AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Wprowadzenie Statystyka opisowa

Statystyka

Dr inż. Janusz Majewski
Katedra Informatyki

Literatura

- Tadeusiewicz R., Izworski A., Majewski J.: „Biometria”, Skrypt AGH, Kraków 1993
- Armitage P.: „Metody statystyczne w badaniach medycznych”, PZWL, Warszawa 1978
- Greń J.: „Statystyka matematyczna. Modele i zadania”, PWN, Warszawa 1982
- Parker R.E.: „Wprowadzenie statystki dla biologów”, PWN, Warszawa 1978
- Żuk B.: „Biometria stosowana”, PWN, Warszawa 1989
- Blalock H.M.: „Statystyka dla socjologów”, PWN, Warszawa 1977
- Zieliński R, Zieliński W.: „Tablice statystyczne” PWN, 1990
- Aczel A.: „Statystyka w zarządzaniu”, PWN, Warszawa, 2007

Literatura

- Prezentacja wykorzystuje fragmenty książki: Amir D. Aczel „Statystyka w zarządzaniu”, PWN, 2007

Przygotowanie danych

Rysunek 1. Podział danych

Przygotowanie danych

Tabela 1. Przykład: Dane jakościowe/skala nominalna

Grupa krwi	Liczba pacjentów	Udział %
A	425	39,5%
B	180	16,7%
AB	84	7,8%
0	388	36,0%
Razem	**1077**	**100,0%**

Przygotowanie danych

Tabela 2. Przykład: Dane jakościowe/skala porządkowa

Stan migdałków	Liczba dzieci	Udział %
niepowiększone	516	36,9%
powiększone	589	42,1%
bardzo powiększone	293	21,0%
Razem	**1398**	**100,0%**

Przygotowanie danych

Tabela 3. Przykład: agregacja danych ilościowych

Wiek	Liczba pacjentów	Udział
25÷34	19	0,018
35÷44	116	0,087
45÷54	493	0,363
55÷64	545	0,401
65÷74	186	0,137
Razem	**1359**	**1,000**

Przygotowanie danych (dane ilościowe, szereg rozdzielczy)

Wiek pacjentek z nowotworem szyjki macicy w pewnym szpitalu w Algierii

wiek	liczba pacjentek	środek przedziału wiekowego
(A)	(B)	(C)
20-25	3	22,5
25-30	10	25,5
30-35	38	32,5
35-40	71	37,5
40-45	117	42,5
45-50	100	47,5
50-55	89	52,5
55-60	75	57,5
60-65	70	62,5
65-70	59	67,5
70-75	21	72,5
75-80	11	77,5
80-85	1	82,5
85-90	2	87,5
Suma	667	

Statystyka opisowa – miary tendencji centralnej

Miary tendencji centralnej (dla próby)

(1) Średnia z próby

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

Gdzie:

x_i- obserwacja wartości badanej cechy dla i-tego elementu populacji generalnej wybranego dla próby

n – liczba obserwacji w próbie

(2) Mediana – wartość obserwacji środkowej, jeżeli obserwacje uporządkowaliśmy w kolejności np. rosnących wartości. Gdy liczba obserwacji w próbie jest parzysta, to jako medianę przyjmujemy średnią z dwu obserwacji w próbie jest parzysta, to jako medianę przyjmujemy średnią z dwu obserwacji środkowych.

(3) Moda (dominanta) – najczęściej występująca wartość obserwacji w próbie

Statystyka opisowa – miary tendencji centralnej

Uwagi:

(a) Średnia jest obliczana na podstawie wszystkich wartości obserwacji.
(b) Mediana nie zależy od obserwacji skrajnych – dlatego lepiej odzwierciedla tendencję centralną przy rozkładach silnie asymetrycznych.
(c) Dominanta w próbie może być jedna, może ich być więcej lub nie być w ogóle

Statystyka opisowa – miary tendencji centralnej

Obliczanie średniej, mediany i mody dla danych w postaci szeregów rozdzielczych

Średnia:

$$\bar{x} \cong \frac{\sum_{i=1}^{k} \dot{x}_i n_i}{\sum_{i=1}^{k} n_i}$$

n_i- liczebność w i-tym przedziale klasowym

k – liczba klas

$\dot{x}_i$- środek i-tego przedziału klasowego

Statystyka opisowa – miary tendencji centralnej

Mediana:

$$M_e \cong x_0 + \frac{l}{n_0}(N_{Me} - N^*)$$

x_0 – dolna granica przedziału klasowego mediany

l – szerokość przedziału klasowego mediany

n_0 - liczebność w przedziale mediany

N_{Me} – numer obserwacji, której wartość jest medianą

N^* - skumulowana liczba obserwacji do klasy mediany (bez klasy mediany)

Statystyka opisowa – miary tendencji centralnej

Moda (dominanta):

$$D \cong x_0 + l\frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})}$$

x_0 – dolna granica przedziału klasowego mody

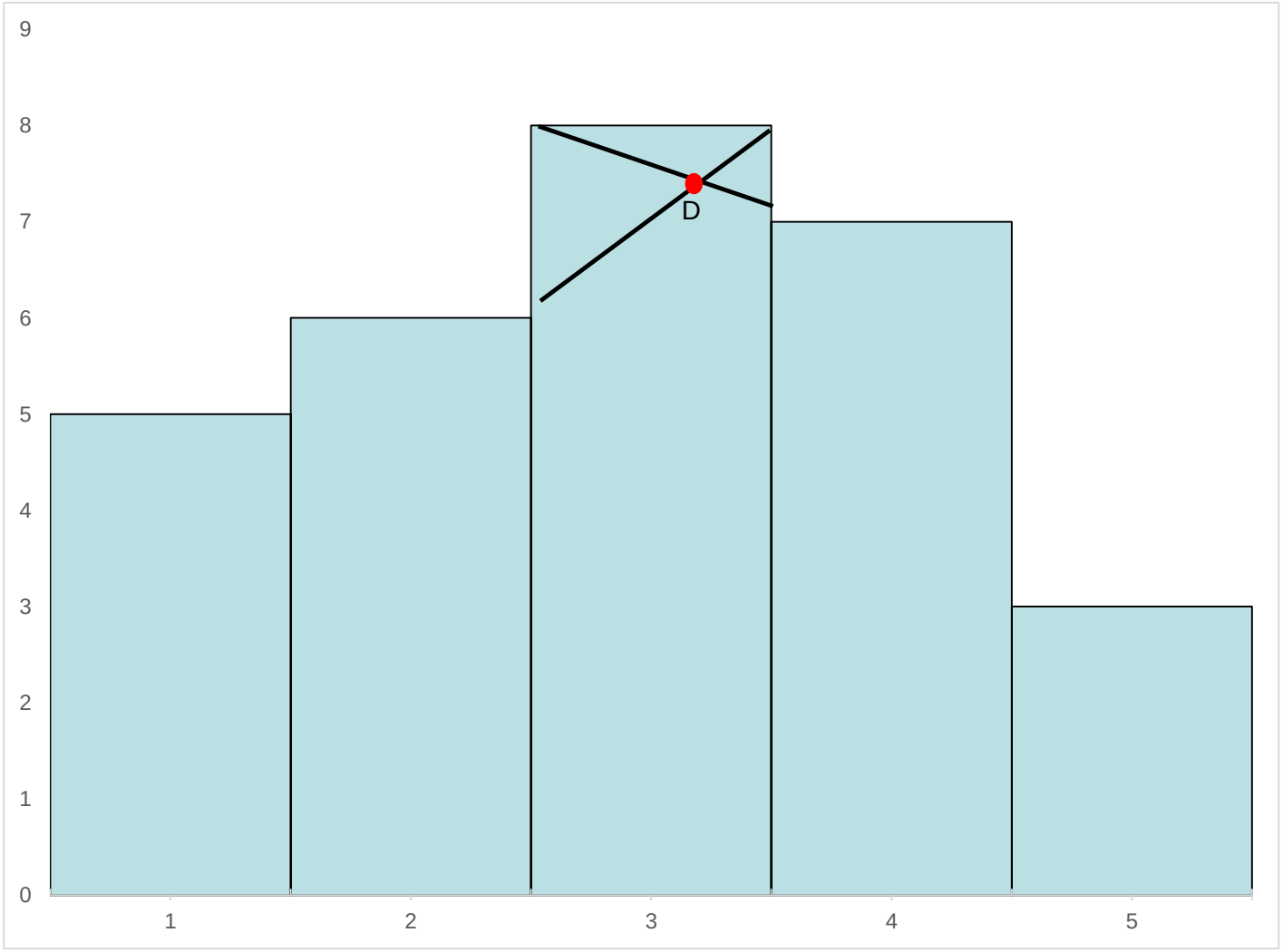
l – szerokość przedziału klasowego mody

n_d - liczebność w przedziale mody

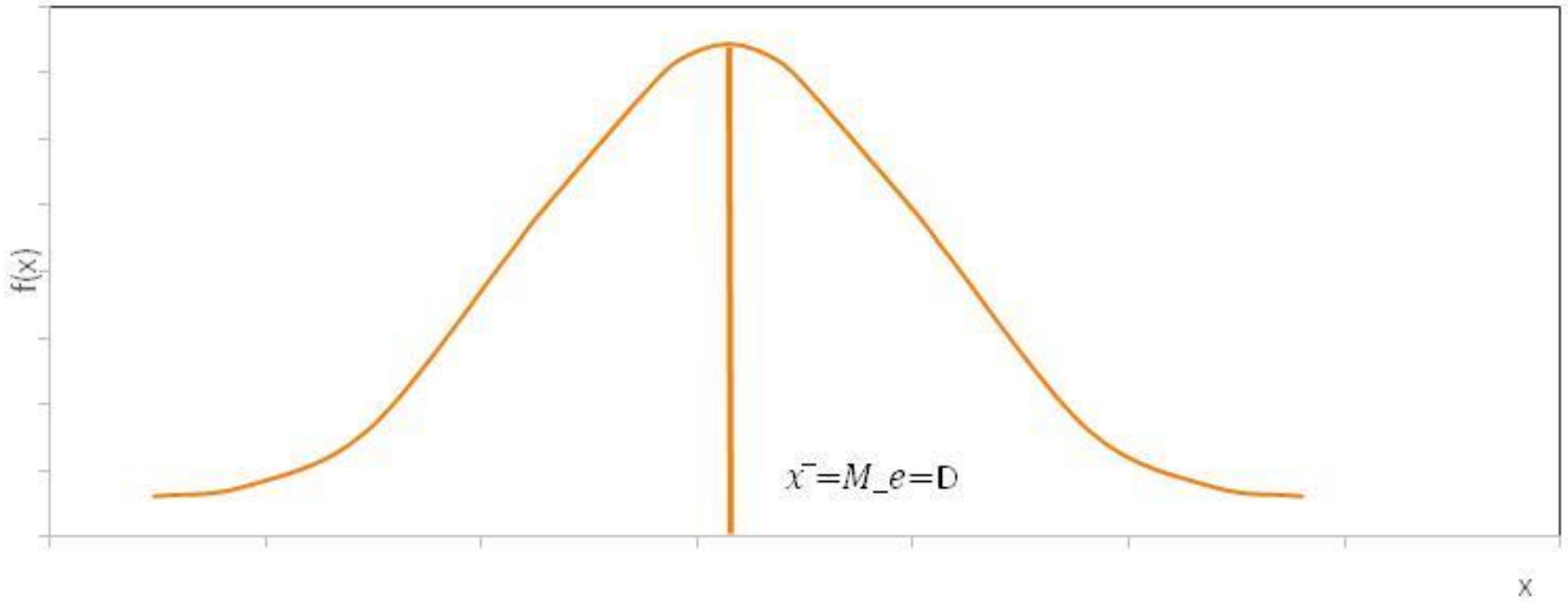
n_{d-1} - liczebność w przedziale poprzedzającym przedział mody

n_{d+1} - liczebność w przedziale następującym po przedziale mody

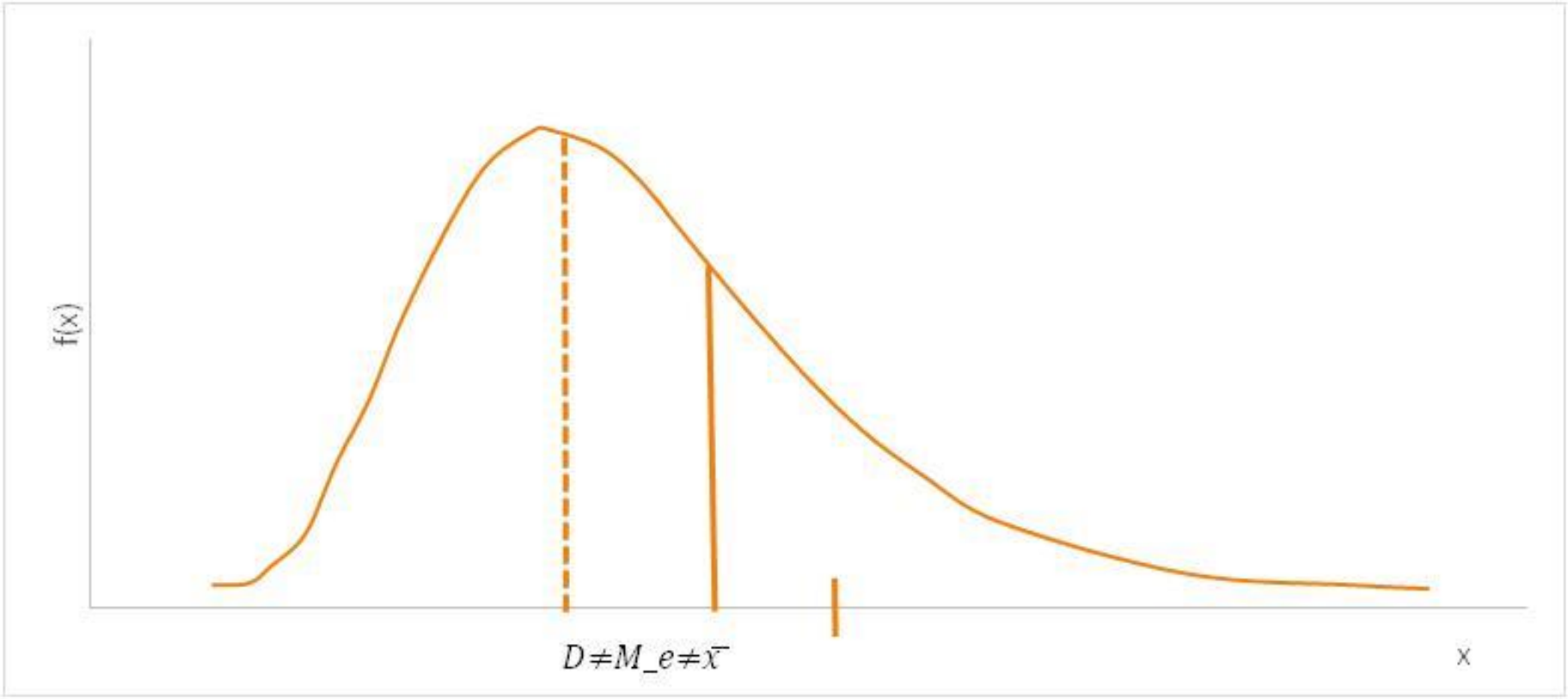
Statystyka opisowa – miary tendencji centralnej



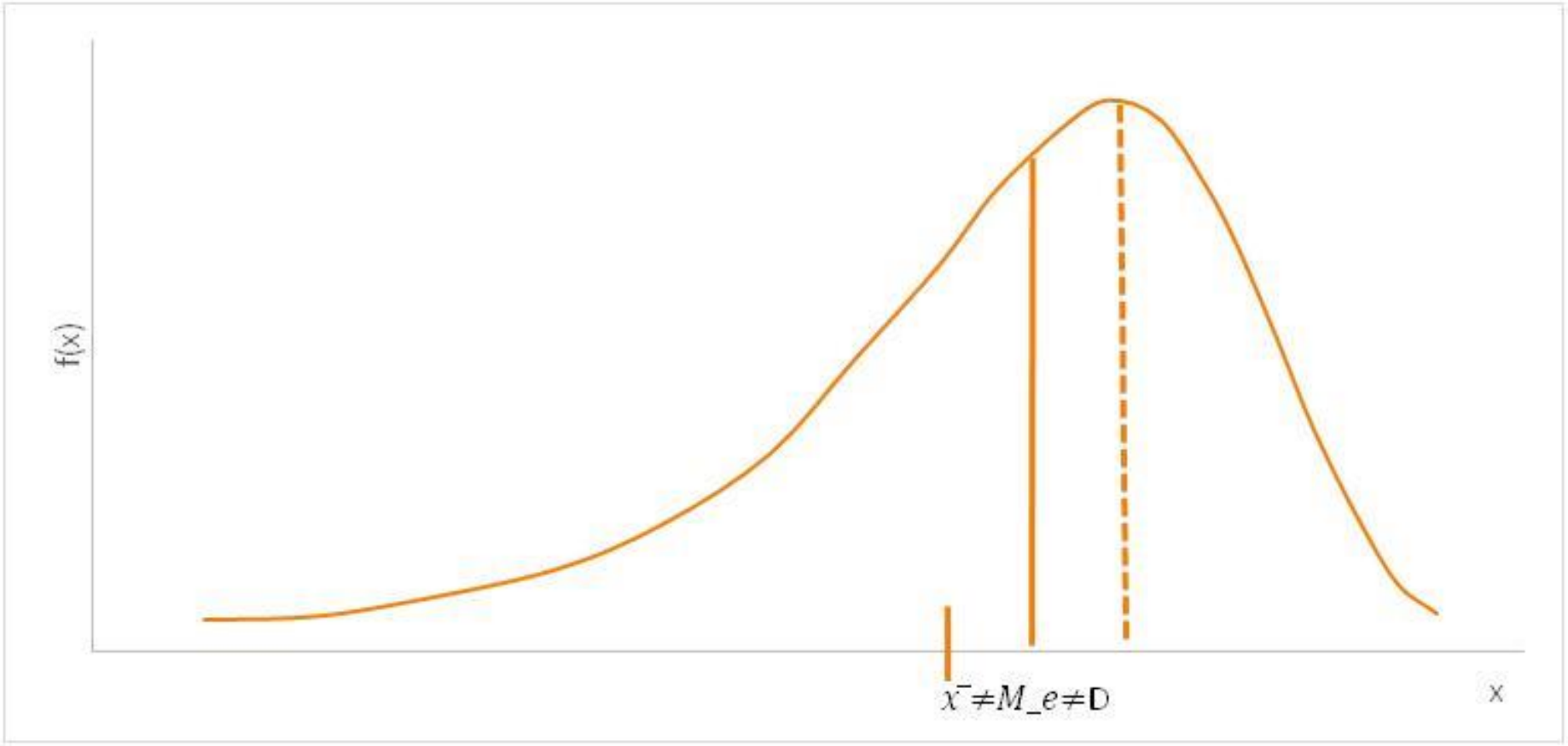
Średnia, mediana i moda rozkładu populacji



Średnia, mediana i moda rozkładu populacji



Średnia, mediana i moda rozkładu populacji

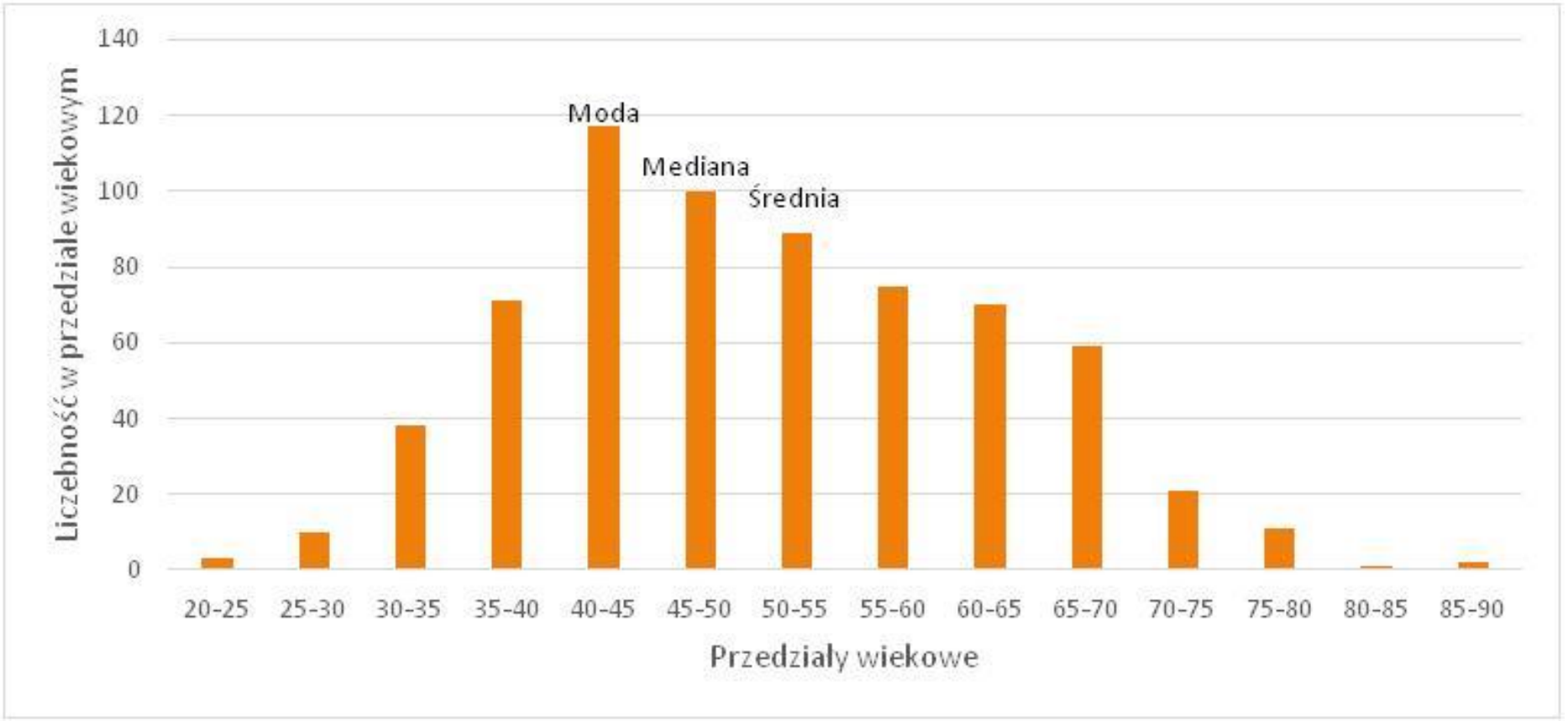


Średnia, mediana i moda rozkładu

Wiek pacjentek z nowotworem szyjki macicy w pewnym szpitalu w Algierii

Wiek	Liczba pacjentek	Środek przedziału wiekowego	(D):=(B)*(C)	Liczba pacjentek narastająco
(A)	(B)	(C)	(D)	(E)
20-25	3	22,5	67,5	3
25-30	10	25,5	255	13
30-35	38	32,5	1235	51
35-40	71	37,5	2662,5	122
40-45	117	42,5	4972,5	239
45-50	100	47,5	4750	339
50-55	89	52,5	4672,5	428
55-60	75	57,5	4312,5	503
60-65	70	62,5	4375	573
65-70	59	67,5	3982,5	632
70-75	21	72,5	1522,5	653
75-80	11	77,5	852,5	664
80-85	1	82,5	82,5	665
85-90	2	87,5	175	667
Suma	667		33917,5	
		Średnia:	**50,9**	
		Mediana:	**49,8**	
		Moda:	**43,7**	

Średnia, mediana i moda rozkładu



Miary pozycyjne - staniny

KARTY WYNIKÓW MATURY 2018	INFORMATYKA

INFORMATYKA na poziomie rozszerzonym (egzamin zdawało 7 310 osób)

Podział wyników na dziewięć klas

klasa	nazwa klasy	wyniki na świadectwie	Komentarz dla zdającego (informację o procentach podano w przybliżeniu)
1	najniższa	0%	4% zdających ma wynik w tej klasie, 96% zdających ma wynik w wyższych klasach
2	bardzo niska	1% – 4%	7% zdających ma wynik w tej klasie, 89% zdających ma wynik w wyższych klasach, 4% w niższej
3	niska	5% – 12%	12% zdających ma wynik w tej klasie, 77% zdających ma wynik w wyższych klasach, 11% w niższych
4	poniżej średniej	13% – 20%	17% zdających ma wynik w tej klasie, 60% zdających ma wynik w wyższych klasach, 23% w niższych
5	średnia	21% – 32%	20% zdających ma wynik w tej klasie, 40% zdających ma wynik w wyższych klasach, 40% w niższych
6	powyżej średniej	33% – 50%	17% zdających ma wynik w tej klasie, 23% zdających ma wynik w wyższych klasach, 60% w niższych
7	wysoka	51% – 68%	12% zdających ma wynik w tej klasie, 11% zdających ma wynik w wyższych klasach, 77% w niższych
8	bardzo wysoka	69% – 86%	7% zdających ma wynik w tej klasie, 4% zdających ma wynik w wyższej klasie, 89% w niższych
9	najwyższa	87% – 100%	4% zdających ma wynik w tej klasie, 96% w niższych

67%

Miary pozycyjne - staniny

KARTY WYNIKÓW MATURY 2018	GEOGRAFIA

GEOGRAFIA na poziomie rozszerzonym (egzamin zdawało 66 119 osób)

Podział wyników na dziewięć klas

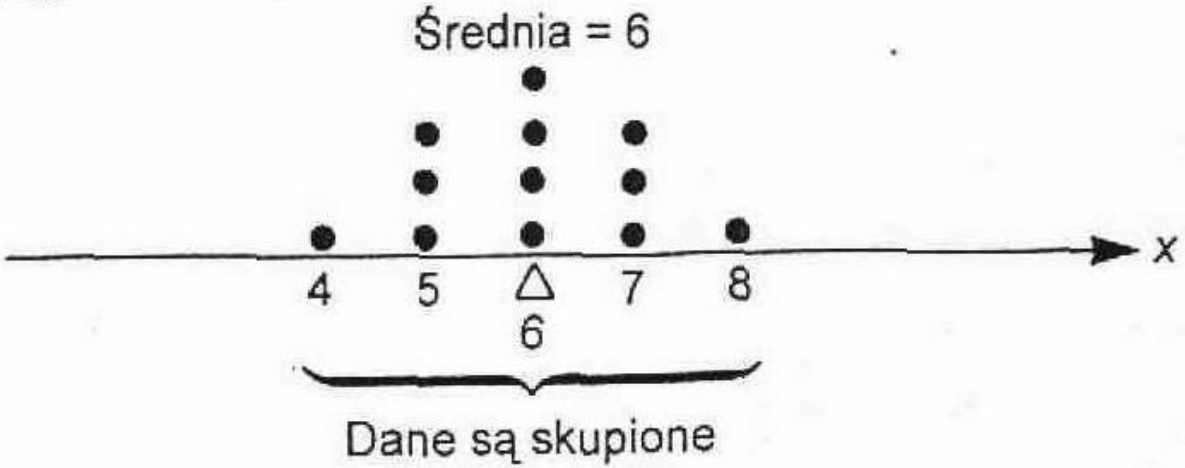
klasa	nazwa klasy	wyniki na świadectwie	Komentarz dla zdającego (informację o procentach podano w przybliżeniu)
1	najniższa	0% – 7%	4% zdających ma wynik w tej klasie, 96% zdających ma wynik w wyższych klasach
2	bardzo niska	8% – 10%	7% zdających ma wynik w tej klasie, 89% zdających ma wynik w wyższych klasach, 4% w niższej
3	niska	11% – 15%	12% zdających ma wynik w tej klasie, 77% zdających ma wynik w wyższych klasach, 11% w niższych
4	poniżej średniej	16% – 22%	17% zdających ma wynik w tej klasie, 60% zdających ma wynik w wyższych klasach, 23% w niższych
5	średnia	23% – 30%	20% zdających ma wynik w tej klasie, 40% zdających ma wynik w wyższych klasach, 40% w niższych
6	powyżej średniej	31% – 40%	17% zdających ma wynik w tej klasie, 23% zdających ma wynik w wyższych klasach, 60% w niższych
7	wysoka	41% – 52%	12% zdających ma wynik w tej klasie, 11% zdających ma wynik w wyższych klasach, 77% w niższych
8	bardzo wysoka	53% – 65%	7% zdających ma wynik w tej klasie, 4% zdających ma wynik w wyższej klasie, 89% w niższych
9	najwyższa	66% – 100%	4% zdających ma wynik w tej klasie, 96% w niższych

67%

Miary rozrzutu (zmienności)

Zbiór I:

Zbiór II:



Statystyka opisowa – miary rozrzutu

(1) Odchylenie przeciętne

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

Bazuje na informacji zawartej we wszystkich obserwacjach, ale trudno poddaje się działaniom matematycznym, stąd nie ma szerszego zastosowania.

Statystyka opisowa – miary rozrzutu

(2) Wariancja i odchylenie standardowe

Wariancja

$$s^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n-1}$$

Odchylenie standardowe

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n-1}}$$

Estymator nieobciążony:

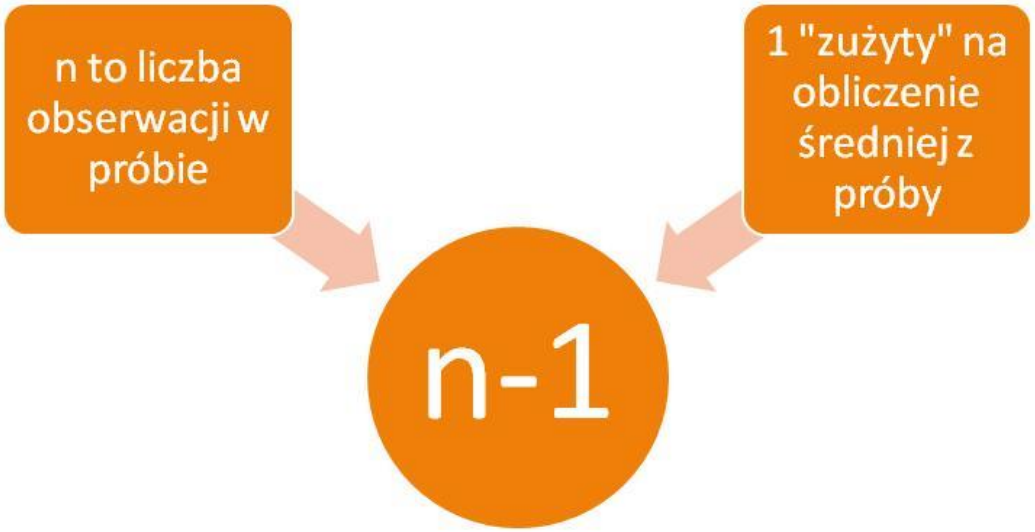
$$s^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n-1}$$

Estymator obciążony:

$$s^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n}$$

Statystyka opisowa – miary rozrzutu

$$Oszacowanie\ wariancji = \frac{suma\ kwadratów\ odchyleń\ od\ pewnej\ wartości}{liczba\ stopni\ swobody}$$



$$s^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n-1}$$

Statystyka opisowa – miary rozrzutu

Wariancja dla szeregu rozdzielczego:

$$s^2 = \frac{\sum_{i=0}^{k}(\dot{x}_i - \bar{x})^2}{\sum_{i=1}^{k} n_i - 1}$$

k – liczba przedziałów klasowych

$\dot{x}_i$ - środek i-tego przedziału klasowego

n_i - liczebność w i-tym przedziale klasowym

Odchylenie standardowe dla szeregu rozdzielczego:

$$s = \sqrt{s^2}$$

Statystyka opisowa – miary rozrzutu

(3) Współczynnik zmienności:

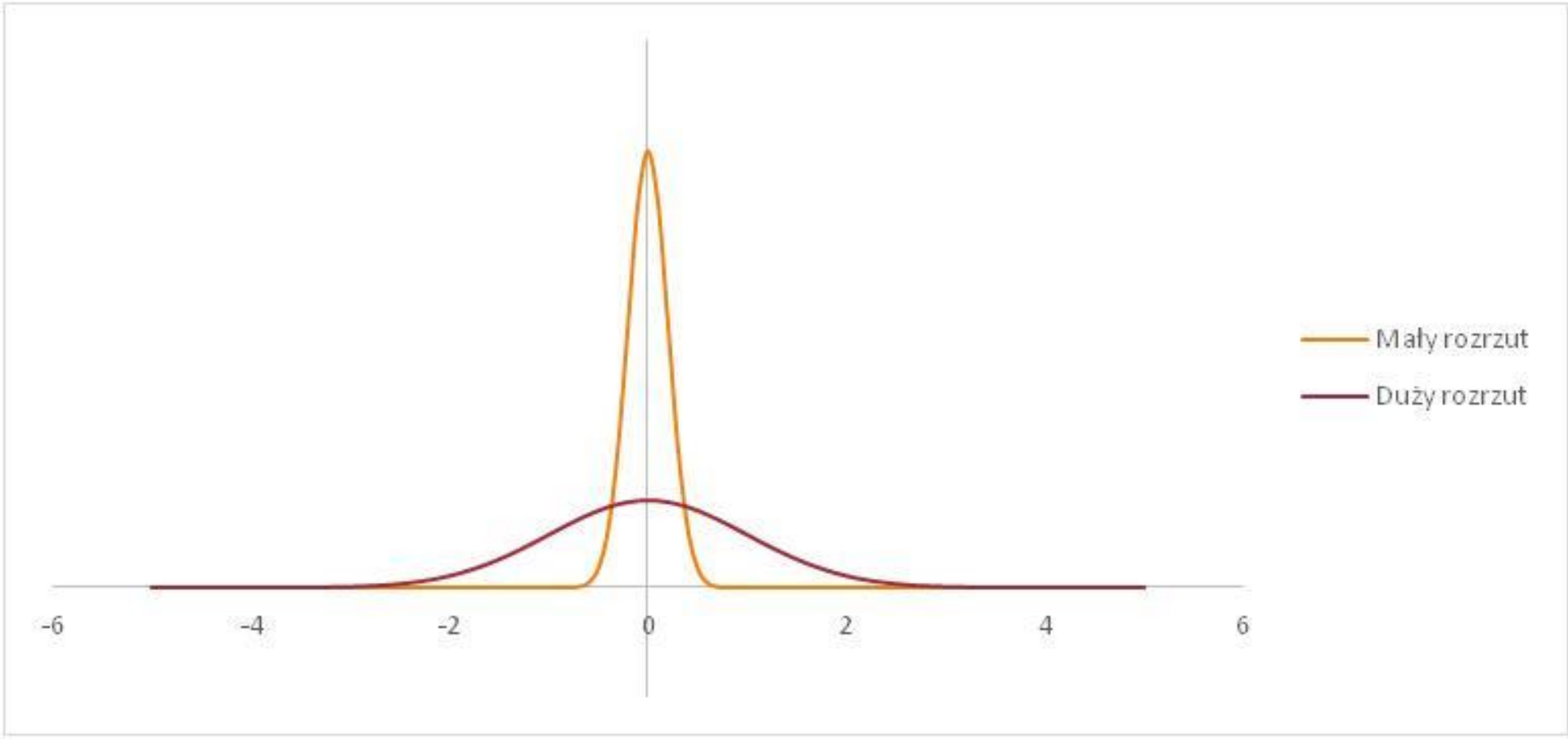
$$v = \frac{s}{\bar{x}} \cdot 100\%$$

(4) Współczynniki asymetrii:

Pierwszy: $a_1 = \frac{\bar{x} - D}{s}$, gdzie: D – Moda

Drugi: $a_2 = \frac{3(\bar{x} - M_e)}{s}$, gdzie: M_e – Mediana

Statystyka opisowa – miary rozrzutu



Rozkład dwumianowy

Prawdopodobieństwo, że wykonując *n* niezależnych doświadczeń, każde o prawdopodobieństwie sukcesu równym π, odniesiemy *r* sukcesów wynosi:

$$P(r) = \binom{n}{r} \pi^r (1-\pi)^{n-r}$$

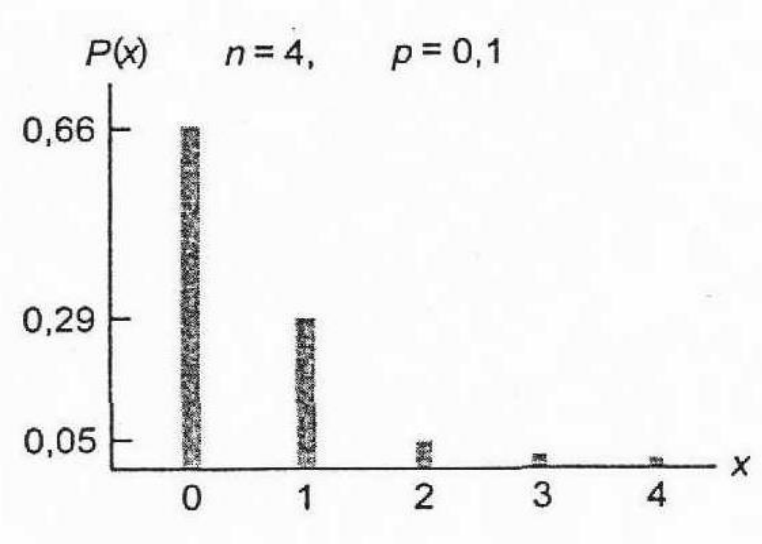
Zmienna losowa przyjmująca wartości dyskretne *r* z przedziału od *0* do *n* z prawdopodobieństwem wyrażonym powyższym wzorem ma rozkład dwumianowy, gdzie:

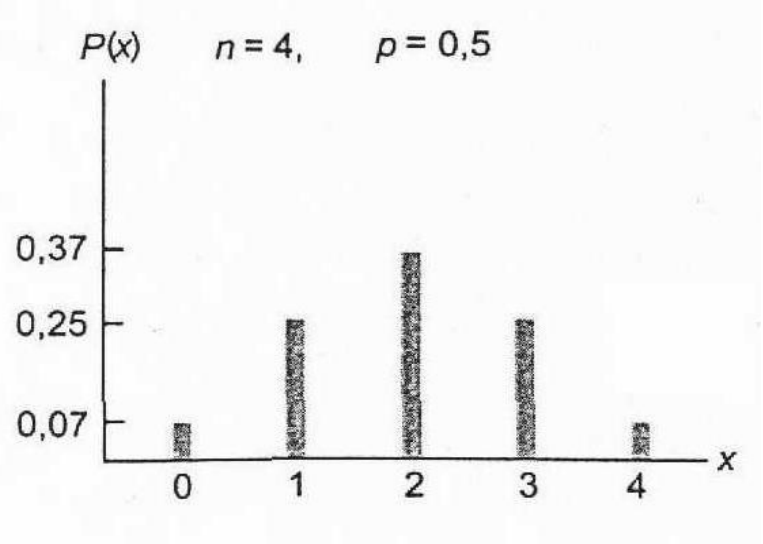
$$E(r) = n\pi$$

$$\sigma^2(r) = n\pi(1-\pi)$$

W zastosowaniach praktycznych na ogół znamy wartość *n*, ale nie znamy π. Prawdopodobieństwo sukcesu π można wyznaczyć na podstawie średniej z próby.

Dwumianowy rozkład prawdopodobieństwa





Rozkład dwumianowy

Przykład: Przeprowadzamy test na zdolność kiełkowania umieszczając na 100 szalkach po 5 nasion (razem 500 nasion). Otrzymano następujące wyniki:

Liczba kiełkujących na szalce	**5**	**4**	**3**	**2**	**1**	**0**	**Suma**
Liczba szalek	17	36	31	12	4	0	100
Ogólna liczba kiełkujących nasion	85	144	93	24	4	0	350

Rozkład dwumianowy

Liczba kiełkujących na szalce	5	4	3	2	1	0	Suma
Liczba szalek	17	36	31	12	4	0	100
Ogólna liczba kiełkujących nasion	85	144	93	24	4	0	350

Liczba szalek N_{szalek} = 100

Liczba nasion na szalce $n = 5$

Łącznie wykiełkowało *350*

Łącznie posiano $N = 500$

Zdolność kiełkowania $350/500 = 0,7 = p$

p jest estymatorem π, $p = 0,7$

Średnia liczba kiełkujących nasion = $np = 5*0,7 = 3,5$

Wariancja liczby kiełkujących nasion = $np(1-p) = s^2 = 5*0,7*0,3 = 1,05$

Odchylenie standardowe = $\sqrt{s^2} = 1,025$

Rozkład dwumianowy

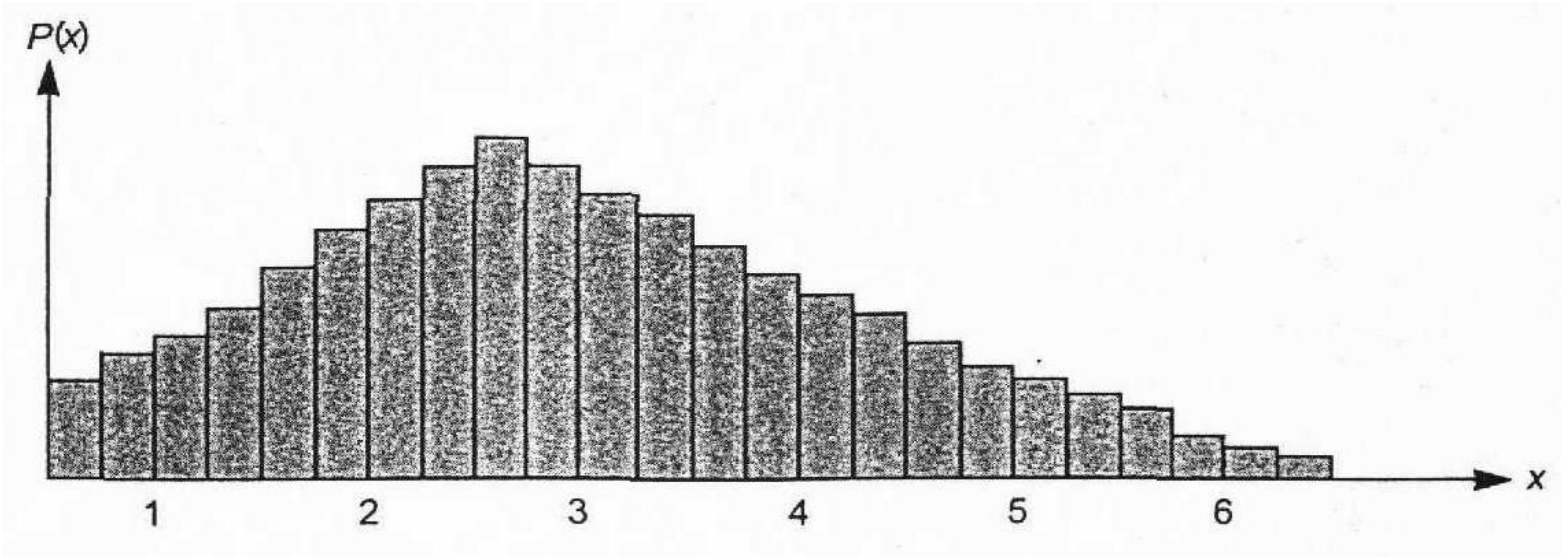
Takie postępowanie badawcze jest słuszne, gdy nasiona kiełkują niezależnie (tzn. gdy kiełkowanie jednego nasienia nie ma wpływu na kiełkowanie żadnego innego). Metody sprawdzania zgodności rozkładu będą podane później. Obecnie jedynie obliczamy „oczekiwane" (teoretyczne) częstości ze wzoru:

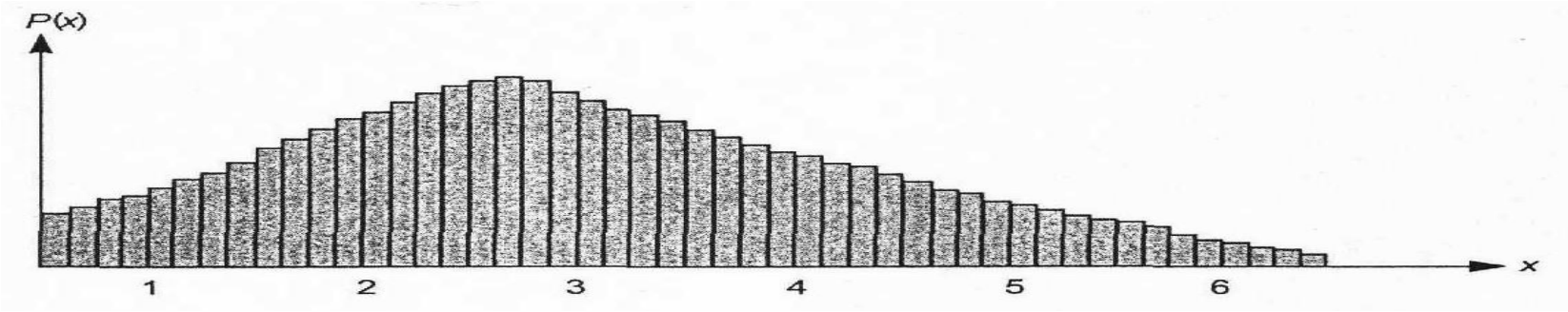
$$E_r = N_{szalek} \binom{n}{r} p^r (1-p)^{n-r}$$

Liczba kiełkujących na szalce	**5**	**4**	**3**	**2**	**1**	**0**	**Suma**
Liczba szalek obserwowana	17	36	31	12	4	0	100
Liczba szalek oczekiwana (teoretyczna)	16,81	36,01	30,87	13,23	2,83	0,25	100

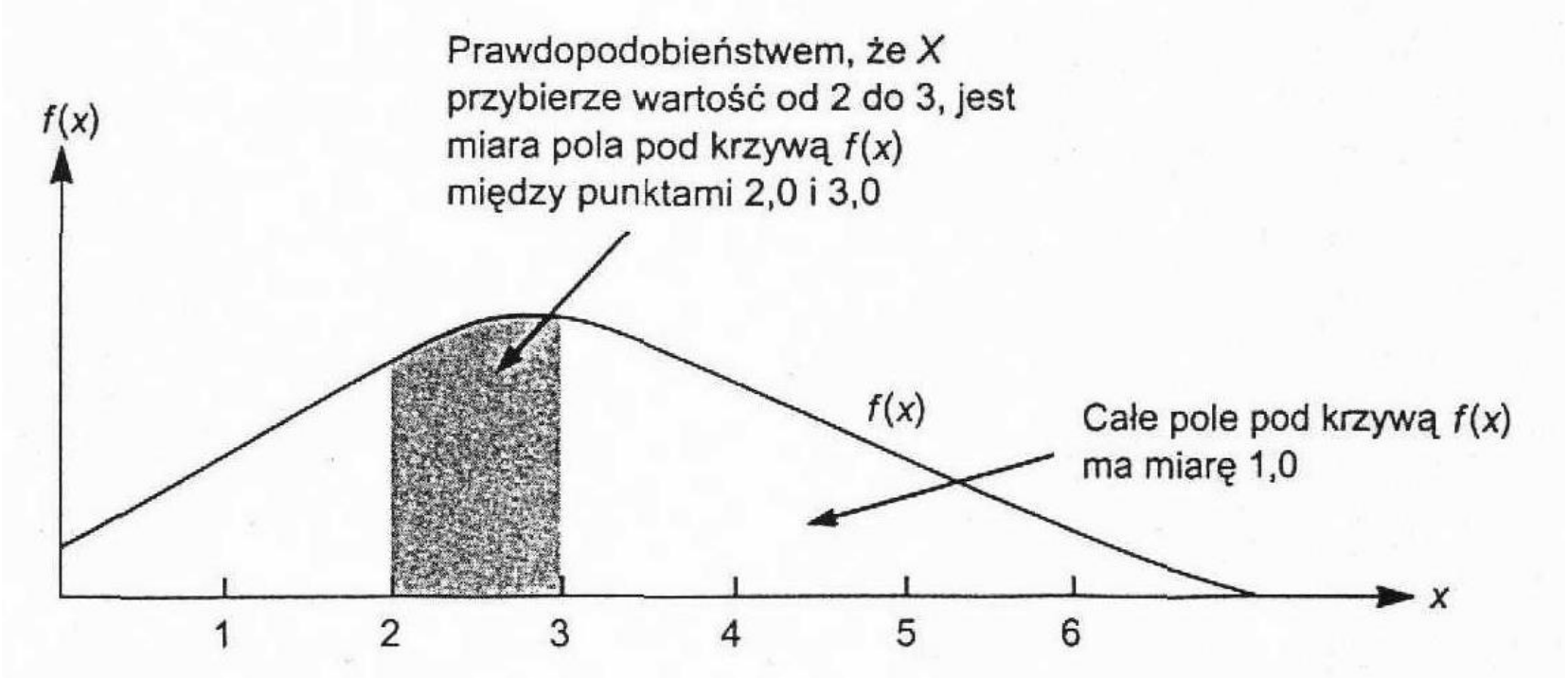
„Na oko" widać dużą zgodność.

Pojęcie ciągłej zmiennej losowej





Pojęcie ciągłej zmiennej losowej



Ciągłe zmienne losowe

Ciągła zmienna losowa to taka zmienna losowa, która może przyjmować dowolne wartości z pewnego przedziału liczbowego.

Prawdopodobieństwa związane z ciągłą zmienną losową X są wyznaczane przez funkcję gęstości prawdopodobieństwa zmiennej losowej. Ta funkcja, oznaczana $f(x)$, ma następujące własności:

1. $f(x) \geq 0$ dla wszystkich x.
2. Prawdopodobieństwo, że X przyjmie wartość między a i b jest równe mierze pola pod krzywą (wykresem) $f(x)$ między punktami a i b.
3. Całe pole pod krzywą (wykresem) $f(x)$ ma miarę 1,0.

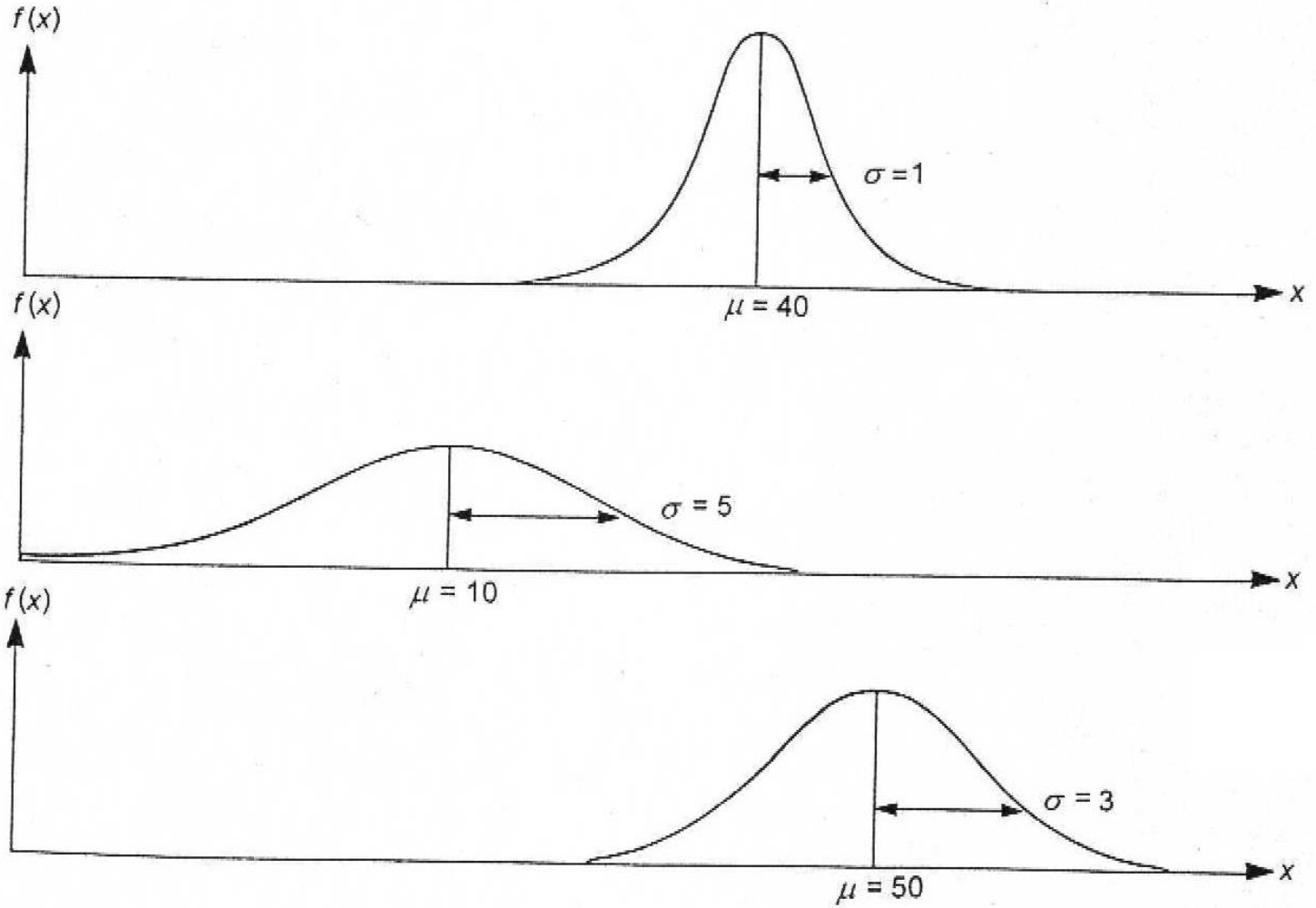
Normalny rozkład prawdopodobieństwa

Funkcja gęstości prawdopodobieństwa normalnej zmiennej losowej o średniej μ i odchyleniu standardowym σ:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{dla} \quad -\infty < x < \infty$$

gdzie e i π są liczbami 2,718… i 3,141…

Normalny rozkład prawdopodobieństwa



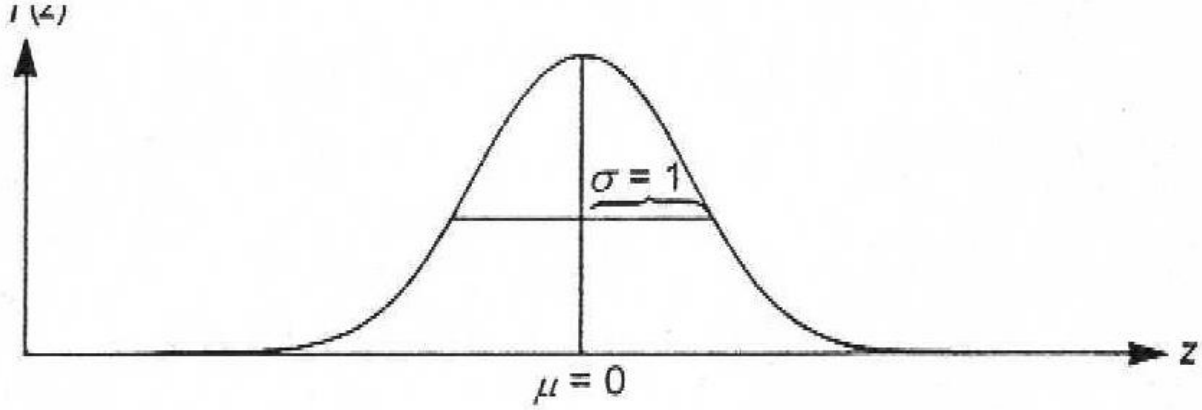
Rysunek 4.2. Rozkład normalny o różnych wartościach średniej (μ) i odchylenia standardowego (σ)

Standaryzowany rozkład normalny

Standaryzowaną normalną zmienną losową Z jest normalna zmienna losowa o średniej $\mu = 0$ i odchyleniu standardowym $\sigma = 1$.

Stosując wprowadzony sposób oznaczania zmiennych losowych zapiszemy:

$$Z \sim N(0, 1^2). \tag{4.3}$$



Rysunek 4.3. Standaryzowana normalna funkcja gęstości

Rozkład dwumianowy dla coraz dłuższych serii zmierza do rozkładu normalnego