Statystyka opisowa

Definicje statystyki i rachunku prawdopodobieństwa

Statystyka – nauka traktująca o metodach ilościowych wykorzystywanych w celu poszukiwania prawidłowości w pozornie chaotycznych zjawiskach masowych.

Rachunek prawdopodobieństwa – dział matematyki zajmujący się zdarzeniami losowymi i badaniem abstrakcyjnych pojęć matematycznych stworzonych do opisu zjawisk, które nie są deterministyczne: zmiennych losowych w przypadku pojedynczych zdarzeń oraz procesów stochastycznych w przypadku zdarzeń powtarzających się (w czasie).

Zbiorowość statystyczna (populacja/masa statystyczna) – zbiór dowolnych elementów objęty badaniem statystycznym.

Jednostka statystyczna (jednostka badania lub obserwacji) – element składowy badanej zbiorowość.

Cecha statystyczna – właściwość jednostki statystycznej:

- cecha stała i zmienna;
- cecha jakościowa (kolor oczu, płeć) i ilościowa (mierzalna; kg, cm, czyli waga, wzrost, dochód);
- cecha skokowa (dyskretna) i ciągła.

Badanie statystyczne – zespół czynności zmierzających do określenia prawidłowości w badanej zbiorowości.

Rodzaje:

- ciągłe (cały czas przeprowadzane), cykliczne (okresowe), doraźne (na potrzebach chwili);
- pełne (spis powszechny), częściowe (badanie ankietowe, badanie reprezentacyjne), szacowanie(interpolacja (jeśli w szeregu czasowym brakuje danych), ekstrapolacja(w przyszłość)).

Rejestracja bieżąca, badanie monograficzne

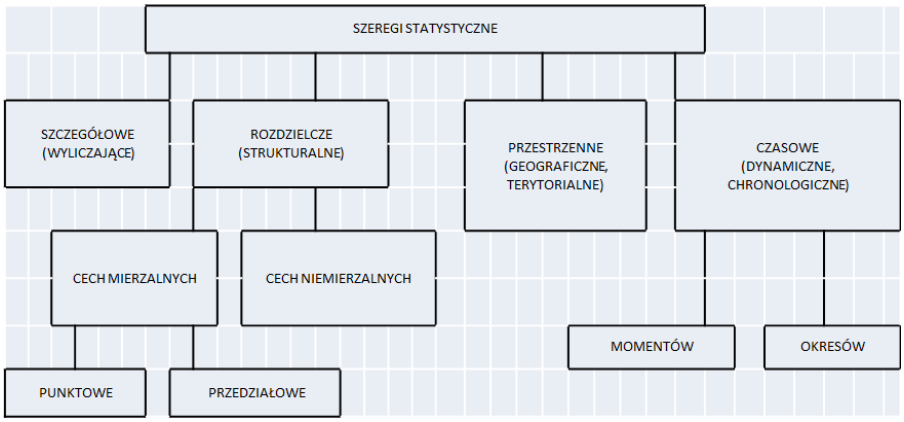
Etapy badania statystycznego:

1. Przygotowanie: cel, jednostka, przedmiot i metoda badania (czyli rodzaj);
2. Obserwacja statystyczna: materiał statystyczny – pierwotny(ankiety) i wtórny (dane wzięte z Internetu i innych źródł);
3. Kontrola materiału statystycznego: formalna, merytoryczna (logiczna, arytmetyczna). Błędy losowe i systematyczne;
4. Przetwarzanie i prezentacja materiału statystycznego: tabele, wykresy, szeregi;
5. Analiza statystyczna.

Objaśnienia znaków umownych

Kreska	(—)	— zjawisko nie wystąpiło.
Zero	(0,0)	— zjawisko istniało, jednakże w ilościach mniejszych od liczb, które mogły być wyrażone uwidocznionymi w tablicy znakami cyfrowymi.
Kropka	(.)	— zupełny brak informacji albo brak informacji wiarygodnych.
Znak	x	— wypełnienie pozycji, ze względu na układ tablicy, jest niemożliwe lub niecelowe.
„W tym”		— oznacza, że nie podaje się wszystkich składników sumy.
Znak	#	— oznacza, że dane nie mogą być opublikowane ze względu na konieczność zachowania tajemnicy statystycznej w rozumieniu ustawy o statystyce publicznej.

Szereg statystyczny – zbiór wyników obserwacji jednostek pod względem pewnej cechy.



Szereg szczegółowy:

Liczba dzieci	0	1	2	3	4	5
Liczba rodzin	3	5	5	4	2	1

Szereg strukturalny:

Oceny x_i	Liczba studentów studiów zaocznych n_i	Liczba studentów studiów dziennych n_i
2	600	100
3	1200	300
4	900	400
5	300	200
Ogółem	3000	1000

Oceny x_i	Liczba studentów studiów zaocznych n_i	Liczba studentów studiów dziennych n_i
niedostateczny	600	100
dostateczny	1200	300
dobry	900	400
bardzo dobry	300	200
Ogółem	3000	1000

Zużycie energii	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12	12 – 14
Liczba rodzin	6	10	30	40	10	4

Szereg geograficzny:

Kraj	Przeciętna miesięczna pensja brutto (w USD)
Słowenia	935
Chorwacja	620
Polska	380
Czechy	370
Węgry	322
Estonia	305
Litwa	265
Słowacja	250
Łotwa	240
Rumunia	115
Bułgaria	110
Rosja	60

Szereg czasowy:

1.01. danego roku	1950	1951	1952
liczba jednostek chorych	100	120	200

lata	<1950-1955)	<1955-1960)	<1960-1965)
liczba nowych zachorowań	80	40	60

Pomiar – proces empiryczny, w którym przyporządkowuje się liczby poszczególnym kategoriom cechy w taki sposób, aby relacje między liczbami odzwierciedlały relacje między kategoriami cechy.

Rodzaje skal pomiarowych:

1. Nominalna
2. Porządkowa (rangowa)
3. Przedziałowa
4. Ilorazowa
5. Absolutna

Analiza statystyczna:

1. Badanie struktury zjawisk i procesów
2. Badanie zależności zjawisk i procesów
3. Badanie dynamiki zjawisk.

Charakterystyki rozkładu jednej cechy

Rozkład Empiryczny

Rozkład empiryczny (cechy) **zmiennej** – przyporządkowanie kolejnym wartościom lub wariantom (cechy) zmiennej x_i odpowiadających im liczb lub częstości w_i jednostek posiadających daną wartość lub wariant x_i

Charakterystyki rozkładu:

1. Miary położenia rozkładu
2. Miary zmienności
3. Miary asymetrii
4. Miary koncentracji

Miary położenia rozkładu:

- Przeciętne:
  - średnie: arytmetyczna, harmoniczna, geometryczna, potęgowa
  - przeciętne pozycyjne: modalna, mediana
- Kwantyle:
  - kwartyle, kwintyle, decyle, centyle

Średnia arytmetyczna:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i\, x_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i\, x_i;$$

Średnia geometryczna:

$$\bar{i} = \sqrt[n-1]{i_{2/1} \cdot i_{3/2} \cdot i_{4/3} \cdot \ldots \cdot i_{n/n-1}} = \sqrt[n-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \ldots} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

Mediana

$$M_e = \begin{cases} x_{\frac{n+1}{2}} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}; \end{cases} \qquad M_e = x_{me} + \frac{h_{me}}{f_{me}}\left[\frac{n}{2} - \sum_{i=1}^{me-1} f_i\right]$$

Modalna

$$M_o = x_m + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} h$$

Kwantyl rzędu p

$$q_p = x_p + \frac{h_p}{f_p}\left[pn - \sum_{i=1}^{q-1} f_i\right]$$

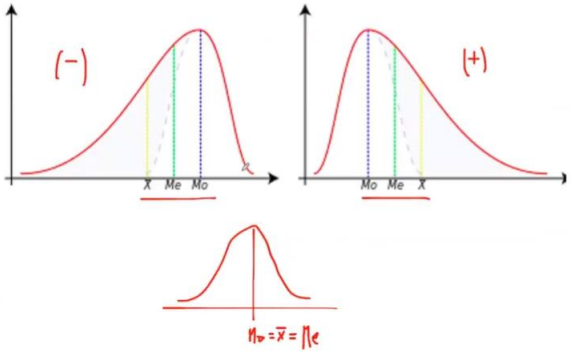
Kwartyl – 3 *Kwintyl* – 4 *Decyl* – 9 *Centyl* – 99

Miary zmienności:

- bezwzględne:
  - rozstęp: $R = \max_i\{x_i\} - \min_i\{x_i\} \qquad R_Q = Q_3 - Q_1$
  - 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: $s = \sqrt{s^2}$
- względne

- współczynnik zbieżności: $V_S = \frac{s}{|\bar{x}|} \cdot 100\%$ $V_D = \frac{D}{|\bar{x}|} \cdot 100\%$ $V_Q = \frac{Q}{Me}$, gdzie:
  $Q = \frac{Q_3 - Q_1}{2}$; Uwaga na moduł w mianowniku.
-

Asymetria rozkładu

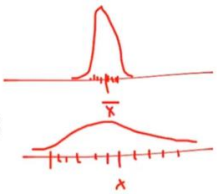


$A_S = \frac{\bar{x} - M_o}{S}$; $A_S = \frac{3(\bar{x} - M_e)}{S}$;

$A_S = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - Q_1)}$

Miary koncentracji:

Kurtoza/Eksces: $K = \frac{M_4}{s^4} - 3$



gdzie: $M_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$

$M_2 = \frac{1}{n}\sum |x_i - \bar{y}|^2 = S^2$

Współczynnik koncentracji Lorenza – współczynnik Giniego 1912 1905

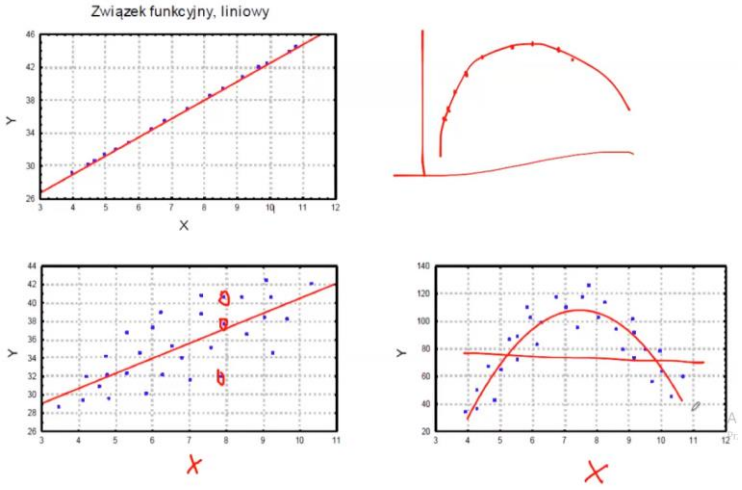
W przypadku symetrii Q_2 znajduje się w środku

Badanie zależności

Zależność statystyczna występuje wówczas, gdy istnieje logiczny związek między dwiema lub więcej cechami w badanej zbiorowości potwierdzony danymi statystycznymi.

Rodzaje:

- Zależność funkcyjna (zmiany wartości jednej cechy są ściśle determinowane zmianami wartości drugiej; w badaniach społeczno-ekonomicznych występują)
- Zależność stochastyczna
- Zależność korelacyjna (jak zmieniają się średnie wartości cechy y pod wpływem cechy x)



Skala nominalna - pomiar zależności dla zmiennych mierzonych na słabych skalach.

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Tablica kontyngencji

Zmienna X		Zmienna Y				suma
		y_1	y_2	…	y_s	
Kategorie zmiennej X	x_1	n_{11}	n_{12}	…	n_{1s}	$n_{1.}$
	x_2	n_{21}	n_{22}	…	n_{2s}	$n_{2.}$
	…	…	…	…	…	…
	x_k	n_{k1}	n_{k2}	…	n_{ks}	$n_{k.}$
Suma		$n_{.1}$	$n_{.2}$	…	$n_{.s}$	n

$$\hat{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{s} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$ Yule'a

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1; k-1)}}$$ V - Cramera

Przykład

Wykształcenie (X)	Rodzaj programu (Y)				Ogółem
	film	teatr	programy rozrywkowe	programy publicystyczne	
Podstawowe	105	10	75	10	200
Średnie	120	60	80	40	300
Wyższe	35	30	15	20	100
Ogółem	260	100	170	70	600

$\chi^2 = 62{,}04$ V = 0,2274 φ = 0, 3216

Skala porządkowa

Współczynniki Spearmana i tau Kendalla

Współczynnik Spearmana

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Współczynnik tau Kendalla

$$K_{tq} = \frac{2R}{1/2 * n(n-1)} - 1$$

gdzie:

R – suma not +1
n – liczba ocenianych obiektów

Przykład

Zbadano 5 uczelni ekonomicznych w Polsce ze względu na orientację na studenta oraz selektywność. Wyniki w punktach przedstawia tabela:

Uczelnia	Liczba punktów	
	Orientacja na studenta (X)	Selektywność (Y)
SGH	84	67
UEP	66	62
UEK	76	58
UEW	61	61
UEKat	55	48

Wyznaczyć wartość współczynnika korelacji Spearmana i tau Kendalla.

r Spearmana:

W pierwszym kroku nadajemy rangi uczelniom osobno ze względu na jedną i drugą cechę

X: 1; 3; 2; 4; 5

Y: 1; 2; 4; 3; 5

Później obliczamy różnice między rangami d_i i podnosimy je do kwadratu, następnie sumujemy i podstawiamy do wzoru:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 * 6}{5(25-1)} = 1 - \frac{36}{120} = 0,7$$

tau-Kendalla:

W pierwszym kroku nadajemy rangi uczelniom osobno ze względu na jedną i drugą cechę:

X: 1; 3; 2; 4; 5

Y: 1; 2; 4; 3; 5

Następnie porządkujemy rangi ze względu na jedną z cech:

X: 1; 2; 3; 4; 5
Y: 1; 4; 2; 3; 5

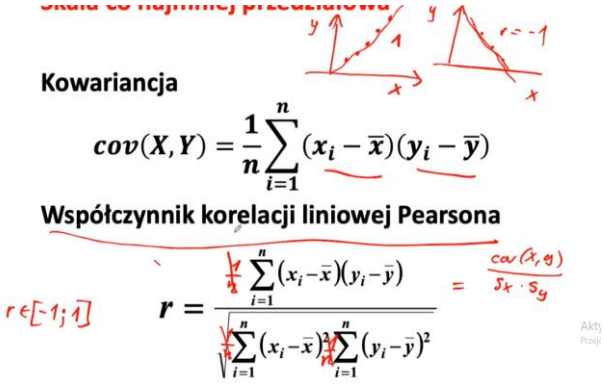
Po tym dla każdej rang cechy Y tworzymy pary z następującymi po niej rangami:

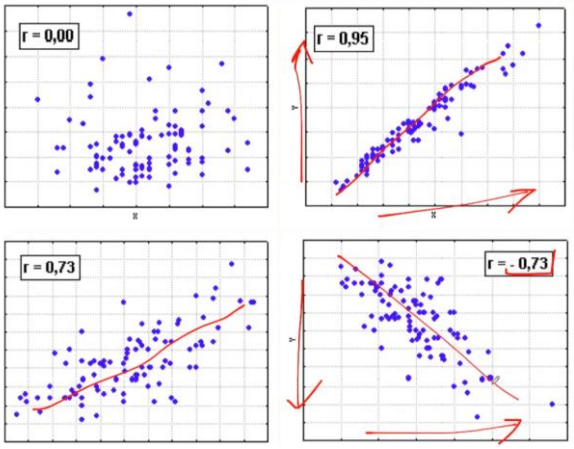
1 (1;4), (1;2), (1;3), (1;5)
4 (4;2), (4;3), (4;5)
2 (2;3), (2;5)
3 (3;5)

Jeśli poprzednik jest mniejszy od następnika to nadajemy notę 1 (**pary oznaczone na zielono**) w przeciwnym przypadku notę -1 (dla rang powiązanych mogłoby zaistnieć zero). Zliczamy jedynki, jest ich 8

$$K_{tq} = \frac{2*8}{1/2*5*(5-1)} - 1 = 0{,}6$$

Skala co najmniej przedziałowa





Funkcje regresji

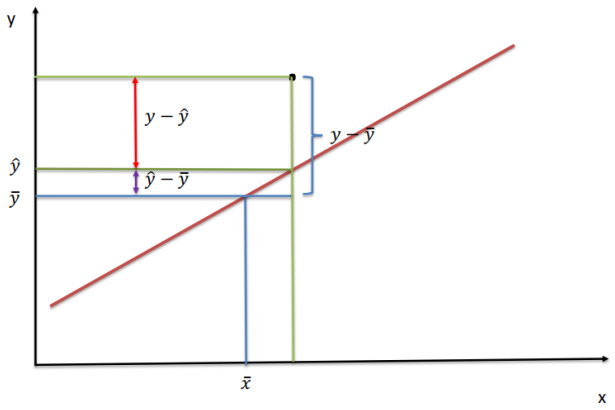
Funkcja regresji – analityczny wyraz przyporządkowania średnich wartości zmiennej objaśnianej (zależnej) konkretnym wartościom zmiennych objaśniających (niezależnych).

Funkcja regresji *I rodzaju* – to funkcja, która realizacjom zmiennych objaśniających przypisuje średnie warunkowe zmiennej objaśnianej.

Dla jednej zmiennej objaśniającej:

$E(Y|X = x_i) = g(x_1)$, gdzie E – wartość oczekiwana, Y – objaśniana, X – objaśniająca

Funkcja regresji *II rodzaju* – $\hat{y} = a_0 + a_1 x + \varepsilon$



$$\varphi^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$\varphi^2 + R^2 = 1 \qquad s_\varepsilon = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

MNK – Metoda Najmniejszych Kwadratów

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i \\ a_0\sum_{i=1}^{n}x_i + a_1\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_iy_i \end{cases} \rightarrow \begin{cases} a_1 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ a_0 = \bar{y} - a_1\bar{x} \end{cases} \rightarrow \begin{cases} a_1 = r\dfrac{s_y}{s_x} \\ a_0 = \bar{y} - a_1\bar{x} \end{cases}$$

Zależność wielu cech

$$R = \begin{bmatrix} 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{bmatrix}$$

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \qquad r_{yz.x} = \frac{r_{yz} - r_{zx}}{\sqrt{(1-r_{yx}^2)(1-r_{zx}^2)}} \qquad r_{xz.y} = \frac{r_{xz} - r_{xy}r_{zy}}{\sqrt{(1-r_{xy}^2)(1-r_{zy}^2)}}$$

$$R_{y.xz} = \sqrt{\frac{r_{xy}^2 + r_{yz}^2 - 2r_{xy}r_{xz}r_{yz}}{1-r_{xz}^2}}$$

$$\hat{y} = a_{y0} + a_{yx}x + a_{yz}z$$

$$\begin{cases} a_{yx} = \frac{r_{yx}-r_{yz}r_{xz}}{1-r_{xz}^2} \cdot \frac{s_y}{s_x} \\ a_{yz} = \frac{r_{yz}-r_{yx}r_{zx}}{1-r_{zx}^2} \cdot \frac{s_y}{s_z} \\ a_{yo} = \bar{y} - a_{yx}\bar{x} - a_{yz}\bar{z} \end{cases}$$

Rachunek prawdopodobieństwa

Zdarzenie elementarne – wyniki doświadczenia losowego.

Własności zdarzeń elementarnych:

1. Dane zdarzenie elementarne może zaistnieć lub nie
2. Jedno ze zdarzeń elementarnych na pewno zaistnieje
3. Zaistnienie jednego zdarzenia elementarnego wyklucza zaistnienie innego.

Przestrzeń zdarzeń elementarnych – *oznaczenie E lub Ω.*

Rodzaje zdarzeń elementarnych:

1. Skończona
2. Nieskończona, ale przeliczalna
3. Nieskończona i nieprzeliczalna
   - zbiór wszystkich podzbiorów zbioru E jest mocy $2n$.
   - zbiór wszystkich podzbiorów zbioru E jest mocy $2^{\aleph 0}$, czyli continuum.
   - zbiór wszystkich podzbiorów zbioru E jest mocy $2^c > c$.

W sytuacji 3.3 należy ograniczyć się do rozważania klasy zbiorów borelowskich (**σ** ciała – przeliczalnie addytywnym ciałem zbiorów). Jest to niepusta klasa Z podzbiorów przestrzeni zdarzeń elementarnych E spełniająca *trzy warunki*:

1. $E \in Z$
2. $A \in Z \Rightarrow \bar{A} \in Z$
3. $A1 \in Z, A2 \in Z, \ldots \Rightarrow (A1 \cup A2 \cup \ldots) \in Z$

Wybiera się więc spośród wielu możliwych ciał podzbiorów E ciało najmniejsze, które nazywamy **σ** ciałem. **σ** ciało istnieje i nie zawiera zbiorów niemierzalnych.

Zdarzenie losowe – każdy element **σ** ciała podzbiorów przestrzeni zdarzeń elementarnych.

- E – zdarzenie pewne;
- $\varnothing$ - zdarzenie niemożliwe;
- $\bar{A}$ – zdarzenie przeciwne do zdarzenia A.

Zdarzenia A i B są parami rozłączne, gdy $A \cap B = \varnothing$. Jeśli $A \subset B$ to zdarzenie A pociąga za sobą zdarzenie B. Suma zdarzeń $A \cup B$. Równość zdarzeń $A = B$.

Definicji prawdopodobieństwa:

- **Laplace'a:**
  Jeśli zdarzenie E rozkłada się na n wykluczających się wzajemnie i jednakowo możliwych zdarzeń elementarnych, spośród których m sprzyja zaistnieniu interesującego nas zdarzenia A, to prawdopodobieństwo zaistnienia zdarzenia A nazywamy ułamek: $P(A) = \frac{m}{n}$.
- **Geometryczna:**
  Jeśli Q i q są to dwa zbiory w przestrzeni r wymiarowej oraz jeśli $q \subset Q$ to prawdopodobieństwo, że dowolny punkt należący do Q będzie również należał do q równa się stosunkowi miary zbioru q do miary zbioru Q.

- **Statystyczna:**
  Jeśli przy wielokrotnej realizacji doświadczenia, w wyniku którego może wystąpić zdarzenie A, częstość tego zdarzenia przejawia wyraźną prawidłowość oscylując wokół pewnej nieznanej liczby p i jeśli wahania częstości przejawiają tendencję malejącą w miarę wzrostu liczby doświadczeń to liczba p nazywa się prawdopodobieństwem zdarzenia A.
- **Aksjomatyczna** (Kołmogorow, 1931)**:**
  Niech E będzie przestrzenią zdarzeń elementarnych doświadczenia losowego D, Z – jego zbiorem zdarzeń losowych. Prawdopodobieństwem nazywamy funkcję P przyporządkowującą każdemu zdarzeniu A ∈ Z liczbę P(A) zgodnie z warunkami:
  1. $P(A) \geq 0$
  2. $P(\mathrm{E}) = 1$
  3. Jeśli $\mathrm{A}_1, \mathrm{A}_2, \ldots$ jest dowolnym ciągiem parami rozłącznych zdarzeń ze zbioru Z to $\mathrm{P}(\mathrm{A}_1 \cup \mathrm{A}_2 \cup \ldots) = \mathrm{P}(\mathrm{A}_1) + \mathrm{P}(\mathrm{A}_2) + \ldots$

Własności prawdopodobieństwa:

1. $\mathrm{P}(\varnothing) = 0$;
2. Jeśli $\mathrm{A} \subset \mathrm{B}$ to $\mathrm{P}(\mathrm{A}) \leq \mathrm{P}(\mathrm{B})$;
3. $\mathrm{P}(\mathrm{A}) \leq 1$;
4. $\mathrm{P}(\mathrm{A}) + \mathrm{P}(\bar{\mathrm{A}}) = 1$;
5. $\mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B}) - \mathrm{P}(\mathrm{A} \cap \mathrm{B})$

Przestrzeń probabilistyczna: (E, Z, P)

Prawdopodobieństwo warunkowe: $P(A|B) = \frac{P(A \cap B)}{P(B)}$, $\mathrm{A}, \mathrm{B} \in \mathrm{Z}, \mathrm{P}(\mathrm{B}) > 0$

Prawdopodobieństwo iloczynu zdarzeń:

$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{A}) * \mathrm{P}(\mathrm{B}|\mathrm{A}) = \mathrm{P}(\mathrm{B}) * \mathrm{P}(\mathrm{A}|\mathrm{B})$, gdy odpowiednio $\mathrm{P}(\mathrm{A})$ i $\mathrm{P}(\mathrm{B}) > 0$
$\mathrm{P}(\mathrm{A} \cap \mathrm{B} \cap \mathrm{C}) = \mathrm{P}(\mathrm{A}) * \mathrm{P}(\mathrm{B}|\mathrm{A}) * \mathrm{P}(\mathrm{C}|\mathrm{A} \cap \mathrm{B})$, gdy $\mathrm{P}(\mathrm{A} \cap \mathrm{B}) > 0$.

Niezależność dwóch zdarzeń: $\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{A}) * \mathrm{P}(\mathrm{B})$

Niezależność zespołowa zdarzeń:
Zdarzenia $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n$ są niezależne zespołowo, gdy prawdopodobieństwo łącznego zaistnienia dowolnych $\mathrm{m} \leq \mathrm{n}$ różnych zdarzeń spośród nich jest równe iloczynowi prawdopodobieństwa tych zdarzeń.

$\mathrm{P}(\mathrm{A} \cap \mathrm{B} \cap \mathrm{C}) = \mathrm{P}(\mathrm{A}) * \mathrm{P}(\mathrm{B}) * \mathrm{P}(\mathrm{C})$ oraz $\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{A}) * \mathrm{P}(\mathrm{B})$ oraz $\mathrm{P}(\mathrm{A} \cap \mathrm{C}) = \mathrm{P}(\mathrm{A}) * \mathrm{P}(\mathrm{C})$ oraz $\mathrm{P}(\mathrm{B} \cap \mathrm{C}) = \mathrm{P}(\mathrm{B}) * \mathrm{P}(\mathrm{C})$.

Twierdzenie o prawdopodobieństwie zupełnym
Jeśli B jest dowolnym zdarzeniem, zdarzenia $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n$ stanowią układ zupełny zdarzeń (wykluczają się parami i wypełniają całą przestrzeń zdarzeń elementarnych) to prawdopodobieństwo zdarzenia B wyraża się wzorem:

$\mathrm{P}(\mathrm{B}) = \mathrm{P}(\mathrm{A1}) * \mathrm{P}(\mathrm{B}|\mathrm{A1}) + \mathrm{P}(\mathrm{A2}) * \mathrm{P}(\mathrm{B}|\mathrm{A2}) + \ldots + \mathrm{P}(\mathrm{An}) * \mathrm{P}(\mathrm{B}|\mathrm{An})$

Twierdzenie Bayes'a
Jeśli zdarzenie B jest dowolnym zdarzeniem o dodatnim prawdopodobieństwie, zdarzenia $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n$ stanowią układ zupełny zdarzeń to: $P(Ai|B) = \frac{P(Ai) * P(B|Ai)}{P(B)}$

Przykład

70 kobiet i 30 mężczyzn; 35 kobiet i 10 mężczyzn uzyskało ocenę 5,0
A_1 – kobieta; A_2 – mężczyzna; B – otrzymanie oceny 5,0

$$P(B) = P(A_1) * P(B|A_1) + P(A_2) * P(B|A_2) = 70/100 * 35/70 + 30/100 * 10/30 = 45/100$$

$$P(A_i|B) = \frac{P(A_i)*P(B|A_i)}{P(B)} = \frac{\frac{70}{100}*\frac{35}{70}}{\frac{45}{100}} = 0{,}78$$

Zmienna losowa

Zmienna losowa – taka wielkość, która w wyniku doświadczenia losowego przyjmuje określoną wartość, znaną po zrealizowaniu doświadczenia, ale nie dającą się przewidzieć przed realizacją tego doświadczenia.

Zmienną losową X nazywa się funkcję $X=X(e)$ określoną na zbiorze zdarzeń elementarnych E, o wartościach ze zbioru liczb rzeczywistych taką, że dla każdej liczby rzeczywistej x zbiór A zdarzeń elementarnych e∈E, dla których X(e) < x spełnia warunek A ∈ Z.

Zmienna dyskretna i ciągła

Dystrybuanta zmiennej losowej: $F(x) = P(X < x)$

Własności:

1. $0 \leq F(x) \leq 1$
2. $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$
3. $F(x)$ jest funkcją niemalejącą
4. $F(x)$ jest funkcją co najmniej lewostronnie ciągłą
5. $P(a \leq X < b) = F(b) - F(a)$

Rozkład prawdopodobieństwa zmiennej losowej skokowej

$P(X = x_i) = p_i$

$$\sum_{i=1}^{n} p_i = 1$$

Charakterystyki rozkładu:

1. Wartość oczekiwana - $E(X) = \sum_{i=1}^{n} x_i p_i$
2. $E(a) = a$
3. $E(X+a) = E(X) + a$; $E(X*a) = E(X) * a$
4. $E(X-E(X)) = 0$
5. $E(X+Y) = E(X) + E(Y)$; $E(X*Y) = E(X) * E(Y)$, gdy X, Y są niezależne

Wariancja

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

$D^2(a) = 0$
$D^2(a*X) = a^2 * D^2(X)$

D^2 (X+a) = D^2 (X)
D^2 (X+Y) = D^2 (X) + D^2(Y), gdy X, Y są niezależne

Rozkłady prawdopodobieństwa

Rozkład jednopunktowy

- P(X=x) = 1
- E(X)=x
- D^2(X) = 0

Rozkład dwupunktowy

- P(X=x_1) = p_1
- P(X=x_2) = p_2

Rozkład zero-jedynkowy

- P(X=0) = 1-p=q
- P(X=1) = p
- E(X)=p
- (X) = p*q = p*(1-p)

Rozkład jednostajny skokowy

P(X=x_1) = 1/n; P(X=x_2) = 1/n; ..., P(X=x_n) = 1/n

E(X) = $\frac{1}{n}\sum_{i=1}^{n} X_i$; D^2(X) = $\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X))^2$

Rozkład Bernoulliego (dwumianowy): X~B(n,p)

Założenia:

1. n niezależnych doświadczeń losowych
2. każde z tych doświadczeń może zakończyć się sukcesem lub porażką
3. prawdopodobieństwo sukcesu jest jednakowe dla każdego doświadczenia i wynosi p

P(X = k) = $\binom{n}{k} p^k (1-p)^{n-k}$ E(X) = np D^2(X) = npq

Rozkład Poissona: X~P(λ)

Warunki jak w przypadku r. Bernoulliego i jeden dodatkowy:

- Liczba doświadczeń jest duża (zmierza do nieskończoności)

P(X = k) = $\frac{\lambda^k}{k!} e^{-\lambda}$ E(X) = **λ** D_2(X) = **λ**

Rozkłady zmiennej losowej ciągłej

Gęstość prawdopodobieństwa: P($x_1 \leq X < x_2$) = F(x_2) – F(x_1), gdy $\Delta x = x_2 - x_1 \rightarrow 0$ to
P($x_1 \leq X < x_2$) = dF(x) + r(x)

dF(x) = F'(x)dx, więc

Jeśli dystrybuanta $\mathrm{F}(\mathrm{x})$ ma pochodną w punkcie x to pochodna ta nazywa się gęstością prawdopodobieństwa zmiennej losowej X w punkcie x, więc

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

Własności:

1. $\mathrm{f}(\mathrm{x}) \geq 0$
2. $P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} f(t)dt$
3. $\int_{-\infty}^{\infty} f(x)dx = 1$
4. $\mathrm{P}(\mathrm{X} = \mathrm{x}) = 0$

Zmienną losową X przyjmującą wszystkie wartości z pewnego przedziału, dla której istnieje nieujemna funkcja f taka, że dystrybuantę F zmiennej losowej X można przedstawić w postaci $F(x) = \int_{-\infty}^{x} f(t)dt$ nazywamy zmienną losową ciągłą, a funkcję f jej gęstością.

Zmienna losowa X jest zmienną ciągłą w danym przedziale, jeśli w tym przedziale gęstość $\mathrm{f}(\mathrm{x})$ istnieje i jest funkcją ciągłą względem x w całym przedziale z wyjątkiem co najwyżej skończonej liczby punktów.

Wartość oczekiwana: $E(x) = \int_{-\infty}^{\infty} xf(x)dx$

Wariancja: $D^2(x) = \int_{-\infty}^{\infty} \big(x - E(x)\big)^2 f(x)dx = E(X^2) - E(X)^2$

Rozkład jednostajny ciągły:

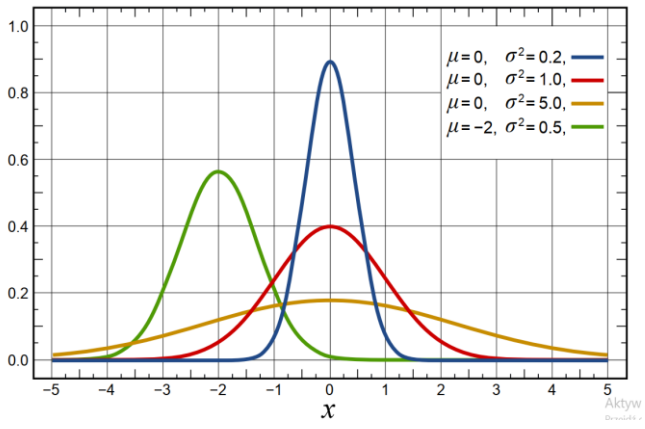
$$\mathrm{F'(x)} = \lim_{\Delta x \to 0} \frac{F(x_2) - F(x_1)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P(x_1 \leq X < x_2)}{\Delta x} \qquad f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & poza\ tym \end{cases} \qquad \mathrm{F}(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1 & x > b \end{cases}$$

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

Rozkład normalny: $\mathrm{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$

$$f(x) = \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mathrm{E}(\mathrm{X}) = \mathrm{M_o} = \mathrm{M_e} = \boldsymbol{\mu}$ $\qquad$ $\mathrm{D}^2(\mathrm{X}) = \boldsymbol{\sigma}^2$



Reguła trzech sigm:

- $P\ (\boldsymbol{\mu}\text{-}\boldsymbol{\sigma} < X < \boldsymbol{\mu}+\boldsymbol{\sigma}) = 0{,}6826$
- $P\ (\boldsymbol{\mu}\text{-}2\boldsymbol{\sigma} < X < \boldsymbol{\mu}+2\boldsymbol{\sigma}) = 0{,}9545$
- $P\ (\boldsymbol{\mu}\text{-}3\boldsymbol{\sigma} < X < \boldsymbol{\mu}+3\boldsymbol{\sigma}) = 0{,}9973$

Standaryzacja:

gdy $X \sim N\ (\boldsymbol{\mu}, \boldsymbol{\sigma})$ $\quad U = \frac{X-\mu}{\delta}$

więc $U \sim N\ (0, 1)$ $\quad P(X < x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(u)$

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

Dwuwymiarowa i wielowymiarowa zmienna losowa

(X, Y) oraz $(X_1, X_2, \ldots, X_n)$

Badanie wielowymiarowych zmiennych losowych jest interesujące z powodu zależności między zmiennymi składowymi.

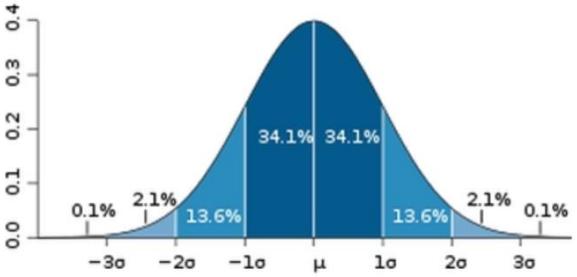
- Rozkłady brzegowe
- Rozkłady warunkowe
- Korelacja, regresja
- Wielowymiarowy rozkład normalny

Centralne twierdzenie graniczne Linderberg'a-Levy'ego

Jeżeli {Xn} jest losowym ciągiem niezależnych zmiennych o jednakowym rozkładzie, o wartości przeciętnej $\boldsymbol{\mu}$ i skończonej wariancji $\boldsymbol{\sigma}_2 > 0$ to ciąg (Fn) dystrybuant standaryzowanych średnich arytmetycznych $\overline{X}_n$ (lub standaryzowanych sum $\sum_{i=1}^{n} X_i$)

$$Y_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

jest zbieżny do dystrybuanty $\boldsymbol{\phi}$ rozkładu N (0, 1).



Przykład

Przy założeniu, że rozkład wydajności pracy (w szt. na dzień) jest rozkładem jednostajnym o parametrach a=10 i b=40, obliczyć prawdopodobieństwo, że średnia wydajność pracy 360 pracowników będzie większa od 26 szt. w ciągu dnia roboczego.

$E(X) = \frac{10+40}{2} = 25$ $\qquad D^2(X) = \frac{(40-10)^2}{12} = 75$ $\qquad D(X) = 8{,}66$

$\sum X = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N(25, 0{,}456)$

$$P(\overline{x} > 26) = 1 - P\left(\frac{26-25}{0{,}456}\right) = 1 - \Phi(2{,}19) = 1 - 0{,}986 = 0{,}014$$

Twierdzenie graniczne Moivre'a-Laplace'a

Jeżeli {Xn} jest losowym ciągiem niezależnych zmiennych o rozkładzie dwumianowym z parametrami n, p to ciąg (Fn) dystrybuant standaryzowanych sum $\sum_{i=1}^{n} X_i$)

$$Y_n = \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{npq}}$$

jest zbieżny do dystrybuanty **Φ** rozkładu N (0, 1).

Przykład

PZU ocenia, że każdego roku 1% ubezpieczonych mężczyzn traci życie w wypadkach. Jakie jest prawdopodobieństwo, że w danym roku PZU będzie musiało wypłacić odszkodowanie więcej niż trzy razy, jeśli ubezpieczyło się od wypadków 100 mężczyzn?

$X \sim N(np, \sqrt{npq}) = X \sim N(100 * 0{,}01, 100 * 0{,}01 * 0{,}99) = X \sim N(1, 0{,}995)$

$P(X>3) = 1\text{-}\ P(X\leq 3) = 1\text{-}\ \Phi\left(\frac{3-1}{0{,}995}\right) = 1 - \Phi\,(2{,}01) = 1 - 0{,}97778 = 0{,}02222$

Statystyka matematyczna

Zagadnienia wstępne

W rachunku prawdopodobieństwa zakłada się znajomość rozkładu prawdopodobieństwa zmiennej losowej. W statystyce matematycznej nie zakłada się pełnej znajomości rozkładu – poznaje się go w ramach tzw. wnioskowania statystycznego, czyli:

- Estymacji parametrycznej (punktowej lub przedziałowej) lub nieparametrycznej.
- Weryfikacji hipotez parametrycznych lub nieparametrycznych

Wnioskowanie statystyczne oparte jest na częściowej informacji.

Próba statystyczna – podzbiór populacji generalnej.

Próba jest reprezentatywna, gdy jej struktura ze względu na badane cechy jest co najmniej zbliżona do struktury populacji generalnej, z której ona pochodzi. Próba reprezentatywna: losowa i odpowiednio liczna.

Próba jest losowa, gdy dobór jednostek do próby został dokonany w drodze doboru losowego (losowania).
Schematy losowania:

1. losowanie niezależne i zależne
2. losowanie indywidualne i zespołowe
3. losowanie jednostopniowe i wielostopniowe
4. losowanie nieograniczone i ograniczone (tu warstwowe i systematyczne)

Losowanie proste: indywidualne, nieograniczone i niezależne.

Próba prosta – ciąg niezależnych zmiennych losowych X_1, ..., X_n o jednakowym rozkładzie takim jaki ma cecha X w populacji.

Model statystyczny to przestrzeń próby χ (chi) wraz z rodziną rozkładów P.

Niech rozkład badanej cechy X zależy od nieznanego parametru Θ, który będzie szacowany o n-elementową próbę prostą pobraną z populacji.

Każdą funkcję g $(X_1, ..., X_n)$, będącą funkcją próby losowej $X_1, ..., X_n$ nazywamy statystyką.

Rozkład statystyki zależy od:

- postaci funkcji g,
- rozkładu zmiennych losowych $X_1, ..., X_n$,
- liczebności próby

Rozkład statystyki $\Theta\hat{}_n$ - rozkład z próby

- dokładny
- graniczny

Rozkłady statystyk z próby

1. Rozkład średniej arytmetycznej z próby dla zmiennej X o rozkładzie normalnym

   $X \sim N(\mu, \sigma)$ $\quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

   $$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$
   $$D^2(\bar{X}) = D^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D^2(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$
   $$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

2. Rozkład średniej arytmetycznej z próby dla zmiennej X o rozkładzie normalnym z nieznanym odchyleniem standardowym σ

   $X \sim N(\mu, \sigma)$ $\quad \sigma$ - ? $\quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

   $$t = \frac{\bar{X} - \mu}{S}\sqrt{n-1} = \frac{\bar{X} - \mu}{\hat{S}}\sqrt{n} \sim t - Studenta$$
   $$o\ n-1\ stopniach\ swobody$$
   $$E(t) = 0 \qquad D^2(t) = \sqrt{\frac{n-1}{n-3}}$$

3. Rozkład różnicy średnich arytmetycznych z próby dla X o rozkładzie normalnym

   $X_1 \sim N(\mu_1, \sigma_1)$, $\quad X_2 \sim N(\mu_2, \sigma_2)$

   $$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right), \qquad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

   $$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

4. Rozkład różnicy średnich arytmetycznych z próby dla zmiennej X o rozkładzie normalnym z nieznanymi, ale jednakowymi odchyleniami standardowymi

   $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$ $\sigma_1 = \sigma_2$ - ?

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t - Studenta\ o\ n_1 + n_2 - 2$$

   stopniach swobody

5. Rozkład wariancji z próby dla zmiennej X o rozkładzie normalnym $X \sim N(\mu, \sigma)$

$$\chi^2 = \sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sum(X_i - \mu)^2}{\sigma^2} = \frac{nS_*^2}{\sigma^2} \sim \chi^2 o\ n\ st.swob.$$

   gdy $\boldsymbol{\mu}$ - ?

$$\chi^2 = \frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2 o\ v = n - 1\ st.swob. \qquad E(\chi^2) = v;\ D^2(\chi^2) = 2v$$

   gdy v → ∞ to $\sqrt{2\chi^2} \sim N\left(\sqrt{2v - 1}, 1\right)$

6. Rozkład ilorazu wariancji z próby dla zmiennej X o rozkładzie normalnym dla dwóch populacji

   $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$

$$F = \frac{\frac{n_1 S_1^2}{\sigma_1^2(n_1 - 1)}}{\frac{n_2 S_2^2}{\sigma_2^2(n_2 - 1)}} = \frac{n_1 S_1^2}{\sigma_1^2(n_1 - 1)} * \frac{\sigma_2^2(n_2 - 1)}{n_2 S_2^2} =$$

$$= \frac{S_1^2}{S_2^2} * \frac{\sigma_2^2}{\sigma_1^2} * \frac{(n_2 - 1)}{(n_1 - 1)} * \frac{n_1}{n_2} \sim rozkład\ F - Snedecora\ o\ n_1 - 1\ i\ n_2 - 1\ st.swob.$$

$$E\left(\frac{v_2}{v_2 - 2}\right) \qquad D^2\left(\frac{2v_2^2(v_1 + v_2 - 2}{v_1(v_2 - 2)^2(v_2 - 4)^2}\right)$$

7. Graniczny rozkład frakcji z próby dla zmiennej X o rozkładzie Bernoulliego

   $X_n \sim B(n, p)$,

$$X_n = \sum X_i \qquad Y_n = \frac{X_n}{n}$$

   Z twierdzenia Moivre'a-Laplace'a

$$Y_n \sim N\left(p, \sqrt{\frac{pq}{n}}\right) \qquad X_n \sim N\left(np, \sqrt{npq}\right)$$

8. Rozkład średniej arytmetycznej z próby dla zmiennej X o dowolnym rozkładzie

   X~rozkład dowolny $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

   Z twierdzenia Lindeberga-Levy'ego wynika, że:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

   Podobnie dla różnicy średnich

9. Rozkład różnicy frakcji z próby dla zmiennej X o rozkładzie Bernoulliego
   $X_{n1} \sim B(n_1, p_1)$, $X_{n2} \sim B(n_2, p_2)$

$$X_{n1} = \sum X_{i1} \qquad Y_{n1} = \frac{X_{n1}}{n_1} \qquad X_{n2} = \sum X_{i2} \qquad Y_{n2} = \frac{X_{n2}}{n_2}$$

$$Y_{n1} - Y_{n2} \sim N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

Estymacja

Estymacja – szacowanie, ocenianie, przybliżanie

Estymator – każda statystyka $\Theta\hat{}_n\,(X_1, \ldots, X_n)$, która służy do oszacowania danego parametru. Estymator – zmienna losowa, jego rozkład jest zależny od rozkładu zmiennej losowej X w populacji oraz od szacowanego parametru.

Ocena parametru – wartość estymatora obliczona na bazie jednej próby

Błąd estymacji (szacunku) parametru $\boldsymbol{\theta}$ – różnica między wartością estymatora a wartością parametru: $d = \Theta\hat{}_n - \Theta$.

Błąd szacunku jest zmienną losową, więc: $\Delta = E\,(\Theta\hat{}_n - \Theta)^2$.

Jeśli $E\,(\Theta\hat{}_n) = \Theta$ to $\Delta = E\,(\Theta\hat{}_n - E\,(\Theta\hat{}_n))^2 = D(\Theta\hat{}_n)^2$

$D(\Theta\hat{}_n)$ - standardowy błąd szacunku

$D(\Theta\hat{}_n) / \Theta$ - względny błąd szacunku

Własności estymatora:

1. Nieobciążoność

$$E(\hat{\Theta}_n) = \Theta$$

Obciążenie: $B_n(\Theta) = E(\hat{\Theta}_n) - \Theta$

Estymator asymptotycznie nieobciążony

$$\lim_{n\to\infty} B_n(\Theta) = 0$$

2. Zgodność estymatora

$$\lim_{n\to\infty} P\left(\left|\hat{\Theta}_n - \Theta\right| < \varepsilon\right) = 1, \boldsymbol{\varepsilon} > 0.$$

Jeśli estymator jest zgodny to jest **asymptotycznie nieobciążony**.
Jeśli estymator jest nieobciążony lub asymptotycznie nieobciążony oraz jego wariancja spełnia warunek $\lim_{n\to\infty} D^2(\hat{\Theta}_n) = 0$ to estymator ten jest estymatorem zgodnym.

3. Efektywność estymatora

Estymatorem efektywnym (najefektywniejszym) nazywamy nieobciążony estymator $\Theta\hat{}_n$ parametru Θ, który ma najmniejszą wariancję spośród wszystkich nieobciążonych estymatorów tego parametru wyznaczonych z prób n-elementowych.

Nierówność Rao-Cramera

$$D^2(\hat{\Theta}_n) \geq \frac{1}{nE\left(\frac{\partial}{\partial\Theta} lnf(X,\Theta)\right)^2}$$

Miarą efektywności estymatora $\Theta\hat{}_n$ jest liczba

$$ef(\hat{\Theta}_n) = \frac{D^2(\tilde{\Theta}_n)}{D^2(\hat{\Theta}_n)}$$

$$0 < ef(\hat{\Theta}_n) < 1$$

Estymator asymptotycznie efektywny

$$\lim_{n\to\infty} ef(\hat{\Theta}_n) = 1$$

Metody wyznaczania estymatorów:

1. Metoda momentów – przyrównanie momentów z próby do momentów rozkładu
   Własności takich estymatorów:
   - Na ogół niska efektywność
   - Są na ogół zgodne
2. Metoda największej wiarygodności (MNW)
   Niech rozkład zmiennej X zależy od k nieznanych parametrów $\boldsymbol{\theta}1$, $\boldsymbol{\theta}2$, …, $\boldsymbol{\theta}$k, które chcemy oszacować na podstawie n-elementowej próby.

   $$L = \prod_{i=1}^{n} f(X_i, \Theta_1, \dots, \Theta_k)$$

   Własności takich estymatorów:
   - zgodne
   - asymptotycznie nieobciążone
   - asymptotycznie efektywne
   - mają asymptotyczne rozkłady normalne

Przegląd estymatorów

Szanowanie wartości przeciętnej $\boldsymbol{\mu}$:

Estymator - średnia arytmetyczna z próby: nieobciążony, zgodny dla rozkładu dowolnego, efektywny dla rozkładu normalnego,

Estymator – mediana z próby: asymptotycznie nieobciążony, zgodny dla rozkładu dowolnego, efektywność równa $2/\boldsymbol{\pi}$ = 0,64 dla rozkładu normalnego

Szanowanie wariancji $\boldsymbol{\sigma}2$:

Estymator - $S_*^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$: nieobciążony, zgodny dla rozkładu dowolnego, efektywny dla rozkładu normalnego,

Estymator – $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$: asymptotycznie nieobciążony, zgodny dla rozkładu dowolnego

Estymator – $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$: nieobciążony, zgodny dla rozkładu dowolnego, efektywność równa (n-1)/n dla rozkładu normalnego

Szanowanie odchylenia standardowego $\boldsymbol{\sigma}$

Estymator - $S*$, S , $S\hat{}$: zgodny dla rozkładu dowolnego

Estymator – b_nS, c_nS: zgodny, asymptotycznie nieobciążony, efektywny dla rozkładu normalnego

$$b_n = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma(n)}\sqrt{\frac{n}{2}} \qquad c_n = \sqrt{\frac{n-1}{n}}\,b_n$$

dla n = 20 b_n = 1,04, a c_n = 1,014

Szanowanie wskaźnika struktury p

Estymator - $\hat{p}$ = m/n: nieobciążony, zgodny i efektywny dla rozkładu Bernoulliego

Estymacja przedziałowa

Przedziałem ufności dla parametru $\boldsymbol{\theta}$ na poziomie ufności 1 – α nazywamy przedział (θ_1, θ_2) spełniający warunki:

- jego końce θ_1 (X_1, …, X_n) i θ_2 (X_1, …, X_n) są funkcjami próby losowej i nie zależą od szacowanego parametru,
- prawdopodobieństwo pokrycia przez ten przedział nieznanego parametru $\boldsymbol{\theta}$ jest równe 1 - α , tzn.

$$P\left(\theta_1(X_1, \dots, X_n) < \theta < \theta_2(X_1, \dots, X_n)\right) = 1 - \alpha$$

Dokładność estymacji – różnica między górną i dolną granicą przedziału – długość przedziału ufności.

Zależy ona od:

- współczynnika ufności – im wyższy tym długość przedziału większa a mniejsza dokładność,
- liczebności próby – im większa tym długość przedziału mniejsza a dokładność większa

Konstrukcja przedziału ufności dla wartości przeciętnej μ w populacji, w której badana cecha ma rozkład N(μ, σ), gdy σ jest znane.

1. Wiadomo, że statystyka $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ma rozkład $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
2. Standaryzujemy: $U = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$
3. Ponieważ rozkład U nie zależy od szacowanego parametru możemy wykorzystać go do konstrukcji przedziału ufności
4. Dla danego α można znaleźć takie wartości u1 i u2 , aby:

$$P(u_1 < U < u_2) = \Phi(u_2) - \Phi(u_1) = 1 - \alpha$$

$\alpha_1 + \alpha_2 = \alpha$; $\alpha_1 > 0$; $\alpha_2 < \alpha$

$u_1 = U(\alpha_1)$; $u_2 = U(1-\alpha_2)$

5. Podstawiamy

$$P\left(u(\alpha_1) < \frac{\bar{X} - \mu}{\sigma}\sqrt{n} < u(1-\alpha_2)\right) = 1 - \alpha$$

6. Rozwiązując nierówność wewnątrz nawiasu względem μ mamy:

$$P\left(\bar{X} - u(1-\alpha_2)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - u(\alpha_1)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

7. Przyjmując, że $\alpha 1 = \alpha 2 = ½\alpha$ to przedział ufności wygląda następująco:

$$P\left(\bar{X} - u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - u\left({}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

8. Ze względu na symetrię rozkładu N(0, 1), gdzie $-u(\alpha/2) = u(1-\alpha/2)$:

$$P\left(\bar{X} - u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Wyznaczanie minimalnej liczebności próby

Przyjmijmy, że estymować chcemy wartość przeciętną μ w populacji, w której badana cecha ma rozkład $N(\mu, \sigma)$, gdy σ jest znane. Wtedy przedział ufności jest następujący:

$$P\left(\bar{X} - u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Długość tego przedziału wynosi:

$$2d = 2u\left(1 - {}^{\alpha}/_{2}\right)\frac{\sigma}{\sqrt{n}}$$

Po przekształceniu

$$n = \left[\frac{u^2\left(1 - {}^{\alpha}/_{2}\right)\sigma^2}{d^2}\right] + 1$$

Analogiczny wzór uzyskujemy dla w przypadku estymacji wskaźnika struktury:

$$n = \left[\frac{u^2\left(1 - {}^{\alpha}/_{2}\right)\hat{p}(1-\hat{p})}{d^2}\right] + 1$$

Weryfikacja hipotez

Hipoteza statystyczna – każde przypuszczenie dotyczące nieznanego rozkładu badanej cechy populacji o prawdziwości lub fałszywości, którego wnioskuje się na podstawie próby losowej.

Hipotezy:

- parametryczne
- nieparametryczne
- proste

- złożone

Test statystyczny – metoda postępowania (procedura), która każdej możliwej realizacji próby losowej X1, …, Xn przyporządkowuje (z ustalonym prawdopodobieństwem) decyzję przyjęcia lub odrzucenia sprawdzanej hipotezy.

Konstrukcja testu statystycznego

1. Formułuje się hipotezę, którą weryfikujemy (hipotezę zerową – H0)
2. Formułuje się hipotezę alternatywną
3. Wybieramy odpowiednią statystykę testową
4. Konstruuje się tzw. zbiór krytyczny

Błędy przy weryfikacji hipotezy statystycznej

Decyzja	Hipoteza H_0	
	prawdziwa	fałszywa
Przyjąć H_0	Decyzja poprawna	Błąd II rodzaju
Odrzucić H_0	Błąd I rodzaju	Decyzja poprawna

$$P(\delta(X_1, \dots, X_n) \in W | H_0) = \alpha$$

$$P(\delta(X_1, \dots, X_n) \in W' | H_1) = \beta$$

Testy konstruuje się tak, aby zminimalizować prawdopodobieństwo popełnienia błędu II rodzaju przy ustalonym poziomie prawdopodobieństwie popełnienia błędu pierwszego rodzaju (α).

Takie testy nazywamy testami najmocniejszymi, ponieważ przy ustalonym α odpowiada im największa moc, tzn. prawdopodobieństwo odrzucenia fałszywej hipotezy H0 i przyjęcia hipotezy H1.

Test jednostajnie najmocniejszy – najmocniejszy względem każdej hipotezy H1

Moc testu

$$M(W, \Theta) = P(\delta(X_1, \dots, X_n) \in W | \Theta) = 1 - P(\delta(X_1, \dots, X_n) \in W' | \Theta)$$

$$M(W, \Theta_0) = P(\delta(X_1, \dots, X_n) \in W | \Theta_0) = P(\delta(X_1, \dots, X_n) \in W | H_0) = \alpha$$

$$M(W, \Theta_1) = P(\delta(X_1, \dots, X_n) \in W | \Theta_1) = P(\delta(X_1, \dots, X_n) \in W | H_1)$$
$$= 1 - P(\delta(X_1, \dots, X_n) \in W' | H_1) = 1 - \beta$$

Funkcja operacyjno-charakterystyczna (charakterystyka testu)

$$L(W, \Theta) = P(\delta(X_1, \dots, X_n) \in W' | \Theta) = 1 - M(W, \Theta)$$

$$L(W, \Theta_0) = 1 - \alpha$$

$$L(W, \Theta_1) = \beta$$

Testy nieparametryczne

- testy zgodności
- testy niezależności

Test zgodności $\chi 2$ – Pearsona
X ~ dowolny rozkład o określonej postaci dystrybuanty
Duża próba prosta
H0: F(x) = F0(x)
H1: F(x) ≠ F0

Statystyka testowa

$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2$ o k-1 stopniach swobody przy hipotezie prostej, natomiast przy hipotezie złożonej liczba stopni swobody wynosi k-s-1.

Warunki stosowalności:

- Taki podział na klasy, by npi ≥ 5
- Gdy liczba stopni swobody jest > 5, to npi w dwóch klasach może być mniejsze od 5, ale większe bądź równe od 1
- W rozkładach jednomodalnych o klasach tej samej długości w skrajnych klasach mogą być mniejsze liczebności teoretyczne

Test niezależności $\chi 2$

Cel: weryfikacja hipotezy o związkach między dwiema zmiennymi

H0: P (X = xi, Y = yj) = P(X = xi) * P(Y = yj)

H1: P (X = xi, Y = yj) ≠ P(X = xi) * P(Y = yj)

Statystyka testowa:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - \widehat{n_{ij}})^2}{\widehat{n_{ij}}}$$
$$P\left(\chi^2 \geq \chi^2_{\alpha;(r-1)*(s-1)}\right) = \alpha$$

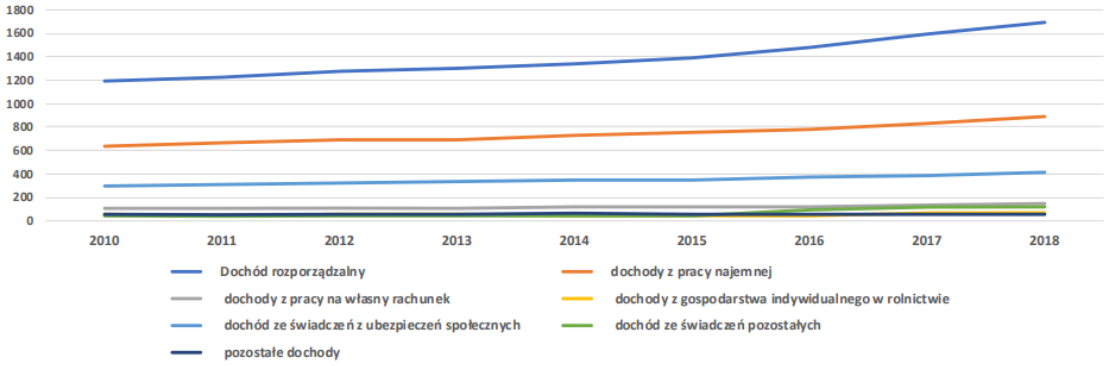
Proste metody analizy dynamiki zjawisk

Szereg czasowy (dynamiczny, chronologiczny)

W ujęciu tradycyjnym szereg czasowy jest ciąg wielkości charakteryzujących element lub zbiorowość, uporządkowanych według jednostek czasu. Jednostka czasu: punkt lub okres

y_t – wielkość zjawiska w t-tej jednostce czasu

Kategoria dochodu / Rok	2010	2011	2012	2013	2014	2015	2016	2017	2018
Dochód rozporządzalny	1192,82	1226,95	1278,43	1299,07	1340,44	1386,16	1474,56	1598,13	1693,46
dochody z pracy najemnej	636,56	667,03	685,77	690,31	723,57	757,89	778,01	831,94	889,99
dochody z pracy na własny rachunek	109,26	109,01	108,46	111,81	114,51	120,14	124,38	134,58	147,29
dochody z gospodarstwa indywidualnego w rolnictwie	50,32	44,61	52,84	57,24	44,65	44,22	48,21	67,45	63,64
dochód ze świadczeń z ubezpieczeń społecznych	297,03	309,41	327,21	333,44	343,09	353,35	369,67	381,52	414,16
dochód ze świadczeń pozostałych	40,00	39,45	42,04	44,77	44,61	43,95	89,47	117,58	114,22
pozostałe dochody	55,27	52,27	56,75	55,33	65,03	59,76	59,15	59,67	56,80



Indeksowe metody analizy dynamiki

Proste metody analizy dynamiki:

- przyrosty:
  - absolutne
  - względne

  Jedne i drugie mogą być:
  - o podstawie stałej
  - łańcuchowe
- indeksy:
  - indywidualne
  - agregatowe

  Jedne i drugie mogą być o podstawie stałej lub łańcuchowe

Przyrosty:

- absolutne o podstawie stałej:
  $$\Delta_{t/0} = y_t - y_0$$
- absolutne o podstawie zmiennej (łańcuchowe):
  $$\Delta_{t/t-1} = y_t - y_{t-1}$$
- względne o podstawie stałej:
  $$\Delta'_{t/0} = \frac{y_t - y_0}{y_0}$$
- względne o podstawie zmiennej (łańcuchowe):
  $$\Delta'_{t/t-1} = \frac{y_t - y_{t-1}}{y_{t-1}}$$

Indeksy indywidualne:

- podstawie stałej:

$$i_{t/0} = \frac{y_t}{y_0} = \Delta'_{t/0} + 1 = \frac{y_t - y_0}{y_0} + \frac{y_0}{y_0}$$

- o podstawie zmiennej (łańcuchowe):

$$i_{t/t-1} = \frac{y_t}{y_{t-1}} = \Delta'_{t/t-1} + 1 = \frac{y_t - y_{t-1}}{y_{t-1}} + \frac{y_{t-1}}{y_{t-1}}$$

P.S.: Moi drodzy, na tym egzaminie nic strasznego nie ma. Dał egzamin na 20 minut (20 pytań testowych) i cały czas chodził po sali (i tam jeszcze jakąś studentkę poprosił, żeby śledziła za nami). Ale jak się nauczycie tych odpowiedzi ze spikera, to pójdzie wam na 4 min. Więc trzymam kciuki za was ^^