STATYSTYKA to nauka, której przedmiotem zainteresowania są metody pozyskiwania i prezentacji, a przede wszystkim analizy danych opisujących zjawiska masowe. Metody statystyczne oparte są na rachunku prawdopodobieństwa.

STATYSTYCZNA ANALIZA DANYCH:
etap badania statystycznego polegający na wykrywaniu
- przy użyciu odpowiednich metod - prawidłowości
kształtowania się zjawisk statystycznych oraz
związków i zależności miedzy nimi, a także na interpretacji
wyników badań i formułowaniu wniosków

ZDARZENIE ELEMENTARNE to możliwy wynik doświadczenia losowego. Wszystkie takie możliwe wyniki tworzą zbiór zdarzeń elementarnych.

ZMIENNA LOSOWA, to funkcja, która zdarzeniom losowym przypisuje liczby. Na przykład, losując z pewnej populacji jednego osobnika przypisujemy mu jego wagę.

Rodzaje zmiennych losowych:
1) skokowa (dyskretna)
2) ciągła

PRAWDOPODOBIEŃSTWEM (wg Laplace) zajścia zdarzenia A nazywamy iloraz liczby zdarzeń sprzyjających zdarzeniu A do liczby wszystkich możliwych przypadków, zakładając, że wszystkie przypadki wzajemnie się wykluczają i są jednakowo możliwe.

$$P(A) = \frac{|A|}{|\Omega|}$$

PRAWDOPODOBIEŃSTWO - definicja częstościowa (Von Mises)

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

gdzie n_A to liczba rezultatów sprzyjających zdarzeniu A po n próbach

AKSJOMATYCZNA DEFINICJA PRAWDOPODOBIEŃSTWA (KOŁMOGOROWA)

1) Dla danego zbioru E zachodzi:

$$0 \leq P(E) \leq 1$$

Oznacza to, że prawdopodobieństwo zbioru zdarzeń E jest liczbą rzeczywistą większą lub równą 0 i mniejszą lub równą

2)

$$P(\Omega) = 1$$

prawdopodobieństwo, że wystąpi jakieś zdarzenie elementarne w przestrzeni wynosi 1. Innymi słowy: nie ma zdarzeń elementarnych poza zbiorem Ω.

3)

Każdy przeliczalny ciąg parami rozłącznych zdarzeń elementarnych E1, E2, ... spełnia własność:

$$P(E_1 \cup E_2 \cup E_3 \cup \ldots) = \sum_i P(E_i)$$

To znaczy: prawdopodobieństwo zdarzenie, które jest sumą rozłącznych zdarzeń, obliczamy jako sumę prawdopodobieństw tych zdarzeń.

Sumą (alternatywą) dwóch zdarzeń *A* i *B* nazywamy zdarzenie $A \cup B$ zawierające wszystkie zdarzenia elementarne należące do *A* lub *B* - zajdzie przynajmniej jedno ze zdarzeń.

Iloczynem (koniunkcją) dwóch zdarzeń *A* i *B* nazywamy zdarzenie $A \cap B$ zawierające wszystkie zdarzenia elementarne należące do *A* i do *B* - zajdą równocześnie zdarzenia *A* i *B*.

Różnicą dwóch zdarzeń *A* i *B* nazywamy zdarzenie *A* - *B*, składające się ze zdarzeń elementarnych należących do *A* i nie należących do *B* - zajdzie zdarzenie *A* i nie zajdzie *B*.

Zdarzeniem przeciwnym do *A* nazywamy zdarzenie $\overline{A}$ zawierające wszystkie zdarzenia elementarne nnienależącedo *A*, tzn. $\overline{A} = E - A$.

Przykład

Rzucamy kostką do gry: $E = \{e_1,e_2,e_3,e_4,e_5,e_6\}$.
Zdarzenie *A* polega na wyrzuceniu nieparzystej liczby oczek: $A = \{e_1,e_3,e_5\}$, a zdarzenie *B* - liczba oczek jest mniejsza od 4: $B = \{e_1,e_2,e_3\}$.

$$A \cup B = \{e_1,e_3,e_5\} \cup \{e_1,e_2,e_3\} = \{e_1,e_2,e_3,e_5\}$$

$$A \cap B = \{e_1,e_3,e_5\} \cap \{e_1,e_2,e_3\} = \{e_1,e_3\}$$

$$A - B = \{e_1,e_3,e_5\} - \{e_1,e_2,e_3\} = \{e_5\}$$

$$\overline{A} = \{e_1,e_2,e_3,e_4,e_5,e_6\} - \{e_1,e_3,e_5\} = \{e_2,e_4,e_6\}$$

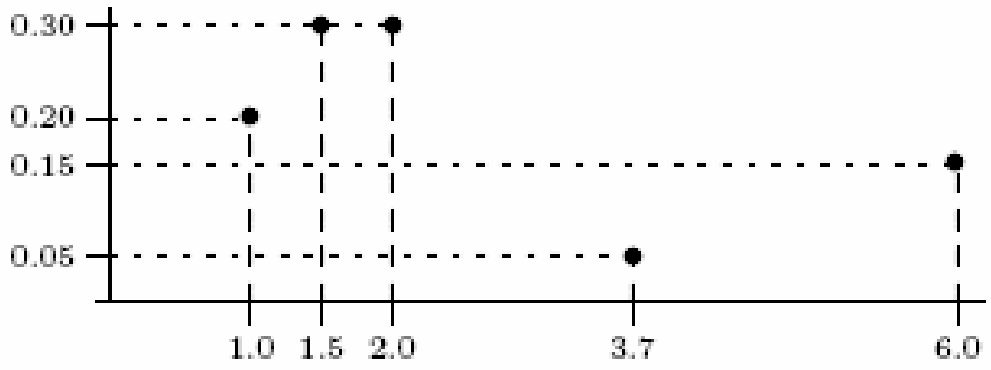
ROZKŁAD ZMIENNEJ LOSOWEJ zbiór wartości zmiennej losowej oraz prawdopodobieństwa, z jakimi są te wartości przyjmowane.

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

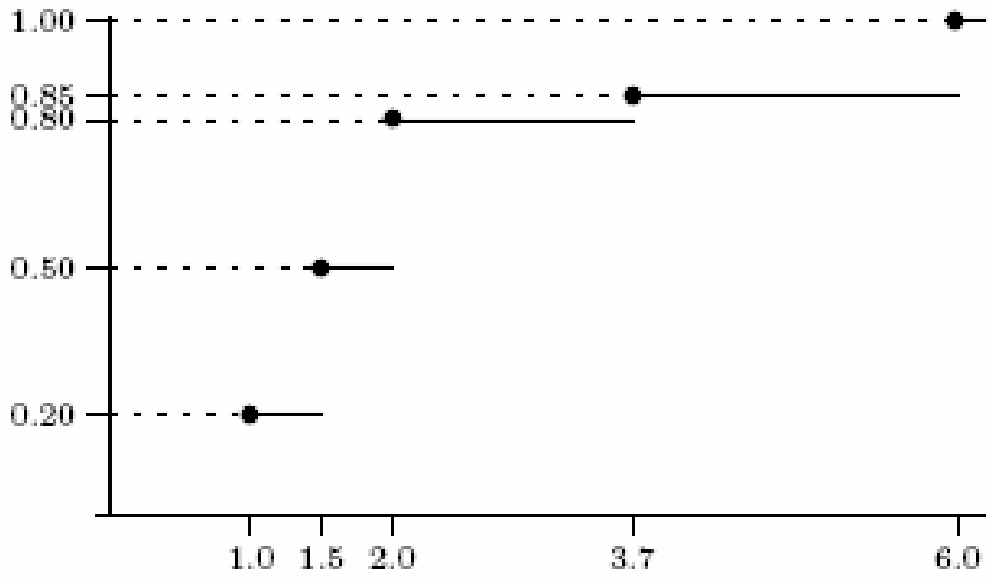
x_i	1	2	3	4	5	6
p_i	1/6	1/6	1/6	1/6	1/6	1/6

DYSTRYBUANTA to funkcja $F(x)=P(X\leq x)$
Najważniejsze własności dystrybuanty
1. $0 \leq F(x) \leq 1$
2. $F(-1) = 0$, $F(1) = 1$
3. dystrybuanta jest funkcja niemalejącą
4. $P\{a < X \leq b\} = F(b) - F(a)$

Skokowa zmienna losowa



Funkcja rozkładu prawdopodobieństwa



Dystrybuanta

Zmienna losowa ciągła

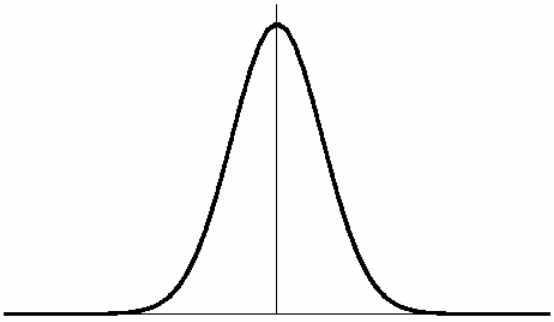
Funkcja (gęstości) rozkładu prawdopodobieństwa

f jest funkcja określona na zbiorze liczb rzeczywistych **R** wzorem

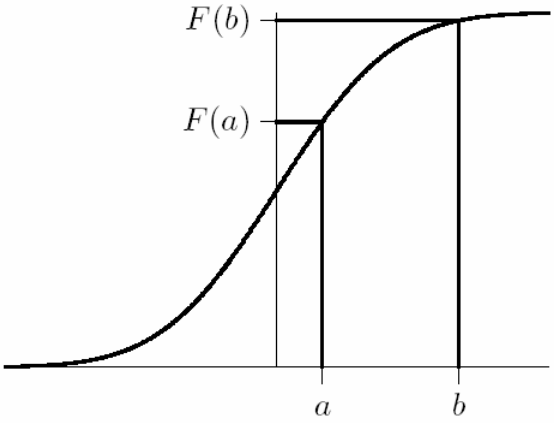
$$f(x) = \begin{cases} F'(x), & \text{jeżeli } F'(x) \text{ istnieje} \\ 0, & \text{w przeciwnym przypadku} \end{cases}$$

Najważniejsze własności funkcji gęstości

1. $f(x) \geq 0$
2. $P\{a < X \leq b\} = \int_a^b f(x)dx$



Funkcja gęstości



Dystrybuanta

Podstawowe parametry rozkładu zmiennej losowej skokowej

W rachunku prawdopodobieństwa **wartość oczekiwana** (inaczej wartość przeciętna, wartość średnia, nadzieja matematyczna) skokowej (dyskretnej) zmiennej losowej jest sumą iloczynów wartości tej zmiennej losowej oraz prawdopodobieństw, z jakimi te wartości są przyjmowane.

$$E(X) = \sum_{i=1}^{n} x_i \bullet p_i$$

Wariancja to klasyczna miara zmienności. Wyraża zróżnicowanie zbiorowości, jest średnią arytmetyczną kwadratów odchyleń poszczególnych wartości cechy od średniej arytmetycznej zbiorowości.

$$D^2(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 \bullet p_i$$

Odchylenie standardowe

$$D^2(X) = \sqrt{D^2(X)}$$

Przykładowe rozkłady zmiennych losowych skokowych

1) Rozkład dwupunktowy

Z rozkładem dwupunktowym mamy do czynienia wówczas, gdy w wyniku doświadczenia możemy uzyskać tylko jedną z dwóch wartości zmiennej losowej: x_1 lub x_2 z prawdopodobieństwami odpowiednio p oraz 1-p. W szczególnym przypadku, gdy $x_1 = 0$ oraz $x_2 = 1$ rozkład ten nazywany jest rozkładem zero-jedynkowym.

2) Rozkład dwumianowy (Bernouliego)

Rozkład dwumianowy występuje wówczas, gdy przeprowadza się n jednakowych doświadczeń, z których każde może zakończyć się jednym z dwóch wyników: „sukcesem” z prawdopodobieństwem p lub „porażką” z

prawdopodobieństwem 1-p. Zmienną losową X w tym eksperymencie jest liczba sukcesów w n próbach.

Rozkład prawdopodobieństwa w rozkładzie Bernoulliego jest określony wzorem:

$$P(X=k)=\binom{n}{k}\cdot p^{k}\cdot(1-p)^{n-k}$$

$$\mathbf{E(X)=n\cdot p \quad D^2(X)=n\cdot p\cdot(1-p)}$$

$$D(X)=\sqrt{n\bullet p\bullet(1-p)}$$

3) Rozkład Poissona

jest rozkładem zmiennej losowej skokowej, z którym mamy do czynienia w przypadku określania prawdopodobieństwa zajścia zdarzeń stosunkowo rzadkich i niezależnych od siebie, takich jak np. liczba usterek w produkowanej partii materiału. Rozkład Poissona jest przybliżeniem rozkładu Bernoulliego dla dużych

prób i przy małym prawdopodobieństwie zajścia zdarzenia („sukcesu”).

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

e - podstawa logarytmów naturalnych,

λ - stała, która jest wartością oczekiwaną i równocześnie wariancją rozkładu,

Przykładowe rozkłady zmiennych losowych ciągłych

1) Rozkład jednostajny

Jest to najprostszy z rozkładów zmiennej losowej ciągłej. Mamy z nim do czynienia wtedy, gdy prawdopodobieństwo zajścia zdarzenia jest stałe w pewnym przedziale <a, b>. Funkcja gęstości tego rozkładu jest dana wzorem:

$$f(x)=\begin{cases}\frac{1}{b-a} & \text{dla } x\in[a,b]\\ 0 \text{ dla pozostałych } x\end{cases}$$

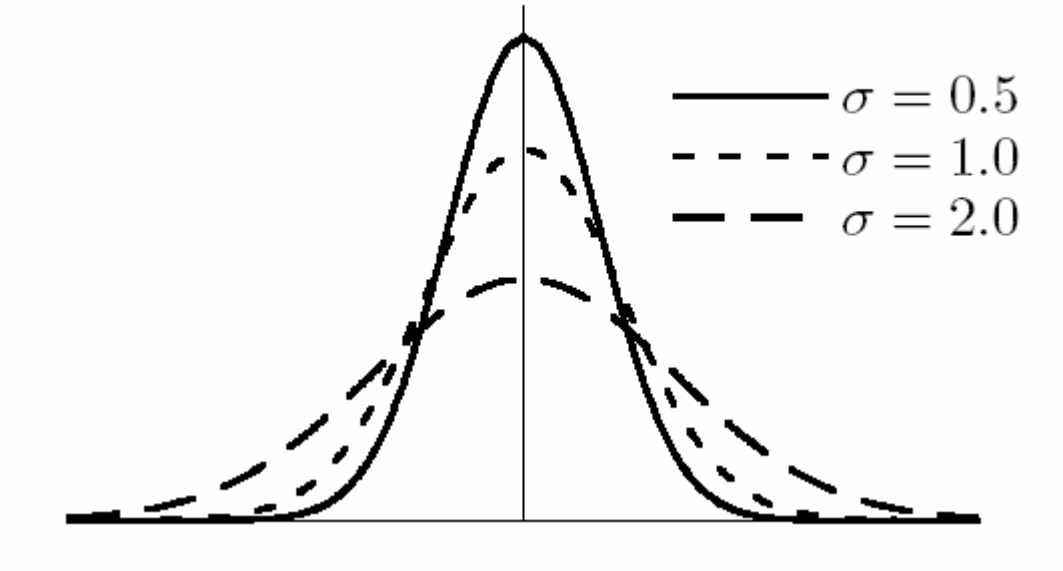
$$E(X)=\frac{a+b}{2} \qquad D^2(X)=\frac{(b-a)^2}{12}$$

$$D(X)=\frac{b-a}{2\sqrt{3}}$$

2) **Rozkład normalny**, zwany także rozkładem Gaussa-Laplace'a jest najczęściej spotykanym w naturze rozkładem zmiennej losowej ciągłej. Ciągła zmienna losowa X ma rozkład normalny o wartości oczekiwanej μ i odchyleniu standardowym σ co oznaczamy $X \sim N(\mu,\sigma^2)$, jeśli jej funkcja gęstości – określona dla wszystkich rzeczywistych wartości x – da się przedstawić za pomocą wzoru:

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

μ=E(X) , σ=D(X)



Standaryzacja

Jeżeli $X \sim N(\mu, \sigma^2)$

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$$P\{X \in (a, b)\} = P\left\{Z \in \left(\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma}\right)\right\}$$
$$= F\left(\frac{b - \mu}{\sigma}\right) - F\left(\frac{a - \mu}{\sigma}\right)$$

Prawo trzech sigm

$$P\{|X - \mu| < \sigma\} = 0.68268 \approx 0.68$$

$$P\{|X - \mu| < 2\sigma\} = 0.95450 \approx 0.95$$

$$P\{|X - \mu| < 3\sigma\} = 0.99730 \approx 0.997$$

ORGANIZACJA BADANIA STATYSTYCZNEGO

Określenie:

a) populacji
b) jednostki populacji
c) cechy populacji

Metody badania statystycznego

1) Badanie pełne (badanie obejmuje całą populację)
2) Badanie częściowe (badanie odbywa się na pewnych losowo wyodrębnionych elementach populacji, czyli **próbie losowej**)
   a) metoda reprezentacyjna
   b) metoda monograficzna
   c) metoda ankietowa

OPRACOWANIE MATERIAŁU STATYSTYCZNEGO

Charakterystyki położenia

- Średnia arytmetyczna:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- inne średnie:
średnia harmoniczna
średnia geometryczna

- Mediana

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{dla } n \text{ nieparzystych} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{dla } n \text{ parzystych} \end{cases}$$

Charakterystyki rozproszenia

- Odchylenie przeciętne

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

- Wariancja

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

- Odchylenie standardowe

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

wartości typowe: ($\bar{x}$-s,$\bar{x}$+s)

- współczynnik zmienności

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

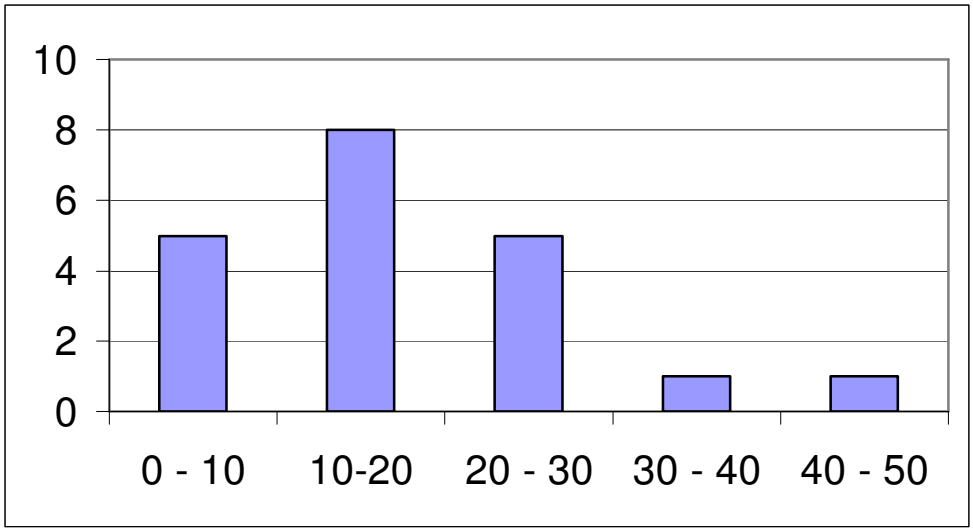
- kwartyle, decyle, centyle

Grupowanie danych

- proste (jedna cecha np. wg wieku)
- złożone (wiele cech np. wg wieku i płci)

Wartości cechy (np. wiek)	Liczebność	Częstość
0 - 10	5	0.25
10 - 20	8	0.40
20 - 30	5	0.25
30 - 40	1	0.05
40 - 50	1	0.05

Przedstawianie graficzne za pomocą **histogramu**



Estymacja - to dział wnioskowania statystycznego będący zbiorem metod pozwalających na uogólnianie wyników badania próby losowej na nieznaną postać i parametry rozkładu zmiennej losowej całej populacji oraz szacowanie błędów wynikających z tego uogólnienia.

Metody estymacji parametrycznej można w zależności od sposobu szacowania szukanego parametru podzielić na dwie grupy:
- **estymacja punktowa** (szacowanie wartości)

- **estymacja przedziałowa** (szacowanie przedziałów)

Estymatory punktowe

- estymator wariancji

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

suma kwadratów odchyleń od średniej

$$\operatorname{var} X = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

- estymator odchylenia standardowego

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

Estymatory przedziałowe

Przedział ufności dla średniej

$$\left(\bar{X} - t(\alpha; n-1)\frac{S}{\sqrt{n}} \;,\quad \bar{X} + t(\alpha; n-1)\frac{S}{\sqrt{n}} \right)$$

$t(\alpha; n-1)$: wartość krytyczna rozkładu t - Studenta
z n-1(v) stopniami swobody

Poziom ufności: $1-\alpha$ ustalone z góry prawdopodobieństwo z jakim ten przedział pokrywa nieznaną wartość parametru np. średniej

Przedział ufności dla wariancji

(Średnia μ jest nieznana)

$$\left(\frac{\operatorname{var} X}{\chi^2\left(\frac{\alpha}{2}; n-1\right)} \; , \quad \frac{\operatorname{var} X}{\chi^2\left(1-\frac{\alpha}{2}; n-1\right)} \right)$$

$\chi 2\,(\alpha;\, n-1)$ jest wartością krytyczna rozkładu chi–kwadrat z v stopniami swobody.

Przedział ufności dla odchylenia standardowego

$$\left(\sqrt{\frac{\operatorname{var} X}{\chi^2(\frac{\alpha}{2}; n-1)}} \; , \quad \sqrt{\frac{\operatorname{var} X}{\chi^2(1-\frac{\alpha}{2}; n-1)}} \right)$$

ESTYMACJA (ciąg dalszy – dwie populacje)

Przedział ufności dla różnicy średnich

Dla 2 populacji o rozkładzie normalnym

$$\{(\overline{x}_1 - \overline{x}_2) - t_{\alpha,\nu,} \cdot S_r ; (\overline{x}_1 - \overline{x}_2) + t_{\alpha,\nu,} \cdot S_r\}$$

jest to przedział w którym z prawdopodobieństwem 1-α zawiera się różnica średnich dla 2 populacji (m_1-m_2).

Zakładamy, że wariancje dla tych populacji są równe tj. $\sigma_1^2 = \sigma_2^2$

gdzie:

$$S_r = \sqrt{S_e^2 \frac{1}{n_1} + \frac{1}{n_2}}$$ - *błąd różnicy średnich*

$$S_e^2 = \frac{\operatorname{var} X_1 + \operatorname{var} X_2}{(n_1 - 1) + (n_2 - 1)}$$ - *wariancja wspólna*

varX – suma kwadratów odchyleń od średniej

$t_{\alpha,\nu}$- *wartość dla rozkładu t-studenta przy ustalonym α (najczęściej 0,05) oraz v (liczba stopni swobody, czyli n_1+n_2-2).*

ESTYMACJA – ROZKŁAD DWUPUNKTOWY

Przedział ufności dla wskaźnika struktury w rozkładzie dwupunktowym

$$\{\frac{m}{n} - z_\alpha \cdot \sqrt{\frac{\frac{m}{n} \cdot (1 - \frac{m}{n})}{n}}; \frac{m}{n} + z_\alpha \cdot \sqrt{\frac{\frac{m}{n} \cdot (1 - \frac{m}{n})}{n}}\}$$

jest to przedział ufności, w którym wskaźnik struktury w rozkładzie dwupunktowym zawiera się z prawdopodobieństwem 1-α, gdzie:

m- liczba elementów wyróżnionych znalezionych w próbie

n- liczebność próby

z_α - wartość z tablic rozkładu normalnego N(1;0) dla ustalonej wartości α

Przykłady rozkładu dwupunktowego:

1) udział nasion kiełkujących i niekiełkujących w materiale siewnym

2) udział produktów sprawnych i wadliwych w produkowanej serii

Przedział ufności dla różnicy dwóch frakcji (rozkład dwupunktowy)

$$\{(\frac{m_A}{n_A}-\frac{m_B}{n_B})-z_\alpha \cdot SP_r;(\frac{m_A}{n_A}-\frac{m_B}{n_B})+z_\alpha \cdot SP_r\}$$

W przedziale ufności z prawdopodobieństwem 1-α zawiera się wartość różnicy prawdopodobieństw dwóch rozkładów dwupunktowych (p_A-p_B).

m_A, m_B – liczby elementów wyróżnionych w próbach

n_A, n_B – liczebności prób

$$SP_r=\sqrt{\overline{p}\cdot(1-\overline{p})}\cdot(\frac{1}{n_A}+\frac{1}{n_B})$$

$$\overline{p}=\frac{m_A+m_B}{n_A+n_B}$$

Powyższe wzory można zastosować tylko dla prób o dużej liczebności > 100 elementów

HIPOTEZY STATYSTYCZNE I ICH WERYFIKACJA

Weryfikacja (testowanie) hipotez statystycznych, czyli sprawdzenie określonych przypuszczeń (założeń) wysuniętych w stosunku do parametrów lub rozkładu populacji generalnej na podstawie próby.

Podział hipotez:

Hipotezy statystyczne – dotyczące rozkładu populacji

Hipotezy parametryczne – dotyczące parametrów rozkładu (który jest znany)

Test statystyczny – reguła postępowania, która pozwala na przyjęcie (nieodrzucenie) bądź odrzucenie sprawdzanej hipotezy

Błąd I rodzaju– błąd odrzucenia, występuje gdy odrzucamy hipotezę, natomiast jest ona prawdziwa

Błąd II rodzaju – błąd przyjęcia, występuje gdy przyjmujemy hipotezę, natomiast jest ona fałszywa

Prawdopodobieństwo popełnienia błędu I rodzaju nazywamy **poziomem istotności (α)**

Hipotezy dla cech mających rozkład normalny

1) Porównanie średniej z normą

Ho: $\mu = \mu_0$

Funkcja testowa $t_{emp} = \frac{\overline{x} - \mu_0}{S_{\overline{x}}}$

Gdzie S_x – błąd standardowy $S_{\overline{x}} = \frac{s}{\sqrt{n}}$

Wartość krytyczna $\mathbf{t_{\alpha,\nu}}$, dla rozkładu t-studenta, gdzie α jest przyjętym poziomem istotności (najczęściej 0,05), a ν liczbą stopni swobody, czyli liczebność próby pomniejszona o 1 (n-1)

Jeżeli $\mathbf{|t_{emp}| > t_{\alpha,\nu}}$ to hipotezę **H_0 odrzucamy** i przyjmujemy **hipotezę alternatywną H_1: $\mu \neq \mu_0$**

2) Porównanie średnich 2 populacji

Ho: $\mu_1 = \mu_2$

założenie $\sigma_1^2 = \sigma_2^2$

Funkcja testowa $t_{emp} = \frac{\overline{x} - \overline{y}}{S_r}$

Gdzie S_r – błąd różnicy średnich

Wartość krytyczna $\mathbf{t_{\alpha,\nu}}$, dla rozkładu t-studenta, gdzie α jest przyjętym poziomem istotności (najczęściej 0,05), a ν liczbą stopni swobody, czyli liczebność 2 prób pomniejszona o 2 ($n_1 + n_2 - 2$)

Jeżeli $|t_{emp}| > t_{\alpha,\nu}$ to hipotezę H_0 odrzucamy i przyjmujemy **hipotezę alternatywną H_1: $\mu_1 \neq \mu_2$**

3) Porównanie wariancji 2 populacji

Ho: $\sigma_1^2 = \sigma_2^2$

Funkcja testowa $F_{emp} = \frac{s_1^2}{s_2^2}$

Wartość krytyczna $\mathbf{F_{\alpha,\nu,u}}$ dla rozkładu F-Fishera, gdzie α jest przyjętym poziomem istotności (najczęściej 0,05), a ν i u liczbami stopni swobody, czyli liczebnością próby pierwszej (n_1-1) i drugiej (n_2 -1)

Wartość $s_1^2 > s_2^2$

Jeżeli $\mathbf{F_{emp} > F_{\alpha,\nu,u}}$ to **H_0 odrzucamy**

Porównanie średnich w wielu populacjach o rozkładzie normalnym – Analiza wariancji (ANOVA)

Założenia:

$X_i \sim N(\mu,\sigma^2)$

$\sigma_1 = \sigma_2 = \sigma_3 = ... = \sigma_i$

model analizy wariancji:

$$\mathbf{y_{ij} = \mu + a_i + e_{ij}}$$

gdzie:

y_{ij} – wielkość cechy
μ – średnia ogólna
a_i – efekt *i*-tego poziomu czynnika
e_{ij} – błędy losowe, o rozkładzie $N(0, \sigma_e)$

Hipoteza: $a_1 = a_2 = a_3 = ... = a_i$
Tabela analizy wariancji:

Źródło zmienności	Stopnie swobody	Sumy kwadratów	Średnie kwadraty	F_{emp}
Czynnik	$k-1$	$\text{var}A$	$S_a^2 = \frac{\text{var}A}{k-1}$	S_a^2/S_e^2
Błąd losowy	$N-k$	$\text{var}E$	$S_e^2 = \frac{varE}{N-k}$	
Ogółem	$N-1$	$\text{var}T$		

$$\text{var}A = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2, \text{var}E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2,$$

$$\text{var}T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2,$$

$$\text{var}A + \text{var}E = \text{var}T$$

Funkcja testowa $\mathbf{F_{emp}}$

Wartość krytyczna $\mathbf{F_{\alpha,k-1,n-k}}$

α – poziom istotności (najczęściej przyjmujemy 0,05)

k – liczba poziomów czynnika

n – liczebność prób

Jeżeli $\mathbf{F_{emp} > F_{\alpha,k-1,n-k}}$ to **H_0 odrzucamy**

Porównania wielokrotne (szczegółowe)

Grupy jednorodne — podzbiory średnich, które można uznać za takie same

Procedury porównań wielokrotnych — postępowanie statystyczne zmierzające do podzielenia zbioru średnich na grupy jednorodne

Procedury: Tukeya, Scheff´ego, Bonfferroniego, Duncana, Newmana–Kuelsa i inne.

NIR — najmniejsza istotna różnica

$$\text{Jeżeli } |\bar{X}_i - \bar{X}_j| < NIR, \text{ to uznajemy, że } \mu_i = \mu_j.$$

Procedura Tukeya

$$NIR = t_{\alpha,k,n-k} \cdot S_e \sqrt{\frac{1}{n}}$$

$t_{\alpha,k,n-k}$ – wartość krytyczna studentyzowanego rozstępu

WSPÓŁCZYNNIK KORELACJI

Współczynnik korelacji liniowej Pearsona

(oznaczany najczęściej symbolem - *r*) określa poziom zależności liniowej między zmiennymi losowymi.

$$r = \frac{\operatorname{cov}(X,Y)}{s_x \cdot s_y}$$

gdzie, wartość kowariancji (cov) na podstawie próby liczymy wg następującego wzoru:

$$\operatorname{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

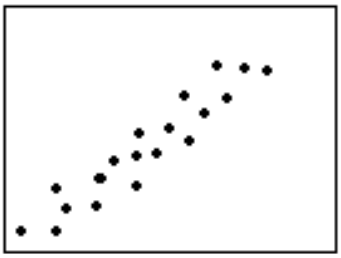
natomiast s_x i s_y są odchyleniami standardowymi dla zmiennej X i Y

Współczynnik korelacji liniowej dwóch zmiennych jest, zatem ilorazem kowariancji i iloczynu odchyleń standardowych.

Współczynnik korelacji liniowej przyjmuje zawsze wartości w zakresie [-1,1].

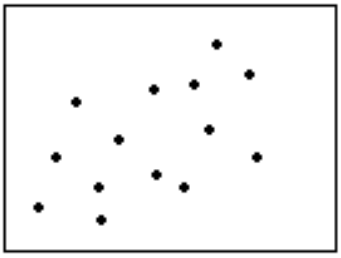
Im większa wartość bezwzględna współczynnika, tym większa jest zależność liniowa między zmiennymi. $r_{xy} = 0$ oznacza brak korelacji, $r_{xy} = 1$ oznacza silną korelację dodatnią, jeżeli jedna zmienna (x) rośnie to również rośnie druga zmienna (y), natomiast $r_{xy} = -1$ oznacza korelację ujemną (jeżeli zmienna x rośnie, to y maleje i na odwrót).

Stopień korelacji

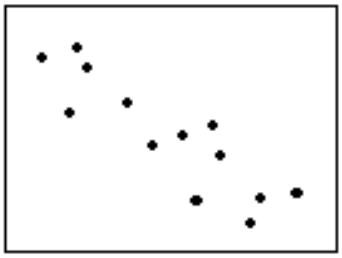


silna dodatnia (r = 0,8)

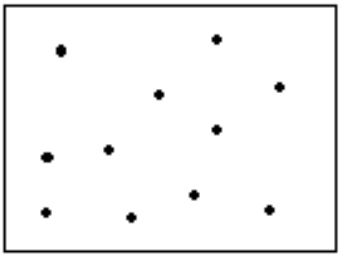
silna ujemna (r = -0,8)



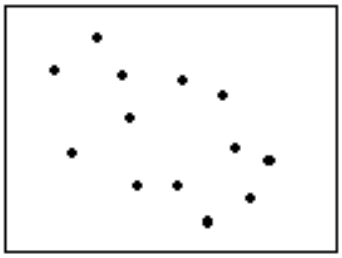
słaba dodatnia (r = 0,3)



umiarkowana ujemna (r = -0,5)



brak korelacji (r = 0,0)



słaba ujemna (r = -0,3)

Testowanie istotności korelacji

Hipoteza zerowa: $H_0:\rho=0$

ρ- wartość współczynnika korelacji dla całej populacji

Jeżeli $|r_{emp}|>r_{\alpha,2,n-2}$ to H_0 odrzucamy.

$r_{\alpha,2,n-2}$ – jest wartością krytyczną współczynnika korelacji prostej Pearsona

Regresja prosta liniowa

Regresja liniowa to metoda estymowania wartości oczekiwanej jednej zmiennej (Y) znając wartości innej zmiennej (X). Szukana zmienna, Y, jest nazywana zmienną zależną, zmienna X nazywa się zmienną niezależną.

Model regresji prostej liniowej:

$$y_i = a + bx_i + e_i$$

gdzie:

b- współczynnik regresji

a – stała regresji

e_i – błędy losowe o rozkładzie $N(0;\sigma_e^2)$

Estymację współczynników równania regresji prowadzi się zwykle metodą najmniejszych kwadratów, która polega na minimalizacji następującej sumy kwadratów:

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Estymatory wartości współczynników a i b oblicza się ze wzorów:

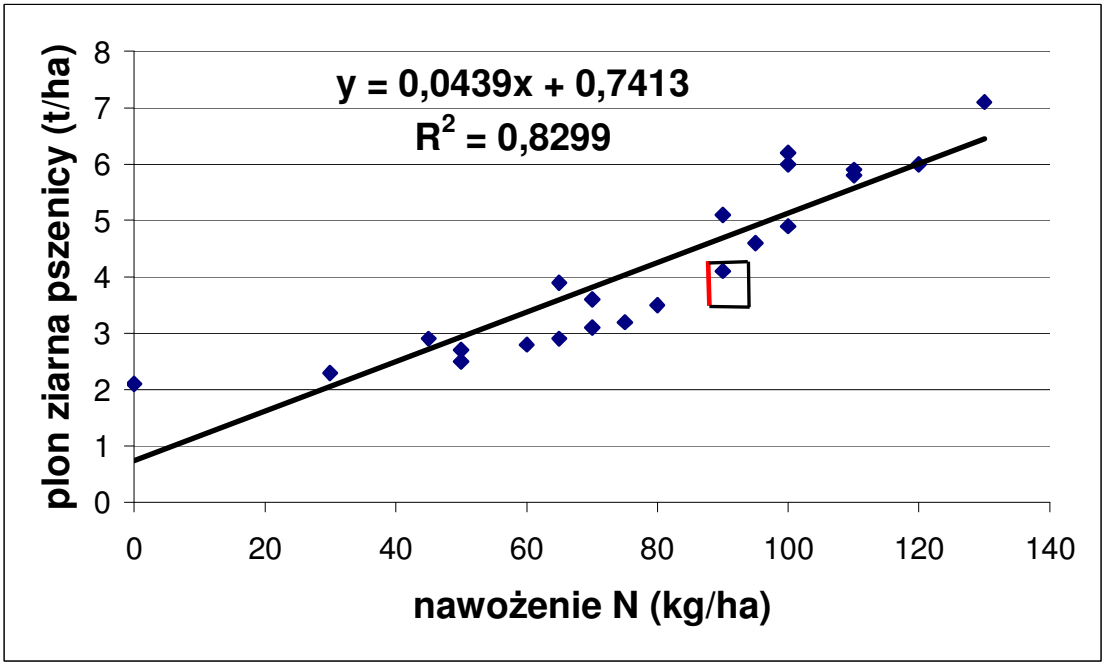
$$b = \frac{\operatorname{cov}(X,Y)}{s_x^2} \qquad a = \bar{y} - b\bar{x}$$

Przedział ufności dla współczynnika regresji:

$$(b - t_{\alpha;n-2} \cdot S_b;\ b + t_{\alpha;n-2} \cdot S_b)$$

gdzie wariancja estymatora b $S_b = \frac{S^2}{\text{var } X}$

Testowanie hipotezy H_0: b=0 jest równoważne z testowaniem hipotezy o istotności korelacji



R^2 – współczynnik determinacji , który określa stosunek zmienności wyjaśnianej przez model regresji do zmienności całkowitej. W przypadku regresji prostej liniowej $R^2=r_{xy}^2$

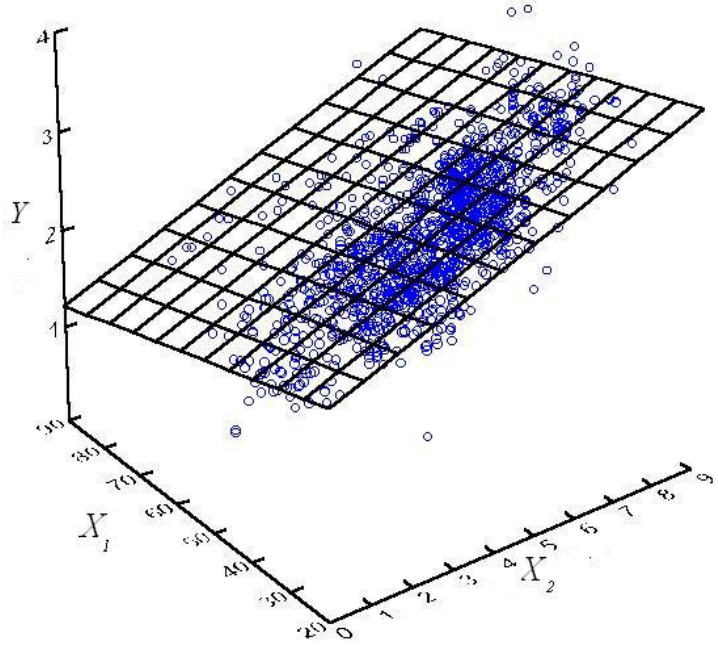
Regresja wielokrotna liniowa

Jeżeli zmienna zależna (Y) jest determinowana przez więcej niż jedną zmienną niezależną (X_i) to estymowany model regresji możemy zapisać równaniem:

$$\mathbf{Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_k \cdot X_k}$$

W przypadku regresji wielokrotnej zastosowanie metody najmniejszych kwadratów to minimalizowanie sumy:

$$\sum_{i=1}^{n} (y_i - a - b_1 x_{i1} - b_2 x_{i2} - ... - b_k x_{ik})^2$$



Graficzne przedstawienie regresji z 2 zmiennymi niezależnymi (X_1, X_2)

Test niezależności cech jakościowych - Test χ^2

Rozważając liczbę obserwacji sklasyfikowanych wg dwóch kryteriów, np. ludzi wg koloru oczu i koloru włosów (kolory oczu: brązowy, niebieski; kolory włosów: blondyni, szatyni, bruneci) lub np. rośliny pszenicy wg odmiany i stopnia porażenia chorobą (odmiany: Olimpia, Eta, Kontesa; stopień porażenia: brak, słaby, średni, duży, bardzo duży) w każdej z klas liczymy liczbę osobników i przedstawiamy w postaci tablicy dwudzielnej zwanej tablica kontyngencji

Tablica kontyngencji

Klasy cechy *Y*	**Klasy cechy *X***						
	A_1	A_2	A_3	A_4	…	A_m	**razem**
B_1	n_{11}	n_{21}	n_{31}	n_{41}	…	n_{m1}	Σn_{i1}
B_2	n_{12}	n_{22}	n_{32}	n_{42}	…	n_{m2}	Σn_{i2}
B_3	n_{13}	n_{23}	n_{33}	n_{43}	…	n_{m3}	Σn_{i3}
…					…		
B_k	n_{1k}	n_{2k}	n_{3k}	n_{4k}	…	n_{mk}	
razem	Σn_{1j}	Σn_{2j}	Σn_{3j}	Σn_{4j}	…		Σn_{ij}

n- liczebności osobników zaliczonych do określonej klasy

H_o: Cechy X i Y są niezależne

Statystyka testowa

$$\chi^2_{\text{emp}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - n^t_{ij})^2}{n^t_{ij}}$$

$$n^t_{ij} = \frac{n_{i.}n_{.j}}{N},\ N = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}$$

$$n_{i.} = \sum_{j=1}^{m} n_{ij},\ n_{.j} = \sum_{i=1}^{k} n_{ij}$$

Jeżeli $\chi^2_{\text{emp}} > \chi^2(\alpha; (k-1)(m-1))$,
to hipotezę H_0 odrzucamy