20 Testy nieparametryczne

Testy nieparametryczne

- Dotychczas omawiane testy i przedziały ufności zakładały znajomość typu rozkładu badanej cechy w populacji z którego pochodziła próba losowa.
- W przypadku gdy nie wiemy jaki jest rozkład badanej cechy w populacji możemy oprzeć testy o statystykę opartą o estymator mediany.
- W przypadku rozkładów symetrycznych mediana i wartość oczekiwana są sobie równoważne.
- Dla rozkładów asymetrycznych mediana lepiej oddaje rozkład prawdopodobieństwa niż wartość oczekiwana.

Przedział ufności dla mediany.

Procedura określania przedziału ufności α dla mediany M dla próby prostej z n-elementowej próby z rozkładu ciągłego.

$$P(X \leqslant M) = P(X \geqslant M) = 1/2 = p$$

- Uporządkuj dane od najmniejszej do najwiekszej $X_{(1)}, X_{(2)}, ..., X_{(n)}$
- Niech N_- oznacza liczbę elementów próby losowej mniejszych od M.
- N_- podlega rozkładowi Bernuliego o parametrach $p = 1/2$ i n-liczebność próby
- Z tablic rozkładu dwumianowego o $p = 1/2$ znajdź wartość k_- dla której

$$P(N_- \leqslant k_-) = \sum_{i=0}^{k_-} \left(\begin{array}{c} n \\ i \end{array} \right) \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{n-i} = \sum_{i=0}^{k_-} \left(\begin{array}{c} n \\ i \end{array} \right) \left(\frac{1}{2}\right)^n = (1-\alpha)/2$$

  lub najbliżej $(1-\alpha)/2$

- Niech $k_+ = n + 1 - k_-$
- Przedział ufności jest taki, ze dolną granicą jest element o indeksie k_- a górną o indeksie k_+ szeregu utworzonego w pierwszym punkcie.

$$X_{(k_-)} \leqslant M \leqslant X_{(k_+)}$$

Przedział ufności dla mediany

Przykład 20.1. W dużej firmie uzyskano następujące wyniki w rezultacie losowego sprawdzenia wieku 20 pracowników: 24, 31, 28, 43, 28, 56, 48, 39, 52, 32, 38, 49, 51, 49, 62, 33, 41, 58, 63, 56. Znajdź na poziomie ufności 95% przedział ufności dla mediany wieku pracowników tej firmy.

- Porządkujemy rosnąco listę wieku 20 losowo wybranych pracowników:

  24 28 28 31 32 33 38 39 41 43 48 49 49 51 52 56 56 58 62 63

- Z tablic rozkładu dwumianowego dla $n = 20$ i $p = 1/2$ odczytujemy

$$P(N_- \leqslant 5) = 0.0207$$

  A więc dla $k_- = 5$ osiągamy wartość prawdopodobieństwa najbliższą $(1-\alpha)/2 = 0.025$

- Jako dolną granicę przedziału ufności należy wybrać piąty element w uporządkowanej rosnąco liście wieku pracowników, $X_{(5)} = 32$

- Górną granicę przedziału ufności wyznaczamy z warunku $X_{(n+1-k_-)} = X_{(16)} = 56$
- Przedział ufności na poziomie ufności 95% dla mediany wieku pracowników to:

$$32 < M < 56$$

Test hipotezy dla mediany.

- Rozważmy na poziomie istotności α test hipotezy dotyczącej mediany M. Niech $H_0 : M = M_0$ przy hipotezie alternatywnej $H_1 : M > M_0$.
- Zakładamy, że próba losowa, $x_1, x_2, ..., x_n$, pochodzi z populacji o rozkładzie ciągłym, tak aby $P(X \leqslant M) = 1/2$
- Jako statystykę testową wybieramy, N^+, oznaczającą liczbę elementów próby losowej większych od M_0.
- Hipotezę H_0 odrzucamy na rzecz hipotezy H_1, jeśli wartość n^+ statystyki testowej N^+ spełnia relację $n^+ \geqslant k$, gdzie k wyznaczamy z warunku

$$P(N_+ \geqslant k) = \sum_{i=k}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n = \alpha \quad \text{lub najbliżej } \alpha$$

- Zbiory krytyczne dla innych hipotez alternatywnych mają odpowiednio postać:

  Dla $H_1 : M < M_0$:

$$P(N_+ \leqslant k) = \sum_{i=0}^{k} \binom{n}{i} \left(\frac{1}{2}\right)^n = \alpha \quad \text{lub najbliżej } \alpha$$

  Dla $H_1 : M \neq M_0$:

$$P(N_+ \geqslant k_+) = \sum_{i=k_+}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n = \alpha/2 \quad \text{lub najbliżej } \alpha/2$$

$$P(N_+ \leqslant k_-) = \sum_{i=0}^{k_-} \binom{n}{i} \left(\frac{1}{2}\right)^n = \alpha/2 \quad \text{lub najbliżej } \alpha/2$$

Test znaków dla mediany.

- Test znaków: na przykładzie $H_0 : M = M_0$ oraz $H_1 : M \neq M_0$
- Zastępujemy wszystkie wartości w próbie losowej większe od M_0 przez znak +, a mniejsze od M_0 przez znak -
- Jeśli w próbie któraś z wartości jest równa M_0, to ją usuwamy, zmniejszając jednocześnie liczebność próby n
- Niech n^+ oznacza liczbę znaków +. Obliczamy prawdopodobieństwo z rozkładu dwumianowego o parametrach n oraz $p = 1/2$

$$\gamma_+ = P(N^+ \geqslant n^+) \quad \gamma_- = P(N^+ \leqslant n^+)$$

  jeśli γ_+ lub γ_- są mniejsze od $\alpha/2$ to odrzucamy hipotezę H_0 na rzecz H_1.

- Zbiory krytyczne dla innych hipotez alternatywnych mają, odpowiednio postać:

  Dla $H_1 : M > M_0$:
  $$\gamma = P(N^+ \geqslant n^+)$$
  Dla $H_1 : M < M_0$:
  $$\gamma = P(N^+ \leqslant n^+)$$
  jeśli γ jest mniejsza od α to odrzucamy hipotezę H_0 na rzecz H_1.

Test znaków dla mediany.

Przykład 20.2. Dla następującego zbioru danych eksperymentalnych:
1.51, 1.35, 1.69, 1.48, 1.29, 1.27, 1.54, 1.39, 1.45
przeprowadź na poziomie istotności $\alpha = 0.05$ test hipotezy $H_0 : M = 1.4$ przy hipotezie alternatywnej $H_1 : M > 1.4$.

- Zastępujemy wartości w próbie losowej większe od 1.4 przez znak plus, a mniejsze przez znak minus:

  + - + + - - + - +

- Dla $n^+ = 5, n = 9$ oraz $p = 1/2$ obliczamy $\gamma = P(N^+ \geqslant 5) = 0.5$
- Ponieważ $\gamma > 0.05$, więc nie ma powodu do odrzucenia hipotezy zerowej, tzn. mediana nie jest większa od 1.4.

Test znaków dla mediany: duża próba

- Jeśli liczebność próby losowej jest duża, $n > 10$, można przybliżyć rozkład dwumianowy rozkładem normalnym.
- Statystyka testowa, N^+, przy założeniu prawdziwości hipotezy zerowej H_0 ma więc rozkład normalny o wartości oczekiwanej $\mu = np = n/2$ oraz wariancji $\sigma^2 = np(1-p) = n/4$.
- Zmienna losowa
  $$z = \frac{N^+ - n/2}{\sqrt{n/4}} = \frac{2N^+ - n}{\sqrt{n}}$$
  ma więc standardowy rozkład normalny.
- W zależności od hipotezy alternatywnej, zbiory krytyczne przyjmują postać:
  $$H_1 : M > M_0, \quad C = (z(\alpha), +\infty)$$
  $$H_1 : M < M_0, \quad C = (-\infty, -z(\alpha))$$
  $$H_1 : M \neq M_0, \quad C = (-\infty, -z(\alpha)) \cup (z(\alpha), +\infty)$$

Test znaków dla mediany: duża próba

Przykład 20.3. 15 pracownikom zlecono przycięcie krzewów winogron. Efektywność wykonanej przez nich pracy mierzona liczbą osobo-godzin/akr wynosiła:
5.2, 5.0, 4.8, 3.9, 6.1, 4.2, 4.4, 5.5, 5.8, 4.5, 4.2, 5.3, 4.9, 4.7, 4.9.
Przetestuj na poziomie istotności $\alpha = 0.05$ hipotezę, że mediana czasu potrzebnego na przycięcia 1 akra winogron wynosi 4.5 h, wobec hipotezy alternatywnej, że ta mediana jest większa.

- $H_0 : M = 4.5, H_1 : M > 4.5, \quad \alpha = 0.05$
- Zapisujemy próbę losową w postaci:

  + + + - + - - + + - + + ++
- Statystyka testowa:
$$z = \frac{2N^+ - n}{\sqrt{n}} = 1.6 \notin C = (z(\alpha), \infty) = (1.645, \infty)$$
  Ponieważ $z \notin C$, więc na poziomie istotności $\alpha = 0.05$ nie ma powodu do odrzucenia hipotezy H_0 na rzecz hipotezy H_1.

Test Wilcoxona

- Test Wilcoxona służy do testowania hipotezy dotyczącej mediany i jest ulepszeniem testu znaków.
- W teście Wilcoxona zastępujemy poszczególne elementy próbki losowej przez rangi uporządkowane zgodnie z wartościami $z_i = |x_i - M_0|$.
- Najmniejszej wartości z_i przypisujemy rangę 1, itd. Jeśli dwa różne z_i mają tą samą wartość to obu przypisujemy rangi średnie.
- Zastępujemy wszystkie z_i przez ich rangi, a wartościom w próbie losowej większym od M_0 przypisujemy znak +, a mniejszym od M_0 znak -
- Jeśli w próbie któraś z wartości jest równa M_0, to ją usuwamy, zmniejszając jednocześnie liczebność próby n.
- Niech W^+ będzie sumą rang elementów próbki losowej którym przypisano znak +, a W^- sumą rang elementów próbki, którym przypisano znak -.
- Test Wilcoxona może być stosowany tylko do rozkładów ciągłych i symetrycznych.
- W zależności od hipotezy alternatywnej, zbiory krytyczne przyjmują postać:

  $H_1 : M > M_0, \quad C = (c, \infty)$, gdzie $P(W^+ \geqslant c) = \alpha$

  $H_1 : M < M_0, \quad C = (-\infty, c)$, gdzie $P(W^+ \leqslant c) = \alpha$

  $H1 : M \neq M_0, \quad C = (-\infty, c_1) \cup (c_2, \infty)$, gdzie $P(W^+ \leqslant c_1) = P(W^+ \geqslant c_2) = \alpha/2$

Test Wilcoxona

Przykład 20.4. W eksperymencie uzyskano dane:
1.51, 1.35, 1.69, 1.48, 1.29, 1.27, 1.54, 1.39, 1.45.
Przetestuj na poziomie istotności $\alpha = 0.05$ hipotezę, że mediana wynosi $M = 1.4$ wobec hipotezy alternatywnej, że jest różna od tej wartości.

-
$$H0 : M = 1.4, H_1 : M \neq 1.4, \alpha = 0.05$$

Wartość	1.51	1.35	1.69	1.48	1.29	1.27	1.54	1.39	1.41
zi = |xi - 1.4|	0.11	0.05	0.29	0.08	0.11	0.13	0.14	0.01	0.01
Znak	+	-	+	+	-	-	+	-	+
Ranga	5.5	3	9	4	5.5	7	8	1.5	1.5

- $W^+ = 29, n = 9, C = (\infty, 6) \cup (38, \infty)$
- $W^+ \notin C$ - nie ma powodu do odrzucenia hipotezy H_0

Test Wilcoxona: duża próba

- Dla dużej próbki losowej (n > 20), rozkład W^+ przy założeniu prawdziwości hipotezy $H_0 : M = M_0$ ma rozkład normalny o wartości oczekiwanej i wariancji:

$$E[W^+] = \frac{1}{4n(n+1)} \qquad Var[W^+] = \frac{1}{24n(n+1)(2n+1)}$$

- Zmienna losowa

$$z = \frac{W^+ - E[W^+]}{\sqrt{Var[W^+]}}$$

  ma więc standardowy rozkład normalny.

- W zależności od hipotezy alternatywnej, zbiory krytyczne przyjmują postać:

$$H_1 : M > M_0, \quad C = (z(\alpha), +\infty)$$

$$H_1 : M < M_0, \quad C = (-\infty, -z(\alpha))$$

$$H_1 : M \neq M_0, \quad C = (-\infty, -z(\alpha)) \cup (z(\alpha), +\infty)$$

Testy nieparametryczne równości median dla dwóch niezależnych prób losowych

- Często zachodzi potrzeba dokonania testu równości median dla dwóch niezależnych prób z dwóch populacji
- Jedna próba jest próbą kontrolną a druga eksperymentalną np. grupa pacjentów której podano testowany lek lub grupa osób poddana nowej metodzie nauki, ...
- Chcemy dokonać testu hipotezy o braku wpływu czynnika obecnego w grupie eksperymentalnej na mediane rozkładu jakiejś pożądanej cechy wobec hipotezy alternatywnej mówiącej o wpływie tego czynnika na medianę.
- Niech m_1 i m_2 będą medianami dwóch populacji. Dysponujemy próbami losowymi o liczebności n_1 z populacji 1 i liczebności n_2 z populacji 2.

Test znaków - równości median dla dwóch niezależnych prób losowych

$H_0 : m_1 = m_2$ na poziomie istotności α

- Łączymy obie próby w jedną o liczebności $n_1 + n_2$
- Znajdujemy medianę z łacznej próby
- W sytuacji gdy mediana jest równa jednej z wartości odrzucamy ją (odpowiednio zmniejszając o jeden n_1 lub n_2. Niech $n_1 + n_2 = 2k$
- Statystyka testowa: N_1^+ - liczba obserwacji pochodzących z pierwszej próby większych od mediany połączonej próby (powinno ich być okolo $n_1/2$
- Rozkład statystyki N_1^+ jest rozkładem hipergeometrycznym

$$P(N_1^+ = n_1^+) = \frac{\begin{pmatrix} n_1 \\ n_1^+ \end{pmatrix} \begin{pmatrix} n_2 \\ k - n_1^+ \end{pmatrix}}{\begin{pmatrix} n_1 + n_2 \\ k \end{pmatrix}}$$

Zbiorami krytycznymi są:

- dla $H_1 : m_1 \neq m_2$ $C = (0, c_1) \cup (c_2, k)$ $\quad P(N_1^+ \leqslant c_1) = P(N_1^+ \geqslant c_2) = \alpha/2$
- $H_1 : m_1 > m_2$ $C = (c, k)$ $\quad P(N_1^+ \geqslant c) = \alpha$
- $H_1 : m_1 < m_2$ $C = (0, c)$ $\quad P(N_1^+ \leqslant c) = \alpha$

Test znaków - równości median dla dwóch niezależnych prób losowych

$H_0 : m_1 = m_2$ na poziomie istotności α

- Dla dużych próbek ($n_1 > 5,\ \ n_2 > 5$) rozkład hipergeometryczny przybliżamy rozładem normalnym.
- $$E[N_1^+] = \frac{N^+ n_1}{n} \quad Var[N_1^+] = \frac{N^+ N^- n_1 n_2}{n^2(n-1)}$$
  gdzie $N^+ = N_1^+ + N_2^+, \quad N^- = N_1^- + N_2^-, \quad n = n_1 + n_2$
- $$Z = \frac{N_1^+ - E[N_1^+]}{\sqrt{Var[N_1^+]}}$$

Zbiorami krytycznymi są:

- dla $H_1 : m_1 \neq m_2$ $C = (-\infty, -z(\alpha/2)) \cup (z(\alpha/2), \infty)$
- $H_1 : m_1 > m_2$ $C = (z(\alpha), \infty)$
- $H_1 : m_1 < m_2$ $C = (-\infty, -z(\alpha))$

Test znaków - równości median dla dwóch niezależnych prób losowych

Przykład 20.5. Dane na temat zużycia opon (w tys. km) dwóch producentów I i II:
I : 34, 32, 37, 35, 42, 43, 47, 58, 59, 62, 69, 71, 78, 84
II : 39, 48, 54, 65, 70, 76, 87, 90, 111, 118, 126, 127
Przetestuj na poziomie istotności $\alpha = 0.05$ hipotezę, że mediana zużycia opon od producenta II jest większa niż opon od producenta I.

- $H_0 : m_1 = m_2, H_1 : m_1 < m_2, \alpha = 0.05$
- Ponieważ $n_1 = 14 > 5$ oraz $n_2 = 12 > 5$, więc stosujemy przybliżenie normalne.
- mediana z połączonych próbek : 63.5
- Znajdujemy: $N_1^- = 10, N_1^+ = 4, N_2^- = 3, N_2^+ = 9, \rightarrow N^- = 13, N^+ = 13$
- Wyznaczamy: $E[N_1^+] = 7, Var[N_1^+] = 1.68, z = -0.023, C = (= \infty, -1.645)$
- Ponieważ $z \notin C$, więc nie ma powodu do odrzucenia hipotezy o równości median w obu próbach.

Test Wilcoxona - równości median dla dwóch niezależnych prób losowych

- Załóżmy, że mamy n_1 elementową próbkę z populacji I oraz n_2 elementową próbkę z populcji II.
- Testujemy hipotezę o równości median ($H_0 : m_1 = m_2$).
- Łączymy próbki w jedną o liczebności $n_1 + n_2$ i porządkujemy rosnąco, ale pamiętając, który element pochodzi z której próbki.
- Sumujemy rangi elementów z próbki II - niech ta suma będzie równa R.
- Obliczamy wartość statystyki testowej Wilcoxona

$$W = R - \frac{1}{2}n_2(n_2 + 1)$$

- Zbiorami krytycznymi są:
  - dla $H_1 : m_1 \neq m_2$ $C = (0, c_1) \cup (c_2, n_1 n_2)$ $P(W \leqslant c_1) = P(W \geqslant c_2) = \alpha/2$
  - $H_1 : m_1 > m_2$ $C = (c, \infty)$ $P(W \geqslant c) = \alpha$
  - $H_1 : m_1 < m_2$ $C = (0, c)$ $P(W \leqslant c) = \alpha$

Test Wilcoxona - równości median dla dwóch niezależnych prób losowych

Przykład 20.6. Ceny opon (w USD) podobnego typu dwóch producentów wynoszą odpowiednio:
I : 85, 99, 100, 110, 105, 87
II : 67, 69, 70, 93, 105, 90, 110, 115
Wykonaj test sumy rang Wilcoxona z $\alpha = 0.05$ do zweryfikowania hipotezy o równości median, przy hipotezie alternatywnej, że mediany są różne.

- $H_0 : m_1 = m_2, H_1 : m1 \neq m2, \alpha = 0.05, n_1 = 6, n_2 = 8$

Wartość	67	69	70	85	87	90	93	99	100	105	105	110	110	115
Populacja	II	II	II	I	I	II	II	I	I	I	II	I	II	II
Ranga	1	2	3	4	5	6	7	8	9	10.5	10.5	12.5	12.5	14

- $R = 56 \to W = 20 \to W \notin C = (0, 9) \cup (38, 48)$