Statystyka opisowa

Robert Pietrzykowski

email: robert_pietrzykowski@sggw.pl

www.ekonometria.info

Statystyka jest bardziej sposobem myślenia lub wnioskowania niż pęczkiem recept na młócenie danych w celu odsłonięcia odpowiedzi

C. R. Rao

Kłamstwo, wierutne kłamstwo, statystyka

Liczby nie kłamią ale kłamcy liczą

Ch. H. Grosvenor

Na dziś…

- Sprawy bieżące
  - Przypominam, że 14.11.2015 pierwszy sprawdzian
- Konsultacje
  - Sobota 9:00 – 10:00 pok. 13A b. 5
- Inne

Na dziś…

- Powtórzenie z poprzedniego wykładu
- Wykład 2:
  - rozkłady prawdopodobieństwa
  - rachunek prawdopodobieństwa
  - miary koncentracji
  - miary skośności


ciągła
jednowymiarowa
Skokowa
wielowymiarowa
Zmienna losowa

Zmienna losowa i jej rozkład

- Rozkłady teoretyczne
  - Normalny
  - Dwumianowy
  - Poissona
- Funkcje rozkładów
  - Funkcja gęstości
  - dystrybuanta

Zmienna losowa (cecha)

Funkcja o wartościach rzeczywistych określona na zbiorze zdarzeń elementarnych.

Rozkład zmiennej losowej

Zbiór wartości zmiennej losowej oraz prawdopodobieństwa z jakimi są te wartości przyjmowane.

Przykład. Jednokrotny rzut kostką.
Zmienna losowa: ilość wyrzuconych oczek.
Zbiór wartości: $\{1, 2, 3, 4, 5, 6\}$

x_i	1	2	3	4	5	6
p_i	1/6	1/6	1/6	1/6	1/6	1/6

CECHA X: powierzchnia mieszkania w m^2

32,45	33,21	34,36	35,78	37,79
39,80	41,45	41,55	42,27	4
44,90	45,10	45,90	46,52	4
49,55	49,65	49,70	49,90	5
51,98	52,00	52,10	52,30	5
55,30	55,56	55,62	56,00	5
57,70	57,90	58,00	58,50	5
64,30	64,60	65,00	66,29	6
76,80	77,10	77,80	78,90	7

38,54	38,91	38,96	39,50	39,67

przedział	środek	liczba mieszkań	częstości
$30-40$	35	11	0.11
$40-50$	45	23	0.23
$50-60$	55	33	0.33
$60-70$	65	12	0.12
$70-80$	75	6	0.06
$80-90$	85	8	0.08

Rozkład zmiennej losowej

Zbiór wartości zmiennej losowej oraz prawdopodobieństwa z jakimi są te wartości przyjmowane.

Popyt	1000	2000	3000	4000
Prawdopodobieństwo	0.4	0.3	0.2	0.1

Własności $f(x)$:

- $P(X \in (a,b)) = \int_a^b f(x)dx$ dla $a < b$,
- $\int_{-\infty}^{\infty} f(x)dx = 1$,
- $f(x) \geq 0$ dla dowolnego x.

Własności $F(x)$:

- $0 \leq F(x) \leq 1$ dla dowolnego $x \in R$,
- $F(x) = \int_{-\infty}^{x} f(t)dt$,
- $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$.

Wartość oczekiwana (średnia). Wartość oczekiwana EX zmiennej losowej X jest liczbą charakteryzującą położenie zbioru jej wartości

- Wartość oczekiwana zmiennej losowej X

  $EX = \int_{-\infty}^{\infty} x f(x) dx$, w przypadku zmiennej ciągłej.

  $EX = \sum_i x_i p_i$, w przypadku zmiennej skokowej.

Własności wartości oczekiwanej:

- $E(aX + b) = aEX + b$,
- $E(X + Y) = EX + EY$, dla dowolnych X i Y,
- $E(XY) = EX * EY$, dla X i Y niezależnych.

Wariancja. Wariancja D^2X zmiennej losowej jest liczbą charakteryzującą rozrzut zbioru jej wartości wokół wartości średniej EX

- ◇ Wariancja zmiennej losowej X
  $D^2X = E(X - EX)^2$

Własności wariancji:

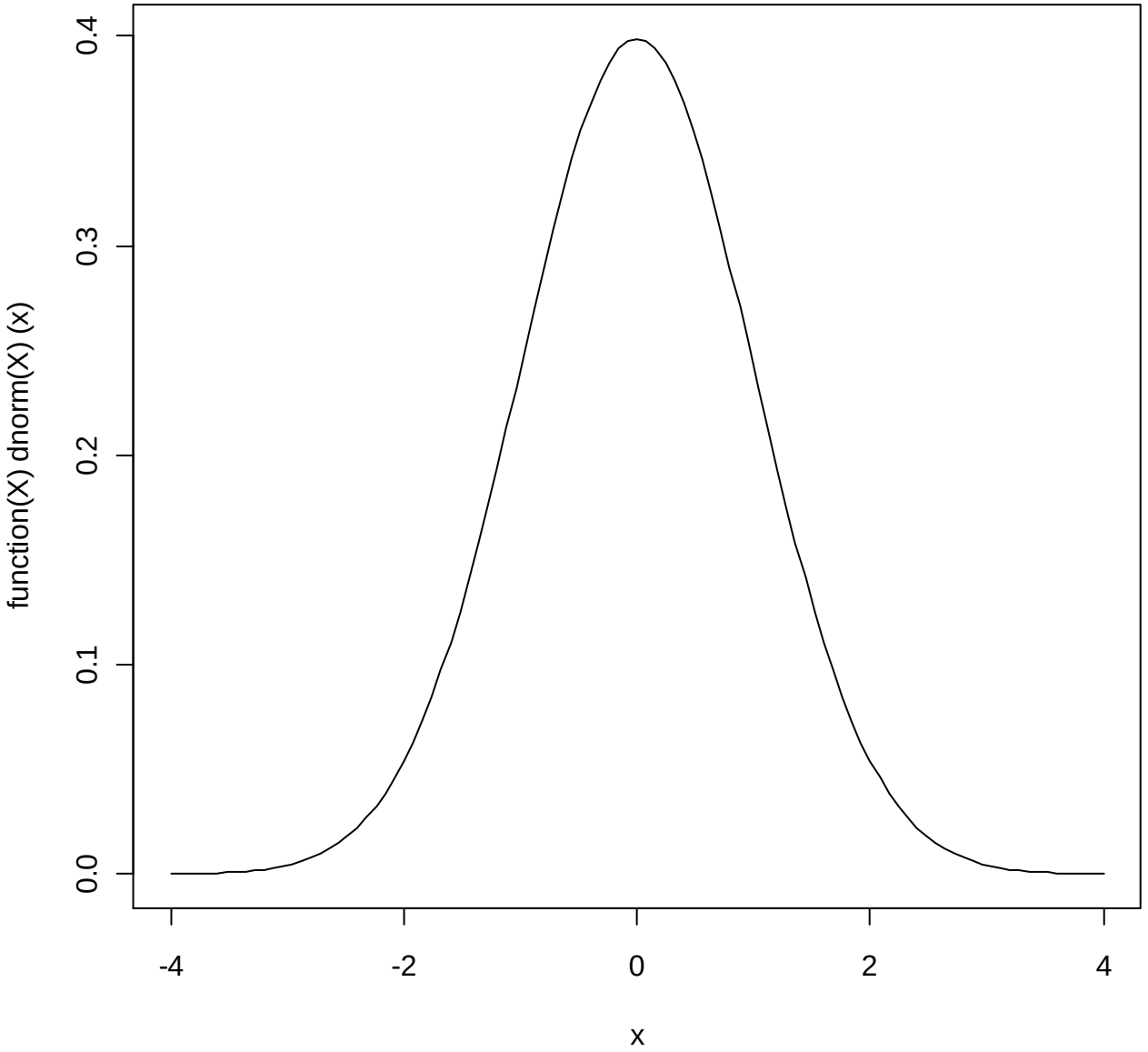
- ○ $D^2X = EX^2 - (EX)^2$,
- ○ $D^2(aX + b) = a^2D^2X$,
- ○ $E(aX + bY) = a^2D^2X + b^2D^2Y$, dla niezależnych X i Y.

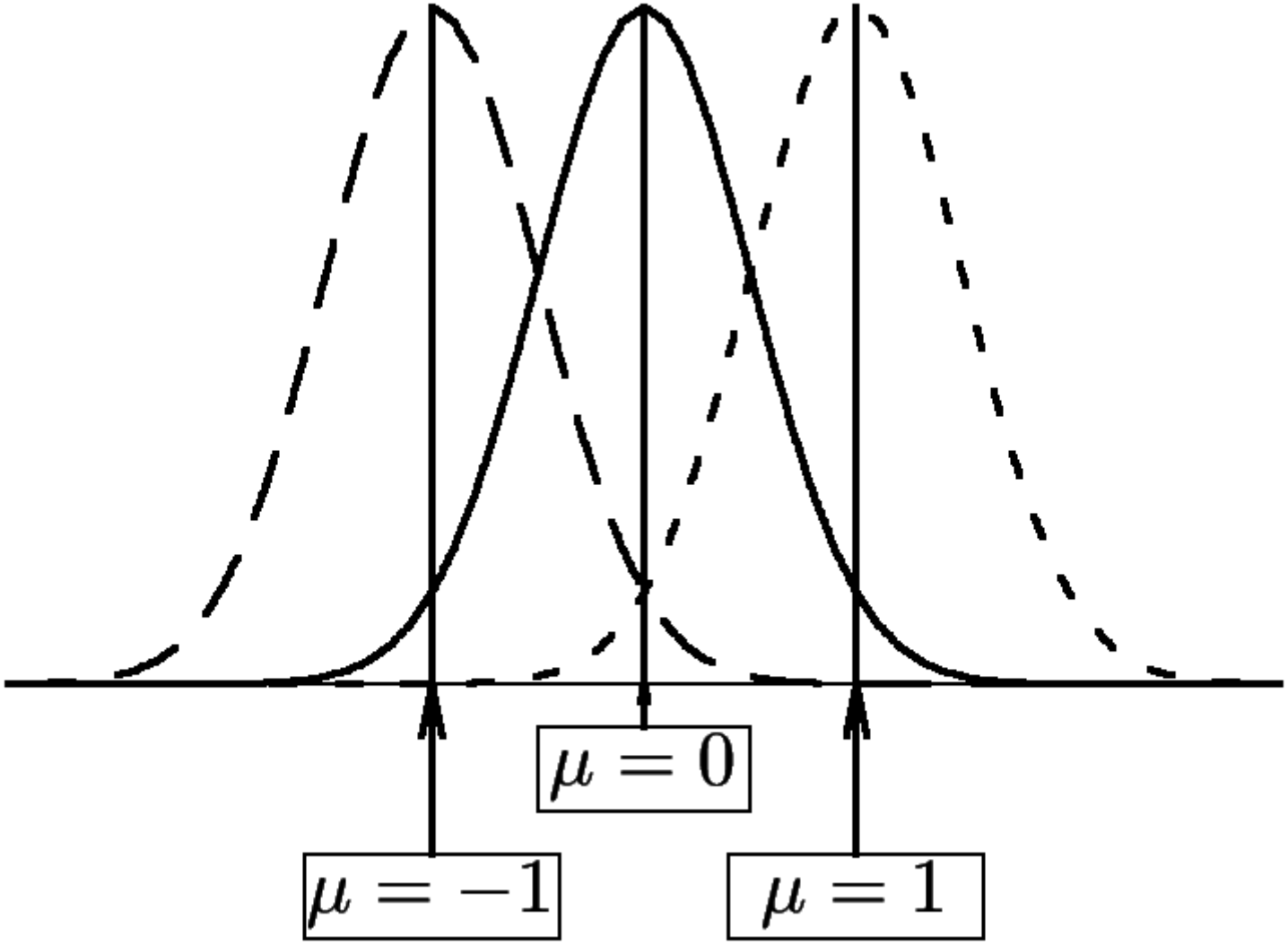
Rozkład normalny

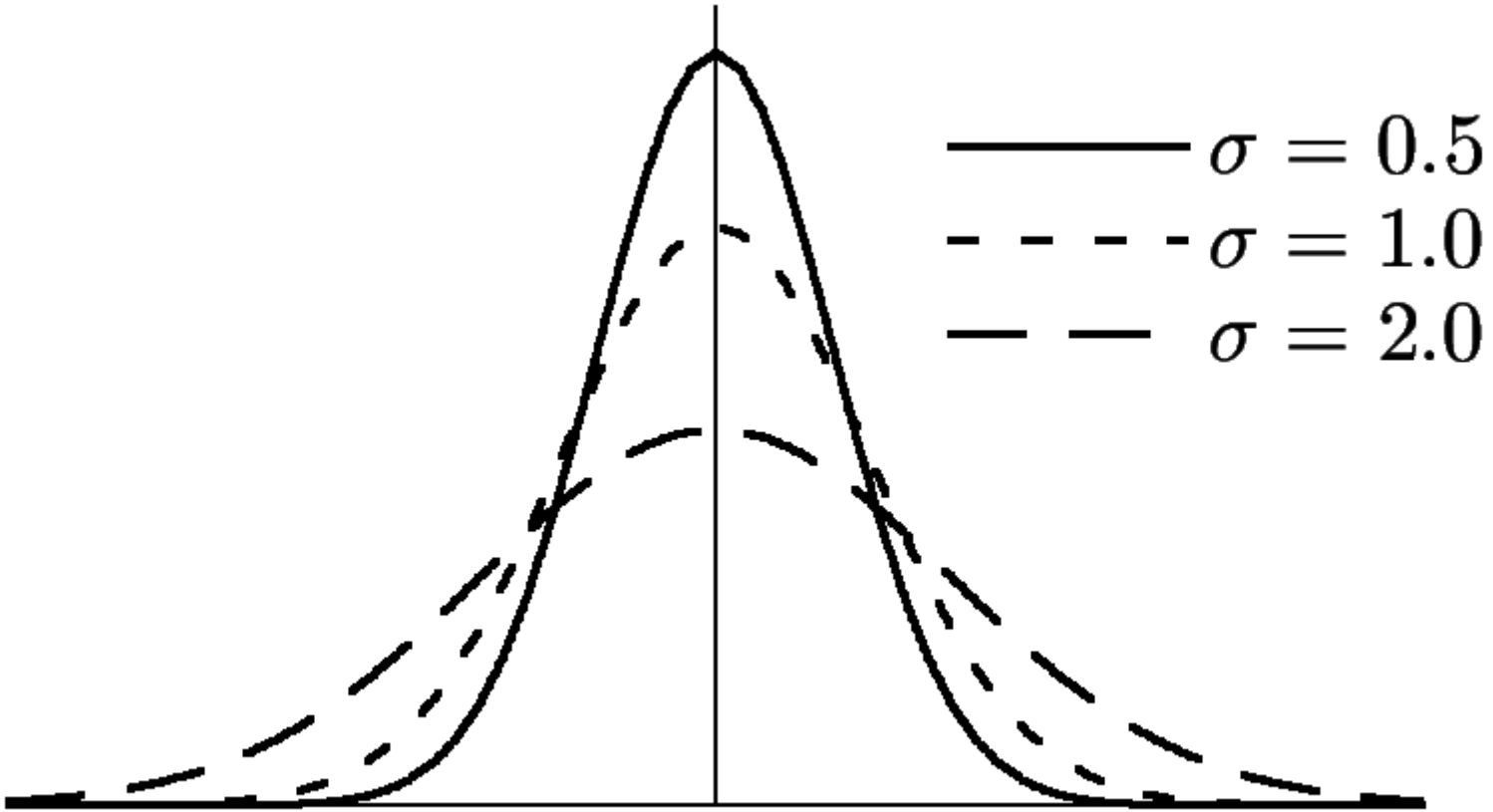
Zmienna losowa X ma rozkład normalny $N(\mu, \sigma^2)$ o wartości średniej μ i wariancji σ^2, jeżeli jej funkcja gęstości wyraża się wzorem

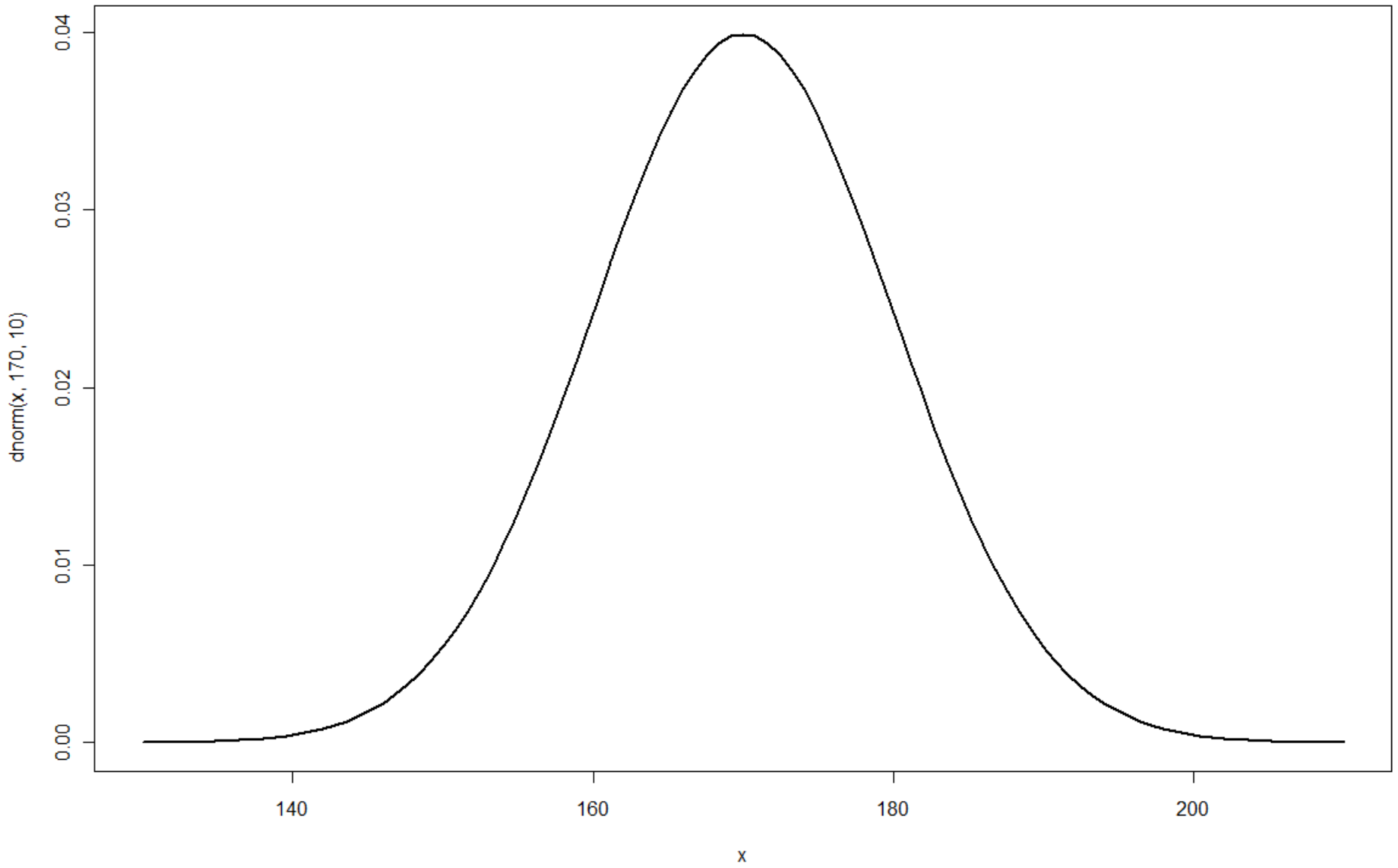
$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

$$EX = \mu \qquad D^2X = \sigma^2.$$

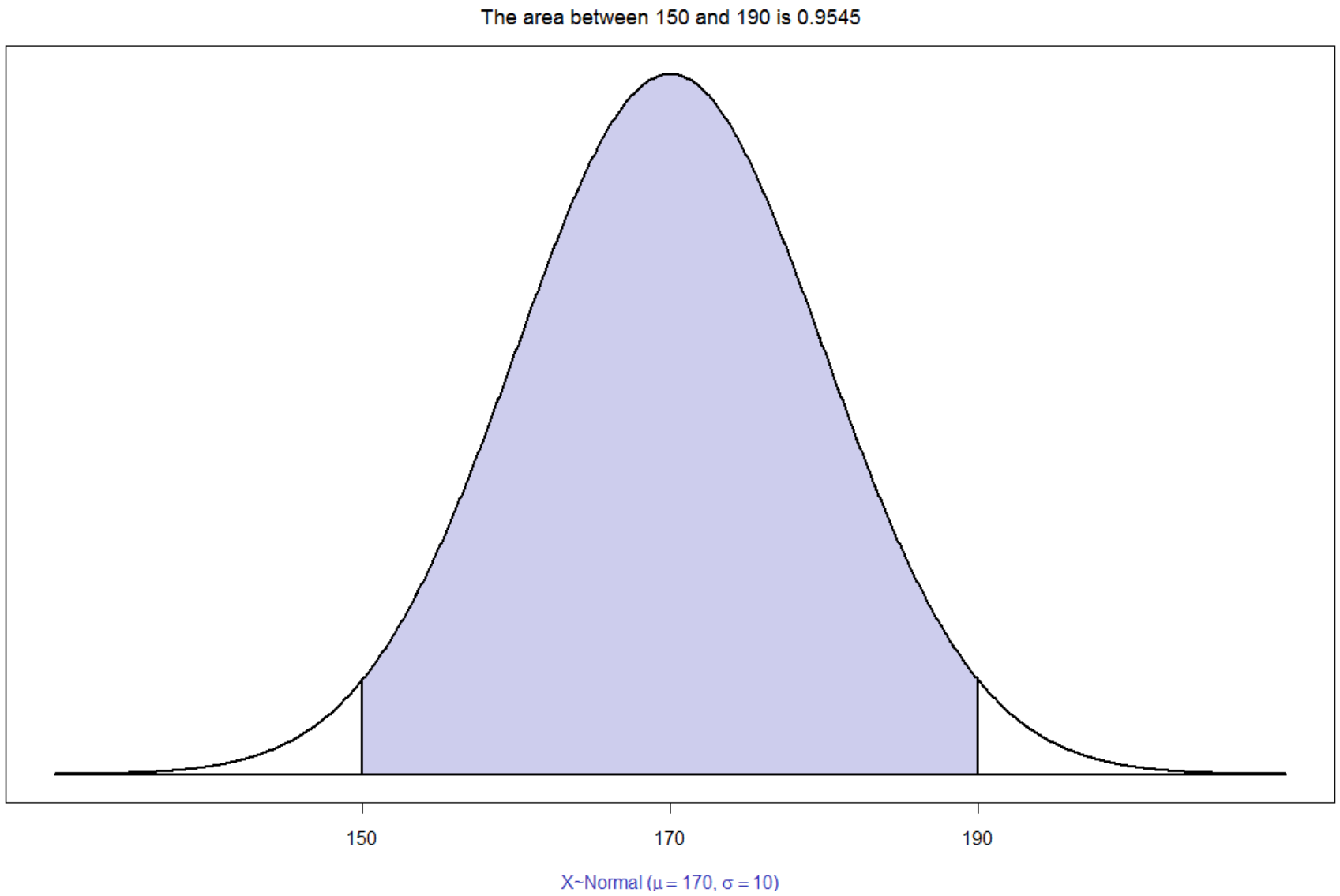

function(X) dnorm(X) (x)
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

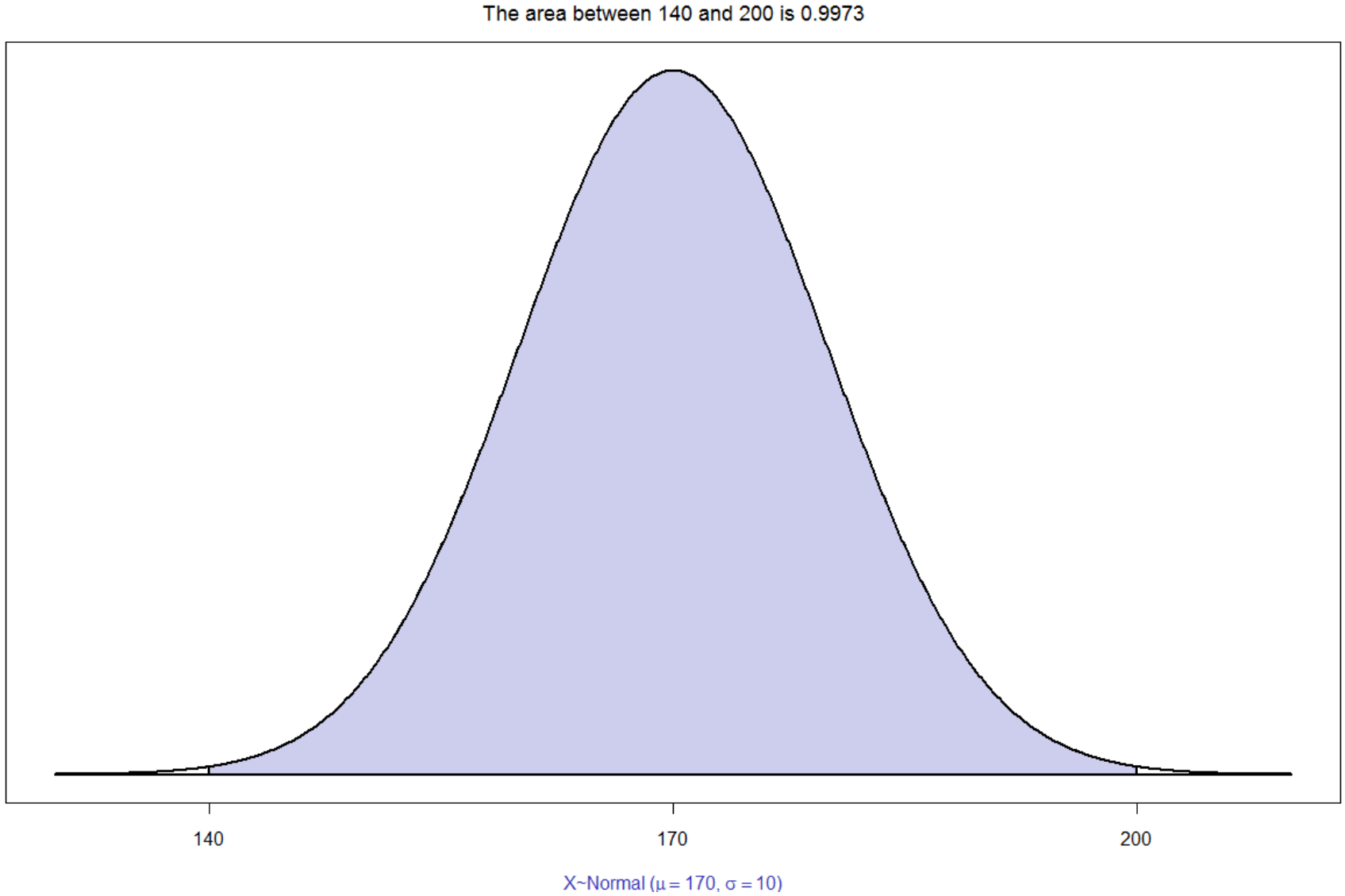

$\mu = 0$
$\mu = -1$
$\mu = 1$

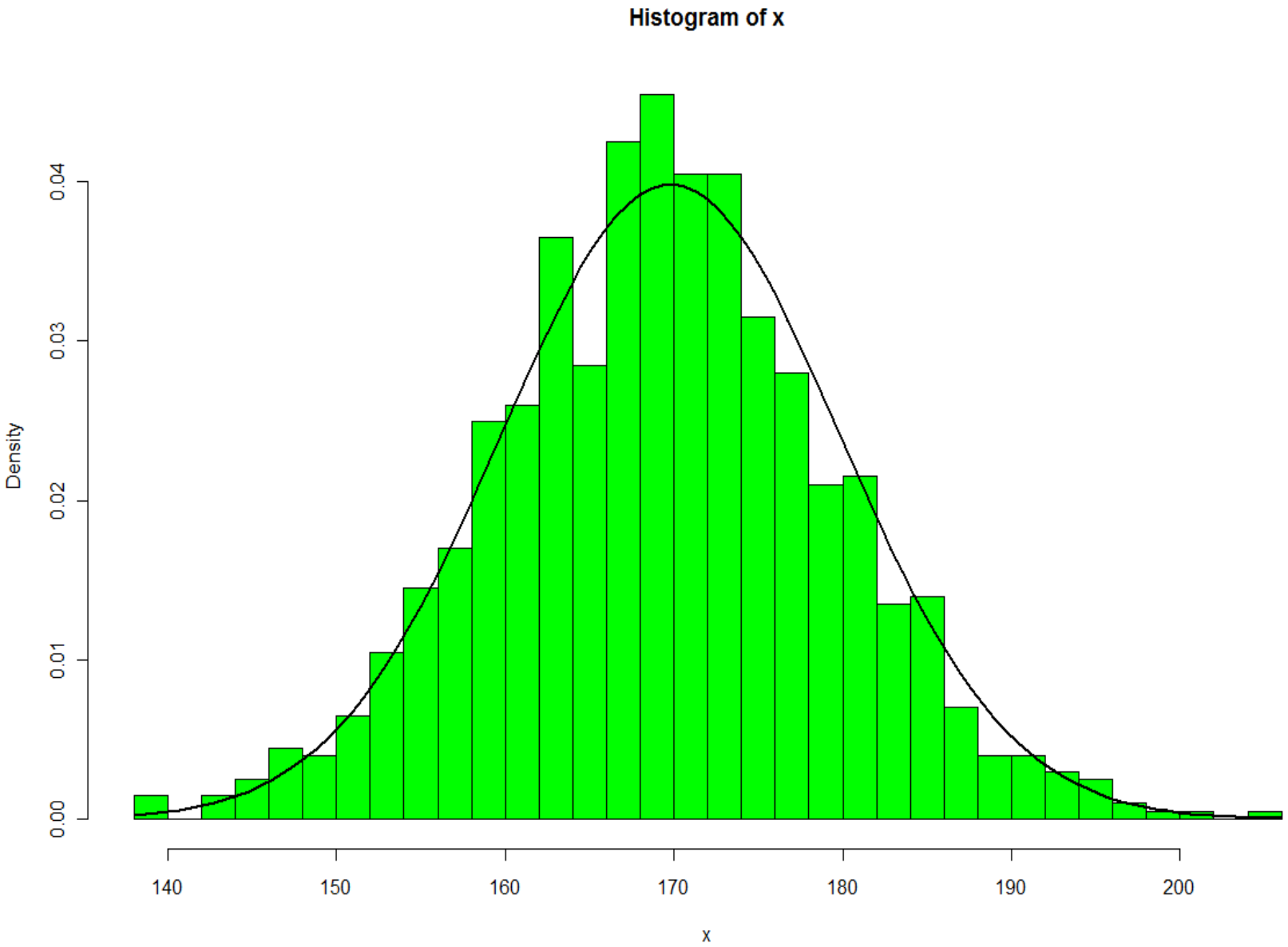

$\sigma = 0.5$
$\sigma = 1.0$
$\sigma = 2.0$

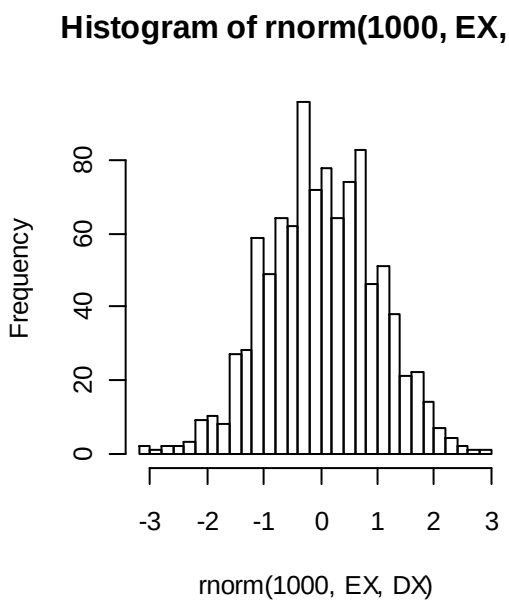

dnorm(x, 170, 10)
0.00
0.01
0.02
0.03
0.04
140
160
180
200
x

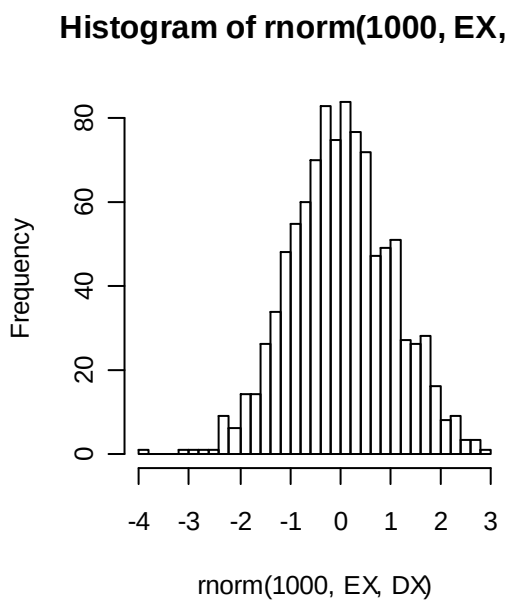

The area between 160 and 180 is 0.6827
160
170
180
X~Normal (μ = 170, σ = 10)

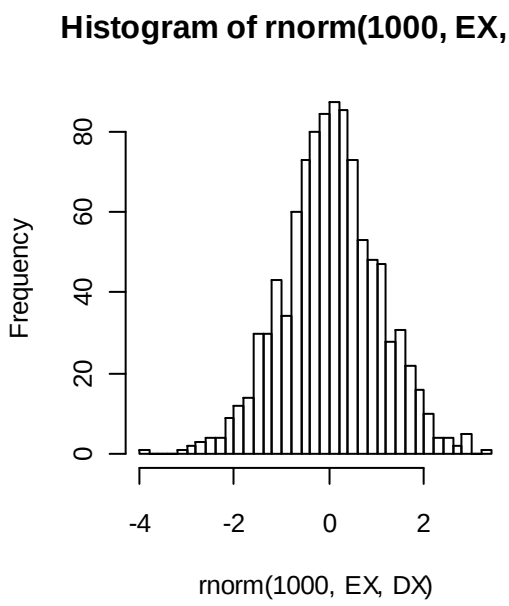

The area between 150 and 190 is 0.9545
150
170
190
X~Normal (μ = 170, σ = 10)

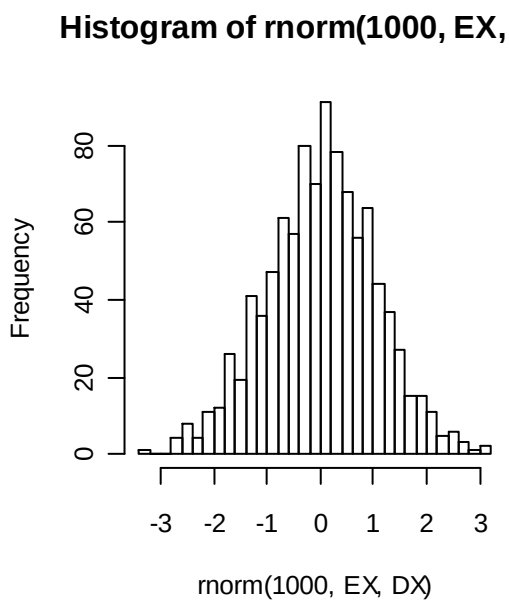

The area between 140 and 200 is 0.9973
140
170
200
X~Normal (μ = 170, σ = 10)

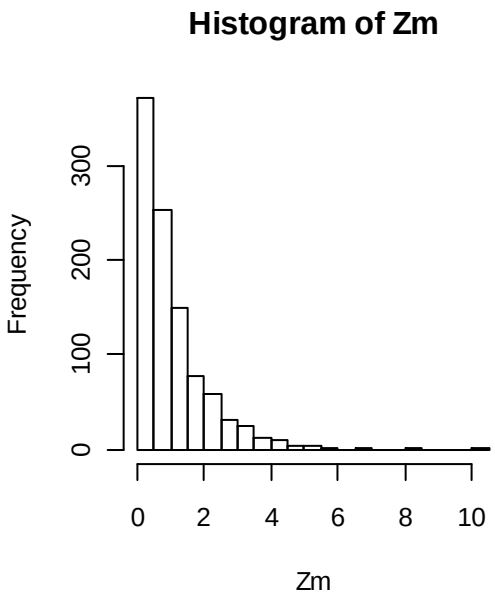

Histogram of x
Density
0.00
0.01
0.02
0.03
0.04
140
150
160
170
180
190
200
x

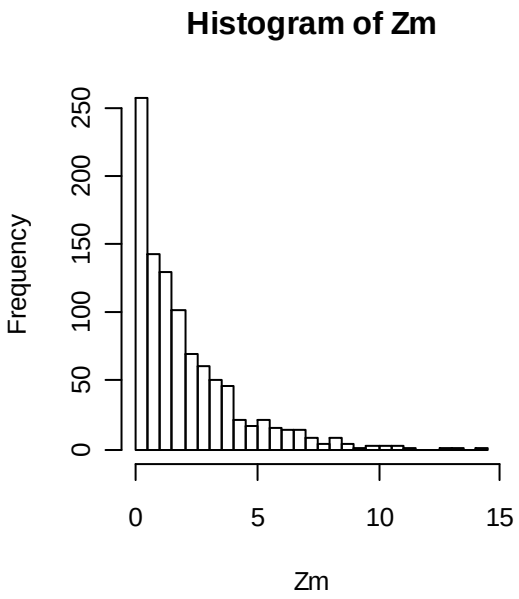

Histogram of rnorm(1000, EX, DX)
Frequency
0
20
40
60
80
-3
-2
-1
0
1
2
3
rnorm(1000, EX, DX)

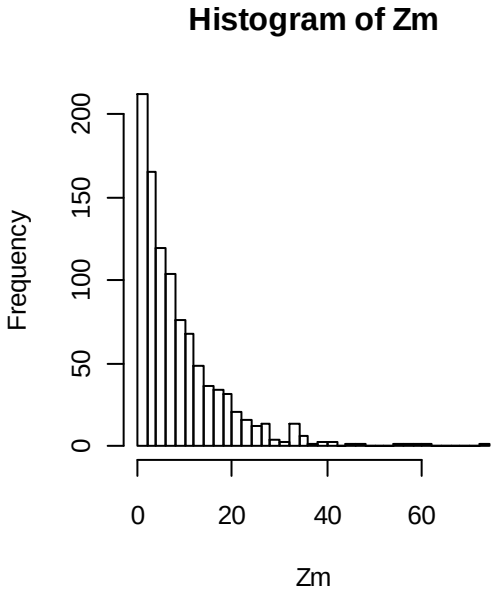

Histogram of rnorm(1000, EX, DX)
Frequency
0
20
40
60
80
-4
-3
-2
-1
0
1
2
3
rnorm(1000, EX, DX)

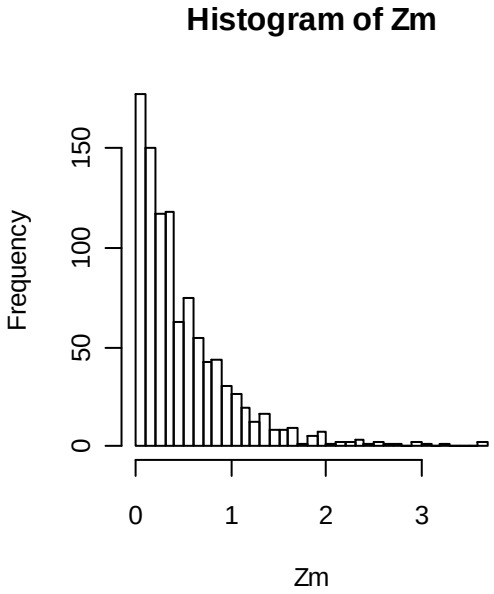

Histogram of rnorm(1000, EX, DX)
Frequency
0
20
40
60
80
-4
-2
0
2
rnorm(1000, EX, DX)

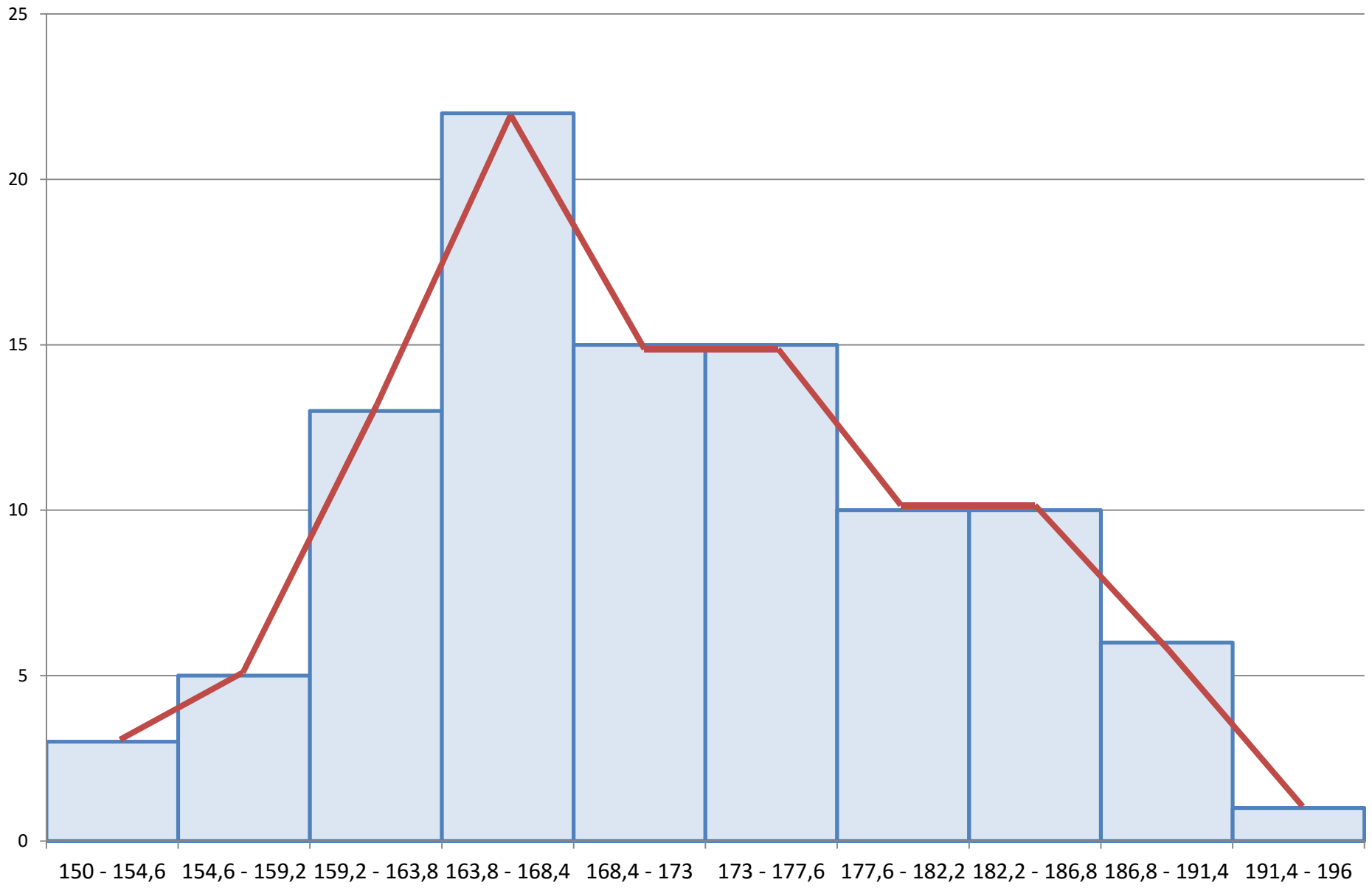

Histogram of rnorm(1000, EX, DX)
Frequency
0
20
40
60
80
-3
-2
-1
0
1
2
3
rnorm(1000, EX, DX)

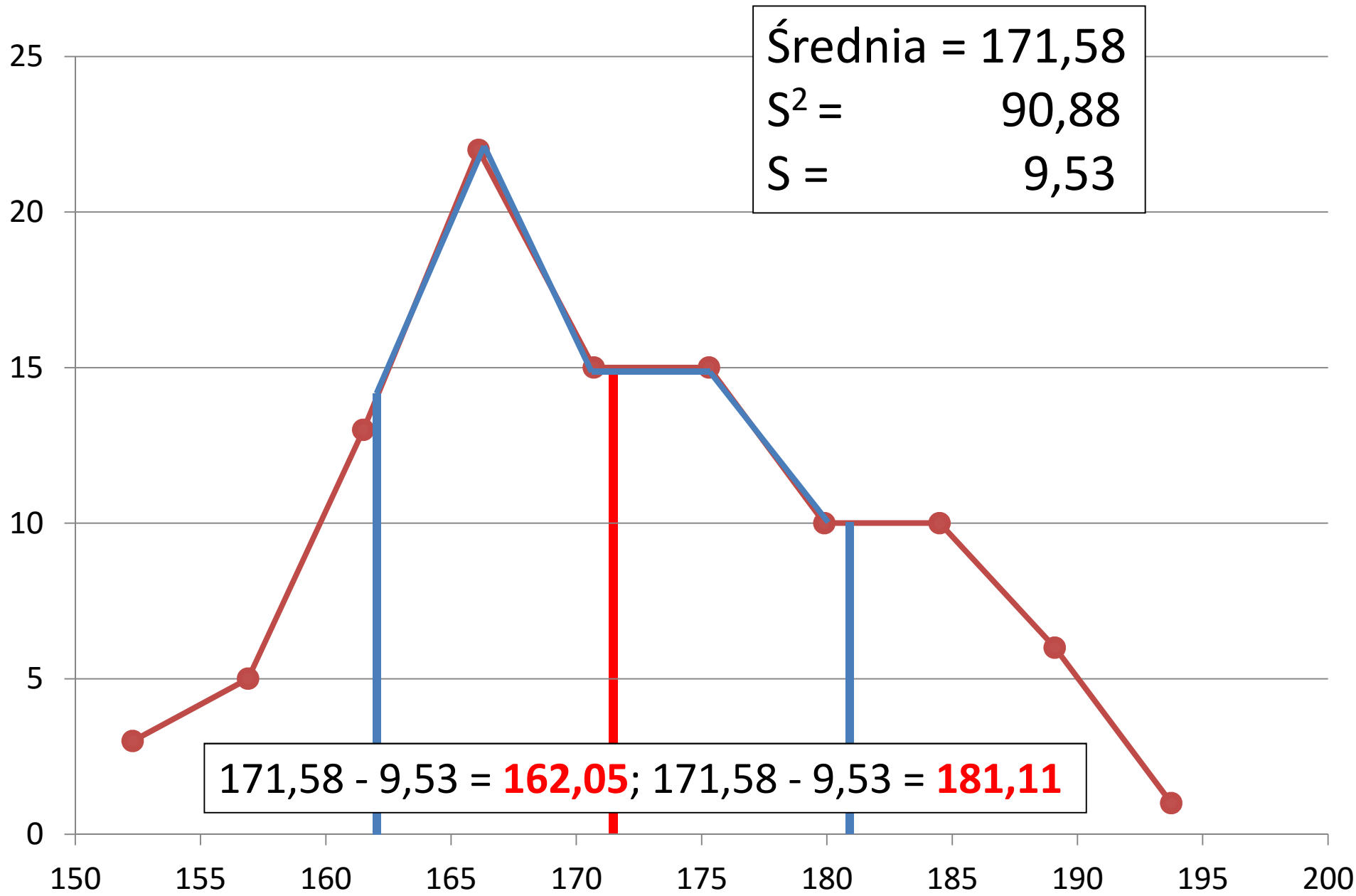

Histogram of Zm
Frequency
0
100
200
300
0
2
4
6
8
10
Zm

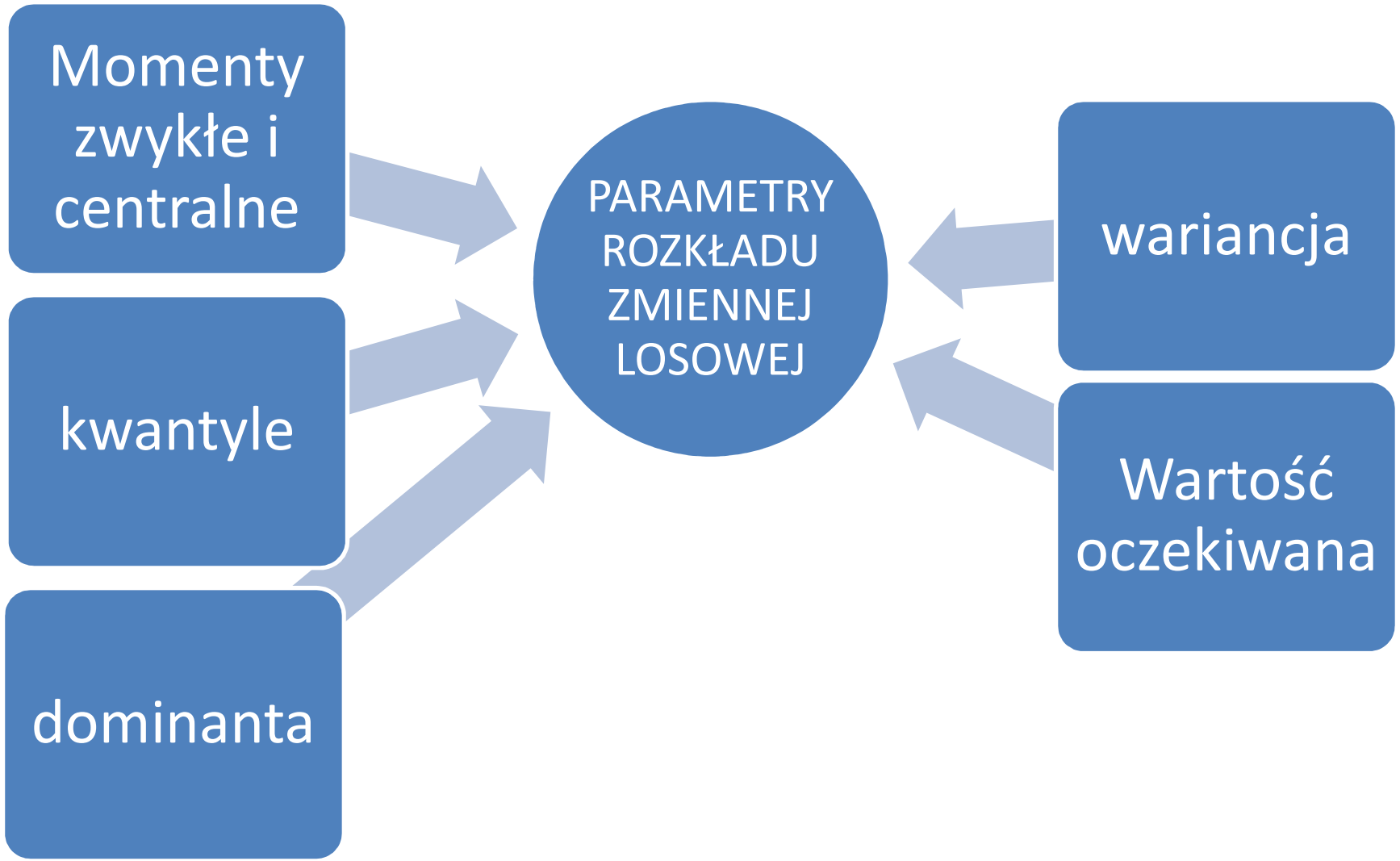

Histogram of Zm
Frequency
0
50
100
150
200
250
0
5
10
15
Zm


Histogram of Zm
Frequency
0
50
100
150
200
0
20
40
60
Zm


Histogram of Zm
Frequency
0
50
100
150
0
1
2
3
Zm


25
20
15
10
5
0
150 - 154,6
154,6 - 159,2
159,2 - 163,8
163,8 - 168,4
168,4 - 173
173 - 177,6
177,6 - 182,2
182,2 - 186,8
186,8 - 191,4
191,4 - 196


Średnia = 171,58
S2 = 90,88
S = 9,53
171,58 - 9,53 = 162,05; 171,58 - 9,53 = 181,11
25
20
15
10
5
0
150
155
160
165
170
175
180
185
190
195
200


Momenty zwykłe i centralne
kwantyle
dominanta
PARAMETRY ROZKŁADU ZMIENNEJ LOSOWEJ
wariancja
Wartość oczekiwana

Momenty zwykłe i centralne

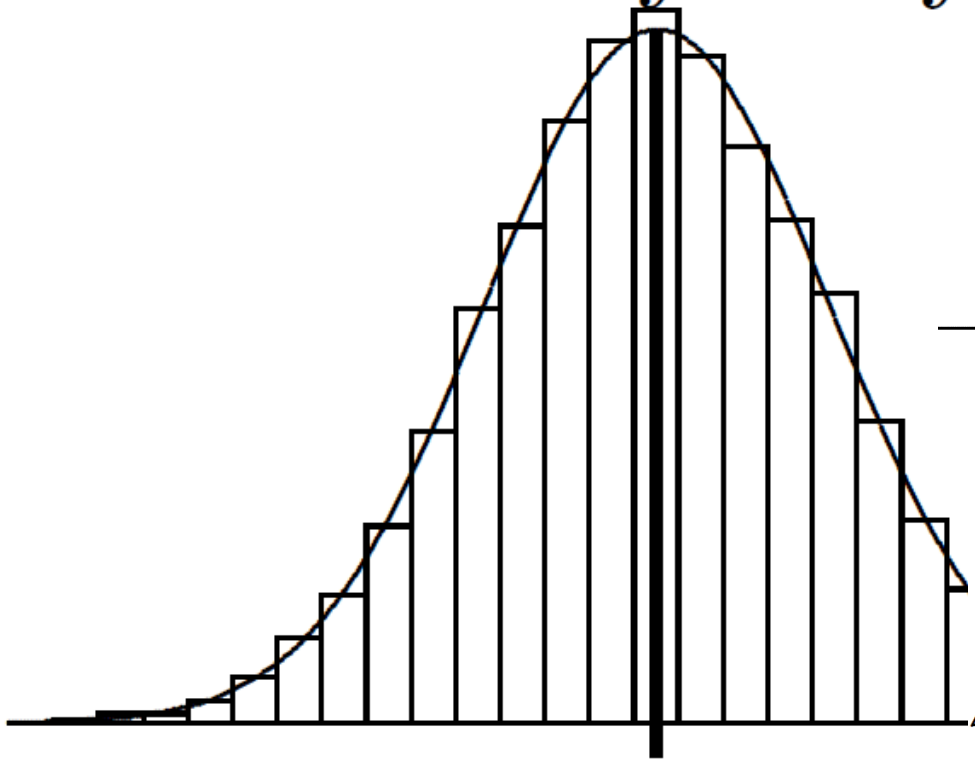
$$m_r = E(X^r) = \int_{-\infty}^{\infty} x^r \, dF(x) = \begin{cases} \sum_i x_i^r \, p_i \\ \int_{-\infty}^{\infty} x^r f(x) dx \end{cases}$$

$$\mu_r = E(X - E(X))^r = \begin{cases} \sum_i (x_i - E(x))^r \, p_i \\ \int_{-\infty}^{\infty} (x - E(x))^r f(x) dx \end{cases}$$

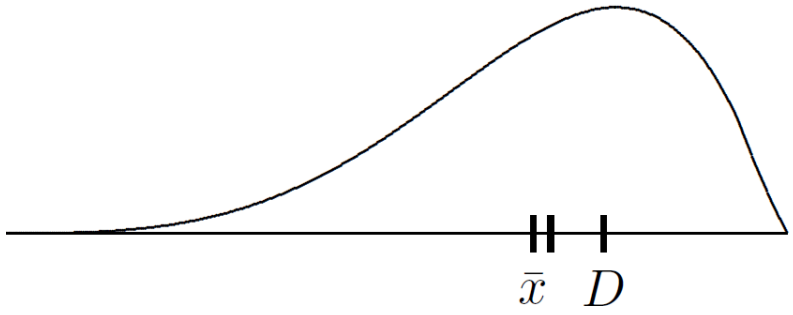
Momenty centralne

- Momentem centralnym nazywamy średnią arytmetyczną z odchyleń poszczególnych wartości zmiennej od średniej arytmetycznej podniesionych do r-tej potęgi.
- Moment centralny drugiego rzędu nazywamy wariancję
- Moment centralny trzeciego rzędu nazywamy współczynnik asymetrii obserwacji (współczynnik skośności)
- Moment centralny czwartego rzędu nazywamy miarę koncentracji obserwacji (współczynnik kurtozy)

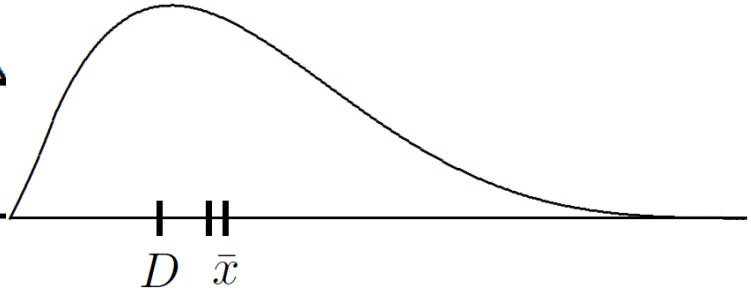
Rozkład symetryczny



Asymetria lewostronna



Asymetria prawostronna

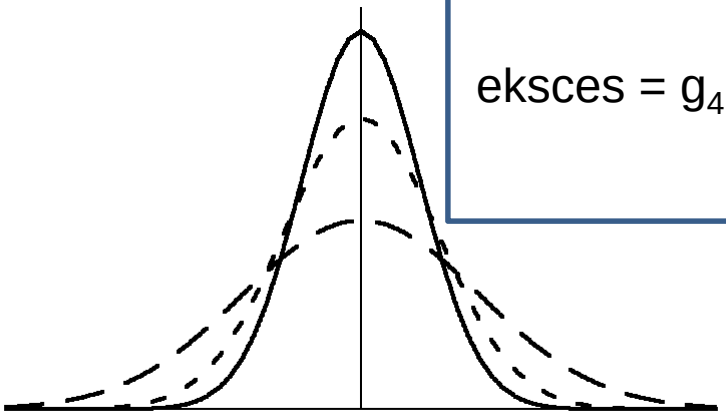


$$\bar{x} = Me = D$$

$$As = \frac{\bar{x} - D}{s} \qquad As_Q = \frac{Q_3 + Q_1 - 2Me}{2Q}$$

Kurtoza

$$g_4 = \frac{\frac{1}{N}\sum(x_i - \bar{x})^4}{S^4}$$



eksces = g_4 - 3

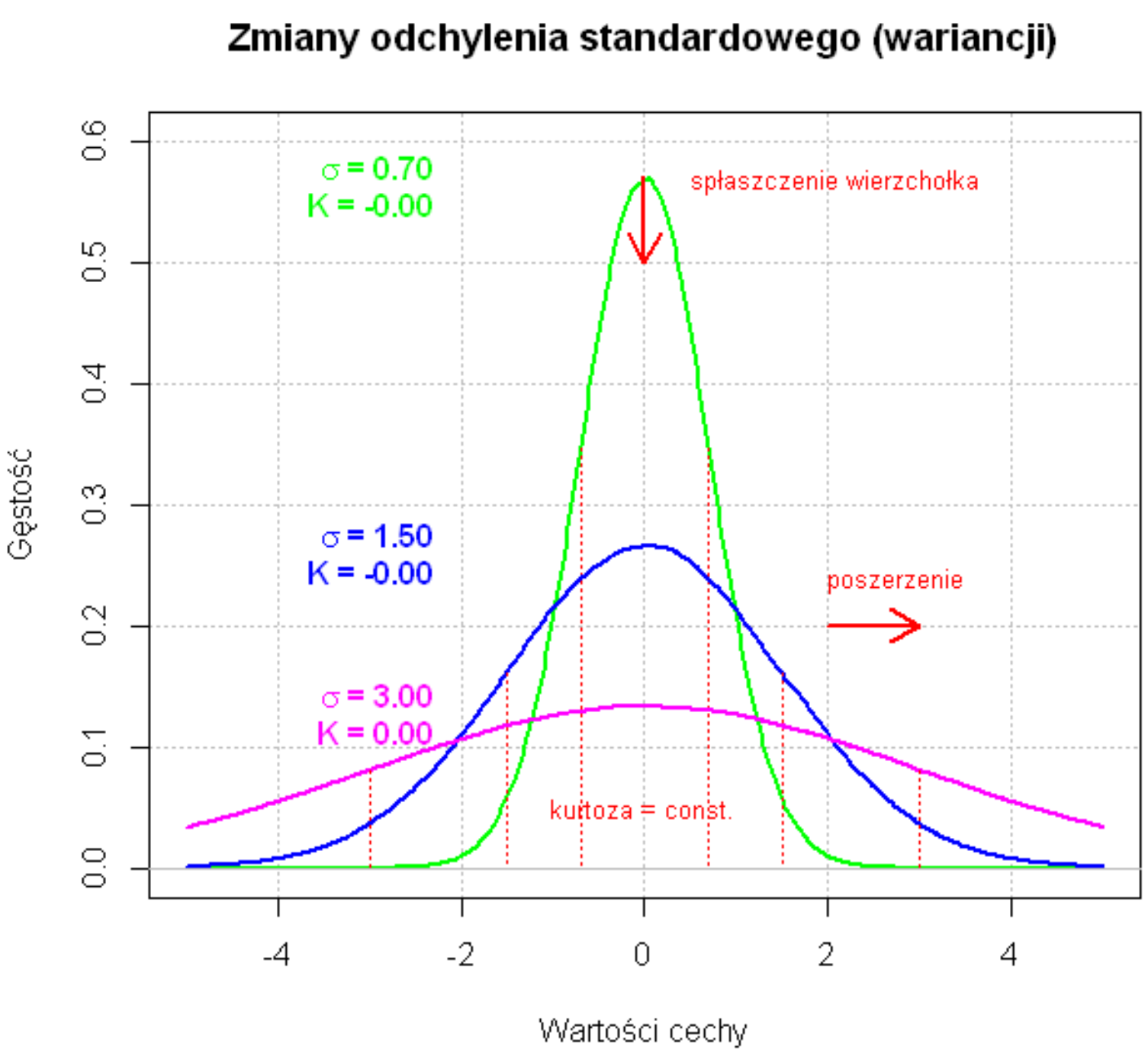
Kurtoza informuje właściwie o tym czy dane są bardziej w centralnej części rozkładu, czy w ogonach.

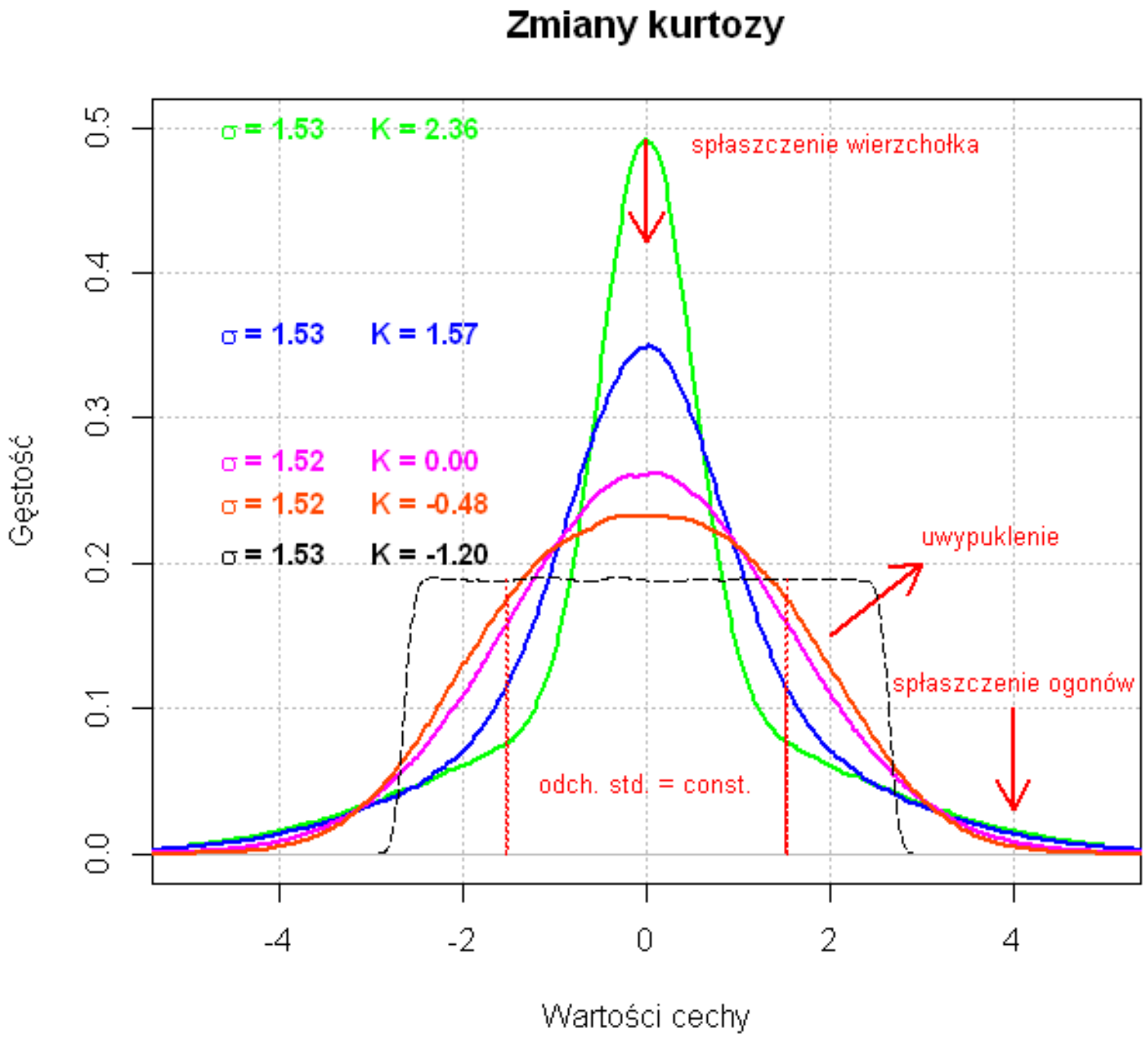
Rozkłady prawdopodobieństwa można podzielić ze względu na wartość kurtozy na rozkłady:

mezokurtyczne - wartość kurtozy wynosi 3, spłaszczenie rozkładu jest podobne do spłaszczenia rozkładu normalnego (dla którego ekscesu wynosi dokładnie 0)

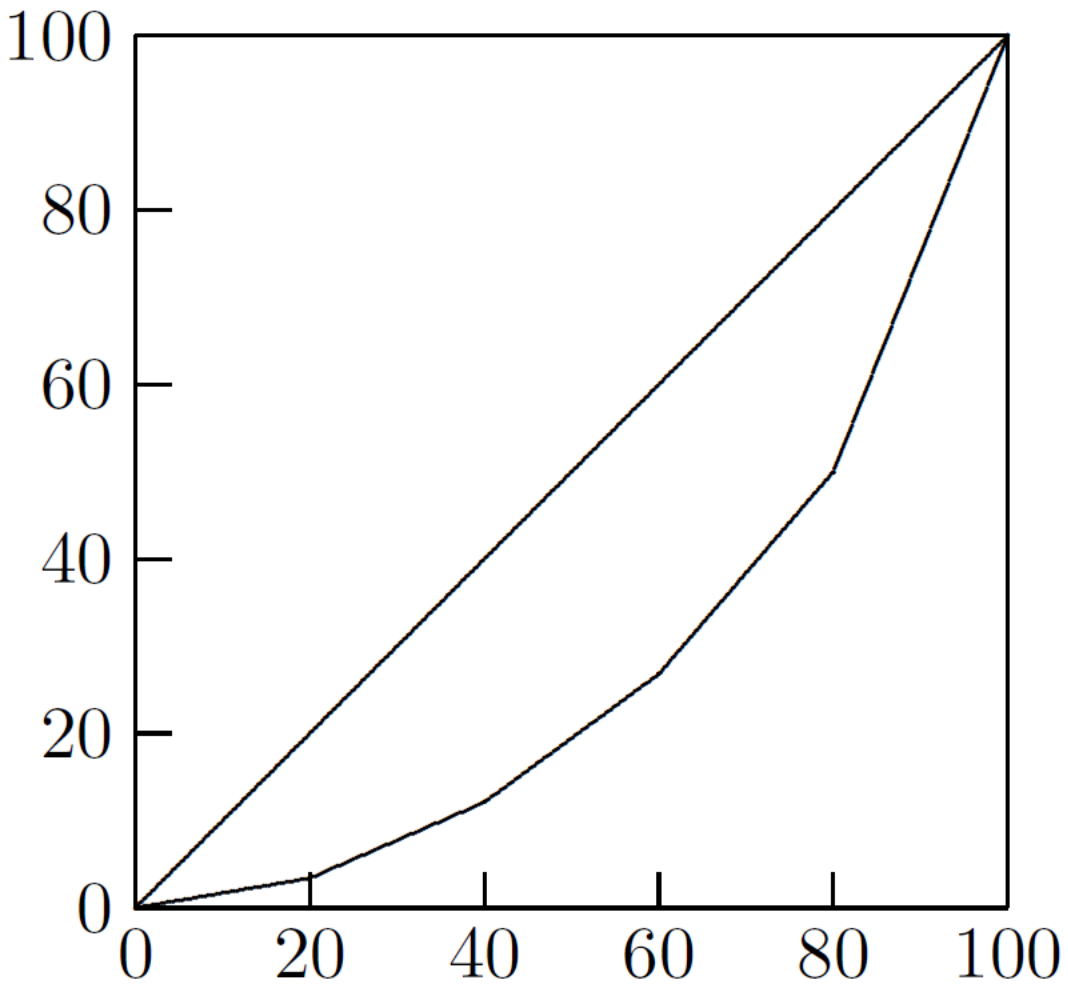
leptokurtyczne – wartość kurtozy większa od 3 (eksces jest dodatni), wartości cechy bardziej skoncentrowane niż przy rozkładzie normalnym

platokurtyczne - wartość kurtozy mniejsza od 3 (eksces jest ujemny), wartości cechy mniej skoncentrowane niż przy rozkładzie normalnym


Zmiany odchylenia standardowego (wariancji)
σ = 0.70
K = -0.00
spłaszczenie wierzchołka
σ = 1.50
K = -0.00
poszerzenie
σ = 3.00
K = 0.00
kurtoza = const.
Gęstość
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-4
-2
0
2
4
Wartości cechy


Zmiany kurtozy
σ = 1.53 K = 2.36
σ = 1.53 K = 1.57
σ = 1.52 K = 0.00
σ = 1.52 K = -0.48
σ = 1.53 K = -1.20
spłaszczenie wierzchołka
uwypuklenie
spłaszczenie ogonów
odch. std. = const.
Gęstość
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
4
Wartości cechy

Krzywa koncentracji Lorentza



Współczynnik Giniego

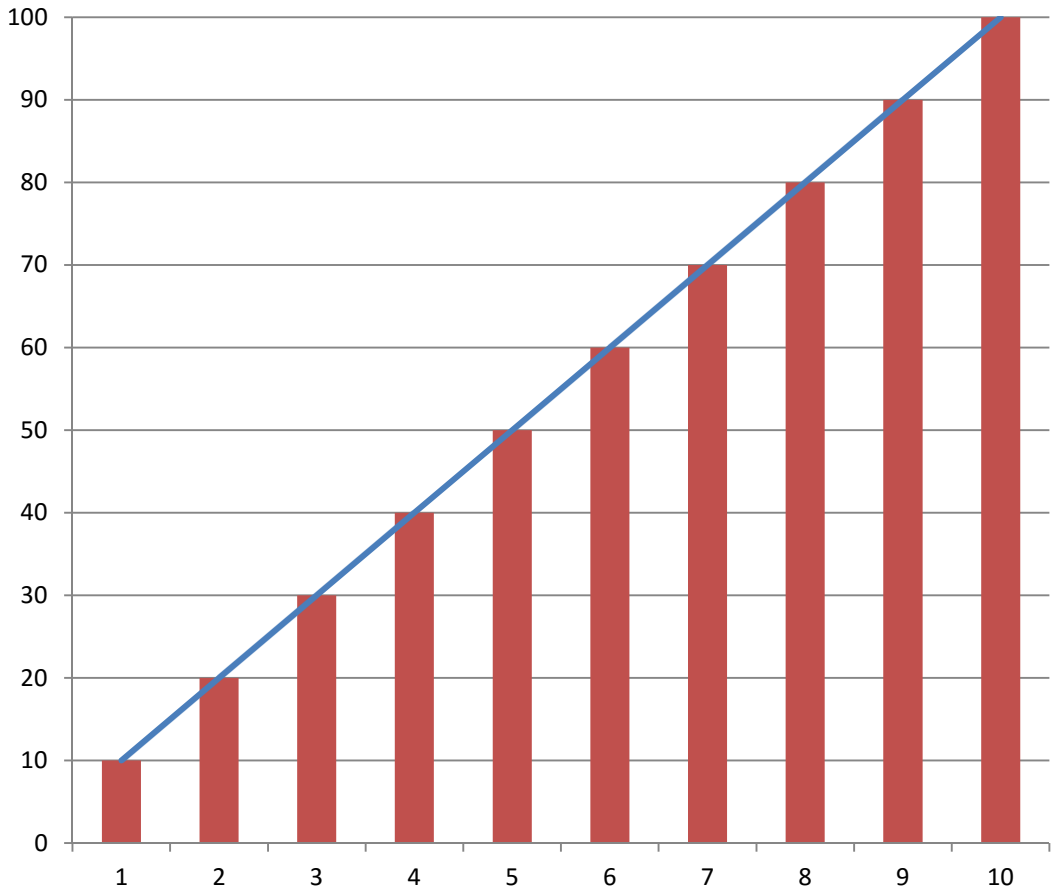
$$G = |1 - \sum_{i=1}^{k} (x_i - x_{i-1})(y_i + y_{i-1})|$$

x_i skumulowany udział w zbiorowości

y_i skumulowany udział w sumie wartości

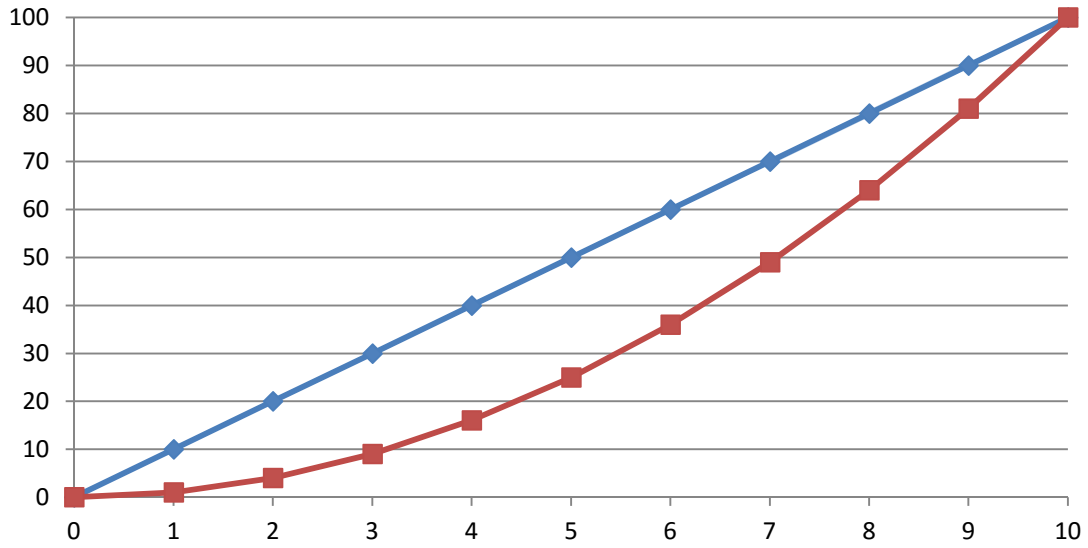
Przykład

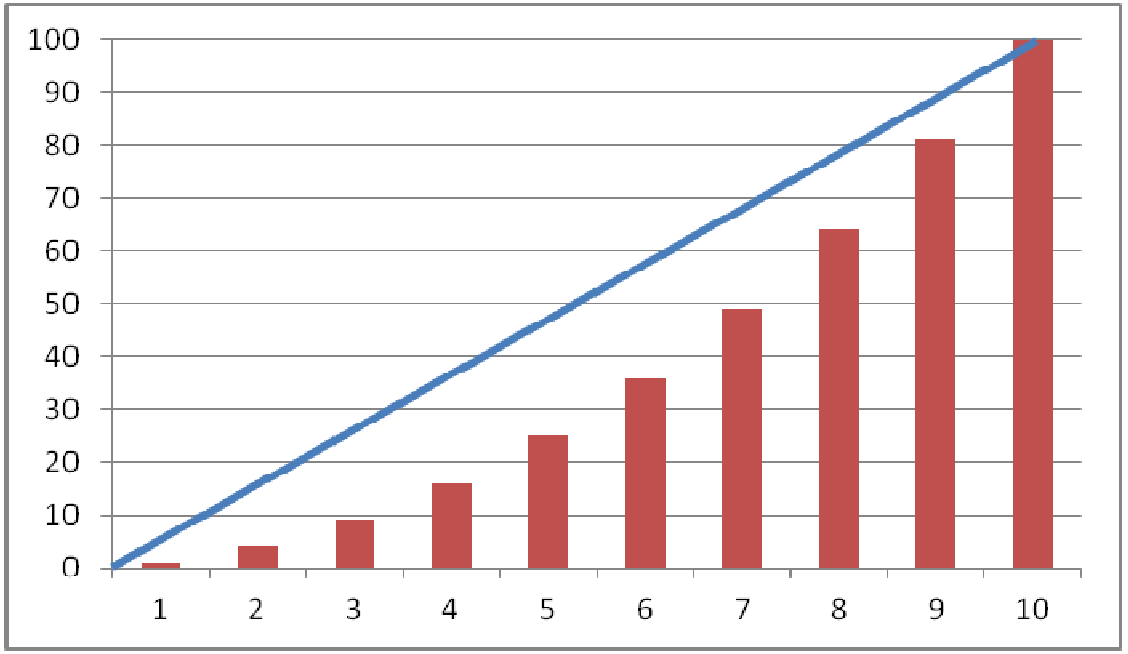
Decyle	Dochód [%]	Dochód skumulowany [%]
1	10	10
2	10	20
3	10	30
4	10	40
5	10	50
6	10	60
7	10	70
8	10	80
9	10	90
10	10	100



Przykład

Decyle	Dochód [%]	Dochód skumulowany [%]
1	1	1
2	3	4
3	5	9
4	7	16
5	9	25
6	11	36
7	13	49
8	15	64
9	17	81
10	19	100





PYTANIA

1. Wymień znane Ci teoretyczne rozkłady prawdopodobieństwa wykorzystywane w statystyce matematycznej.
2. Znając rozkład popytu na pewien towar określ oczekiwany zysk wiedząc, że cena sprzedaży wynosi 10, a koszty stałe 10000.

Popyt	1000	2000	3000	4000
P(popytu)	0,4	0,3	0,2	0,1

3. Jaka jest interpretacja pojecia kurtozy?
4. Wyjaśnij pojęcie asymetria prawostronna.
5. Dwaj niezależni analitycy badali zużycie paliwa w pewnej firmie. Stwierdzili, że badana cecha ma rozkład normalny. W wyniku obliczeń analityk A stwierdził, że zużycie paliwa charakteryzuje się silną prawostronną asymetrią, a analityk B który liczył pozycyjny współczynnik asymetrii stwierdził, że jest tam silna asymetria lewostronna. Który z nich miał rację ? Odpowiedź uzasadnij.
6. Badając rozkład dochodów w pewnym powiecie uzyskano następujące udziały w łącznych dochodach dla kolejnych części zbiorowości (równych pod względem liczebności): 5%, 10%, 20%, 65%. Ile wynosi współczynnik Giniego?
7. Wykreśl krzywą Lorenza w oparciu o dane z pytania 6. Jak wyglądałby taki wykres gdyby w całej zbiorowości tylko jedna osoba miała dochody?
8. Co oznacza określenie rozkład leptokurtyczny?
9. W pewnej zbiorowości wyznaczono średnia wartość badanej zmiennej, odchylenie standardowe i dominantę uzyskując odpowiednio: 100, 9 i 113.5. Oblicz wartość współczynnika asymetrii.
10. Wzrost kobiet jest zmienną losową o rozkładzie normalnym ze średnią 166 i wariancją 400. Jaki procent kobiet będzie miał wzrost z przedziału od 146 do 186 centymetrów?