STATYSTYKA MATEMATYCZNA

1. Pojęcia podstawowe

Populacja (***populacja generalna***) – zbiór elementów (osób, rzeczy, zjawisk), podlegających badaniu ze względu na jedną lub więcej cech.
Cechy statystyczne mogą być:

- *mierzalne* (*ilościowe*) – przyjmują wartości ze zbioru liczbowego,
  np.: długość, waga
- *niemierzalne* (*jakościowe*) – cechy których nie można wyrazić ilościowo, są opisywane słownie lub wyrażane przy pomocy wybranej skali,
  np.: płeć, kolor, funkcjonalność.

Próba (***populacja próbna***) – wybrany w określony sposób (np. przez losowanie) podzbiór populacji generalnej.
Wartości prób mogą być prezentowane w formie tzw. ***szeregów***.
Szereg prosty – wartości porządkowane są rosnąco lub malejąco.

długość	2.9	3.0	3.2	3.3	3.4	3.5	3.5	3.6	4.0	4.1

Szereg rozdzielczy – wartości dzielone są na *klasy* (kategorie), dla każdej klasy podawana jest jej *liczebność* lub *częstość* (stosunek liczebności klasy do liczebności całej próby).

długość	[2.5 3.0)	[3.0 3.5)	[3.5 4.0)	[4.0 4.5]
liczebność	1	4	3	2
częstość	0.1	0.4	0.3	0.2

Zmienna – to wielkość, która może przyjmować wartości z określonego zbioru.

Zmienna losowa – to zmienna, która w wyniku pewnego doświadczenia przyjmuje wartość z określonego zbioru z pewnym prawdopodobieństwem.

Skokowa (dyskretna) zmienna losowa – zmienna losowa która przyjmuje skończoną lub przeliczalną liczbę wartości.

Ciągła zmienna losowa – zmienna losowa której zbiór wartości jest nieskończony i nieprzeliczalny, może być np. przedstawiony w postaci przedziału liczbowego.

Przykład 1. Doświadczenie polega na kontroli jakości 6 wybranych produktów z linii produkcyjnej.

Zmienna losowa „*Liczba wadliwych produktów*” jest zmienną skokową

(może przyjmować wartości 0, 1, ..., 6)

Przykład 2. Doświadczenie polega na rejestracji dziennej ilości sprzedanych sztuk wybranego produktu.

Zmienna losowa „*Liczba sprzedanych sztuk*” jest zmienną skokową

(może przyjmować wartości 0, 1, ...)

Przykład 3. Doświadczenie polega na pomiarze długości wybranych detali z linii produkcyjnej.

Zmienna losowa „*Długość detalu*” jest zmienną ciagłą

(może przyjmować wartości np.: 19.9..30.9).

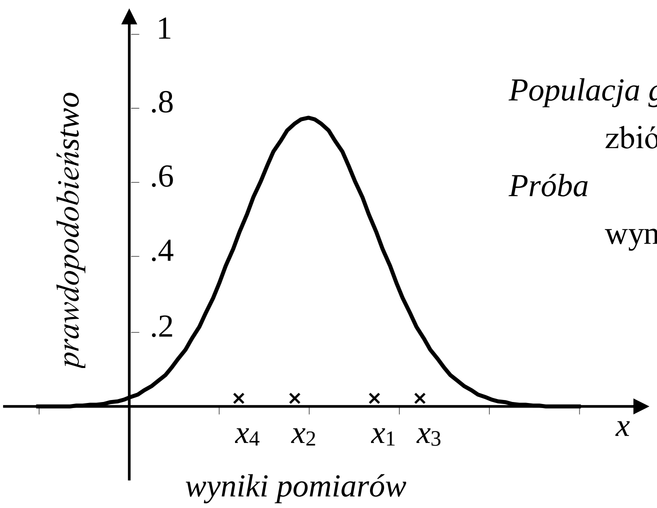
Pomiar jako zmienna losowa

Pomiar – czynności mające na celu wyznaczenie wartości wielkości fizycznej (*Encyklopedia PWN*).

Celem pomiaru jest określenie wartości liczbowej mierzonej wielkości.

Wynik pomiaru jest ustalany poprzez porównanie wielkości mierzonego obiektu z wielkością przyjętą za jednostkę miary tej wielkości. Wyniki pomiarów tej samej wielkości fizycznej różnią się. Różnice te są spowodowane niedokładnościami przyrządów, niedokładnościami metod pomiarowych itd.

Wynik pomiaru jest tylko przybliżeniem rzeczywistej wielkości mierzonej. Ze względu na występowanie błędów i niepewności pomiarowych ***wyniki pomiarów*** mogą być traktowane jako ***zmienne losowe*** (wyniki pomiarów przyjmują określone wartości liczbowe z pewnym prawdopodobieństwem).



Populacja generalna

zbiór wszystkich możliwych wyników pomiarów

Próba

wyniki faktycznie wykonanych pomiarów

2. Jednowymiarowe zmienne losowe

Jeżeli znany jest zbiór możliwych wartości zmiennej losowej oraz prawdopodobieństwa przyjęcia tych wartości przez zmienną losową (bądź też prawdopodobieństwa, że zmienna przyjmie wartość z określonego przedziału) to mówimy, że znany jest ***rozkład tej zmiennej losowej****.
* (Z.Pawłowski, *Wstęp do statystyki matematycznej)*.

Rozkładem prawdopodobieństwa zmiennej losowej X nazywana jest funkcja $P(S)$ oznaczająca prawdopodobieństwo tego, że zmienna losowa przyjmie wartość z S (funkcja ta przedstawia związek między wartościami zmiennej losowej a prawdopodobieństwami, z jakimi te wartości występują). Sposób przedstawiania rozkładu prawdopodobieństwa zależy od typu zmiennej losowej:

- dla *zmiennej losowej skokowej* podaje się wartości tej zmiennej wraz z odpowiadającymi im prawdopodobieństwami,
- dla *zmiennej losowej ciągłej* rozkład zmiennej losowej podaje się za pomocą *funkcji gęstości prawdopodobieństwa*.

Dystrybuanta zmiennej losowej X: F(x) – to funkcja opisująca prawdopodobieństwo wystąpienia wartości zmiennej X mniejszych od x:

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

Uwaga! $F(\infty) = 1$.

Do opisania ***rozkładu skokowej zmiennej losowej*** wystarczy podać wszystkie prawdopodobieństwa:

$$p_i = P(X = x_i), \quad i = 1, 2, \ldots$$

gdzie: X – zmienna losowa; x_i – i-ta wartość zmiennej losowej X; $P(X = x_i)$ – prawdopodobieństwo, że zmienna X przyjmie wartość xi; $\sum P(X = x_i) = 1$.

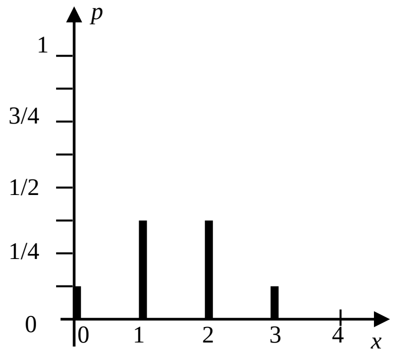
Dystrybuantę dyskretnej zmiennej losowej można zapisać wzorem:

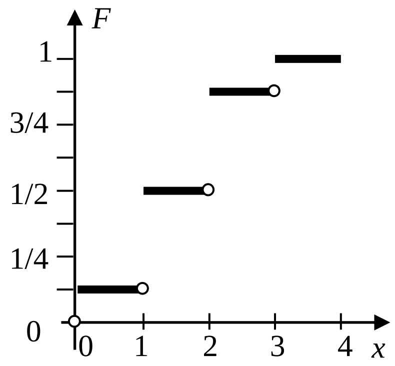
$$F(x) = P(X < x) = \sum_{x_i < x} P(X = x_i)$$

Funkcja rozkładu prawdopodobieństwa i dystrybuanta dyskretnej zmiennej losowej przedstawiane są w formie tabelarycznej lub w postaci wykresu.

Przykład 4.

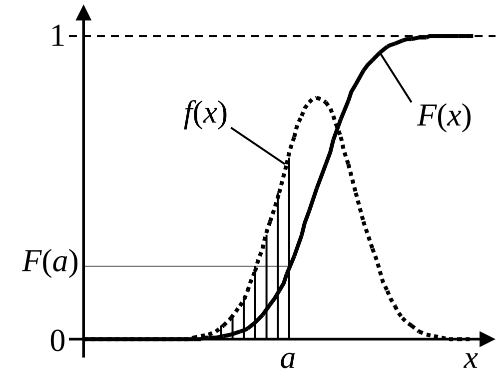
x_i	0	1	2	3	4
p_i	1/8	3/8	3/8	1/8	0
F	0	1/8	4/8	7/8	1



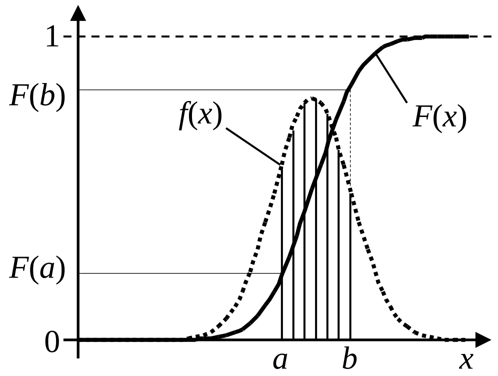


Do opisania ***rozkładu ciągłej zmiennej losowej*** wykorzystywana jest ***funkcja gęstości prawdopodobieństwa f***, dla której spełniona jest zależność:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)\,dt$$



$$F(a) = P(X < a) = \int_{-\infty}^{a} f(x)\,dx$$



$$F(b) - F(a) = P(a \le X < b) = \int_{a}^{b} f(x)\,dx$$

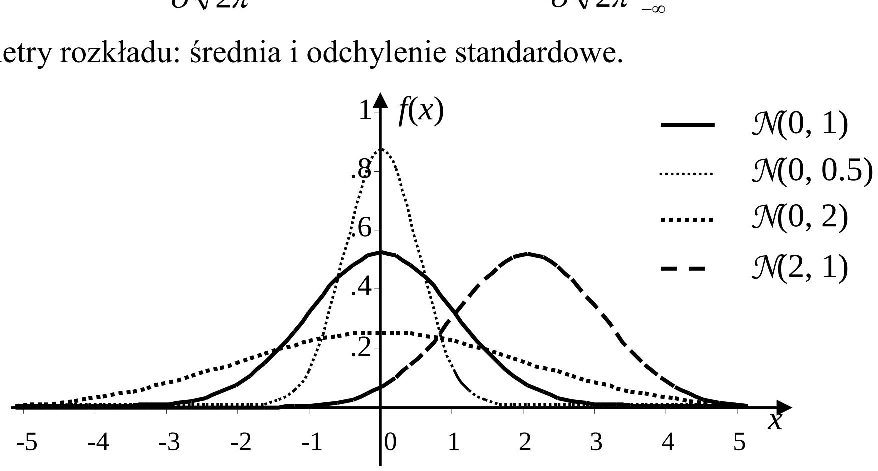
2.1. Rozkład normalny

Rozkład normalny (***rozkład Gaussa***) jest jednym z częściej spotykanych rozkładów zmiennych losowych ciągłych (wiele zjawisk fizycznych ma rozkład normalny).

Funkcja gęstości rozkładu f i ***dystrybuanta F*** rozkładu normalnego N(μ, σ) opisane są zależnościami:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx.$$

gdzie: μ, σ – parametry rozkładu: średnia i odchylenie standardowe.



Rys.1. Wykresy gęstości prawdopodobieństwa f(x) rozkładów normalnych
$\mathcal{N}(0, 1)$, $\mathcal{N}(0, 0.5)$, $\mathcal{N}(0, 2)$, $\mathcal{N}(2, 1)$

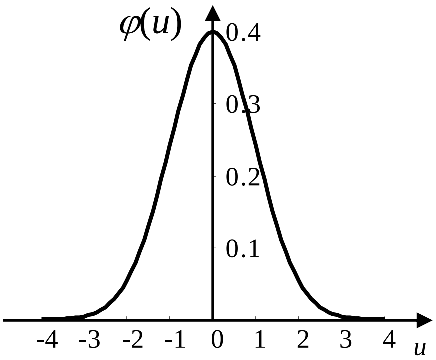
Zmienna losowa U utworzona ze zmiennej losowej X o rozkładzie normalnym $\mathcal{N}(\mu, \sigma)$ za pomocą przekształcenia *(średnia populacji μ jest odejmowana od każdej wartości cechy x, każda wyznaczona różnica dzielona jest przez odchylenie standardowe populacji σ)*:

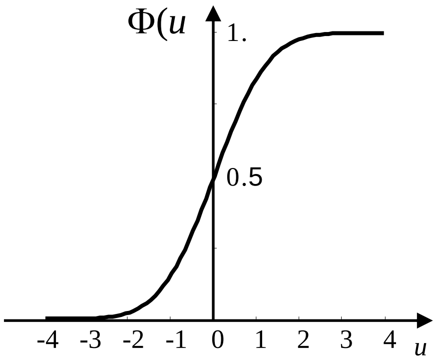
$$U = \frac{X - \mu}{\sigma}$$

ma rozkład normalny $\mathcal{N}(0, 1)$.

Zmienna U jest nazywana ***zmienną losową normalną standaryzowaną***, a rozkład $\mathcal{N}(0, 1)$ jest nazywany jest ***rozkładem normalnym standaryzowanym***. Funkcja gęstości prawdopodobieństwa i dystrybuanta rozkładu opisane są zależnościami:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{u^2}{2}}, \qquad \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{u^2}{2}}\, dt.$$





Rys.2. Wykresy gęstości prawdopodobieństwa φ(u) i dystrybuanty Φ(u) rozkładu normalnego $\mathcal{N}(0, 1)$.

W rzeczywistości wiele wielkości losowych ma w przybliżeniu rozkład normalny – rozkład ten ma bardzo duże znaczenie w statystyce i w zastosowaniach praktycznych.

Uzasadnieniem powszechności występowania rozkładów zbliżonych do normalnego jest *centralne twierdzenie graniczne*.

Jeżeli X_1, X_2, ..., X_n są niezależnymi zmiennymi losowymi o jednakowym rozkładzie o wartości oczekiwanej μ i wariancji σ^2 to dla $n \to \infty$ zmienna losowa:

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

ma w przybliżeniu rozkład $\mathcal{N}(0, 1)$.

Konsekwencją tego twierdzenia są wnioski:

Rozkład zmiennej losowej: $X_1 + X_2 + \ldots + X_n$ dla $n \to \infty$ jest zbieżny do $\mathcal{N}\left(n\mu, \sigma\sqrt{n}\right)$.

Rozkład zmiennej losowej: $\dfrac{X_1 + X_2 + \ldots + X_n}{n}$ dla $n \to \infty$ jest zbieżny do $\mathcal{N}\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

Dowodzi się także, że jeżeli X_1, X_2, ..., X_n są niezależnymi zmiennymi losowymi o rozkładach $\mathcal{N}_i(\mu_i, \sigma_i)$ to dla $n \to \infty$ zmienna losowa: $a_1X_1 + a_2X_2 + \ldots + a_nX_n$ ma rozkład normalny:

$$\mathcal{N}\left(a_1\mu_1 + \ldots + a_n\mu_n,\ \sqrt{a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2}\right).$$

Przykład 5.

Na podstawie pomiarów długości dużej partii detali wykonywanych na pewnym stanowisku stwierdzono, że rozkład długości jest rozkładem $\mathcal{N}(20,\ 1.5)$. Obliczyć prawdopodobieństwo, że długość losowo wybranego detalu:

a) jest mniejsza lub równa 20.5,

b) jest większa od 21.5,

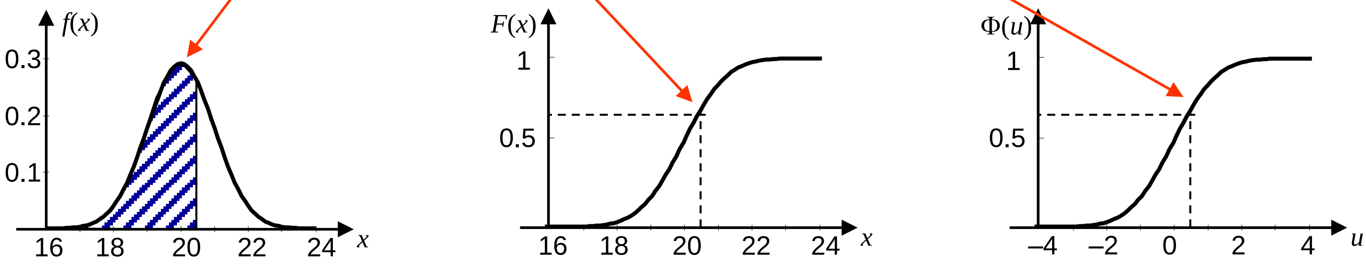
c) mieści się w przedziale (20.5 21.5],

d) co najmniej o 2 jednostki różni się od średniej,

e) obliczyć odchylenie od średniej dla którego prawdopodobieństwo wystąpienia detali o długości przekraczającej wyznaczone odchylenie wyniesie 0.1.

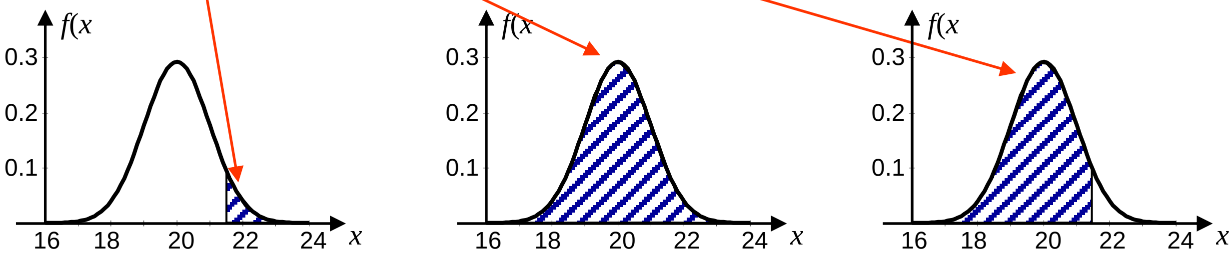
a)

$$P(x \le 20.5) = F_{\mathcal{N}(20,1.5)}(20.5) = \Phi\left(\frac{20.5-20}{1.5}\right) = \Phi(0.3333) = 0.6306$$



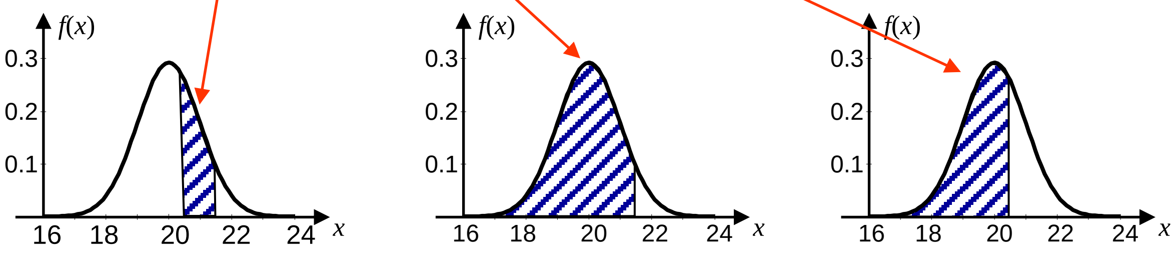
b)

$$P(x > 21.5) = 1 - P(x \le 21.5) = 1 - F_{\mathcal{N}(20,1.5)}(21.5) = 1 - \Phi\left(\frac{21.5-20}{1.5}\right) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$$



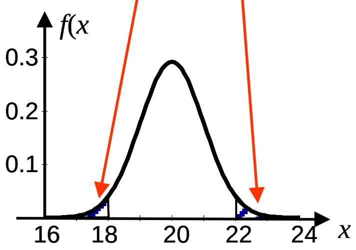
c)

$$P(20.5 < x \le 21.5) = P(x \le 21.5) - P(x \le 20.5) = F_{\mathcal{N}(20,1.5)}(21.5) - F_{\mathcal{N}(20,1.5)}(20.5) = \Phi(1) - \Phi(0.3333) = 0.2108$$

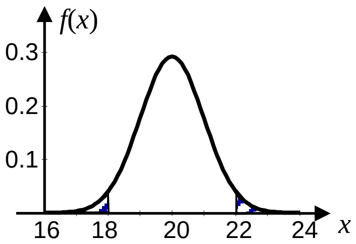


d)

$$P(|x-20|\ge 2)=P(x\le 18)+P(x\ge 22)=2\,P(x\le 18)=2F_{\mathcal{N}(20,1.5)}(18)=2\Phi\left(\frac{18-20}{1.5}\right)=2\Phi(-1.3333)=0.1824$$



e)



$$\left.\begin{array}{l} P(|x-20|\ge odl)=0.1 \\ P(|x-20|\ge odl)=2\,P(x\le 20-odl)=2\,F_{\mathcal{N}(20,1.5)}(20-odl) \end{array}\right\}$$

$$\rightarrow F_{\mathcal{N}(20,1.5)}(20-odl)=0.05 \longrightarrow 20-odl=F^{-1}_{\mathcal{N}(20,1.5)}(0.05)$$

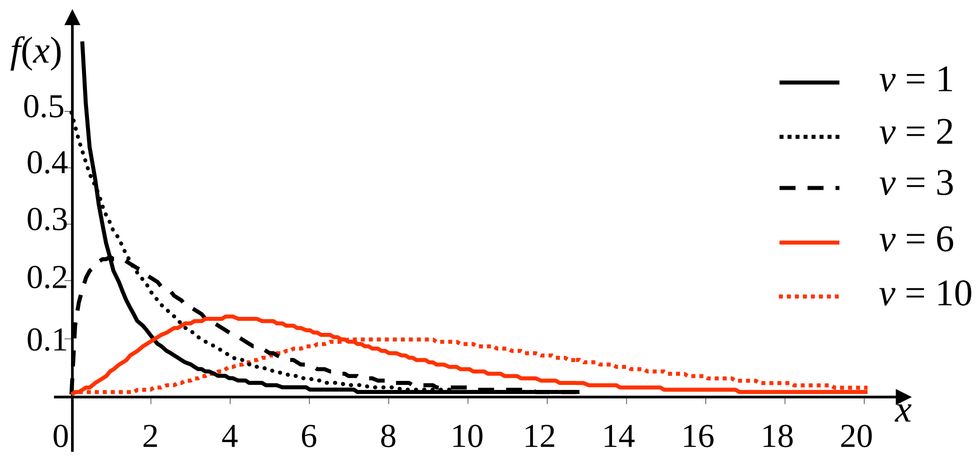
$$\rightarrow odl=20-F^{-1}_{\mathcal{N}(20,1.5)}(0.05) \longrightarrow odl=2.4673$$

2.2. Rozkład $\chi 2$

Rozkład χ2 (chi kwadrat). Zmienną o rozkładzie χ^2 o *n* stopniach swobody nazywana jest zmienna zdefiniowana w postaci sumy kwadratów *n* niezależnych zmiennych o rozkładzie normalnym standaryzowanym:

$$\chi^2 = X_1^2 + X_2^2 + \ldots + X_n^2$$

gdzie: X_1, X_2, ..., X_n – zmienne o rozkładzie $\mathcal{N}(0,\,1)$; *n* – liczba zmiennych niezależnych X_i w sumie; parametr rozkładu (jedyny) nazywany *liczbą stopni swobody*; *liczba stopni swobody* oznaczana jest także symbolem *v*.



Rys.3. Wykresy gęstości prawdopodobieństwa rozkładu χ^2 dla v = 1, 2, 3, 6, 10 stopni swobody.

Dla $v \to \infty$ rozkład χ^2 o jest zbieżny do rozkładu normalnego.

Zmienne losowe o rozkładzie χ^2

Zmienna losowa $$\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 \qquad (*)$$

gdzie: $\overline{X}=\frac{1}{n}\left(X_1+X_2+\ldots+X_n\right)$ ma rozkład χ^2 o $v=(n-1)$ *stopniach swobody*.

Zmienna ta, po przekształceniach, zapisywana jest także w postaci:

$$\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\frac{ns^2}{\sigma^2}$$

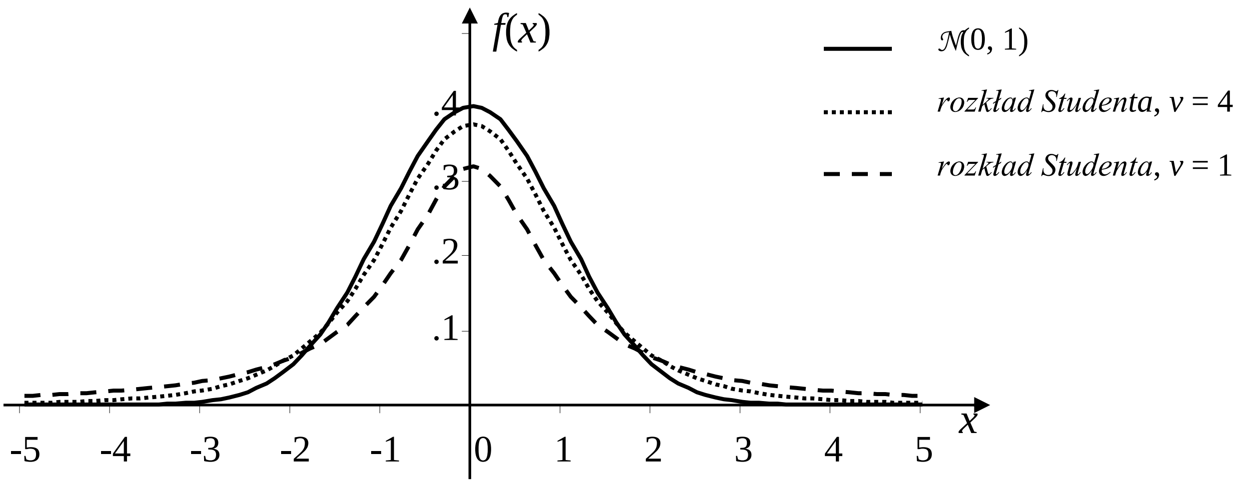
Zmienna (*) ma (n – 1) *stopni swobody* ponieważ tylko (n – 1) spośród n zmiennych X_1, X_2, ..., X_n jest liniowo niezależnych. Wartość jednej ze zmiennych można wyznaczyć wykorzystując pozostałe zmienne i średnią $\overline{X}$.

2.3. Rozkład t–Studenta

Rozkład t – Studenta. Zmienną o rozkładzie *t – Studenta* o n stopniach swobody nazywana jest zmienna zdefiniowana w postaci ilorazu zmiennej o rozkładzie normalnym standaryzowanym i zmiennej o rozkładzie χ^2 o n stopniach swobody:

$$t=\frac{U\sqrt{n}}{\sqrt{\chi^2}}$$

gdzie: U – zmienna o rozkładzie $\mathcal{N}(0,1)$; χ^2 – zmienna o rozkładzie χ^2 o n stopniach swobody; $v=n$ – liczba stopni swobody.



Rys.4. Wykresy gęstości prawdopodobieństwa rozkładu $\mathcal{N}(0, 1)$ i rozkładu t – Studenta dla v = 1, 4 stopni swobody.

Dla $v > 30$ rozkład t – Studenta pokrywa się z rozkładem $\mathcal{N}(0, 1)$.

Zmienne losowe o rozkładzie t – Studenta

Zmienna losowa $$\frac{\overline{X}-\mu}{s}\sqrt{n-1} \qquad (*)$$

gdzie: $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$, $s^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, $X_1, X_2, ..., X_n$ – zmienne o rozkładzie $\mathcal{N}(\mu, \sigma)$;

ma rozkład t – *Studenta* o $v = (n – 1)$ *stopniach swobody*.

Można pokazać, że:

1. zmienna: $\frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ ma rozkład $\mathcal{N}(0, 1)$;
2. zmienna: $\frac{n s^2}{\sigma^2}$ ma rozkład ma rozkład rozkładzie χ^2 o n –1 stopniach swobody.

Podstawiając zmienne (1) i (2) do definicji zmiennej o rozkładzie t – *Studenta* otrzymuje się zmienną (*). Zmienna ta ma więc rozkład t – *Studenta* o n – 1 *stopniach swobody*:

$$t=\frac{U\sqrt{n-1}}{\sqrt{\chi^2}}=\frac{\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\sqrt{n-1}}{\sqrt{\frac{n s^2}{\sigma^2}}}=\frac{\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\sqrt{n-1}\,\sigma}{\sqrt{n}\,s}=\frac{\left(\overline{X}-\mu\right)\sqrt{n-1}}{s}.$$

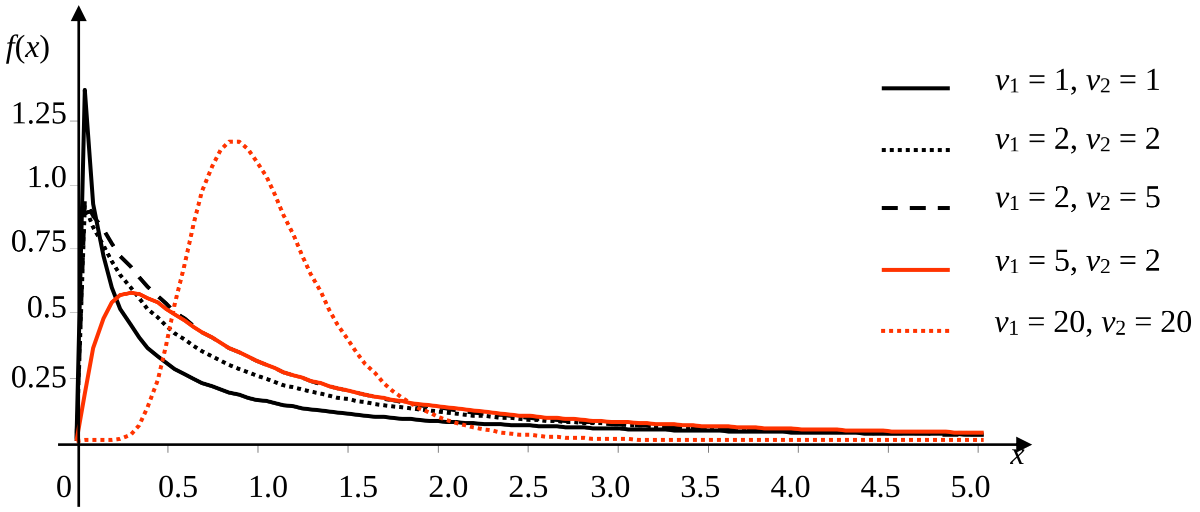
2.5. Rozkład F Snedecora (Fishera).

Zmienną o rozkładzie F i stopniach swobody v_1 i v_2 nazywana jest zmienna zdefiniowana w postaci ilorazu zmiennych o rozkładzie χ^2:

$$F=\frac{\chi_1^2}{v_1}:\frac{\chi_2^2}{v_2}$$

gdzie:

χ_1^2, χ_2^2 – zmienne o rozkładzie χ^2 z odpowiednio $v1$, $v2$ stopniami swobody.



Rys.5. Wykresy gęstości prawdopodobieństwa rozkładu F .

DZIEDZINY ZASTOSOWAŃ

Statystyka matematyczna zajmuje się *wnioskowaniem statystycznym*, tzn. wnioskowaniem o *populacji generalnej* na podstawie znajomości *próby*.

Podstawowymi działami statystyki są:

- *teoria estymacji*

  zajmuje się wnioskowaniem o własnościach rozkładu prawdopodobieństwa populacji generalnej na podstawie próby;

  estymacja parametryczna zajmuje się wyznaczaniem (szacowaniem) wartości nieznanych parametrów rozkładu, *estymacja nieparametryczna* – poszukuje postaci funkcyjnej rozkładu;

  szacowanie wartość parametru rozkładu populacji na podstawie próby nazywane jest *estymacją punktową*, *estymacja przedziałowa* wyznacza pewien przedział, do którego z określonym prawdopodobieństwem należy szacowana wartość parametru rozkładu;

- *teoria weryfikacji hipotez statystycznych*

  zajmuje się tworzeniem reguł umożliwiających rozstrzyganie o słuszności sądów (hipotez statystycznych);

  testy parametryczne służą do weryfikacji hipotez o nieznanych parametrach rozkładu ale znanym samym rozkładzie, *testy nieparametryczne* weryfikują hipotezy w których nie ma założeń o postaci rozkładu.

3. ESTYMACJA PUNKTOWA

3.1. Estymacja punktowa – miary położenia

W praktyce rozkład prawdopodobieństwa badanej zmiennej losowej może nie być znany – mogą być mierzone natomiast pewne wielkości wyznaczające przybliżony opis rozkładu. Miary położenia stosowane są do oceny miejsca skupienia wyników.

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

Średnia geometryczna $G = \sqrt[n]{\prod_{i=1}^{n} x_i}$,

Średnia harmoniczna $H = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$,

gdzie: n – liczebność próby; x_i, – i-ta wartość badanej cechy.

Moda (*wartość modalna, wartość najczęstsza*) M_0

wartość najczęściej występująca w próbie

***Kwantyl rzędu p** (0<p<1)*

wartość cechy x_p, która dzieli szereg na dwie części w taki sposób, że:

- w pierwszej części znajduje się 100p [%] elementów próbki (wartości tych elementów są mniejsze lub równe *kwantylowi* x_p),
- w drugiej części znajduje się 100(1–p) [%] elementów (wartości tych elementów są większe bądź równe *kwantylowi* x_p).

Kwartale to kwantyle rzędu 1/4, 2/4, ¾,

- *kwartyl dolny (pierwszy)* Q_1 (*kwantyl rzędu* p = ¼),
- *mediana* Q_2, *Me* (*kwantyl rzędu* p = ½),
- *kwartyl górny (trzeci)* Q_3 (*kwantyl rzędu* p = ¾).

Percentyle to kwantyle rzędu 1/100, 2/100,..., 99/100.

3.2. Estymacja punktowa – miary rozproszenia

Miary rozproszenia (rozrzutu) stosowane są do oceny stopnia rozproszenia wartości badanej cechy.

Odchylenie standardowe s

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}} \ (*) \qquad \text{lub} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}} ,$$

Wariancja s²

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \ (*) \qquad \text{lub} \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 ,$$

gdzie: n – liczebność próby; x_i – i-ta wartość badanej cechy; $\bar{x}$ – średnia arytmetyczna; * – małe próby.

Rozstęp r

różnica pomiędzy wartością największą i najmniejszą: $r = x_{max} - x_{min}$

Rozstęp międzykwartylowy IQR

$$IQR = Q_3 - Q_1$$

gdzie: Q_3, Q_1 – kwartyl górny i dolny.

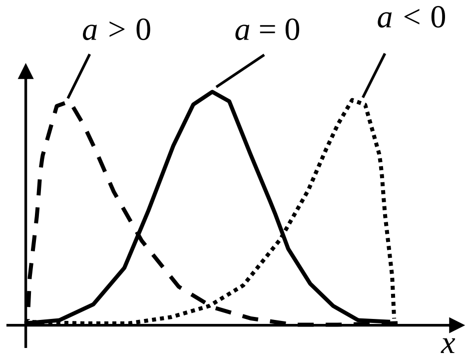
3.3. Estymacja punktowa – miary zniekształcenia

Miary zniekształcenia_stosowane są do oceny asymetrii i stopnia spłaszczenia rozkładu w stosunku do rozkładu normalnego.

Współczynnik skośności

$$a = \frac{M_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{s^3}$$

gdzie: n, x_i, $\overline{x}$ – jw.; M_3 – moment centralny rzędu 3.



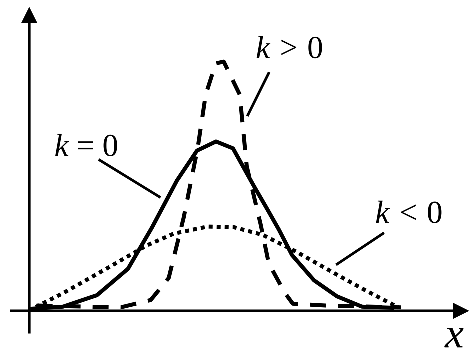
Jeśli współczynnik jest:

- = 0 – rozkład jest symetryczny,
- > 0 – prawa strona rozkładu jest wydłużona,
- < 0 – lewa strona rozkładu jest wydłużona.

Współczynnik spłaszczenia (kurtoza)

$$k = \frac{M_4}{s^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{s^4} - 3$$

gdzie: n, x_i, $\overline{x}$ – jw.; M_4 – moment centralny rzędu 4.



Współczynnik wykorzystywany do porównania rozkładu z rozkładem normalnym. Jeśli jest:

- = 0 – rozkład jest podobny do r. normalnego,
- > 0 – rozkład jest bardziej stromy od normalnego,
- < 0 – rozkład jest bardziej spłaszczony od normalnego,