

Biostatystyka

Zespół biostatystyków ZPZŚiA

WNZ

8 marca 2019

Biostatystyka Dziedzina naukowa, w której stosuje się metody statystyczne, aby odpowiedzieć na pytania dotyczące biologii człowieka i medycyny

- Biostatystyka** Dziedzina naukowa, w której stosuje się metody statystyczne, aby odpowiedzieć na pytania dotyczące biologii człowieka i medycyny
- Statystyka** Nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych; posługuje się rachunkiem prawdopodobieństwa.

Biostatystyka Dziedzina naukowa, w której stosuje się metody statystyczne, aby odpowiedzieć na pytania dotyczące biologii człowieka i medycyny

Statystyka Nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych; posługuje się rachunkiem prawdopodobieństwa.

Pojęcia podstawowe

Biostatystyka Dziedzina naukowa, w której stosuje się metody statystyczne, aby odpowiedzieć na pytania dotyczące biologii człowieka i medycyny

Statystyka Nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych; posługuje się rachunkiem prawdopodobieństwa.

Pojęcia podstawowe

Populacja Zbiór obiektów z wyróżnioną cechą (cechami). Obiektami mogą być przedmioty lub wartości cechy

Biostatystyka Dziedzina naukowa, w której stosuje się metody statystyczne, aby odpowiedzieć na pytania dotyczące biologii człowieka i medycyny

Statystyka Nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych; posługuje się rachunkiem prawdopodobieństwa.

Pojęcia podstawowe

Populacja Zbiór obiektów z wyróżnioną cechą (cechami). Obiektami mogą być przedmioty lub wartości cechy

Próba Wybrana część populacji podlegająca badaniu. Próba powinna stanowić reprezentację populacji w tym sensie, że częstości występowania w próbie każdej z badanych cech nie powinny się znacznie różnić od częstości występowania tych cech w populacji

Statystyka Nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych; posługuje się rachunkiem prawdopodobieństwa.

Populacja Zbiór obiektów z wyróżnioną cechą (cechami).Obiektami mogą być przedmioty lub wartości cechy

Cecha losowa Wielkość losowa charakteryzująca obiekty danej populacji

Rodzaje cech

niemierzalna – zwana też jakościową – przyjmuje wartości nie będące liczbami
(np. *kolor, płeć, smak*)

Rodzaje cech

- niemierzalna** – zwana też jakościową – przyjmuje wartości nie będące liczbami (np. *kolor, płeć, smak*)
- mierzalna** – zwana też ilościową – przyjmuje pewne wartości liczbowe (np. *długość, wytrzymałość, ciężar*)

Rodzaje cech

niemierzalna – zwana też jakościową – przyjmuje wartości nie będące liczbami (np. *kolor, płeć, smak*)

mierzalna – zwana też ilościową – przyjmuje pewne wartości liczbowe (np. *długość, wytrzymałość, ciężar*)

Rodzaje cech mierzalnych

skokowa – zwana też dyskretną – nie przyjmuje wartości pośrednich (np. *ilość bakterii, ilość pracowników, ilość pasażerów,*).

Rodzaje cech

niemierzalna – zwana też jakościową – przyjmuje wartości nie będące liczbami (np. *kolor, płeć, smak*)

mierzalna – zwana też ilościową – przyjmuje pewne wartości liczbowe (np. *długość, wytrzymałość, ciężar*)

Rodzaje cech mierzalnych

skokowa – zwana też dyskretną – nie przyjmuje wartości pośrednich (np. *ilość bakterii, ilość pracowników, ilość pasażerów,*).

ciągła – przyjmuje wartości z pewnego przedziału liczbowego (np. *wzrost, waga, ciśnienie, czas obsługi*)

Przykłady parametrów charakteryzujących populację

Mediana badanej cechy to wartość, która dzieli populację na dwie części. Połowa obiektów w populacji ma cechę o wartości poniżej mediany, a połowa powyżej.

Przykłady parametrów charakteryzujących populację

Mediana badanej cechy to wartość, która dzieli populację na dwie części. Połowa obiektów w populacji ma cechę o wartości poniżej mediany, a połowa powyżej.

Kwartył dolny badanej cechy to wartość, która dzieli populację w stosunku 1:3. Jedna czwarta obiektów w populacji ma cechę o wartości poniżej kwartyła dolnego, a pozostałe trzy czwarte powyżej.

Przykłady parametrów charakteryzujących populację

Mediana badanej cechy to wartość, która dzieli populację na dwie części. Połowa obiektów w populacji ma cechę o wartości poniżej mediany, a połowa powyżej.

Kwartył dolny badanej cechy to wartość, która dzieli populację w stosunku 1:3. Jedna czwarta obiektów w populacji ma cechę o wartości poniżej kwartyła dolnego, a pozostałe trzy czwarte powyżej.

Kwartył górny badanej cechy to wartość, która dzieli populację w stosunku 3:1. Trzy czwarte obiektów w populacji ma cechę o wartości poniżej kwartyła dolnego, a pozostała jedna trzecia powyżej.

Przykłady parametrów charakteryzujących populację

Mediana badanej cechy to wartość, która dzieli populację na dwie części. Połowa obiektów w populacji ma cechę o wartości poniżej mediany, a połowa powyżej.

Kwartył dolny badanej cechy to wartość, która dzieli populację w stosunku 1:3. Jedna czwarta obiektów w populacji ma cechę o wartości poniżej kwartyła dolnego, a pozostałe trzy czwarte powyżej.

Kwartył górny badanej cechy to wartość, która dzieli populację w stosunku 3:1. Trzy czwarte obiektów w populacji ma cechę o wartości poniżej kwartyła dolnego, a pozostała jedna trzecia powyżej.

Średnia ...

Wariancja ...

Mierniki położenia i rozproszenia próby; przykłady

Niech X oznacza cechę losową. Niech wartości $x_1, x_2, \dots, x_n$ oznaczają n realizacji tej cechy. Przez $x_{1:n}, x_{2:n}, \dots, x_{n:n}$ będą oznaczane realizacje tej cechy w kolejności od najmniejszej do największej.

Mierniki położenia	Oznaczenia	Wzór
średnia	$\bar{x}$	$\frac{1}{n} \sum_{i=1}^n x_i$
mediana	Me	$x_{(n+1)/2:n}$ (gdy n nieparzyste) $(x_{n/2:n} + x_{n/2+1:n})/2$ (gdy n parzyste)
dolny kwartyl	Q_1	$x_{[n/4]:n}$
górny kwartyl	Q_3	$x_{[3n/4]:n}$
dominanta (moda)	D	najczęściej występująca wartość
minimum	Min	$x_{1:n}$
maksimum	Max	$x_{n:n}$

$[a]$ – część całkowita liczby a

Mierniki położenia i rozproszenia próby; przykłady

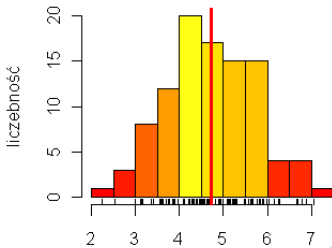
Niech X oznacza cechę losową. Niech wartości $x_1, x_2, \dots, x_n$ oznaczają n realizacji tej cechy. Przez $x_{1:n}, x_{2:n}, \dots, x_{n:n}$ będą oznaczane realizacje tej cechy w kolejności od najmniejszej do największej.

Mierniki rozproszenia	Oznaczenia	Wzór
rozstęp	R	$Max - Min$
wariancja	S^2	$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
odchylenie standardowe	S	$\sqrt{S^2}$
odchylenie przeciętne	d	$\frac{1}{n} \sum_{i=1}^n x_i - \bar{x} $
odchylenie ćwiartkowe	Q	$\frac{Q_3 - Q_1}{2}$
współczynnik zmienności	V	$\frac{S}{\bar{x}} 100\%$

Wnioskowanie statystyczne
polega na wnioskowaniu o
populacji na podstawie próby

Przykład

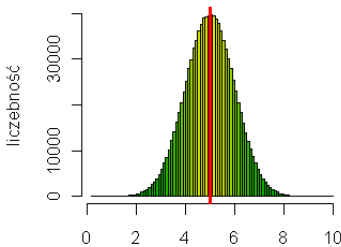
Badamy rozkład wartości cechy w populacji na podstawie próby 100 elementowej. Średnia z próby $\bar{x} = 4.717641$. W tym przykładzie wiadomo, że średnia populacyjna $\mu = 5$



Wnioskowanie statystyczne
polega na wnioskowaniu o
populacji na podstawie próby

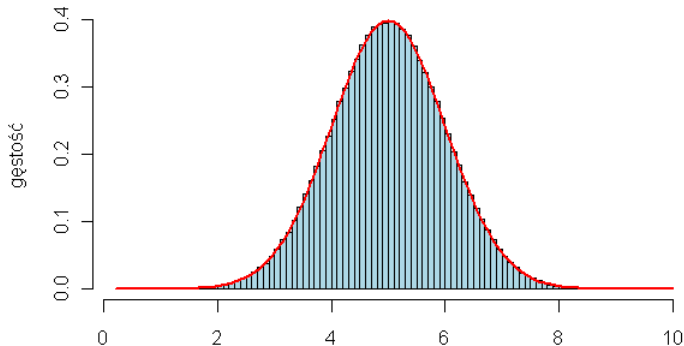
Przykład

Badamy rozkład wartości cechy w populacji na podstawie próby 1000000 elementowej. Średnia z próby $\bar{x} = 5.000585$. W tym przykładzie wiadomo, że średnia populacyjna $\mu = 5$

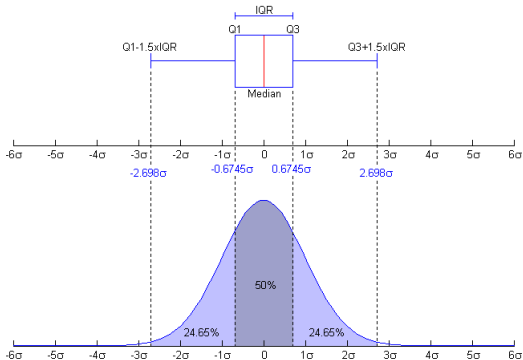


Rozkład normalny

Ten sam histogram po unormowaniu. Pole wszystkich słupków wynosi jeden.
Pole pod **krzywą** jest też równe jeden. Jest to krzywa Gaussa

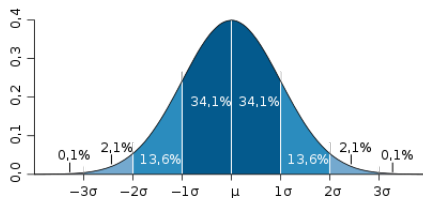


Rozkład normalny



$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} (x/\sigma)^2 \right\}$$

Rozkład normalny



$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right\}, \quad X \sim N(\mu, \sigma^2)$$

Prawo trzech sigm

$$P\{|X - \mu| < \sigma\} = 0.68268 \approx 0.68 \quad (1)$$

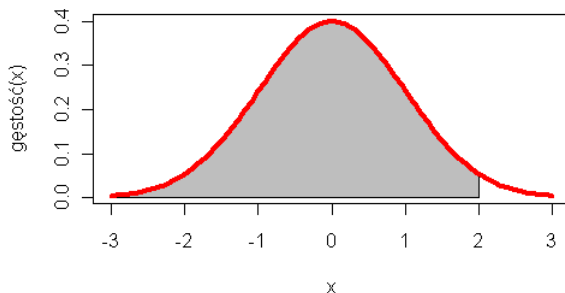
$$P\{|X - \mu| < 2\sigma\} = 0.95550 \approx 0.96 \quad (2)$$

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

Dystrybuanta

$$\forall t \in R \quad F(t) = P(X \leq t)$$

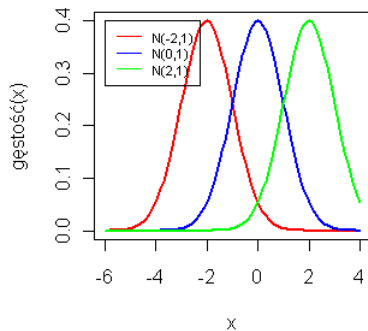
Pole=F(2), gdzie F-dystrybuanta



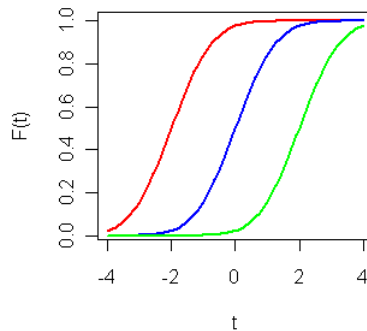
Rozkład normalny

Rozkład a parametr położenia

Gęstości



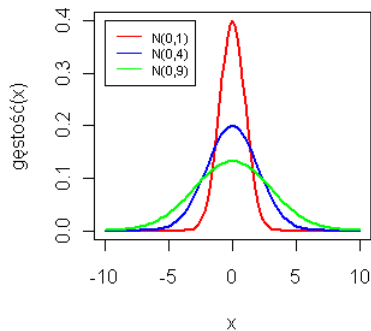
Dystrybuanty



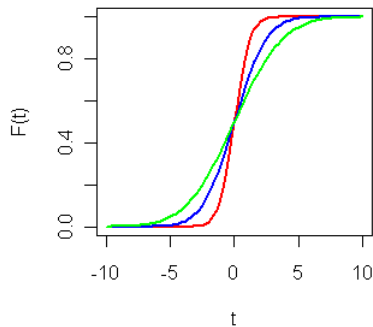
Rozkład normalny

Rozkład a parametr skali

Gęstości



Dystrybuanty



Rozkład dwupunktowy i dwumianowy

Zmienna losowa X ma **rozkład dwupunktowy** ($X \sim D(p)$), jeżeli z dodatnimi prawdopodobieństwami przyjmuje jedynie dwie wartości x_1 i x_2 :

$$P(X = x_2) = p, \quad P(X = x_1) = 1 - p, \quad 0 < p < 1.$$

Zmienna losowa X ma **rozkład dwumianowy** ($X \sim B(n, p)$), jeżeli

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad 0 < p < 1, \quad k = 0, 1, \dots, n.$$

Schemat Bernoulliego. Wykonujemy dwuwynikowe doświadczenie. Wyniki nazywane są umownie *sukces* oraz *porażka*. Prawdopodobieństwo sukcesu wynosi p (porażki $1 - p$). Doświadczenie wykonujemy w sposób niezależny n krotnie. Niech zmienną losową X będzie ilość sukcesów. Zmienna losowa X ma rozkład $B(n, p)$.

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.
- ▶ Podać znane nazwy rozkładów cech i jakiego typu są to cechy.

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.
- ▶ Podać znane nazwy rozkładów cech i jakiego typu są to cechy.
- ▶ Podać dwa przykłady cech o rozkładzie dwumianowym.

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.
- ▶ Podać znane nazwy rozkładów cech i jakiego typu są to cechy.
- ▶ Podać dwa przykłady cech o rozkładzie dwumianowym.
- ▶ Podać dwa przykłady cech o rozkładzie normalnym.

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.
- ▶ Podać znane nazwy rozkładów cech i jakiego typu są to cechy.
- ▶ Podać dwa przykłady cech o rozkładzie dwumianowym.
- ▶ Podać dwa przykłady cech o rozkładzie normalnym.
- ▶ Zmienna losowa ma rozkład $N(20, 4)$. Ile wynosi $P\{X \in (16, 24)\}$?

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.
- ▶ Podać znane nazwy rozkładów cech i jakiego typu są to cechy.
- ▶ Podać dwa przykłady cech o rozkładzie dwumianowym.
- ▶ Podać dwa przykłady cech o rozkładzie normalnym.
- ▶ Zmienna losowa ma rozkład $N(20, 4)$. Ile wynosi $P\{X \in (16, 24)\}$?
- ▶ Omówić pojęcie populacji w badaniach statystycznych.

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.
- ▶ Podać znane nazwy rozkładów cech i jakiego typu są to cechy.
- ▶ Podać dwa przykłady cech o rozkładzie dwumianowym.
- ▶ Podać dwa przykłady cech o rozkładzie normalnym.
- ▶ Zmienna losowa ma rozkład $N(20, 4)$. Ile wynosi $P\{X \in (16, 24)\}$?
- ▶ Omówić pojęcie populacji w badaniach statystycznych.
- ▶ Co to jest próba reprezentatywna?

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.
- ▶ Podać znane nazwy rozkładów cech i jakiego typu są to cechy.
- ▶ Podać dwa przykłady cech o rozkładzie dwumianowym.
- ▶ Podać dwa przykłady cech o rozkładzie normalnym.
- ▶ Zmienna losowa ma rozkład $N(20, 4)$. Ile wynosi $P\{X \in (16, 24)\}$?
- ▶ Omówić pojęcie populacji w badaniach statystycznych.
- ▶ Co to jest próba reprezentatywna?
- ▶ Jakie są zasady pobierania prób reprezentatywnych?

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.
- ▶ Podać znane nazwy rozkładów cech i jakiego typu są to cechy.
- ▶ Podać dwa przykłady cech o rozkładzie dwumianowym.
- ▶ Podać dwa przykłady cech o rozkładzie normalnym.
- ▶ Zmienna losowa ma rozkład $N(20, 4)$. Ile wynosi $P\{X \in (16, 24)\}$?
- ▶ Omówić pojęcie populacji w badaniach statystycznych.
- ▶ Co to jest próba reprezentatywna?
- ▶ Jakie są zasady pobierania prób reprezentatywnych?
- ▶ Co to jest wnioskowanie statystyczne?

Pytania

- ▶ Jaka jest różnica między cechą skokową i ciągłą — podać przykłady każdej z nich.
- ▶ Wymienić typy cech i podać po jednym przykładzie.
- ▶ Podać znane nazwy rozkładów cech i jakiego typu są to cechy.
- ▶ Podać dwa przykłady cech o rozkładzie dwumianowym.
- ▶ Podać dwa przykłady cech o rozkładzie normalnym.
- ▶ Zmienna losowa ma rozkład $N(20, 4)$. Ile wynosi $P\{X \in (16, 24)\}$?
- ▶ Omówić pojęcie populacji w badaniach statystycznych.
- ▶ Co to jest próba reprezentatywna?
- ▶ Jakie są zasady pobierania prób reprezentatywnych?
- ▶ Co to jest wnioskowanie statystyczne?
- ▶ Populacja i próba: wymienić przynajmniej dwie zasadnicze różnice.

Estymacja punktowa

Niech $X_1, X_2, \dots, X_n$ oznacza próbę z populacji oraz θ parametr charakteryzujący tę populację. Na podstawie próby chcemy oszacować (przybliżyć) wartość parametru θ .

Estymator punktowy jest funkcją próby. Przybliża wartość parametru θ :

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$$

Estymacja punktowa parametrów cechy $X \sim N(\mu, \sigma^2)$

Estymator średniej μ — średnia arytmetyczna

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + \cdots + X_n}{n}$$

Estymator wariancji σ^2 — wariancja próbkowa

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Estymator odchylenia standardowego σ

$$S = \sqrt{S^2}$$

Estymacja punktowa parametru p cechy $X \sim D(p)$

Niech n oznacza liczbę obiektów wylosowanych z populacji, wśród których znalazło się k obiektów, które posiadają wyróżnioną właściwość. Przyjmując, że p oznacza prawdopodobieństwo wylosowania z populacji obiektu o wyróżnionej właściwości mamy:

$$\hat{p} = \frac{k}{n}$$

Uwaga. Przyjmując dla $i = 1, 2, \dots, n$, że $P(X_i = 1) = p = 1 - P(X_i = 0)$, mamy $\hat{p} = \bar{X}$.

Estymacja przedziałowa

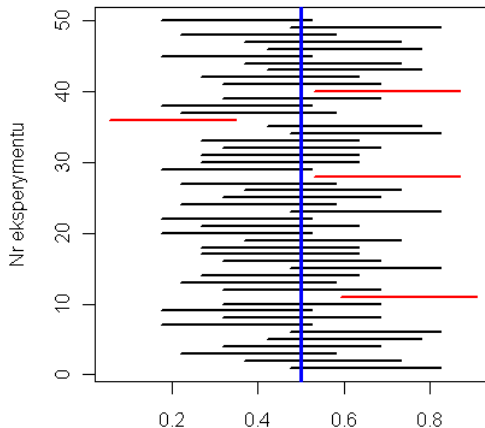
Przedział ufności (estymator przedziałowy) jest przedziałem o końcach zależnych od próby, który z pewnym z góry zadany prawdopodobieństwem $1 - \alpha$ pokrywa nieznaną wartość parametru θ :

$$P\{\theta \in (\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n))\} = 1 - \alpha.$$

Poziom ufności jest to ustalone prawdopodobieństwo $1 - \alpha$.

Definicje

Ilustracja estymacji przedziałowej parametru $\theta = 0.5$ (oznaczonego pionową niebieską linią) na poziomie ufności $1 - \alpha = 0.9$



Jedna populacja

Populacja z wyróżnioną cechą X

Przedział ufności dla średniej μ w rozkładzie normalnym $N(\mu, \sigma^2)$

Wariancja σ^2 jest nieznana

Poziom ufności: $1 - \alpha$

$$\left(\bar{X} - t(1 - \alpha/2; n - 1) \frac{S}{\sqrt{n}}, \bar{X} + t(1 - \alpha/2; n - 1) \frac{S}{\sqrt{n}} \right)$$

$t(\gamma; \nu)$ jest stabilizowanym kwantylem rzędu γ rozkładu t (t -Studenta) z ν stopniami swobody.

ν	Kwantyl rozkładu t -Studenta			
	γ			
	0.9500	0.9750	0.9875	0.9950
8	1.8595	2.3060	2.7515	3.3554
9	1.8331	2.2622	2.6850	3.2498
10	1.8125	2.2281	2.6338	3.1690

Przykład

Na podstawie próby 1.1, 1.2, 0.8, 0.9, 1.2, 1.3, 1.0, 0.7, 0.8, 1.0 oszacować wartość średnią μ rozkładu obserwowanej cechy $X \sim N(\mu, \sigma^2)$, na poziomie ufności $1 - \alpha = 0.95$.

$$\bar{x} = \frac{1.1 + 1.2 + \cdots + 1.0}{10} = 1.0$$

$$\sum (x_i - \bar{x})^2 = (1.1 - 1.0)^2 + \cdots + (1.0 - 1.0)^2 = 0.36$$

$$s^2 = \frac{0.36}{10 - 1} = 0.04, \quad s = \sqrt{s^2} = 0.2$$

$$t(0.975; 9) = 2.2622$$

$$t(0.975; 9) \frac{s}{\sqrt{n}} = 2.2622 \frac{0.2}{\sqrt{10}} = 0.14$$

$$(1 - 0.14, 1 + 0.14) = (0.86, 1.14)$$

Wniosek. Średnia wartość cechy jest jakąś liczbą z przedziału $(0.86, 1.14)$.
Zaufanie do tego wniosku wynosi 95%.

Przedział ufności dla wariancji w rozkładzie normalnym

Średnia μ jest nieznaną

Poziom ufności: $1 - \alpha$

$$\left(\frac{\sum_i (X_i - \bar{X})^2}{\chi^2(\frac{\alpha}{2}; n - 1)}, \frac{\sum_i (X_i - \bar{X})^2}{\chi^2(1 - \frac{\alpha}{2}; n - 1)} \right)$$

Jedna populacja

$\chi^2(\alpha; \nu)$ jest tablicowaną wartością krytyczną rozkładu chi-kwadrat z ν stopniami swobody.

ν	Wartości krytyczne $\chi^2(\alpha; r)$			
	α			
	0.975	0.950	0.050	0.025
8	2.1797	2.7326	15.5073	17.5345
9	2.7004	3.3251	16.9190	19.0228
10	3.2470	3.9403	18.3070	20.4832

Przykład

Na podstawie próby 1.1, 1.2, 0.8, 0.9, 1.2, 1.3, 1.0, 0.7, 0.8, 1.0 oszacować zróżnicowanie rozkładu obserwowanej cechy.

$$\bar{x} = \frac{1.1 + 1.2 + \cdots + 1.0}{10} = 1.0$$

$$\sum_i (x_i - \bar{x})^2 = (1.1 - 1.0)^2 + \cdots + (1.0 - 1.0)^2 = 0.36$$

$$s^2 = \frac{0.36}{10 - 1} = 0.04, \quad s = \sqrt{s^2} = 0.2$$

Poziom ufności $1 - \alpha = 0.95$, czyli $\alpha = 0.05$.

$$\chi^2\left(\frac{\alpha}{2}; n - 1\right) = \chi^2(0.025; 9) = 19.0228$$

$$\chi^2\left(1 - \frac{\alpha}{2}; n - 1\right) = \chi^2(0.975; 9) = 2.7004$$

$$\left(\frac{0.36}{19.0228}, \frac{0.36}{2.7004} \right) = (0.019, 0.133)$$

Wniosek. Wariancja cechy jest liczbą z przedziału $(0.019, 0.133)$. Zaufanie do tego wniosku wynosi 95%.

Estymacja prawdopodobieństwa sukcesu

Przedział przybliżony

$$\left(\hat{p} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

	Kwantyle u_{α} rozkładu normalnego $N(0, 1)$				
α	0.002	0.003	0.004	0.005	0.006
0.96	1.7744	1.7866	1.7991	1.8119	1.8250
0.97	1.9110	1.9268	1.9431	1.9600	1.9774
0.98	2.0969	2.1201	2.1444	2.1701	2.1973
0.99	2.4089	2.4573	2.5121	2.5758	2.6521

Na przykład $u_{0.975} = 1.96$

Dwie populacje

Populacja 1, cecha X_1

Populacja 2, cecha X_2

Oznaczenia

Próby: $X_{11}, \dots, X_{1n_1}; X_{21}, \dots, X_{2n_2}$

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad s_i^2 = \frac{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{n_i - 1}$$

$$s_e^2 = \frac{\sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2}{n_1 + n_2 - 2}, \quad s_r^2 = s_e^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Ocena różnicy między średnimi $\mu_1 - \mu_2$

Ocena punktowa: $\bar{X}_1 - \bar{X}_2$

Założenia:

1. $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$
2. X_1, X_2 są niezależne
3. $\sigma_1^2 = \sigma_2^2$

Przedział ufności (poziom ufności $1 - \alpha$)

$$(\bar{X}_1 - \bar{X}_2 - t(1 - \alpha/2; n_1 + n_2 - 2)s_r, \bar{X}_1 - \bar{X}_2 + t(1 - \alpha/2; n_1 + n_2 - 2)s_r)$$

Dwie populacje

Przykład. Z dwóch populacji pobrano próby: 60, 62, 65, 63, 60 oraz 58, 53, 57, 56, 61. Ocenic różnicę średnich.

$$\bar{x}_1 = 62, \sum_{i=1}^5 (x_{1i} - \bar{x}_1)^2 = 18, \bar{x}_2 = 57, \sum_{i=1}^5 (x_{2i} - \bar{x}_2)^2 = 34$$

$$s_r^2 = \frac{18 + 34}{5 + 5 - 2} \left(\frac{1}{5} + \frac{1}{5} \right) = 2.6$$

$$t(0.975; 8) = 2.3060; t(0.975; 8)s_r = 3.72$$

$$(62 - 57 - 3.72, 62 - 57 + 3.72) = (1.28, 8.72)$$

Wniosek. Różnica średnich jest liczbą z przedziału (1.28, 8.72)

Ocena różnicy frakcji $p_1 - p_2$

Założenia: $X_1 \sim D(p_1)$, $X_2 \sim D(p_2)$

Cechy X_1, X_2 są niezależne

Próba 1: $X_{11}, X_{12}, \dots, X_{1n_1}$ ($X_{1i} = 0$ lub 1)

Próba 2: $X_{21}, X_{22}, \dots, X_{2n_2}$ ($X_{2i} = 0$ lub 1)

$$k_1 = \sum_{i=1}^{n_1} X_{1i}$$

$$k_2 = \sum_{i=1}^{n_2} X_{2i}$$

Dwie populacje

Ocena punktowa: $\hat{p}_1 - \hat{p}_2 = \frac{k_1}{n_1} - \frac{k_2}{n_2}$

Przybliżony przedział ufności (poziom ufności $1 - \alpha$)

$$\hat{p}_1 - \hat{p}_2 \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Iloraz frakcji: $\frac{p_1}{p_2}$ (ryzyko względne)

$$\ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right) \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1} + \frac{1-\hat{p}_2}{n_2\hat{p}_2}}$$

Przykład: Porównanie lekarstw ze względu na odsetek osób, które nie reagują na podany lek

p_1	p_2	$p_1 - p_2$	p_1/p_2
0.01	0.001	0.009	10
0.410	0.401	0.009	1.02

Dwie populacje

Rozkład prawdopodobieństwa oraz dane

	Y			Y	
X	p_{11}	p_{12}	X	n_{11}	n_{12}
	p_{21}	p_{22}		n_{21}	n_{22}

Iloraz szans $\theta = \frac{p_{11}/p_{12}}{p_{21}/p_{22}}$

Estymator ilorazu szans $\hat{\theta} = \frac{\hat{p}_{11}/\hat{p}_{12}}{\hat{p}_{21}/\hat{p}_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

Przedział ufności dla $\ln(\theta)$

$$\ln(\hat{\theta}) \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Pytania

- ▶ Co to jest estymator?

Pytania

- ▶ Co to jest estymator?
- ▶ Co to znaczy, że estymator jest precyzyjny?

Pytania

- ▶ Co to jest estymator?
- ▶ Co to znaczy, że estymator jest precyzyjny?
- ▶ Podać przynajmniej dwa różne oszacowania średniej wartości cechy?

Pytania

- ▶ Co to jest estymator?
- ▶ Co to znaczy, że estymator jest precyzyjny?
- ▶ Podać przynajmniej dwa różne oszacowania średniej wartości cechy?
- ▶ Co to jest przedział ufności?

Pytania

- ▶ Co to jest estymator?
- ▶ Co to znaczy, że estymator jest precyzyjny?
- ▶ Podać przynajmniej dwa różne oszacowania średniej wartości cechy?
- ▶ Co to jest przedział ufności?
- ▶ Co to jest poziom ufności?

Pytania

- ▶ Co to jest estymator?
- ▶ Co to znaczy, że estymator jest precyzyjny?
- ▶ Podać przynajmniej dwa różne oszacowania średniej wartości cechy?
- ▶ Co to jest przedział ufności?
- ▶ Co to jest poziom ufności?
- ▶ Jaka jest interpretacja poziomu ufności?

Pytania

- ▶ Co to jest estymator?
- ▶ Co to znaczy, że estymator jest precyzyjny?
- ▶ Podać przynajmniej dwa różne oszacowania średniej wartości cechy?
- ▶ Co to jest przedział ufności?
- ▶ Co to jest poziom ufności?
- ▶ Jaka jest interpretacja poziomu ufności?
- ▶ Od jakich czynników i jak zależy długość przedziału ufności? Czy prowadzący doświadczenie może mieć wpływ na długość przedziału ufności?

Pytania

- ▶ Co to jest estymator?
- ▶ Co to znaczy, że estymator jest precyzyjny?
- ▶ Podać przynajmniej dwa różne oszacowania średniej wartości cechy?
- ▶ Co to jest przedział ufności?
- ▶ Co to jest poziom ufności?
- ▶ Jaka jest interpretacja poziomu ufności?
- ▶ Od jakich czynników i jak zależy długość przedziału ufności? Czy prowadzący doświadczenie może mieć wpływ na długość przedziału ufności?
- ▶ Na podstawie badań uzyskano dla średniej następujący przedział ufności (2, 13). Czy można uznać, że średnia w populacji jest równa 7 i dlaczego?

Hipoteza statystyczna

Dowolne przypuszczenie dotyczące rozkładu prawdopodobieństwa cechy (oznaczenie H_0).

Hipoteza statystyczna

Dowolne przypuszczenie dotyczące rozkładu prawdopodobieństwa cechy (oznaczenie H_0).

Test hipotezy statystycznej

Postępowanie mające na celu odrzucenie lub nie odrzucenie hipotezy statystycznej.

Hipoteza statystyczna

Dowolne przypuszczenie dotyczące rozkładu prawdopodobieństwa cechy (oznaczenie H_0).

Test hipotezy statystycznej

Postępowanie mające na celu odrzucenie lub nie odrzucenie hipotezy statystycznej.

Statystyka testowa

Funkcja próby, na podstawie której wnioskuje się o odrzuceniu lub nie hipotezy statystycznej.

Błąd I rodzaju

Błąd wnioskowania polegający na odrzuceniu hipotezy, gdy w rzeczywistości jest ona prawdziwa.

Błąd II rodzaju

Błąd wnioskowania polegający na nieodrżuceniu hipotezy, gdy w rzeczywistości jest ona fałszywa.

Hipoteza	Decyzja o hipotezie	
	nie odrzucić	odrżucić
prawdziwa	prawidłowa	błędna
fałszywa	błędna	prawidłowa

Błąd I rodzaju kontroluje się przez zadanie małej wartości dla poziomu istotności. **Poziom istotności** jest to górne ograniczenie prawdopodobieństwa popełnienia błędu I rodzaju.

Błądu II rodzaju nie można kontrolować w taki sposób, jak błąd I rodzaju. W praktyce nie wiadomo, ile dokładnie wynosi prawdopodobieństwo popełnienia tego błędu.

Cecha X ma rozkład normalny $N(\mu, \sigma^2)$

Średnia μ oraz wariancja σ^2 są nieznane

$$H_0 : \mu = \mu_0$$

Test Studenta (poziom istotności α)

Próba: $X_1, \dots, X_n$

Statystyka testowa

$$t_{\text{emp}} = \frac{\bar{X} - \mu_0}{S} \sqrt{n} .$$

Jeżeli $|t_{\text{emp}}| > t(1 - \alpha/2; n - 1)$, to hipotezę odrzucamy.

Przykład. W biochemicznym doświadczeniu badano czas życia komórek w pewnym środowisku. Dokonano ośmiu pomiarów uzyskując wyniki (w godzinach): 4.7, 5.3, 4.0, 3.8, 6.2, 5.5, 4.5, 6.0. Czy można uznać, że średni czas życia komórek w badanym środowisku wynosi 4 godziny?

Cecha X — czas życia komórki ($X \sim N(\mu, \sigma^2)$)

$$H_0 : \mu = 4$$

Test Studenta; poziom istotności $\alpha = 0.05$

$$\bar{x} = 5, s = 0.891227, t_{\text{emp}} = 3.1736, t(0.975; 7) = 2.3646$$

Weryfikacja: Ponieważ $t_{\text{emp}} > t(0.975; 7)$, odrzucamy hipotezę

Wniosek: średni czas życia komórek w badanym środowisku nie wynosi 4 godziny.

Cecha X ma rozkład normalny $N(\mu, \sigma^2)$

Średnia μ oraz wariancja σ^2 są nieznane

$$H_0 : \sigma^2 = \sigma_0^2$$

Statystyka chi-kwadrat (poziom istotności α)

Próba: $X_1, \dots, X_n$

Statystyka testowa $\chi_{\text{emp}}^2 = \frac{\sum_i (X_i - \bar{X})^2}{\sigma_0^2}$

Wartości krytyczne $\chi^2(1 - \frac{\alpha}{2}; n - 1)$, $\chi^2(\frac{\alpha}{2}; n - 1)$

Jeżeli $\chi_{\text{emp}}^2 < \chi^2(1 - \frac{\alpha}{2}; n - 1)$ lub $\chi_{\text{emp}}^2 > \chi^2(\frac{\alpha}{2}; n - 1)$ to hipotezę $H_0 : \sigma^2 = \sigma_0^2$ odrzucamy.

Cecha $X \sim D(p)$

p nie jest znane

$$H_0 : p = p_0$$

Statystyka testowa

$$u_{\text{emp}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Wartość krytyczna: $u_{1-\frac{\alpha}{2}}$

Jeżeli $|u_{\text{emp}}| > u_{1-\frac{\alpha}{2}}$, to hipotezę odrzucamy.

Przykład. Dziesięć lat temu odsetek dzieci chorych na astmę wynosił 4%. Czy odsetek ten uległ zmianie, jeżeli w próbie dwustu dzieci rozpoznano osiemnaście przypadków astmy?

Niech X oznacza liczbę przypadków astmy wśród wylosowanych dzieci. Możemy założyć, że $X \sim B(200, p)$, gdzie p oznacza prawdopodobieństwo wylosowania dziecka chorego na astmę.

Cel: Zweryfikować hipotezę $H_0 : p = 0.04$

Zadaję poziom istotności $\alpha = 0.05$.

$$\text{Wyznaczam } \hat{p} = 0.09, u_{\text{emp}} = \frac{0.09 - 0.04}{\sqrt{\frac{0.04(1-0.04)}{200}}} = 2.887, u_{0.975} = 1.96$$

Ponieważ $|u_{\text{emp}}| > u_{0.975}$, hipotezę odrzucamy.

Wniosek: Odsetek dzieci chorych na astmę uległ zmianie.

Porównanie średnich

Cecha X_1 ma rozkład normalny $N(\mu_1, \sigma_1^2)$

Cecha X_2 ma rozkład normalny $N(\mu_2, \sigma_2^2)$

Średnia μ_1 oraz wariancja σ_1^2 są nieznane

Średnia μ_2 oraz wariancja σ_2^2 są nieznane

$\sigma_1^2 = \sigma_2^2$

$$H_0 : \mu_1 = \mu_2$$

test t-Studenta

$$t_{\text{emp}} = \frac{\bar{X}_1 - \bar{X}_2}{S_r}$$

Wartość krytyczna $t(1 - \alpha/2; n_1 + n_2 - 2)$

Jeżeli $|t_{\text{emp}}| > t(1 - \alpha/2; n_1 + n_2 - 2)$, to hipotezę $H_0 : \mu_1 = \mu_2$ odrzucamy

Porównanie frakcji

Cecha X_1 ma rozkład dwupunktowy $D(p_1)$

Cecha X_2 ma rozkład dwupunktowy $D(p_2)$

$$H_0 : p_1 = p_2$$

Statystyka testowa

$$u_{\text{emp}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

gdzie

$$\hat{p}_1 = \frac{k_1}{n_1}, \quad \hat{p}_2 = \frac{k_2}{n_2}, \quad \hat{p} = \frac{(k_1 + k_2)}{(n_1 + n_2)}$$

Jeżeli $|u_{\text{emp}}| \geq u_{1-\alpha/2}$, to hipotezę $H_0 : p_1 = p_2$ odrzucamy