Uniwersytet Ekonomiczny w Katowicach

Statystyka opisowa – Wzory

I. Analiza struktury

1. Miary tendencji centralnej (średnie, przeciętne)

Średnia arytmetyczna		
Dla szeregu wyliczającego:	Dla sz. ważonego dla zmiennej skokowej	Dla sz. ważonego dla zm. ciągłej (przedziałowego)
$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i w_i$	$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} \hat{x}_i n_i = \sum_{i=1}^{k} \hat{x}_i w_i$
gdzie n_i – liczebności, w_i – częstości, $\hat{x}_i$ – środek przedziału k, – liczba przedziałów (grup)		

Dominanta	
Dla sz. ważonego dla zmiennej skokowej	Dla sz. ważonego dla zm. ciągłej (przedziałowego)
$D = x_D$ dla której $n_D = \max\{n_i\}$	$$D = x_D + \frac{g_D - g_{D-1}}{(g_D - g_{D-1}) + (g_D - g_{D+1})} \cdot \Delta_D,$$ gdzie x_D – lewy koniec przedziału z D (tj. przedziału o największej gęstości), g_D – gęstość przedziału z D, g_{D-1} – gęstość przedziału poprzedzającego przedział D, g_{D+1} – gęstość przedziału następującego po przedziale D, Δ_D – długość przedziału D.
	Jeśli *szereg ma przedziały o równej długości*, to można korzystać ze wzoru: $$D = x_D + \frac{n_D - n_{D-1}}{(n_D - n_{D-1}) + (n_D - n_{D+1})} \cdot \Delta_D,$$ gdzie n_D – liczebność przedziału z D, n_{D-1} – liczebność przedziału poprzedzającego przedział D, n_{D+1} – liczebność przedziału następującego po przedziale D,

Kwantyl rzędu p	
Dla sz. ważonego dla zmiennej skokowej	Dla sz. ważonego dla zm. ciągłej (przedziałowego)
$Q_p = x_{[N\cdot p]+1}$ gdzie p to rząd kwantyla **Mediana** to $Me = Q_{0,5}$ jeśli N jest parzyste, to $Me = \frac{x_{0,5N} + x_{0,5N+1}}{2}$ jeśli N jest nieparzyste,to $Me = x_{0,5(N+1)}$	$$Q_p = x_{Q_p} + \frac{p \cdot N - cum\, n_{Q_p - 1}}{n_{Q_p}} \cdot \Delta_{Q_p}$$ lub $$Q_p = x_{Q_p} + \frac{p - cum\, w_{Q_p - 1}}{w_{Q_p}} \cdot \Delta_{Q_p}$$ gdzie x_{Q_p} – lewy koniec przedziału z Q_p $cum\, n_{Q_p-1}$ – skumulowana liczebność do przedziału poprzedzającego przedział z Q_p, $cum\, w_{Q_p-1}$ – skumulowana częstość do przedziału poprzedzającego przedział z Q_p, n_{Q_p} – liczebność przedziału z Q_p, w_{Q_p} – częstość przedziału z Q_p, Δ_{Q_p} – długość przedziału z Q_p,

2. Miary zróżnicowania (rozproszenia, zmienności, dyspersji)

Wariancja

Dla szeregu wyliczającego:	Dla sz. ważonego dla zmiennej skokowej	Dla sz. ważonego dla zm. ciągłej (przedziałowego)
$S^2(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$	$S^2(x) = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i$	$S^2(x) = \frac{1}{N}\sum_{i=1}^{k}(\hat{x}_i - \bar{x})^2 n_i$

Odchylenie standardowe

$$s(x) = \sqrt{S^2(x)}$$

Odchylenie przeciętne

Dla szeregu wyliczającego:	Dla sz. ważonego dla zmiennej skokowej	Dla sz. ważonego dla zm. ciągłej (przedziałowego)
$d(x) = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i - \bar{x}\rvert$	$d(x) = \frac{1}{N}\sum_{i=1}^{k}\lvert x_i - \bar{x}\rvert n_i$	$d(x) = \frac{1}{N}\sum_{i=1}^{k}\lvert\hat{x}_i - \bar{x}\rvert n_i$

Klasyczny współczynnik zmienności

$$V(x) = \frac{s(x)}{\bar{x}} \cdot 100\%$$

Rozstęp

$$R = x_{max} - x_{min}$$

Odchylenie ćwiartkowe

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

Pozycyjny współczynnik zmienności

$$V_Q = \frac{Q}{Me}$$

3. Miary asymetrii (skośności)

Moment centralny trzeciego rzędu

Dla szeregu wyliczającego:	Dla sz. ważonego dla zmiennej skokowej	Dla sz. ważonego dla zm. ciągłej (przedziałowego)
$M_3(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3$	$M_3(x) = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i$	$M_3(x) = \frac{1}{N}\sum_{i=1}^{k}(\hat{x}_i - \bar{x})^3 n_i$

Zestandaryzowany moment centralny trzeciego rzędu

$$\lambda_3(x) = \frac{M_3(x)}{(s(x))^3}$$

Wskaźnik asymetrii

$$W_s = \bar{x} - D$$

Współczynnik asymetrii Pearsona

$$\gamma = \frac{\bar{x} - D}{s(x)}$$

Współczynnik asymetrii Yule'a-Kendall'a (pozycyjny)

$$A = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1}$$

Kowariancja

$$\operatorname{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})\cdot(y_i-\bar{y})$$

Współczynnik korelacji liniowej Pearsona

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{s(x)\cdot s(y)}$$

gdzie
$s(x)$, $s(y)$ – odchylenia standardowe zmiennych x, y.

- r_{xy} bada tylko **liniową** zależność między zmiennymi,
- $r_{xy} \in \langle -1, 1\rangle$ i pozwala określić siłę i kierunek zależności liniowej
  - jeśli $|r_{xy}|$ jest bliskie 0, to mamy słabą zależność liniową między zmiennymi,
  - jeśli $|r_{xy}|$ jest bliskie 1, to mamy silną zależność liniową między zmiennymi,
  - jeśli $r_{xy} > 0$, to zależność między zmiennymi jest dodatnia,
  - jeśli $r_{xy} < 0$, to zależność między zmiennymi jest ujemna.

Funkcja regresji (II rodzaju)

$$\hat{y}_i = ax_i + b$$

gdzie

$$a = \frac{\operatorname{cov}(x,y)}{S^2(x)}, \qquad b = \bar{y} - a\bar{x}$$

$\hat{y}_i$ to wartości teoretyczne zm. Y, $S^2(x)$ – wariancja zm. X, zaś $\bar{x}$, $\bar{y}$ – średnie dla zm. X i Y.

Miary dopasowania funkcji regresji

- odchylenie standardowe reszt:

$$s_u = \sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{N-k}}$$

  gdzie k to liczba parametrów funkcji regresji;
- współczynnik zbieżności:

$$\varphi^2 = \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}$$

- współczynnik determinacji:

$$R^2 = 1-\varphi^2$$

  - $R^2 \in \langle 0, 1\rangle$,
  - im większe R^2 tym lepsze dopasowanie funkcji regresji do danych.

Standardowy błąd prognozy

$$s_u \cdot \sqrt{1 + \frac{1}{N} + \frac{(x_{new}-\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}}$$

gdzie x_{new} to nowa obserwacja, dla której chcemy zrobić prognozę $\hat{f}(x_{new})$;

Tablica korelacyjna

$X \setminus Y$	y_1	y_2	$\ldots$	y_m	$n_{i\cdot}$
x_1	n_{11}	n_{12}	$\ldots$	n_{1m}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\ldots$	n_{2m}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_n	n_{n1}	n_{n2}	$\ldots$	n_{nm}	$n_{n\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot m}$	N

Stosunki korelacyjne (wskaźniki siły korelacyjnej)

$*$ zm. Y względem zm. X

$$e_{y/x} = \sqrt{\frac{S^2(\overline{y}|_{x_i})}{S^2(y)}}$$

gdzie $S^2(\overline{y}|_{x_i}) = \frac{1}{N}\sum_{i=1}^{n}(\overline{y}|_{x_i} - \bar{y})^2 n_{i\cdot}$ oraz $S^2(y) = \frac{1}{N}\sum_{j=1}^{m}(y_j - \bar{y})^2 n_{\cdot j}$

$*$ zm. X względem zm. Y

$$e_{x/y} = \sqrt{\frac{S^2(\overline{x}|_{y_j})}{S^2(x)}}$$

gdzie $S^2(\overline{x}|_{y_j}) = \frac{1}{N}\sum_{j=1}^{m}\left(\overline{x}|_{y_j} - \bar{x}\right)^2 n_{\cdot j}$ oraz $S^2(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2 n_{i\cdot}$

$e_{y/x}, e_{y/x} \in \langle 0, 1\rangle$, $\overline{x}|_{y_j}$, $\overline{y}|_{x_i}$ – średnie warunkowe.

III. Analiza dynamiki zjawisk

Przyrosty absolutne

$*$ o stałej podstawie: $\Delta_{t/c} = y_t - y_c$ $\quad$ $*$ łańcuchowe: $\Delta_{t/t-1} = y_t - y_{t-1}$

Przyrosty względne

$*$ o stałej podstawie: $d_{t/c} = \frac{y_t - y_c}{y_c}$ $\quad$ $*$ łańcuchowe: $d_{t/t-1} = \frac{y_t - y_{t-1}}{y_{t-1}}$

Indeksy indywidualne

$*$ o stałej podstawie: $i_{t/c} = \frac{y_t}{y_c}$ $\quad$ $*$ łańcuchowe: $i_{t/t-1} = \frac{y_t}{y_{t-1}}$

Średni indeks zmian

$$\bar{i}_G = \sqrt[n-1]{\prod_{t=2}^{n} i_{t/t-1}} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

Prognoza na k okresów w przyszłość

$$\tilde{y}_{t+k} = y_t \cdot (\bar{i}_G)^k$$

Agregatowy indeks wartości

$$I_w = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{i0}}$$

gdzie p_{i0}, p_{it} – ceny w okresie bazowym i badanym, q_{i0}, q_{it} – ilości w okresie bazowym i badanym.

Agregatowe indeksy cen

$*$ formuła Laspeyresa $I_{p/q_0} = \frac{\sum_{i=1}^{n} p_{it}q_{i0}}{\sum_{i=1}^{n} p_{i0}q_{i0}}$ $\quad$ $*$ formuła Paaschego: $I_{p/q_t} = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{it}}$

Agregatowe indeksy ilości

$*$ formuła Laspeyresa $I_{q/p_0} = \frac{\sum_{i=1}^{n} p_{i0}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{i0}}$ $\quad$ $*$ formuła Paaschego: $I_{q/p_t} = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{it}q_{i0}}$