzędu $1/5$, $2/5$, $3/5$, $4/5$ to inaczej kwintyle.

Kwantyle rzędu $1/10$, $2/10$, ..., $9/10$ to inaczej decyle.

Kwantyle rzędu $1/100$, $2/100$, ..., $99/100$ to inaczej percentyle.

Zmierzono inteligencję 20 osób:

74, 80, 80, 85, 92, 94, 97, 98, 98, 100, 101, 101, 104, 104, 106, 109, 112, 115, 128, 137

Kwantylem rzędu 0,25 (czyli pierwszym kwartylem) jest tutaj liczba 92, gdyż dokładnie pięć próbek (czyli $1/4$ z populacji 20 próbek) ma wartość mniejszą lub równą 92. Kwantylem rzędu 0,75 (czyli trzecim kwartylem) jest tu liczba 106.

Miary rozkładu

Wyznacz miary rozproszenia 25-elementowej uporządkowanej próbki:

2,15	2,18	2,19	2,27	2,29	2,31	2,33	2,41	2,43	2,47	2,52	2,54	2,55
2,61	2,62	2,73	2,73	2,75	2,81	2,85	2,97	3,00	3,01	3,08	3,11	

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 2,5964, \quad m_e = x_{13} = 2,55$$

$$s^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

$$Q_1 = x_7 = 2,33$$

$$Q_3 = x_{19} = 2,81$$

Zmienna losowa

- E – zbiór zdarzeń elementarnych danego doświadczenia
- Zmienna losowa – funkcja $X(e)$ przyporządkowująca każdemu zdarzeniu elementarnemu e ze zbioru E tylko jedną liczbę $X(e) = x$

Rozkład prawdopodobieństwa

- Rozkład zmiennej dyskretnej
- Rozkład zmiennej ciągłej

Funkcja prawdopodobieństwa

- Funkcja prawdopodobieństwa dla zmiennej skokowej X przyjmującej wartości x_i , $i=1,2,\dots$

$$P(X = x_i) = p_i$$

$$\sum_{i=1}^n p_i = 1$$

$$\sum_{i=1}^{\infty} p_i = 1$$

Funkcja gęstości prawdopodobieństwa

- Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej – funkcja $f(x)$, określona na zbiorze liczb rzeczywistych taka, że:
 - $f(x) \geq 0$

$$\int_a^b f(x)dx = P(a < X \leq b) \text{ dla } a < b$$

Parametry rozkładu

- Wartość oczekiwana

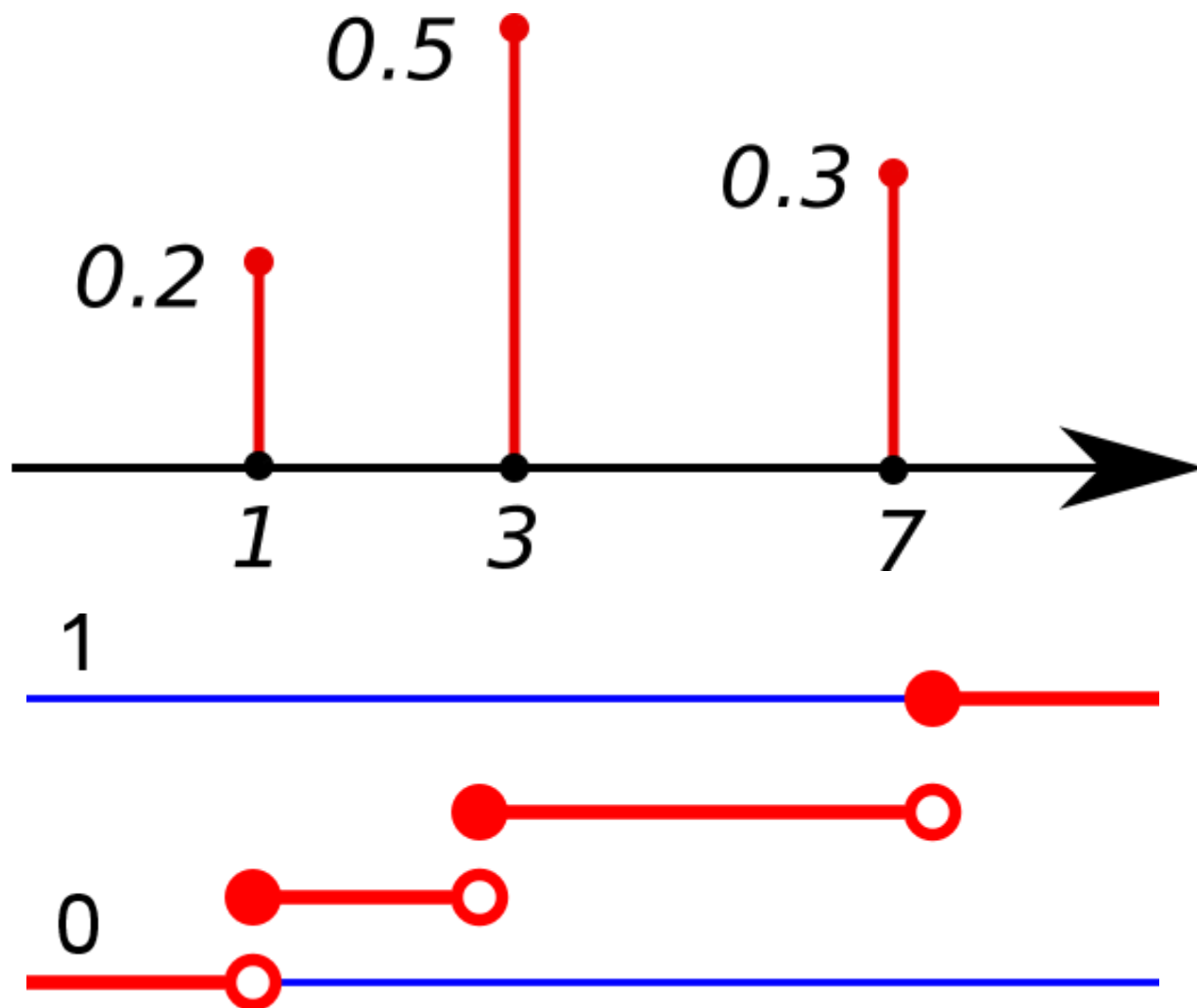
$$E(X) = \left\{ \begin{array}{ll} \sum_i x_i p_i & \text{dla zmiennej losowej skokowej} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{dla zmiennej losowej ciągłej} \end{array} \right\}$$

Parametry rozkładu

- Wariancja zmiennej losowej X

$$D^2(X) = \left\{ \begin{array}{ll} \sum_i [x_i - E(X)]^2 p_i & \text{dla zmiennej losowej skokowej} \\ \int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx & \text{dla zmiennej losowej ciągłej} \end{array} \right\}$$

Rozkład zmiennej dyskretnej

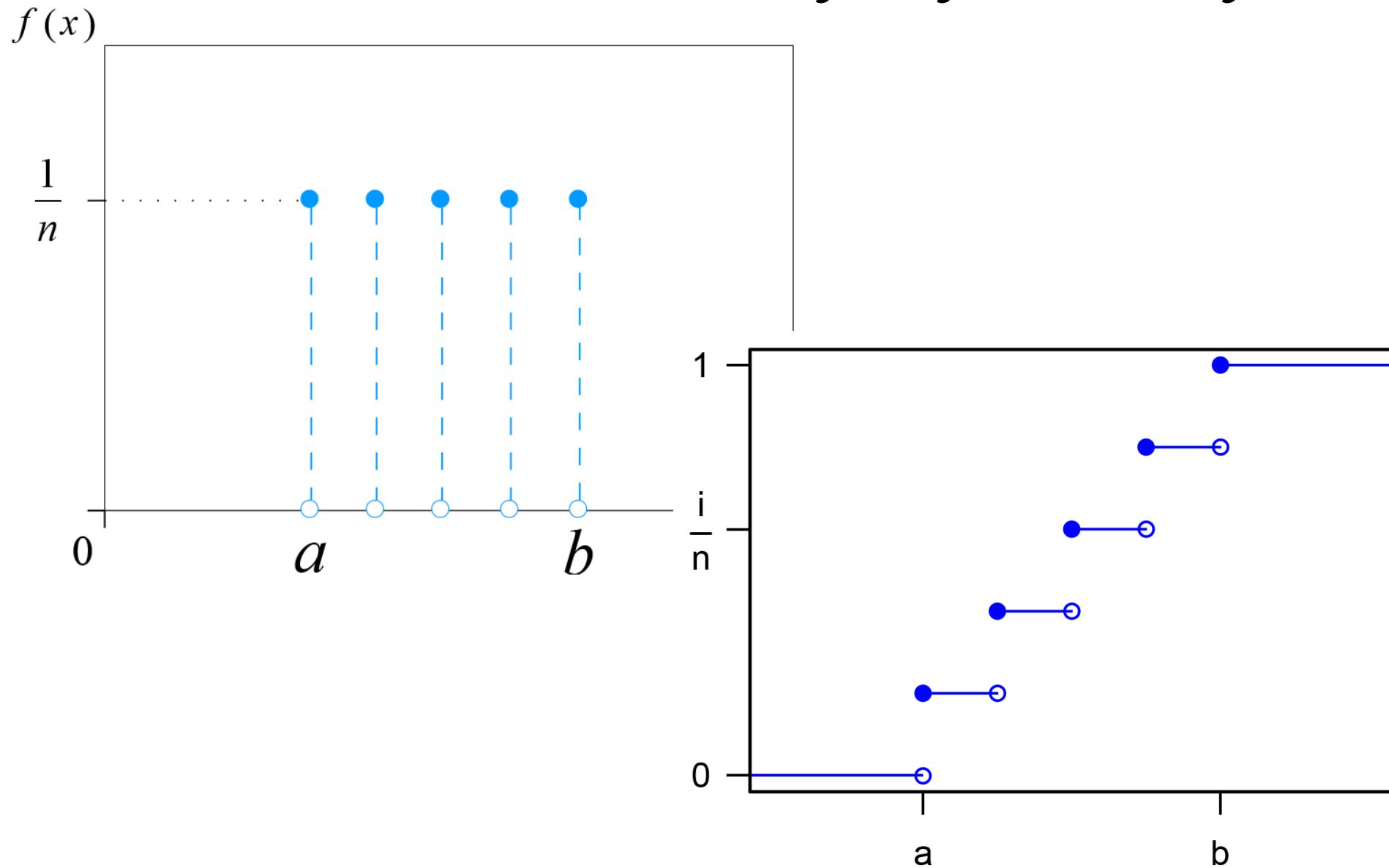


wartość oczekiwana i wariancja zmiennej losowej X

x_i	1	3	7
p_i	0.2	0.5	0.3

- $E(X)=3.80$
- $\text{Var}(X)=4.96$

Rozkład jednostajny dyskretny



Rozkład jednostajny dyskretny - przykład

- Rozkład wyników rzutu jedną kostką.
 - $E(X)=3.5$
 - $\text{Var}(X)=2.92$

Rozkład zero-jedynkowy

- Zmienna losowa przyjmuje wartość 1 z prawdopodobieństwem $0 < p < 1$ oraz wartość 0 z prawdopodobieństwem $q = 1 - p$
- Doświadczenie, którego rezultatem są dwa wzajemnie wykluczające się zdarzenia losowe:
 - Strzał do celu – trafienie lub jego brak
 - Ziarno wykiełkuje lub nie

Rozkład dwumianowy (Bernoulliego)

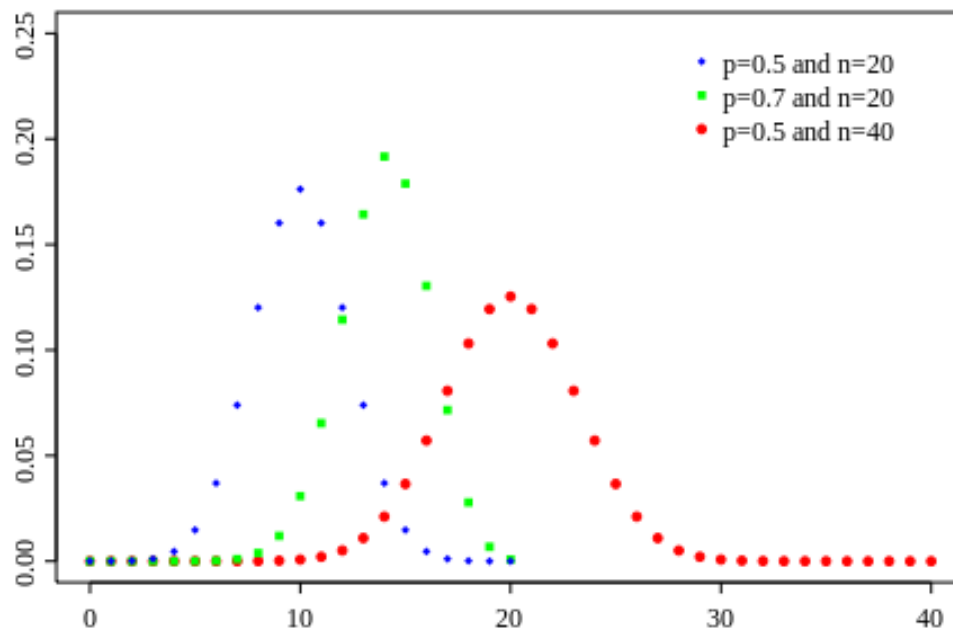
- Zmienna losowa ma rozkład dwumianowy, jeśli przyjmuje wartości $k=0, 1, 2, \dots, n$ z prawdopodobieństwem:

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

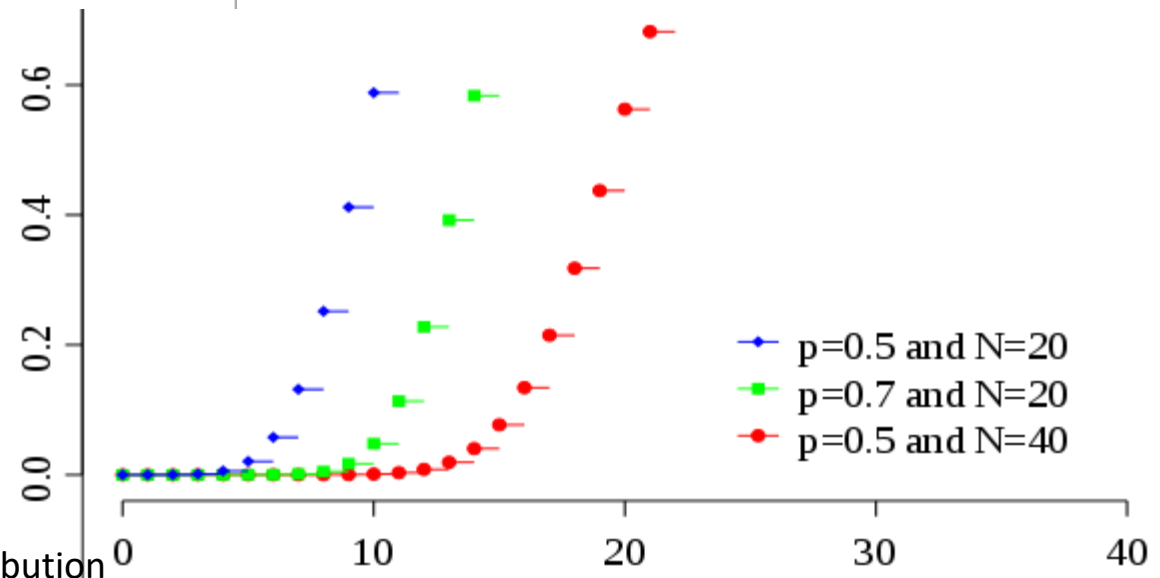
Rozkład dwumianowy

- Wynik doświadczenia – zdarzenie A z prawdopodobieństwem p lub przeciwne z prawdopodobieństwem $q=1-p$ (sukces – porażka)
- Doświadczenie powtórzone n - krotnie w sposób niezależny, czyli prawdopodobieństwo sukcesu jest w pojedynczych próbach stałe i równe p .

Rozkład dwumianowy



$$f(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$



Rozkład dwumianowy

- $E(k) = np$
- $Var(k) = np(1-p)$

Rozkład dwumianowy - przykład

Z partii nasion o sile kiełkowania 90% wylosowano 10 nasion.
Podaj rozkład liczby nasion, które wykiełkują.

Rozkład dwumianowy - przykład

Śmiertelność szczurów zarażonych pewną chorobą wynosi 85%. Obserwacji będzie podlegać grupa 20 zarażonych szczurów. Wykorzystując wartości dystrybuanty, oblicz prawdopodobieństwo zdarzenia, że na 20 zarażonych szczurów przeżyje:

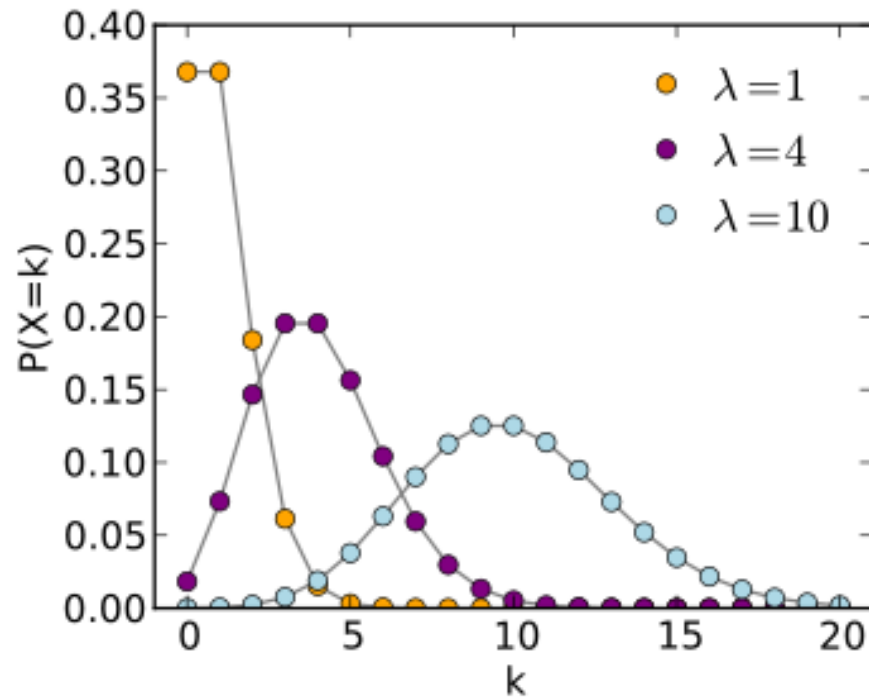
- a) więcej niż 12,
- b) co najmniej 5,
- c) dokładnie połowa,
- d) nie więcej niż połowa,
- e) mniej niż połowa.

Rozkład Poissona o parametrze $\lambda > 0$

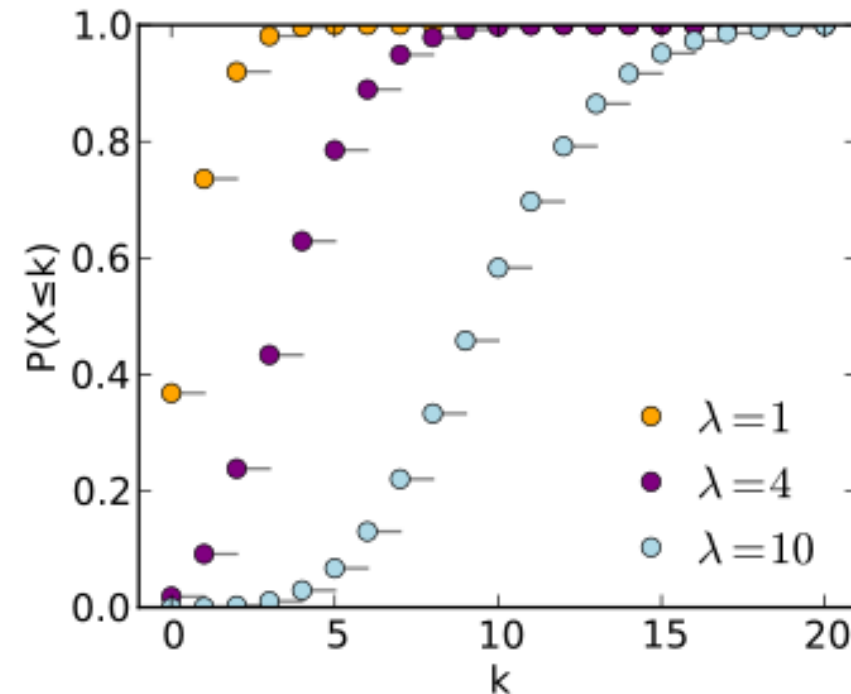
- Zmienna losowa X przyjmująca wartości $k = 0, 1, 2, \dots$, której funkcja prawdopodobieństwa:

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

Rozkład Poisson'a



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



Rozkład Poisson'a

- $E(k) = \lambda$
- $\text{Var}(k) = \lambda$

Rozkład Poissona - przykłady

- Liczba usterek w produkowanych urządzeniach
- Liczba zgłoszeń szkód ubezpieczeniowych w określonym czasie
- Liczba cząstek alfa emitowanych przez substancję radioaktywną w krótkim czasie
- Liczba błędów drukarskich na jednej stronie

Rozkład Poissona

- Dystrybuanta

$$F(x) = \sum_{k \leq x} \frac{\lambda^k}{k!} e^{-\lambda}, x \in R$$

- Wartość oczekiwana $E(X)=\lambda$
- Wariancja zmiennej losowej $D^2(X)=\lambda$

Rozkład normalny (Gaussa)

- Zmienna losowa X ma rozkład normalny o parametrach m oraz σ – $X:N(m, \sigma)$, jeśli funkcja gęstości:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, -\infty < x < \infty, \sigma > 0$$

Rozkład normalny - przykłady

- Waga lub wzrost osobników jednorodnych populacji ludzkich lub zwierzęcych
- Plon na jednakowych poletkach doświadczalnych
- Losowe błędy pomiarów

Rozkład normalny

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

$$E(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(x-m)^2}{2\sigma^2}} dx = m$$

$$D^2(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \sigma^2$$

