



Andrzej Krajka

Metody probabilistyczne i statystyka

Literatura

- ❶ Fisz M., Rachunek prawdopodobieństwa i statystyka matematyczna, Warszawa, 1986
- ❷ Gesternkorn T., Śródka T., Kombinatoryka i rachunek prawdopodobieństwa, PWN, 1983
- ❸ Greń J., Modele i zadania statystyki matematycznej, PWN, Warszawa, 1968
- ❹ Kryszicki W., Bartos J., Dyczka W., Królikowska K., Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach. Cz. I i II, Warszawa, 1997
- ❺ Krzyśko M., Wykłady z teorii prawdopodobieństwa, WNT, 2000
- ❻ Krzyśko M., Statystyka matematyczna, Uniwersytet Adama Mickiewicza, 2004
- ❼ Feller W., Wstęp do rachunku prawdopodobieństwa i jego zastosowań, PWN, Warszawa, 1987
- ❽ Kubik L. M., Zastosowanie elementarnego rachunku prawdopodobieństwa do wnioskowania statystycznego, PWN, Warszawa, 1995
- ❾ Plucińska A., Pluciński E., Rachunek prawdopodobieństwa. Statystyka matematyczna. Procesy stochastyczne, WNT, Warszawa, 2000
- ❿ Biecek P., Przewodnik po pakiecie R, Wrocław 2008

Rachunek prawdopodobieństwa

- 1 Pojęcia wstępne: doświadczenie losowe, przestrzeń zdarzeń elementarnych, działania na zdarzeniach.
- 2 Definicje prawdopodobieństwa: klasyczna, geometryczna, aksjomatyczna. Wnioski z aksjomatów rachunku prawdopodobieństwa.
- 3 Prawdopodobieństwo warunkowe. Twierdzenie o prawdopodobieństwie całkowitym. Twierdzenie Bayesa. Niezależność zdarzeń.
- 4 Zmienna losowa. Rozkład prawdopodobieństwa. Zmienne losowe typu skokowego i ciągłego. Dystrybuanta. Gęstość prawdopodobieństwa.
- 5 Rozkłady dyskretne (dwumianowy, Poissona, geometryczny, ujemnie dwumianowy) i absolutnie ciągłe (prostokątny, wykładniczy, gamma, normalny, Cauchy'ego, Laplace'a). Rozkłady funkcji zmiennej losowej.
- 6 Wartość oczekiwana, wariancja i inne charakterystyki rozkładów prawdopodobieństwa (zmiennych losowych).

Statystyka

- 1 Statystyka opisowa, miary statystyczne, szeregi rozdzielcze
- 2 Estymacja punktowa i przedziałowa
- 3 Weryfikacja hipotez statystycznych (parametrycznych i nieparametrycznych)
- 4 Jednoczynnikowa Analiza Wariancji
- 5 Regresja

- 1 EXCEL (VBA)
- 2 MATLAB (Statistics toolbox)
- 3 STATISTICA (Statistica Basic i R)
- 4 SPSS
- 5 PYTHON (pandas)
- 6 R

Rachunek prawdopodobieństwa

Kombinatoryka

Kombinatoryka jest działem matematyki zajmującym się obliczaniem liczby zbiorów jakie można utworzyć z danej liczby elementów.

Kombinatoryka - metody tworzenia zbiorów

	Liczba elementów	Liczba wybranych	Czy zwracamy	Czy ważna jest kolejność
permutacje	n	n	nie	tak
permutacje z powtórzeniami	n część identyczna	n	nie	tak
wariacje	n	$k \leq n$	nie	tak
wariacje z powtórzeniami	n	k	tak	tak
kombinacje	n	$k \leq n$	nie	nie
kombinacje z powtórzeniami	n	k	tak	nie

Kombinatoryka - wzory

	oznaczenie	liczba zbiorów
permutacje	P_n	$n!$
permutacje z powtórzeniami	$\overline{P}_n^{n_1, n_2, \dots, n_k}$	$\frac{n!}{n_1! n_2! \dots n_k!}$
wariacje	V_n^k	$\frac{n!}{(n-k)!}$
wariacje z powtórzeniami	\overline{V}_n^k	n^k
kombinacje	C_n^k	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$
kombinacje z powtórzeniami	\overline{C}_n^k	$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$

Rachunek prawdopodobieństwa

Rachunek prawdopodobieństwa jest dziedziną matematyki zajmującą się badaniem doświadczeń losowych, czyli zdarzeń, których wyniki są trudne do przewidzenia.

Przestrzeń zdarzeń elementarnych

Przestrzenią zdarzeń elementarnych, nazywamy zbiór wszystkich możliwych wyników doświadczenia losowego. Przestrzeń tę oznaczamy Ω . Poszczególne elementy przestrzeni zdarzeń elementarnych nazywamy zdarzeniami elementarnymi i oznaczamy ω_i .

Przestrzeń zdarzeń elementarnych - przykład

- ❶ Pojedynczy rzut kostką: $\Omega = \{1, 2, 3, 4, 5, 6\}$ $\omega_1 = 1$ itd.
- ❷ Trzykrotny rzut monetą:

$$\Omega = \{(o, o, o), (o, o, r), (o, r, o), (o, r, r), (r, o, o), (r, o, r), (r, r, o), (r, r, r)\}$$

$$\omega_1 = (o, o, o) \text{ itd.}$$

Zdarzenia losowe

Zdarzeniami losowymi nazywamy podzbiory przestrzeni zdarzeń elementarnych. Zdarzenia losowe oznaczamy dużymi literami.

Zdarzenia losowe - przykład

- ① Zdarzenie A - w pojedynczym rzucie monetą wypadła liczba oczek większa od 2 $A = \{3, 4, 5, 6\}$
- ② Zdarzenie B - w trzech rzutach monetą wypadł co najmniej dwa razy orzeł

$$B = \{(o, o, o), (o, o, r), (o, r, o), (r, o, o)\}$$

Działania na zbiorach (zdarzeniach)

- ❶ $A \cup B$ - suma zbiorów (zdarzeń) $x \in (A \cup B) = (x \in A) \vee (x \in B)$
- ❷ $A \cap B$ - iloczyn zbiorów (zdarzeń) $x \in (A \cap B) = (x \in A) \wedge (x \in B)$
- ❸ $A \setminus B$ - różnica zbiorów (zdarzeń) $x \in (A \setminus B) = (x \in A) \wedge (x \notin B)$
- ❹ $\Omega \setminus A = A' = \bar{A}$ - dopełnienie zbioru A, zdarzenie przeciwne do A
 $x \in A' = x \notin A$
- ❺ Ω - zdarzenie pewne
- ❻ \emptyset - zdarzenie niemożliwe
- ❼ $A \cap B = \emptyset$ - zdarzenia rozłączne
- ❽ $A \subset B$ - zdarzenie A pociąga za sobą zdarzenie B ($x \in A \Rightarrow x \in B$)

Definicja prawdopodobieństwa Laplace'a

Jeżeli spośród wszystkich $n = |\bar{\Omega}|$ jednakowo możliwych zdarzeń elementarnych $m = |\bar{A}|$ sprzyja zajściu zdarzenia losowego A , to liczbę $P[A] = \frac{m}{n} = \frac{|\bar{A}|}{|\bar{\Omega}|}$ nazywamy prawdopodobieństwem zajścia zdarzenia A .

Rzucono dwukrotnie kostką. Jakie jest prawdopodobieństwo, że suma wyrzuconych oczek jest nie większa do 4?

Ω - zbiór wszystkich par liczb z których każda jest równa 1,2,3,4,5 lub 6.

$\overline{\overline{\Omega}} = \overline{V}_6^2 = 36$, $A = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)\}$ $\overline{\overline{A}} = 6$. Zatem

$$P[A] = \frac{6}{36} = \frac{1}{6}$$

W urnie są 3 kule czarne i 5 białych. Jakie jest prawdopodobieństwo w 2 losowaniach, wylosowania 2 kul białych?

$$\overline{\overline{\Omega}} = C_8^2 = \binom{8}{2} = 28, \overline{\overline{A}} = C_5^2 = \binom{5}{2} = 10. \text{ Zatem } P[A] = \frac{10}{28}$$

Wady klasycznej definicji prawdopodobieństwa

- ① Tautologia - prawdopodobieństwo definiujemy przez prawdopodobieństwo ("jednakowo możliwych")
- ② $\overline{\Omega} = n$ musi być skończone
- ③ $\overline{A} = m$ często ciężko jest obliczyć

Geometryczna definicja prawdopodobieństwa

Geometryczna definicja prawdopodobieństwa

Jeżeli istnieje miara μ którą możemy zmierzyć zdarzenie A , oraz przestrzeń zdarzeń elementarnych Ω , to prawdopodobieństwo zajścia zdarzenia A jest równe $P[A] = \frac{\mu(A)}{\mu(\Omega)}$.

Linia telefoniczna zerwała się na trasie Puławy - Lublin (50 km). Oblicz prawdopodobieństwo, że zerwanie nastąpiło w odległości do 10 km od jednego z miast.

$$P[A] = \frac{\mu(A)}{\mu(\Omega)} = \frac{20}{50} = \frac{2}{5}.$$

Definicja prawdopodobieństwa von Misesa

Jeżeli częstość występowania zdarzenia losowego A oscyluje wokół pewnej liczby p i ta oscylacja maleje, to liczbę p nazywamy prawdopodobieństwem zdarzenia losowego A .

$$P[A] = p = \lim_{n \rightarrow \infty} \frac{m_n}{n}$$

m_n - liczba wystąpień zdarzenia

n - liczba prób losowych

Wady definicji von Misesa

- 1 Aby wyznaczyć p trzeba wykonać doświadczenie.
- 2 W definicji występuje granica (przy skończonej liczbie prób p nie jest wyznaczone jednoznacznie)

Definicja prawdopodobieństwa Kołmogorowa

Niech Ω będzie przestrzenią zdarzeń elementarnych, zaś $A, A_i \subset \Omega$, będą dowolnymi zdarzeniami losowymi. Prawdopodobieństwem nazywamy funkcję spełniającą warunki:

- 1 $0 \leq P[A] \leq 1$,
- 2 $P(\Omega) = 1$,
- 3 Jeżeli $A_1, A_2, \dots, A_n, \dots$ jest ciągiem zdarzeń parami rozłącznych, to

$$P[A_1 \cup A_2 \cup \dots \cup A_n \cup \dots] = \sum_{i=1}^{\infty} P[A_i]$$

(przeliczalna addytywność)

Własności prawdopodobieństwa

- ❶ $P[\emptyset] = 0$,
- ❷ Jeżeli $A \cap B = \emptyset$ to $P[A \cup B] = P[A] + P[B]$,
- ❸ Jeżeli $A \subset B$ to $P[B \setminus A] = P[B] - P[A]$,
- ❹ $P[A \cup B] = P[A] + P[B] - P[A \cap B]$,
- ❺ Jeżeli $A \subset B$, to $P[A] \leq P[B]$,
- ❻ $P[A'] = 1 - P[A]$,
- ❼ $P[A_1 \cup A_2 \cup \dots \cup A_n \cup \dots] \leq P[A_1] + P[A_2] + \dots + P[A_n] + \dots$
(subaddytywność miary)

Własności prawdopodobieństwa - przykład

W urnie są 3 kule czarne i 5 białych. Jakie jest prawdopodobieństwo w 2 losowaniach, wylosowania co najmniej 1 kuli czarnej?

A - wylosowano co najmniej 1 kulę czarną A' - wylosowano 2 kule białe

$$P[A] = 1 - P[A'] = 1 - \frac{10}{28} = \frac{18}{28}$$

Prawdopodobieństwo warunkowe

Prawdopodobieństwem zajścia zdarzenia A pod warunkiem, że wiemy, że zaszło zdarzenie B nazywamy liczbę

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

o ile $P[B] > 0$.

Prawdopodobieństwo warunkowe - zadanie

Rzucamy dwukrotnie kostką. Obliczyć prawdopodobieństwo tego, że wypadnie suma oczek mniejsza niż 4, o ile w pierwszym rzucie otrzymaliśmy 1.

A - zdarzenie polegające na wyrzuceniu w dwóch rzutach sumy oczek mniejszej od 4

B – zdarzenie polegające na wyrzuceniu w pierwszy rzucie 1 oczka

Wzór na prawdopodobieństwo warunkowe:

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

Wszystkich zdarzeń elementarnych jest:

$$\overline{\Omega} = 6^2 = 36$$

Wypiszmy zdarzenia sprzyjające zdarzeniu B:

$$B = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)\}, \quad \overline{\overline{B}} = 6$$

$$P[B] = \frac{\overline{\overline{B}}}{\overline{\overline{\Omega}}} = \frac{6}{36} = \frac{1}{6}$$

Wypiszmy zdarzenia sprzyjające zdarzeniu $A \cap B$:

$$A \cap B = \{(1, 1), (1, 2)\}, \quad \overline{\overline{A \cap B}} = 2$$

$$P[A \cap B] = \frac{\overline{\overline{A \cap B}}}{\overline{\overline{\Omega}}} = \frac{2}{36} = \frac{1}{18}$$

Wstawiając do wzoru na prawdopodobieństwo warunkowe mamy:

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{\frac{1}{18}}{\frac{1}{6}} = \frac{1}{3}$$

Wzór na prawdopodobieństwo całkowite i wzór Bayesa

Wzór na prawdopodobieństwo całkowite i wzór Bayesa

Jeśli $\{H_i, 1 \leq i \leq n\}$ jest zbiorem zdarzeń elementarnych takich, że:

① $H_i \cap H_j = \emptyset$ dla $i \neq j, 1 \leq i, j \leq n$,

② $\Omega = \bigcup_{i=1}^n H_i$,

to dla dowolnego zdarzenia A

$$P[A] = \sum_{i=1}^n P[A|H_i]P[H_i]$$

oraz

$$P[H_i|A] = \frac{P[A|H_i]P[H_i]}{P[A]} = \frac{P[A|H_i]P[H_i]}{\sum_{j=1}^n P[A|H_j]P[H_j]}$$

Wzór Bayesa - zadanie

Dane są dwie urny z kulami: urna A zawierająca 6 czarnych i 9 białych kul i urna B o zawartości 5 czarnych i 15 białych kul. Wylosowano białą kulę.

Jakie jest prawdopodobieństwo, że pochodzi ona z urny A?

A - zdarzenie polegające na wylosowaniu kuli białej

H_1 - zdarzenie polegające na wylosowaniu urny A

H_2 - zdarzenie polegające na wylosowaniu urny B

Mamy policzyć prawdopodobieństwo $P[H_1|A]$. Ze wzoru Bayesa:

$$P[H_1|A] = \frac{P[A|H_1] \cdot P[H_1]}{P[A]}$$

Do mianownika stosujemy wzór na prawdopodobieństwo całkowite:

$$P[A] = P[A|H_1] \cdot P[H_1] + P[A|H_2] \cdot P[H_2]$$

Wzór Bayesa - zadanie c.d.

Liczymy prawdopodobieństwa:

$$P[H_1] = \frac{1}{2}, \quad P[A|H_1] = \frac{9}{15}$$

$$P[H_2] = \frac{1}{2}, \quad P[A|H_2] = \frac{15}{20}$$

Wstawiamy do wzoru na prawdopodobieństwo całkowite:

$$P[A] = \frac{9}{15} \cdot \frac{1}{2} + \frac{15}{20} \cdot \frac{1}{2} = \frac{3}{10} + \frac{15}{40} = \frac{12 + 15}{40} = \frac{27}{40}$$

i wstawiamy do wzoru Bayesa:

$$P[H_1|A] = \frac{P[A|H_1] \cdot P[H_1]}{P[A]} = \frac{\frac{9}{15} \cdot \frac{1}{2}}{\frac{27}{40}} = \frac{\frac{3}{10}}{\frac{27}{40}} = \frac{4}{9}$$

Niezależność zdarzeń

Zdarzenia **A** i **B** nazywamy niezależnymi, jeśli

$$P[A \cap B] = P[A]P[B]$$

Niezależność zdarzeń - przykład

Czy zdarzenia A - na kostce wypadła nieparzysta liczba oczek i B - Wypadły co najmniej 4 oczka są niezależne?

$$P[A] = 0.5, \quad P[B] = 0.5, \quad P[A \cap B] = \frac{1}{6}$$

$$\frac{1}{6} \neq 0.5 \cdot 0.5,$$

zatem te zdarzenia nie są niezależne.

Zmienna losowa

Zmienną losową nazywamy dowolną funkcję $X : \Omega \rightarrow \mathbb{R}$ przyporządkowującą każdemu elementowi przestrzeni zdarzeń elementarnych Ω pewną liczbę rzeczywistą.

Zmienne losowe oznaczamy zwykle dużymi literami z końca alfabetu (X, Y, Z) i zamiast pisać $X(\omega)$, często piszemy po prostu X pomijając argument ω . Prawdopodobieństwo, że zmienna losowa X przyjmuje wartości z pewnego zbioru A oznaczamy przez $P[\omega : X(\omega) \in A] = P[X \in A]$.

- 1 Rzucamy kostką do gry. Uzyskanym wynikom przypiszmy funkcję która przyjmuje wartość 1 dla wyników nieparzystych i 0 dla parzystych.
 $X(\{2\}) = X(\{4\}) = X(\{6\}) = 0$ i $X(\{1\}) = X(\{3\}) = X(\{5\}) = 1$.
Takie przyporządkowanie jest zmienną losową.
- 2 Rzucamy dwukrotnie monetą i jeśli wypadną 2 orły otrzymujemy 10 zł w jeśli wypadną 2 reszki tracimy 5 zł, natomiast gdy wypadnie po jednym orle i reszce tracimy 3zł.
Funkcja przypisująca uzyskanemu wynikowi wygraną jest zmienną losową. $X(\{o, o\}) = 10$, $X(\{r, r\}) = -5$ zaś
 $X(\{o, r\}) = X(\{r, o\}) = -3$.

Zmienne losowe *przenoszą* nas z przestrzeni zdarzeń elementarnych Ω do zbioru \mathbb{R} , który z matematycznego punktu widzenia jest "wygodniejszy".

Rodzaje zmiennych losowych

Istnieją 3 wzorcowe rodzaje zmiennych losowych:

- ➊ Zmienne losowe dyskretne (skokowe)
- ➋ Zmienne losowe ciągłe
 - ➊ Zmienne losowe bezwzględnie ciągłe
 - ➋ Zmienne losowe osobliwe

Dystrybuanta

Dystrybuantą zmiennej losowej X nazywamy funkcję $F : \mathbb{R} \rightarrow [0, 1]$ spełniającą następujące warunki:

- 1 $\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow +\infty} F(x) = 1,$
- 2 niemalejąca $x < y, \quad F(x) \leq F(y),$
- 3 lewostronnie ciągła $\lim_{x \rightarrow x_0^-} F(x) = F(x_0)$

Dystrybuanta jednoznacznie wyznacza prawdopodobieństwo tego, że zmienna losowa przyjmuje wartości z konkretnego przedziału:

- ① $F(x) = P[X < x]$
- ② $P[X \geq x] = 1 - F(x),$
- ③ $P[a \leq X < b] = F(b) - F(a)$
- ④ $P[X = c] = \lim_{x \rightarrow c^+} F(x) - F(c)$

Jednoznaczność dystrybuanty

Każda zmienna losowa ma swoją dystrybuantę i każda dystrybuanta jednoznacznie wyznacza zmienną losową.

Twierdzenie Lebesgue'a

Twierdzenie Lebesgue'a

Każda dystrybuanta F daje się jednoznacznie przedstawić w postaci:

$$F(x) = a_1 F_d(x) + a_2 F_c(x) + a_3 F_o(x),$$

gdzie $a_1 + a_2 + a_3 = 1, 0 \leq a_i \leq 1, i = 1, 2, 3$ a F_d, F_c, F_o są dystrybuantami zmiennych losowych odpowiednio dyskretnej, bezwzględnie ciągłej i osobliwej. W tym sensie każda zmienna losowa ma część dyskretną, bezwzględnie ciągłą i osobliwą.

Dyskretne zmienne losowe

Zmienna losowa jest dyskretna (ma rozkład dyskretny), jeśli jest skupiona na co najwyżej zbiorze przeliczalnym (może być skończony) swoich atomów (wartości które może przyjmować).

Jeśli dana zmienna losowa może przyjmować wartości x_i odpowiednio z prawdopodobieństwami p_i , ($P[X = x_i] = p_i$) to musi być spełniony warunek $\sum_{i=1}^n p_i = 1$, dla zmiennej losowej prostej i $\sum_{i=1}^{\infty} p_i = 1$, dla zmiennej losowej dyskretnej o przeliczalnym zbiorze wartości.

Sposoby określania prawdopodobieństwa dla dyskretnych zmiennych losowych:

- 1 Za pomocą tabeli
- 2 Za pomocą wzoru $P[X = x_i] = f(x_i)$
- 3 Za pomocą dystrybuantry

Sposoby określania prawdopodobieństwa zmiennej losowej typu dyskretnego

Wracając do przykładu 2 ze slajdu 34 możemy tę zmienną losową określić:

1 Gęstość

x_i	10	-5	-3
p_i	0.25	0.25	0.5

2 Dystrybuanta

$$F(x) = \begin{cases} 0 & x \in (-\infty, -5], \\ 0.25 & x \in (-5, -3], \\ 0.75 & x \in (-3, 10], \\ 1 & x \in (10, +\infty) \end{cases}$$

Dystrybuanta zmiennej losowej skokowej jest nieciągła.

Zmienna losowa ciągła (bezwzględnie ciągła)

Zmienna losowa jest ciągła (ma rozkład ciągły), jeśli istnieje nieujemna funkcja rzeczywista $f : \mathbb{R} \rightarrow \mathbb{R}$ (zwana gęstością) taka, że dla każdego $x \in \mathbb{R}$ mamy $P[X < x] = \int_{-\infty}^x f(t)dt$.

Funkcja gęstości musi spełniać warunki:

- 1 $f(x) \geq 0$,
- 2 $\int_{-\infty}^{\infty} f(x)dx = 1$

Dystrybuanta zmiennej losowej typu ciągłego jest ciągła.

Funkcję gęstości można wyznaczyć z dystrybuanty: $f(x) = F'(x)$.

Przykład ciągłej zmiennej losowej - rozkład jednostajny $U(0, 1)$

Losujemy punkt z odcinka $[0, 1]$ i X jest wylosowanym punktem.

Zauważmy, że $P[X = x] = 0$ dla każdego $x \in [0, 1]$. Gdyby bowiem było $P[X = x_0] = p > 0$ dla pewnego x_0 , to z symetrii musiałoby też być $P[X = x] = p$ dla każdego x (dlaczego x_0 miałby być bardziej prawdopodobny?), co prowadzi do sprzeczności z $P[X \in [0, 1]] = 1$.

Zmienna losowa X jest więc zmienną typu ciągłego a z symetrii jej funkcja gęstości musi spełniać warunek $f(x) = c$ dla $x \in [0, 1]$ i $f(x) = 0$ dla $x \notin [0, 1]$. Ponieważ

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^1 f(x) dx = c = 1$$

więc

$$f(x) = \begin{cases} 0, & \text{jeśli } x \notin [0, 1], \\ 1, & \text{jeśli } x \in [0, 1], \end{cases}$$
$$F(x) = \begin{cases} 0, & \text{jeśli } x \in (-\infty, 0], \\ x, & \text{jeśli } x \in (0, 1], \\ 1, & \text{jeśli } x \in (1, \infty), \end{cases}$$

Zmienna losowa osobliwa

Zmienna losowa jest osobliwa (ma rozkład osobliwy), jeśli jest skupiona na nieprzeliczalnym zbiorze o mierze 0 i nie ma atomów (tzn. punktów takich, że $P[X = x] > 0$).

Dystrybuanta zmiennej osobliwej jest ciągła jednak nie istnieje gęstość (precyzyjniej - istnieje, jest równa zero wszędzie, poza zbiorem miary 0, matematycznie - nie jest mierzalna wg. miary Lebesgue'a) Przykładem takiej dystrybuanty jest zbiór Cantora.

Funkcja Cantora

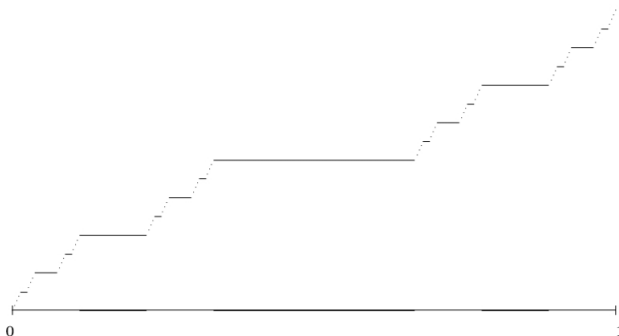
Skonstruujemy zmienną X taką, że:

- $P[X = x] = 0$ dla każdego $x \in \mathbb{R}$ (czyli X nie ma części dyskretnej),
- na żadnym przedziale $[a, b]$ dla którego $P[X \in [a, b]] > 0$, nie da się zdefiniować funkcji $f : [a, b] \rightarrow \mathbb{R}$ takiej, że $P[X \in I] = \int_I f(t)dt$ dla każdego przedziału $I \subseteq [a, b]$ (a więc X nie ma gęstości na żadnym przedziale o niezerowym prawdopodobieństwie, czyli nie ma części ciągłej)

Niech $F()$ będzie funkcją taką, że $F(x) = 0$ dla $x < 0$ i $F(x) = 1$ dla $x > 1$ natomiast na przedziale $[0, 1]$ definiujemy F następująco

- Dzielimy przedział $I = [0, 1]$ na trzy części $I = [0, 1/3] \cup [1/3, 2/3] \cup [2/3, 1]$ i przyjmujemy, że $F(x) = 1/2$ dla $x \in [1/3, 2/3]$
- Dzielimy każdy z pozostałych przedziałów $[0, 1/3] = [0, 1/9] \cup [1/9, 2/9] \cup [2/9, 1/3]$ oraz $[2/3, 1] = [2/3, 7/9] \cup [7/9, 8/9] \cup [8/9, 1]$ i przyjmujemy, że $F(x) = 1/4$ dla $x \in [1/9, 2/9]$ oraz $F(x) = 3/4$ dla $x \in [7/9, 8/9]$.
- Podobnie postępujemy z każdym z dalszych odcinków

Funkcja Cantora



Niech X będzie zmienną losową o dystrybucji F . Łatwo pokazać, że F jest ciągła, a zatem dla każdego x zachodzi $P[X = x] = 0$. Można też pokazać (co jest nieco trudniejsze), że F_X ma zerową pochodną wszędzie poza zbiorem miary zero (jest to tzw. zbiór Cantora), a zatem nie ma gęstości na żadnym przedziale $[a, b]$ dla którego $P[[a, b]] > 0$.

Nie będziemy się w dalszym wykładzie zajmować tego typu zmiennymi losowymi.

Wspólne określenie zmiennych ciągłych i dyskretnych

Problem. Czy da się połączyć określenie zmiennej losowej za pomocą rozkładu $\{(x_i, p_i), i = 1, 2, 3, \dots\}$ zmiennych losowych typu dyskretnego z przedstawieniem za pomocą funkcji gęstości f zmiennych typu ciągłego?

Jak powinna wyglądać funkcja gęstości dla rozkładu dyskretnego?

$$f(x) = \begin{cases} 0, & x \notin \{x_i, i = 1, 2, 3, \dots\}, \\ p_i, & i \in \mathcal{N} \end{cases}$$

Całka Lebesgue'a-Stieltjesa

Problem. Jak powinna wyglądać całka określająca dystrybucję dla zmiennych losowych typu dyskretnego?

Jeśli punkty x_i są uporządkowane tak, że $x_1 \leq x_2 \leq x_k \leq x \leq x_{k+1} \leq \dots$ to

$$\begin{aligned} F(x) &= \int_{-\infty}^x f(t) dt = \int_{-\infty}^{x_1} f(t) dt + f(x_1) \\ &\quad + \sum_{i=2}^k \int_{x_{i-1}}^{x_i} f(t) dt + \sum_{i=2}^k f(x_i) + \int_{x_k}^x f(t) dt \end{aligned} \quad (1)$$

Oczywiście wtedy dla zmiennych typu skokowego mamy wtedy

$$F(x) = \sum_{\{i: x_i < x\}} f(x_i) = \sum_{\{i: x_i < x\}} p_i$$

jednak wzór (1) jest uniwersalny (dla zmiennych losowych będących mieszaniną zmiennych typu ciągłego i skokowego. Tak rozumianą całkę nazywamy całką Riemanna-Stieltjesa.

Charakterystyki liczbowe zmiennych losowych

Nazwa	Wzór	dla zm. los. skokowych	ogólnie
Wartość oczekiwana	EX	$\sum_{i=1}^{\infty} x_i p_i$	$\int_{-\infty}^{\infty} t f(t) dt$
Wariancja	DX	$\sum_{i=1}^{\infty} (x_i - EX)^2 p_i$	$\int_{-\infty}^{\infty} (t - EX)^2 f(t) dt$
Odchylenie standardowe	ρX		\sqrt{DX}
Moment rzędu k	EX^k	$\sum_{i=1}^{\infty} x_i^k p_i$	$\int_{-\infty}^{\infty} t^k f(t) dt$
Moment centralny rzędu k	$E(X - EX)^k$	$\sum_{i=1}^{\infty} (x_i - EX)^k p_i$	$\int_{-\infty}^{\infty} (t - EX)^k f(t) dt$
Moda/Dominanta	M	$\{x_i : p_i = \sup_{j \in \mathcal{N}} p_j\}$	$\{x : f(x) = \sup_{y \in \mathcal{R}} f(y)\}$
Mediana	Me	$\{x : P[X \leq x] \geq \frac{1}{2} \wedge P[X \geq x] \geq \frac{1}{2}\}$	
Kwantyl rzędu p	Q_p	$\{x : P[X \leq x] \geq p \wedge P[X \geq x] \geq 1 - p\}$	
Odstęp międzykwartylowy	IRQ	$Q_{0.75} - Q_{0.25}$	
Skośność	SKE	$\frac{E(X-EX)^3}{D^{3/2} X}$	
Kurtoza	KRT	$\frac{E(X-EX)^4}{D^2 X} - 3$	

O ILE TE WARTOŚCI ISTNIEJĄ!!!

Mamy: $DX = E(X - EX)^2 = EX^2 - (EX)^2$, $Me = Q_{0.5}$.



Dyskretny rozkład jednostajny $U(\{1, 2, \dots, n\})$

$$P[X = i] = \frac{1}{n}, \quad i = 1, 2, \dots, n,$$

$$f(x) = \begin{cases} \frac{1}{n}, & x \in \{1, 2, \dots, n\} \\ 0, & \text{poza tym} \end{cases}$$

$$F(x) = \begin{cases} 0, & x < 1, \\ \frac{\lfloor x \rfloor}{n}, & 1 \leq x \leq n, \\ 1, & x > n, \end{cases}$$

$$EX = \frac{n+1}{2},$$

$$DX = \frac{n^2 - 1}{12},$$

Ciągły rozkład jednostajny $U([a, b])$

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

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x \leq b, \\ 1, & x > b, \end{cases}$$

$$EX = \frac{a+b}{2},$$

$$DX = \frac{(b-a)^2}{12},$$

Rozkład Bernoulliego $B(n, p)$

$$P[X = k] = \frac{n}{k} p^k q^{n-k}, k = 0, 1, 2, \dots, n,$$

$$f(x) = \begin{cases} \frac{n}{x} p^x q^{n-x}, & x \in \{0, 1, 2, \dots, n\}, \\ 0, & \text{poza tym} \end{cases}$$

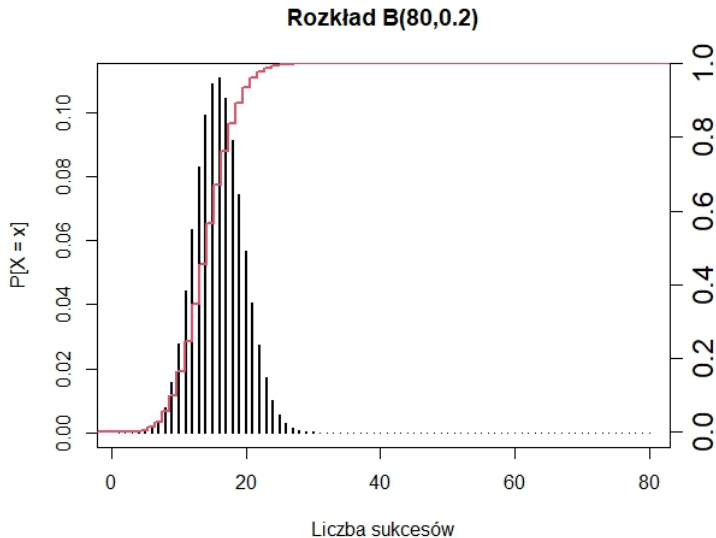
$$F(x) = \begin{cases} 0, & x < 0, \\ \sum_{k < x} \frac{n}{k} p^k q^{n-k}, & x \in [0, n], \\ 1, & n \leq x, \end{cases}$$

$$EX = np,$$

$$DX = npq,$$

gdzie $q = 1 - p, 0 < p < 1$.

Rozkład $B(80, 0.2)$



Rozkład geometryczny $G(p)$

$$P[X = k] = q^k p, k = 0, 1, 2, \dots, n,$$

$$f(x) = \begin{cases} q^x p, & x \in \{0, 1, 2, \dots, n\}, \\ 0, & \text{poza tym} \end{cases}$$

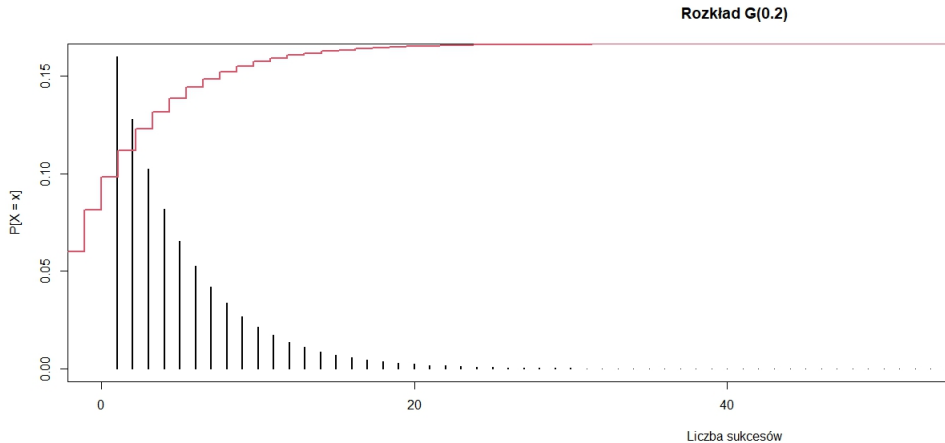
$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - q^{\lceil x \rceil}, & 0 \leq x, \end{cases}$$

$$EX = \frac{1}{p},$$

$$DX = \frac{1 - p}{p^2},$$

gdzie $q = 1 - p, 0 < p < 1$.

Rozkład $G(0.2)$



Rozkład Poissona $Pois(\lambda)$

$$P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, \dots$$

$$f(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x \in \{0, 1, 2, \dots, n\}, \\ 0, & \text{poza tym} \end{cases}$$

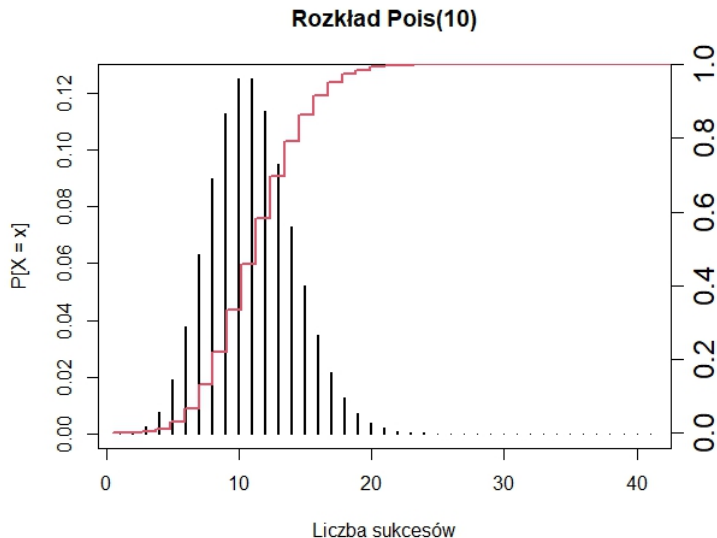
$$F(x) = \begin{cases} 0, & x < 0, \\ \sum_{i: 0 \leq i \leq x} \frac{\lambda^i e^{-\lambda}}{i!}, & 0 \leq x, \end{cases}$$

$$EX = \lambda,$$

$$DX = \lambda,$$

gdzie $\lambda > 0$.

Rozkład $Pois(10)$



Rozkład wykładniczy $Exp(\lambda)$

$$f(x) = \begin{cases} 0, & x < 0, \\ \lambda e^{-\lambda x}, & x \geq 0, \end{cases}$$

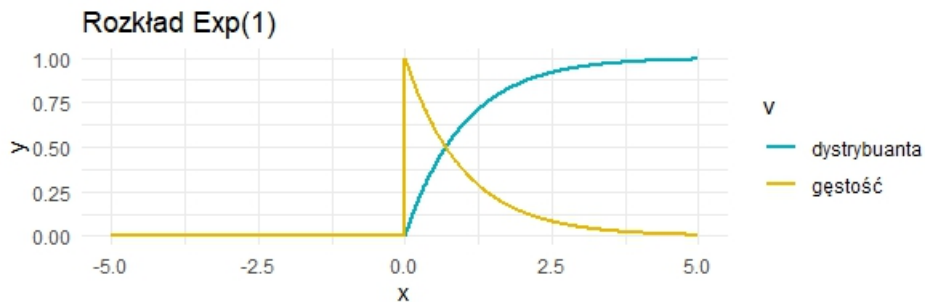
$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & 0 \leq x, \end{cases}$$

$$EX = \frac{1}{\lambda},$$

$$DX = \frac{1}{\lambda^2},$$

gdzie $\lambda > 0$.

Rozkład $Exp(1)$



Rozkład normalny $N(\mu, \sigma)$

$$f(x) = \phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

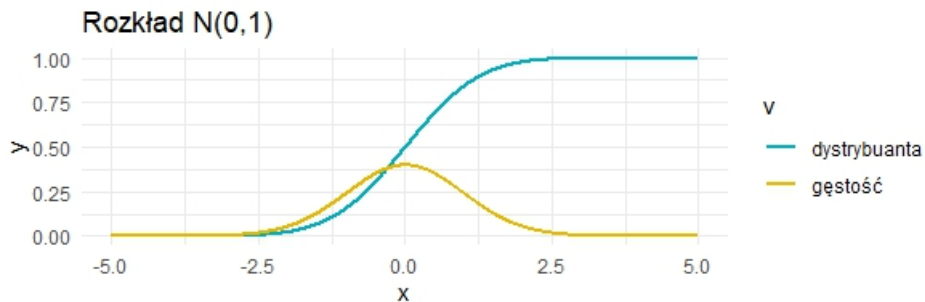
$$F(x) = \Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x t e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt,$$

$$EX = \mu,$$

$$DX = \sigma^2,$$

gdzie $\sigma > 0$.

Rozkład $N(0, 1)$



Standaryzacja rozkładu normalnego

Jeśli zmienna losowa X ma rozkład $N(\mu, \sigma)$, to zmienna losowa

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

ma standardowy rozkład normalny $N(0, 1)$.

Standaryzacja - przykład

W pewnej społeczności średni wzrost jest równy 175 cm, a jego odchylenie standardowe 10 cm. Zakładając, że wzrost ma rozkład normalny, oblicz jakie jest prawdopodobieństwo spotkania osoby o wzroście między 170 a 175 cm.

Jest $X \sim N(175, 10)$ to $Y = \frac{X-175}{10} \sim N(0, 1)$ więc dla $x = 175$ mamy $y = 0$, a dla $x = 170$ mamy $y = -0.5$ Stąd z tablic rozkładu normalnego

$$\begin{aligned} P[170 \leq X \leq 175] &= \Phi_{175,10}(175) - \Phi_{175,10}(170) \\ &= \Phi(0) - \Phi(-0.5) \\ &= 0.5 - 0.3085 = 0.1915 \end{aligned}$$

Typy zbieżności w rachunku prawdopodobieństwa

Rodzaje zbieżności

Zbieżność prawie pewna (silna, z prawdopodobieństwem 1) $X_n \xrightarrow{a.s.} X, n \rightarrow \infty$

$$P[\{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}] = 1$$

Zbieżność wg. prawdopodobieństwa $X_n \xrightarrow{P} X, n \rightarrow \infty$

$$\bigwedge_{\epsilon > 0} \lim_{n \rightarrow \infty} P[\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}] = \lim_{n \rightarrow \infty} P[|X_n - X| > \epsilon] = 0.$$

Zbieżność wg. rozkładu (słaba) $X_n \xrightarrow{w} X, n \rightarrow \infty$ Dla każdego $x \in \mathfrak{R}$ dla którego $F(x) = P[X \leq x]$ jest ciągła zachodzi

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

gdzie $F_n(x) = P[X_n \leq x]$.

Rodzaje zbieżności

Niech $\{X, X_n, n \geq 1\}$ będą zmiennymi losowymi. Wtedy

$$(1) \quad X_n \xrightarrow{a.s.} X \quad \implies \quad X_n \xrightarrow{P} X$$

$$(2) \quad X_n \xrightarrow{P} X \quad \implies \quad X_n \xrightarrow{w} X$$

Dowód (1). Niech $A = \{\omega : X_n(\omega) \rightarrow X(\omega), n \rightarrow \infty\}$. Z założenia $P[A] = 1$. Ustalmy $\epsilon > 0$. Wtedy dla każdego $N \in \mathcal{N}$ mamy $A \subset A_\epsilon$ gdzie $A_\epsilon = \bigcup_{i=N+1}^{\infty} A_{\epsilon,i}$ gdzie $A_{\epsilon,i} = \{\omega : |X_j(\omega) - X(\omega)| < \epsilon, \text{ dla każdego } j \geq i\}$. Wtedy

$$\begin{aligned} 1 = P[A] &= \lim_{N \rightarrow \infty} P[A_{\epsilon,N}] \\ &\leq \lim_{N \rightarrow \infty} P[|X_N - X| \leq \epsilon] = 1 - \lim_{N \rightarrow \infty} P[|X_N - X| > \epsilon] = 1. \end{aligned}$$

Dowód (2). Niech a będzie punktem ciągłości dystrybuanty F i niech $\epsilon > 0$ będzie ustalone. Dla każdego $n \in \mathbb{N}$ zachodzi

$$\{X \leq a - \epsilon\} \subset \{|X_n - X| \geq \epsilon\} \cup \{X_n \leq a\}$$

oraz

$$\{X_n \leq a\} \subset \{|X_n - X| \geq \epsilon\} \cup \{X \leq a + \epsilon\}$$

zatem

$$F_X(a - \epsilon) \leq P[|X_n - X| \geq \epsilon] + F_{X_n}(a) \leq 2P[|X_n - X| \geq \epsilon] + F_X(a + \epsilon)$$

czyli

$$F_X(a - \epsilon) \leq \liminf_n F_{X_n}(a) \leq \limsup_n F_{X_n}(a) \leq F_X(a + \epsilon)$$

i przechodząc z ϵ do 0 oraz korzystając z ciągłości F_X w punkcie a otrzymujemy tezę.

Przykład

Niech $X, Y \sim B(100, 0.5)$ będą dwiema niezależnymi zmiennymi losowymi. Wtedy ciąg $X_n = X, n \geq 1$ jest oczywiście $X_n \xrightarrow{d} Y \sim X$ jednak

$$P[|X_n - Y| \geq 1] = P[|X - Y| \geq 1] = \frac{1}{2}$$

i ciąg ten nie jest zbieżny wg. pdp.

Przykład

Niech naszą przestrzenią probabilistyczną będzie $([0, 1], \mathcal{B}([0, 1]), P)$ gdzie P - miara Lebesgue'a na $[0, 1]$. Zdefiniujmy tablicę

$$X_{k,l} = \begin{cases} 1, & \text{dla } \frac{l-1}{2^k} < \omega \leq \frac{l}{2^k} \\ 0, & \text{dla pozostałych } \omega \end{cases}$$

i ustawmy tę tablicę w ciąg $Y_n = X_{n+1-2^{\lfloor \log_2 n \rfloor}, 2^{\lfloor \log_2 n \rfloor}}$ tak więc ponieważ $P[|Y_n| > \frac{1}{2^{\lfloor \log_2 n \rfloor}}] = 0$ więc dla n takich że $\epsilon > \frac{1}{2^{\lfloor \log_2 n \rfloor}}$ mamy $P[|Y_n| > \epsilon] = 0$ więc $Y_n \xrightarrow{P} 0, n \rightarrow \infty$. Jednak ponieważ dla każdego $\omega \in [0, 1]$ granica $X_n(\omega)$ nie istnieje, więc $Y_n \not\xrightarrow{\text{a.s.}} 0, n \rightarrow \infty$.

Funkcja charakterystyczna

Funkcja charekterystyczna

Funkcją charakterystyczną dowolnej (nie zawierającej składnika osobliwego) zmiennej losowej X jest

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} dF(x) = \int_{-\infty}^{\infty} \cos(tx) dF(x) + i \int_{-\infty}^{\infty} \sin(tx) dF(x), t \in \mathbb{R}.$$

- 1 Funkcja charakterystyczna jednoznacznie wyznacza rozkład zmiennej losowej.
- 2 $\phi_X(0) = 1$, $\phi'_X(0) = iEX$, $\phi''_X(0) = -DX$, o ile te wartości istnieją.
- 3 $\phi_{N(0,1)}(t) = e^{-t^2/2}$
- 4 Zbieżność $F_{X_n}(x) \rightarrow F_X(x)$ w każdym punkcie ciągłości \times dystrybucyjności F_X jest równoważna zbieżności $\phi_{X_n}(t) \rightarrow \phi_X(t)$ jednostajnej na zwartych podzbiorach \mathbb{R} .
- 5 Jeżeli $\{X_i, i \geq 1\}$ są niezależnymi zmiennymi losowymi to $\phi_{X_1+X_2+\dots+X_n}(t) = \prod_{i=1}^n \phi_{X_i}(t)$.
- 6 Jeżeli X jest zmienną losową dla której DX istnieje to

$$\phi_X(t) = \phi_X(0) + t\phi'_X(0) + \frac{t^2}{2}\phi''_X(0) + r(t^t),$$

gdzie $r(t)$ jest funkcją taką, że $\frac{r(t)}{t} \rightarrow 0$ gdy $t \rightarrow 0$.

Centralne Twierdzenie Graniczne

Jeżeli $\{X_n, n \geq 1\}$ jest ciągiem niezależnych zmiennych losowych takich, pochodzącymi z tej samej populacji o wartości oczekiwanej μ oraz dodatniej i skończonej wariancji σ^2 to ciąg zmiennych losowych, w postaci znormalizowanych wartości oczekiwanych U_n :

$$U_n = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)}{\sigma / \sqrt{n}}$$

zbieżny jest według rozkładu do standardowego rozkładu normalnego $N(0, 1)$, gdy $n \rightarrow \infty$.

Dowód Centralnego Twierdzenia Granicznego

Z niezależności $\{X_n, n \geq 1\}$ mamy

$$\phi_{U_n}(t) = E e^{i \frac{t}{\sqrt{n}\sigma} \sum_{i=1}^n (X_i - \mu)} = \prod_{i=1}^n \phi_{X_i - \mu}\left(\frac{t}{\sqrt{n}\sigma}\right) = \phi_{X_1 - \mu}^n\left(\frac{t}{\sqrt{n}\sigma}\right)$$

a stąd

$$\begin{aligned} n \ln \phi_{U_n}(t) &= n \ln \phi_{X_1 - \mu}\left(\frac{t}{\sqrt{n}\sigma}\right) \\ &= n \ln\left(\phi_{X_1 - \mu}(0) + \frac{t}{\sqrt{n}\sigma} \phi'_{X_1 - \mu}(0) + \frac{1}{2} \frac{t^2}{n\sigma^2} \phi''_{X_1 - \mu}(0) + r\left(\frac{t^2}{n\sigma^2}\right)\right) \\ &= n \ln\left(1 - \frac{1}{2} \frac{t^2}{n} + r\left(\frac{t^2}{n\sigma^2}\right)\right) \\ &\approx n\left(-\frac{1}{2} \frac{t^2}{n} + r\left(\frac{t^2}{n\sigma^2}\right)\right) \\ &\rightarrow -\frac{t^2}{2} \end{aligned}$$

a to jest funkcja charakterystyczna rozkładu $N(0, 1)$.

Centralne Twierdzenie Graniczne - przykład

Czas kontroli jednego elementu podlega rozkładowi jednostajnemu na przedziale od 10 do 16 sekund. Oszacuj prawdopodobieństwo, że czas kontroli 100 elementów przekroczy 22 minuty.

$$\begin{aligned}\mu &= EX = \frac{a+b}{2} = \frac{10+16}{2} = 13, \\ \sigma^2 &= \frac{(b-a)^2}{12} = \frac{36}{12} = 3, \quad \sigma = \sqrt{3} \approx 1,732\end{aligned}$$

Z Centralnego Twierdzenia Granicznego mamy $\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{\mathcal{D}} N(0, 1)$ czyli

$$P\left[\sum_{i=1}^{100} X_i > 22 \cdot 60\right] = P\left[\frac{\sum_{i=1}^{100} X_i - 1300}{10 \cdot 1,732} > \frac{1320 - 1300}{17,32}\right] \approx P[N(0, 1) > 1,15]$$

Stąd

$$P\left[\sum_{i=1}^{100} X_i > 22 \cdot 60\right] \approx P[N(0, 1) > 1,15] = 1 - \Phi(1,15) = 1 - 0,8749 = 0,1251$$

Mocne Prawo Wielkich Liczb

Jeżeli $\{X_n, n \geq 1\}$ jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie takim, że $E|X_1|$ istnieje, to

$$\frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{p.p.} EX_1,$$

gdy $n \rightarrow \infty$ gdzie $Y_n \xrightarrow{p.p.} Y$ oznacza, że dla pewnego zdarzenia $A \in \mathcal{A}$ takiego, że $P[A] = 1$ mamy $Y_n(\omega) \rightarrow Y(\omega)$ dla każdego $\omega \in A$.

Badano czasy reakcji elementu technicznego w konstrukcji otrzymując wartości 1.2, 1.7, 2.1, 1.5, 1.2, 1.3, 1.2, 1.5, 1.1, 1.2 Jak można przybliżyć wartość oczekiwaną czasu reakcji tego elementu?

Z Mocnego Prawa Wielkich Liczb

$$\frac{X_1(\omega) + \dots + X_{10}(\omega)}{10} = \frac{1.2 + 1.7 + 2.1 + \dots + 1.2}{10} = 1.4$$

przybliża EX_1 o ile oczywiście ω nie należy do zbioru $\Omega \setminus A$.

Niezależność

Mówimy, że zmienne losowe X i Y są niezależne, jeżeli dla każdego dwóch zbiorów A i B zdarzenia $[X \in A]$ oraz $[Y \in B]$ są niezależne, tzn.

$$P[X \in A, Y \in B] = P[X \in A]P[Y \in B].$$

Jeżeli zmienne losowe X i Y są niezależne, to

$$\begin{aligned}f_{(X,Y)}(x,y) &= f_X(x)f_Y(y), \\F_{(X,Y)}(x,y) &= F_X(x)F_Y(y), \\EXY &= EXEY\end{aligned}$$

Korelacja

Kowariancją zmiennych losowych X i Y nazywamy wartość

$$\text{Cov}(X, Y) = E[XY] - [EX][EY].$$

Korelacją (Pearsona) zmiennych losowych X i Y to

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

$$-1 \leq \rho_{X,Y} \leq 1$$

$$\rho_{X,Y} = 1 \iff X = aY + b, a > 0, b \in \mathbb{R},$$

$$\rho_{X,Y} = -1 \iff X = aY + b, a < 0, b \in \mathbb{R},$$

$$\rho_{X,Y} = 0 \iff X, Y \text{ są niezależne,}$$

$$\rho_{aX+b,Y} = \rho_{X,Y}, a > 0, b \in \mathbb{R}$$

Korelacja - Przykład

Niech zmienne losowe X i Y będą zadane tabelką:

$x \backslash y$	-3	0	
-2	0.3	0.3	0.6
0	0	0.3	0.3
2	0.1	0	0.1
	0.4	0.6	1

wtedy

$$EXY = (-2) * (-3) * 0.3 + (-3) * 2 * 0.1 = 1.2$$

$$EX = (-2) * 0.6 + 2 * 0.1 = -1$$

$$EY = (-3) * 0.4 = -1.2$$

$$\text{Cov}(X, Y) = 1.2 - (-1) * (-1.2) = 0$$

$$P[X = -2, Y = -3] = 0.3 \neq 0.24 = 0.4 * 0.6 = P[X = -2] * P[Y = -3]$$

Regresja

Prostą regresji zmiennej losowej Y względem X nazywamy prostą $y = ax + b$ taką, że wartość oczekiwana $E(Y - aX - b)^2$ osiąga wartość minimalną.

- Równanie prostej regresji Y względem X ma postać:

$$\frac{y - EY}{\sigma Y} = \rho_{X,Y} \frac{x - EX}{\sigma X}$$

czyli $a = \rho_{X,Y} \frac{\sigma Y}{\sigma X}$, $b = EY - aEX$.

- Analogicznie definiujemy równanie prostej regresji X względem Y i ma ona postać:

$$\frac{x - EX}{\sigma X} = \frac{1}{\rho_{X,Y}} \frac{y - EY}{\sigma Y}.$$

Regresja - Przykład 1

Niech wektor losowy (X, Y) ma gęstość

$$f_{X,Y}(x,y) = \begin{cases} x+y, & \text{gdy } x,y \in [0,1] \\ 0, & \text{w przeciwnym wypadku.} \end{cases}$$

Wtedy

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy = x + \frac{1}{2},$$

$$f_Y(y) = y + \frac{1}{2},$$

$$EX = EY = \int_{-\infty}^{\infty} x(x + \frac{1}{2}) dx = \frac{7}{12},$$

$$EX^2 = EY^2 = \frac{5}{12}$$

$$\sigma X = \sigma Y = \frac{\sqrt{11}}{12}$$

Regresja - Przykład 1

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{3} - \frac{7}{12} * \frac{7}{12} = -\frac{1}{144} \\ \rho_{X,Y} &= -\frac{1}{11} \end{aligned}$$

zatem prosta regresji ma równanie

$$y = -\frac{1}{11}x + \frac{7}{12}.$$

Regresja - Przykład 2

Niech zmienne losowe X i Y będą zadane tabelką:

$x \backslash y$	-1	0	1	
-1	0.3	0.1	0.2	0.6
0	0	0.1	0.1	0.2
1	0.1	0	0.1	0.2
	0.4	0.2	0.4	1

Wtedy

$$EX = -1 * 0.6 + 0 * 0.2 + 1 * 0.1 = 0.4$$

$$EX^2 = 1 * 0.6 + 0 * 0.2 + 1 * 0.2 = 0.8$$

$$\sigma X = \sqrt{0.8 - 0.4^2} = 0.8$$

$$EY = -1 * 0.6 + 0 * 0.2 + 1 * 0.2 = -0.4$$

$$EY^2 = 1 * 0.6 + 0 * 0.2 + 1 * 0.2 = 0.8$$

$$\sigma Y = \sqrt{0.8 - 0.4^2} = 0.8$$

Regresja - Przykład 2

$$\begin{aligned}EXY &= 0.3 - 0.2 - 0.1 + 0.1 = 0.1 \\Cov(X, Y) &= 0.1 - 0.4^2 = -0.06 \\ \rho_{X,Y} &= \frac{-0.06}{0.8 * 0.8} = -0.094\end{aligned}$$

i prosta regresji ma postać

$$y = -0.094x + (-0.4 - -0.094 * 0.4) = -0.094x - 0.362$$

Statystyka

Statystyka zajmuje się:

- 1 Gromadzeniem i porządkowaniem danych (statystyka opisowa)
- 2 Wyznaczaniem parametrów rozkładów (estymacja)
- 3 Stawianiem i weryfikowaniem hipotez
- 4 Wyznaczaniem zależności pomiędzy badanymi cechami (regresja)
- 5 Badaniem wpływu jednego czynnika na inny (analiza wariancji)

Próba

Próbą nazywamy układ wartości x_1, x_2, \dots, x_n , otrzymanych z obserwacji lub eksperymentów. W praktyce próbą jest ciąg n zmiennych losowych X_1, X_2, \dots, X_n , przyjmujących wartości równe wartościom badanej cechy.

Populacja

Populacją nazywamy zbiór wszystkich elementów spośród których losujemy próbę. W praktyce oznacza to, iż populacją jest zbiór wszystkich możliwych wartości, jakie może przyjąć zmienna losowa.

Statystyka

Statystyką nazywamy dowolną funkcję z próby np:

$$\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i, \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}, R = \max_{1 \leq i \leq n} x_i - \min_{1 \leq i \leq n} x_i.$$

Słowem statystyka określa się zarówno

- ❶ zbiór liczb (przedmiot statystyki)
- ❷ zbiór metod przetwarzania ciągu (ciągów) liczb (metody statystyki)
- ❸ naukę obejmującą to wszystko powyżej (nauka Statystyka)

Metody losowania to:

- **Prosty dobór losowy próby badawczej** - jedna z najbardziej podstawowych technik. Badacz, na podstawie ponumerowanych elementów listy wchodzących w skład populacji, dobiera próbę korzystając z tablicy liczb losowych.
- **Systematyczny dobór próby** - do próby badacz typuje co któryś punkt z listy.
- **Dobór warstwowy** - populacja jest dzielona na warstwy (badacz bierze pod uwagę ważne dla badania czynniki), a następnie z każdej losowana jest próba. Przykład: jeżeli chcesz otrzymać próbę warstwową nauczycieli w konkretnym regionie kraju, możesz posegregować populację według wieku, zarobków, poglądów politycznych itp.

- **Grupowy wielostopniowy dobór próby** - najpierw badacz przygotowuje listę grup losowania i losuje z nich próbę. Kolejnym krokiem jest wykonanie listy elementów, które wchodzi w skład wylosowanych do próby grup i kolejne losowanie.
- **Celowy dobór próby badawczej** - badacz w procesie wyboru elementów do próby wykorzystuje własną wiedzę, a sam proces przebiega bardzo subiektywnie,
- **Dobór kwotowy** - dla doboru kluczowe znaczenie mają cechy populacji; próba musi mieć taki sam ich rozkład i na tej podstawie dobierane są jej elementy,
- **Kula śnieżna** - metoda polegająca na przepytaniu kilku osób z konkretnej populacji i poproszenie respondentów o pomoc w dotarciu do kolejnych,
- **Dobór przypadkowy** - przypadkowy dobór elementów do próby, np. sondaże uliczne.

- 1 Nominalna - stosowana jest w mierzeniu cech jakościowych np. płeć, marka samochodu, stan cywilny
- 2 Porządkowa - służy do mierzenia cech porządkowych np. jak często ogląda pan/pani telewizję? b. często, często, rzadko, nigdy. Jest ona subiektywna i różni respondenci mogą w tych samych sytuacjach udzielać różnych odpowiedzi.
- 3 Interwałowa - służy do mierzenia cech ilościowych w skali mającej stałą jednostkę, ale bez arbitralnie przyjętego 0 np. jak ciekawa była ta książka w skali 1-10?
- 4 Ilorazowa - jak poprzednio ale z arbitralnie przyjętym 0 np. wiek, wzrost, dochody, ceny.

Dane

Jakościowe

Ilościowe

Skala
nominalna

Skala
porządkowa

Skala
interwałowa

Skala
ilorazowa

Podział na kategorie
(wyczerpujący i
rozłączny)

Podział na kategorie
dające się
uporządkować

Można określić
"odległość" między
danymi

Można określić
"punkt zerowy" skali

Cele przekształcania danych

- 1 **Stabilizacja wariancji** (w ramach kilku grup danych różniących się średnimi mogą być różne rozrzuty, podczas gdy wymagany jest stały rozrzut w grupach). Jeśli znana jest zależność wariancji od średniej w grupach: $\sigma^2(x) = \Phi[E(x)]$ to funkcję $y = f(x)$, według której należy przekształć dane można w przybliżeniu wyznaczyć z zależności: $\frac{dy}{dx} = \frac{\text{const}}{\sqrt{\sigma^2(x)}}$
- 2 **Linearyzacja zależności między dwiema cechami** jeżeli linia regresji między dwoma zmiennymi ilościowymi wyraźnie nie posiada charakteru liniowego, to można konstruować linię regresji w postaci krzywoliniowej lub przekształcić wstępnie dane, aby otrzymać zależność liniową. W tym drugim przypadku metody analizy są szczególnie proste, a wyniki intuicyjnie zrozumiałe i łatwe do interpretacji.
- 3 **Normalizacja rozkładu.** Często metoda analizy, którą zamierzamy stosować wymaga, aby dana próba została pobrana ze zbiorowości o rozkładzie normalnym (Gaussa). Niektóre metody są mało wrażliwe na nienormalność rozkładu, więc cel (3) jest na ogół mniej ważny od celów (1) i (2). Niektóre przekształcenia normalizują rozkład *przy okazji* stabilizacji wariancji czy linearyzacji linii regresji

(1) Przekształcenie logarytmiczne $y = \log_a(x)$ gdzie $a = e$ lub $a = 10$.
Przekształcenie to

- **Stabilizuje wariancję**, gdy $\sigma^2(x)$ rośnie znacznie w zależności od wartości średniej
- **Linearyzuje zależności zbliżone do wykładniczych**
- **Zmniejsza dodatnią asymetrię rozkładu**

Przekształcenie logarytmiczne używane jest np. wówczas, gdy średni przyrost efektu ΔE jest proporcjonalny do średniego względnego przyrostu przyczyny $\frac{\Delta P}{P}$ tzn. $\Delta E = k \frac{\Delta P}{P} \rightarrow E = k \ln(P) + C$. Również stosowane, gdy dane przyjmują wartości ciągu geometrycznego, np. w przypadku szeregu rozcieńczeń.

(2) Przekształcenie *odwrotnościowe* $y = \frac{1}{x}$

- **Stabilizuje wariancję**, gdy $\sigma^2(x)$ jest proporcjonalna do czwartej potęgi średniej
- **Dużym wartościom pierwotnym odpowiadają małe wartości po przekształceniu**

Średnia arytmetyczna danych przekształconych jest średnią harmoniczną danych pierwotnych. Stosowane dla danych w postaci czasów przeżycia.

(a) **Przekształcenia kątowe** $y = \arcsin(p)$

Dobrze stabilizuje wariancję, linearyzacja nie jest idealna.

(b) **Przekształcenie logitowe** $y = \ln \frac{p}{1-p}$

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

(c) **Przekształcenie probitowe** $y = \Phi^{-1}(p) + 5$

Ma właściwości podobne do logitowego.

Wartości pobranej próby których często jest bardzo dużo, zwykle wygodnie jest uporządkować w tzw. szeregu rozdzielczym. Jeśli liczba możliwych przyjmowanych wartości jest niewielka stosuje się szereg rozdzielczy jednostopniowy (prosty, punktowy), w przypadku gdy liczba możliwych przyjmowanych wartości jest duża stosuje się szereg rozdzielczy wielostopniowy (złożony, przedziałowy).

Szereg rozdzielczy umożliwia oszacowanie i zobrazowanie graficzne (za pomocą histogramu) rozkładu z którego została wylosowana próba.

Etapy tworzenia szeregu rozdzielczego złożonego

- 1 Wyznaczamy $\min_i x_i$ oraz $\max_i x_i$
- 2 Wyznaczamy rozstęp z próby $R = \max_i x_i - \min_i x_i$
- 3 Ustalamy liczbę przedziałów klasowych k
- 4 Wyznaczamy lewy koniec pierwszego przedziału klasowego $\min_i x_i - \frac{\alpha}{2}$, gdzie α - dokładność pomiaru,
- 5 Wyznaczamy długość przedziału klasowego: $h \geq \frac{R+\alpha}{k}$
- 6 Rozkład empiryczny obserwacji zebranych w utworzonym szeregu rozdzielczym, możemy zobrazować wykresem (histogramem).

Wyznaczanie liczby przedziałów klasowych

- ❶ $k \cong \sqrt{n}$
- ❷ $k \cong 5 \ln(n)$
- ❸ $k \cong 1 + 3,3 \ln(n)$
- ❹ Za pomocą tabeli:

Liczba próbek	Liczba przedziałów
30 – 60	6 – 8
60 – 100	7 – 10
100 – 200	9 – 12
200 – 500	11 – 17
500 – 1500	16 – 25

Jednostopniowy szereg rozdzielczy - przykład

W pewnej klasie uczniowie otrzymali następujące oceny:

3	4	3,5	4	2	4
2	3,5	5	3	4,5	5
2	2	4,5	3	4	3,5
2	4	4	5	4,5	2
5	4,5	3	2	4	3
4	2	4,5	3,5	3	3

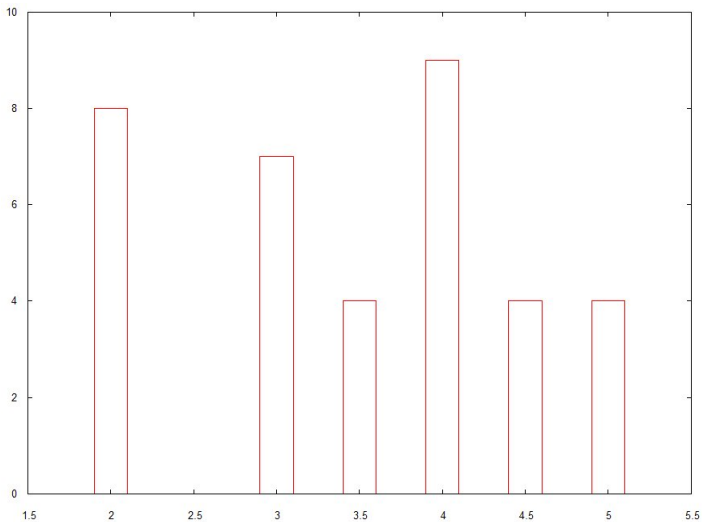
Sporządzić jednostopniowy szereg rozdzielczy.

Jednostopniowy szereg rozdzielczy - przykład

ocena	liczebność w próbie	częstość $f_i = \frac{n_i}{n}$	częstość skumulowana
2	8	$\frac{8}{36}$	$\frac{8}{36}$
3	7	$\frac{7}{36}$	$\frac{15}{36}$
3,5	4	$\frac{4}{36}$	$\frac{19}{36}$
4	9	$\frac{9}{36}$	$\frac{28}{36}$
4,5	4	$\frac{4}{36}$	$\frac{32}{36}$
5	4	$\frac{4}{36}$	$\frac{36}{36}$
Σ	36	1	-

Jednostopniowy szereg rozdzielczy - histogram

Histogram



Wielostopniowy szereg rozdzielczy - przykład

Dane dotyczące zarobków w pewnej grupie zawodowej wśród 40 wybranych w sposób losowy osób kształtowały się następująco (z dokładnością do 1 zł):

1307	1560	1101	1617	1251	1481	1513	1390
1451	1612	1470	1680	1480	1413	1240	1580
1298	1690	1597	1273	1154	1312	1220	1683
1570	1635	1470	1535	1372	1590	1506	1455
1576	1341	1620	1357	1423	1170	1685	1473

Zbudować wielostopniowy (przedziałowy) szereg rozdzielczy. Na podstawie szeregu rozdzielczego utworzyć histogram.

Wielostopniowy szereg rozdzielczy - przykład

$$\min_i x_i = 1101, \quad \max_i x_i = 1690, \quad R = 1690 - 1101 = 589$$

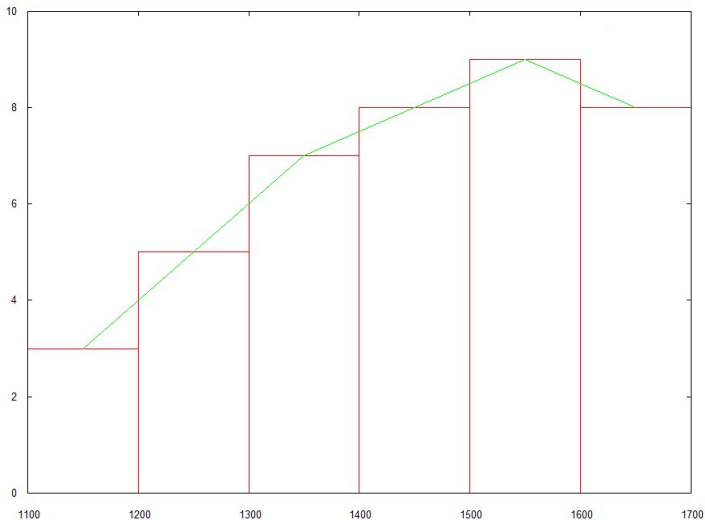
$$k = \sqrt{n} = \sqrt{40} \approx 6, \quad h = \frac{R}{k} = \frac{589}{6} = 98,167 \approx 100$$

$$x_0 = \min_i x_i - \frac{\alpha}{2} = 1101 - 0,5 = 1100,5$$

Przedział ($x_{i-1}, x_i]$	Środek x_i^*	Liczebność n_i	Częstość f_i	Częstość skumulowana f_i^*
1100,5 – 1200	1150,25	3	$\frac{3}{40}$	$\frac{3}{40}$
1200 – 1300	1250	5	$\frac{5}{40}$	$\frac{8}{40}$
1300 – 1400	1350	7	$\frac{7}{40}$	$\frac{15}{40}$
1400 – 1500	1450	8	$\frac{8}{40}$	$\frac{23}{40}$
1500 – 1600	1550	9	$\frac{9}{40}$	$\frac{32}{40}$
1600 – 1700	1650	8	$\frac{8}{40}$	$\frac{40}{40}$
Σ	-	40	1	-

Wielostopniowy szereg rozdzielczy - histogram

Histogram



1 Średnia z próby:

- 1 Dane nieogrupowane : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$,
- 2 Szereg rozdzielczy punktowy : $\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i$
- 3 Szereg rozdzielczy przedziałowy : $\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i^* n_i$

2 Moda (dominanta) (M_o) - wartość najczęstsza

- 1 Dane nieogrupowane - wartość najczęstsza o ile nie jest to wartość skrajna (wówczas moda jest nieokreślona)
- 2 Szereg rozdzielczy punktowy - wartość najczęstsza o ile nie jest to wartość skrajna (wówczas moda jest nieokreślona)
- 3 Szereg rozdzielczy przedziałowy -

$$M_o = x_m + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})} h_m$$

3 Mediana - wartość środkowa w uporządkowanej próbie:

1 Dane niegrupowane :

$$M_e = \begin{cases} x_{(\frac{n+1}{2})}, & n = 2k + 1 \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & n = 2k \end{cases}$$

2 Szereg rozdzielczy punktowy :

$$M_e = \begin{cases} x_{(\frac{n+1}{2})}, & n = 2k + 1 \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & n = 2k \end{cases}$$

3 Szereg rozdzielczy przedziałowy :

$$M_e = x_m + \frac{\frac{n}{2} - \sum_{i=1}^{m-1} n_i}{n_m} h_m$$

x_m - lewy koniec przedziału z medianą, n_m - liczebność przedziału z medianą, h_m - długość przedziału z medianą

- 4 Kwartyle (dolny Q_1 i górny Q_3) : wartości które dzielą uporządkowaną próbę w stosunku 1:3 i 3:1.
- 5 Kwantyle rzędu $p \in [0; 1]$: wartości które dzielą uporządkowaną próbę w stosunki $p : 1-p$.
- 6 Inne średnie (geometryczna, harmoniczna).

Miary statystyczne - miary rozproszenia (zmienności)

1 Rozstęp

$$R = \max_i x_i - \min_i x_i$$

2 Wariancja:

1 Dane nie pogrupowane :

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

2 Szereg rozdzielczy punktowy :

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

3 Szereg rozdzielczy przedziałowy :

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

3 Odchylenie standardowe

$$s = \sqrt{s^2},$$

typowy przedział zmienności ($\bar{x} - s, \bar{x} + s$)

4 Odchylenie przeciętne od średniej arytmetycznej

- 1 Dane nie pogrupowane : $d_1 = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$
- 2 Szereg rozdzielczy punktowy : $d_1 = \frac{1}{n} \sum_{i=1}^k |x_i - \bar{x}| n_i$
- 3 Szereg rozdzielczy przedziałowy : $d_1 = \frac{1}{n} \sum_{i=1}^k |x_i^* - \bar{x}| n_i$

5 Odchylenie przeciętne od mediany

- 1 Dane nie pogrupowane : $d_2 = \frac{1}{n} \sum_{i=1}^n |x_i - M_e|$
- 2 Szereg rozdzielczy punktowy : $d_2 = \frac{1}{n} \sum_{i=1}^k |x_i - M_e| n_i$
- 3 Szereg rozdzielczy przedziałowy : $d_2 = \frac{1}{n} \sum_{i=1}^k |x_i^* - M_e| n_i$

Miary statystyczne - miary rozproszenia (zmienności)

- 6 Odchylenie ćwiartkowe $Q = \frac{Q_3 - Q_1}{2}$
- 7 Współczynnik zmienności $V = \frac{s}{\bar{x}} \cdot 100\%$
- 8 Współczynnik nierównomierności $H = \frac{d_1}{\bar{x}} \cdot 100\%$
- 9 Pozycyjny współczynnik nierównomierności $H_2 = \frac{d_2}{m_0}$

1 Współczynnik asymetrii

1 Dane niepogrupowane : $A = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$

2 Szereg rozdzielczy punktowy : $A = \frac{\frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^3 n_i}{s^3}$

3 Szereg rozdzielczy przedziałowy : $A = \frac{\frac{1}{n} \sum_{i=1}^k (x_i^* - \bar{x})^3 n_i}{s^3}$

2 Wskaźnik asymetrii $w_1^s = \bar{x} - M_O$, $w_2^s = \bar{x} - M_e$

3 Współczynniki skośności $A_1 = \frac{\bar{x} - M_O}{s}$, $A_2 = \frac{\bar{x} - M_O}{d_1}$, $A_3 = \frac{Q_3 + Q_1 - 2M_e}{2Q}$

1 Współczynnik skupienia (kurtoza)

- 1 Dane niepogrupowane : $K = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4}$
- 2 Szereg rozdzielczy punktowy : $K = \frac{\frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^4 n_i}{s^4}$
- 3 Szereg rozdzielczy przedziałowy : $K = \frac{\frac{1}{n} \sum_{i=1}^n (x_i^* - \bar{x})^4 n_i}{s^4}$

2 Eksces $q = K - 3$

Miary statystyczne - przykład 1.

ocena	n_i	f_i	f_i^{sk}	$x_i n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
2	8	$\frac{8}{36}$	$\frac{8}{36}$	16	-1,47222	2,16743	17,33944
3	7	$\frac{7}{36}$	$\frac{15}{36}$	21	-0,47222	0,22299	1,56093
3,5	4	$\frac{4}{36}$	$\frac{19}{36}$	14	0,02778	0,00077	0,00308
4	9	$\frac{9}{36}$	$\frac{28}{36}$	36	0,52778	0,27855	2,50695
4,5	4	$\frac{4}{36}$	$\frac{32}{36}$	18	1,02778	1,05633	4,22532
5	4	$\frac{4}{36}$	$\frac{36}{36}$	20	1,52778	2,33411	9,33644
Σ	36	1	-	125	-	-	34,97216

$$\bar{x} = \frac{1}{n} \sum_{i=1}^6 x_i n_i = \frac{125}{36} = 3,47222$$

$$s^2 = \frac{1}{n} \sum_{i=1}^6 (x_i - \bar{x})^2 n_i = \frac{34,97216}{36} = 0,97145$$

$$s = \sqrt{0,97145} = 0,98562$$

$$M_e = \frac{x_{18} + x_{19}}{2} = \frac{3,5 + 3,5}{2} = 3,5, \quad M_o = 4$$

Miary statystyczne - przykład 2.

$(x_{i-1}, x_i]$	x_i^*	n_i	f_i	f_i^{sk}	$x_i^* n_i$	$(x_i^*)^2 n_i$
1100,5 – 1200	1150,25	3	$\frac{3}{40}$	$\frac{3}{40}$	3450,75	3969225
1200 – 1300	1250	5	$\frac{5}{40}$	$\frac{8}{40}$	6250	7812500
1300 – 1400	1350	7	$\frac{7}{40}$	$\frac{15}{40}$	9450	12757500
1400 – 1500	1450	8	$\frac{8}{40}$	$\frac{23}{40}$	11600	16820000
1500 – 1600	1550	9	$\frac{9}{40}$	$\frac{32}{40}$	13950	21622500
1600 – 1700	1650	8	$\frac{8}{40}$	$\frac{40}{40}$	13200	21780000
Σ	-	40	1	-	57900,75	84761725

$$\bar{x} = \frac{1}{n} \sum_{i=1}^6 x_i^* n_i = \frac{57900,75}{40} = 1447,52$$

$$s^2 = \frac{1}{n} \sum_{i=1}^6 (x_i^*)^2 n_i - (\bar{x})^2 = \frac{84761725}{40} - 2095310,53 = 23732,6$$

$$s = \sqrt{23732,6} = 154,05, \quad M_e = 1400 + \frac{20 - 15}{8} \cdot 100 = 1462,5$$

$$M_o = 1500 + \frac{9 - 8}{(9 - 8) + (9 - 8)} \cdot 100 = 1550$$

Jednym z najważniejszych zadań statystyki matematycznej jest ocena parametrów rozkładu z którego została wylosowana próba na podstawie wylosowanych wartości próby. Przykładowo jeśli badamy wzrost w populacji osób, który ma rozkład $N(\mu, \sigma)$, to na podstawie wylosowanej próbki chcielibyśmy oszacować średni wzrost populacji (μ) i o ile średnio odchyła się wzrost w populacji od tej wyznaczonej wartości średniej (σ). Jeśli poszukujemy oszacowania parametru Θ , to szukamy takiej statystyki (funkcji z próby) $T(X_1, X_2, \dots, X_n)$, która w sposób najlepszy przybliży nam parametr Θ . Statystyki szacujące dany parametr Θ nazywamy estymatorami. Jeśli poszukujemy pojedynczego parametru Θ , to taką estymację nazywamy estymacją punktową.

Estymator nieobciążony

Estymator nazywamy nieobciążonym, jeśli

$$E(T(X_1, X_2, \dots, X_n)) = \Theta,$$

czyli "na średnio" wartości estymatora są zbliżone do szacowanego parametru. Jeśli tak nie jest to estymator $T(X_1, X_2, \dots, X_n)$ jest obciążony, a różnicę $E(T(X_1, X_2, \dots, X_n)) - \Theta$, nazywamy jego obciążeniem.

Błąd estymatora

O jakości estymatora świadczy średnie odchylenie kwadratowe (im mniejsze, tym dany estymator jest lepszy)

$$BSK_{\Theta}(T(X_1, X_2, \dots, X_n)) = E(T(X_1, X_2, \dots, X_n) - \Theta)^2.$$

Estymatory są z reguły tym lepsze im z większej liczby próbek wyznaczają szukany parametr.

Estymator zgodny

Estymator $T(X_1, X_2, \dots, X_n)$ jest zgodny z Θ , jeśli $T(X_1, X_2, \dots, X_n) \rightarrow \Theta$, gdy $n \rightarrow \infty$, czyli dla dostatecznie dużej próby wartości estymatora leżą "blisko" faktycznej wartości parametru Θ .

- 1 Metoda momentów - wyznaczamy momenty z próby

$$M_k = \frac{1}{n} \sum_{i=1}^n x_i^k,$$

które są estymatorami momentów EX^k , a te z kolei zależą od wyznaczanych parametrów.

- 2 Metoda największej wiarygodności - jeżeli badana populacja ma gęstość rozkładu $f(x, \Theta)$, to tworzymy funkcję wiarygodności

$$L(X_1, X_2, \dots, X_n) = \prod_{j=1}^n f(x_j, \Theta)$$

i znajdujemy parametr Θ , dla którego osiąga ona maksimum.

Metoda momentów - przykład

Jeśli szukanym parametrem jest $\Theta = EX$, to

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

jest najlepszym estymatorem Θ .

Jeśli szukanym parametrem jest $\Theta = \sigma^2 X$, to

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

jest najlepszym estymatorem Θ .

Estymacja przedziałowa

W przypadku estymacji punktowej wyznaczaliśmy pojedynczą wartość, która stanowiła przybliżenie poszukiwanego parametru Θ . Jednakże prawdopodobieństwo, że wyznaczona z próbki wartość szukanego parametru jest równa wartości tego parametru w populacji jest bliskie 0. Znacznie lepszym podejściem jest estymacja przedziałowa. W estymacji przedziałowej ustalamy prawdopodobieństwo $1 - \alpha$ tzw. poziom ufności oraz wyznaczamy przedział do którego z prawdopodobieństwem $1 - \alpha$ należy wyznaczany parametr, czyli wyznaczamy takie a i b , że $P[a \leq \Theta \leq b] = 1 - \alpha$. Oczywiście im większy przyjmiemy poziom ufności, tym większy przedział uzyskamy. Najczęściej jako poziom ufności przyjmuje się wartość 0,95 (typowy poziom ufności), albo 0,9; 0,99, czy też 0,975.

Oznaczmy:

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

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

Przedziały ufności dla średniej

- ❶ MODEL I - Badana cecha ma rozkład normalny $N(\mu, \sigma)$, o nieznanym μ i znanym σ . Z tablic standardowego rozkładu normalnego odczytujemy wartość $u(1 - \alpha/2)$. Przedział ufności:

$$\mu \in [\bar{x} - u(1 - \frac{\alpha}{2}) \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + u(1 - \frac{\alpha}{2}) \cdot \frac{\sigma}{\sqrt{n}}]$$

- ❷ MODEL II - Badana cecha ma rozkład normalny $N(\mu, \sigma)$, o nieznanym μ i σ . Z tablic rozkładu t-studenta z $n - 1$ stopniami swobody odczytujemy wartość $t(1 - \frac{\alpha}{2}, n - 1)$. Przedział ufności:

$$\mu \in [\bar{x} - t(1 - \frac{\alpha}{2}, n - 1) \cdot \frac{s}{\sqrt{n - 1}}, \bar{x} + t(1 - \frac{\alpha}{2}, n - 1) \cdot \frac{s}{\sqrt{n - 1}}]$$

- ❸ MODEL III - Badana cecha ma dowolny rozkład o nieznanym μ i nieznanym odchyleniu standardowym σ , zaś liczebność próby jest duża ($n \geq 100$). Z tablic standardowego rozkładu normalnego odczytujemy wartość $u(1 - \alpha/2)$. Przedział ufności:

$$\mu \in [\bar{x} - u(1 - \frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}, \bar{x} + u(1 - \frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}]$$

Twierdzenie o średniej

Jeżeli $\{X_n, n \geq 1\}$ jest ciągiem niezależnych zmiennych losowych o rozkładzie $N(\mu, \sigma)$ to

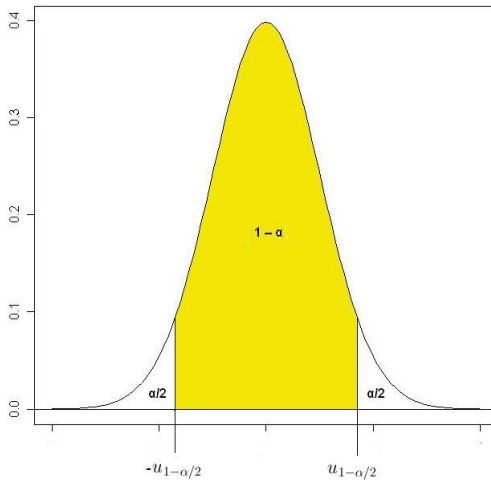
$$\frac{\frac{\sum_{i=1}^n X_i}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} N(0, 1), \quad n \rightarrow \infty,$$

$$\frac{\frac{\sum_{i=1}^n X_i}{n} - \mu}{\frac{s}{\sqrt{n-1}}} \stackrel{D}{\sim} t(n-1), \quad n < 100,$$

$$\frac{\frac{\sum_{i=1}^n X_i}{n} - \mu}{\frac{s}{\sqrt{n}}} \xrightarrow{D} N(0, 1), \quad n \rightarrow \infty,$$

gdzie $t(n)$ jest rozkładem t-Studenta z parametrem n .

Kwantyle rzędu $1 - \frac{\alpha}{2}$



Kwantyle rozkładu normalnego $u(p)$

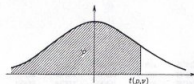
T a b l i c a 6. Kwantyle $u(p)$ rzędu p rozkładu normalnego $N(0, 1)$

p	0,90	0,95	0,975	0,99	0,995
$u(p)$	1,28	1,64	1,96	2,33	2,58

Kwantyle rozkładu t-studenta $t(p, n)$

Tablice statystyczne

287



Tablica 7. Kwantyle $t(p, v)$ rzędu p rozkładu Studenta o v stopniach swobody

v	p				
	0,90	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	,638	,353	3,182	4,541	5,841
4	,533	,332	2,776	3,747	4,604
5	,476	,315	,271	,365	,332
6	1,440	1,943	2,447	3,143	3,707
7	,415	,295	,265	2,998	,499
8	,397	,289	,266	,287	,355
9	,383	,283	,262	,282	,350
10	,372	,281	,228	,264	,345
11	1,363	1,795	2,201	2,718	3,106
12	,356	,282	,261	,264	,344
13	,350	,271	,260	,260	,342
14	,345	,261	,259	,259	,341
15	,341	,253	,258	,258	,340
16	1,337	1,746	2,120	2,583	2,921
17	,333	,240	,257	,257	,338
18	,330	,234	,252	,252	,336
19	,328	,229	,253	,253	,335
20	,325	,225	,252	,252	,334
21	1,322	1,721	2,080	2,518	2,831
22	,321	,217	,251	,251	,333
23	,319	,214	,250	,250	,332
24	,318	,211	,249	,249	,331
25	,316	,208	,248	,248	,330
26	1,315	1,706	2,055	2,479	2,779
27	,314	,203	,247	,247	,329
28	,312	,201	,246	,246	,328
29	,311	,199	,245	,245	,327
30	,310	,197	,244	,244	,326
31	1,309	1,695	2,039	2,453	2,744
32	,309	,194	,243	,243	,325
33	,308	,192	,242	,242	,324
34	,307	,191	,241	,241	,323
35	,306	,190	,240	,240	,322

288

Tablice statystyczne

Tablica 7 (od.)

v	p				
	0,90	0,95	0,975	0,99	0,995
36	1,305	1,688	2,028	2,434	2,720
37	,305	,687	,025	,431	,715
38	,304	,686	,024	,429	,712
39	,304	,685	,023	,425	,708
40	,303	,684	,021	,423	,704
41	1,303	1,683	2,019	2,421	2,701
42	,302	,682	,018	,418	,698
43	,302	,681	,017	,416	,695
44	,301	,680	,015	,414	,692
45	,301	,679	,014	,412	,690
46	1,300	1,679	2,013	2,410	2,687
47	,300	,678	,012	,408	,685
48	,299	,677	,011	,407	,682
49	,299	,677	,010	,405	,680
50	,299	,676	,009	,403	,678
55	1,297	1,673	2,004	2,396	2,668
60	,295	,671	,000	,390	,660
65	,295	,669	,997	,385	,654
70	,294	,667	,994	,381	,648
75	,293	,665	,992	,377	,643
80	1,292	1,664	1,990	2,374	2,639
90	,291	,662	,987	,369	,632
100	,290	,660	,984	,364	,626
120	,289	,658	,980	,358	,617
150	,287	,655	,976	,351	,609
200	1,286	1,653	1,972	2,345	2,601
300	,284	,650	,968	,339	,592
500	,283	,648	,965	,334	,586
1000	,282	,646	,962	,330	,581
∞	,282	,645	,960	,326	,576

Przedziały ufności dla średniej

Przykład

W celu analizy miesięcznych wydatków mieszkaniowych wylosowano $n = 25$ rodzin i wyznaczono średnie wydatki $\bar{x} = 530$ z miesięcznie. Z poprzednich badań wiadomo, że rozkład tych wydatków jest w przybliżeniu normalny oraz $\sigma = 60$ z. Oblicz przedział ufności dla średniej wydatków mieszkaniowych dla współczynnika ufności $1 - \alpha = 0.01$.

Wiemy, że $\bar{x} = \sum_{i=1}^{25} X_i / n = 530$ oraz $\sigma = 60$, korzystamy z MODELU I. Z Tablic rozkładu normalnego: odczytujemy, że $u(1 - \frac{\alpha}{2}) = 2.58$ więc

$$\mu \in [530 - 2.58 * \frac{60}{5}, 530 + 2.58 * \frac{60}{5}] = [499, 561].$$

Przykład

W wyniku badania stanu zdrowia 1000 losowo wybranych dzieci zamieszkałych w Lublinie u 250 stwierdzono wady wzroku. Jak liczna powinna być próba, aby przy współczynniku ufności $1 - \alpha = 0.95$ oszacować odsetek ogółu dzieci z wadami wzroku w Lublinie, jeśli nie chcemy się pomylić o więcej niż 4%?

Przedziały ufności dla wariancji

- ❶ MODEL I - Badana cecha ma rozkład normalny $N(\mu, \sigma)$, o nieznanym parametrach μ i σ , zaś próba jest mała ($n \leq 50$). Z tablic rozkładu χ^2 odczytujemy wartości $\chi^2(1 - \frac{\alpha}{2}, n - 1)$ i $\chi^2(\frac{\alpha}{2}, n - 1)$. Przedział ufności:

$$\sigma^2 \in \left[\frac{ns^2}{\chi^2(1 - \frac{\alpha}{2}, n - 1)}, \frac{ns^2}{\chi^2(\frac{\alpha}{2}, n - 1)} \right]$$

- ❷ MODEL II - Badana cecha ma rozkład normalny $N(\mu, \sigma)$, o nieznanym parametrach μ i σ , zaś próba jest duża ($n > 50$). Z tablic standardowego rozkładu normalnego odczytujemy wartość $u(1 - \alpha/2)$. Przedział ufności:

$$\sigma^2 \in \left[\frac{2ns^2}{(\sqrt{2n-3} + u(1 - \frac{\alpha}{2}))^2}, \frac{2ns^2}{(\sqrt{2n-3} - u(1 - \frac{\alpha}{2}))^2} \right]$$

Twierdzenie o wariancji

Jeżeli $\{X_n, n \geq 1\}$ jest ciągiem niezależnych zmiennych losowych o rozkładzie $N(\mu, \sigma)$ to

$$\frac{\sum_{i=1}^n X_i^2}{\sigma^2} \stackrel{D}{\sim} \chi^2(n-1), \quad n \leq 50,$$
$$\frac{\sqrt{2 \sum_{i=1}^n X_i^2}}{\sigma} - \sqrt{2n-3} \stackrel{D}{\rightarrow} N(0,1), \quad n \rightarrow \infty.$$

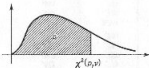
gdzie $\chi^2(n)$ jest rozkładem χ^2 z n stopniami swobody.

Kwantyle rozkładu χ^2 $\chi^2(p, n)$

Tablice statystyczne

289

Tablica 8. Kwantyle $\chi^2(p, v)$ rzędu p rozkładu χ^2 o v stopniach swobody



v	p							
	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
1	—	—	0,001	0,004	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	11,071	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,299
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319
15	4,601	5,229	6,263	7,261	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	27,587	30,191	33,405	35,718
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,336	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	40,113	43,194	46,963	49,645
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,993
29	13,121	14,257	16,047	17,708	42,557	45,722	49,588	52,336
30	13,787	14,954	16,791	18,493	43,773	46,979	50,892	53,672

290

Tablice statystyczne

Tablica 8 (cd)

v	p							
	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
31	14,458	15,655	17,539	19,281	44,985	48,232	52,191	55,003
32	15,134	16,362	18,291	20,072	46,194	49,480	53,486	56,338
33	15,815	17,074	19,047	20,867	47,400	50,725	54,776	57,648
34	16,501	17,789	19,806	21,664	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	49,802	53,203	57,342	60,275
36	17,887	19,233	21,336	23,269	50,998	54,437	58,619	61,581
37	18,586	19,960	22,106	24,075	52,192	55,668	59,892	62,883
38	19,289	20,691	22,878	24,884	53,384	56,896	61,162	64,181
39	19,996	21,426	23,654	25,695	54,572	58,120	62,428	65,476
40	20,707	22,164	24,433	26,509	55,758	59,342	63,691	66,766
41	21,421	22,906	25,215	27,326	56,942	60,561	64,950	68,053
42	22,138	23,650	25,999	28,144	58,124	61,777	66,206	69,336
43	22,859	24,398	26,785	28,965	59,304	62,990	67,459	70,616
44	23,584	25,148	27,575	29,787	60,481	64,201	68,710	71,893
45	24,311	25,901	28,366	30,612	61,656	65,410	69,957	73,166
46	25,041	26,657	29,160	31,439	62,830	66,617	71,201	74,437
47	25,775	27,416	29,956	32,268	64,001	67,821	72,443	75,704
48	26,511	28,177	30,755	33,098	65,171	69,023	73,683	76,969
49	27,249	28,941	31,555	33,930	66,339	70,222	74,919	78,231
50	27,991	29,707	32,357	34,764	67,505	71,420	76,154	79,490

- 1 W MODELU III dla średniej, w przypadku, gdy parametr σ jest znany w miejsce s wstawiamy σ .
- 2 W MODELACH dla wariancji, jeśli parametr μ jest znany, wówczas we wzorze na s^2 zastępujemy μ , czyli

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

a w przypadku MODELU I także liczbę stopni swobody $(n - 1)$, zastępujemy wartością n .

Przedziały ufności dla wariancji

Przykład

Na podstawie losowej próby 20 tabliczek czekolady otrzymano odchylenie standardowe $s = 10\text{g}$ (zakładamy, że rozkład wagi tabliczki czekolady jest w przybliżeniu $N(\mu, \sigma)$). Jaki jest przedział ufności dla odchylenia standardowego w rozkładzie wagi wszystkich produkowanych tabliczek czekolady przy współczynniku ufności $1 - \alpha = 0.9$?

Ponieważ liczebność próby jest $20 < 50$ a μ nieznane więc korzystamy z MODELU I. Mamy $ns^2 = 2000$ oraz

$\chi^2(0.95, 19) = 10.12$, $\chi^2(0.05, 19) = 30.14$ więc

$$\sigma^2 \in \left[\frac{2000}{30.14}, \frac{2000}{10.12} \right] = [66.3, 197.7]$$

a stąd

$$\sigma \in [\sqrt{66.3}, \sqrt{197.7}] = [8.1, 14.1].$$

Przedziały ufności dla frakcji (wskaźnika struktury)

Założmy, iż pobierając próbkę interesuje nas w obserwacjach tylko posiadanie pewnej cechy, lub jej brak (czyli próbki mogą przyjmować tylko 2 wartości). Wskaźnik struktury p określa jaka część populacji posiada tę cechę.

- 1 Próbka jest duża ($n \geq 100$). Z tablic standardowego rozkładu normalnego odczytujemy wartość $u(1 - \alpha/2)$, oraz wyznaczamy wartość $\hat{p} = \frac{m}{n}$, gdzie m jest liczbą elementów w próbie, które posiadają badaną cechę.

Przedział ufności:

$$p \in \left[\hat{p} - u(1 - \frac{\alpha}{2}) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + u(1 - \frac{\alpha}{2}) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right]$$

Twierdzenie o frakcji

Niech $\{X, X_n, n \geq 1\}$ ciągiem niezależnych zmiennych losowych o rozkładzie normalnym a A będzie pewnym zdarzeniem o prawdopodobieństwie $P[X \in A] = p$. Wtedy

$$\frac{p - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \xrightarrow{D} N(0, 1), \quad n \rightarrow \infty,$$

gdzie $\hat{p} = \frac{\sum_{i=1}^n I[X_i \in A]}{n}$.

Graniczny błąd pomiaru

Zagadnieniem zbliżonym do estymacji przedziałowej jest wyznaczanie granicznego błędu pomiaru ($e_g = \pm ks$). Szukamy zatem takiej wartości e_g , że

$$P[|x - \mu| \leq e_g] = P[x - e_g \leq \mu \leq x + e_g] = p,$$

czyli wyznaczona doświadczalnie wielkość x różni się od faktycznej μ o e_g z prawdopodobieństwem p . Jeśli próba jest duża lub badana wielkość ma rozkład normalny o znanej wariancji, to $e_g = u(\frac{1}{2} + \frac{p}{2}) \frac{\sigma}{\sqrt{n}}$. Jeśli rozkład badanej wielkości jest normalny a wariancja nieznana, to

$$e_g = t(\frac{1}{2} + \frac{p}{2}, n - 1) \frac{s}{\sqrt{n-1}}.$$

Najczęściej w badaniach według norm międzynarodowych przyjmuje się $k = 2$, $k = 2,6$ lub $k = 3$.

Estymacja - przykład 1.

Przeprowadzono badania wśród studentów, zadając pytanie o czas (w godzinach) poświęcony na naukę w tygodniu i otrzymano wyniki:

2,5	4	7	3,5	8
4	5,5	6,5	5,5	3,5
6	3	2,5	7,5	5
5,5	4,5	8,5	6,5	7

Na typowym poziomie ufności ($1 - \alpha = 0,95$), zakładając, że czas poświęcony na naukę ma rozkład normalny wyznaczyć przedział ufności dla wartości średniej liczby godzin poświęconych przez studentów na naukę w ciągu tygodnia.

Estymacja - przykład 1.

Parametry μ i σ , są nieznane, rozkład jest normalny - wybieramy MODEL II dla średniej.

$$1 - \frac{\alpha}{2} = 0,975, \quad n - 1 = 19, \quad t(0,975; 19) = 2,093$$

$$\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = 5,3$$

$$s^2 = \frac{1}{20} \sum_{i=1}^{20} x_i^2 - (\bar{x})^2 = 3,16, \quad s = 1,78$$

$$\mu \in \left[5,3 - 2,093 \cdot \frac{1,78}{\sqrt{19}}, 5,3 + 2,093 \cdot \frac{1,78}{\sqrt{19}} \right]$$

$$\mu \in [4,4453; 6,1547]$$

Estymacja - przykład 2.

Wylosowano 300 osób zamieszkałych w Lublinie i okazało się, że 220 spośród nich pali papierosy. Na typowym poziomie ufności $1 - \alpha = 0,95$, wyznaczyć przedział ufności dla odsetka mieszkańców Lublina palących papierosy.

Szukamy przedziału ufności dla wskaźnika struktury, próba jest duża ($n > 100$) - wybieramy MODEL dla wskaźnika struktury.

$$1 - \frac{\alpha}{2} = 0,975, \quad \hat{p} = \frac{m}{n} = \frac{220}{300} = \frac{11}{15}, \quad u(0,975) = 1,96$$

$$p \in \left[\frac{11}{15} - 1,96 \sqrt{\frac{\frac{11}{15} \cdot \frac{4}{15}}{300}}, \frac{11}{15} + 1,96 \sqrt{\frac{\frac{11}{15} \cdot \frac{4}{15}}{300}} \right]$$

$$p \in [0,6833; 0,7834]$$

Estymacja - przykład 3.

Zakładamy, że roczne dochody pracownika z pewnej grupy zawodowej są zmienną losową o rozkładzie normalnym $N(\mu, \sigma)$ z nieznanymi parametrami μ i σ . W celu oszacowania parametru σ wylosowano próbę $n = 17$ osób i obliczono ich średnie dochody $\bar{x} = 12$ tys. zł. i odchylenie standardowe ich dochodów $s = 4,72$ tys. zł. Wyznacz przedział ufności dla odchylenia standardowego σ .

Estymacja - przykład 3.

Szukamy przedziału ufności dla wariancji parametry μ i σ są nieznane, rozkład jest normalny, próba jest mała ($n < 50$) - wybieramy MODEL I dla wariancji.

$$1 - \alpha = 0,95, \quad 1 - \frac{\alpha}{2} = 0,975$$

$$\chi^2(0,975, 16) = 28,845, \quad \chi^2(0,025, 16) = 6,908$$

$$\sigma^2 \in \left[\frac{17 \cdot (4,72)^2}{28,845}, \frac{17 \cdot (4,72)^2}{6,908} \right]$$

$$\sigma^2 \in [13,13; 54,825]$$

$$\sigma \in [3,6235; 7,4]$$

Kombinatoryka

Hipotezą statystyczną nazywamy dowolne stwierdzenie lub przypuszczenie, które może być zweryfikowane metodami statystycznymi w oparciu o losowo wybraną próbę.

Rodzaje hipotez statystycznych

Hipotezy dzielimy na:

- 1 Parametryczne - wymagają normalności rozkładu badanej cechy (a co za tym idzie badana cecha musi być na skali ilorazowej) i dotyczą parametrów tego rozkładu
- 2 Nieparametryczne - nie wymagają normalności rozkładu badanej cechy

Podział ten jest w dużej mierze historyczny a granice między poszczególnymi rodzajami hipotez coraz bardziej się zacierają.

Etapy weryfikacji hipotezy statystycznej (testu statystycznego)

- 1 Wybór testu.
- 2 Ustalenie postaci hipotez zerowej H_0 i alternatywnej H_1 .
- 3 Wyznaczenie obszaru krytycznego W na podstawie przyjętego poziomu istotności testu (α).
- 4 Wyznaczenie wartości statystyki testowej i ustalenie czy należy ona do obszaru krytycznego
- 5 Pozostawienie hipotezy H_0 jako nieodrzuconej, lub odrzucenie hipotezy H_0 na rzecz hipotezy alternatywnej H_1 .

Rodzaje błędów weryfikacji hipotez statystycznych

	Hipoteza prawdziwa	Hipoteza nieprawdziwa
przyjąć	+	błąd 2-rodzaju (β)
odrzuć	błąd 1-rodzaju (α)	+

Przy weryfikacji hipotezy możemy popełnić 2 rodzaje błędów:

- 1 Błąd 1-rodzaju - odrzucenie hipotezy prawdziwej
- 2 Błąd 2-rodzaju - przyjęcie hipotezy fałszywej

Rodzaje błędów weryfikacji hipotez statystycznych

Nie da się zmniejszyć prawdopodobieństwa popełnienia tych błędów jednocześnie. Zmniejszając prawdopodobieństwo popełnienia błędu pierwszego rodzaju (α), zwiększamy prawdopodobieństwo popełnienia błędu drugiego rodzaju (β) i odwrotnie. Większość testów statystycznych zbudowana jest w ten sposób, że testujący wybiera poziom istotności testu (α), a sam test minimalizuje wartość β . Podobnie jak w wypadku estymacji najczęściej przyjmowanym poziomem istotności jest $\alpha = 0,05$.

Prawdopodobieństwo popełnienia błędu pierwszego rodzaju (α) nazywamy poziomem istotności testu a w naukach technicznych jest to ryzyko wytwórcy.

Wartość $1 - \alpha$ nazywamy poziomem ufności testu.

Wartość $1 - \beta$ nazywamy mocą testu.

Prawdopodobieństwo popełnienia błędu drugiego rodzaju (β) w naukach technicznych nazywane jest ryzykiem odbiorcy.

Alternatywnie zamiast sprawdzać, czy wyznaczona wartość statystyki testowej znajduje się w obszarze krytycznym W , można wyznaczyć taką wartość prawdopodobieństwa popełnienia błędu pierwszego rodzaju p , przy której wartość statystyki testowej znajduje się na brzegu obszaru krytycznego. Jest to zatem najmniejszy poziom istotności, który pozwala na odrzucenie hipotezy H_0 .

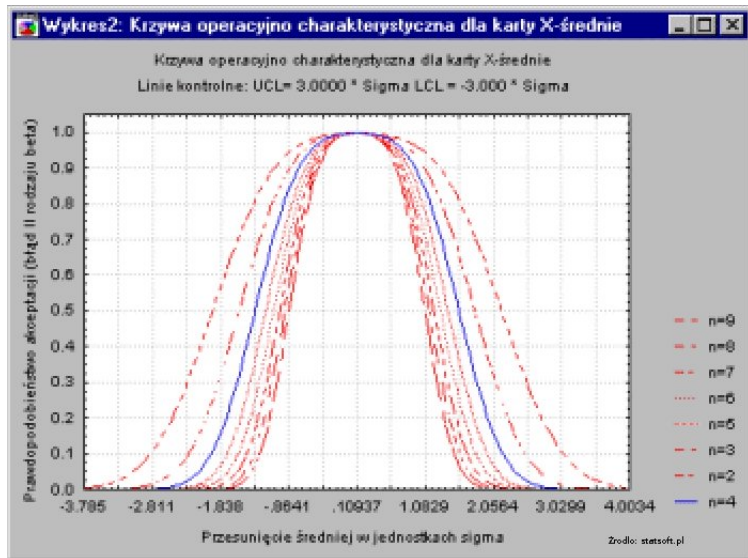
Czyli jeśli α jest wybranym poziomem istotności, to:

- 1 Jeśli $p \leq \alpha$ wówczas hipotezę H_0 odrzucamy
- 2 Jeśli $p > \alpha$ wówczas nie ma podstaw do odrzucenia hipotezy H_0

Taka metoda stosowana jest w wielu programach statystycznych np. *STATISTICA*.

Ryzyko odbiorcy (prawdopodobieństwo popełnienia błędu drugiego rodzaju) można przedstawić na wykresie w zależności od wartości średniej z wylosowanej próby. Wykres taki jest nazywany krzywą operacyjno-charakterystyczną lub w skrócie krzywą OC.

Krzywe operacyjne OC



Testy parametryczne stosuje się w sytuacjach w których stawiamy hipotezy dotyczące parametrów rozkładu. Są 2 rodzaje testów parametrycznych:

- 1 Testy porównań z wartością referencyjną
- 2 Testy porównań międzygrupowych (wymagają jednorodności wariancji).

Hipotezy parametryczne - średnia

Test z

Założenia:

- 1 Badana cecha ma rozkład normalny $N(\mu, \sigma)$,
- 2 Parametr σ jest znany,
- 3 Dane na skali interwałowej lub ilorazowej.

Stawiamy hipotezę zerową $H_0 : \mu = \mu_0$

Statystyka testowa	H_1	Obszar krytyczny W
$U = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$	$\mu \neq \mu_0$	$(-\infty, -u(1 - \frac{\alpha}{2})) \cup [u(1 - \frac{\alpha}{2}, +\infty)$
	$\mu < \mu_0$	$(-\infty, -u(1 - \alpha)]$
	$\mu > \mu_0$	$[u(1 - \alpha), +\infty$

Hipotezy parametryczne - średnia

Test t

Założenia:

- 1 Badana cecha X ma rozkład normalny $N(\mu, \sigma)$
- 2 Parametr σ jest nieznanym,
- 3 Dane na skali interwałowej lub ilorazowej.

Stawiamy hipotezę zerową $H_0 : \mu = \mu_0$

Statystyka testowa	H_1	Obszar krytyczny W
$t = \frac{\bar{x} - \mu_0}{s} \sqrt{n - 1}$	$\mu \neq \mu_0$	*
	$\mu < \mu_0$	$(-\infty, -t(1 - \alpha, n - 1)]$
	$\mu > \mu_0$	$[t(1 - \alpha, n - 1), +\infty)$

$$* = (-\infty, -t(1 - \frac{\alpha}{2}, n - 1)] \cup [t(1 - \frac{\alpha}{2}, n - 1), +\infty)$$

Test z

Założenia:

- 1 Badana cecha X ma dowolny rozkład
- 2 Próba jest duża ($n \geq 100$),
- 3 Dane na skali interwałowej lub ilorazowej.

Stawiamy hipotezę zerową $H_0 : \mu = \mu_0$

Statystyka testowa	H_1	Obszar krytyczny W
$U = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$	$\mu \neq \mu_0$	$(-\infty, -u(1 - \frac{\alpha}{2})) \cup [u(1 - \frac{\alpha}{2}), +\infty)$
	$\mu < \mu_0$	$(-\infty, -u(1 - \alpha)]$
	$\mu > \mu_0$	$[u(1 - \alpha), +\infty)$

Hipotezy parametryczne - wariancja

Założenia:

- 1 Badana cecha X ma rozkład normalny $N(\mu, \sigma)$,
- 2 Próba jest mała ($n < 50$),
- 3 Dane na skali interwałowej lub ilorazowej.

Stawiamy hipotezę zerową $H_0 : \sigma^2 = \sigma_0^2$

Statystyka testowa	H_1	Obszar krytyczny W
$\chi^2 = \frac{ns^2}{\sigma_0^2}$	$\sigma^2 \neq \sigma_0^2$	$(0, \chi^2(\frac{\alpha}{2}, n-1)] \cup [\chi^2(1 - \frac{\alpha}{2}, n-1), +\infty)$
	$\sigma^2 < \sigma_0^2$	$(0, \chi^2(\alpha, n-1)]$
	$\sigma^2 > \sigma_0^2$	$[\chi^2(1 - \alpha, n-1), +\infty)$

Hipotezy parametryczne - wariancja

Założenia:

- 1 Badana cecha X ma rozkład normalny $N(\mu, \sigma)$,
- 2 Próba jest duża ($n \geq 50$),
- 3 Dane na skali interwałowej lub ilorazowej.

Stawiamy hipotezę zerową $H_0 : \sigma^2 = \sigma_0^2$

Statystyka testowa	H_1	Obszar krytyczny W
$U = \sqrt{\frac{2ns^2}{\sigma_0^2}} - \sqrt{2n-3}$	$\sigma^2 \neq \sigma_0^2$	$(-\infty, -u(1 - \frac{\alpha}{2})) \cup [u(1 - \frac{\alpha}{2}), +\infty)$
	$\sigma^2 < \sigma_0^2$	$(-\infty, u(1 - \alpha)]$
	$\sigma^2 > \sigma_0^2$	$[u(1 - \alpha), +\infty)$

Hipotezy parametryczne - wskaźnik struktury

Założenia:

- 1 Próba jest mała $n < 100$.

Stawiamy hipotezę zerową $H_0 : p = p_0$

Statystyka testowa	H_1	Obszar krytyczny W
U^*	$p \neq p_0$	$(-\infty, -u(1 - \frac{\alpha}{2})) \cup [u(1 - \frac{\alpha}{2}), +\infty)$
	$p < p_0$	$(-\infty, -u(1 - \alpha))$
	$p > p_0$	$[u(1 - \alpha), +\infty)$

gdzie

$$U^* = (2 \arcsin \sqrt{\frac{m}{n}} - 2 \arcsin \sqrt{p_0}) \sqrt{n}$$

Hipotezy parametryczne - wskaźnik struktury

Założenia

- 1 Próba jest duża $n \geq 100$.

Stawiamy hipotezę zerową $H_0 : p = p_0$

Statystyka testowa	H_1	Obszar krytyczny W
$U = \frac{m - np_0}{\sqrt{np_0(1-p_0)}}$	$p \neq p_0$	$(-\infty, -u(1 - \frac{\alpha}{2})) \cup [u(1 - \frac{\alpha}{2}), +\infty)$
	$p < p_0$	$(-\infty, -u(1 - \alpha)]$
	$p > p_0$	$[u(1 - \alpha), +\infty)$

Hipotezy parametryczne - równość średnich

Test z

Założenia:

- 1 Badana cecha X_1, X_2 ma w dwóch populacjach rozkłady normalne $X_1 = N(\mu_1, \sigma_1)$ i $X_2 = N(\mu_2, \sigma_2)$ (skala interwałowa lub ilorazowa),
- 2 σ_1 i σ_2 są znane.

Stawiamy hipotezę zerową $H_0 : \mu_1 = \mu_2$

Statystyka testowa	H_1	Obszar krytyczny W
$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\mu_1 \neq \mu_2$	$(-\infty, -u(1 - \frac{\alpha}{2})] \cup [u(1 - \frac{\alpha}{2}), +\infty)$
	$\mu_1 < \mu_2$	$(-\infty, -u(1 - \alpha)]$
	$\mu_1 > \mu_2$	$[u(1 - \alpha), +\infty)$

Hipotezy parametryczne - równość średnich

Test z

Założenia:

- 1 Próbę jest duża ($n_1 \geq 100, n_2 \geq 100$),
- 2 Skala interwałowa lub ilorazowa

Stawiamy hipotezę zerową $H_0 : \mu_1 = \mu_2$

Statystyka testowa	H_1	Obszar krytyczny W
$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$\mu_1 \neq \mu_2$	$(-\infty, -u(1 - \frac{\alpha}{2})] \cup [u(1 - \frac{\alpha}{2}), +\infty)$
	$\mu_1 < \mu_2$	$(-\infty, -u(1 - \alpha)]$
	$\mu_1 > \mu_2$	$[u(1 - \alpha), +\infty)$

Hipotezy parametryczne - równość średnich

Test t

Założenia:

- 1 Badana cecha X_1, X_2 ma w dwóch populacjach rozkłady normalne $X_1 = N(\mu_1, \sigma_1)$ i $X_2 = N(\mu_2, \sigma_2)$ (skala interwałowa lub ilorazowa),
- 2 σ_1 i σ_2 są nieznane, ale $\sigma_1 = \sigma_2$ (homogeniczność wariancji)

Stawiamy hipotezę zerową $H_0 : \mu_1 = \mu_2$

Statystyka testowa	H_1	Obszar krytyczny W
$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$	$\mu_1 \neq \mu_2$	*
	$\mu_1 < \mu_2$	$(-\infty, -t(1 - \alpha, n_1 + n_2 - 2)]$
	$\mu_1 > \mu_2$	$[t(1 - \alpha, n_1 + n_2 - 2), +\infty)$

$$* = (-\infty, -t(1 - \frac{\alpha}{2}, n_1 + n_2 - 2)] \cup [t(1 - \frac{\alpha}{2}, n_1 + n_2 - 2), +\infty)$$

Hipotezy parametryczne - równość średnich

Test C Cochra i Coxa

Założenia:

- 1 Badana cecha X_1, X_2 ma w dwóch populacjach rozkłady normalne $X_1 = N(\mu_1, \sigma_1)$ i $X_2 = N(\mu_2, \sigma_2)$ (skala interwałowa lub ilorazowa),
- 2 σ_1 i σ_2 są nieznane, ale $\sigma_1 \neq \sigma_2$ (heterogeniczność wariancji)

Stawiamy hipotezę zerową $H_0 : \mu_1 = \mu_2$

Statystyka testowa	H_1	Obszar krytyczny W
$C = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$	$\mu_1 \neq \mu_2$	*
	$\mu_1 < \mu_2$	$(-\infty, -c(1 - \alpha, n_1, n_2)]$
	$\mu_1 > \mu_2$	$[c(1 - \alpha, n_1, n_2), +\infty)$

$$* = (-\infty, -c(1 - \frac{\alpha}{2}, n_1, n_2)] \cup [c(1 - \frac{\alpha}{2}, n_1, n_2), +\infty),$$

gdzie kwantyle rozkładu C dane są przybliżonym wzorem:

$$c(p, n_1, n_2) \approx \frac{\frac{s_1^2}{n_1 - 1} t(p, n_1 - 1) + \frac{s_2^2}{n_2 - 1} t(p, n_2 - 1)}{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

Hipotezy parametryczne - równość wskaźników struktury

Założenia:

- ❶ Próba jest duża $n_1 \geq 100$, $n_2 \geq 100$.

Stawiamy hipotezę zerową $H_0 : p_1 = p_2$

Statystyka testowa	H_1	Obszar krytyczny W
$U = \frac{p_1^* - p_2^*}{\sqrt{\frac{p^*(1-p^*)}{n^*}}}$	$p_1 \neq p_2$	$(-\infty, -u(1 - \frac{\alpha}{2})] \cup [u(1 - \frac{\alpha}{2}), +\infty)$
	$p_1 < p_2$	$(-\infty, -u(1 - \alpha)]$
	$p_1 > p_2$	$[u(1 - \alpha), +\infty)$

gdzie

$$p_1^* = \frac{m_1}{n_1}, \quad p_2^* = \frac{m_2}{n_2}, \quad p^* = \frac{m_1 + m_2}{n_1 + n_2}, \quad n^* = \frac{n_1 n_2}{n_1 + n_2}$$

Hipotezy parametryczne - równość wskaźników struktury

Założenia:

- ❶ Próba jest mała $n_1 < 100$, lub $n_2 < 100$.

Stawiamy hipotezę zerową $H_0 : p_1 = p_2$

Statystyka testowa	H_1	Obszar krytyczny W
U^*	$p_1 \neq p_2$	$(-\infty, -u(1 - \frac{\alpha}{2})] \cup [u(1 - \frac{\alpha}{2}), +\infty)$
	$p_1 < p_2$	$(-\infty, -u(1 - \alpha)]$
	$p_1 > p_2$	$[u(1 - \alpha), +\infty)$

gdzie

$$U^* = (2 \arcsin \sqrt{p_1^*} - 2 \arcsin \sqrt{p_2^*}) \sqrt{n^*},$$

oraz

$$p_1^* = \frac{m_1}{n_1}, \quad p_2^* = \frac{m_2}{n_2}, \quad n^* = \frac{n_1 n_2}{n_1 + n_2}$$

Problemy

Mamy dwie grupy badanych osób: Grupa A z rakiem: $n_1 = 200$

Grupa B zdrowi: $n_2 = 300$

Liczba palaczy w każdej grupie wynosi odpowiednio:

Grupa A $m_1 = 183$

Grupa B $m_2 = 156$

Chcemy wiedzieć czy proporcje palaczy w obu grupach są takie same?

1. $H_o : p_A = p_B$

2. $H_o : p_A \leq p_B$

3. $H_o : p_A \geq p_B$

1. $H_1 : p_A \neq p_B$

2. $H_1 : p_A > p_B$

3. $H_1 : p_A < p_B$

Hipotezy parametryczne - przykłady

Czas pracy baterii pewnego rodzaju ma rozkład $N(\mu, 70)$. Na poziomie istotności $1 - \alpha = 0,95$ zweryfikować hipotezę, że przeciętny czas pracy tego typu baterii wynosi ponad 500 godzin, jeśli dla 16 losowo wybranych baterii otrzymano $\bar{x} = 560$.

- ❶ Hipotezy o średniej - porównanie z wartością referencyjną, badana cecha ma rozkład normalny o znanym σ - stosujemy test z.
- ❷ $H_0 : \mu = \mu_0 = 500, \quad H_1 : \mu > \mu_0 = 500$
- ❸ $W = [u(1 - \alpha), \infty) = [1,64; \infty)$
- ❹ $U = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} = \frac{560 - 500}{70} \sqrt{16} = 3,43$
- ❺ $U \in W$ - odrzucamy hipotezę H_0 na rzecz hipotezy H_1 .

Hipotezy parametryczne - przykłady

Dokonano 100 pomiarów opóźnień autobusów w stosunku do rozkładu. Otrzymano $\bar{x} = 8$, $s = 4$. Zakładając, że czas opóźnień ma rozkład normalny, zweryfikować na poziomie istotności $1 - \alpha = 0,99$ hipotezę, że wariancja opóźnień jest równa 9.

- 1 Hipotezy o wariancji - porównanie z wartością referencyjną, badana cecha ma rozkład normalny, próba jest duża - stosujemy test 2 dla wariancji.
- 2 $H_0 : \sigma^2 = \sigma_0^2 = 9, \quad H_1 : \sigma^2 \neq \sigma_0^2 = 9$
- 3 $W = (-\infty; -u(1 - \frac{\alpha}{2})] \cup [u(1 - \frac{\alpha}{2}; \infty) = (-\infty; -2,33] \cup [2,33; \infty)$
- 4 $U = \sqrt{\frac{2ns^2}{\sigma_0^2}} - \sqrt{2n-3} = \sqrt{\frac{2 \cdot 100 \cdot 16}{9}} - \sqrt{197} = 4,82$
- 5 $U \in W$ - odrzucamy hipotezę H_0 na rzecz hipotezy H_1 .

Hipotezy parametryczne - przykłady

W pewnym przedsiębiorstwie wylosowano niezależnie do próby 250 kobiet i 380 mężczyzn, i spytano ich, czy są skłonni zmienić swoje miejsce pracy. Twierdząco odpowiedziało 100 kobiet i 280 mężczyzn. Czy na poziomie istotności $1 - \alpha = 0,99$ można stwierdzić, że kobiety w większym stopniu boją się zmiany pracy niż mężczyźni?

① Hipotezy o wskaźniku struktury - porównania międzygrupowe, próby są duże - stosujemy test 1 dla wskaźników struktury

② $H_0 : p_1 = p_2, \quad H_1 : p_1 < p_2$

③ $W = (-\infty; -u(1 - \alpha)] = (-\infty; -2,33]$

④

$$p_1^* = \frac{100}{250} = 0,4, \quad p_2^* = \frac{280}{380} = 0,737, \quad p^* = \frac{380}{630} = 0,603, \quad n^* = \frac{250 \cdot 380}{630}$$

$$n^* = 150,8, \quad U = \frac{p_1^* - p_2^*}{\sqrt{\frac{p^*(1-p^*)}{n^*}}} = \frac{0,4 - 0,737}{\sqrt{\frac{0,603 \cdot 0,397}{150,8}}} = -8,458$$

⑤ $U \in W$ - odrzucamy hipotezę H_0 na rzecz hipotezy H_1 .

Hipotezy nieparametryczne - rodzaje

- ❶ Testy jednorodności wariancji: test F Fishera-Snedecora, test Levene'a, test Browna-Forsythe'a
- ❷ Testy zgodności rozkładów: test Kołmogorowa-Smirnowa, test Shapiro-Wilka, test Lilieforsa, test χ^2
- ❸ Testy istotności różnic
 - ❶ Dla 2 grup: test U Manna-Whitneya, test Fishera, testy parametryczne (z,t i Cohrana i Coxa)
 - ❷ Dla wielu grup: ANOVA i MANOVA, test Kruskala-Wallisa, test χ^2
 - ❸ Testy dla danych zależnych: test Wilcoxona, test McNemary, test Friedmana, test Kendalla
- ❹ Inne testy: test losowości próby (test serii), test znaków, test niezależności χ^2

Rangi

Podczas przeprowadzania testów nieparametrycznych, często zachodzi konieczność nadania rang uzyskanym obserwacjom (rangowania). Rangowane dane muszą być na skali co najmniej porządkowej. W procedurze rangowania, należy najpierw uporządkować rosnąco zebrane dane, a następnie nadaje się im wartości (zwykle są to kolejne liczby całkowite).

Rangowanie - przykład

Na przykład dla zmiennej o następujących wartościach: 8.6, 5.3, 8.6, 7.1, 9.3, 7.2, 7.3, 7.4, 7.3, 5.2, 7, 9.9, 8.6, 5.7 przypisywane są następujące rangi:

posortowane wartości zmiennej	rangi
5.2	1
5.3	2
5.7	3
7	4
7.1	5
7.2	6
7.3	7.5
7.3	7.5
7.4	9
8.6	11
8.6	11
8.6	11
9.3	13
9.9	14

Testy jednorodności wariancji

Jednym z istotnych warunków stosowalności niektórych testów porównań międzygrupowych jest równość wariancji badanej cechy w porównywanych grupach (jednorodność wariancji, homoskedastyczność). Testami weryfikującymi hipotezę o jednorodności wariancji w porównywanych grupach są testy:

- ❶ Dla 2 grup porównawczych:
 - ❶ Test F Fishera-Snedecora
- ❷ Dla więcej niż 2 grup porównawczych:
 - ❶ Test Levene'a
 - ❷ Test Browna-Forsythe'a
 - ❸ Test Bartletta

Testy jednorodności wariancji

Test F Fishera-Snedecora

Założenia:

- 1 Badana cecha X_1, X_2 ma w dwóch populacjach rozkłady normalne $X_1 = N(\mu_1, \sigma_1)$ i $X_2 = N(\mu_2, \sigma_2)$,
- 2 Dane na skali interwałowej lub ilorazowej

Stawiamy hipotezę zerową $H_0 : \sigma_1^2 = \sigma_2^2$

Statystyka testowa	H_1	Obszar krytyczny W
$F = \frac{s_1^2}{s_2^2}$	$\sigma_1^2 \neq \sigma_2^2$	$(0, \min\{F^*, \frac{1}{F^*}\}) \cup [\max\{F^*, \frac{1}{F^*}\}, +\infty)$
	$\sigma_2^2 < \sigma_1^2$	$[F(1 - \alpha, n_1 - 1, n_2 - 1), +\infty)$
$F = \frac{s_2^2}{s_1^2}$	$\sigma_2^2 > \sigma_1^2$	$[F(1 - \alpha, n_2 - 1, n_1 - 1), +\infty)$

$$F^* = F\left(\frac{\alpha}{2}, n_1 - 1, n_2 - 1\right).$$

Test Levene'a

Założenia:

- 1 Dane na skali interwałowej lub ilorazowej
- 2 Porównujemy ze sobą $k \geq 2$ grup

Stawiamy hipotezy:

$$H_0 : \sigma_1 = \sigma_2 = \dots = \sigma_k$$

$$H_1 : \sigma_i \neq \sigma_j \text{ dla dowolnego } i \text{ oraz } j$$

Wyznaczamy wartości średnie w każdej z grup ($x_{.j}$). Wartości obserwacji x_{ij} zastępujemy wartościami $z_{ij} = |x_{ij} - x_{.j}|$.

Test Levene'a

Statystyka testowa:

$$F = \frac{n - k}{k - 1} \cdot \frac{\sum_{j=1}^k N_j (z_{.j} - z_{..})^2}{\sum_{j=1}^k \sum_{i=1}^{n_j} (z_{ij} - z_{.j})^2}$$

gdzie: $z_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} z_{ij}$ - średnia wartość z_{ij} , w j -tej grupie

$z_{..} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} z_{ij}$ - średnia wartość spośród wszystkich wartości z_{ij}

Obszar krytyczny:

$$W = [F(1 - \alpha, k - 1, n - k); \infty)$$

Test Browna - Forsythe'a

Założenia:

- 1 Dane na skali interwałowej lub ilorazowej
- 2 Porównujemy ze sobą $k \geq 2$ grup

Stawiamy hipotezy:

$$H_0 : \sigma_1 = \sigma_2 = \dots = \sigma_k$$

$$H_1 : \sigma_i \neq \sigma_j \text{ dla dowolnego } i \text{ oraz } j$$

Wyznaczamy wartości mediany w każdej z grup ($x_{.j}$). Wartości obserwacji x_{ij} zastępujemy wartościami $z_{ij} = |x_{ij} - x_{.j}|$.

Testy jednorodności wariancji

Test Browna - Forsythe'a

Statystyka testowa:

$$F = \frac{n - k}{k - 1} \cdot \frac{\sum_{j=1}^k N_j (z_{.j} - z_{..})^2}{\sum_{j=1}^k \sum_{i=1}^{n_j} (z_{ij} - z_{.j})^2}$$

gdzie: $z_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} z_{ij}$ - średnia wartość z_{ij} , w j -tej grupie

$z_{..} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} z_{ij}$ - średnia wartość spośród wszystkich wartości z_{ij}

Obszar krytyczny:

$$W = [F(1 - \alpha, k - 1, n - k); \infty)$$

Test Browna - Forsythe'a jest bardziej odporny względem odstających obserwacji, od testu Levena. Z tego względu najczęściej stosuje się go przy asymetrycznych rozkładach.

Listing 1: Testy wariancji

```
# Wizualizacja
boxplot(utrata_wagi ~ dieta, data = data,
        main = "Utrata wagi w dietach",
        xlab = "Dieta", ylab = "Utrata wagi",
        col = "steelblue", border = "black")

# Test Browna–Forsythe'a
library(onewaytests)
bf.test(utrata_wagi ~ dieta, data = data)
```

Testy zgodności

Testy zgodności, są grupą testów, w których porównujemy rozkład badanej cechy do dowolnego rozkładu. Hipoteza zerowa i alternatywna w tego typu testach mają postać:

$$H_0 : F_n(x) \equiv F(x), \quad H_1 : F_n(x) \not\equiv F(x)$$

gdzie $F_n(x)$ jest dystrybuantą empiryczną wyznaczoną w oparciu o pobraną próbę, a $F(x)$ jest dystrybuantą rozkładu do którego porównujemy badany rozkład, lub

$$H_0 : P[X = x_i] = p_i, \quad i = 1, 2, \dots, k$$

w przypadku rozkładów nie posiadających dystrybuanty (np. danych na skali nominalnej).

Szczególną podgrupą testów zgodności są testy normalności rozkładu, w których rozkładem do którego porównujemy rozkład badanej cechy jest rozkład normalny $N(\mu, \sigma)$, zatem $F(x) = \Phi(\frac{x-\mu}{\sigma})$ jest dystrybuantą tego rozkładu.

Test Kołmogorowa

Założenia:

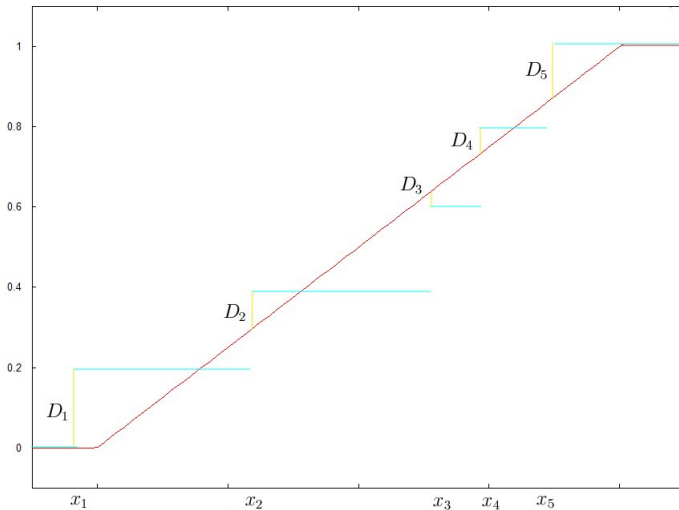
- 1 Badana cecha ma rozkład ciągły
- 2 Dane na skali interwałowej lub ilorazowej
- 3 Znane powinny być parametry μ i σ

Stawiamy hipotezę zerową $H_0 : F_n(x) \equiv F(x)$

Statystyka testowa	H_1	Obszar krytyczny W
$D = \sup_x F_n(x) - F(x) \sqrt{n}$	$F_n(x) \not\equiv F(x)$	$[\lambda(1 - \alpha), \infty)$

gdzie $\lambda(p)$ jest kwantylem rzędu p rozkładu λ Kołmogorowa - Smirnova.

Test Kołmogorowa - Smirnova



Test Kołmogorowa - Smirnova dla 2 rozkładów

Założenia:

- 1 Badana cecha ma rozkład ciągły
- 2 Dane na skali interwałowej lub ilorazowej
- 3 Znane powinny być parametry μ_1 , σ_1 , μ_2 i σ_2

Stawiamy hipotezę zerową $H_0 : F_{n_1}(x) \equiv F_{n_2}(x)$

Statystyka testowa	H_1	Obszar krytyczny W
$D = \sup_x F_n(x) - F(x) \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$	$F_{n_1}(x) \not\equiv F_{n_2}(x)$	$[\lambda(1 - \alpha), \infty)$

gdzie $\lambda(p)$ jest kwantylem rzędu p rozkładu λ Kołmogorowa - Smirnova.

Test Lillieforsa

Założenia:

- 1 Porównujemy rozkład z próby z rozkładem normalnym
- 2 Dane na skali interwałowej lub ilorazowej
- 3 Nie są znane parametry μ i σ .

Stawiamy hipotezę zerową $H_0 : F_n(x) \equiv F(x)$, gdzie $F(x)$ jest dystrybuantą rozkładu normalnego $N(\mu, \sigma)$.

Statystykę testową D wyznacza się tak samo jak w przypadku testu Kołmogorowa-Smirnova, ale do wyznaczenia obszaru krytycznego zamiast rozkładu Kołmogorowa-Smirnova stosuje się rozkład Lillieforsa.

Testy zgodności

Test Shapiro-Wilka

Założenia:

- 1 Porównujemy rozkład z próby z rozkładem normalnym
- 2 Dane na skali interwałowej lub ilorazowej

Stawiamy hipotezę zerową $H_0 : F_n(x) \equiv F(x)$, gdzie $F(x)$ jest dystrybuantą rozkładu normalnego $N(\mu, \sigma)$.

W pierwszym kroku porządkujemy rosnąco wartości pobranej próby $(x_1 \leq x_2 \leq \dots \leq x_n)$, a następnie wyznaczamy wartość statystyki W

$$W = \frac{(\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} a_i(n)(x_{n-i+1} - x_i))^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

gdzie $a_i(n)$ są współczynnikami dla testu Shapiro-Wilka (można odczytać je z tabel współczynników testu Shapiro-Wilka).

Statystyka testowa	H_1	Obszar krytyczny W
W	$F_n(x) \not\equiv F(x)$	$[W(\frac{\alpha}{2}, n), W(1 - \frac{\alpha}{2}, n)]$

Testy zgodności

Test χ^2 Pearsona

Założenia:

- 1 Dane na skali co najmniej nominalnej (dowolna skala)
- 2 Liczebności i liczebności empiryczne dla każdej wartości ≥ 5

Stawiamy hipotezę zerową $H_0 : P[X = x_i] = p_i$ lub $F_n(x) = F(x)$

Statystyka testowa	H_1	Obszar krytyczny W
$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$	$P[X = x_i] \neq p_i$	$[\chi^2(1 - \alpha, k - r - 1); \infty)$

gdzie o_i jest liczbą obserwacji dla wartości x_i , a e_i jest teoretycznie wyznaczoną liczbą obserwacji dla wartości x_i ($e_i = n \cdot p_i$), k jest liczbą różnych wartości x_i a r liczbą parametrów rozkładu estymowanych z próby.

Test χ^2 Pearsona - przykład

W celu sprawdzenia, czy kostka jest dobrze wyważona wykonano 120 rzutów i uzyskano wyniki zawarte w tabeli. Na poziomie istotności $1 - \alpha = 0,95$, zweryfikować hipotezę, że rzut tą kostką jest uczciwy.

	x_i					
	1	2	3	4	5	6
o_i	20	22	17	18	19	24
e_i	20	20	20	20	20	20

$$\chi^2 = \frac{(20 - 20)^2}{20} + \frac{(22 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(18 - 20)^2}{20} + \frac{(19 - 20)^2}{20} + \frac{(24 - 20)^2}{20} = 1,7$$

$$W = [\chi^2(1 - \alpha, k - r - 1); \infty) = [\chi^2(0,95; 5); \infty) = [11,07; \infty)$$

$\chi^2 \notin W$ zatem nie ma podstaw do odrzucenia hipotezy, iż rzut tą kostką jest uczciwy.

Test χ^2 Pearsona - przykład

Na podstawie zawartych w tabeli danych dotyczących czasu działania 40 baterii, na typowym poziomie istotności zweryfikować hipotezę, że ten czas ma rozkład $N(3,5; 0,7)$.

Czas działania	o_i	e_i
$1,45 - 1,95$ $1,95 - 2,45$ $2,45 - 2,95$	$\left. \begin{matrix} 2 \\ 1 \\ 4 \end{matrix} \right\} 7$	$\left. \begin{aligned} 40 \cdot \Phi\left(\frac{1,95-3,5}{0,7}\right) &= 0,5 \\ 40 \cdot \left[\Phi\left(\frac{2,45-3,5}{0,7}\right) - \Phi\left(\frac{1,95-3,5}{0,7}\right)\right] &= 2,1 \\ 40 \cdot \left[\Phi\left(\frac{2,95-3,5}{0,7}\right) - \Phi\left(\frac{2,45-3,5}{0,7}\right)\right] &= 5,9 \end{aligned} \right\} 8,5$
$2,95 - 3,45$	15	$40 \cdot \left[\Phi\left(\frac{3,45-3,5}{0,7}\right) - \Phi\left(\frac{2,95-3,5}{0,7}\right)\right] = 10,3$
$3,45 - 3,95$	10	$40 \cdot \left[\Phi\left(\frac{3,95-3,5}{0,7}\right) - \Phi\left(\frac{3,45-3,5}{0,7}\right)\right] = 10,7$
$3,95 - 4,45$ $4,45 - 4,95$	$\left. \begin{matrix} 5 \\ 3 \end{matrix} \right\} 8$	$\left. \begin{aligned} 40 \cdot \left[\Phi\left(\frac{4,45-3,5}{0,7}\right) - \Phi\left(\frac{3,95-3,5}{0,7}\right)\right] &= 7,0 \\ 40 \cdot \left[\Phi\left(\frac{4,95-3,5}{0,7}\right) - \Phi\left(\frac{4,45-3,5}{0,7}\right)\right] &= 3,5 \end{aligned} \right\} 10,5$

Test χ^2 Pearsona - przykład c.d.

$$H_0 : F_n(x) \equiv \Phi\left(\frac{x - 3,5}{0,7}\right), \quad H_1 : F_n(x) \not\equiv \Phi\left(\frac{x - 3,5}{0,7}\right)$$

$$W = [\chi^2(1 - \alpha, k - r - 1); \infty) = [\chi^2(0,95; 3); \infty) = [7,815; \infty)$$

$$\chi^2 = \frac{(7 - 8,5)^2}{8,5} + \frac{(15 - 10,3)^2}{10,3} + \frac{(10 - 10,7)^2}{10,7} + \frac{(8 - 10,5)^2}{10,5} = 3,05$$

$\chi^2 \notin W$, zatem nie ma podstaw do odrzucenia hipotezy, że rozkład czasu pracy baterii tego typu ma rozkład normalny ze średnią $\mu = 3,5$ i odchyleniem standardowym $\sigma = 0,7$.

Inne testy zgodności:

- 1 Test Cramera von Misesa
- 2 Test Andersona - Darlinga
- 3 Test Watsona
- 4 Test Jarque - Bera
- 5 Test Shapiro - Francia
- 6 Test D'Agostino

Testy istotności różnic

Testy istotności różnic są testami w których porównujemy ze sobą 2 lub więcej badanych grup. W przypadku danych na skali interwałowej lub ilorazowej porównujemy parametry (najczęściej średnią) w badanych grupach. Testy te dzielą się na 2 rodzaje, w zależności od tego, czy porównujemy ze sobą 2, czy więcej grup. W testach istotności różnic porównujących ze sobą 2 grupy hipoteza zerowa i alternatywna mają postać:

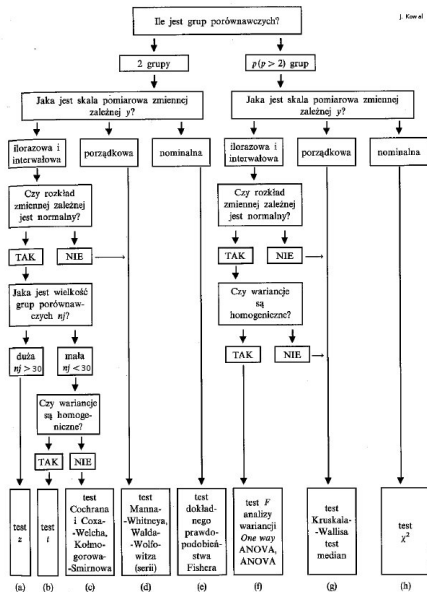
$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2 \quad (\mu_1 > \mu_2 \text{ lub } \mu_1 < \mu_2),$$

a w testach porównujących więcej niż 2 grupy:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k, \quad H_1 : \mu_i \neq \mu_j \text{ dla pewnego } i \text{ i } j.$$

W przypadku danych na skali nominalnej lub porządkowej porównujemy rozkłady badanej cechy w obu grupach, a zatem są to wówczas testy zgodności (postaci hipotez są takie jak w testach zgodności).

Algorytm wyboru testu istotności różnic



Testy istotności różnic

Test serii Walda - Wolfowitza

Założenia:

- 1 Dane na skali co najmniej porządkowej
- 2 Porównujemy ze sobą rozkłady w 2 grupach

Weryfikujemy hipotezę

$$H_0 : P[X_1 = x_i] = P[X_2 = x_i] \text{ dla wszystkich } i$$

$$H_1 : P[X_1 = x_i] \neq P[X_2 = x_i] \text{ dla pewnego } i$$

$\max(n_1, n_2)$	Statystyka testowa	Obszar krytyczny W
≤ 20	U_{n_1, n_2} - liczba serii	$(0; k]$
> 20	$U_{n_1, n_2}^* = \frac{U_{n_1, n_2} - 1 - \frac{2n_1 n_2}{n_1 + n_2}}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - (n_1 + n_2))}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}}$	$(-\infty, -u(1 - \alpha)]$

Wartość k odczytujemy z tablic dla liczby serii

Test serii Walda - Wolfowitza – przykład

Podczas połowu ryb wyciągnięto 12 sieci w rejonie A i 12 sieci w rejonie B. Odsetki złowionych śledzi kształtowały się następująco:

A	60,2	70,4	36,0	44,0	42,0	68,0	70,0	70,1	53,0
B	30,4	20,6	70,4	60,3	36,1	49,1	69,1	73,2	44,1
A	59,0	68,9	70,0						
B	32,6	48,0	40,0						

Zweryfikować hipotezę, że odsetki złowionych łososi w obu rejonach są jednakowe. Sortujemy rosnąco uzyskane wyniki z uwzględnieniem grup do jakich należały:

B	B	B	A	B	B	A	A
20,6	30,4	32,6	36,0	36,1	40,0	42,0	44,0
B	B	B	A	A	A	B	A
44,1	48,0	49,1	53,0	59,0	60,2	60,3	68,0
A	B	A	A	A	A	B	B
68,9	69,1	70,0	70,0	70,1	70,4	70,5	73,2

Test serii Walda - Wolfowitza – przykład C.D.

Ostatecznie uzyskujemy ciąg serii:

BBB A BB AA BBB AAA B AA B AAAA BB

Zatem liczba serii $U_{12,12} = 11$ Wartość k dla $n_1 = 12$, $n_2 = 12$, $\alpha = 0,05$ jest równa 8.

$$W = (0; 8]$$

$U_{12,12} \notin W$ zatem nie ma podstaw do odrzucenia hipotezy H_0 , że odsetki złowionych łososi w obu rejonach są jednakowe.

Testy istotności różnic

Test U Manna - Whitneya

Założenia:

- 1 Dane na skali co najmniej porządkowej

Postaci hipotez:

$$H_0 : P[X_1 = x_i] = P[X_2 = x_i] \text{ dla wszystkich } i,$$

$$H_1 : P[X_1 = x_i] \neq P[X_2 = x_i] \text{ dla pewnego } i$$

Nadajemy rangi zebranych obserwacjom i sumujemy rangi w pierwszej i drugiej grupie uzyskując wartości R_1 i R_2 .

Statystyka testowa:

$$U_i = n_1 n_2 + \frac{n_i(n_i + 1)}{2} - R_i, \quad U = \min(U_1, U_2)$$

$\max(n_1, n_2)$	Statystyka testowa	Obszar krytyczny W
≤ 20	U	Odczytujemy z tablic
> 20	$U^* = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$	$(-\infty, -u(1 - \frac{\alpha}{2})] \cup [u(1 - \frac{\alpha}{2}), \infty)$

Testy istotności różnic

Test dokładnego prawdopodobieństwa Fishera

Założenia:

- 1 Dane na skali nominalnej
- 2 Badana cecha jest dychotomiczna (przyjmuje tylko 2 możliwe wartości)

Postaci hipotez badawczych:

$$H_0 : P[X_1 = x_1] = P[X_2 = x_1], \quad H_1 : P[X_1 = x_1] \neq P[X_2 = x_1]$$

Tworzymy tabelę licznosci:

	x_1	x_2	
X_1	a	b	$a + b$
X_2	c	d	$c + d$
	$a + c$	$b + d$	N

Obszar krytyczny: $W = [0; \alpha]$

Statystyka testowa: $p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$

Porównania wielu grup

W testach porównujących ze sobą wiele grup, oprócz wartości badanej cechy, mamy także zmienne, których wartości określają przydział obserwacji do jednej z porównywanych grup.

Przykładowo można porównywać wyniki badań pacjentów w zależności od ich grupy krwi. Zmienną określającą przydział do grup jest wówczas grupa krwi (mamy 4 grupy: 0, A, B, AB). Taką analizę nazywamy jednoczynnikową (tylko jedna zmienna decyduje o przydziale do konkretnej grupy - grupa krwi).

Może się również zdarzyć, iż o przydziale do grup decyduje wartość więcej niż jednej zmiennej, wówczas mamy do czynienia z analizą wieloczynnikową. Przykładowo, gdybyśmy w badaniach oprócz grupy krwi pacjentów uwzględnili także czynnik Rh mielibyśmy wówczas 8 grup zdeteminowanych przez 2 zmienne (grupa krwi i czynnik Rh).

Porównania wielu grup

W przypadku porównań wielu grup, gdy wartości badanej cechy są na skali co najmniej interwałowej (test parametryczny) stosujemy analizę wariancji ANOVA (jednoczynnikową lub wieloczynnikową).

W przypadku, gdy badana cecha jest na skali porządkowej stosujemy test Kruskala - Wallisa, (szczególnym przypadkiem tego testu jest test U Manna - Whitneya który stosuje się w przypadku porównań 2 grup).

W przypadku danych na skali nominalnej stosujemy test χ^2 .

ANOVA jednoczynnikowa

Założenia:

- 1 Badana cecha ma rozkład normalny
- 2 Dane na skali interwałowej lub ilorazowej
- 3 Wariancje są homogeniczne
- 4 Liczebności obserwacji w grupach nie powinny się istotnie różnić (o więcej niż 10%)

Postaci hipotez:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k, \quad H_1 : \mu_i \neq \mu_j \text{ dla któregoś } i \text{ i } j,$$

gdzie k jest liczbą grup porównawczych.

Wyznaczamy średnią wartość cechy \bar{x} , oraz średnie w każdej z grup \bar{x}_j .

ANOVA jednoczynnikowa

Wyznaczamy wariancje:

❶ Międzygrupową

$$SSE = \frac{\sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2}{k - 1}$$

❷ Wewnątrzgrupową

$$SSB = \frac{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n - k}$$

Statystyka testowa	Obszar krytyczny W
$F = \frac{SSE}{SSB}$	$W = [F(1 - \alpha; k - 1; n - k), \infty)$

W przypadku odrzucania hipotezy H_0 , wyznacza się tzw. kontrasty, czyli określa pomiędzy którymi grupami różnice są statystycznie istotne. W tym celu stosuje się testy: test NIR (najmniejszych istotnych różnic ang. LSD), Duncana, Tukeya, Scheffego lub Bonferoniego.

ANOVA dwuczynnikowa

Założenia:

- 1 Badana cecha ma rozkład normalny
- 2 Dane na skali interwałowej lub ilorazowej
- 3 Wariancje są homogeniczne
- 4 Liczebności obserwacji w poszczególnych grupach są równe.

Postaci hipotez:

$$H_0^A : \mu_{1.} = \mu_{2.} = \dots = \mu_{w.}, \quad H_1^A : \mu_{i.} \neq \mu_{j.} \text{ dla któregoś } i \text{ i } j,$$

$$H_0^B : \mu_{.1} = \mu_{.2} = \dots = \mu_{.k}, \quad H_1^B : \mu_{.i} \neq \mu_{.j} \text{ dla któregoś } i \text{ i } j,$$

$$H_0^{AB} : \mu_{11} = \mu_{12} = \dots = \mu_{wk}, \quad H_1^{AB} : \mu_{ij} \neq \mu_{pr} \text{ dla któregoś } ij \text{ i } pr,$$

Wyznaczamy średnie wartości cechy \bar{x} , oraz średnie w każdej z grup $\bar{x}_{i.}$, $\bar{x}_{.j}$ i \bar{x}_{ij} .

ANOVA dwuczynnikowa

Wyznaczamy wariancje:

- ① Międzygrupową względem A

$$SSE_A = \frac{km \sum_{i=1}^w (\bar{x}_{i..} - \bar{x})^2}{w - 1}$$

- ② Międzygrupową względem B

$$SSE_B = \frac{wm \sum_{j=1}^k (\bar{x}_{.j.} - \bar{x})^2}{k - 1}$$

- ③ Międzygrupową łączną

$$SSE_{A \times B} = \frac{m \sum_{i=1}^w \sum_{j=1}^k (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2}{(w - 1)(k - 1)}$$

- ④ Wewnątrzgrupową

$$SSB = \frac{\sum_{i=1}^w \sum_{j=1}^k \sum_{s=1}^m (x_{ijs} - \bar{x}_{ij.})^2}{wk(m - 1)}$$

ANOVA dwuczynnikowa

Hipoteza	Statystyka	Obszar krytyczny W
H_0^A	$F = \frac{SSE_A}{SSB}$	$W = [F(1 - \alpha; w - 1; n - wk), \infty)$
H_0^B	$F = \frac{SSE_B}{SSB}$	$W = [F(1 - \alpha; k - 1; n - wk), \infty)$
H_0^{AB}	$F = \frac{SSE_{A \times B}}{SSB}$	$W = [F(1 - \alpha; (w - 1)(k - 1); n - wk), \infty)$

Porównania wielu grup

Test Kruskala - Wallisa

Założenia:

- 1 Dane na skali co najmniej porządkowej

Postaci hipotez:

H_0 : Wartości badanej cechy nie różnią się istotnie między grupami

H_1 : Wartości badanej cechy różnią się istotnie między grupami

Zebrany danym nadajemy rangi.

Wartość statystyki testowej:

$$\chi^2 = \frac{12}{n(n+1)} \sum_{i=1}^p \frac{R_i^2}{n_i} - 3(n+1),$$

gdzie R_i jest sumą rang w i -tej grupie

Obszar krytyczny:

$$W = [\chi^2(\alpha; p-1), \infty)$$

Porównania wielu grup

Test χ^2 (niezależności)

Założenia

- 1 Dane na skali co najmniej nominalnej (dowolna skala)
- 2 Liczebności i liczebności empiryczne dla każdej wartości ≥ 5

Stawiamy hipotezy:

$$H_0 : P[X_i = x_k] = P[X_j = x_k] \text{ dla wszystkich } i, j \text{ i } k$$

$$H_1 : P[X_i = x_k] \neq P[X_j = x_k] \text{ dla którychkolwiek } i, j \text{ i } k$$

Statystyka testowa	Obszar krytyczny W
$\chi^2 = \sum_{i=1}^p \sum_{k=1}^t \frac{(o_{ik} - e_{ik})^2}{e_{ik}}$	$[\chi^2(1 - \alpha, (p - 1)(t - 1)); \infty)$

gdzie o_{ik} jest liczbą obserwacji dla wartości x_k w i -tej grupie, a e_{ik} jest teoretycznie wyznaczoną liczbą obserwacji dla wartości x_k w i -tej grupie ($e_{ik} = n_{.k} \cdot \frac{n_{i.}}{n}$), t jest liczbą różnych wartości x_i a p liczbą grup porównawczych.

Test χ^2 dla wielu grup - przykład

Zbadano poglądy wśród amerykańskich wyborców na temat prawa do aborcji i uzyskano następujące wyniki:

Prawo do aborcji	Republikanie	Demokraci	Niezależni
Za	82	70	62
Przeciw	93	62	67
Niezdecydowani	25	18	21

Na poziomie istotności $1 - \alpha = 0,95$ zwerfikować hipotezę, że nie ma istotnych różnic w poglądach na temat prawa do aborcji między zwolennikami różnych partii.

$$W = [\chi^2(0,95; 4); \infty) = [9,488; \infty)$$

Test χ^2 dla wielu grup - przykład c.d.

Prawo do aborcji	Republikanie		Demokraci		Niezależni		Razem
Za	82	85,6	70	64,2	62	64,2	214
Przeciw	93	88,8	62	66,6	67	66,6	222
Niezdecydowani	25	25,6	18	19,2	21	19,2	64
Razem	200		150		150		500

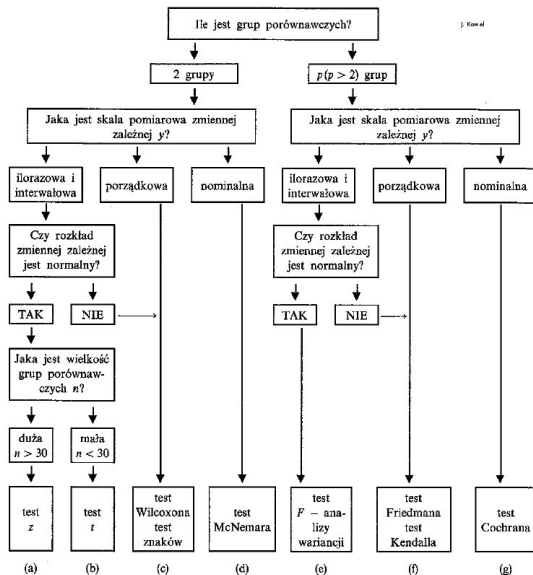
$$\chi^2 = \frac{(82 - 85,6)^2}{85,6} + \frac{(70 - 64,2)^2}{64,2} + \frac{(62 - 64,2)^2}{64,2} + \frac{(93 - 88,8)^2}{88,8} +$$
$$\frac{(62 - 66,6)^2}{66,6} + \frac{(67 - 66,6)^2}{66,6} + \frac{(25 - 25,6)^2}{25,6} + \frac{(18 - 19,2)^2}{19,2} + \frac{(21 - 19,2)^2}{19,2}$$
$$\chi^2 = 1,53 \notin W$$

Zatem nie ma podstaw do odrzucenia hipotezy, że nie ma istotnych różnic w poglądach na temat prawa do aborcji między zwolennikami różnych partii.

Testy dla danych zależnych

Przeprowadzając analizy statystyczne, zdarzają się sytuacje w których mamy do czynienia z danymi zależnymi. Najczęściej są to sytuacje w których badamy te same obiekty, poddane wpływowi pewnego czynnika. Przykładem jest badanie stanu pacjentów przed i po podaniu pewnego leku. Weryfikując hipotezę, że podanie tego leku wpływa na stan pacjentów, porównujemy ich wyniki przed i po podaniu leku.

Wybór testu istotności różnic dla danych zależnych



Testy istotności różnic dla danych zależnych

Test z dla danych zależnych

Założenia:

- 1 Dane na skali interwałowej lub ilorazowej
- 2 Rozkład zmiennej zależnej jest normalny
- 3 Próba jest duża ($n \geq 30$)
- 4 Dane są zależne

Dla wartości x_i i y_i (obserwacje przed i po doświadczeniu) wyznaczamy wartości $d_i = y_i - x_i$ i wyznaczamy z nich średnią i wariancję.

Stawiamy hipotezę zerową $H_0 : \mu_d = 0$ (doświadczenie nie ma wpływu na badane obiekty).

Statystyka testowa	H_1	Obszar krytyczny W
$U = \frac{\bar{d}}{s_d} \sqrt{n}$	$\mu_d \neq 0$	$(-\infty, -u(1 - \frac{\alpha}{2})) \cup [u(1 - \frac{\alpha}{2}), +\infty)$
	$\mu_d < 0$	$(-\infty, -u(1 - \alpha)]$
	$\mu_d > 0$	$[u(1 - \alpha), +\infty)$

Testy istotności różnic dla danych zależnych

Test t dla danych zależnych

Założenia:

- 1 Dane na skali interwałowej lub ilorazowej
- 2 Rozkład zmiennej zależnej jest normalny
- 3 Próba jest mała ($n < 30$)
- 4 Dane są zależne

Dla wartości x_i i y_i (obserwacje przed i po doświadczeniu) wyznaczamy wartości $d_i = y_i - x_i$ i wyznaczamy z nich średnią i wariancję.

Stawiamy hipotezę zerową $H_0 : \mu_d = 0$ (doświadczenie nie ma wpływu na badane obiekty).

Statystyka testowa	H_1	Obszar krytyczny W
$t = \frac{\bar{d}}{s_d} \sqrt{n-1}$	$\mu_d \neq 0$	*
	$\mu_d < 0$	$(-\infty, -t(1-\alpha; n-1)]$
	$\mu_d > 0$	$[t(1-\alpha; n-1), +\infty)$

$$* = (-\infty, -t(1 - \frac{\alpha}{2}; n - 1)] \cup [t(1 - \frac{\alpha}{2}; n - 1), +\infty)$$

Testy istotności różnic dla danych zależnych

Test T Wilcoxona

Założenia:

- 1 Dane na skali co najmniej porządkowej
- 2 Dane są zależne

Dla wartości x_i i y_i (obserwacje przed i po doświadczeniu) wyznaczamy wartości $d_i = y_i - x_i$. Dokunujemy rangowania wartości $|d_i|$ po odrzuceniu wszystkich $d_i = 0$ i sumujemy rangi dla $d_i < 0$ uzyskując wartość T_1 oraz rangi dla $d_i > 0$, uzyskując wartość T_2 . $T = \min(T_1, T_2)$

Stawiamy hipotezy $H_0 : \mu_d = 0$, $H_1 : \mu_d \neq 0$

Liczba $d_i \neq 0$	Statystyka testowa	Obszar krytyczny W
6 – 25	T	Odczytujemy z tablic
> 25	$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$	$(-\infty; -u(1 - \frac{\alpha}{2})] \cup [u(1 - \frac{\alpha}{2}); \infty)$

Testy istotności różnic dla danych zależnych

Test McNemary

Założenia:

- 1 Dane na skali nominalnej dychotomicznej (przyjmujące tylko 2 wartości)
- 2 Dane są zależne

Tworzymy tabelę licznosci względem wartości badanej cechy przed i po wykonaniu doświadczenia.

	"0"	"1"
"0"	a	b
"1"	c	d

H_0 : Doświadczenie nie miało wpływu na badaną cechę

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

$$W = [\chi^2(1 - \alpha; 1); \infty)$$

Testy istotności różnic dla danych zależnych

Test S Friedmana

Założenia:

- 1 Dane na skali co najmniej porządkowej
- 2 Dane są zależne
- 3 Porównujemy $p > 2$ grup (np. badamy te same obiekty pod wpływem p różnych czynników)

Tworzymy tablicę w której w wierszach umieszczamy obserwacje dotyczące kolejnych badanych obiektów, a w kolumnach wartości obserwacji pod wpływem kolejnych czynników. Rangujemy wyniki dla każdego obiektu z osobna (w wierszach). Obliczamy sumy rang dla każdego czynnika (w kolumnach) uzyskując wartości R_j .

Postaci hipotez:

H_0 : Żaden czynnik nie wpływa na wartość obserwacji

H_1 : Przynajmniej dla jednego czynnika różnica jest statystycznie istotna

Testy istotności różnic dla danych zależnych

Test S Friedmana

	Czynnik 1		Czynnik 2		...	Czynnik p	
obiekt	Wart1	Ranga1	Wart2	Ranga2	...	Wartp	Rangap
obiekt 1	a_{11}	R_{11}	a_{12}	R_{12}	...	a_{1p}	R_{1p}
obiekt 2	a_{21}	R_{21}	a_{22}	R_{22}	...	a_{2p}	R_{2p}
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
obiekt n	a_{n1}	R_{n1}	a_{n2}	R_{n2}	...	a_{np}	R_{np}
Suma rang		R_1		R_2	...		R_p

Statystyka testowa:

$$S = \frac{12}{np(p+1)} \sum_{i=1}^p R_i^2 - 3n(p+1)$$

Testy istotności różnic dla danych zależnych

Test S Friedmana

Wartości k i n	Obszar krytyczny W
$k = 3, n = 2, \dots, 13$	$[S(1 - \alpha, p, n); \infty)$
$k = 4, n = 2, \dots, 8$	
$k = 5, n = 3, 4, 5$	
Pozostałe	$[\chi^2(1 - \alpha; p - 1); \infty)$

Testy istotności różnic dla danych zależnych

Test Q Cochрана

Założenia:

- 1 Dane na skali nominalnej dychotomicznej (przyjmujące tylko 2 wartości: "0" i "1")
- 2 Dane są zależne
- 3 Porównujemy $p > 2$ grup (np. badamy te same obiekty pod wpływem p różnych czynników)

Tworzymy tablicę w której w wierszach umieszczamy obserwacje dotyczące kolejnych badanych obiektów, a w kolumnach wartości obserwacji pod wpływem kolejnych czynników.

Postaci hipotez:

H_0 : Żaden czynnik nie wpływa na wartość obserwacji

H_1 : Przynajmniej dla jednego czynnika różnica jest statystycznie istotna

Testy istotności różnic dla danych zależnych

Test Q Cochрана

	Czynnik 1	Czynnik 2	...	Czynnik p	Suma
obiekt 1	a_{11}	a_{12}	...	a_{1p}	$y_{1\cdot}$
obiekt 2	a_{21}	a_{22}	...	a_{2p}	$y_{2\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
obiekt n	a_{n1}	a_{n2}	...	a_{np}	$y_{n\cdot}$
Suma	$y_{\cdot 1}$	$y_{\cdot 2}$...	$y_{\cdot p}$	

Statystyka testowa:

$$Q = \frac{(p-1)p \sum_{i=1}^p y_{\cdot i}^2 - (\sum_{i=1}^p y_{\cdot i})^2}{p \sum_{k=1}^n y_{k\cdot} - \sum_{k=1}^n y_{k\cdot}^2}$$

Obszar krytyczny:

$$W = [\chi^2(1 - \alpha, p - 1); \infty)$$

Test serii (Test losowości próby)

Postać testowanej hipotezy:

$$H_0 : (X_1, X_2, \dots, X_n) \text{ jest próbą losową}$$

Wyznaczamy medianę M_e i wyznaczamy wartości $y_i = \text{sign}(x_i - M_e)$.

Wartości 1 i -1 tworzą serie których liczbę wyznaczamy.

Obszar krytyczny:

$$W = [0; R(\alpha, n)]$$

Test serii - przykład

Pobrano próbę w której badana cecha miała kolejno następujące wartości:
1,04; 1,07; 1,08; 0,96; 1,02; 1,03; 1,03; 0,92; 1,00; 0,97; 0,95;
0,99; 0,96; 1,04; 0,98

Czy ta próba jest losowa?

$$M_e = 1,00$$

Wartości y_i :

$$1, 1, 1, -1, 1, 1, 1, -1, 0, -1, -1, -1, -1, 1, -1$$

Mamy $k=7$ serii.

$$W = [0; R(0,05; 15)] = [0; 4]$$

$$k \notin W$$

Zatem nie ma podstaw do odrzucenia hipotezy, iż ta próba jest losowa.