Основные формулы

Число k определяется по формуле Стерджеса

$k = 1 + [3{,}322\ lgn\ ] = 1 + [log_2 n]$, где n – объем выборки.

Степенная средняя порядка p: $\bar{x}_p = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$

$$\bar{x}_p = \left(\frac{1}{n}\sum_{i=1}^{k} m_i x^p{}_{(i)}\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{k} w_i x^p{}_{(i)}\right)^{\frac{1}{p}}.$$

Порядок, p	*Название*	*Формула вычисления для несгруппированного ряда*	*Формула вычисления для сгруппированного ряда (ВР)*
-1	Средняя гармоническая	$\bar{x}_{-1} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$	$\bar{x}_{-1} = \frac{n}{\sum_{i=1}^{k}\frac{m_i}{x_{(i)}}}$
0	Средняя геометрическая	$\bar{x}_0 = \sqrt[n]{x_1 \ldots x_n}$	$\bar{x}_0 = \sqrt[n]{x_1^{m_1} \ldots x_k^{m_k}}$, $n = \sum_{i=1}^{k} m_i$
1	Средняя арифметическая	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i x_{(i)}$
2	Средняя квадратическая	$\bar{x}_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$	$\bar{x}_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{k} m_i x^2{}_{(i)}}$
3	Средняя кубическая	$\bar{x}_3 = \sqrt[3]{\frac{1}{n}\sum_{i=1}^{n} x_i^3}$	$\bar{x}_3 = \sqrt[3]{\frac{1}{n}\sum_{i=1}^{k} m_i x^3{}_{(i)}}$

Медиана

ДВР	*ИВР*
$Me = \begin{cases} x^*_{l+1}, n = 2l+1, \\ \frac{x^*_l + x^*_{l+1}}{2}, n = 2l. \end{cases}$	$Me = x_{Me} + h\frac{\frac{1}{2} - w^c_{Me-1}}{w_{Me}}$ $Me = x_{Me} + h\frac{\frac{n}{2} - m^c_{Me-1}}{m_{Me}}$, где x_{Me} – нижняя граница медианного интервала.

Свойство медианы: $\sum_{i=1}^{n} |x_i - Me| \le \sum_{i=1}^{n} |x_i - L|, \quad \forall L \in \mathrm{R}.$

Квантиль порядка p для ИВР: $Q_p = x_{Q_p} + h\frac{np - m^c_{Q_p-1}}{m_{Q_p}}$, $Q_p = x_{Q_p} + h\frac{p - w^c_{Q_p-1}}{w_{Q_p}}$,

x_{Qp} – нижняя граница квантильного интервала, порядка p;

Квартили: $q_1 = Q_{1/4} = x_{q_1} + h\dfrac{\frac{1}{4}n - m^c_{q_1 - 1}}{m_{q_1}}$, $q_2 = Me$, $q_3 = Q_{3/4} = x_{q_3} + h\dfrac{\frac{3}{4}n - m^c_{q_3 - 1}}{m_{q_3}}$.

Место квартили*:* $N_{q_1} = \dfrac{\sum_{i=1}^{k} m_i}{4}$, $N_{q_2} = \dfrac{2\sum_{i=1}^{k} m_i}{4}$, $N_{q_3} = \dfrac{3\sum_{i=1}^{k} m_i}{4}$.

Мода

ДВР	*ИВР*
Варианта с доминирующей частотой **Применяют соглашения:** ❖ Если все варианты имеют одинаковую частоту, то говорят, что вариационный ряд не имеет моды. ❖ Если две или более соседние варианты имеют одинаковую доминирующую частоту, то мода равна средней арифметической этих вариант. ❖ Если две (или более) несоседние варианты, имеют одинаковую доминирующую частоту, то говорят, что признак имеет две (или более) моды и называется *бимодальным* (*полимодальным*).	$Mo = x_{Mo} + h\dfrac{w_{Mo} - w_{Mo-1}}{2w_{Mo} - w_{Mo-1} - w_{Mo+1}}$, где x_{Mo} – нижняя граница модального интервала.

Предположим, что вся совокупность значений признака X разбита на j групп

$$\underbrace{x_1,\ldots,x_{n_1}}_{(1)},\ \underbrace{x_{n_1+1},\ldots,x_{n_1+n_2}}_{(2)},\ \ldots,\ \underbrace{x_{n_1+\ldots+n_{j-1}+1},\ldots,x_n}_{(j)}$$

и для каждой группы вычисляется групповая средняя $\bar{x}^{(l)}, l = \overline{1, j}$.

Тогда средняя общая $\bar{x} = \dfrac{1}{n}\sum_{l=1}^{j} n_l \bar{x}^{(l)}$.

Правило сложения дисперсий:

$$D_{\text{общ}} = \overline{D} + \bar{\delta}^2,$$

$$\overline{D} = \frac{D_1 n_1 + D_2 n_2 + \ldots + D_j n_j}{n}; \qquad \bar{\delta}^2 = \frac{n_1(\bar{x}^{(1)} - \bar{x})^2 + \ldots + n_j(\bar{x}^{(j)} - \bar{x})^2}{n}.$$

$\overline{D}$ – средняя из внутригрупповых дисперсий; $\bar{\delta}^2$ – межгрупповая дисперсия.

Коэффициент асимметрии: $A_s = \dfrac{\mu_{3,\bar{x}}}{\sigma^3} = r_3$, где $\mu_{3,\bar{x}} = \dfrac{1}{n}\sum_{i=1}^{k} m_i (x_{(i)} - \bar{x})^3$.

Если $A_s > 0$, то распределение скошено вправо, т.е. преобладают положительные отклонения от математического ожидания;

Если $A_s < 0$, то распределение скошено влево, т.е. преобладают отрицательные отклонения от математического ожидания;

Если $A_s = 0$, то распределение симметрично относительно математического ожидания.

Принято считать асимметрию при значении:

$\lvert \boldsymbol{A_s} \rvert \le 0{,}25$	малой,
$0{,}25 < \lvert \boldsymbol{A_s} \rvert \le 0{,}5$	умеренной,
$0{,}5 < \lvert \boldsymbol{A_s} \rvert \le 1{,}5$	большой,
$1{,}5 < \lvert \boldsymbol{A_s} \rvert$	исключительно большой.

Иногда используют первый $A_s \approx \dfrac{\bar{x} - Mo}{\sigma}$ или второй $A_s \approx \dfrac{3(\bar{x} - Me)}{\sigma}$ *приближенные* коэффициенты асимметрии Пирсона.

Об остроте вершины кривой распределения судят по коэффициенту ***эксцесса***:

$$E_x = r_4 - 3 = \frac{\mu_{4,\bar{x}}}{\sigma^4} - 3\,.$$

Если $E_x > 0$, то распределение имеет острый пик (по сравнению с нормальным распределением);

Если $E_x < 0$ (минимальное значение $E_x = -2$), то распределение имеет плосковершинную форму (по сравнению с нормальным распределением, для которого $E_x = 0$).

При $-0{,}5 < E_x < 3$ считают, что распределение приближается к нормальному.

В общем случае:

Условные средние: $\bar{y}_{x_i} = \dfrac{1}{m_{x_i}} \sum\limits_{j=1}^{l} m_{ij} y_j$ и $\bar{x}_{y_j} = \dfrac{1}{m_{y_j}} \sum\limits_{i=1}^{k} m_{ij} x_i$.

Общее среднее $\bar{y} = \dfrac{1}{n} \sum\limits_{j=1}^{l} m_{y_j} y_j$; $\bar{x} = \dfrac{1}{n} \sum\limits_{i=1}^{k} m_{x_i} x_i$;

В частном случае :

Общее среднее: $\boldsymbol{\bar{y}} = \dfrac{1}{\boldsymbol{k}} \sum\limits_{\boldsymbol{j}=1}^{\boldsymbol{k}} \boldsymbol{y_j}$; $\boldsymbol{\bar{x}} = \dfrac{1}{\boldsymbol{k}} \sum\limits_{\boldsymbol{i}=1}^{\boldsymbol{k}} \boldsymbol{x_i}$;

дисперсия: $\sigma_{\boldsymbol{y}}^2 = \dfrac{1}{\boldsymbol{k}} \sum\limits_{\boldsymbol{j}=1}^{\boldsymbol{k}} (\boldsymbol{y_j} - \boldsymbol{\bar{y}})^2$; $\sigma_{\boldsymbol{x}}^2 = \dfrac{1}{\boldsymbol{k}} \sum\limits_{\boldsymbol{i}=1}^{\boldsymbol{k}} (\boldsymbol{x_i} - \boldsymbol{\bar{x}})^2$, *ковариация:* $\sigma_{xy} = \frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\,\bar{y}$.

Коэффициент корреляции Пирсона: $r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$.

Оценка существенности линейного коэффициента корреляции при *большом объеме* выборки (свыше 50) проводится с использованием следующего отношения для наблюдаемого значения критерия: $t_{расч} = \frac{r_{xy}}{\sigma_r}$, где $\sigma_r = \frac{1 - r_{xy}^2}{\sqrt{n-1}}$ – среднеквадратическая ошибка линейного коэффициента корреляции, n – объем выборки. При *недостаточно большом* объеме выборки величину среднеквадратической ошибки коэффициента корреляции определяют по формуле: $\sigma_r = \frac{\sqrt{1 - r_{xy}^2}}{\sqrt{n-2}}$.

Коэффициент Фехнера: $K_{\Phi} = \frac{\Sigma C - \Sigma H}{\Sigma C + \Sigma H}$.

Коэффициент Спирмена: $r_c = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$, $d_i = R_{x_i} - R_{y_i}$ – разность рангов.

В случае присутствия связанных рангов

$$r_c = \frac{\frac{1}{6}(n^3 - n) - \sum_{i=1}^{n} d_i^2 - T_x - T_y}{\sqrt{\{\frac{1}{6}(n^3 - n) - 2T_x\}\{\frac{1}{6}(n^3 - n) - 2T_y\}}},$$

$$T_x = \frac{1}{12}\sum_{i=1}^{p}(t_i^3(x) - t_i(x)), \qquad T_x = \frac{1}{12}\sum_{j=1}^{q}(t_j^3(y) - t_j(y)),$$

p – количество связанных рангов признака X;
$t_i(x)$ – количество значений признака X в i-ом связанном ранге;
q – количество связанных рангов признака Y;
$t_j(y)$ – количество значений признака Y в j-ом связанном ранге.

Коэффициент ассоциации $K_a = \frac{m_{11}m_{22} - m_{12}m_{21}}{\sqrt{m_{x_0}m_{x_1}m_{y_0}m_{y_1}}}$;

Коэффициент контингенции $K_к = \frac{m_{11}m_{22} - m_{12}m_{21}}{m_{11}m_{22} + m_{11}m_{22}}$.

Всегда $|K_a| \le |K_к|$. Статистическая зависимость считается существенной, если $|K_к| \ge 0{,}5$, $|K_a| \ge 0{,}3$.

Коэффициент корреляции Пирсона $\varphi = \dfrac{p_{xy} - p_x p_y}{\sqrt{p_x(1-p_x)p_y(1-p_y)}}$,

где p_x – доля выборочных значений признака x, равных 1;

p_y – доля выборочных значений признака y, равных 1;

p_{xy} – доля вариант (x_i, y_i) с единичными значениями у обоих признаков.

Каждый из двух признаков состоит более чем из двух групп:

Коэффициент взаимной сопряженности Пирсона $K_П = \sqrt{\dfrac{\varphi^2}{1+\varphi^2}}$,

$\varphi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\dfrac{m_{ij}^2}{m_{x_i} m_{y_j}} - 1$, $\qquad$ k,l – число групп по двум признакам.

Коэффициент взаимной сопряженности Чупрова $K_ч = \sqrt{\dfrac{\varphi^2}{\sqrt{(k-1)(l-1)}}}$.

Коэффициенты линейной модели регрессии: $a = \dfrac{\sigma_{xy}}{\sigma_x^2} = r_{xy}\dfrac{\sigma_y}{\sigma_x}$, $\; b = \bar{y} - r_{xy}\dfrac{\sigma_y}{\sigma_x}\bar{x}$.

Дисперсия отклонения ε: $\qquad \sigma_\varepsilon^2 = \sigma_y^2(1 - r_{xy}^2)$.

Если $r_{xy} = \pm 1$ (т.е. $r^2_{xy} = 1$), то $\sigma_\varepsilon^2 = 0$; модель является адекватной.

Если $r_{xy} = 0$, тогда σ_ε^2 совпадает с σ_y^2, т.е. модель не является адекватной.

Коэффициент ***детерминации****:* $\quad \eta_{y(x)}^2 = \dfrac{\overline{\delta^2}}{\sigma_y^2} = \dfrac{\sigma_y^2 - \sigma_{y/x}^2}{\sigma_y^2} = 1 - \dfrac{\sigma_{y/x}^2}{\sigma_y^2}$.

Чем ближе $\dfrac{\sigma_{y/x}^2}{\sigma_y^2}$ к нулю, тем модель адекватнее.

$$\eta_{y(x)} = \sqrt{1 - \frac{\sigma_{y/x}^2}{\sigma_y^2}},\ 0 \le \eta_{y(x)} \le 1 \text{ – \textit{корреляционное отношение}.}$$

В случае линейной зависимости: $\qquad \eta_{y(x)}^2 = r_{xy}^2$.

В остальных случаях $\eta_{y(x)}^2 \ge r_{xy}^2$.

Причем отклонение от линейности считается существенным, если $\eta_{y(x)}^2 - r_{xy}^2 \ge 0{,}1$. Если $\eta_{y(x)}^2 - r_{xy}^2 \le 0{,}1$, то несущественным.

$$\begin{cases} a_1\sigma_{11} + a_2\sigma_{12} ... + a_k\sigma_{1k} = \sigma_{1y} \\ ... \\ a_1\sigma_{k1} + a_2\sigma_{k2} ... + a_k\sigma_{kk} = \sigma_{ky} \end{cases} \qquad a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2 - ... - a_k\bar{x}_k = \bar{y} - \sum_{j=1}^{k} a_j\bar{x}_j$$

$$\sigma_{ij} = \sigma_{x_i x_j}, \quad \sigma_{ii} = \sigma^2{}_{x_i} = \sigma_i^2; \quad \sigma_{jy} = \sigma_{x_j y}.$$

$$a_1 = \frac{\sigma_2^2\sigma_{1y} - \sigma_{12}\sigma_{2y}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, \qquad a_2 = \frac{\sigma_1^2\sigma_{2y} - \sigma_{12}\sigma_{1y}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}.$$

Стандартизированные коэффициенты множественной регрессии $\beta_i = a_i \dfrac{\sigma_i}{\sigma_y}$.

$$\begin{cases} \beta_1 + \beta_2 r_{12} + ... + \beta_k r_{1k} = r_{1y} \\ ... \\ \beta_1 r_{k1} + \beta_2 r_{k2} + ... + \beta_k = r_{ky} \end{cases}$$

Корреляционная матрица $R = \begin{pmatrix} 1 & r_{12} & ... & r_{1k} \\ r_{21} & 1 & & r_{2k} \\ ... & & & ... \\ r_{k1} & r_{k2} & ... & 1 \end{pmatrix}$

Коэффициент множественной детерминации: $r_y^2 = \dfrac{\bar{\delta}^2}{\sigma_y^2} = \sum_{i=1}^{k} \beta_i r_{iy} = -\Delta_{k+1}/\Delta_k$.

$$\Delta_k = \begin{vmatrix} 1 & r_{12}... & r_{1k-1} & r_{1k} \\ r_{21} & 1 & r_{2k-1} & r_{2k} \\ ... & & & ... \\ r_{k1} & r_{k2} & ... & 1 \end{vmatrix}, \qquad \Delta_{k+1} = \begin{vmatrix} 1 & r_{12}... & r_{1k} & r_{1y} \\ r_{21} & 1 & r_{2k} & r_{2y} \\ r_{k1} & r_{k2}... & 1 & r_{ky} \\ r_{1y} & r_{2y}... & r_{ky} & 0 \end{vmatrix}$$

Частный коэффициент детерминации:

$$r^2_{yx_j} \setminus x_1 ... x_{j-1} = \frac{r_y^2 - r^2_{y\setminus x_j}}{1 - r^2_{y\setminus x_j}}, \qquad r^2_{yx_m} \setminus x_1 ... x_{m-1} x_{m+1} ... x_k = \frac{r_y^2 - r^2_{y\setminus x_m}}{1 - r^2_{y\setminus x_m}}.$$

В случае k = 2: $r^2{}_y = \dfrac{r_{1y}^2 + r_{2y}^2 - 2r_{1y}r_{2y}r_{12}}{1 - r_{12}^2}$.

$$r^2_{yx_1} \setminus x_2 = \frac{r_y^2 - r^2_{y\setminus x_1}}{1 - r^2_{y\setminus x_1}} = /\ r^2_{y\setminus x_1} = r^2_{2y},\ r^2_{y\setminus x_2} = r^2_{1y}\ / = \frac{(r_{1y} - r_{12}r_{2y})^2}{(1 - r_{12}^2)(1 - r_{2y}^2)}.$$

$$r^2_{yx_2} \setminus x_1 = \frac{(r_{2y} - r_{12}r_{1y})^2}{(1 - r_{12}^2)(1 - r_{1y}^2)}.$$