

ПРИЛОЖЕНИЕ 2

*Измеряй микрометром. Отмечай мелом. Отрубай топором.
Мерфология, Правило точности Рея*

Предварительная обработка данных (отсев грубых погрешностей)

Предварительная обработка данных (результатов измерений или наблюдений) необходима для того, чтобы в дальнейшем корректно и с наибольшей эффективностью использовать для построения эмпирических зависимостей статистические методы.

Содержание предварительной обработки в основном состоит в отсеивании грубых погрешностей измерения или погрешностей, неизбежно имеющих место при подготовке цифрового материала исследования.

Различных рекомендаций для проведения отсева грубых погрешностей наблюдения (аномальных значений) имеется достаточно много. Для практического использования ниже предлагаются наиболее простые методы их отсева.



*Что для одного ошибка, для другого – исходные данные.
Мерфология, следствие Бермана из аксиомы Робертса*

П2.1. Правило "ящик с усами"

Правило "ящик с усами" получило название от типа соответствующего графика, используемого для наглядного представления разброса эмпирических данных с нанесенными значениями медианы и квартилей.

Порядок вычисления применения правила следующий:

- а) Определяются выборочные значения межквартильного размаха R_μ и медианы μ .
- б) Выборочные значения, меньшие $(\mu - 1,5R_\mu)$ и большие $(\mu + 1,5R_\mu)$, называются мягкими (подозрительными) выбросами.
- в) Выборочные значения, меньшие $(\mu - 3R_\mu)$ и большие $(\mu + 3R_\mu)$, называются экстремальными выбросами и должны быть исключены.

Вычисление межквартильного размаха R_μ упорядоченной по возрастанию количественной выборки производится по формуле:

$$R_\mu = \mu_{3/4} - \mu_{1/4},$$

где $\mu_{1/4}$, $\mu_{3/4}$ – значение нижней и верхней квартилей выборки.

Критерий удобен для автоматической идентификации любого числа экстремально малых и больших значений выборки. Критерий очень популярен, однако его рекомендуется применять только в случае, если численность выборки достаточно велика.

Пример П2.1. В качестве примера рассмотрим отсев грубых погрешностей данных, занесенных в диапазон A63:H68 (рис. П2.1).

1. В ячейки J63: J67 заносятся поясняющие данные.
2. Заполняются формулами следующие ячейки.

адрес	формула	пояснение
K63	=КВАРТИЛЬ(A63:H68;3)- КВАРТИЛЬ(A63:H68;1) {=QUARTILE(A63:H68;3)- QUARTILE(A63:H68;1)}	межквартильный размах R_{μ}
K64	=МЕДИАНА(A63:H68)-1,5*K\$63 {=MEDIAN(A63:H68)-1,5*K\$63}	нижний предел "подозрительных" данных
K65	=МЕДИАНА(A63:H68)+1,5*K\$63 {=MEDIAN(A63:H68)+1,5*K\$63}	верхний предел "подозрительных" данных
K66	=МЕДИАНА(A63:H68)-3*K\$63 {=MEDIAN(A63:H68)-3*K\$63}	нижний предел "экстремальных" данных
K67	=МЕДИАНА(A63:H68)+3*K\$63 {=MEDIAN(A63:H68)+3*K\$63}	верхний предел "экстремальных" данных

3. В ячейку A77 заносится формула =ЕСЛИ (ИЛИ ((A63 < K\$64); (A63 > K\$65)); ""; A63)
{=IF(OR((A63<K\$64);(A63>K\$65));""; A63)}, которая оставляет ячейку пустой, если значение исходной ячейки данных A63 меньше нижнего или больше верхнего предела "подозрительных" данных, и значение исходной ячейки A63 в противоположном случае (т.е. если данное значение не является подозрительным). Механизмом автозаполнения по A77 вводятся значения для оставшихся ячеек диапазона A77:H82.

	A	B	C	D	E	F	G	H	I	J	K	L	M
59	Предварительная обработка данных (отсев грубых погрешностей)												
60	Правило "ящик с усами"												
61													
62	ЯРУС 5												
63	11,8	11,5	16,8	16,1	16,8	16,8	16,3	16,3		$R\mu = 2,9$			
64	12,6	12,0	12,4	5,6	18,9	18,5	18,6	16,6		$\mu - 1,5R\mu = 11,9$			
65	13,1	13,9	13,0	17,7	17,7	17,1	17,1	16,7		$\mu + 1,5R\mu = 20,4$			
66	14,8	14,4	14,3	14,6	14,5	14,4	14,4	16,3		$\mu - 3R\mu = 7,6$			
67	15,1	15,3	15,5	15,5	15,4	15,4	15,4	16,2		$\mu + 3R\mu = 24,7$			
68	19,8	19,1	19,8	17,2	17,4	17,7	20,8	20,6					
69													
70										=МЕДИАНА(A63:H68)-1,5*K\$63			
71										=МЕДИАНА(A63:H68)+3*K\$63			
72										КВАРТИЛЬ(A63:H68;3)-КВАРТИЛЬ(A63:H68;1)			
73													
74										=ЕСЛИ(ИЛИ((A63<\$K\$64);(A63>\$K\$65)); "";			
75													
76										выборка с удаленными "подозрительными" значениями			
77			16,8	16,1	16,8	16,8	16,3	16,3					
78	12,6	12,0	12,4		18,9	18,5	18,6	16,6					
79	13,1	13,9	13,0	17,7	17,7	17,1	17,1	16,7					
80	14,8	14,4	14,3	14,6	14,5	14,4	14,4	16,3					
81	15,1	15,3	15,5	15,5	15,4	15,4	15,4	16,2					
82	19,8	19,1	19,8	17,2	17,4	17,7							
83													
84										=ЕСЛИ(ИЛИ((A63<\$K\$66);(A63>\$K\$67)); "", A63)			
85										выборка с удаленными "экстремальными" данными			
86	11,8	11,5	16,8	16,1	16,8	16,8	16,3	16,3					
87	12,6	12,0	12,4		18,9	18,5	18,6	16,6					
88	13,1	13,9	13,0	17,7	17,7	17,1	17,1	16,7					
89	14,8	14,4	14,3	14,6	14,5	14,4	14,4	16,3					
90	15,1	15,3	15,5	15,5	15,4	15,4	15,4	16,2					
91	19,8	19,1	19,8	17,2	17,4	17,7	20,8	20,6					

Рис. П2.1. Скриншот расчетного листа MS Excel (правило "ящик с усами")

4. В ячейку A86 заносится формула =ЕСЛИ (ИЛИ ((A63<\$K\$66); (A63 > \$K\$67));""; A86) {=IF(OR((A63<\$K\$66);(A63>\$K\$67));""; A86)}, которая оставляет ячейку пустой, если значение исходной ячейки данных A63 меньше нижнего или больше верхнего предела "экстремальных" данных, и значение исходной ячейки A63 в противоположном случае (т.е. если данное значение не является экстремальным). Механизмом автозаполнения по A86 вводятся значения для оставшихся ячеек диапазона A86:H91.
5. Экстремальные значения подлежат безусловной отбраковке, действия по подозрительным – на усмотрение исследователя. В зависимости от этого используется тот или иной из полученных "очищенных" массивов данных.

П2.2. Правило Томпсона (критерий Рошера)

В правиле Томпсона (критерий Рошера) для исключения выбросов используется статистика

$$t = \frac{|x - \bar{x}|}{\sigma}, \quad \begin{array}{l} \text{где } x_i - \text{результаты наблюдений,} \\ \bar{x} - \text{выборочное среднее значение,} \\ \sigma - \text{выборочное стандартное отклонение, корень} \\ \text{квадратный из выборочной дисперсии.} \end{array}$$

Величина статистики критерия сравнивается с критическим значением

$$T = \sqrt{\frac{(n-1)t_{\left(1-\frac{\alpha}{2}\right), (n-2)}^2}{n-2 + t_{\left(1-\frac{\alpha}{2}\right), (n-2)}^2}} \quad \begin{array}{l} \text{где } t_{\left(1-\frac{\alpha}{2}\right), (n-2)} - \text{значение обратной функции } t\text{-распре-} \\ \text{деления с параметрами } \left(1 - \frac{\alpha}{2}\right) \text{ и } (n-2), \\ \alpha - \text{заданный уровень значимости.} \end{array}$$

При величине статистики, большей критического значения, наблюдение исключается. Процедура повторяется для каждого наблюдения.

Пример П2.2. В качестве примера рассматривается алгоритм отсева грубых погрешностей данных, занесенных в диапазон A68:H73 (рис. П2.2).

1. В ячейки J68: J73 заносятся поясняющие данные.
2. По адресу K68 вводится значение уровня значимости.
3. В ячейке K69 формулой =СЧЁТ(A68:H73)-1 {=COUNT(A68:H73)-1} рассчитывается количество степеней свободы.
4. Заполняются формулами следующие ячейки.

адрес	формула	пояснение
K70	=СРЗНАЧ(A68:H73) {=AVERAGE(A68:H73)}	расчет среднеарифметического выборки $\bar{x}$
K71	=СТАНДОТКЛОН(A68:H73) {=STDEV(A68:H73)}	расчет стандартного отклонения σ
K72	=СТЮДРАСПОБР(2*\$K\$68; K69-1) {=TINV(2*\$K\$68;K69-1)}	расчет $t_{(1-\frac{\alpha}{2}), (n-2)}$ -обратного значения t – распределения

	A	B	C	D	E	F	G	H	I	J	K	L
64	Предварительная обработка данных (отсев грубых погрешностей)											
65	Правило Томпсона (критерий Рошера)											
66												
67	ЯРУС 5											
68	11,8	11,5	16,8	16,1	16,8	16,8	16,3	16,3		$\alpha =$	0,05	
69	12,6	12,0	12,4	5,6	18,9	18,5	18,6	16,6		$df =$	47,0	
70	13,1	13,9	13,0	17,7	17,7	17,1	17,1	16,7		$\bar{x} =$	15,8	
71	14,8	14,4	14,3	14,6	14,5	14,4	14,4	16,3		$\sigma =$	2,72	
72	15,1	15,3	15,5	15,5	15,4	15,4	15,4	16,2		$t =$	1,68	
73	19,8	19,1	19,8	17,2	17,4	17,7	20,8	20,6		$T =$	1,65	
74											=СЧЁТ(A68:H73)-1	
75	=ЕСЛИ(ABS(A68-\$K\$70)/\$K\$71>\$K\$73;"";A68)										=СРЗНАЧ(A68:H73)	
76											=СТАНДОТКЛОН(A68:H73)	
77											=СТЮДРАСПОБР(2*\$K\$68;K69-1)	
78											=K72*КОРЕНЬ(K69/(K69-1+K72*K72))	
79												
80												
81												
82	11,8	11,5	16,8	16,1	16,8	16,8	16,3	16,3				
83	12,6	12,0	12,4		18,9	18,5	18,6	16,6				
84	13,1	13,9	13,0	17,7	17,7	17,1	17,1	16,7				
85	14,8	14,4	14,3	14,6	14,5	14,4	14,4	16,3				
86	15,1	15,3	15,5	15,5	15,4	15,4	15,4	16,2				
87	19,8	19,1	19,8	17,2	17,4	17,7						

Рис. П2.2. Скриншот расчетного листа MS Excel (правило Томпсона)

5. В ячейке K73 формулой $=K72 \cdot \sqrt{K69 / (K69 - 1 + K72 \cdot K72)}$ $\{=K72 \cdot \text{SQRT}(K69 / (K69 - 1 + K72 \cdot K72))\}$ рассчитывается значение критерия Томпсона.
6. В ячейку A82 заносится формула $=\text{ЕСЛИ}(\text{ABS}(A68 - \$K\$70) / \$K\$71 > \$K\$73; ""; A68)$ $\{=IF(\text{ABS}(A68 - \$K\$70) / \$K\$71 > \$K\$73; "", A68)\}$, которая оставляет ячейку пустой, если значение статистики t для ячейки исходных данных A68 больше "критического", верхнего предела T , и значение ячейки исходных данных A68 в противоположном случае (т.е. если данное значение не является экстремальным). Механизмом автозаполнения по A82 вводятся значения для оставшихся ячеек диапазона A82:H87.
7. Для дальнейших исследований используется полученный "очищенный" массив данных.

