AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Wprowadzenie Przekształcenia danych

Statystyka w medycynie

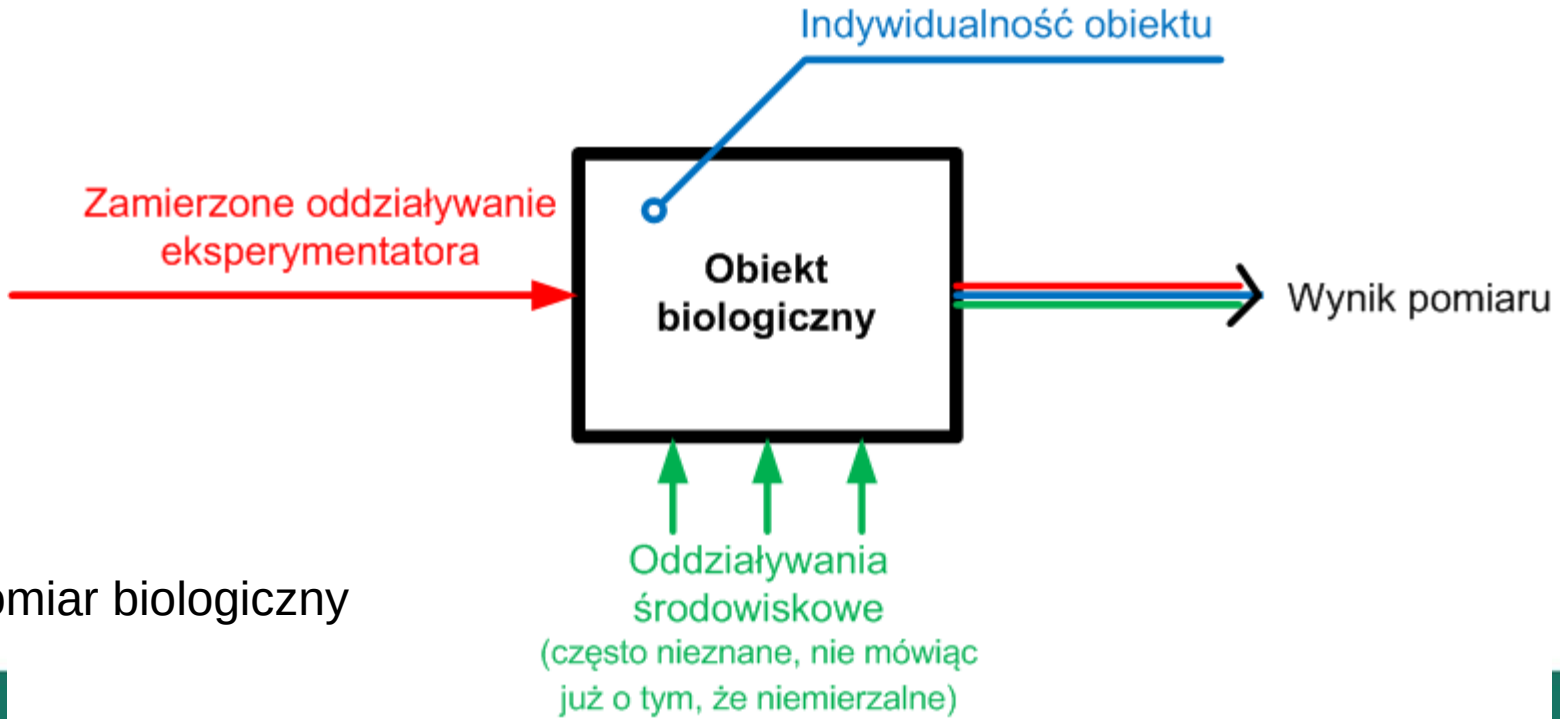
Dr inż. Janusz Majewski
Katedra Informatyki

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

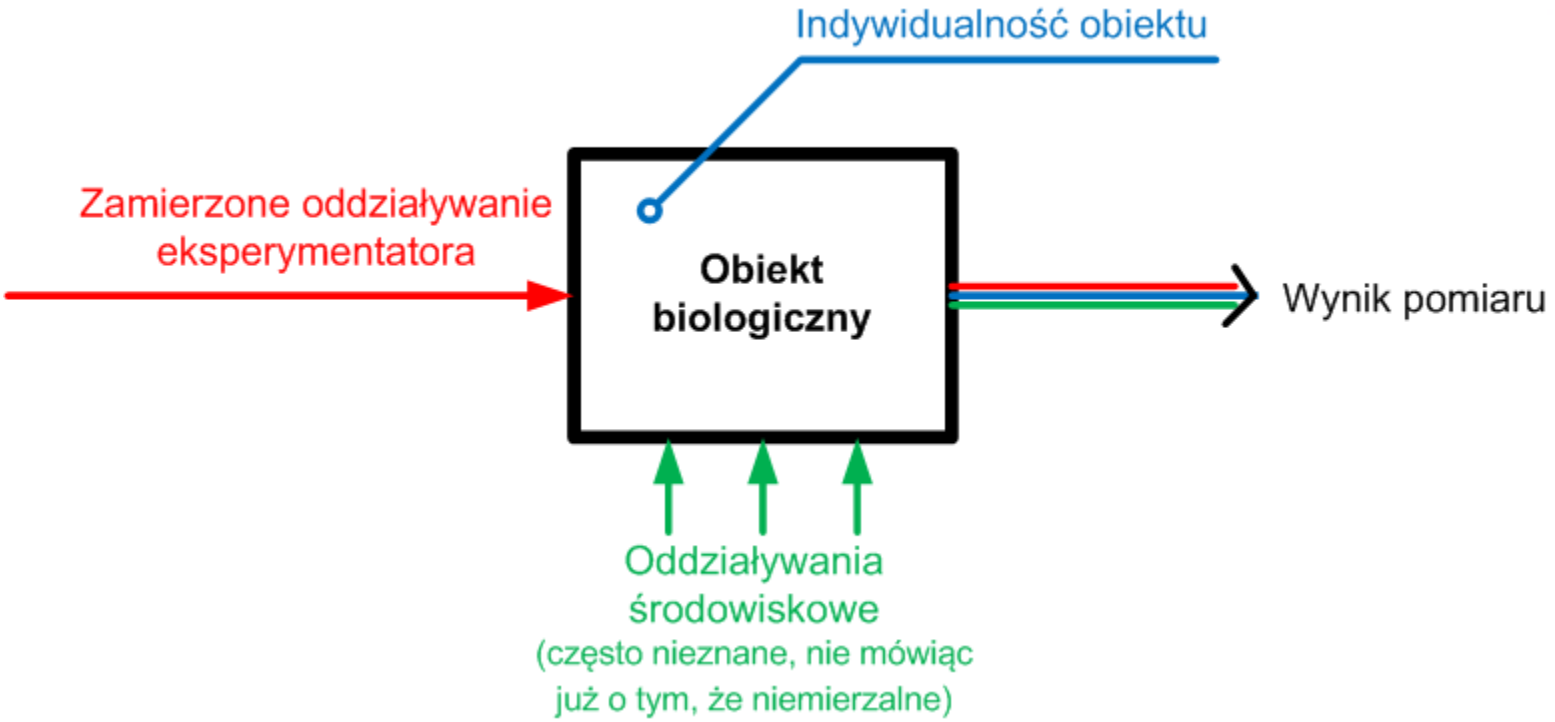
Wprowadzenie - biometria

Biometria – nie jest działem metrologii zajmującym się sposobami przeprowadzania pomiarów parametrów i cech różnych systemów biologicznych. Biometria jest nauką o tym, jak zaplanować eksperyment pomiarowy i co zrobić z wynikami pomiaru systemu biologicznego.



Rys. 1. Pomiar biologiczny

Wprowadzenie - biometria



Wynik pomiaru jest niepewny. Biometria mówi nam o tym, jak precyzyjnie wnioskować na podstawie nieprecyzyjnych danych.

Wprowadzenie - biometria

Środki, którymi dysponujemy:

- Pomiary wykonujemy nie dla pojedynczych obiektów, lecz dla pewnej liczby obiektów (próby) wybranych ze zbiorowości,
- Pomiary wykonujemy wielokrotnie, eksperymenty badawcze dublujemy,
- Wykorzystujemy pomiary prowadzone w grupach kontrolnych dla porównań.

Narzędziem do opracowywania wyników eksperymentu badawczego jest STATYSTYKA. Statystyka jest tylko narzędziem, więc będziemy się uczyć, jak ją umiejętnie i sensownie stosować, nie będziemy zaś udowadniać żadnych twierdzeń i wyprowadzać wzorów.

Wprowadzenie - biometria

Zadaniem naszym, jako inżynierów, jest wspomaganie konsultacją i pomocą obliczeniową badań prowadzonych przez lekarzy i naukowców zatrudnionych w placówkach służby zdrowia.

Powinniśmy umieć zasugerować sensowne wykorzystanie statystyki w medycynie i przeciwstawiać się traktowaniu statystyki, jako niezbędnej „dekoracji” wszelkich prac naukowych z dziedziny medycyny.

Przygotowanie danych

Rysunek 2. 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**

Przekształcanie danych

CELE PRZEKSZTAŁCANIA DANYCH:

1) Stabilizacja wariancji (w ramach kilku grup danych różniących się średnimi mogą być różne rozrzuty, podczas gdy wymagany jest stały rozrzut w grupach). Jeśli znana jest zależność wariancji od średniej w grupach:

$$\boldsymbol{\sigma^2(x) = \Phi[E(x)]}$$

to funkcję *y=f(x)*, według której należy przekształcić dane można w przybliżeniu wyznaczyć z zależności:

$$\frac{\boldsymbol{dy}}{\boldsymbol{dx}} = \frac{\boldsymbol{const}}{\sqrt{\boldsymbol{\sigma^2(x)}}}$$

Przekształcanie danych

CELE PRZEKSZTAŁCANIA DANYCH:

2) Linearyzacja zależności między dwiema cechami: jeżeli linia regresji między dwoma zmiennymi ilościowymi wyraźnie nie posiada charakteru liniowego, to można konstruować linię regresji w postaci krzywoliniowej lub przekształcić wstępnie dane, aby otrzymać zależność liniową. W tym drugim przypadku metody analizy są szczególnie proste, a wyniki intuicyjnie zrozumiałe i łatwe do interpretacji.

Przekształcanie danych

CELE PRZEKSZTAŁCANIA DANYCH:

3) Normalizacja rozkładu. Często metoda analizy, którą zamierzamy stosować wymaga, aby dana próba została pobrana ze zbiorowości o rozkładzie normalnym (Gaussa). Niektóre metody są mało wrażliwe na nienormalność rozkładu, więc cel (3) jest na ogół mniej ważny od celów (1) i (2). Niektóre przekształcenia normalizują rozkład „przy okazji" stabilizacji wariancji czy linearyzacji linii regresji.

Najczęściej stosowane przekształcenia danych ilościowych

(1) Przekształcenie logarytmiczne

$$y = log_a x$$

Gdzie:

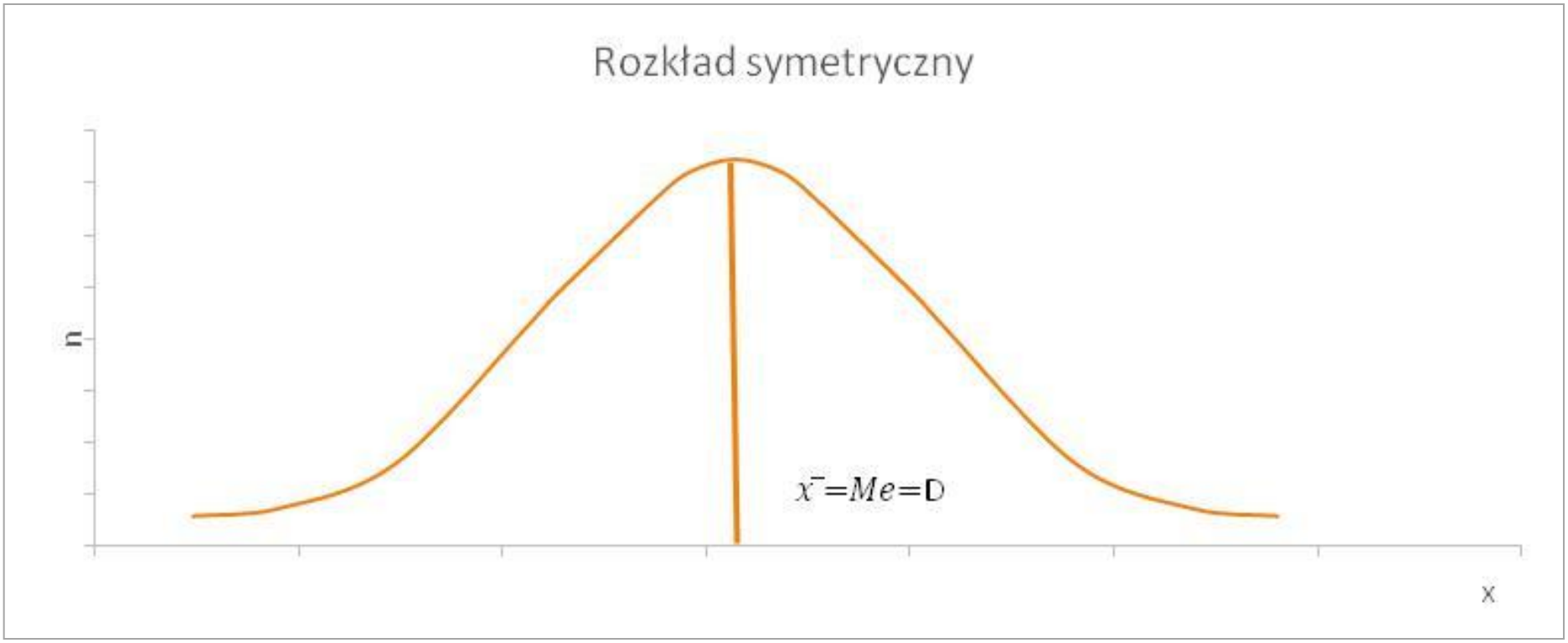
$$a = \begin{cases} 10 \\ e \end{cases}$$

→ Stabilizuje wariancję, gdy $\sigma^2(x)$ rośnie znacznie w zależności od wartości średniej

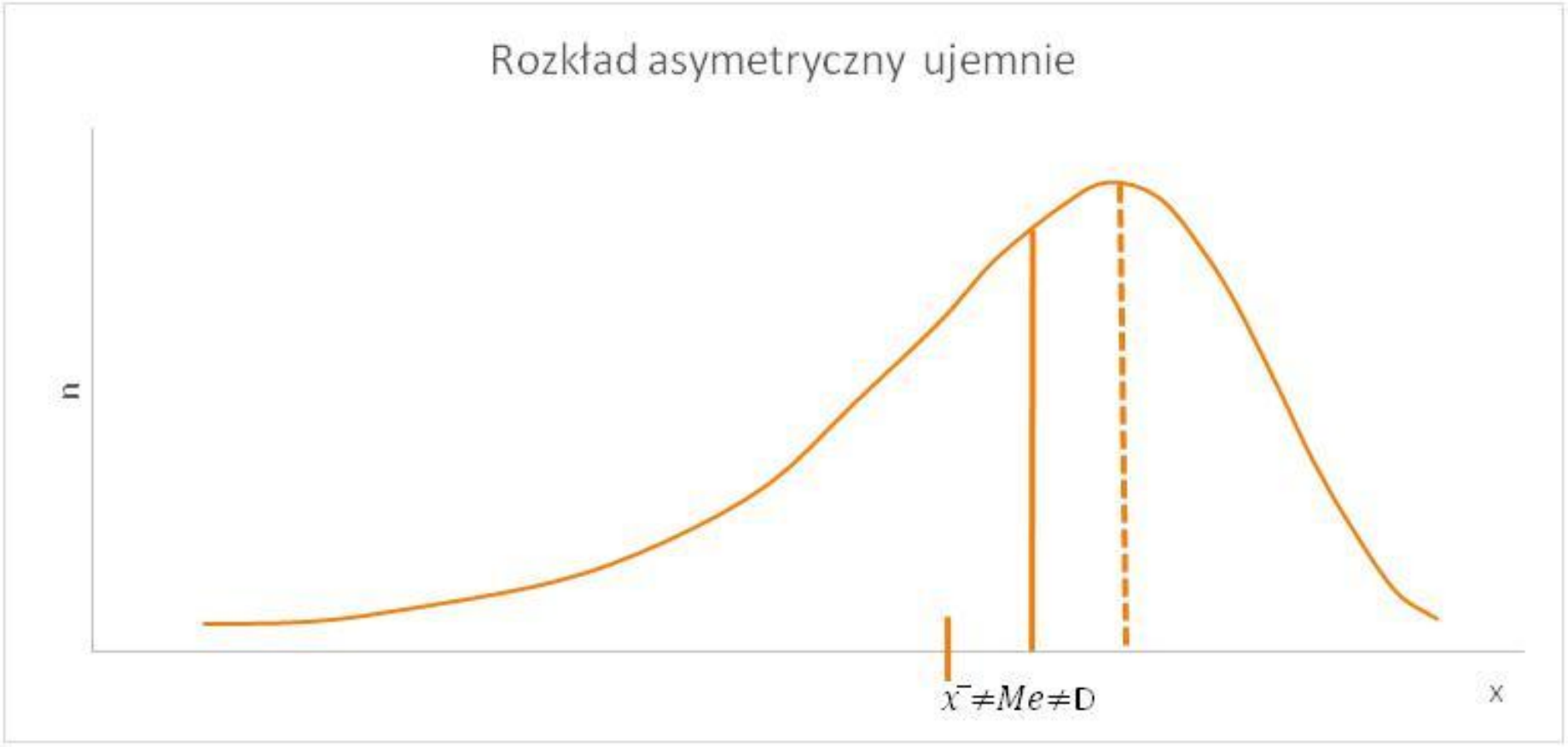
→ Linearyzuje zależności zbliżone do wykładniczych

→ Zmniejsza dodatnią asymetrię rozkładu

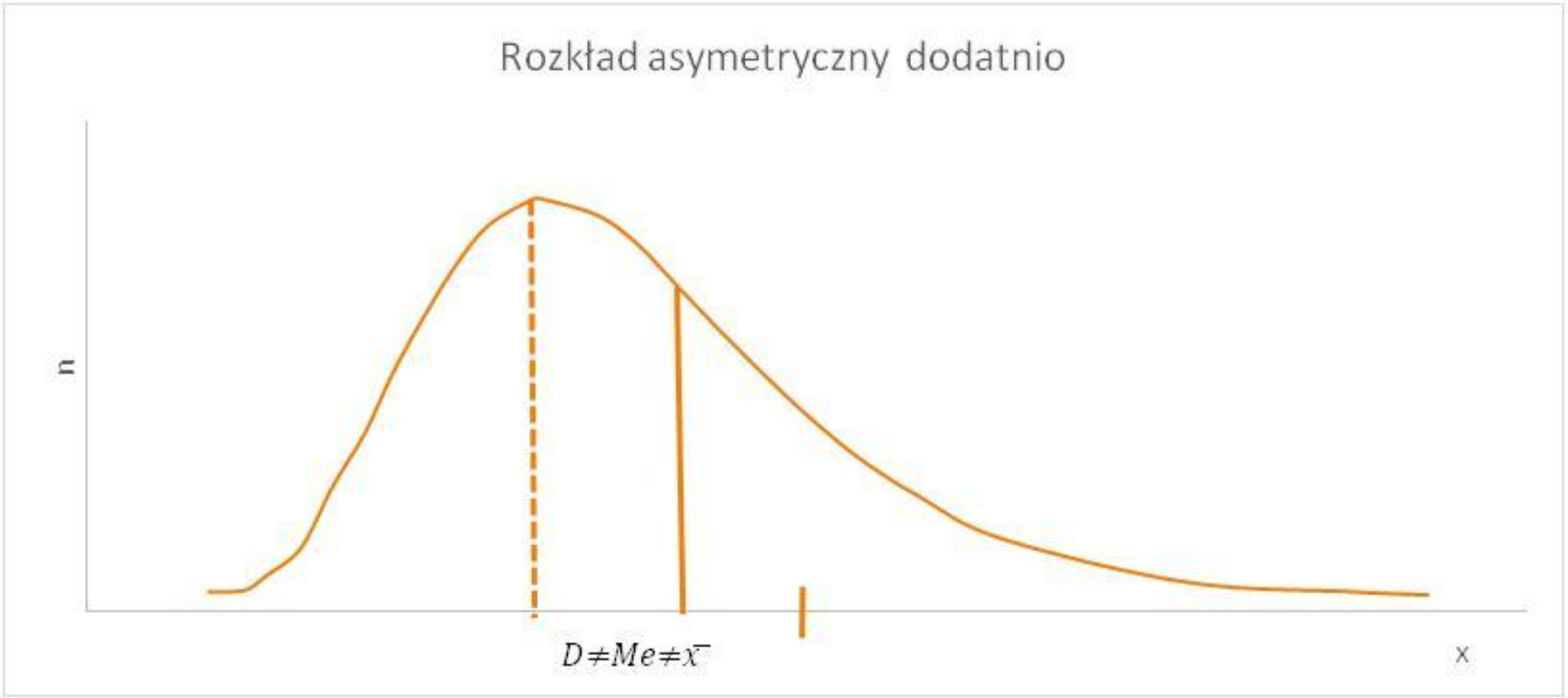
Typy rozkładów



Typy rozkładów



Typy rozkładów



Najczęściej stosowane przekształcenia danych ilościowych

Przekształcenie logarytmiczne używane jest np. wówczas, gdy średni przyrost efektu ΔE jest proporcjonalny do średniego względnego przyrostu przyczyny $\frac{\Delta P}{P}$

$$\Delta E = k\frac{\Delta P}{P} \Longrightarrow E = k\, lnP + C$$

Stosowane, gdy dane przyjmują wartości ciągu geometrycznego, np. w przypadku szeregu rozcieńczeń.

Średnia arytmetyczna danych przekształconych jest średnią geometryczną danych pierwotnych.

Najczęściej stosowane przekształcenia danych ilościowych

(2) Przekształcenie „odwrotnościowe”

$$y = \frac{1}{x}$$

→ Stabilizuje wariancję, gdy $\sigma^2(x)$ jest proporcjonalna do czwartej potęgi średniej

→ Dużym wartościom pierwotnym odpowiadają małe wartości po przekształceniu

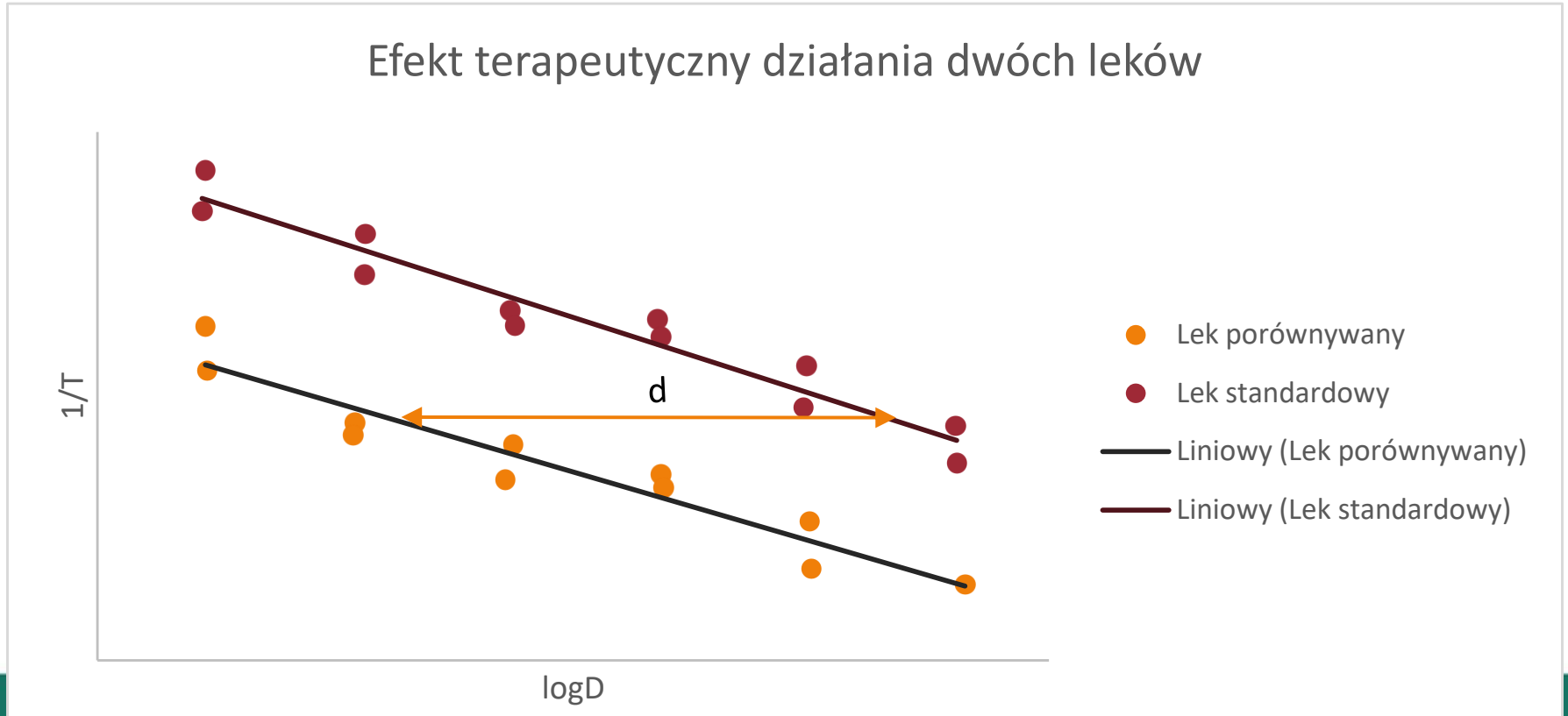
Średnia arytmetyczna danych przekształconych jest średnią harmoniczną danych pierwotnych.

Stosowane dla danych w postaci czasów przeżycia.

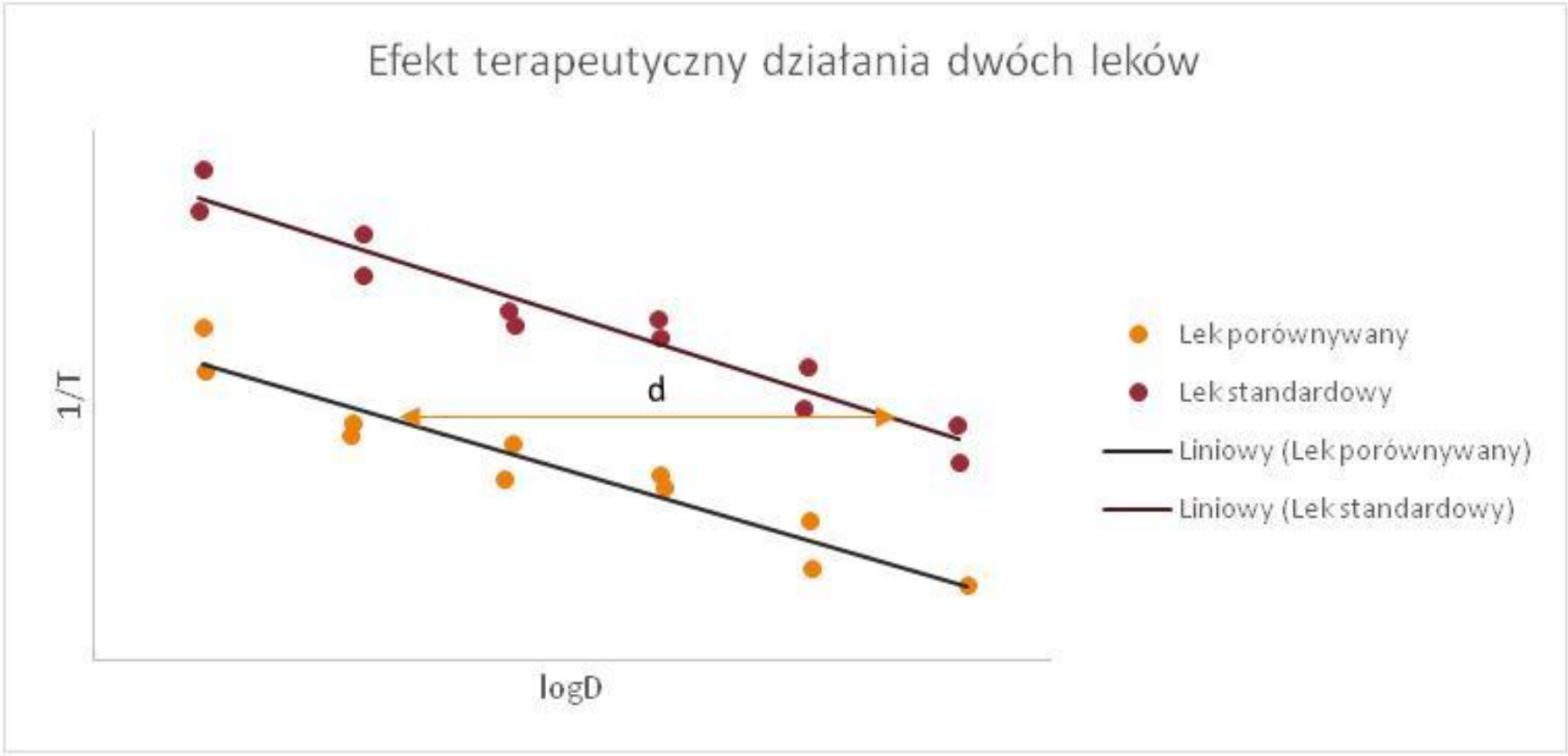
Najczęściej stosowane przekształcenia danych ilościowych

Przykład: Porównywanie efektu terapeutycznego dwu leków.

Zwierzęta doświadczalne zakażone 1000-krotną „50% dawką śmiertelną” pewnego wirusa, leczono lekiem badanym (jedna grupa) i kontrolnym, (druga grupa) o stężeniu D, notując czas przeżycia T



Najczęściej stosowane przekształcenia danych ilościowych

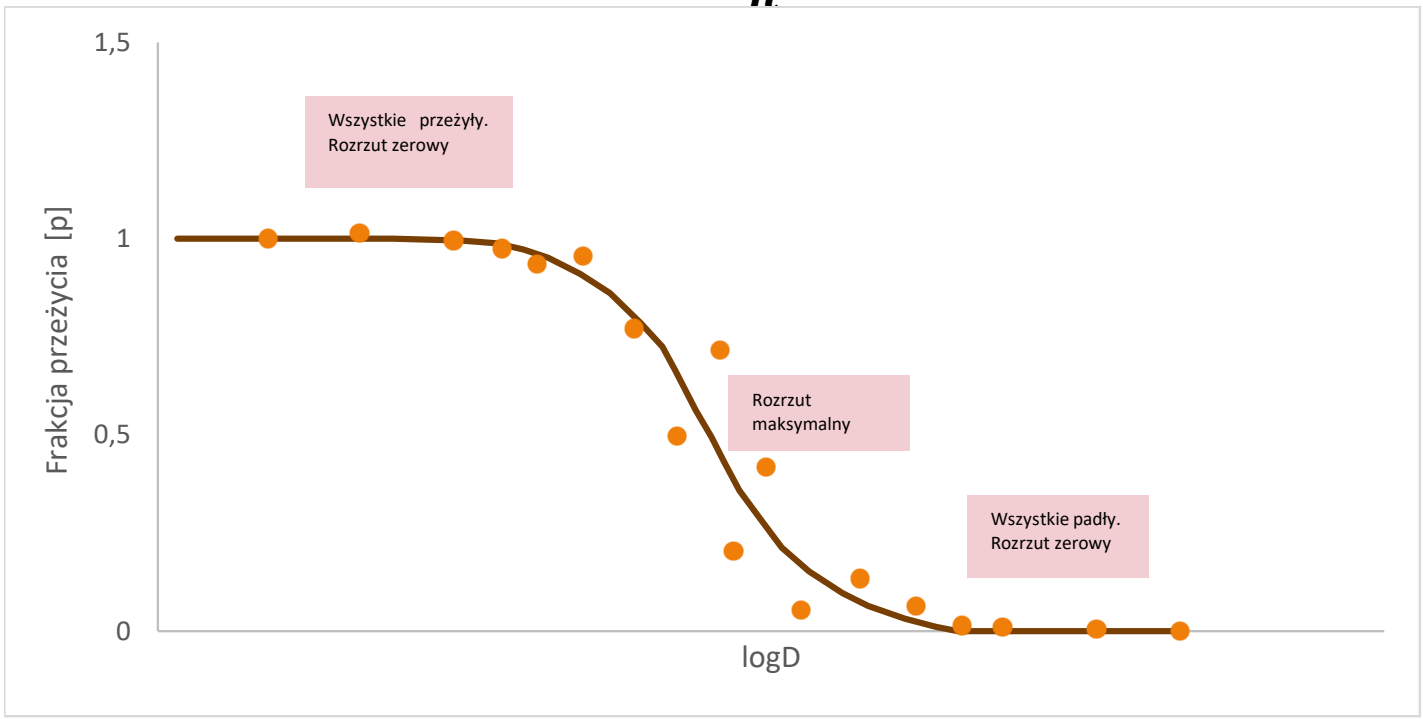


d - Po antylogarytmowaniu mówi, ile razy większe stężenie leku kontrolnego w stosunku do stężenia leku badanego powoduje ten sam efekt terapeutyczny.

Najczęściej stosowane przekształcenia frakcji

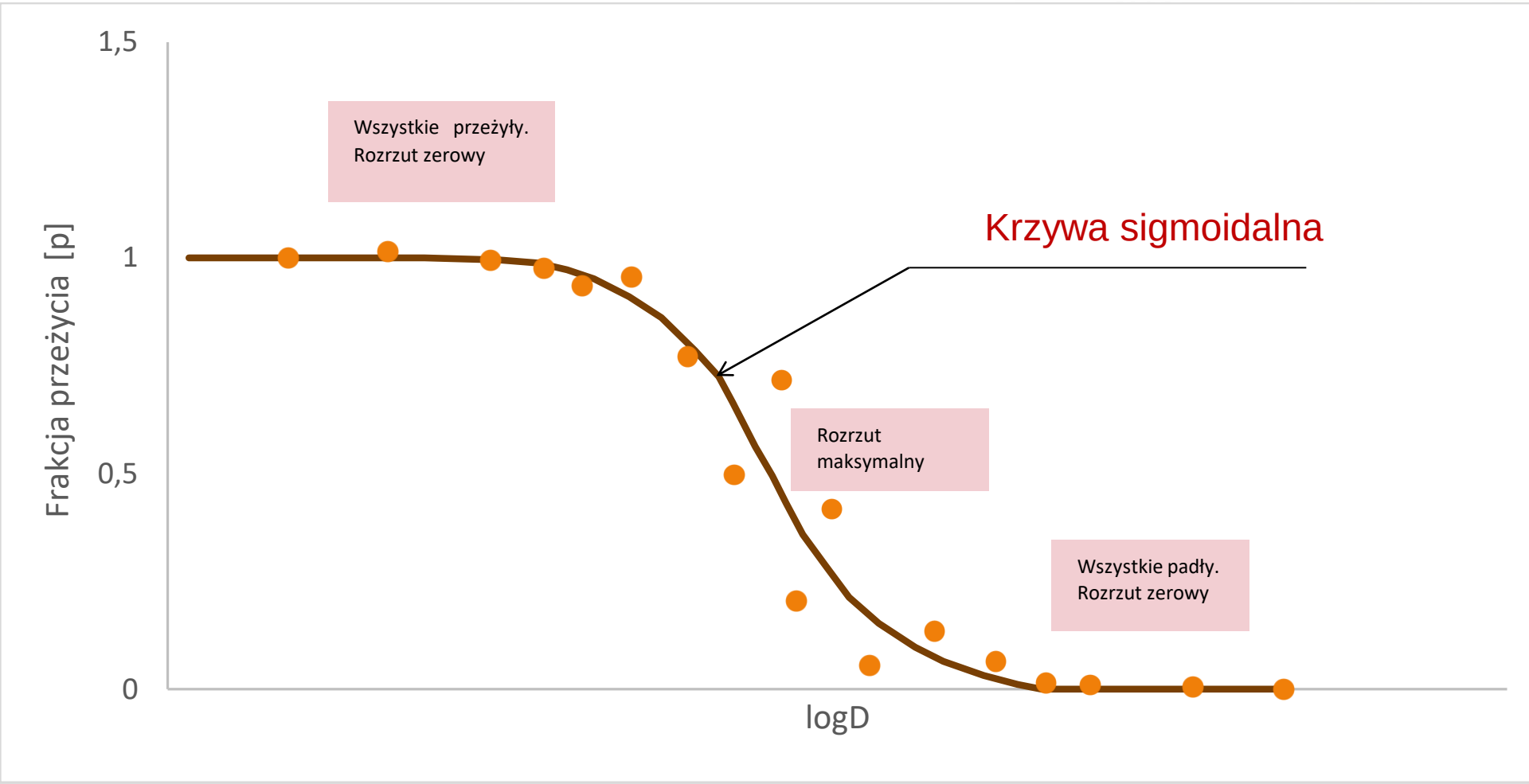
Przykład: ustalanie „50% dawki śmiertelnej” toksyny. Zwierzęta podzielono na wiele grup. Każda grupa otrzymała inne stężenie D trucizny. Notowano frakcję przeżycia

$$p = \frac{r}{n}$$



Wykres 1. Przykładowe wyniki badań skuteczności pewnej trucizny – dane w postaci frakcji wraz z dopasowaną sigmoidalną krzywą regresji.

Najczęściej stosowane przekształcenia frakcji



Najczęściej stosowane przekształcenia frakcji

Problem: linearyzacja sigmoidalnej krzywej z równoczesną stabilizacją rozrzutu

(1) Przekształcenia kątowe

$$y = arcsin\sqrt{p}$$

Dobrze stabilizuje wariancję, linearyzacja nie jest idealna.

(2) Przekształcenie logitowe

$$y = ln\frac{p}{1-p}$$

Lepiej linearyzuje krzywą sigmoidalną, nie zapewniają jednak całkowitej stabilizacji rozrzutu.

Najczęściej stosowane przekształcenia frakcji

(3) Przekształcenie <u>probitowe</u>

Obliczamy *y'* z zależności

$$\boldsymbol{p = F(y')}$$

Gdzie: *F(·)* – dystrybuanta standaryzowanego rozkładu normalnego (korzystamy z tablic)

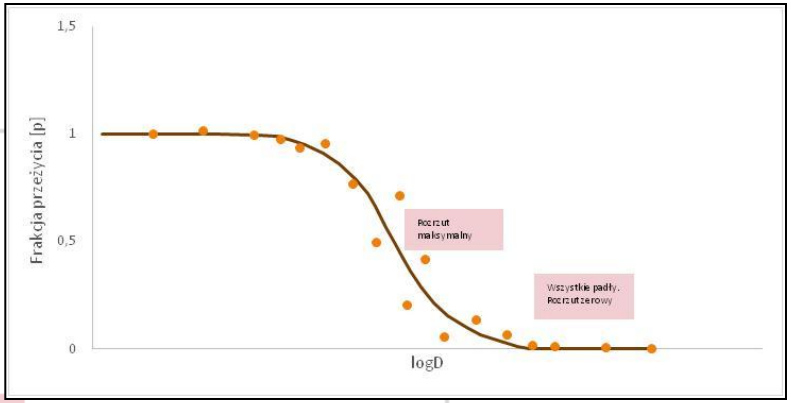
Probit *y* frakcji *p* jest definiowany jako

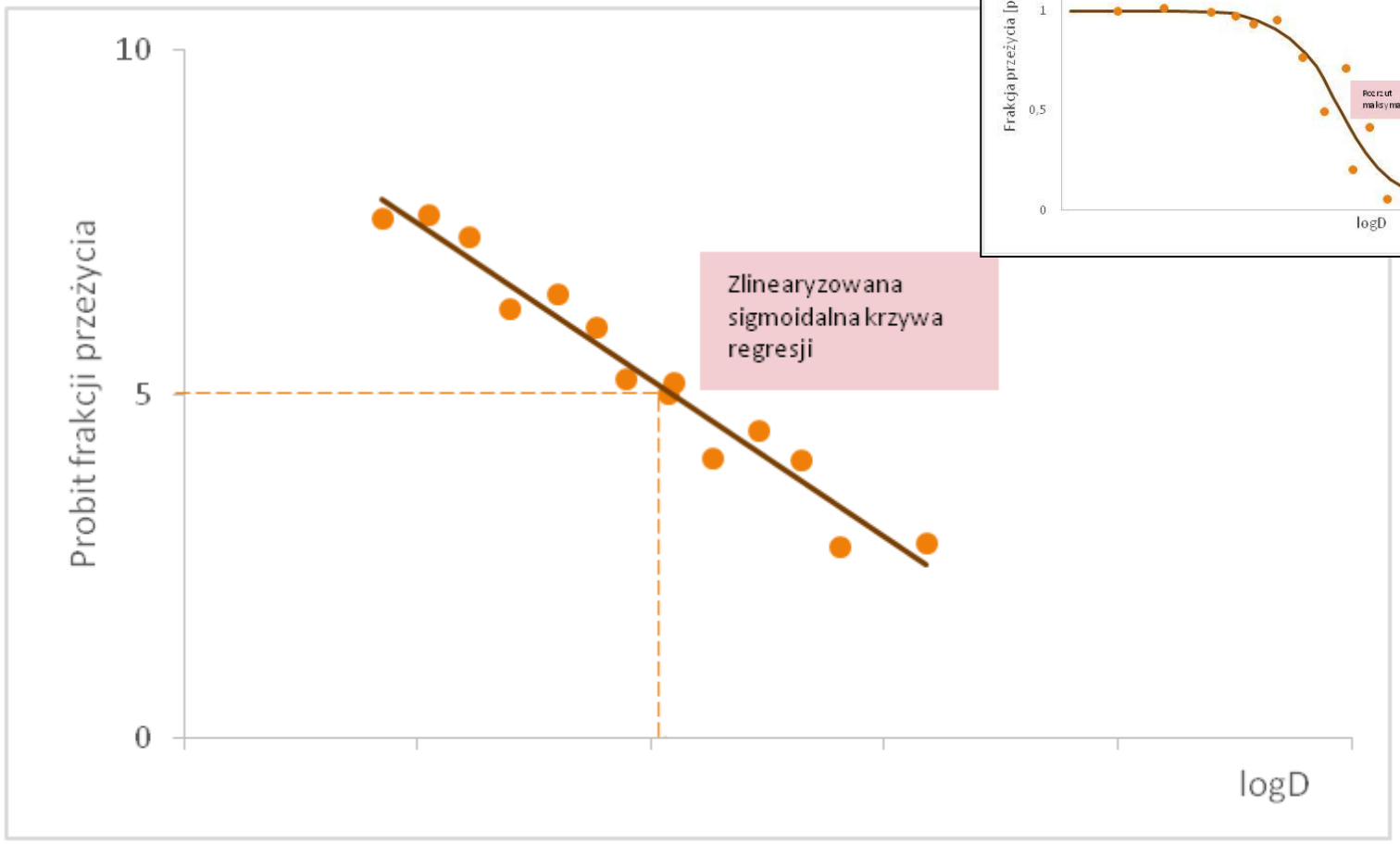
$$\boldsymbol{y = y' + 5}$$

Przekształcenie probitowe ma własności podobne do logitowego.

Najczęściej stosowane przekształcenia frakcji

Przykład c.d.:





Poprawność i kompletność danych

→ Eliminacja ewidentnych błędów grubych, np. Wzrost = 272 cm lub 27 cm

→ Ostrożna analiza obserwacji nietypowych

Reguła trzech sigm:

Niech:

μ – średnia zmiennej losowej o rozkładzie normalnym,
σ – odchylanie standardowe zmiennej losowej o rozkładzie normalnym.

Prawdopodobieństwo tego, że wartość zmiennej losowej o rozkładzie normalnym znajduje się w przedziale (μ-3σ, μ+3σ) wynosi **0,9973.**

Obserwacje spoza przedziału o promieniu 3σ i środku μ są bardzo mało prawdopodobne, jakkolwiek możliwe.

→ Pilnowanie kompletności danych (z reguły brak)

→ Ocena liczebności próby (z reguły danych jest za mało)

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

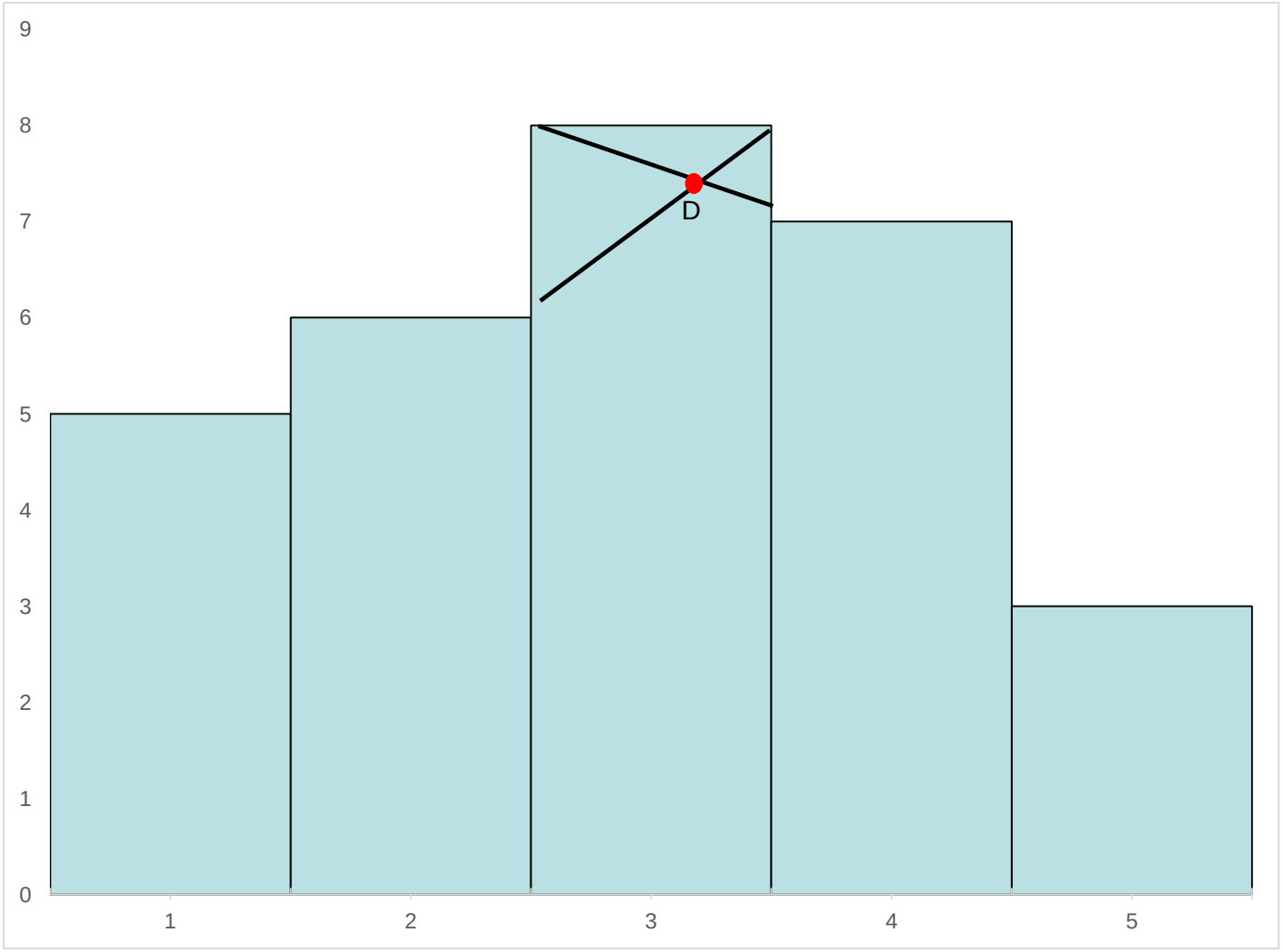
l – szerokość przedziału klasowego mody

n_d - liczebność w przedziale mody

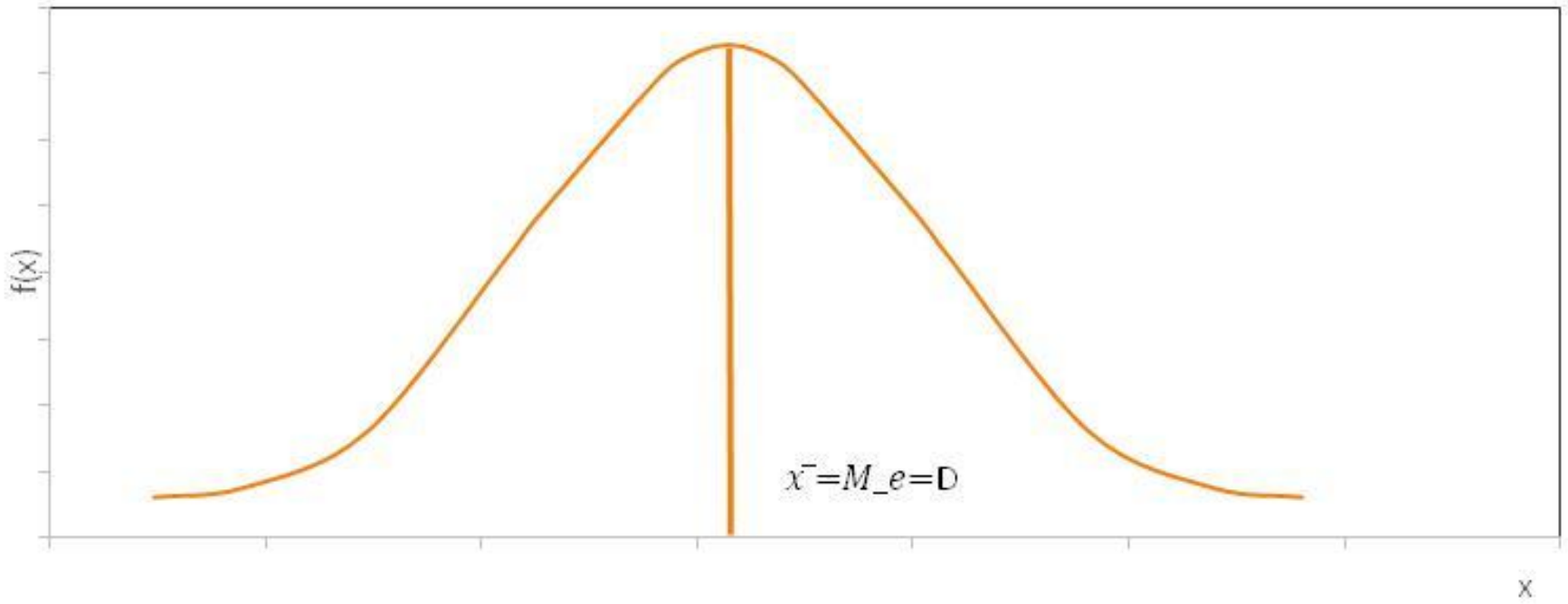
n_{d-1} - liczebność w przedziale poprzedzającym przedział mody

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

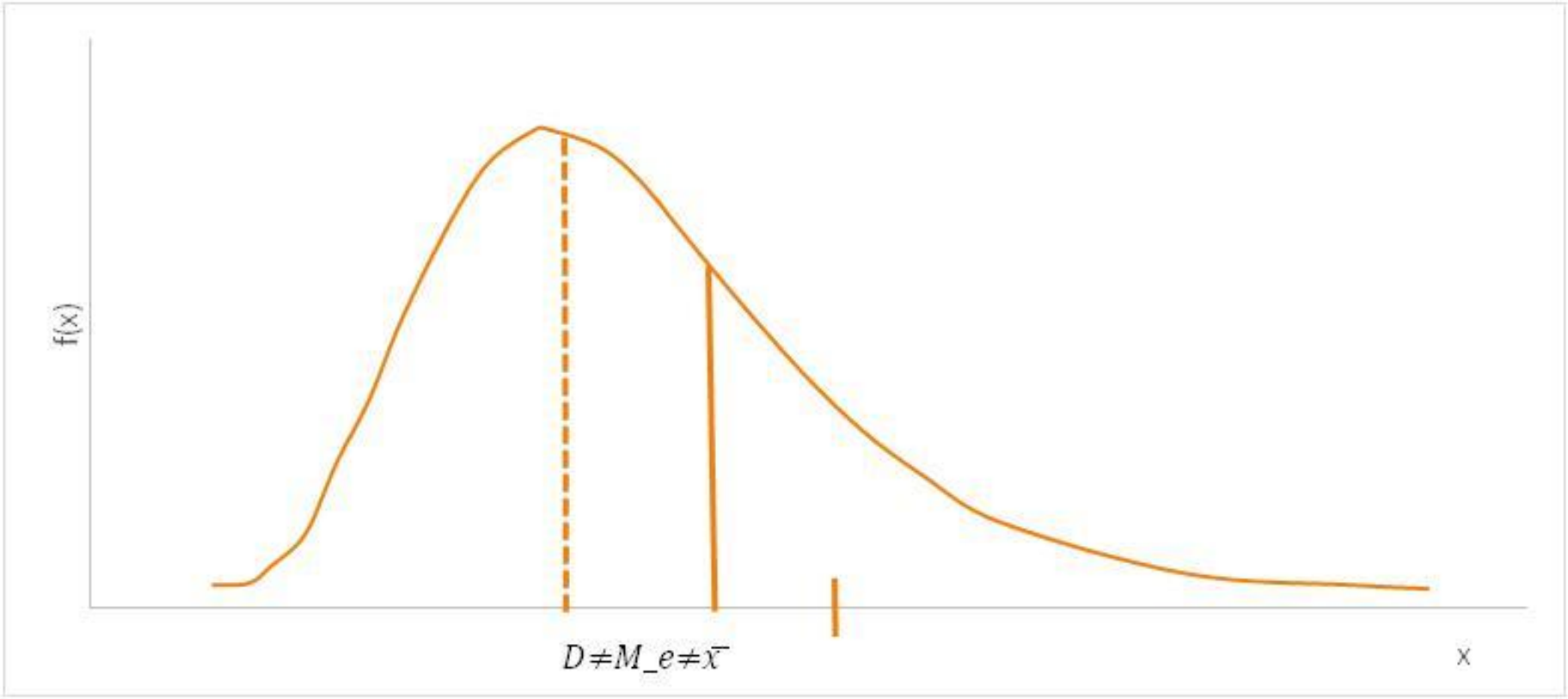
Statystyka opisowa – miary tendencji centralnej



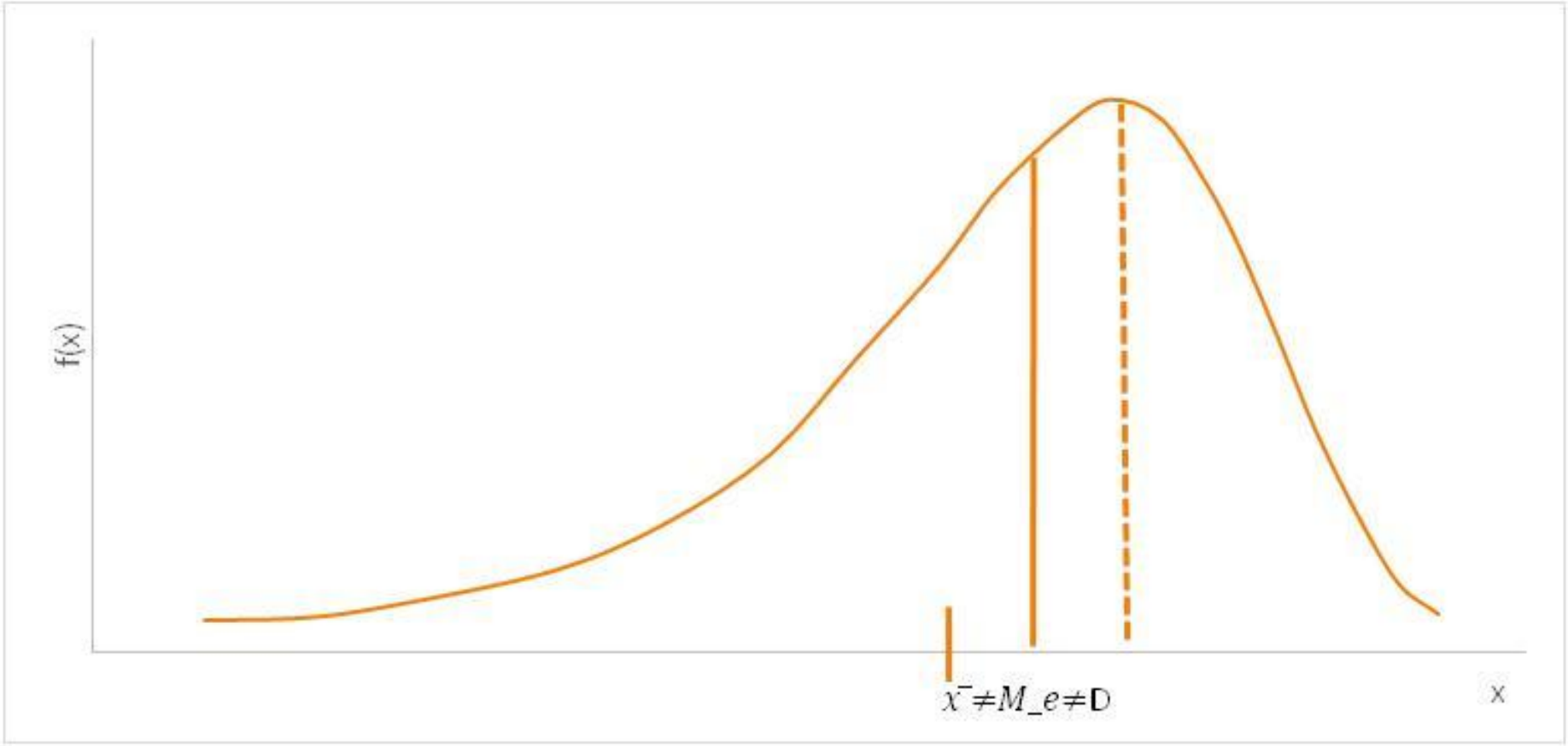
Średnia, mediana i moda rozkładu populacji



Średnia, mediana i moda rozkładu populacji



Średnia, mediana i moda rozkładu populacji

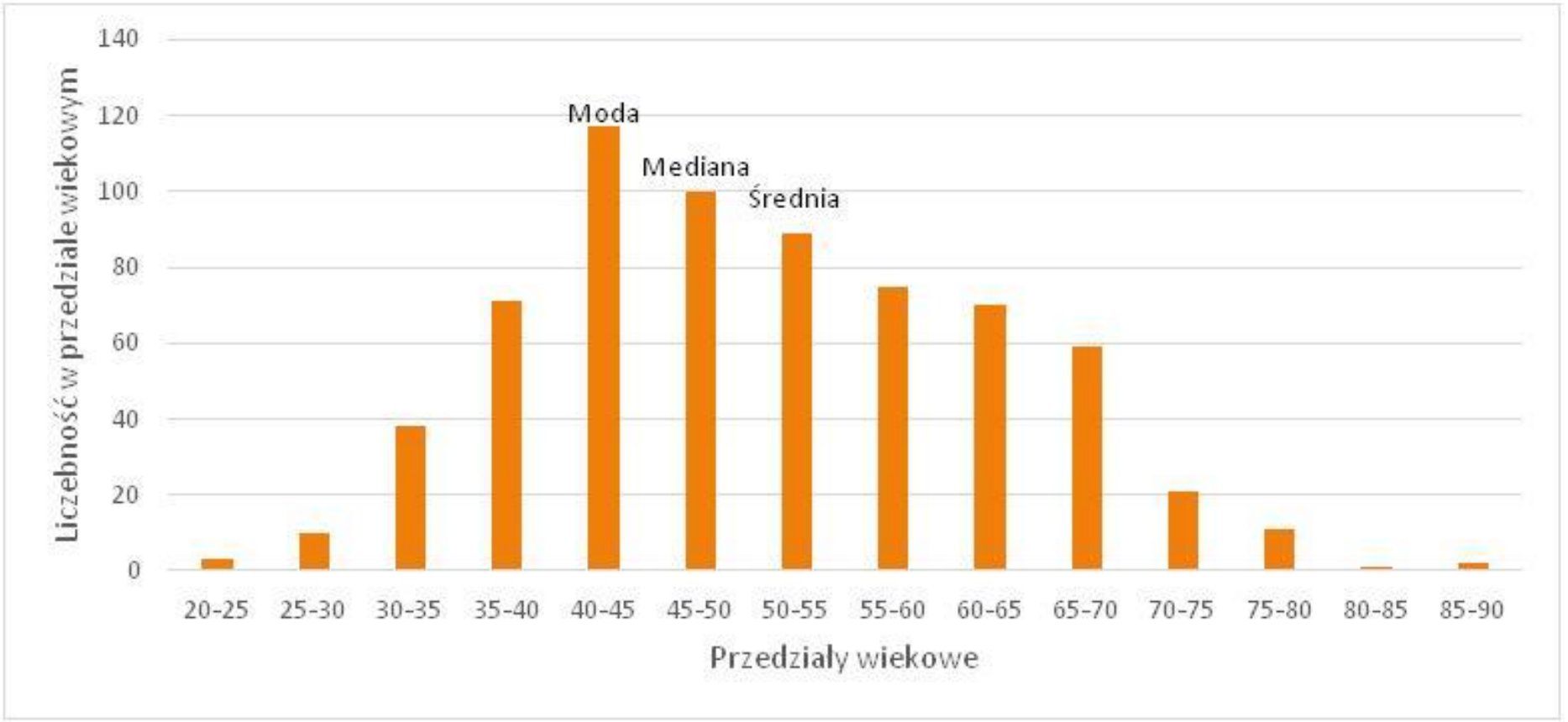


Średnia, mediana i moda rozkładu

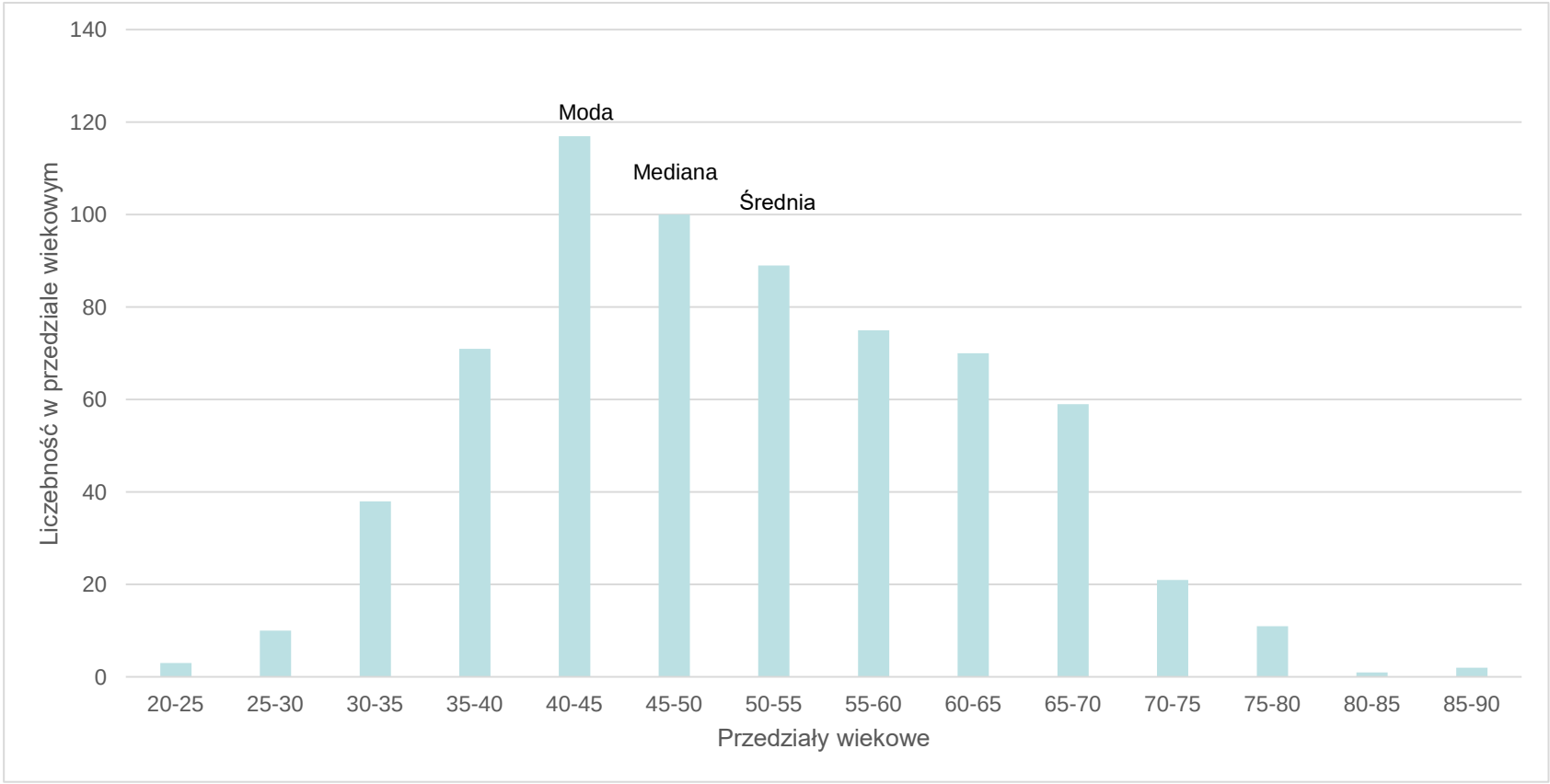
Wiek pacjentek z nowotworem szyjki macicy w pewnym szpitalu w Algierii

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



Średnia, mediana i moda rozkładu



Statystyka opisowa – miary tendencji centralnej

(4) Średnia geometryczna

$$\overline{x_g} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Jeśli dane zostaną poddane przekształceniu logarytmicznemu, po czym zostanie obliczona średnia arytmetyczna danych przekształconych, to ta średnia arytmetyczna po antylogarytmowaniu okaże się średnią geometryczną danych pierwotnych.

$$\overline{x_g} \leq \overline{x}$$

Statystyka opisowa – miary tendencji centralnej

(5) Średnia harmoniczna

$$\overline{x_h} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Jeśli dane będą przekształcone przy zastosowaniu przekształcenia odwrotnościowego, po czym zostanie obliczona średnia arytmetyczna danych przekształconych, to ta średnia arytmetyczna po przekształceniu odwrotnym okaże się średnią harmoniczną danych pierwotnych.

$$\overline{x_h} \leq \overline{x}$$

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:

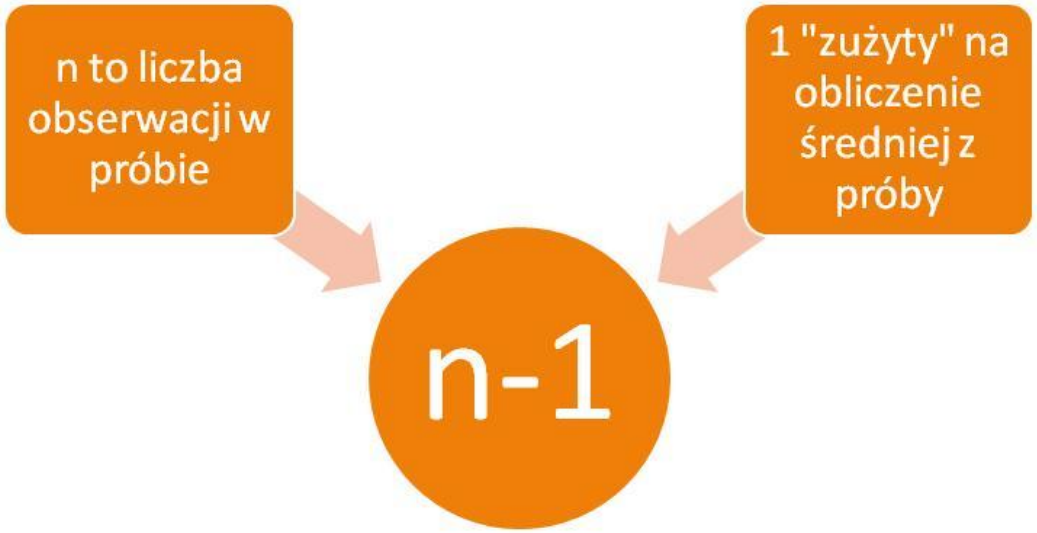
$$s^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n-1}$$

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

Praktyczne przekształcenia:

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

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:

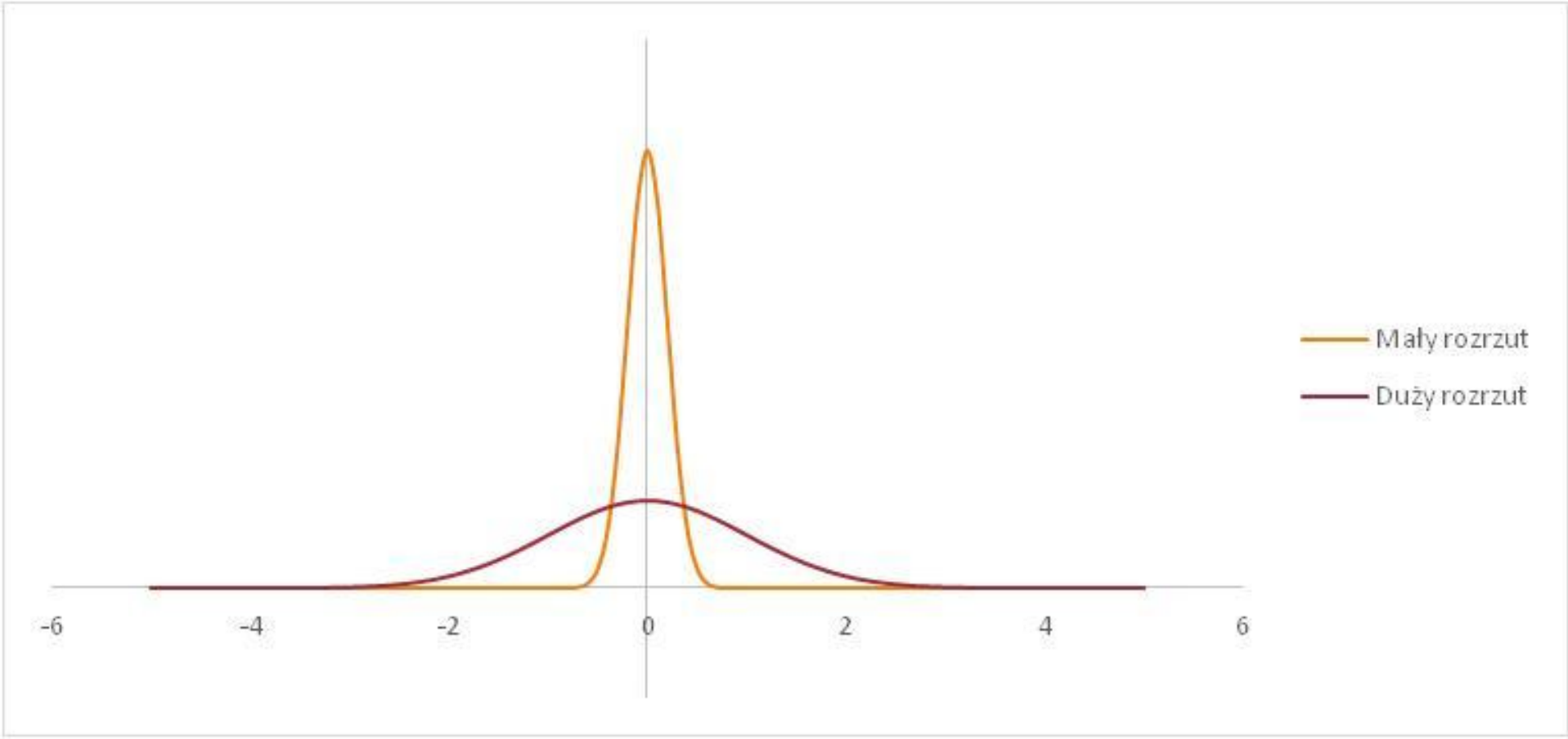
$$v = \frac{s}{\bar{x}} \cdot 100\%$$

(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.

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.

Rozkład Poissona

Z rozkładem Poissona mamy do czynienia, gdy liczymy zdarzenia losowe lub obiekty rozmieszczone losowo w przestrzeni lub czasie. Najczęściej są to zliczenia mikroorganizmów w próbkach objętościowych lub powierzchniowych Prawdopodobieństwo zajścia *r* takich zdarzeń (prawdopodobieństwo zliczenia dającego wartość *r)* wyraża się wzorem

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

Przy czym

$$E(r) = \sigma^2(r) = \lambda$$

Niezachowanie powyższej równości świadczy o niespełnieniu założenia o losowym rozmieszczaniu obiektów i sugeruje ich tendencje do grupowania się i tworzenia skupisk.

Rozkład Poissona

Przykład: Metodą kolejnego zliczania próbowano ocenić gęstość zawiesiny bakterii. Zawiesinę najpierw rozcieńczono w proporcji 1:10000, potem pobrano próbki rozcieńczonej zawiesiny o objętości 0,5cm^3 każda i rozprowadzono w aseptycznych warunkach na płytkach z pożywką agarową. Po inkubacji z każdego drobnoustroju znajdującego się na pożywce rozwinie się kolonia. Kolonie te policzono na każdej z 250 płytek. Wyniki podane są w tabelce częstości na następnym slajdzie.

Rozkład Poissona

x_r	**Liczba kolonii na płytkę**	**0**	**1**	**2**	**3**	**4**	**5**
n_r	Liczba płytek	43	75	56	34	17	12
E_r	Oczekiwana liczba płytek	33,83	67,67	67,67	45,11	22,56	9,02

x_r	**Liczba kolonii na płytkę**	**6**	**7**	**8**	**9**	**10**
n_r	Liczba płytek	9	3	1	0	0
E_r	Oczekiwana liczba płytek	3,01	0,89	0,21	0,05	0,01

Liczba płytek $N = 250$

Liczba kolonii łącznie = 500

Średnia liczba kolonii na płytce $\lambda = \frac{500}{250} = 2$

Rozkład Poissona

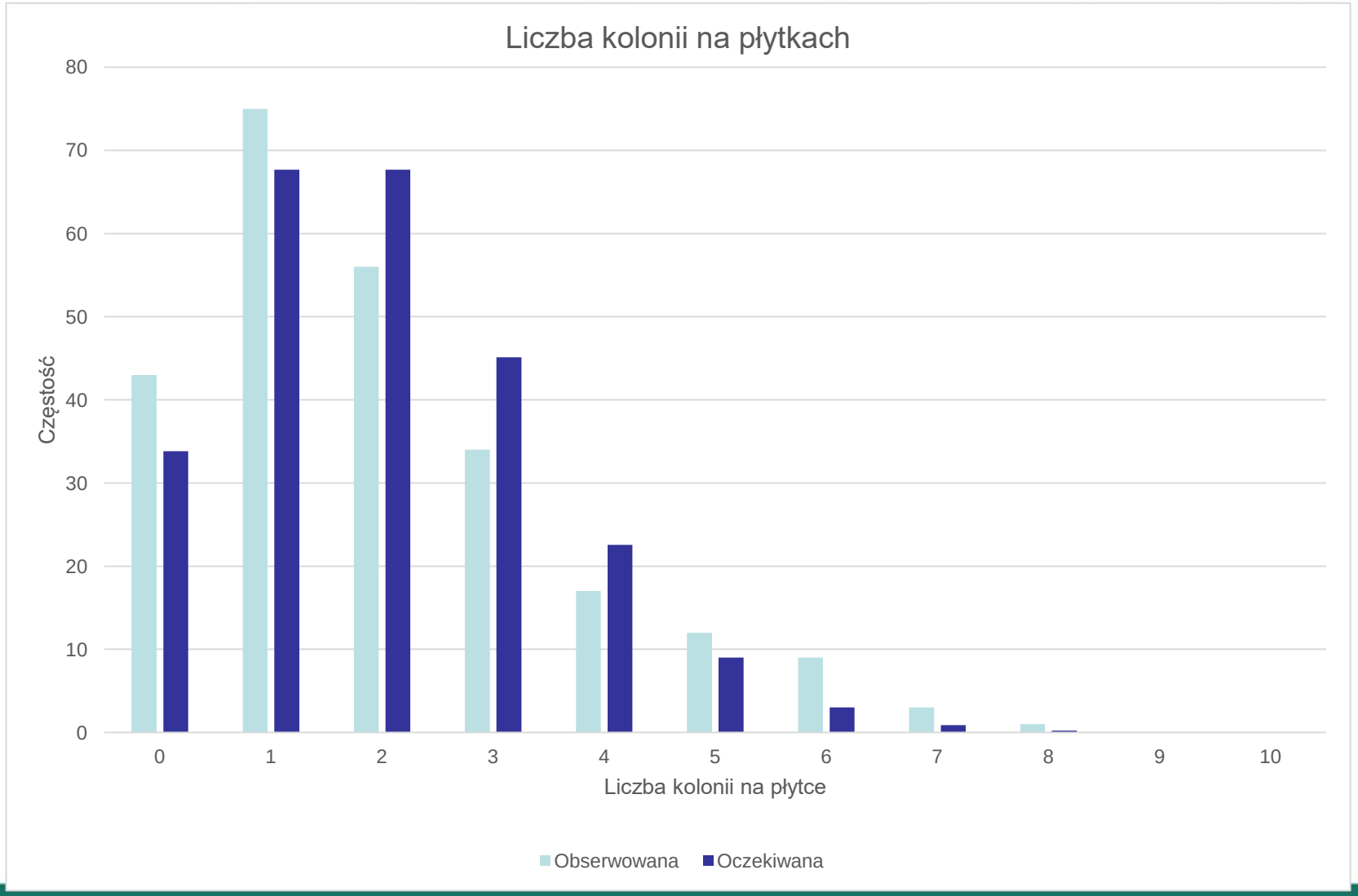
x_r	**Liczba kolonii na płytkę**	**0**	**1**	**2**	**3**	**4**	**5**
n_r	Liczba płytek	43	75	56	34	17	12
E_r	Oczekiwana liczba płytek	33,83	67,67	67,67	45,11	22,56	9,02

x_r	**Liczba kolonii na płytkę**	**6**	**7**	**8**	**9**	**10**
n_r	Liczba płytek	9	3	1	0	0
E_r	Oczekiwana liczba płytek	3,01	0,89	0,21	0,05	0,01

Liczebności oczekiwane liczono według wzoru (zakładającego rozkład Poissona):

$$E_r = N\frac{\lambda^r e^{-\lambda}}{r!}$$

Rozkład Poissona



Rozkład Poissona

Liczebności oczekiwane liczono według wzoru (zakładającego rozkład Poissona):

$$E_r = N\frac{\lambda^r e^{-\lambda}}{r!}$$

Obliczone ze („zwykłego”) wzoru oszacowanie wariancji wynosi:

$$s^2 = \frac{\sum_{r=0}^{10}(\dot{x_r} - \lambda)^2 n_r}{\sum_{r=0}^{10} n_r - 1} = 2{,}86$$

Oszacowanie wariancji $s^2=2{,}86$ znacznie odbiega od wartości teoretycznej $s^2 = \lambda = 2{,}0$. Prawdopodobnie badane drobnoustroje nie są rozmieszczone w zawiesinie losowo i dzięki np. wytworzonej wydzielinie mają tendencję do łączenia się w grupy i zlepiania się.