

Statystyka opisowa i ekonomiczna

Stanisław Jaworski

SGGW

Katedra Ekonometrii i Statystyki

e-mail: stanislaw_jaworski@sggw.pl

Ostatnia modyfikacja: 27 czerwca 2020, Warszawa

Pojęcia statystyczne

Populacja, próba, cecha

Skale pomiarowe

Parametry statystyczne

Parametry populacji

Parametry próby

Miary położenia i zróżnicowania

Miary asymetrii

Miary koncentracji

Indeksy agregatowe

Wnioskowanie statystyczne

Estymacja punktowa

Estymacja przedziałowa

Weryfikacja hipotez statystycznych

Regresja

Model wahań okresowych

Model multiplikatywny (M)

Model addytywny (A)

Metoda średnich ruchomych

Wygładzanie wykładnicze

Modele bez trendu

Modele z trendem liniowym

Modele z trendem wykładniczym

Modele z trendem gasnącym

Wybór wartości początkowych

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*)

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 zapisaliśmy 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).

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.

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.

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.

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 ...

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.

Mienniki 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 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

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.

Mienniki 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.

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

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

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

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

$$Min = x_0$$

$$Max = x_k$$

$$R = \text{Max} - \text{Min}$$

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

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

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

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

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

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

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

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

Miary asymetrii

Próba prosta

trzeci moment centralny: $e_3 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^3$

współczynnik asymetrii: $A = \frac{e_3}{S^3}$

pozycyjny współczynnik asymetrii: $A_1 = \frac{Q_3 - 2Me + Q_1}{2Q}$

współczynnik skośności: $A_3 = \frac{\bar{x} - D}{S}$

Szereg rozdzielczy

$$e_3 = \frac{1}{n-1} \sum_{i=1}^k n_i (X_i - \bar{x})^3$$

$$A = \frac{e_3}{S^3}$$

$$A_1 = \frac{Q_3 - 2Me + Q_1}{2Q}$$

$$A_3 = \frac{\bar{x} - D}{S}$$

Współczynnik koncentracji Lorenza

Przedział	Liczebność	Częstość	Środek	t_i	z_i	$z_{(i)}$
$x_0 - x_1$	n_1	$g_1 = n_1/n$	$\hat{x}_1$	$t_1 = n_1\hat{x}_1$	$w_1 = t_1/t$	$w_{(1)} = w_1$
$x_1 - x_2$	n_2	$g_2 = n_2/n$	$\hat{x}_2$	$t_2 = n_2\hat{x}_2$	$w_2 = t_2/t$	$w_{(2)} = w_1 + w_2$
$\vdots$	$\vdots$	$\vdots$				
$x_{k-1} - x_k$	n_k	$g_k = n_k/n$	$\hat{x}_k$	$t_k = n_k\hat{x}_k$	$w_k = t_k/t$	$w_{(k)} = w_1 + w_2 + \dots + w_k$
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	$20 + \frac{30}{260} \cdot \left(\frac{1000}{4} - 155 \right) = 30.96$
górný 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

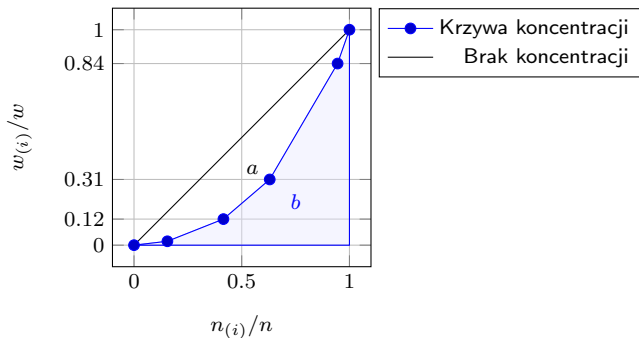
Oś X : skumulowany odsetek dłużników

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 $G = \frac{a}{a+b} = \frac{a}{a + (0.5 - a)} = K$



Indeksy agregatywne

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	$I_p^F = \sqrt{I_p^L \cdot I_p^P}$

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



Rysunek: Jerzy Neyman (1894-1981)

1. 1928: Statystyka ilorazu wiarygodności, zbiór alternatyw, błędy pierwszego i drugiego rodzaju, funkcja mocy
2. 1930: Lemat Neymana-Pearsona
3. 1934: Praca zaliczona do największych dwudziestowiecznych osiągnięć w dziedzinie metodologii statystyki
4. 1935: Przedziały ufności

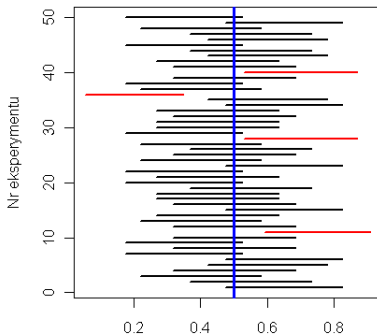
Estymacja przedziałowa

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 θ (oznaczonego pionową **niebieską** linią) na poziomie ufności $1 - \alpha$



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	$t(\gamma; \nu)$			
	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 $1 - \alpha = 0.95$.

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

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

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

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

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

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

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 normalnymŚrednia μ jest nieznanąPoziom 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.

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

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

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

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%.

Przykład

Postanowiono ocenić wartość opałową węgla dostarczonego do elektrowni. W tym celu pobrano losowo dwanaście próbek i określono wartość opałową węgla w tych próbkach. Uzyskano wyniki (w KJ/kg): 30354, 31338, 30680, 30509, 31443, 31372, 29936, 30769, 30396, 30777, 31045, 30480. Wiadomo, że głównym źródłem różnicowania uzyskanych wyników są:

1. niejednorodność partii dostarczonego węgla do elektrowni,
2. błędy pomiarowe powodujące to, że dany pomiar może dawać wynik nieznacznie mniejszy lub większy od rzeczywistej wartości opałowej próbki węgla.

Zatem te dwa źródła zmienności dają wyniki, które mogą być większe lub mniejsze od przeciętnej wartości opałowej całej partii węgla, której ocena jest celem wykonania pomiarów. Chcąc uwzględnić statystyczny błąd oceny, zdecydowano się na wyznaczenie przedziału ufności dla wartości przeciętnej na poziomie ufności 0.95. Proszę wyznaczyć ten przedział oraz go zinterpretować uwzględniając przyjęte założenia.

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 $1 - \alpha$)

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

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 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.

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

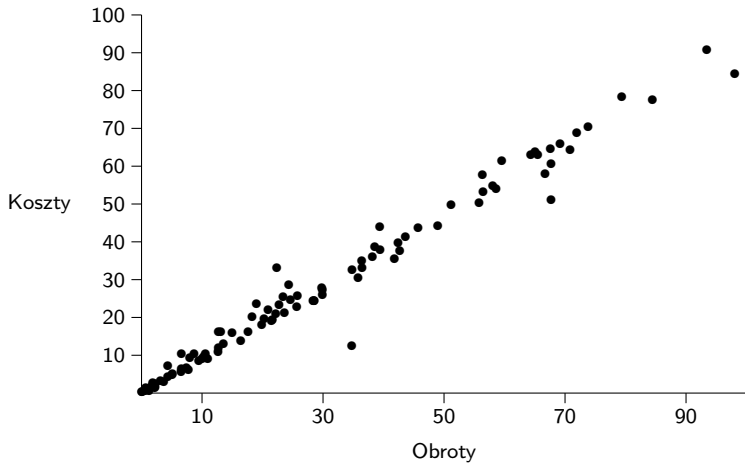
$$u_{\text{emp}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

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

Obroty	Koszty	Obroty	Koszty	Obroty	Koszty	Obroty	Koszty
0.03	0.13	7.74	6.18	22.75	23.14	43.62	41.24
0.08	0.34	7.98	9.19	23.4	25.26	45.73	43.65
0.6	0.57	8.72	10.33	23.63	21.25	48.96	44.08
0.69	0.95	9.45	8.51	24.37	28.55	51.18	49.67
0.73	1.38	10.01	9.09	24.62	24.61	55.79	50.26
0.75	0.75	10.3	9.51	25.68	22.7	56.33	57.61
0.85	0.74	10.59	10.19	25.78	25.68	56.47	53.02
1.04	0.88	10.95	8.99	28.38	24.37	58.05	54.61
1.26	0.64	12.67	10.77	28.55	24.37	58.59	53.93
1.78	1.71	12.72	11.88	28.55	24.37	59.53	61.29
1.85	2.64	12.74	16.14	29.8	27.67	64.37	63.03
1.95	1.86	13.09	16.20	29.89	27.29	65.05	63.80
2.22	1.36	13.55	13.05	29.92	25.81	65.49	62.87
2.25	2.48	15.00	15.85	34.75	12.49	66.69	57.81
2.26	1.57	16.43	13.61	34.8	32.59	67.58	64.38
3.11	3.18	17.61	16.15	35.83	30.31	67.69	51.09
3.66	2.79	18.27	20.19	36.43	34.88	67.72	60.58
4.34	7.05	19.02	23.51	36.47	32.95	69.2	65.82
4.37	4.26	19.86	18.08	38.16	35.87	70.84	64.35
5.04	4.68	20.26	19.51	38.55	38.47	71.97	68.64
5.14	4.89	20.93	21.82	39.38	43.97	73.83	70.26
6.57	5.64	21.49	19.05	39.44	37.9	79.39	78.24
6.58	6.25	21.63	19.25	41.79	35.4	84.45	77.45
6.59	10.27	22.22	20.97	42.42	39.74	93.46	90.57
7.44	6.69	22.37	33.17	42.67	37.54	98.07	84.33



Dane oryginalne – Wzory

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \text{cov}(x, y) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y},$$

$$\text{var}x = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2, \quad \text{var}y = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2,$$

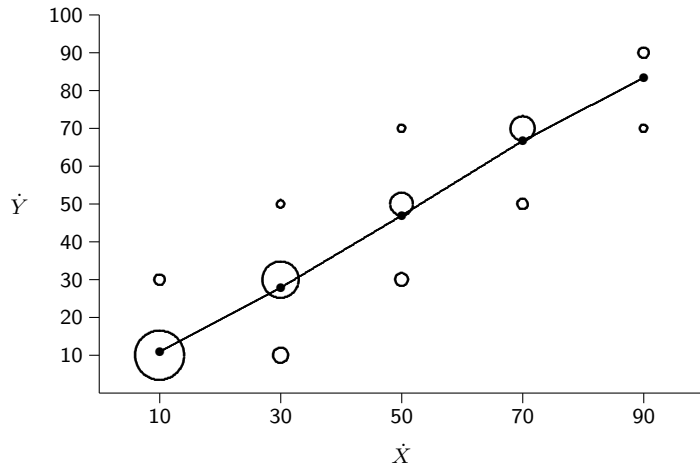
$$C(x, y) = \frac{1}{n} \text{cov}(x, y), \quad S^2(x) = \frac{1}{n} \text{var}x, \quad S^2(y) = \frac{1}{n} \text{var}y,$$

$$r_{xy} = r_{yx} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}x \cdot \text{var}y}} = \frac{C(x, y)}{S(x)S(y)}$$

$$\hat{\beta}_1 = \frac{\text{cov}(x, y)}{\text{var}x}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

x	y					$n_{i.}$	$\bar{y}_i$	$S_i(y)$
	0-20	20-40	40-60	60-80	80-100			
0-20	42	2				44	10.909	4.166
20-40	4	23	1			28	27.857	8.175
40-60		3	9	1		13	46.923	10.659
60-80			2	10		12	66.667	7.454
80-100				1	2	3	83.333	9.428
$n_{.j}$	46	28	12	12	2	100	29.2	22.257
$\bar{x}_j$	11.739	30.714	51.667	70	90	30.4	1372	
$S_j(x)$	5.635	8.421	9.860	8.165	0	22.975		484.320

x	y					$n_{i.}$	$\bar{y}_i$	$S_i(y)$
	0-20	20-40	40-60	60-80	80-100			
$\dot{x}_1 = 10$	42	2				$n_1.$	$\bar{y}_1$	$S_1(y)$
$\dot{x}_2 = 30$	4	23	1			$n_2.$	$\bar{y}_2$	$S_2(y)$
$\dot{x}_3 = 50$		3	9	1		$n_3.$	$\bar{y}_3$	$S_3(y)$
$\dot{x}_4 = 70$			2	10		$n_4.$	$\bar{y}_4$	$S_4(y)$
$\dot{x}_5 = 90$				1	2	$n_5.$	$\bar{y}_5$	$S_5(y)$
$n_{.j}$	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	n	$\bar{y}$	$S(y)$
$\bar{x}_j$	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$	$\bar{x}_5$	$\bar{x}$	$\overline{xy}$	
$S_j(x)$	$S_1(x)$	$S_2(x)$	$S_3(x)$	$S_4(x)$	$S_5(x)$	$S(x)$		$C(x, y)$



Tablica korelacyjna – Wzory

Indeksowanie

$$i = 1, 2, \dots, k, \quad j = 1, 2, \dots, m$$

Średnie

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_{i.} \dot{x}_i \quad \bar{y} = \frac{1}{n} \sum_{j=1}^m n_{.j} \dot{y}_j$$

Średnie warunkowe

$$\bar{x}_j = \frac{1}{n_{.j}} \sum_{i=1}^k n_{ij} \dot{x}_i \quad \bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^m n_{ij} \dot{y}_j$$

Wariancje

$$S^2(x) = \frac{1}{n} \sum_{i=1}^k n_{i.} (\dot{x}_i - \bar{x})^2, \quad S^2(y) = \frac{1}{n} \sum_{j=1}^m n_{.j} (\dot{y}_j - \bar{y})^2$$

Wariancje warunkowe

$$S_j^2(x) = \frac{1}{n_{\cdot j}} \sum_{i=1}^k n_{ij} (\dot{x}_i - \bar{x}_j)^2, \quad S_i^2(y) = \frac{1}{n_{i\cdot}} \sum_{j=1}^m n_{ij} (\dot{y}_j - \bar{y}_i)^2$$

Kowariancja

$$C(x, y) = \frac{1}{n} \sum_{i,j} n_{ij} (\dot{x}_i - \bar{x})(\dot{y}_j - \bar{y}) = \frac{1}{n} \sum_{i,j} n_{ij} \dot{x}_i \dot{y}_j - \bar{x}\bar{y} = \overline{\dot{x}\dot{y}} - \bar{x}\bar{y}$$

Współczynnik korelacji

$$r_{yx} = r_{xy} = \frac{C(x, y)}{S(x)S(y)}$$

Kwadrat stosunku korelacyjnego

$$e_{yx}^2 = 1 - \frac{\frac{1}{n} \sum_{i=1}^k n_i \cdot S_i^2(y)}{S^2(y)} = \frac{\frac{1}{n} \sum_{i=1}^k n_i \cdot (\bar{y}_i - \bar{y})^2}{S^2(y)} \quad (y \text{ względem } x)$$

$$e_{xy}^2 = 1 - \frac{\frac{1}{n} \sum_{j=1}^m n_j \cdot S_j^2(x)}{S^2(x)} = \frac{\frac{1}{n} \sum_{j=1}^m n_j \cdot (\bar{x}_j - \bar{x})^2}{S^2(x)} \quad (x \text{ względem } y)$$

Ocena stopnia krzywoliniowości

$$m_{yx} = e_{yx}^2 - r_{yx}^2$$

$$m_{xy} = e_{xy}^2 - r_{xy}^2$$

Test liniowości regresji

H_0 : regresja jest liniowa (tzn. Y zależy od X liniowo)

Statystyka testowa $F = \frac{e_{yx}^2 - r_{yx}^2}{1 - e_{yx}^2} \frac{n-k}{k-2}$

Hipotezę odrzucamy, gdy $F > F(1 - \alpha, k - 2, n - k)$

Model wahań okresowych

Y – obserwowana cecha

Szereg czasowy:

$$y_0, y_1, \dots, y_{n-1}$$

Składniki szeregu czasowego:

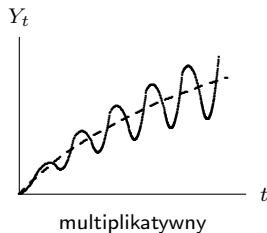
1. **trend** (*średnia wartość cechy w czasie*)
funkcja trendu $f(t)$
2. **okresowość** (*regularne odchylenia wartości cechy od trendu*)
funkcje wahań okresowych $S_i(t)$,
 $i = 0, 1, \dots, r - 1$,
 r – długość okresu,
 $S_i(t + r) = S_i(t)$
3. **szum**
składnik losowy ξ_t ,
 $E\xi_t = 0$, $D^2\xi_t = \sigma^2$.

Model multiplikatywny (M)

$$Y_t = f(t)S_i(t) + \xi_t, \quad t = 0, 1, 2, \dots$$

przy założeniach

$$\sum_{i=0}^{r-1} S_i(t) = r, \quad S_i(t) > 0$$

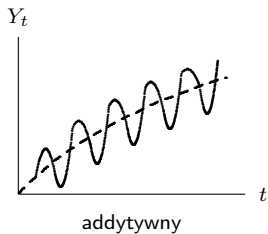


Model addytywny (A)

$$Y_t = f(t) + S_i(t) + \xi_t, \quad t = 0, 1, 2, \dots$$

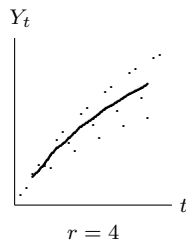
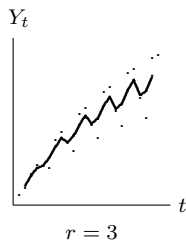
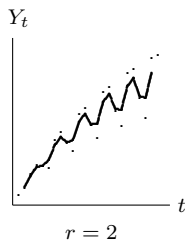
przy założeniu

$$\sum_{i=0}^{r-1} S_i(t) = 0$$



Cel.

1. Wyznaczyć długość okresu.
2. Rozpoznać postać funkcji trendu.



Średnie ruchome $\bar{y}_m$:

r – nieparzyste

$$\bar{y}_m = \frac{1}{r}(y_i + y_{i+1} + \cdots + y_{i+r-1}),$$

$$m = i + \frac{r-1}{2}, \quad i = 0, 1, \dots, n-r+1$$

r – parzyste

$$\bar{y}_m = \frac{1}{r}\left(\frac{1}{2}y_i + y_{i+1} + \cdots + y_{i+r-1} + \frac{1}{2}y_{i+r}\right),$$

$$m = i + \frac{r}{2}, \quad i = 0, 1, \dots, n-r$$

Etapy analizy szeregu czasowego:

1. Estymacja funkcji trendu.
2. Estymacja funkcji okresowości.
3. Prognoza.

Estymacja funkcji trendu

1. Metoda najmniejszych kwadratów.
2. Inne.

Przypadek szczególny: $f(t) = \beta_0 + \beta_1 t$

$$\hat{\beta}_0 = A \sum_{t=0}^{n-1} y_t + C \sum_{t=0}^{n-1} t y_t, \quad \hat{\beta}_1 = C \sum_{t=0}^{n-1} y_t + B \sum_{t=0}^{n-1} t y_t,$$

$$A = \frac{2(2n-1)}{n(n+1)}, \quad B = \frac{12}{n(n+1)(n-1)}, \quad C = \frac{-6}{n(n+1)}.$$

Estymacja funkcji okresowości

(metoda wskaźników)

Surowe wskaźniki okresowości

Model addytywny

$$S_{si} = \frac{1}{m} \sum_{j=0}^{m-1} y_{i+jr} - \frac{1}{m} \sum_{j=0}^{m-1} \hat{f}(i+jr)$$

Model multiplikatywny

$$S_{si} = \frac{\sum_{j=0}^{m-1} y_{i+jr}}{\sum_{j=0}^{m-1} \hat{f}(i+jr)}$$

gdzie $\hat{f}$ – oszacowana funkcja trendu, $m = \left\lfloor \frac{n-i}{r} \right\rfloor + 1$ ($[a]$ oznacza część całkowitą liczby a)

Estymacja funkcji okresowości

(metoda wskaźników)

Oczyszczone wskaźniki okresowości

Model addytywny

$$S_{oi} = S_{si} - \frac{1}{r} \sum_{j=0}^{r-1} S_{sj}$$

Model multiplikatywny

$$S_{oi} = S_{si}k, \text{ gdzie } k = r / \sum_{j=0}^{r-1} S_{sj}$$

Prognoza

dla chwili $n' \geq n$

► Punktowa

model (A)

model (M)

$$\hat{y}_{n'} = \hat{f}(n') + S_{oi} \quad \hat{y}_{n'} = \hat{f}(n') S_{oi}$$

gdzie $\hat{y}_{n'}$ oznacza prognozowaną wartość szeregu oraz $i = (n' \bmod r)$

Prognoza

dla chwili $n' \geq n$

► Przedziałowa

$$(\hat{y}_{n'} - S, \hat{y}_{n'} + S)$$

gdzie

$$S = \sqrt{\frac{1}{n-1} \sum_{t=0}^{n-1} (y_t - \hat{y}_t)^2}$$

oraz

$$\hat{y}_t = \begin{cases} \hat{f}(t) + S_{o(t \bmod r)} & \text{dla modelu (A)} \\ \hat{f}(t) S_{o(t \bmod r)} & \text{dla modelu (M)} \end{cases}$$

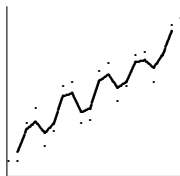
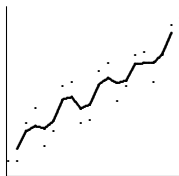
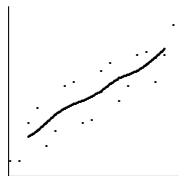
Przykład

Zanotowano ilość zleceń wykonanych przez pewną firmę w kolejnych kwartałach lat 1996 – 2000:

Lata	Kwartały			
	I	II	III	IV
1996	4	4	14	18
1997	8	12	24	25
1998	14	15	28	30
1999	20	24	32	33
2000	25	32	40	42
średnia	14.2	17.4	27.6	29.6

Interesuje nas ilość zleceń w trzecim kwartale 2001 roku.

Metoda średnich ruchomych

 $r = 2$  $r = 3$  $r = 4$

Dla $r = 2$: $m = 1, 2, \dots, 18$

Dla $r = 3$: $m = 1, 2, \dots, 18$

Dla $r = 4$: $m = 2, \dots, 17$

Estymacja funkcji trendu.

$$A = \frac{2(2 \cdot 20 - 1)}{20(20 + 1)} = 0.1857$$

$$B = \frac{12}{20(20 + 1)(20 - 1)} = 0.0015$$

$$C = \frac{-6}{20(20 + 1)} = -0.0143$$

Estymacja funkcji trendu.

$$\sum_{t=0}^{19} y_t = 444, \quad \sum_{t=0}^{19} ty_t = 5311$$

$$\hat{\beta}_0 = A \sum_{t=0}^{19} y_t + C \sum_{t=0}^{19} ty_t = 6.586$$

$$\hat{\beta}_1 = C \sum_{t=0}^{19} y_t + B \sum_{t=0}^{19} ty_t = 1.644$$

$$\hat{f}(t) = 1.644t + 6.586$$

Estymacja funkcji okresowości.

Wartości trendu

Lata	Kwartały			
	I	II	III	IV
1996	6.586	8.229	9.873	11.517
1997	13.160	14.804	16.447	18.091
1998	19.735	21.378	23.022	24.665
1999	26.309	27.953	29.596	31.240
2000	32.884	34.527	36.171	37.814
średnia	19.735	21.378	23.022	24.665

Estymacja funkcji okresowości.

średnia	I	II	III	IV
dane oryginalne	14.200	17.400	27.600	29.600
trend	19.735	21.378	23.022	24.665

Surowe wskaźniki okresowości

$$S_{s0} = 14.200 - 19.735 = -5.535$$

$$S_{s1} = 17.400 - 21.378 = -3.978$$

$$S_{s2} = 27.600 - 23.022 = 4.578$$

$$S_{s3} = 29.600 - 24.665 = 4.935$$

Oczyszczone wskaźniki okresowości

Ponieważ $S_{s0} + S_{s1} + S_{s2} + S_{s3} = 0$, zatem $S_{oi} = S_{si}$, dla $i = 0, 1, 2, 3$.

Prognoza dla trzeciego kwartału 2001 roku $(n' = 22)$ **Prognoza punktowa**

$$\hat{y}_{22} = \hat{f}(22) + S_{o2} = 1.644 \cdot 22 + 6.586 + 4.578 = 47.332$$

Prognoza przedziałowa

$$S^2 = \frac{1}{19} ((4 - 6.586 + 5.535)^2 + (4 - 8.229 + 3.978)^2 + \dots$$

$$\dots + (42 - 37.814 - 4.935)^2) = 2.942$$

$$\left(47.332 - \sqrt{2.942}, 47.332 + \sqrt{2.942} \right) =$$

$$= (45.62, 49.05)$$

Wniosek. W trzecim kwartale 2001 roku należy spodziewać się co najmniej 45. zleceń, ale mniej niż 50.

WYGŁADZANIE WYKŁADNICZE

Metoda obróbki szeregu czasowego zmniejszająca jego wariancję za pomocą ważonej średniej ruchomej z przeszłych wartości, o wagach malejących wykładniczo wraz z odległością w czasie

Modele bez trendu

Oznaczenia

- ▶ y_t – szereg czasowy ($t = 1, \dots, N$)
- ▶ S_t – szereg „wygładzony” ($t = 1, \dots, N$)
- ▶ α – stała wygładzająca
- ▶ p – długość okresu
- ▶ I_t – wskaźnik sezonowości w chwili t
- ▶ δ – stała wygładzająca sezonowość

Wyrównania proste

- ▶ **Bez sezonowości**

$$S_t = \alpha y_t + (1 - \alpha)S_{t-1}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t$

- ▶ **Sezonowość addytywna**

$$S_t = \alpha(y_t - I_{t-p}) + (1 - \alpha)S_{t-1}$$

$$I_t = \delta(y_t - S_t) + (1 - \delta)I_{t-p}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t + I_{t-p+m}$

- ▶ **Sezonowość multiplikatywna**

$$S_t = \alpha(y_t / I_{t-p}) + (1 - \alpha)S_{t-1}$$

$$I_t = \delta(y_t / S_t) + (1 - \delta)I_{t-p}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t I_{t-p+m}$

Modele z trendem liniowym

Oznaczenia

- ▶ y_t – szereg czasowy ($t = 1, \dots, N$)
- ▶ S_t – szereg „wygładzony” ($t = 1, \dots, N$)
- ▶ α – stała wygładzająca
- ▶ p – długość okresu
- ▶ I_t – wskaźnik sezonowości w chwili t
- ▶ δ – stała wygładzająca sezonowość
- ▶ T_t – trend „wygładzony”
- ▶ γ – stała wygładzająca trend
- ▶ Φ – stała „gasząca” trend, $\Phi \in (0, 1)$
- ▶ $e_t = y_t - \hat{y}_{t-1}(1)$

Modele Holta – Wintersa

► Bez sezonowości

$$S_t = \alpha y_t + (1 - \alpha)(S_{t-1} + T_{t-1})$$

$$T_t = \gamma(S_t - S_{t-1}) + (1 - \gamma)T_{t-1}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t + mT_t$

► Sezonowość addytywna

$$S_t = \alpha(y_t - I_{t-p}) + (1 - \alpha)(S_{t-1} + T_{t-1})$$

$$T_t = \gamma(S_t - S_{t-1}) + (1 - \gamma)T_{t-1}$$

$$I_t = \delta(y_t - S_t) + (1 - \delta)I_{t-p}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t + mT_t + I_{t-p+m}$

► Sezonowość multiplikatywna

$$S_t = \alpha(y_t / I_{t-p}) + (1 - \alpha)(S_{t-1} + T_{t-1})$$

$$T_t = \gamma(S_t - S_{t-1}) + (1 - \gamma)T_{t-1}$$

$$I_t = \delta(y_t / S_t) + (1 - \delta)I_{t-p}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = (S_t + mT_t)I_{t-p+m}$

Model Browna

► Bez sezonowości

$$S_t = \alpha y_t + (1 - \alpha)S_{t-1}$$

$$T_t = \gamma(S_t - S_{t-1}) + (1 - \gamma)T_{t-1}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t + ((1 - \alpha)/\alpha)T_t + mT_t$

Modele z trendem wykładniczym

► Bez sezonowości

$$S_t = \alpha y_t + (1 - \alpha)(S_{t-1}T_{t-1})$$
$$T_t = \gamma(S_t/S_{t-1}) + (1 - \gamma)T_{t-1}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t T_t^m$

► Sezonowość addytywna

$$S_t = \alpha(y_t - I_{t-p}) + (1 - \alpha)(S_{t-1}T_{t-1})$$
$$T_t = \gamma(S_t/S_{t-1}) + (1 - \gamma)T_{t-1}$$
$$I_t = \delta(y_t - S_t) + (1 - \delta)I_{t-p}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t T_t^m + I_{t-p+m}$

► Sezonowość multiplikatywna

$$S_t = \alpha(y_t/I_{t-p}) + (1 - \alpha)(S_{t-1}T_{t-1})$$
$$T_t = \gamma(S_t/S_{t-1}) + (1 - \gamma)T_{t-1}$$
$$I_t = \delta(y_t/S_t) + (1 - \delta)I_{t-p}$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = (S_t T_t^m) I_{t-p+m}$

Modele z trendem gasnącym

► Bez sezonowości

$$S_t = S_{t-1} + \Phi T_{t-1} + \alpha e_t$$

$$T_t = \Phi T_{t-1} + \alpha \gamma e_t$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t + \sum_{i=1}^m \Phi^i T_i$

► Sezonowość addytywna

$$S_t = S_{t-1} + \Phi T_{t-1} + \alpha(2 - \alpha)e_t$$

$$T_t = \Phi T_{t-1} + \alpha(\alpha - \Phi + 1)e_t$$

$$I_t = I_{t-p} + \delta[1 - \alpha(2 - \alpha)]e_t$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = S_t + \sum_{i=1}^m \Phi^i T_i + I_{t-p+m}$

► Sezonowość multiplikatywna

$$S_t = S_{t-1} + \Phi T_{t-1} + \alpha(2 - \alpha)e_t / I_{t-p}$$

$$T_t = \Phi T_{t-1} + \alpha(\alpha - \Phi + 1)e_t / I_{t-p}$$

$$I_t = I_{t-p} + \delta[1 - \alpha(2 - \alpha)]e_t / S_t$$

Prognoza w chwili $t + m$: $\hat{y}_t(m) = (S_t + \sum_{i=1}^m \Phi^i T_i) / I_{t-p+m}$

Wybór wartości początkowych

Modele bez trendu

$$S_0 = \bar{y}$$

(wartości w pełnych cyklach sezonowych)

Modele z trendem liniowym

Bez sezonowości

$$T_0 = (y_N - y_1)/(N - 1)$$

$$S_0 = y_1 - T_0/2$$

Z sezonowością

$$T_0 = (M_k - M_1)/((k - 1)p)$$

$$S_0 = M_1 - T_0/2$$

gdzie:

k – liczba pełnych cykli sezonowych

M_k – średnia ostatniego cyklu sezonowego

M_1 – średnia pierwszego cyklu sezonowego

Modele z trendem wykładniczym

Bez sezonowości

$$T_0 = (y_2/y_1)$$

$$S_0 = y_1/\sqrt{T_0}$$

Z sezonowością

$$T_0 = \exp\{(\ln(M_k) - \ln(M_1))/p\}$$

$$S_0 = \exp\{(\ln(M_1) - p \cdot \ln(T_0))/2\}$$

Modele z trendem gasnącym

Bez sezonowości

$$T_0 = (1/\Phi)(y_N - y_1)/(N - 1)$$

$$S_0 = y_1 - T_0/2$$

Z sezonowością

$$T_0 = (1/\Phi)(M_k - M_1)/[(k - 1)p]$$

$$S_0 = M_1 - pT_0/2$$

Dla wszystkich modeli z sezonowością

I_0 : przy pomocy dekompozycji sezonowej

Wybór parametru α :

„odgadnąć”

MSE=Min!, MEA=Min!

Pomiar dokładności prognoz

Średni błąd

$$ME = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)$$

Średni błąd względny

$$MPE = \frac{1}{n} \sum_{i=1}^n \frac{\hat{y}_i - y_i}{y_i}$$

Średni absolutny błąd

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

Średni absolutny błąd względny

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

Pomiar dokładności prognoz

Błąd średniokwadratowy

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Względny błąd średniokwadratowy

$$\text{MSPE} = \frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{y}_i - y_i}{y_i} \right)^2$$

Pierwiastek błędu średniokwadratowego

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

Pierwiastek względnego błędu średniokwadratowego

$$\text{RMSPE} = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{y}_i - y_i}{y_i} \right)^2}$$

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

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

$$\text{gdy } J = \{1, 2, \dots, 10\}, S_j > 1/10, S_j = \frac{\sum_{i \in G_{D_j}} x_i}{\sum_{i=1}^n x_i}$$

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 $(X_1, \dots, X_n)$)

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