

# Elementy statystyki matematycznej

## Literatura:

1. Kordecki Wojciech Rachunek prawdopodobieństwa i statystyka matematyczna. Definicje, twierdzenia, wzory. GIS 2002
2. Kordecki Wojciech Rachunek prawdopodobieństwa i statystyka matematyczna. Przykłady i zadania GIS 2002
3. Fisz M. Rachunek prawdopodobieństwa i statystyka matematyczna. PWN, 1969
4. Klonecki W. Elementy statystyki dla inżynierów. PWr, 1996.
5. Clegg F. Po prostu statystyka. WsiP, 1994.

## Podstawowe pojęcia statystyki matematycznej

Statystyka matematyczna pozwala nam na podstawie informacji uzyskanej dla przebadanego zespołu elementów pochodzących z pewnego zbioru, uzyskać umotywowane informacje o wszystkich elementach należących do tego zbioru.

We **wnioskowaniu statystycznym** nie mają znaczenia same elementy badanego zbioru, a tylko pewne cechy tych elementów.

Przykładowo: w przykładzie drugim, przy badaniach antropometrycznych nie interesują nas osoby jako takie, a jedynie niektóre ich wymiary.

**Statystyka** bada prawidłowości występujące w masowych zjawiskach i opisuje te prawidłowości za pomocą liczb.

**Zjawiska masowe** – zjawiska występujące nieskończoną ilość razy.

**Badanie statystyczne** – proces pozyskiwania danych statystycznych.

Podstawowym pojęciem w statystyce matematycznej jest pojęcie **populacji generalnej**.

**Populacją generalną (statystyczną)** – nazywamy zbiór elementów podlegających badaniu statystycznemu lub szacowaniu (estymowaniu).

Populacja generalna może składać się z **jednoznacznie określonych sztuk** np. ludzie, partia śrub określonego rozmiaru, partia zegarków, bądź też podział populacji na elementy może być **sztuczny** np. węgiel w workach, ołówki w pudełkach po 10 szt., kredki w kompletach, partia zapalek, w której sztuką (elementem) będzie paczka składająca się z 10 pudełek.

Zatem **przed badaniem statystycznym trzeba określić co uważamy za populację generalną oraz w jaki sposób dzielimy ją na elementy**.

Populacje generalne mogą zawierać **skończoną lub nieskończoną liczbę elementów**.

**Zbiór statystyczny** – zbiór elementów rozpatrywanych z punktu widzenia pewnej wspólnej cechy – **populacje generalne**.

**Jednostka statystyczna** – element zbioru statystycznego poddawany bezpośredniej obserwacji lub pomiarowi (obiekt badania).

**Własności zbioru statystycznego** – **cechy statystyczne, argumenty**.

Aby zbadać własności populacji generalnej, należałoby przebadać każdy element populacji. W praktyce jednak nie ma bezbłędnych metod badania. Występują błędy pomiaru, zmęczenie badającego itp. Również badanie wszystkich elementów populacji może być kosztowne, a w przypadku badań niszczących – niemożliwe. Widać więc, że trzeba wprowadzić badanie tylko części populacji zamiast całej

populacji, uwzględniać błędy pomiaru i zdawać sobie sprawę z błędów popełnianych przy ocenie całej populacji. To postępowanie jest właśnie przedmiotem **statystyki matematycznej**.

**Próbka statystyczna – część populacji generalnej przeznaczona do badań, zespół elementów wylosowanych z populacji zgodnie z rozkładem równomiernym.**

Inne określenie:

**Próba statystyczna** – skończony zbiór (elementów) doświadczeń wykonanych w celu określenia kształtu lub parametrów poszukiwanego rozkładu.

O badaniu populacji na podstawie tak pobranej próbki mówimy, że jest **wyrywkowe**.

**Próbka zwrotna** – każdy element populacji po pobraniu do próbki jest po zbadaniu zwracany do populacji.

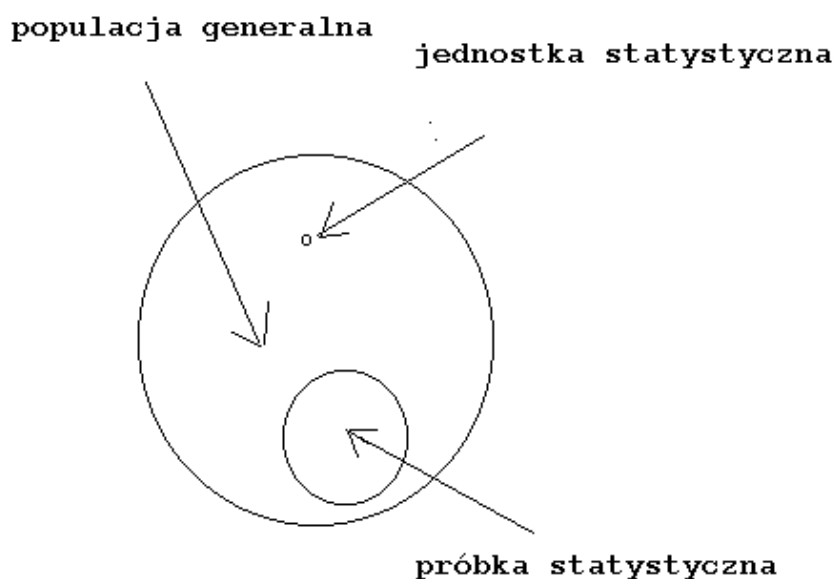
**Próbka bezzwrotna** - każdy element populacji po pobraniu do próbki nie jest po zbadaniu zwracany do populacji.

Próbki mogą być:

- pojedynczymi elementami,
- próbkami bezkształtnymi – np. próbka węgla, mąki, średnie próbki laboratoryjne itp.

W dalszych rozważaniach będziemy zakładać, że próbki będą się składać z elementów równouprawnionych, w których rozkład cechy statystycznej każdego elementu jest taki sam jak rozkład cechy w populacji. Przykładem takiej próbki jest próbka zwrotna pobrana ze skończonej populacji o równouprawnionych elementach.

Na rysunku przedstawiono te podstawowe pojęcia statystyki matematycznej.



**Cechy statystyczne mogą być:**

- **mieralne (liczbowe)** - np. pomiar wagi, wzrostu,
- **niemieralne (jakościowe)** – np. zawód, płeć, sprawdzanie czy cegła jest pokruszona.

Często elementy populacji mogą być badane na kilka cech zarówno mierzalnych jak i niemierzalnych. Cechę niemierzalną (jakościową) – można sprowadzić w sposób sztuczny do cechy mierzalnej (liczbowej). Przykładowo : osobnikowi badanemu na płeć można przyporządkować liczbę:

0 – osobnik męski,

1 – osobnik żeński,

i odwrotnie – cechę liczbową elementu populacji można sprowadzić w sposób sztuczny do cechy jakościowej. Przykładowo;

pręt metalowy w badanej partii na wytrzymałość:

dobry – jeśli jego wytrzymałość  $> A$ ,  
zły – jeśli jego wytrzymałość  $\leq A$ .

W związku z tym, w dalszych rozważaniach zajmiemy się tylko **cechami mierzalnymi (liczbowymi)**.

**Próba n – elementowa** to próba składająca się z  $n$  elementów. Na przykład poniższy zapis przedstawia  $l$  z  $n$  – elementowych prób:

1 próba:  $x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}$

2 próba:  $x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)}$

.

.

$j$  próba:  $x_1^{(j)}, x_2^{(j)}, \dots, x_n^{(j)}$

.

.

$l$  próba:  $x_1^{(l)}, x_2^{(l)}, \dots, x_n^{(l)}$

co w postaci wektorowej możemy zapisać jako:

$$\bar{x}^{(j)} = \{x_1^{(j)}, x_2^{(j)}, \dots, x_n^{(j)}\}.$$

Ponieważ różne wartości tego wektora przyjmowane będą z różnym prawdopodobieństwem, możemy więc powiedzieć, że wektor ten posiada pewną gęstość prawdopodobieństwa:

$$g(\bar{x}) = g(x_1, x_2, \dots, x_n).$$

Próbę nazywamy próbą losową, gdy spełnione są poniższe 2 warunki:

1)  $g(\bar{x}) = g_1(x_1) * g_2(x_2) * \dots * g_n(x_n)$

2)  $g_1(x_1) = g_2(x_2) = \dots = g_n(x_n) = f(x).$

Rozkładem w próbie nazywamy funkcję:

$$W_n(x) = \frac{n_x}{n}$$

gdzie:  $n$  – liczność (wymiar) próby,

$n_x$  – liczba elementów próby, takich, że:  $X < x$ .

Dla tak zdefiniowanego rozkładu zachodzi:

$$W_n(x) \xrightarrow{n \rightarrow \infty} F(x)$$

Zmienną losową będącą funkcją zmiennej  $\bar{x}$  nazywamy **statystyką**.

**Statystyka Z** - zmienna losowa będąca dowolną funkcją wyników próby losowej, tzn. dowolną funkcją:

$$Z = f(\bar{x})$$

Poszukiwanie parametrów rozkładu na podstawie badania rozkładu próby nazywamy **estymacją parametrów**. Poszukiwana wartość parametru jest statystyką zwaną też **estymatorem**.

Podstawowym zagadnieniem pojawiającym się w badaniu statystycznym jest możliwość uogólniania uzyskanych na podstawie próby wyników, na całą populację oraz oszacowanie popełnianych przy tym błędów.

Takie działania nazywa się **wnioskowaniem statystycznym**.

Wyróżnia się dwa podstawowe typy problemów:

- **estymacja (szacowanie)** nieznanych wartości parametrów rozkładu cechy,
- **sprawdzanie (weryfikacja)** hipotez dotyczących wartości parametrów rozkładu lub postaci samego rozkładu.

## Zastosowania statystyki

Opis i parametry zjawisk o charakterze losowym.

- *średni wzrost*,

- *średnia płaca,*
- *rozpiętość temperatur*

Związki i korelacje pomiędzy kilkoma zjawiskami losowymi.

- *czy istnieje związek pomiędzy wzrostem a wynagrodzeniem?*
- *jak zależy jasność gwiazdy od jej średnicy?*

Szacowanie parametrów populacji na podstawie losowo wybranej próbki.

- *procent poparcia dla partii,*
- *średni wzrost Polaków,*
- *średnia temperatura w lipcu,*
- *czas połowicznego rozpadu*

Testowanie hipotez statystycznych.

- *średni wzrost Polaków i Rosjan jest taki sam,*
- *metoda leczenia A daje mniej powikłań niż metoda B,*
- *proton jest cząstką nietrwałą*

Prognozowanie.

- *średnia długość życia w 2010 roku powinna wzrosnąć do 67 lat,*
- *zmniejszenie zużycia węgla o  $X$  powinno spowodować wzrost zużycia ropy o  $Y$*

Organizacja badań statystycznych

**W badaniu statystycznym wyróżniamy następujące etapy:**

- **przygotowanie (programowanie) badania,**
- **obserwację statystyczną,**
- **opracowanie i prezentację materiału statystycznego,**
- **opis lub wnioskowanie statystyczne.**

**Etap pierwszy** - obejmuje czynności przygotowawcze:

- ustalenie celów i metody badania,
- określenie zbiorowości statystycznej i cech podlegających badaniu,
- zdefiniowanie jednostki statystycznej.

**Cele:**

- poznanie rozkładu zbiorowości pod względem wybranych cech,
- ustalenie, jakiego rodzaju związki występują między cechami, ocena współzależności cech,
- porównanie i porządkowanie elementów,
- określenie przedmiotu badania, czyli zbiorowości statystycznej pod względem rzeczowym, czasowym i przestrzennym,
- określenie jednostki statystycznej,
- wybór metody badania - pełne czy częściowe.

**Etap drugi** - polega na prowadzeniu obserwacji w celu ustalenia wartości ilościowych lub odmian cech jakościowych u wszystkich jednostek tworzących zbiorowość statystyczną.

**Obserwacja może odbywać się za pomocą:**

- **pomiaru bezpośredniego,**
- **zbierania informacji od jednostek sprawozdawczych,**

**Materiał statystyczny** - zbiór danych uzyskanych w wyniku obserwacji:

- **materiał pierwotny** - są to dane gromadzone specjalnie do celów badania statystycznego,
- **materiał wtórny** - są to dane gromadzone z innych powodów.

**Surowy materiał statystyczny** - materiały statystyczne zebrane w pierwotnej postaci, obciążone pewnymi błędami.

**Ze względu na przyczynę powodującą błędy dzielimy je na:**

- **błędy systematyczne** - wynikają z jednokierunkowej tendencji do zniekształcania badanej rzeczywistości (np. dane wielkości są zaniżone lub zawyżone),
- **błędy przypadkowe** - popełniane nieumyślnie, wynikające z nieumiejętności podania prawidłowych informacji lub niedbalstwa bądź też są wynikiem błędów pomiarowych.

**Wykrywanie błędów w surowym materiale statystycznym:**

- **kontrola formalna (ilościowa)** - sprawdza kompletność, pełność i zupełność materiału statystycznego,
- **kontrola merytoryczna** - obejmuje kontrolę:
  - *logiczną* - sprawdza, czy treść odpowiada rzeczywistości
  - *arytmetyczną* - sprawdza zgodność wartości liczbowych

**Etap trzeci** - polega na opracowaniu materiału statystycznego, obejmuje dwie zasadnicze czynności:

- **grupowanie** - polega na wyodrębnieniu jednorodnych lub względnie jednorodnych części w ramach większej i zróżnicowanej zbiorowości statystycznej,
- **zliczanie** - czynność ściśle związana z grupowaniem (ręczne, elektroniczne).

**Etap czwarty** - polega na:

- **opisie statystycznym** - dotyczy tylko danej zbiorowości generalnej,
- **wnioskowaniu statystycznym** - kiedy badanie jest reprezentacyjne (próba losowa) i jego wyniki są uogólniane na całą populację generalną.

Uogólnianie wyników z próby losowej na całą populację jest możliwe dzięki zastosowaniu rachunku prawdopodobieństwa, który jest teoretyczną podstawą wnioskowania statystycznego.

Metody używane w opisie statystycznym wchodzą w zakres **statystyki opisowej**, natomiast metody wnioskowania statystycznego - **statystyki matematycznej**.

## Dystrybucja teoretyczna

Jeśli z populacji generalnej wybierzemy 1 element na chybił trafił, to cecha liczbową takiego elementu jest zmienną losową o pewnym rozkładzie – rozkład danej cechy liczbowej w populacji. Niech będzie dana populacja generalna zawierająca  $N$  elementów. Elementy tej populacji są badane na 1 cechę liczbową.

Niech  $N$  – liczność populacji generalnej

Niech cechy liczbowe elementów tej populacji:

$$x_1, x_2, \dots, x_N$$

Po uporządkowaniu ich w porządku rosnącym, otrzymamy:

$$x_1^*, x_2^*, \dots, x_N^*$$

**Dystrybuanta teoretyczna w populacji generalnej:**

$$F(x) = \frac{1}{N} \sum_{k=1}^N \eta(x - x_k^*) \quad \text{gdzie} \quad \eta(x - a) = \begin{cases} 0 & \text{dla } x \leq a \\ 1 & \text{dla } x > a \end{cases}$$

**Funkcja charakterystyczna w populacji generalnej:**

$$\varphi(t) = \sum e^{ix_k^* t} \cdot \frac{1}{N} = \frac{1}{N} \sum_{k=1}^N e^{ix_k^* t}$$

**Moment rzędu r w populacji generalnej:**

$$m_r = \frac{1}{N} \sum_{k=1}^N x_k^r$$

**Wartość średnia i wariancja w populacji generalnej:**

$$\mu = m = \frac{1}{N} \sum_{k=1}^N x_k \quad \sigma^2 = \frac{1}{N} \sum_{k=1}^N (x_k - m)^2$$

**Mediana (wartość środkowa) w populacji generalnej:**

Jeśli N – nieparzyste, czyli  $N = 2k-1$  to:  $m_e = x_k^*$

Jeśli N – parzyste, czyli  $N = 2k$  to  $m_e$  dowolna liczba spełniająca nierówność:

$$x_k^* \leq m_e \leq x_{k+1}^*$$

przy czym zazwyczaj przyjmuje się, że:

$$m_e = \frac{x_k^* + x_{k+1}^*}{2}$$

Inaczej mówiąc, aby obliczyć medianę ze zbioru N obserwacji, sortujemy je w kolejności od najmniejszej do największej i numerujemy od 1 do N. Następnie, jeśli N jest nieparzyste, medianą jest wartość obserwacji w środku (czyli obserwacji numer (N+1)/2). Jeśli natomiast N jest parzyste, wynikiem jest średnia arytmetyczna między dwiema środkowymi obserwacjami, czyli obserwacją numer N/2 i obserwacją numer N/2+1.

**Mediana** jest miarą pozycyjną, która rozdziela całą populację na dwie liczebnie równe części w ten sposób, że w jednej z nich znajdują się jednostki o wartościach niższych lub równych od mediany, a w drugiej o wartościach wyższych lub równych od mediany.

**Kwantyl rzędu p w populacji generalnej:**

**Kwantylem rzędu p**, gdzie  $0 \leq p \leq 1$ , w rozkładzie teoretycznym zmiennej losowej  $x^*$  nazywamy najmniejszą wartość  $x_p^*$ , dla której dystrybuanta teoretyczna  $F$  spełnia nierówność  $F(x_p^*) \leq p$ .

Innymi słowy kwantylem rzędu p jest taka wartość  $x_p^*$  zmiennej losowej, że wartości mniejsze lub równe od  $x_p^*$  są przyjmowane z prawdopodobieństwem co najmniej p.

Kwantyl rzędu 1/2 to inaczej mediana.

Wśród kwantyli wyróżniamy: **kwantyl pierwszy** (dolny), **drugi** (mediana lub wartość środkowa) oraz **trzeci** (górny). Każdy z kwantyli dzieli zbiorowość na dwie części pod względem liczebności.

1. **kwantyl pierwszy  $Q_1$** – (rzędu 1/4) dzieli zbiorowość uporządkowaną na dwie części w ten sposób, że 25% jednostek na wartości cechy niższe i 75% wyższe od kwantyla pierwszego;
2. **kwantyl drugi  $Q_2$**  mediana– (rzędu 1/2) dzieli zbiorowość uporządkowaną na dwie części w ten sposób, że 50% jednostek na wartości cechy niższe i 50% wyższe od mediany;
3. **kwantyl trzeci  $Q_3$** – (rzędu 3/4) dzieli zbiorowość uporządkowaną na dwie części w ten sposób, że 75% jednostek na wartości cechy niższe i 25% wyższe od kwantyla trzeciego.

**Rozstęp w populacji generalnej** – różnica między największą i najmniejszą wartością badanej cechy liczbowej w populacji:

$$x_N^* - x_1^*$$

## Dystrybucja empiryczna

**Liczność próbki** – liczba elementów w próbie –  $n$ .

Cechy elementów próbki są niezależnymi zmiennymi losowymi:  $X_1, X_2, \dots, X_n$

O jednakowym rozkładzie, identycznym z rozkładem cechy w populacji generalnej.

Niech wartości cech elementów w próbie uporządkowane w kolejności niemalejącej:

$$X_1^*, X_2^*, \dots, X_n^*$$

**Dystrybucja empiryczna z próbki:**

$$F^*(x) = \frac{1}{n} \sum_{k=1}^n \eta(x - X_k^*) \quad \text{gdzie} \quad \eta(x - a) = \begin{cases} 0 & \text{dla } x \leq a \\ 1 & \text{dla } x > a \end{cases}$$

**Funkcja charakterystyczna z próbki:**

$$\varphi^*(t) = \frac{1}{n} \sum_{k=1}^n e^{iX_k^* t}$$

**Moment rzędu  $r$  z próbki:**

$$M_r = \frac{1}{n} \sum_{k=1}^n X_k^{*r}$$

**Wartość średnia i wariancja z próbki:**

$$\bar{X} = M_1 = \frac{1}{n} \sum_{k=1}^n X_k^* \quad S^2 = M_2 - M_1^2 = \frac{1}{n} \sum_{k=1}^n (X_k^* - \bar{X})^2$$

**Wariancja empiryczna poprawiona (statystyka „S kwadrat z daszkiem”):**

$$\hat{S}^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k^* - M_1)^2$$

**Mediana z próbki:**

Jeśli  $N$  – nieparzyste, czyli  $N = 2k-1$  to:  $M_e = X_k^*$

Jeśli  $N$  – parzyste, czyli  $N = 2k$  to:

$$M_e = \frac{X_k^* + X_{k+1}^*}{2}$$

Rozstęp w próbie:

$$X_n^* - X_1^*$$

## Rozkłady prawdopodobieństwa występujące w statystyce

### Rozkład normalny (rozkład Gaussa - Laplace'a)

Jest rozkładem, któremu podlega wiele zjawisk świata fizycznego, np. waga oraz wzrost populacji ludzi, szumy, zakłócenia szerokopasmowe itp.

Jego znaczenie metodologiczne i analityczne wynika z trzech jego najważniejszych właściwości:

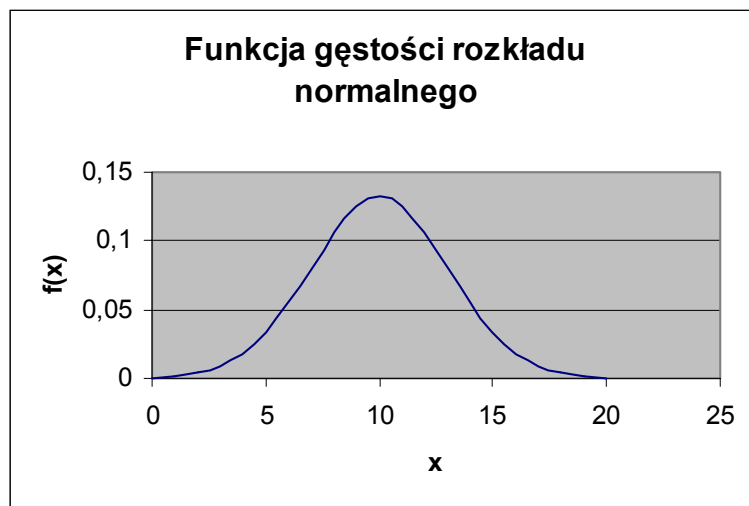
- Przy nieograniczonym wzroście liczby niezależnych doświadczeń statystycznych, wszystkie znane teoretyczne rozkłady zmiennych losowych ciągłych i skokowych są szybko zbieżne do rozkładu normalnego. Stanowi on zatem najbardziej ogólne odniesienie do rozumienia sensu działania **prawa wielkich liczb**,
- W statystycznym wnioskowaniu o parametrach i rozkładach w populacjach generalnych na podstawie wyników badań prób losowych popełniane są błędy przypadkowe, których rozkład jest normalny lub granicznie normalny. Zawiera się w tym merytoryczny sens statystycznej indukcji, czyli wnioskowania. Na podstawie tej prawidłowości, skonstruowane zostały wszystkie metody estymacji parametrów oraz metody weryfikacji hipotez,
- W niektórych sytuacjach badawczych, rozkłady empiryczne obserwowanych zmiennych mogą być zbliżone swoim kształtem do rozkładu normalnego. Wtedy też prawidłowości statystyczne ujawniają się w swojej najczystszej postaci, ale może mieć to miejsce tylko wtedy, kiedy badane zjawisko podlega wpływowi bardzo wielu czynników, działających mniej więcej równomiernie.

**Zmienna losowa  $X$**  ma rozkład normalny z wartością oczekiwaną równą  **$m$**  (czasami oznaczaną jako  $\mu$ ) i odchyleniem standardowym równym  **$\sigma$**   $X : N(m; \sigma)$ , jeśli jej **funkcja gęstości** ma następującą postać:

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

Wykres funkcji gęstości rozkładu normalnego określany jest jako **krzywa normalna**, która przyjmuje następującą postać:

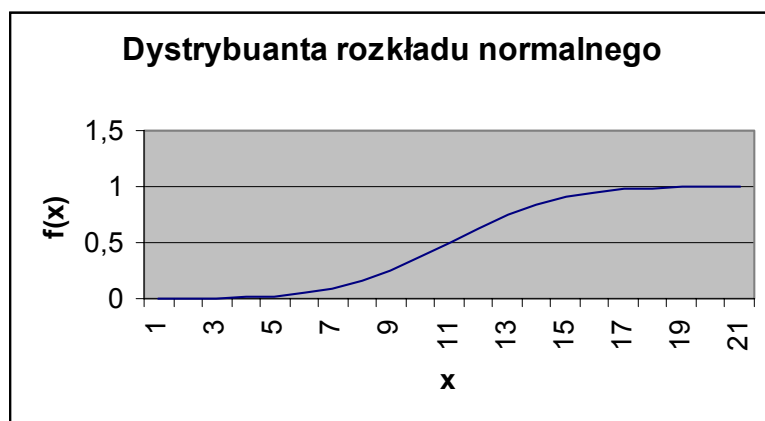




Dystrybuanta rozkładu normalnego ma postać:

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

Wykres dystrybuanty zmiennej losowej  $X : N(m; \sigma)$  przyjmuje następującą postać:



Wartość oczekiwana i wariancja dla rozkładu normalnego wyrażane są następującymi wzorami:

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

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

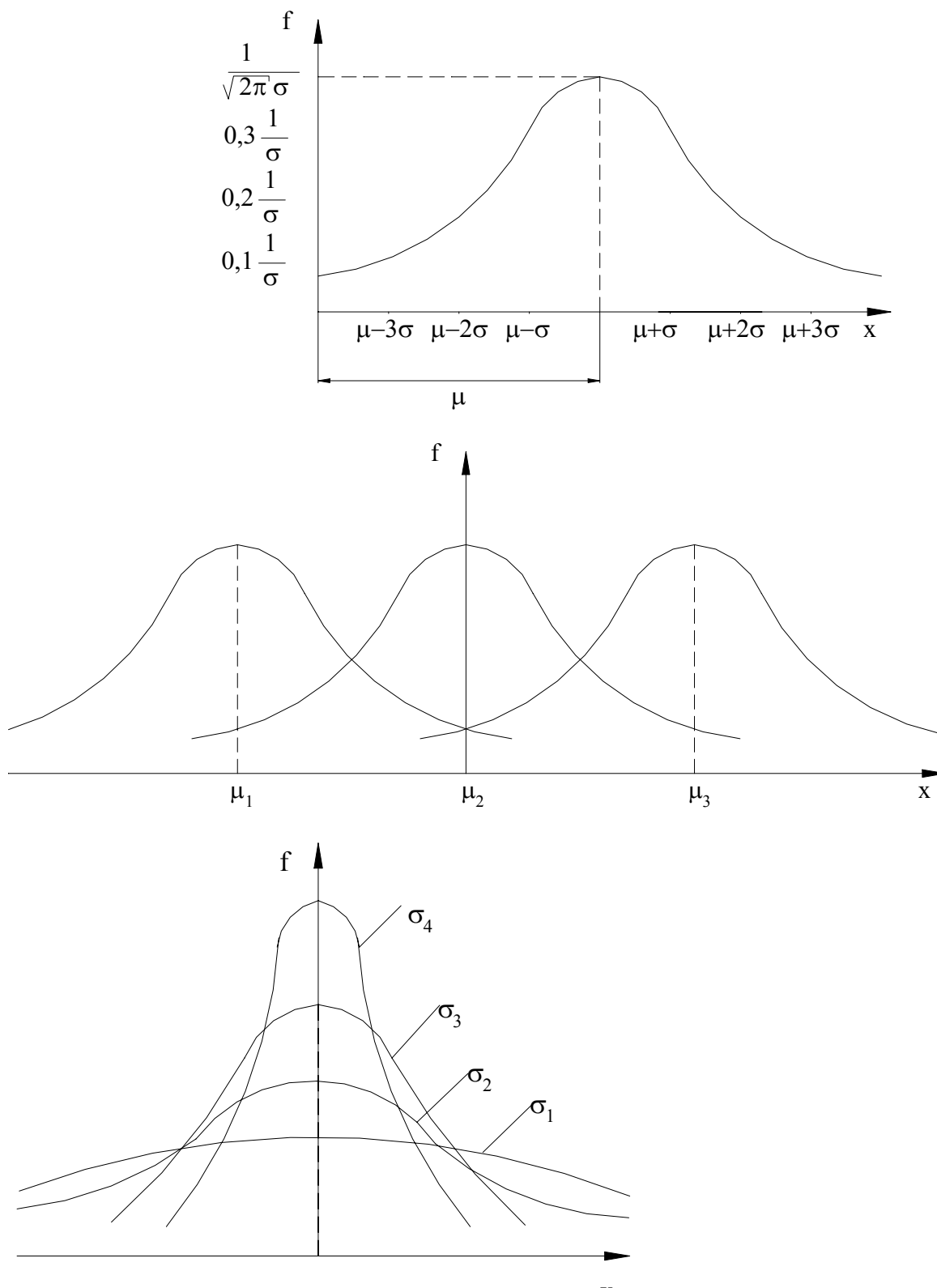
Gdzie:  $m$  - wartość średnia zmiennej losowej  $X$  o rozkładzie normalnym,

$\sigma$  - odchylenie standardowe.

Krzywa gęstości rozkładu normalnego ma następujące własności:

- jest symetryczna względem prostej  $x = m$ ,
- osiąga maksimum równe  $\frac{1}{\sigma\sqrt{2\pi}}$  dla  $x = m$ ,
- jej ramiona mają punkty przegięcia dla  $x = m - \sigma$  oraz  $x = m + \sigma$ .

Wartość parametru  $m$  decyduje o położeniu krzywej normalnej względem osi  $x$ . Im średnia przyjmuje większe wartości, tym krzywa jest bardziej przesunięta w prawo. Wartość parametru  $\sigma$  determinuje natomiast „smukłość” krzywej. Im odchylenie standardowe jest większe, tym krzywa jest bardziej spłaszczona.



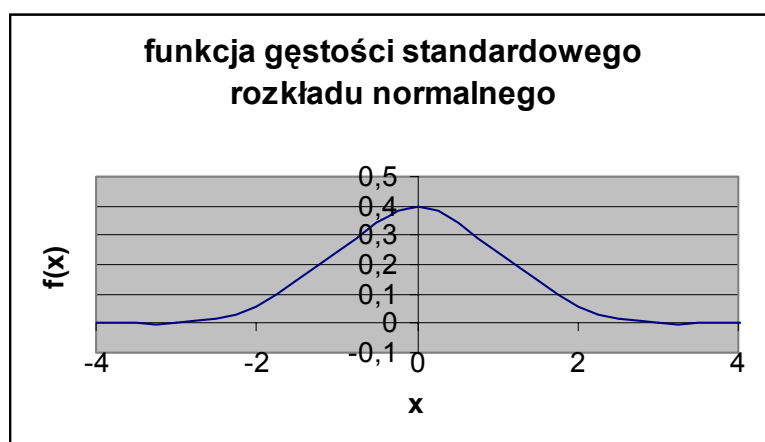
Gdzie:  $\sigma_1 > \sigma_2 > \sigma_3 > \sigma_4$

Możliwość sprowadzenia dowolnego rozkładu normalnego do postaci **standardowego rozkładu normalnego**, którego funkcja gęstości i dystrybuenta zostały umieszczone w tablicy.

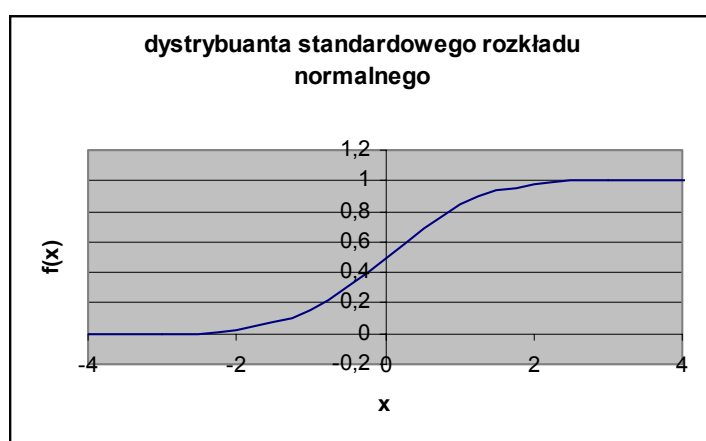
**Standardowym rozkładem normalnym** nazywamy rozkład normalny ze średnią równą 0 oraz odchyleniem standardowym równym 1 i oznaczamy  $N(0;1)$ .

Zmienną losową, która ma standardowy rozkład normalny oznacza się literą  $U$ , jej funkcję gęstości  $\varphi(u)$ , natomiast dystrybuantę  $\Phi(u)$ .

Wykres funkcji gęstości standardowego rozkładu normalnego przyjmuje następującą postać:



Wykres dystrybuanty standardowego rozkładu normalnego przyjmuje postać:



Dystrybuanta standardowego rozkładu normalnego charakteryzuje się następującymi własnościami:

$$P(U \leq u) = \Phi(u)$$

$$P(U \leq -u) = \Phi(-u) = 1 - \Phi(u)$$

$$P(U > u) = 1 - P(U \leq u) = 1 - \Phi(u)$$

$$P(U > -u) = 1 - P(U \leq -u) = 1 - \Phi(-u) = 1 - (1 - \Phi(u)) = 1 - 1 + \Phi(u) = \Phi(u)$$

Ze względu na fakt, że w tablicach najczęściej podawane są wartości tylko dla dodatnich  $u$ , przy wyznaczaniu wartości dla ujemnych  $u$  należy skorzystać z następujących własności funkcji  $\varphi(u)$  i  $\Phi(u)$ :

$$\varphi(u) = \varphi(-u)$$

$$\Phi(u) = 1 - \Phi(-u)$$

W celu obliczenia prawdopodobieństwa  $P(a < X \leq b)$  należy skorzystać z operacji nazywanej **standaryzacją**. Jeśli zmienna losowa  $X$  ma rozkład  $N(m, \sigma)$  to zmienna standaryzowana  $U = \frac{X - m}{\sigma}$  ma rozkład  $N(0; 1)$ . Na tej podstawie można wyznaczyć:

$$P(a < X \leq b) = P\left(\frac{a - m}{\sigma} < \frac{X - m}{\sigma} \leq \frac{b - m}{\sigma}\right) = P\left(\frac{a - m}{\sigma} < U \leq \frac{b - m}{\sigma}\right) = \Phi\left(\frac{b - m}{\sigma}\right) - \Phi\left(\frac{a - m}{\sigma}\right)$$

Wartości  $\Phi\left(\frac{b - m}{\sigma}\right)$  i  $\Phi\left(\frac{a - m}{\sigma}\right)$  odczytuje się z tablic dystrybucyj standardowego rozkładu normalnego.

Z rozkładem normalnym związana jest tzw. **reguła trzech sigm**, zgodnie z którą praktycznie wszystkie obserwacje dokonywane na zmiennej losowej o rozkładzie normalnym mieszczą się w przedziale  $(m - 3\sigma, m + 3\sigma)$ . Reguła trzech sigm jest wykorzystywana w badaniach statystycznych do eliminacji obserwacji niewiarygodnych. Obserwacje niewiarygodne to obserwacje, których wartość różni się od średniej o więcej niż trzy odchylenia standardowe. Przyjmuje się, iż zmienne, które odbiegają tak znacznie od średniej mogą być skutkiem błędu pomiaru. Dla realizacji zmiennej losowej o dowolnym rozkładzie normalnym około 68,3% obserwacji mieści się w granicach jednego odchylenia standardowego wokół średniej, 95,5% obserwacji mieści się w granicach dwóch odchyleń standardowych i 99,7% w granicach trzech odchyleń standardowych.

Zmienne losowe o rozkładzie normalnym mają własności, którą można ująć w postaci twierdzeń:

## Twierdzenia o rozkładzie sumy niezależnych zmiennych losowych

### Twierdzenie 1

Jeżeli zmienne losowe  $X_1, X_2, \dots, X_n$  są niezależne i zmienna losowa  $X_i$  dla  $i = 1, 2, \dots, n$  ma rozkład  $N(m_i; \sigma_i)$  to zmienna losowa  $Y = X_1 + X_2 + \dots + X_n$  ma rozkład:

$$N(m_1 + m_2 + \dots + m_n; \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2})$$

Jeżeli zmienne losowe  $X_1, X_2, \dots, X_n$  są niezależne o takim samym rozkładzie  $X_i : N(m; \sigma)$  dla  $i = 1, 2, \dots, n$ , to zmienna losowa  $Y$  ma rozkład:

$$N(nm; \sigma\sqrt{n})$$

### Twierdzenie 2

Jeżeli zmienne losowe  $X_1, X_2$  są niezależne i zmienna losowa  $X_i$  dla  $i = 1, 2$  ma rozkład  $N(m_i; \sigma_i)$  to zmienna losowa  $Z = X_1 - X_2$  ma rozkład:

$$N(m_1 - m_2; \sqrt{\sigma_1^2 + \sigma_2^2})$$

Jeżeli zmienne losowe  $X_1, X_2$  są niezależne o takim samym rozkładzie  $X_i : N(m; \sigma)$  dla  $i = 1, 2$  to zmienna losowa  $Z = X_1 - X_2$  ma rozkład:

$$N(0; \sigma\sqrt{2})$$

### Twierdzenie 3

Jeżeli zmienne losowe  $X_1, X_2, \dots, X_n$  są niezależne i zmienna losowa  $X_i$  dla  $i = 1, 2, \dots, n$  ma rozkład  $N(m_i; \sigma_i)$  to zmienna losowa  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  ma rozkład:

$$N\left(\frac{1}{n}(m_1 + m_2 + \dots + m_n); \frac{1}{n}\sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}\right)$$

Jeżeli zmienne losowe  $X_1, X_2, \dots, X_n$  są niezależne o takim samym rozkładzie  $X_i : N(m; \sigma)$

dla  $i = 1, 2, \dots, n$ , to zmienna losowa  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  ma rozkład:  $N(m; \frac{\sigma}{\sqrt{n}})$

### Twierdzenia graniczne

Twierdzenia graniczne mówią o zbieżności ciągów zmiennych losowych do pewnych rozkładów, które nazywamy rozkładami granicznymi.

Rozkład Poissona jest rozkładem granicznym dla rozkładu dwumianowego, zmienną losową o rozkładzie dwumianowym, dla  $p < 0,02$  oraz  $k > 30$  można przybliżyć rozkładem Poissona.

### Twierdzenie Moivre'a - Laplace'a

Niech  $X_n$  będzie zmienną losową o rozkładzie dwumianowym  $B(n, p)$  ( $n$  - liczba doświadczeń,  $p$  - prawdopodobieństwo sukcesu) i niech  $X$  będzie zmienną losową o rozkładzie normalnym z wartością oczekiwaną  $m = np$  i odchyleniem standardowym  $\sigma = \sqrt{npq}$ , czyli  $N(np; \sqrt{npq})$ .

Oznaczmy przez  $F_n(x)$  oznacza wartość dystrybuanty zmiennej losowej  $X_n$  w punkcie  $x$  i przez  $F(x)$  wartość dystrybuanty zmiennej losowej  $X$  w punkcie  $x$ .

Między dystrybuantami zachodzi związek:

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

Korzystając z definicji dystrybuanty:

$$\lim_{n \rightarrow \infty} P(X_n < x) = P(X < x)$$

Wiadomo, że liczba doświadczeń jest zawsze skończona, stąd

$$\lim_{n \rightarrow \infty} P(X_n < x) \approx P(X < x)$$

gdzie:

$$P(X_n < x) = \sum_{k < x} C_n^k p^k q^{n-k} \quad P(X < x) = \frac{1}{\sqrt{2\pi} \sqrt{npq}} \int_{-\infty}^x e^{-\frac{(x-np)^2}{2npq}} dx$$

Oznacza to, że jeżeli liczba prób jest duża, to rozkład zmiennej losowej  $X_n$  o rozkładzie  $B(n, p)$  można przybliżyć rozkładem  $N(np; \sqrt{npq})$ , przybliżenie to jest tym lepsze, im  $n$  jest większe (praktycznie  $n > 30$ ).

### Wniosek z twierdzenia Moivre'a-Laplace'a

Rozpatrzmy zmienną losową  $Y_n = \frac{X_n}{n}$  (częstość), gdzie  $X_n$  jest zmienną losową o rozkładzie dwumianowym z parametrami  $n$  i  $p$ . Jeżeli zmienna losowa  $X_n$  przyjmuje wartości  $0, 1, 2, \dots, n$ , to:

$$Y_n : 0, \frac{1}{n}, \frac{2}{n}, \dots, 1.$$

Rozkład zmiennej losowej  $Y_n$ :

$$P(Y_n = \frac{k}{n}) = P(\frac{X_n}{n} = \frac{k}{n}) = P(X = k) \quad k = 1, 2, \dots, n.$$

Wynika stąd, że zmienna  $Y_n$  przyjmuje swoje wartości z prawdopodobieństwami określonymi przez rozkład dwumianowy.

Wartość oczekiwana i wariancja zmiennej losowej  $Y_n$ :

$$E(Y_n) = \frac{1}{n} E(X_n) = \frac{1}{n} np = p$$

$$V(Y_n) = \frac{1}{n^2} V(X_n) = \frac{1}{n^2} npq = \frac{pq}{n}$$

Zmienna losowa  $Y_n$  przy dużych wartościach  $n$  ma rozkład zbliżony do normalnego z wartością oczekiwaną równą  $p$  i odchyleniem standardowym równym  $\sqrt{\frac{pq}{n}}$ , czyli

$$Y_n : N(p; \sqrt{\frac{pq}{n}})$$

### ***Wniosek z centralnego twierdzenia granicznego Lindeberga - Levy'ego***

Najważniejsze twierdzenie statystyki matematycznej, dotyczy zbieżności sum niezależnych zmiennych o takich samych rozkładach (rozkład nie musi być znany) z rozkładem normalnym.

Założmy, że dany jest ciąg  $X_1, X_2, \dots, X_n$  niezależnych zmiennych losowych o jednakowym rozkładzie (oznacza to, że zmienne posiadają jednakowe rozkłady prawdopodobieństwa, wartości oczekiwane i wariancje):

$$E(X_1) = E(X_2) = E(X_3) = \dots = E(X_n) = m$$

$$V(X_1) = V(X_2) = V(X_3) = \dots = V(X_n) = \sigma^2$$

Oznaczmy przez  $Z_n$  następującą zmienną losową:  $Z_n = X_1 + X_2 + \dots + X_n$

Wartość oczekiwana i wariancja zmiennej  $Z_n$ :

$$E(Z_n) = nm$$

$$V(Z_n) = n\sigma^2$$

Centralne twierdzenie graniczne mówi, że jeśli  $n$  jest duże, to rozkład zmiennej losowej  $Z_n$  można przybliżyć rozkładem normalnym z wartością oczekiwaną  $nm$  i odchyleniem standardowym  $\sigma\sqrt{n}$ , czyli

$$Z_n : N(nm; \sigma\sqrt{n})$$

### ***Wniosek z centralnego twierdzenia granicznego Lindeberga - Levy'ego***

Założmy, że ciąg niezależnych zmiennych losowych  $X_1, X_2, \dots, X_n$ , spełnia założenia centralnego twierdzenia granicznego.

Zmienna:  $\bar{X}_n = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$ , wartość oczekiwana  $E(\bar{X}_n) = \frac{1}{n} nm = m$  i wariancja

$$V(\bar{X}_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Z centralnego twierdzenia granicznego wynika, że  $Z_n = X_1 + X_2 + \dots + X_n$  ma w przybliżeniu rozkład normalny, stąd  $\bar{X}_n$  ma również rozkład normalny. Przy dużych wartościach  $n$  rozkład zbliżony jest do

rozkładu:  $N(m; \frac{\sigma}{\sqrt{n}})$ .

### **Rozkład chi - kwadrat**

Przejdźmy do omówienia pewnych rozkładów związanych z rozkładem normalnym.

Wyznaczmy rozkład zmiennej losowej:

$$Y = \chi_n^2 = X_1^2 + X_2^2 + \dots + X_n^2,$$

gdzie  $X_1, X_2, \dots, X_n$  są niezależnymi zmiennymi losowymi, z których każda ma rozkład normalny  $N(0,1)$ . Zmienna losowa  $Y$  ma rozkład gamma o parametrach  $p = \frac{1}{2}n$ ,  $b = \frac{1}{2}$ , a więc ma gęstość

$$f(x) = \begin{cases} 0 & \text{dla } x \leq 0, \\ \frac{1}{2^{\frac{1}{2}n} \Gamma(\frac{1}{2}n)} x^{\frac{1}{2}n-1} e^{-\frac{1}{2}x} & \text{dla } x > 0 \end{cases}$$

Rozkład ten nosi nazwę **rozkładu chi - kwadrat** o  $n$  stopniach swobody. Zmienną losową o tym rozkładzie oznaczamy jako  $\chi_n^2$ . Występujący tu parametr  $n$  oznacza liczbę stopni swobody, tzn. liczbę niezależnych składników, tworzących tę zmienną losową.

Zmienna losowa o rozkładzie chi - kwadrat przyjmuje wartości dodatnie, a jej rozkład zależy od liczby stopni swobody  $n$ . Dla małych wartości  $n$  jest to rozkład silnie asymetryczny, w miarę wzrostu  $n$  asymetria jest coraz mniejsza:  $n$  wyznaczamy najczęściej jako:

$$n = k - 1 \text{ lub } n = k - p - 1$$

gdzie:

$k$  - liczebność próby,

$p$  - liczba szacowanych parametrów z próby.

Wartość oczekiwana w rozkładzie ( $\chi^2$ ):  $E(\chi^2) = n$

Wariancja w rozkładzie ( $\chi^2$ ):  $D^2(\chi^2) = 2n$

Bardzo często interesować nas będą prawdopodobieństwa postaci:

$$P(\chi_n^2 \geq \chi_\alpha^2) = \alpha$$

W tablicach zawarte są wartości prawdopodobieństwa  $\chi_\alpha^2$  dla pewnych wartości  $\alpha$  i  $n \leq 30$ .

Dla  $n > 30$  rozkład chi - kwadrat można z bardzo dobrą dokładnością aproksymować rozkładem normalnym  $N(n; \sqrt{2n})$ .

Jest oczywiste, że rozkład chi - kwadrat o  $n = 2$  stopniach swobody jest **rozkładem wykładniczym**. Ważne dla zastosowań jest następujące twierdzenie:

**Twierdzenie.** Niech  $X_1, X_2, \dots, X_n$  będą niezależnymi zmiennymi losowymi o identycznych rozkładach normalnych  $N(m; \sigma)$ . Niech:

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \quad S^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2,$$

Wówczas

- 1 Zmienne losowe  $\bar{X}$  i  $S^2$  są niezależne.
- 2 Zmienna losowa

$$Z = \frac{nS^2}{\sigma^2}$$

ma rozkład chi - kwadrat o  $n-1$  stopniach swobody.

## Rozkład t Studenta

**Definicja.** Rozkładem t Studenta o  $n$  stopniach swobody nazywamy rozkład prawdopodobieństwa zmiennej losowej:

$$t_n = \frac{X}{\sqrt{\frac{1}{n} \chi_n^2}} \quad \text{gdzie } X \text{ i } \chi_n^2 - \text{niezależne zmienne losowe}$$

$X$  ma rozkład normalny  $N(0,1)$ , a  $\chi_n^2$  ma rozkład chi - kwadrat o  $n$  stopniach swobody. Gęstość prawdopodobieństwa rozkładu  $t$  Studenta ma postać:

$$f_{t_n}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

W tablicach rozkładu  $t$  Studenta podaje się wartości  $t_\alpha$  dla pewnych wartości  $n$  i  $\alpha$  spełniających zależność:

$$P(|t| > t_\alpha) = 2 \int_{t_\alpha}^{\infty} f_{t_n}(t) dt = \alpha$$

Rozkład  $t$  Studenta ma bardzo ważną własność, która zdecydowała o jego szerokich zastosowaniach. Można wykazać, że rozkład prawdopodobieństwa zmiennej losowej:

$$Y = \frac{X}{\sqrt{\frac{1}{n}(X_1^2 + X_2^2 + \dots + X_n^2)}}$$

gdzie  $X, X_1, X_2, \dots, X_n$  - niezależne zmienne losowe o jednakowych rozkładach  $N(0; \sigma)$

ma rozkład  $t$  Studenta i **nie zależy od wartości  $\sigma$** .

Druga ważna własność tego rozkładu, to jego szybka zbieżność do rozkładu normalnego. Oznacza to, że dla dużych  $n$  ( $n > 30$ ) rozkład  $t$  Studenta można aproksymować rozkładem normalnym.

**Twierdzenie.** Jeśli cecha  $X$  elementów populacji ma rozkład normalny  $N(m; \sigma)$ , to:

1. statystyki

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \quad \text{i} \quad S_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 \quad \text{są niezależne,}$$

2. statystyka  $nS_n^2 / \sigma^2$  ma rozkład chi - kwadrat o  $n-1$  stopniach swobody,

3. statystyka  $t = \frac{\bar{X}_n - m}{S_n} \sqrt{n-1}$  ma rozkład  $t$  Studenta o  $n-1$  stopniach swobody.

## Rozkład $F$ Snedecora

**Definicja.** Rozkładem  $F$  Snedecora o  $(m, n)$  stopniach swobody nazywamy rozkład prawdopodobieństwa ilorazu zmiennych losowych:

$$F = \frac{X/m}{Y/n} \quad \text{gdzie } X, Y \text{ są niezależnymi zmiennymi losowymi o}$$

rozkładach chi - kwadrat, odpowiednio z  $m$  i  $n$  stopniach swobody.

Gęstość prawdopodobieństwa tej zmiennej losowej wyraża się wzorem:



$$f_F(z) = \begin{cases} 0 & \text{dla } z \leq 0, \\ \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{n/2} \frac{z^{m/2-1}}{(z+n/m)^{(m+n)/2}} & \text{dla } z > 0. \end{cases}$$

W zależności od  $m$  i  $n$  wartości zmiennej losowej  $F_\alpha$  umieszczono w tablicy, w taki sposób, że dla danych wartości prawdopodobieństw  $\alpha$  zależność:

$$P(F_{m,n} \geq F_\alpha) = \alpha$$

## Szereg rozdzielczy. Histogram

**Szereg statystyczny** - jest to zbiór wyników obserwacji uporządkowanych według określonych cech (kryteriów), których miernikiem są zmienne.

Inaczej mówiąc, **szeregiem statystycznym** nazywamy ciąg liczbowy monotoniczny, ograniczony z góry i z dołu (tj. taki, którego wyrazy występują tylko w pewnym przedziale wartości). Szereg składa się zazwyczaj z dwóch kolumn, z których jedna podaje wielkości cechy lub czas, druga zaś informuje o liczbie jednostek przypadających na daną kategorię przedmiotów lub zjawisk lub mówi o ich natężeniu występującym w danym czasie.

Najczęściej wyróżnia się dwa kryteria podziału szeregów:

- **kryterium formalne** - związane z budową szeregu, na podstawie którego możemy wyodrębnić: *szeregi szczegółowe, szeregi rozdzielcze i szeregi skumulowane*,
- **kryterium merytoryczne** - wynikające z typu badanej cechy zbiorowości, według którego wyróżnia się: *szeregi czasowe i szeregi przestrzenne*.

**Sposób grupowania** cech zależy od: rodzaju badania (przekrojowe, czasowe), rodzaju cechy statystycznej, sposobu pomiaru oraz liczby obserwacji (*szeregi szczegółowe, rozdzielcze*).

| Szeregi statystyczne |   |   |
|----------------------|---|---|
| szczegółowe          | <b>rozdzielcze z cechą mierzalną (ilościową):</b> <ul style="list-style-type: none"><li>- punktowe (proste, skumulowane),</li><li>- przedziałowe (proste, skumulowane),</li></ul> <b>rozdzielcze z cechą niemierzalną (jakościową):</b> <ul style="list-style-type: none"><li>- geograficzne</li><li>- inne</li></ul> | <b>czasowe</b> <ul style="list-style-type: none"><li>- momentów</li><li>- okresów</li></ul> |

Zbiory obserwacji na populacji generalnej mogą więc tworzyć:

**Szereg szczegółowy** - uporządkowany ciąg wartości badanej cechy statystycznej, stosowany, gdy przedmiotem badania jest niewielka liczba jednostek, np. zmienna  $X$  przyjmuje wartości:  $x_1, x_2, \dots, x_n$ , wartości cechy porządkujemy rosnąco:  $x_1 \leq x_2 \leq \dots \leq x_n$  lub malejąco  $x_1 \geq x_2 \geq \dots \geq x_n$ .

**Szereg rozdzielczy** - stanowi zbiorowość statystyczną, podzieloną na części (**klasy**) według określonej cechy jakościowej lub ilościowej z podaniem liczebności lub częstości każdej z wyodrębnionych klas. Szeregi rozdzielcze mogą dotyczyć zarówno cechy jakościowej, jak i ilościowej. Charakteryzują one strukturę danej zbiorowości stąd nazywane są czasem *szeregi strukturalnymi*.

**Rozkład empiryczny** - zestawienie wyników w postaci szeregu rozdzielczego z cechą mierzalną, odzwierciedla strukturę badanej zbiorowości z punktu widzenia określonej cechy statystycznej.

Przy badaniu zbioru statystycznego, w wielu wypadkach grupujemy elementy zbioru w **klasy szeregu**, w których elementy mają argumenty równe lub należące do przedziałów o równych długościach.

Wspólną długość klas nazywamy – **szerokością klasy**.

Wartości argumentów rozdzielaających sąsiednie klasy nazywamy – **granicami przedziału klasowego**.

Przykład szeregu rozdzielczego (rozkład empiryczny):

| Nr klasy szeregu rozd. | Wiek uczniów    | Liczebność |
|------------------------|-----------------|------------|
| 1                      | $x \leq 7$      | 50         |
| 2                      | $7 < x < 8$     | 200        |
| 3                      | $8 \leq x < 9$  | 521        |
| 4                      | $9 \leq x < 10$ | 385        |

**Częstość klasy** – ilość elementów w poszczególnych klasach.

**Częstość względna klasy** – stosunek częstości argumentów w danej klasie  $f_i$  do ogólnej liczebności zbioru  $N$ :

$$\frac{f_1}{N}, \frac{f_2}{N}, \dots, \frac{f_i}{N}$$

Oczywiście, suma wszystkich częstości względnych równa się 1:

$$\frac{f_1}{N} + \frac{f_2}{N} + \dots + \frac{f_k}{N} = 1 \quad \text{gdzie: } k - \text{liczba klas}$$

Częstości względne odgrywają rolę przybliżonych wartości prawdopodobieństw, których w praktyce nie można obliczyć. Im większe są liczebności zbiorów, tym bardziej możemy uważać te częstości względne za przybliżone wartości prawdopodobieństw i stosować prawo wielkich liczb ze wszystkimi jego wnioskami.

**Częstość skumulowana** – suma częstości poprzednich klas i danej klasy.

Przykład:

| Nr klasy | Przedziały klasowe (wiek uczniów) | Środek przedziału x | Częstość | Częstość skumulowana | Częstość względna g(x) | Częstość względna skumulowana G(x) |
|----------|-----------------------------------|---------------------|----------|----------------------|------------------------|------------------------------------|
| 1        | $5.5 \leq x < 6.5$                | 6                   | 20       | 20                   | 20/668                 | 20/668                             |
| 2        | $6.5 \leq x < 7.5$                | 7                   | 252      | 272                  | 252/668                | 272/668                            |
| 3        | $7.5 \leq x < 8.5$                | 8                   | 386      | 658                  | 386/668                | 658/668                            |
| 4        | $8.5 \leq x < 9.5$                | 9                   | 10       | 668                  | 10/668                 | 668/668 = 1                        |

Sposoby prezentacji danych.

**1. Tablice statystyczne** - są wykorzystywane do prezentacji danych statystycznych według określonego kryterium.

**Podział tablic statystycznych:**

- **proste** - charakteryzują strukturę lub dynamikę jednej zbiorowości pod względem jednej cechy (ilościowej lub jakościowej),
- **złożone** - opisują badaną zbiorowość według kilku cech lub kilka zbiorowości według jednej cechy (szczególna rola **tablic dwudzielnych - korelacyjnych**).

**Przykład:**

**Województwa Polski według liczby gmin i powierzchni**

| Powierzchnia (w tys. km <sup>2</sup> ) | Liczba gmin |       |       |       |       |       | Razem |
|--|-------------|-------|-------|-------|-------|-------|-------|
|  | 17-27       | 28-38 | 39-49 | 50-60 | 61-71 | 72-82 |       |
| 1,5-3,1                                | 1           |       |       |       |       |       | 1     |
| 3,1-4,7                                |             | 3     | 8     | 2     |       |       | 13    |
| 4,7-6,3                                |             |       | 8     | 4     |       |       | 12    |
| 6,3-7,9                                |             | 1     | 2     | 3     | 3     |       | 10    |
| 7,9-9,5                                |             |       | 3     | 1     | 2     | 2     | 8     |
| 9,5-11,1                               |             |       | 1     | 3     |       |       | 4     |
| 11,1-12,7                              |             |       |       | 1     |       |       | 1     |
| Razem                                  | 1           | 4     | 22    | 14    | 5     | 2     | 49    |

Każda liczba w wewnętrznej części tabeli określa częstotliwość występowania dwóch cech.

**2. Wykres** - jest graficzną formą rejestracji danych oraz narzędziem prezentacji i analizy uogólnionych informacji statystycznych.

### Najczęściej stosowane typy wykresów:

- **histogramy (wykresy słupkowe)** - zbiór przylegających prostokątów, których podstawy, równe rozpiętości przedziałów klasowych - znajdują się na osi odciętych, a wysokości są liczebnościami (częstościami) przedziałów, w przypadku nierównych szerokości przedziałów - gęstościami liczebności (częstości).
- **diagramy, wykresy liniowe (wielobok liczebności)** - jest łamaną, powstałą przez połączenie punktów, których współrzędnymi są środki przedziałów klasowych i odpowiadające im liczebności (częstości lub gęstości).
- **krzywe liczebności (częstości) dla cechy ciągłej** - gęsta siatka punktów wyznaczająca wielobok liczebności, w konsekwencji wygładzona krzywa otrzymana przy zmniejszaniu rozpiętości przedziałów klasowych

### Tworzenie histogramu.

Jeżeli próba dotycząca jednej cechy mierzalnej nie jest zbyt liczna to dokonuje się jej wstępnego opracowania polegającego na uszeregowaniu w porządku rosnącym wyników próby. Otrzymany w ten sposób ciąg liczb – nazywa się szeregiem pozycyjnym.

Przykład:

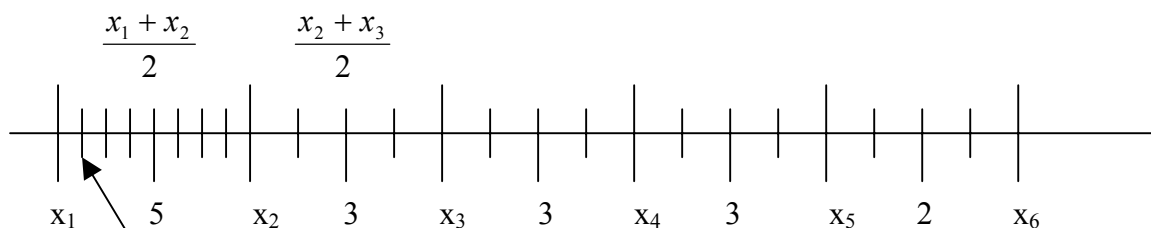
Wyniki „surowe” : 4, 5, 0, 1, 2, 4, 0, 9, 4, 5

Po uporządkowaniu – otrzymujemy szereg pozycyjny: 0, 0, 1, 2, 4, 4, 4, 5, 5, 9;

Jeżeli liczebność próby jest duża (orientacyjnie  $>30$ ) to pierwszym etapem jest uszeregowanie szeregu pozycyjnego, a dopiero drugim etapem dokonanie grupowania, czyli klasyfikacji.

Grupowanie polega na podziale próby na podzbiory zwane grupami lub klasami, a wartościami reprezentacyjnymi poszczególnych klas są ich środki.

Przedziały klasowe oraz ich liczebności, czyli liczby jednostek próby należących do danej klasy tworzą razem szereg rozdzielczy.



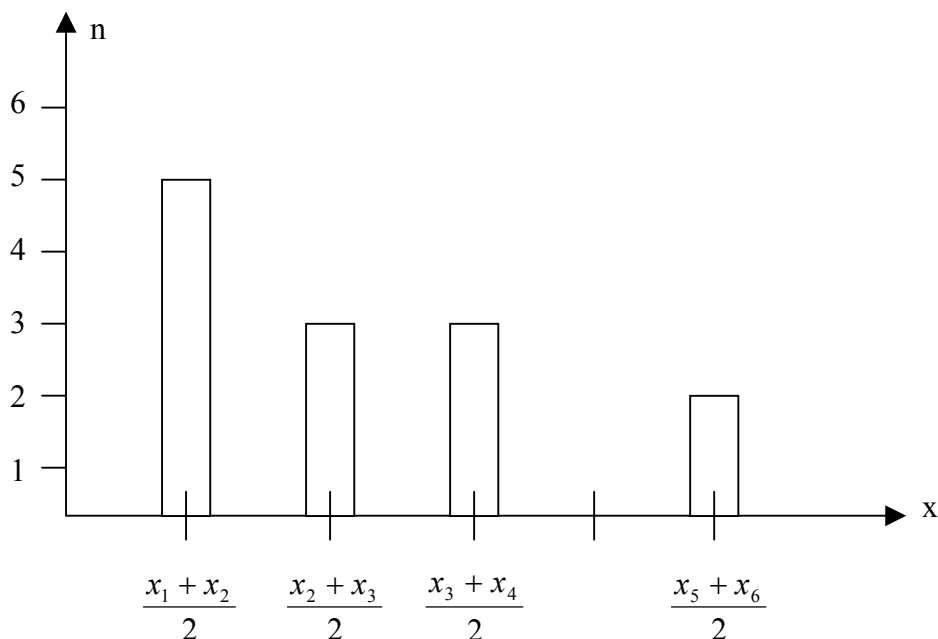
poszczególne wartości klasowe dotyczące  
np. badania (wyników) statystycznych

$x_1 - x_2$  – klasa

$\frac{x_1 + x_2}{2}$  - środek klasy

$\frac{x_1 + x_2}{2} ; 5 \quad \frac{x_2 + x_3}{2} ; 3 \quad \dots \quad \frac{x_5 + x_6}{2} ; 2$

## Histogram



Aby utworzyć szereg rozdzielczy należy:

1. ustalić obszar zmienności  $R$  badanej cechy czyli przedział ograniczony najmniejszym i największym elementem próby  $R = X_{\max} - X_{\min}$
2. wyznaczyć liczbę przedziałów klasowych  $m$  próby o liczebności  $n$ :

| Liczby klas w zależności od liczebności badanej zbiorowości |                               |
|---|-------------------------------|
| Liczba obserwacji<br>$n$                                    | Liczba zalecanych klas<br>$m$ |
| 40-60   | 6-8                           |
| 60-100  | 7-10                          |
| 100-200   | 9-12                          |
| 200-500   | 11-17                         |

Liczbę klas wyliczamy z jednego ze wzorów:

$$0,5\sqrt{n} \leq m \leq \sqrt{n}$$

$$m = 1 + 3,322 \log(n)$$

$$m < 5 \log(n)$$

$m$  – musi być zawsze liczbą całkowitą !

3. podzielić obszar zmienności na klasy i ustalić reprezentację klasy (środek podziału klasowego) oraz końce przedziałów klasowych
4.  $dd = \frac{X_{\max} - X_{\min}}{m}$  - szerokość przedziału (rozstęp przedziałowy).

Przy wyznaczaniu szerokości przedziału należy zawsze wartość niepełną zaokrąglić w górę, gdyż w przeciwnym przypadku nie wszystkie wartości zmiennej zostaną włączone do przedziałów.

wyznaczyć liczebność w klasach (od 0 do  $n$ )

5. wyznaczyć prawdopodobieństwo empiryczne:

$$p_j = \frac{f_j}{n} \quad j = 1 \dots m \quad m - \text{liczba przedziałów}$$

6. zbudować histogram

**Przykład:**

Podane niżej liczby informują o cenach pewnej akcji notowanych w ciągu 36 dni:

10, 13, 14, 16, 11, 18, 19, 20, 18, 15, 20, 21, 22, 25, 23, 22, 26, 27, 29, 28, 31, 30, 32, 34, 33, 38, 41, 40, 42, 51, 50, 46, 37, 29, 23, 28.

Zbuduj szereg statystyczny rozdzielczy przedziałowy.

$N=36$     $x_{\max} = 51$     $x_{\min} = 10$

$$m = 1 + 3,322 \cdot \log 36 = 6,17 \approx 6$$

stąd:

$$dd = \frac{51-10}{6,17} = 6,65 \approx 7$$

Wartości akcji grupujemy w 6 przedziałów o rozstępie równym 7 każdy.

| $x_i$   | Liczebności $f_i$ |
|---------|-------------------|
| 10 – 16 | 6                 |
| 17 – 23 | 10                |
| 24 – 30 | 8                 |
| 31 – 37 | 5                 |
| 38 – 44 | 4                 |
| 45 – 51 | 3                 |
|         | 36                |

**Przykład .**

Województwa Polski w układzie przestrzennym sprzed 1999 r. charakteryzuje między innymi: liczba gmin znajdująca się na terenie województwa (cecha skokowa X)

Struktura województw wg liczby gmin – dla cechy skokowej

**Szereg szczegółowy:**

17, 30, 32, 37, 37, 39, 40, 40, 40, 40, 41, 41, 42, 42, 43, 43, 43, 44, 45, 46, 46, 47, 47, 47, 48, 48, 49, 51, 54, 54, 55, 55, 55, 56, 57, 57, 58, 58, 58, 59, 59, 62, 63, 63, 65, 69, 74, 78, 91.

W przykładzie:

$$R = 91 - 17 = 74,$$

m

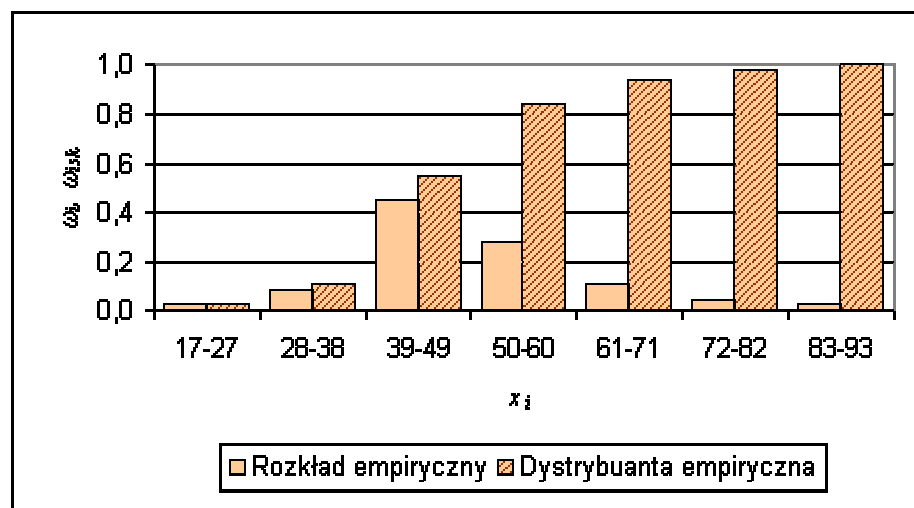
$$dd = 74/7 \approx 10,57 \approx 11$$

początek pierwszego przedziału klasowego  $x_{\min} = 17$

(przyjmujemy, że rozpiętość przedziałów klasowych jest taka sama dla wszystkich klas)

Rozkład empiryczny i dystrybuanta empiryczna – Struktura województw wg liczby gmin:

| Numer klasy | Liczba gmin | Liczba województw | Częstość   | Liczebność skumulowana | Częstość skumulowana |
|-------------|-------------|-------------------|------------|------------------------|----------------------|
| $i$         | $x_i$       | $n_i$             | $\omega_i$ | $n_{isk}$              | $\omega_{isk}$       |
| 1           | 17 - 27     | 1                 | 0,02       | 1                      | 0,02                 |
| 2           | 28 - 38     | 4                 | 0,08       | 5                      | 0,10                 |
| 3           | 39 - 49     | 22                | 0,45       | 27                     | 0,55                 |
| 4           | 50 - 60     | 14                | 0,29       | 41                     | 0,84                 |
| 5           | 61 - 71     | 5                 | 0,10       | 46                     | 0,94                 |
| 6           | 72 - 82     | 2                 | 0,04       | 48                     | 0,98                 |
| 7           | 83 - 93     | 1                 | 0,02       | 49                     | 1,00                 |
|             | $n =$       | 49                |            |                        |                      |



## Zadania

### Własności dystrybucyjny przydatne przy rozwiązywaniu zadań:

$F(-x) = 1 - F(x)$  wynika z symetrii funkcji gęstości względem  $x = 0$

$P(X < x) = F(x)$

$P(x_1 < X < x_2) = F(x_2) - F(x_1)$

$P(X > x) = 1 - F(x)$

$P(x_1 > X > x_2) = F(x_1) + F(x_2)$

Standaryzowany rozkład normalny  $N(0; 1)$   $u = \frac{x - m}{\sigma}$

### Zadanie 1:

Obliczyć prawdopodobieństwo  $P(|X| > 2)$  jeśli zmienna losowa  $X$  ma rozkład normalny  $N(-1; 2)$ .

### Zadanie 2:

W 3 grupach studenckich II roku zbadano przeciętną ilość nieobecności na zajęciach z fizyki i stwierdzono, że podlega ona rozkładowi normalnemu. Obliczone wartości średnie i wariancje ilości nieobecności dla każdej z grup kształtują się następująco:

| Grupa | Ilość osób | Wartość średnia | Wariancja |
|-------|------------|-----------------|-----------|
| 1     | 30         | 12              | 2,25      |
| 2     | 25         | 14              | 9         |
| 3     | 20         | 16              | 9         |

Jakie jest prawdopodobieństwo, że łączna liczba nieobecności na zajęciach dla wszystkich studentów całego roku jest zawarta między 40 a 45?

### Zadanie 3:

W pewnym zakładzie przetwórczym do wypełniania kartonów z sokiem wykorzystywany jest automat. Waga soku w wypełnianych pojemnikach ma rozkład normalny  $N(1 \text{ kg}; 0,05 \text{ kg})$ . Jakie jest prawdopodobieństwo tego, że:

- waga losowo wybranego kartonu jest mniejsza niż 1 kg,
- waga losowo wybranego kartonu jest zawarta w przedziale 0,95 – 1,05 kg,
- waga losowo wybranego kartonu przekroczy 1,05 kg.

### Zadanie 4:

W Katowicach zbadano liczbę dni z przekroczonym zapyleniem w ciągu trzech kolejnych miesięcy kwartału. Stwierdzono, że liczba tych dni podlega rozkładowi normalnemu. Obliczone wartości średnie i wariancje tej zmiennej dla każdego z miesięcy kształtują się następująco:

| Miesiąc  | Wart. średnia | Wariancja |
|----------|---------------|-----------|
| kwiecień | 15            | 4         |
| maj      | 9             | 1         |
| czerwiec | 16            | 4         |

Jakie jest prawdopodobieństwo, że łączna liczba dni z przekroczonym zapyleniem w ciągu badanego kwartału nie przekracza 30 dni?

### Zadanie 5:

W województwach A i B zbadano roczną ilość pożarów. Okazało się, że zarówno w jednym jak i w drugim z województw badana zmienna podlega rozkładowi normalnemu.



Dla województwa A jest to rozkład  $X: N(120;12)$ , a dla województwa B  $Y: N(180;16)$ . Jakie jest prawdopodobieństwo, że w ciągu roku ilość pożarów będzie w obu województwach łącznie niższa niż 280?

### Zadanie 6:

Wzrost mężczyzn w pewnej populacji ma rozkład normalny  $N(180 \text{ cm} ; 12)$ . Jaki jest udział w populacji mężczyzn o wzroście:

- do 170 cm,
- w przedziale 175 –180 cm,
- powyżej 185 cm.

### Zadanie 7:

Rozkład płac pracowników w firmie A jest normalny z wartością oczekiwaną  $m = 2000,00$  zł. Wybrano losowo 25 pracowników. Obliczyć prawdopodobieństwo, że średnia płaca wylosowanych pracowników jest większa od 1800,00 zł, jeśli wariancja płacy pracowników firmy A jest równa:  $\sigma^2 = 1,44$ .

### Zadanie 8

#### Wyniki kolokwium

Problem:

Poniższa tabela zawiera zestawienie wyników z kolokwium z fizyki w grupach I, II i III pierwszego roku. Opiszmy uzyskany rozkład wyników.

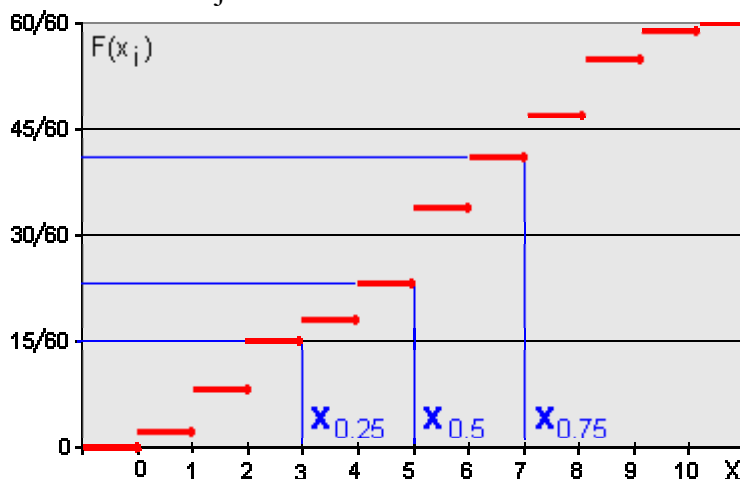
|            |   |   |   |   |   |    |   |   |   |   |    |
|------------|---|---|---|---|---|----|---|---|---|---|----|
| I. punktów | 0 | 1 | 2 | 3 | 4 | 5  | 6 | 7 | 8 | 9 | 10 |
| I. osób    | 2 | 6 | 7 | 3 | 5 | 11 | 7 | 6 | 8 | 4 | 1  |

#### Rozwiązanie:

Jak łatwo policzyć kolokwium pisało 60 osób. Jeżeli jako zmienną losową  $X$  zdefiniujemy wynik kolokwium przypadkowo wybranego studenta pierwszego roku to rozkład gęstości prawdopodobieństwa zmiennej losowej  $X$  oraz jej dystrybuenta będą następujące:

|                  |        |        |        |         |         |         |         |         |         |         |         |         |
|------------------|--------|--------|--------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| $x_i$            | 0      | 1      | 2      | 3       | 4       | 5       | 6       | 7       | 8       | 9       | 10      | powyżej |
| $p_i = P(X=x_i)$ | 2 / 60 | 6 / 60 | 7 / 60 | 3 / 60  | 5 / 60  | 11 / 60 | 7 / 60  | 6 / 60  | 8 / 60  | 4 / 60  | 1 / 60  | 0       |
| $F(x_i)$         | 0 / 60 | 2 / 60 | 8 / 60 | 15 / 60 | 18 / 60 | 23 / 60 | 34 / 60 | 41 / 60 | 47 / 60 | 55 / 60 | 59 / 60 | 60 / 60 |

Z powyższej tabeli odczytujemy rozstęp  $R = (10 - 0) = 10$ . Z wykresu dystrybenty natomiast bez trudu możemy odczytać podstawowe informacje.



Największą wartością zmiennej losowej  $X$  dla której wartość dystrybenty jeszcze nie przekroczyła

połowy jest pięć, tak więc mediana  $x_{0,5} = 5$ . Podobnie wyznaczamy dolny i górny kwartyl:

$$x_{0,25} = 3, \quad x_{0,75} = 7.$$

| A        | B     | C     | D               | E                             | F                             | G                             |
|----------|-------|-------|-----------------|-------------------------------|-------------------------------|-------------------------------|
| $x_i$    | $n_i$ | $p_i$ | $x_i \cdot p_i$ | $(x_i - \bar{x})^2 \cdot p_i$ | $(x_i - \bar{x})^3 \cdot p_i$ | $(x_i - \bar{x})^4 \cdot p_i$ |
| 0        | 2     | 0.033 | 0.00            | 0.82                          | -4.05                         | 20.1                          |
| 1        | 6     | 0.100 | 0.10            | 1.58                          | -6.26                         | 24.8                          |
| 2        | 7     | 0.117 | 0.23            | 1.03                          | -3.07                         | 9.1                           |
| 3        | 3     | 0.050 | 0.15            | 0.19                          | -0.38                         | 0.8                           |
| 4        | 5     | 0.083 | 0.33            | 0.08                          | -0.08                         | 0.1                           |
| 5        | 11    | 0.183 | 0.92            | 0.00                          | 0.00                          | 0.0                           |
| 6        | 7     | 0.117 | 0.70            | 0.12                          | 0.13                          | 0.1                           |
| 7        | 6     | 0.100 | 0.70            | 0.41                          | 0.84                          | 1.7                           |
| 8        | 8     | 0.133 | 1.06            | 1.22                          | 3.70                          | 11.2                          |
| 9        | 4     | 0.067 | 0.60            | 1.09                          | 4.39                          | 17.7                          |
| 10       | 1     | 0.017 | 0.17            | 0.43                          | 2.16                          | 10.9                          |
| $\Sigma$ | 60    | 1.000 | 4.97            | 6.97                          | -2.62                         | 96.5                          |

#### Zadanie 9.

Zmienna  $\chi^2$  ma rozkład o  $k = 25$  stopni swobody. Wyznaczyć  $\chi_1^2$  wiedząc, że  $F(\chi_1^2) = 0,95$ .

#### Zadanie 10.

Niech  $X_1, X_2, \dots, X_n$  będzie próbą prostą z populacji normalnej o rozkładzie  $N(m; 2)$ .

Obliczyć:

a)  $P(S^2 > 4.3)$  dla liczebności próby  $n=18$ .

b)  $P(S^2 < 3.9)$  dla liczebności próby  $n=51$ .

#### Zadanie 11.

Wiadomo, że błąd pomiaru pewnego przyrządu ma rozkład normalny  $N(0; \sigma)$  i z prawdopodobieństwem 0,95 nie wychodzi poza przedział  $(-1, 1)$ .

- Dokonano 10 niezależnych pomiarów tym przyrządem. Obliczyć prawdopodobieństwo, że wariancja empiryczna dla tej próby mieści się między 0,2 a 0,3.
- Dokonano 100 niezależnych pomiarów tym przyrządem. Obliczyć prawdopodobieństwo, że wariancja empiryczna dla tej próby jest większa od 0,28.

#### Zadanie 12.

Zużycie wody w pewnym osiedlu w ciągu dnia ma rozkład normalny  $N(m; 11)$ . Obliczyć prawdopodobieństwo, że empiryczna wariancja zużycia wody w losowo wybranym kwartale nie przekroczy 100 hektolitrów.

#### Zadanie 13.

OBOP ocenia, że 50% rodzin w Polsce żyje w ubóstwie. Wybrano losowo 100 rodzin. Jakie jest prawdopodobieństwo, że liczba rodzin żyjących w ubóstwie: przekracza 40?

#### Zadanie 14.

Wadliwość produktu A wynosi 5%. Pobrano losowo 100 sztuk. Jakie jest prawdopodobieństwo, że udział wadliwych sztuk jest większa od 4%?

**Zadanie 15.**

Średnia waga człowieka wynosi 75 kg z odchyleniem standardowym 3 kg. Samolot zabiera 81 pasażerów. Jakie jest prawdopodobieństwo, że łączna waga pasażerów przekroczy 6 ton?