Analiza danych

Postać danych (próba prosta, dane indywidualne): $X_1, X_2, \ldots, X_n$. Niech $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ będzie uporządkowanym ciągiem danych.
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	$n_{(k)} = n_1 + n_2 + \cdots + n_k (= n)$

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.

Mierniki położenia	Próba prosta	Szereg rozdzielczy
średnia $\bar{x}$	$\frac{1}{n}\sum_{i=1}^{k} X_i$	$\frac{1}{n}\sum_{i=1}^{k} \dot{x}_i n_i$
mediana Me	$X_{(n+1)/2:n}$(n nieparzyste) $(X_{n/2:n} + X_{n/2+1:n})/2$($n$ parzyste)	$x_{0.5} + \frac{h_{0.5}}{n_{0.5}}\left(\frac{n}{2} - n_{(0.5)}\right)$
dolny kwartyl Q_1	$X_{[n/4]:n}$	$x_{0.25} + \frac{h_{0.25}}{n_{0.25}}\left(\frac{n}{4} - n_{(0.25)}\right)$
górny kwartyl Q_3	$X_{[3n/4]:n}$	$x_{0.75} + \frac{h_{0.75}}{n_{0.75}}\left(\frac{3n}{4} - n_{(0.75)}\right)$
dominanta (moda) D	najczęściej występująca wartość	$x_D + h_D \frac{n_D - n_{D-1}}{2n_D - n_{D+1} - n_{D-1}}$
minimum Min	$X_{1:n}$	x_0
maksimum Max	$X_{n:n}$	x_k

Mierniki rozproszenia	Próba prosta	Szereg rozdzielczy
rozstęp R	$Max - Min$	$Max - Min$
wariancja $\mathcal{S}^2$	$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2$	$\frac{1}{n}\sum_{i=1}^{k} n_i(X_i - \bar{x})^2$
odchylenie standardowe $\mathcal{S}$	$\sqrt{\mathcal{S}^2}$	$\sqrt{\mathcal{S}^2}$
odchylenie przeciętne d	$\frac{1}{n}\sum_{i=1}^{n}\lvert X_i - \bar{x}\rvert$	$\frac{1}{n}\sum_{i=1}^{k} n_i\lvert X_i - \bar{x}\rvert$
odchylenie ćwiartkowe Q	$\frac{Q_3 - Q_1}{2}$	$\frac{Q_3 - Q_1}{2}$
współczynnik zmienności V	$\frac{\mathcal{S}}{\bar{x}}100\%$	$\frac{\mathcal{S}}{\bar{x}}100\%$

Koncentracja Lorentza

Przedział	Liczebn.	Częstość	Środek	t_i	z_i	$z_{(i)}$
$x_0 - x_1$	n_1	$w_1 = n_1/n$	$\dot{x}_1$	$t_1 = n_1\dot{x}_1$	$z_1 = t_1/t$	$z_{(1)} = z_1$
$x_1 - x_2$	n_2	$w_2 = n_2/n$	$\dot{x}_2$	$t_2 = n_2\dot{x}_2$	$z_2 = t_2/t$	$z_{(2)} = z_1 + z_2$
$\vdots$	$\vdots$	$\vdots$				
$x_{k-1} - x_k$	n_k	$w_k = n_k/n$	$\dot{x}_k$	$t_k = n_k\dot{x}_k$	$z_k = t_k/t$	$z_{(k)} = z_1 + \cdots + z_k$
Razem	n	1		t	1	

Współczynnik koncentracji Lorentza

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

Dynamika zjawisk (indeksy)

Numer artykułu	Ilość Rok 0	Ilość Rok 1	Cena jednostkowa Rok 0	Cena jednostkowa 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_{1k}	p_{k0}	p_{1k}

Numer	Wartość	Wartość	Wartość	Wartość
1	$w_{1,00} = q_{10}p_{10}$	$w_{1,11} = q_{11}p_{11}$	$w_{1,01} = q_{10}p_{11}$	$w_{1,10} = q_{11}p_{10}$
2	$w_{2,00} = q_{20}p_{20}$	$w_{2,11} = q_{21}p_{21}$	$w_{2,01} = q_{20}p_{21}$	$w_{2,10} = q_{21}p_{20}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$w_{k,00} = q_{k0}p_{k0}$	$w_{k,11} = q_{k1}p_{k1}$	$w_{k,01} = q_{k0}p_{k1}$	$w_{k,10} = q_{k1}p_{k0}$
Razem	w_{00}	w_{11}	w_{01}	w_{10}

Indeks zmian wartości $I_w = w_{11}/w_{00}$
Indeks Laspayresa zmian ilości ${}_LI_{qp} = w_{10}/w_{00}$
Indeks Laspayresa zmian cen ${}_LI_{pq} = w_{01}/w_{00}$
Indeks Paaschego zmian ilości ${}_PI_{qp} = w_{11}/w_{01}$
Indeks Paaschego zmian cen ${}_PI_{pq} = w_{11}/w_{10}$
Indeks Fishera zmian ilości ${}_FI_q = \sqrt{{}_LI_{qp} \cdot {}_PI_{qp}}$

Indeks Fishera zmian cen ${}_F I_p = \sqrt{{}_L I_{pq} \cdot {}_P I_{pq}}$

Czas	Obserwacja	Indeksy jednopodstawowe			Indeksy łańcuchowe		
		absolutne	względne	$i_{t\|c}$	absolutne	względne	$i_{t\|t-1}$
t_0	y_0	0	$\cdot$	1	0	$\cdot$	$\cdot$
t_1	y_1	$y_1 - y_0$	$(y_1 - y_0)/y_0$	y_1/y_0	$y_1 - y_0$	$(y_1 - y_0)/y_0$	y_1/y_0
t_2	y_2	$y_2 - y_0$	$(y_2 - y_0)/y_0$	y_2/y_0	$y_2 - y_1$	$(y_2 - y_1)/y_1$	y_2/y_1
t_3	y_3	$y_3 - y_0$	$(y_3 - y_0)/y_0$	y_3/y_0	$y_3 - y_2$	$(y_3 - y_2)/y_2$	y_3/y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_k	y_k	$y_k - y_0$	$(y_k - y_0)/y_0$	y_k/y_0	$y_k - y_{k-1}$	$(y_k - y_{k-1})/y_{k-1}$	y_k/y_{k-1}

Średnie tempo zmian $\bar{i}_{t|t-1} = \sqrt[k-1]{i_{2|1} i_{3|2} \dots i_{k|k-1}}$

Inne

- $\mathrm{var}x = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2$
- $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
- $\mathrm{cov}(x, y) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$
- $(\bar{X} - t(\alpha; n-1)\frac{S}{\sqrt{n}}, \bar{X} + t(\alpha; n-1)\frac{S}{\sqrt{n}})$
- $t_{\mathrm{emp}} = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$
- $(\bar{X}_1 - \bar{X}_2 - t(\alpha; n_1 + n_2 - 2)s_r, \bar{X}_1 - \bar{X}_2 + t(\alpha; n_1 + n_2 - 2)s_r)$,

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

- $t_{\mathrm{emp}} = \frac{\bar{X}_1 - \bar{X}_2}{S_r}$
- $\left(\hat{p} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$
- $u_{\mathrm{emp}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

- $\hat{p}_1 - \hat{p}_2 \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- $u_{\text{emp}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$
- **Regresja.**

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x}_i)^2}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$r_{emp} = r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (y_i - \bar{y}_i)^2 \sum_i (x_i - \bar{x})^2}}$$

Oszacowania przedziałowe dla β_0 oraz β_1 są postaci

$$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n-2) S_{\beta_1}, \hat{\beta}_1 + t(\alpha; n-2) S_{\beta_1})$$

$$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n-2) S_{\beta_0}, \hat{\beta}_0 + t(\alpha; n-2) S_{\beta_0})$$

gdzie

$$S^2_{\beta_1} = \frac{S^2}{\text{var}x}, \quad S^2_{\beta_0} = \frac{S^2}{\text{var}x} \left(\frac{\text{var}x}{n} + \bar{x}^2 \right)$$

$$S^2 = \frac{\text{var}y - \hat{\beta}_1 \text{cov}(x, y)}{n-2} = \frac{\text{var}y(1 - r^2)}{n-2}$$

Obszar ufności dla prostej regresji

$$f(x) \in (\hat{f}(x) - t(\alpha; n-2) S_Y; \hat{f}(x) + t(\alpha; n-2) S_Y)$$

gdzie

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$S^2_Y = S^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\text{var}x} \right)$$

Obszar predykcji

$$Y(x) \in (\hat{f}(x) - t(\alpha; n-2) S_{Y(x)}; \hat{f}(x) + t(\alpha; n-2) S_{Y(x)})$$

gdzie $Y(x)$ oznacza wartość zmiennej Y dla wybranej wartości x zmiennej X oraz

$$S^2_{Y(x)} = S^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\text{var}x} \right)$$

- Test chi-kwadrat na zgodność:

$$\chi^2_{emp} = \sum_i \frac{(n_i - n_i^t)^2}{n_i^t}$$

$$n_i^t = N \cdot \mathrm{P}\{X \in (i - \text{ta klasa})\},\ N = \sum_i n_i$$

$$\chi^2(\alpha, k-1-u),\ k - \text{liczba klas},\ u - \text{liczba nieznanych parametrów}$$

- Test chi-kwadrat na niezależność:

$$\chi^2_{emp} = \sum_i \sum_j \frac{(n_{ij} - n_{ij}^t)^2}{n_{ij}^t}$$

$$n_{ij}^t = n_{i\cdot} n_{\cdot j}/N,\ N = \sum_i \sum_j n_{ij}$$

$$\chi^2(\alpha, (m-1)(n-1));\ m, n - \text{liczby klas/poziomów}$$