

Statystyka opisowa i ekonomiczna

Stanisław Jaworski

SGGW

Katedra Ekonometrii i Statystyki

e-mail: stanislaw_jaworski@sggw.pl

Ostatnia modyfikacja: 16 marca 2020, Warszawa

Spis treści

1	Pojęcia statystyczne	1
1.1	Populacja, próba, cecha	1
1.2	Skale pomiarowe	2
2	Parametry statystyczne	3
2.1	Parametry populacji	3
2.2	Parametry próby	3
2.2.1	Miary położenia i zróżnicowania	3
2.2.2	Miary asymetrii	6
2.2.3	Miary koncentracji	6
3	Indeksy agregatowe	8
4	Estymacja	10
4.1	Estymacja punktowa	10
4.2	Estymacja przedziałowa	11
4.3	Weryfikacja hipotez statystycznych	16
5	Statystyczne miary nierówności oparte na kwantylach i de-	
	cylach	18

1 Pojęcia statystyczne

1.1 Populacja, próba, cecha

Pojęcia

Statystyka

Nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych oraz przedstawianiu wyników w postaci zestawień tabelarycznych, wykresów, itp.; posługuje się rachunkiem prawdopodobieństwa.

Statystyczna analiza danych

Etap badania statystycznego polegający na wykrywaniu – przy użyciu odpowiednich metod – prawidłowości kształtowania się zjawisk statystycznych oraz związków i zależności między nimi, a także na interpretacji wyników badań i formułowaniu wniosków.

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.

Cecha

Wielkość charakteryzująca obiekty danej populacji

Cecha niemierzalna (jakościowa)

Cecha przyjmująca wartości nie będące liczbami (np. *kolor*, *płeć*, *smak*)

Cecha mierzalna (ilościowa)

Cecha przyjmująca pewne wartości liczbowe (np. *długość*, *wytrzymałość*, *ciężar*)

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

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

1.2 Skale pomiarowe

Skale pomiarowe.

- **Skala nominalna.** W nominalnej skali pomiarowej wartości liczbowe są używane jako etykiety dla grup lub klas. Obiekty są rozróżniane za pomocą przyporządkowanych im wartościom. Interpretacja wartości polega zatem na zidentyfikowaniu klasy lub grupy, do której obiekt należy. Np. 1–mężczyzna, 2–kobieta.
- **Skala porządkowa.** W porządkowej skali pomiarowej obiekty mogą być uporządkowane według rozmiaru lub znaczenia. Obiekty mogą być uporządkowane według przypisanych do nich wartości. Np. przyporządkowujemy mrówce 1, słoniowi 2, wielorybowi 3, chcąc wyrazić, że najmniejsza to mrówka, a największy to wieloryb. Równie dobrze możemy przyporządkować inne wartości, np. -6 , $\sqrt{2}$, 2π .
- **Skala przedziałowa.** W skali przedziałowej wartość zero jest przyporządkowana dowolnie wybranemu obiektowi (także klasie lub grupie). Np. kobietom przypiszemy wartość 0, a mężczyznom 20 chcąc wyrazić to, że mężczyźni są średnio wyżsi od kobiet o 20 cm. Nie możemy jednak wtedy stwierdzić, o ile procent mężczyźni są wyżsi od kobiet.
- **Skala ilorazowa.** Jeśli powiemy, że wyniki dwóch pomiarów zapisaaliśmy w skali ilorazowej, to możemy zinterpretować ich stosunek. Np. wzrost Ani 1, a Zosi 2. Zatem Zosia jest dwa razy wyższa od Ani. Nadal jednak możemy powiedzieć, że Zosia jest wyższa od Ani o $2 - 1 = 1$ (o jedną jednostkę, którą przyjęliśmy).

2 Parametry statystyczne

2.1 Parametry populacji

Parametry populacji

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 ...

2.2 Parametry próby

2.2.1 Miary położenia i zróżnicowania

Parametry próby

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	Oznaczenie	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 kwartył	Q_1	$x_{[n/4]:n}$
górny kwartył	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 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	$\frac{S}{\bar{x}} 100\%$

Postać danych (szereg rozdzielczy, dane skumulowane):

Przedział klasowy	Liczebność	Liczebność skumulowana
$x_0 - x_1$	n_1	$n_{(1)} = n_1$
$x_1 - x_2$	n_2	$n_{(2)} = n_1 + n_2$
$\vdots$	$\vdots$	$\vdots$
$x_{k-1} - x_k$	n_k	$n_{(k)} = n_1 + n_2 + \dots + n_k (= n)$

Dla liczby p takiej, że $0 \leq p \leq 1$ niech x_p, n_p, h_p oznaczają początek, liczebność i długość przedziału zawierającego obserwację o numerze $p \cdot n$ oraz niech $n_{(p)}$ oznacza liczebność skumulowaną przedziału poprzedzającego przedział o początku x_p . Niech $\dot{x}_i$ oznacza środek i -tego przedziału.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k \dot{x}_i n_i$$

$$Me = x_{0.5} + \frac{h_{0.5}}{n_{0.5}} \left(\frac{n}{2} - n_{(0.5)} \right)$$

$$Q_1 = x_{0.25} + \frac{h_{0.25}}{n_{0.25}} \left(\frac{n}{4} - n_{(0.25)} \right)$$

$$Q_3 = x_{0.75} + \frac{h_{0.75}}{n_{0.75}} \left(\frac{3n}{4} - n_{(0.75)} \right)$$

$$Min = x_0$$

$$Max = x_k$$

$$R = Max - Min$$

$$\mathbb{S}^2 = \frac{1}{n} \sum_{i=1}^k n_i (X_i - \bar{x})^2$$

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

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

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

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

$$Q = \frac{Q_3 - Q_1}{2}$$

$$\mathbb{V} = \frac{\mathbb{S}}{\bar{x}} 100\% \text{ (współczynnik zmienności)}$$

$$V = \frac{S}{\bar{x}} 100\% \text{ (współczynnik zmienności)}$$

2.2.2 Miary asymetrii

Miary asymetrii

	Próba prosta	Szereg rozdzielczy
trzeci moment centralny: $e_3 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^3$		$e_3 = \frac{1}{n-1} \sum_{i=1}^k n_i (X_i - \bar{x})^3$
współczynnik asymetrii: $A = \frac{e_3}{S^3}$		$A = \frac{e_3}{S^3}$
pozycyjny współczynnik asymetrii: $A_1 = \frac{Q_3 - 2Me + Q_1}{2Q}$		$A_1 = \frac{Q_3 - 2Me + Q_1}{2Q}$
współczynnik skośności: $A_3 = \frac{\bar{x} - D}{S}$		$A_3 = \frac{\bar{x} - D}{S}$

2.2.3 Miary koncentracji

Współczynnik koncentracji Lorenza

Przedział	Liczebność	Częstość	Środek	t_i	z_i	$z_{(i)}$
$x_0 - x_1$	n_1	$g_1 = n_1/n$	$\dot{x}_1$	$t_1 = n_1 \dot{x}_1$	$w_1 = t_1/t$	$w_{(1)} = w_1$
$x_1 - x_2$	n_2	$g_2 = n_2/n$	$\dot{x}_2$	$t_2 = n_2 \dot{x}_2$	$w_2 = t_2/t$	$w_{(2)} = w_1 + w_2$
$\vdots$	$\vdots$	$\vdots$				
$x_{k-1} - x_k$	n_k	$g_k = n_k/n$	$\dot{x}_k$	$t_k = n_k \dot{x}_k$	$w_k = t_k/t$	$w_{(k)} = w_1 + w_2 + \dots + w_k$
Razem	n	1		t	1	

$$K = 1 - \sum_{i=1}^k [w_{(i)} + w_{(i-1)}] g_i$$

Przykład

Pożyczka		
od	do	Liczba dłużników
0	20	155
20	50	260
50	100	215
100	200	315
200	300	55
Suma		1000

i	x_{i-1}	x_i	$\dot{x}_i$	n_i	$n_{(i)}$
1	0	20	10	155	155
2	20	50	35	260	415
3	50	100	75	215	630
4	100	200	150	315	945
5	200	300	250	55	1000
SUMA				1000	

Mierniki położenia		
średnia $\bar{x}$	$\frac{87775}{1000} = 87.775$	
mediana Me	$50 + \frac{50}{215} \cdot \left(\frac{1000}{2} - 415\right) = 69.77$	
dolny kwartyl Q_1	$20 + \frac{30}{260} \cdot \left(\frac{1000}{4} - 155\right) = 30.96$	
górnny kwartyl Q_3	$100 + \frac{100}{315} \cdot \left(\frac{3 \cdot 1000}{4} - 630\right) = 138.10$	
dominanta (moda) D	$100 + 100 \cdot \frac{315-215}{2 \cdot 315 - 55 - 215} = 127.78$	
minimum Min	0	
maksimum Max	300	

Mierniki rozproszenia

$$R = 300 - 0 = 300$$

$$S^2 = \frac{1}{999} (155(10 - 87.775)^2 + 260(35 - 87.775)^2 + \dots \\ \dots + 55(250 - 87.775)^2) = 7567.11$$

$$S = \sqrt{7567.11} = 86.989$$

$$V = \frac{86.989}{87.775} 100\% = 99.1\%$$

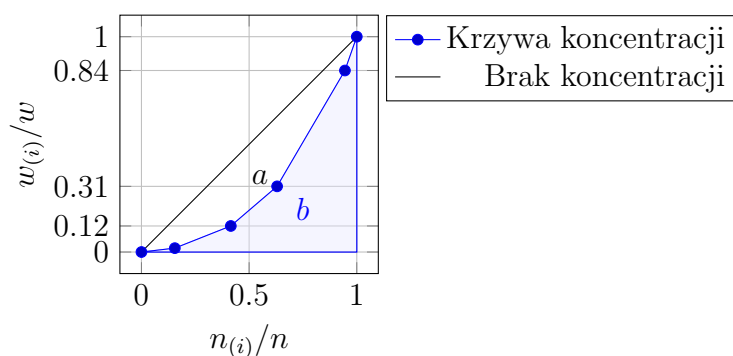
Wielobok liczebności: $\{(\dot{x}_i, n_i) : i = 1, 2, \dots, k\}$

Wielobok skumulowanej liczebności: $\{(\dot{x}_i, n_{(i)}) : i = 1, 2, \dots, k\}$

Obliczenia

i	$\dot{x}_i$	n_i	$n_{(i)}$	$n_{(i)}/n$	w_i	$w_{(i)}$	$w_{(i)}/w$
1	10	155	155	0.155	1550	1550	0.018
2	35	260	415	0.415	9100	10650	0.121
3	75	215	630	0.630	16125	26775	0.305
4	150	315	945	0.945	47250	74025	0.843
5	250	55	1000	1.000	13750	87775	1.000

gdzie $w_i = n_i \dot{x}_i$, $w_{(i)} = w_1 + w_2 + \dots + w_i$ oraz $w = \sum_i w_i$
 Krzywa koncentracji (Lorenza): $\{(\frac{n_{(i)}}{n}, \frac{w_{(i)}}{w}) : i = 0, \dots, k\}$
 Krzywa koncentracji obrazuje nierównomierność podziału ogólnej sumy wartości cechy pomiędzy poszczególne jednostki zbiorowości
 Oś X: skumulowany odsetek dłużników
 Oś Y: skumulowany odsetek sumy pożyczek
 Brak koncentracji: $\frac{n_{(i)}}{n} = \frac{w_{(i)}}{w}$, dla $i = 1, \dots, k$.
 Współczynnik koncentracji Lorenza $K = a/0.5 = 1 - 2b$
 Współczynnik koncentracji Giniego $G = \frac{a}{a+b} = \frac{a}{a+(0.5-a)} = K$



3 Indeksy agregatowe

Indeksy agregatowe

artykułu	Ilość		Cena jednostkowa	
	Rok 0	Rok 1	Rok 0	Rok 1
1	q_{10}	q_{11}	p_{10}	p_{11}
2	q_{20}	q_{21}	p_{20}	p_{21}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	q_{k0}	q_{k1}	p_{k0}	p_{k1}
Numer	Wartość	Wartość	Wartość	Wartość
1	$q_{10}p_{10}$	$q_{11}p_{11}$	$q_{10}p_{11}$	$q_{11}p_{10}$
2	$q_{20}p_{20}$	$q_{21}p_{21}$	$q_{20}p_{21}$	$q_{21}p_{20}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$q_{k0}p_{k0}$	$q_{k1}p_{k1}$	$q_{k0}p_{k1}$	$q_{k1}p_{k0}$
Suma	w_{00}	w_{11}	w_{01}	w_{10}

Indeks zmian wartości	$I_w = w_{11}/w_{00}$
Indeks Laspayresa zmian ilości	$I_q^L = w_{10}/w_{00}$
Indeks Laspayresa zmian cen	$I_p^L = w_{01}/w_{00}$
Indeks Paaschego zmian ilości	$I_q^P = w_{11}/w_{01}$
Indeks Paaschego zmian cen	$I_p^P = w_{11}/w_{10}$
Indeks Fishera zmian ilości	$I_q^F = \sqrt{I_q^L \cdot I_q^P}$
Indeks Fishera zmian cen	$I_p^F = \sqrt{I_p^L \cdot I_p^P}$

4 Estymacja

4.1 Estymacja punktowa

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

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

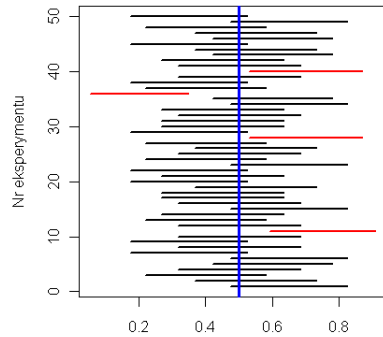
Estymator wariancji σ^2 — wariancja próbkowa

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

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 $\hat{p} = \bar{X}$.

4.2 Estymacja przedziałowa

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 :

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

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

Ilustracja estymacji przedziałowej parametru $\theta = 0.5$ (oznaczonego pionową [niebieską](#) linią) na poziomie ufności $1 - \alpha = 0.9$. Populacja z wyróżnioną cechą X .

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

Wariancja σ^2 jest nieznana

Poziom ufności: $1 - \alpha$

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

gdzie $t(\alpha; n - 1)$ jest wartością krytyczną rozkładu t -Studenta o $n - 1$ stopniach swobody.

ν	Kwantyl rozkładu t - Studenta			
	γ			
	0.1	0.05	0.025	0.01
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 $1 - \alpha = 0.95$.

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

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

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

$$t(0.05; 9) = 2.2622$$

$$t(0.05; 9) \frac{s}{\sqrt{n}} = 2.2622 \frac{0.2}{\sqrt{10}} = 0.14$$

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

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 nieznana

Poziom ufności: $1 - \alpha$

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

$\chi^2(\alpha; \nu)$ 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.

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

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

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

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 $u_{0.975} = 1.96$

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 $1 - \alpha$)

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

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

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

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

$$t(0.05; 8) = 2.3060; t(0.05; 8)s_r = 3.72$$

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

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)

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

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

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

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

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

4.3 Weryfikacja hipotez statystycznych

Weryfikacja hipotez statystycznych

Błąd I rodzaju

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

Błąd II rodzaju

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

Hipoteza	Decyzja o hipotezie	
	nie odrzucić	odrzucić
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.

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

Porównanie średniej z normą

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

Porównanie zróżnicowania z normą

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: $X_1, \dots, X_n$

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

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.

Porównanie frakcji z normą

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.

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 $t(\alpha; n_1 + n_2 - 2)$

Jeżeli $|t_{\text{emp}}| > t(\alpha; n_1 + n_2 - 2)$, 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

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

gdzie

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

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

5 Statystyczne miary nierówności oparte na kwantylach i decylach

Statystyczne miary nierówności oparte na kwantylach i decylach

Wskaźnik zróżnicowania kwintylowego

$$W_{20:20}^{(1)} = \frac{Q_{0.8}}{Q_{0.2}}$$

gdzie $Q_{0.8}$ oraz $Q_{0.2}$ są kwintylami, odpowiednio czwartym i pierwszym

$$W_{20:20}^{(2)} = \frac{\sum_{i \in GK_5} x_i}{\sum_{i \in GK_1} x_i}$$

gdzie GK_i jest i -tą grupą kwintylową. Wskaźnik zróżnicowania decylowego

$$W_{10:10}^{(1)} = \frac{Q_{0.9}}{Q_{0.1}}$$

gdzie $Q_{0.9}$ oraz $Q_{0.1}$ są decylami, odpowiednio dziewiątym i pierwszym

$$W_{10:10}^{(2)} = \frac{\sum_{i \in GD_{10}} x_i}{\sum_{i \in GD_1} x_i}$$

gdzie GD_i jest i -tą grupą decylową. Wskaźnik zróżnicowania skrajnych części rozkładu

$$K_{1/10} = \frac{\sum_{i \in GD_1} x_i}{\sum_{i \in GD_{10}} x_i} = \frac{1}{W_{10:10}^{(2)}}$$

Im wskaźnik $K_{1/10}$ jest bliższy wartości 1, tym mniejsze są nierównomierności. Wskaźnik maksymalnego wyrównania

$$E = \sum_{j \in J} 100(S_j - 1/10)$$

gdzie $J = \{1, 2, \dots, 10\}$, $S_j > 1/10$, $S_j = \frac{\sum_{i \in GD_j} x_i}{\sum_{i=1}^n x_i}$

Przykładowo, jeżeli x oznacza dochód gospodarstwa domowego, to wskaźnik maksymalnego wyrównywania wskazuje jaki procent sumy dochodów wszystkich badanych gospodarstw domowych powinien być transferowany z grup decylowych posiadających więcej niż 10% sumy dochodów do grup decylowych, których udział w sumie dochodów jest mniejszy niż 10%, aby uzyskać całkowitą równomierność rozkładu dochodów gospodarstw domowych Granica ubóstwa

$$z_u = 0.6 \cdot Me$$

Wskaźnik zagrożenia ubóstwem

$$W_{zg.ub.} = F(z_u)$$

Empiryczny wskaźnik zagrożenia ubóstwem (na podstawie próby $(X_1, \dots, X_n)$)

$$\hat{W}_{zg.ub.} = \frac{\#\{X_i \leq 0.6Me\}}{n}$$