Описательная статистика

Грауэр Л.В.

Описательная статистика

Цель

- обработка
- систематизация
- графическое представление
- расчет числовых статистических характеристик

эмпирических данных

Зачем нужна описательная статистика?

Выявить ошибки в данных

Увидеть структуру данных

Найти нарушения в статистических предположениях

Сгенерировать гипотезы

Порядковые статистики. Вариационный ряд

ξ, $X_{[n]} = (X_1, \ldots, X_n)$

Порядковые статистики:
$X_{(1)} = \min\{X_1, \ldots, X_n\}$ — первая порядковая статистика,
$X_{(2)} = \min\left\{\{X_1, \ldots, X_n\} \setminus X_{(1)}\right\}$ — вторая порядковая статистика,
$X_{(3)} = \min\left\{\{X_1, \ldots, X_n\} \setminus \left\{X_{(1)}, X_{(2)}\right\}\right\}$ — третья порядковая статистика,
...
$X_{(n)} = \max\{X_1, \ldots, X_n\}$ — n-ая порядковая статистика.

Вариационный ряд: $X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$.

Примеры

Рост баскетболистов
$X_{[10]}$=(205, 184, 207, 198, 195, 187, 201, 177, 191, 194)

Количество попаданий в мишень из 5 выстрелов
$X_{[10]}$= (5, 3, 5, 3, 4, 5, 4, 5, 3, 3)

Статистический ряд

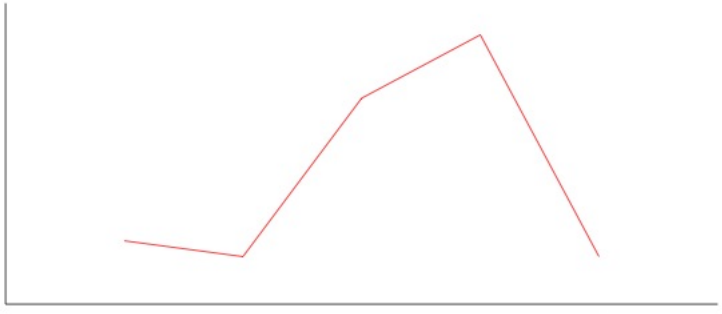
$(X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}) \Rightarrow (Z_{(1)} < Z_{(2)} < \ldots < Z_{(k)})$

x_i	$Z_{(1)}$	$Z_{(2)}$	$\ldots$	$Z_{(k)}$
n_i	n_1	n_2	$\ldots$	n_k
n_i/n	n_1/n	n_2/n	$\ldots$	n_k/n
$\sum_{j=1}^{i} n_j/n$	n_1/n	$\sum_{j=1}^{2} n_j/n$	$\ldots$	1

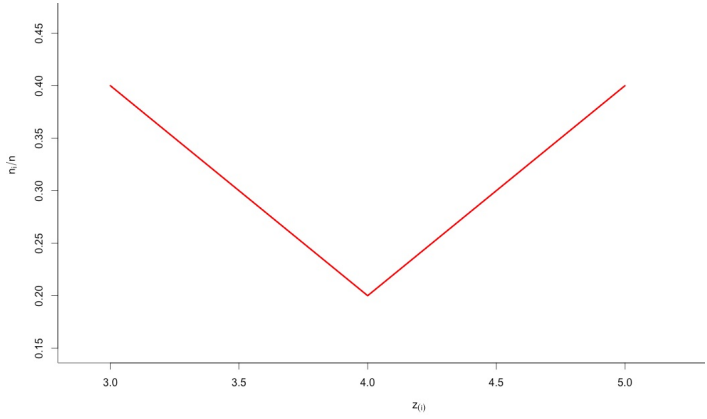
Пример

$X_{[10]}$= (5, 3, 5, 3, 4, 5, 4, 5, 3, 3)

Полигон частот



$X_{[10]}$= (5, 3, 5, 3, 4, 5, 4, 5, 3, 3)



Группированный статистический ряд. Гистограмма

Интервал (a, b), где $a \leq X_{(1)}$ и $X_{(n)} \leq b$ разобьем

$$a_0 = a < a_1 < a_2 < \ldots < a_r = b,$$

$$(a_{i-1}, a_i], i = 1, \ldots, r.$$

n_i — количество элементов выборки, попавших в $(a_{i-1}, a_i]$.

$$n_1 + n_2 + \ldots + n_r = n,$$

$$\Delta_i = a_i - a_{i-1},$$

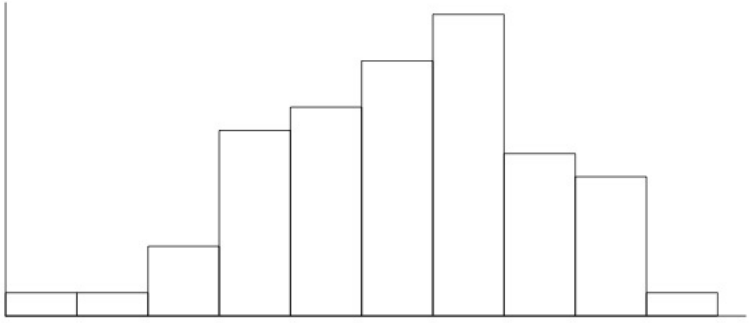
$$h_i = \frac{n_i}{\Delta_i n}.$$

Группированный статистический ряд

x_i	$[a_0, a_1]$	$(a_1, a_2]$	$\ldots$	$(a_{r-1}, a_r]$
n_i	n_1	n_2	$\ldots$	n_r
n_i/n	n_1/n	n_2/n	$\ldots$	n_r/n

Гистограмма

$$f_n^*(x) = \begin{cases} 0, & \text{если } x \leqslant a_0; \\ h_1, & \text{если } a_0 < x \leqslant a_1; \\ \ldots & \\ h_r, & \text{если } a_{r-1} < x \leqslant a_r; \\ 0, & \text{если } x > a_r. \end{cases}$$

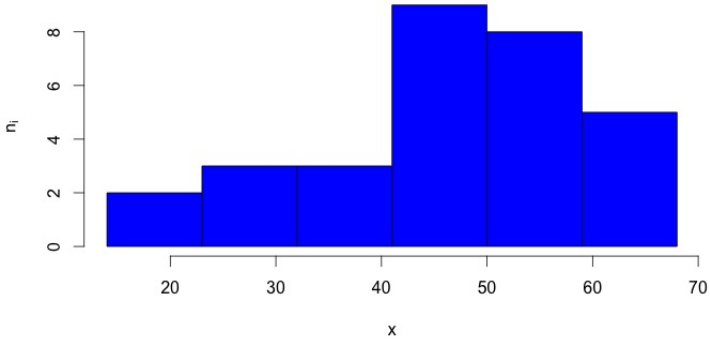


Пример

$X_{[n]}$:

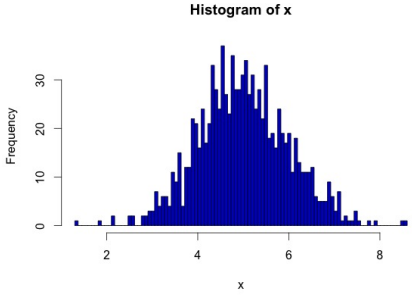
38	60	41	51	33	42
45	21	53	60	68	52
47	46	49	49	14	57
54	59	67	47	28	48
58	32	42	58	61	30

x_i	$[14, 23]$	$(23, 32]$	$(32,41]$	$(41, 50]$	$(50,59]$	$(59,68]$
n_i						
$\frac{n_i}{n}$						

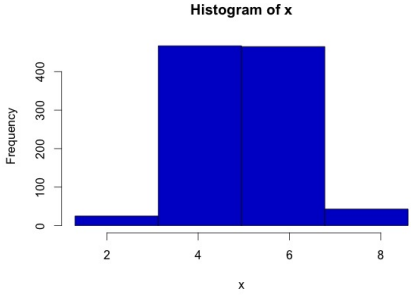

n_i
0
2
4
6
8
20
30
40
50
60
70
x

Как выбрать K?

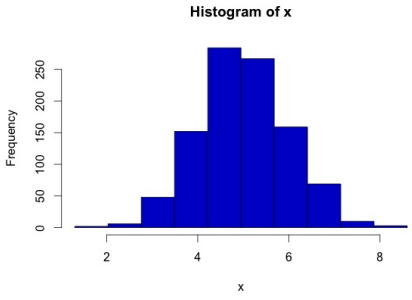
$X_{[1000]} \propto N(5,1)$

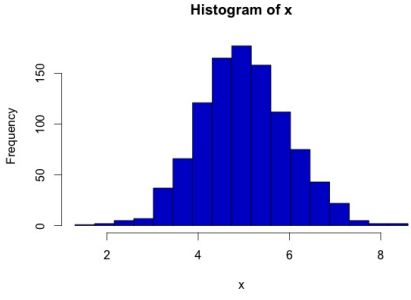


$r = 100$



r=4





$r = [1 + 3.2 \lg n]$

$r = \left[1.72 n^{1/3}\right]$

Выборочные числовые характеристики

Выборочное среднее

$$\bar{X} = a_1^* = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Выборочный начальный момент r-го порядка

$$a_r^* = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

Выборочная дисперсия

$$D^* = D^* X_{[n]} = \frac{1}{n} \sum_{i=1}^{k} \left(X_i - \bar{X}\right)^2$$

Выборочный центральный момент r-го порядка

$$\mu_r^* = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^r$$

Выборочная квантиль x_p порядка p —
$([np] + 1)$ элемент $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$.

Квартили $Q_1,\ Q_2,\ Q_3$ — квантили порядков 0.25, 0.5, 0.75

Выборочная медиана

$$x^*_{med} = \begin{cases} X_{(k+1)}, & n = 2k+1 \\ \dfrac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k \end{cases}$$

Пример

$X_{[10]} = (5, 3, 5, 3, 4, 5, 4, 5, 3, 3)$

Выборочные характеристиками положения

- *выборочное среднее*
- *выборочная медиана*
- *выборочная мода*

Выборочные меры рассеяния

- *размах* $R = X_{\max} - X_{\min}$
- *средний межквартильный размах*
- *персентильный размах* $P_{90} - P_{10}$,
- *выборочная дисперсия*
- исправленная дисперсия $\tilde{s}^2 = nD^*X_{[n]}/(n-1)$
- *среднее квадратическое отклонение* $s = \sqrt{s^2}$

Коэффициент вариации $v = s/\bar{X}$

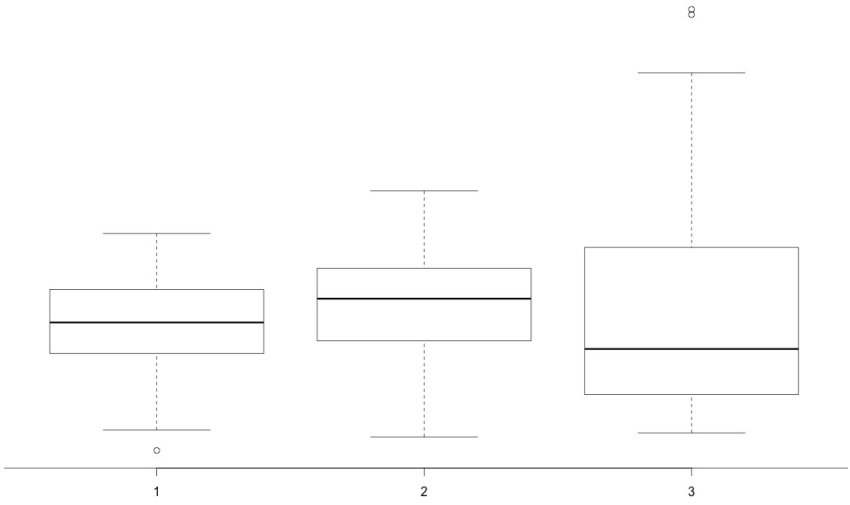
Оценка формы распределения

- *коэффициент асимметрии* $S_{k1} = \mu_3^*/s^3$
- *коэффициент эксцесса* $K = \mu_4^*/s^4 - 3$

Квантильный коэффициент асимметрии

$$S_{k2} = (Q_3 - Q_1 - 2Q_2)/(Q_3 - Q_1)$$

Ящики с усами



Выборочные характеристики многомерных выборок

$(\xi, \eta)^T$

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

Выборочный коэффициент корреляции

$$r_{\xi,\eta} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}}{\tilde{s}_X \tilde{s}_Y}$$

Диаграммы рассеивания