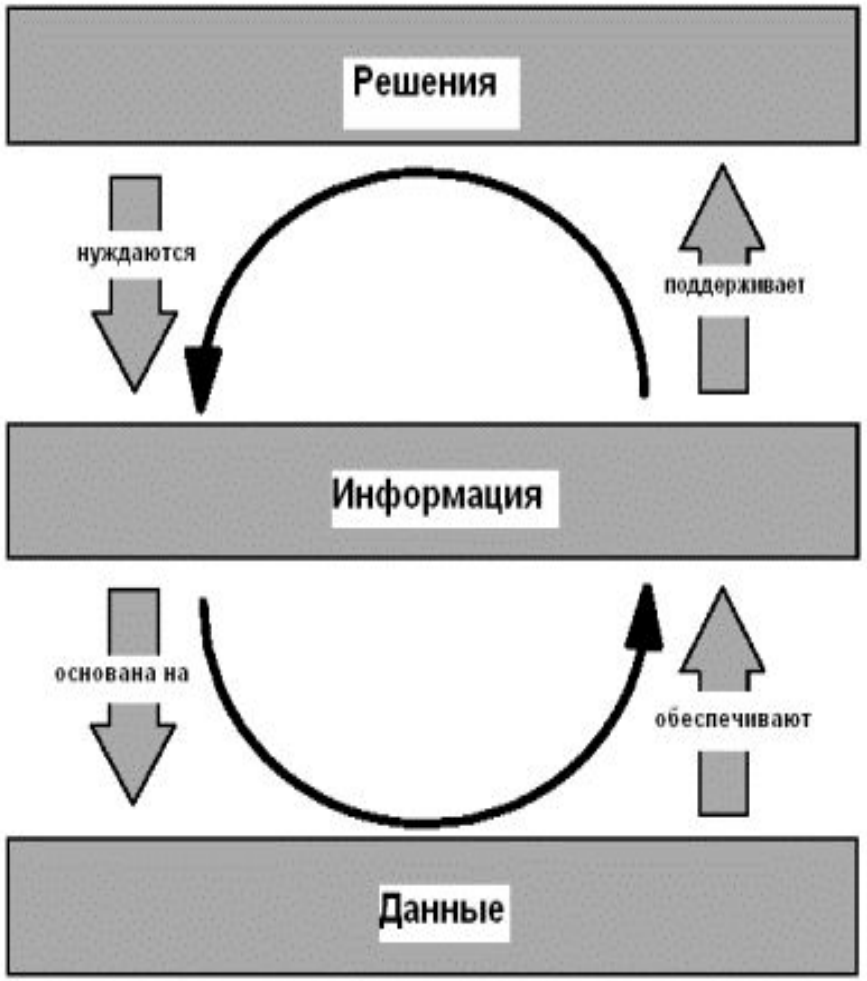
Технологии обработки информации

Лекция 1
Описательная статистика: основные понятия

Преподаватель: Тазиева Рамиля Фаридовна

Информационная пирамида



Методы анализа данных

Статистические:

- ✔ Дескриптивный анализ.
- ✔ Анализ природы данных (проверка гипотез стационарности, нормальности, однородности, оценка вида функции распределения).
- ✔ Анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ).
- ✔ Многомерный статистический анализ .

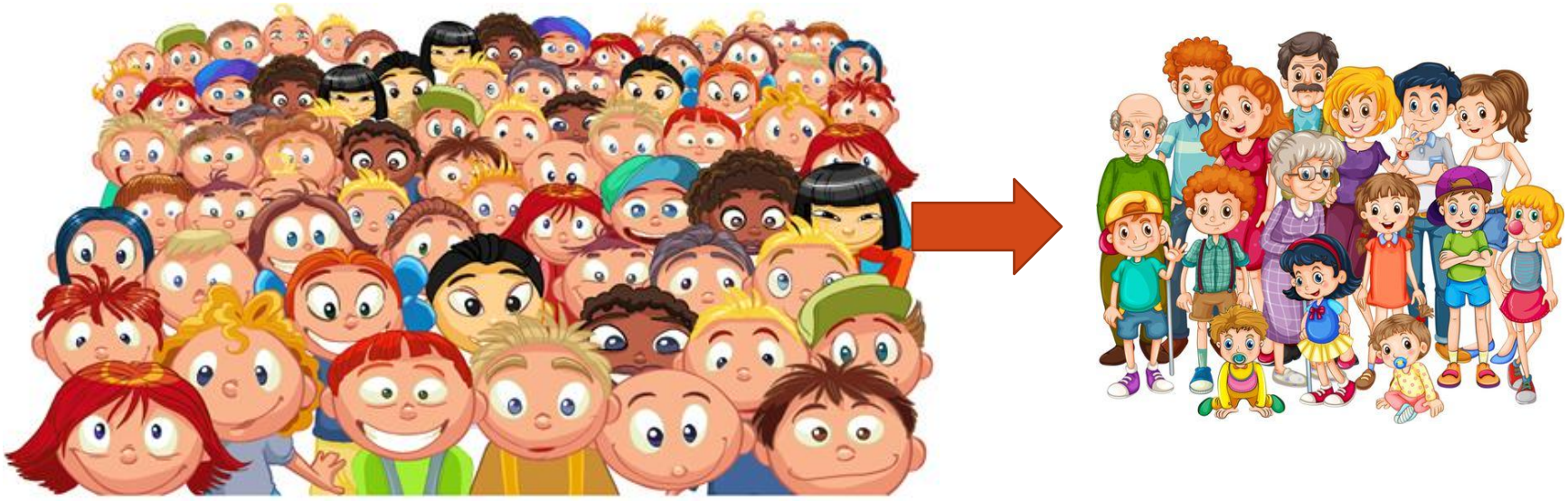
Кибернетические:

- ✔ Методы классификации.
- ✔ Кластерный анализ.
- ✔ Искусственные нейронные сети (распознавание, прогноз).
- ✔ Деревья решений.
- ✔ Методы ближайшего соседа и k-ближайшего соседа
- ✔ Системы обработки экспертных знаний.

Генеральная и выборочная совокупности

Генеральная совокупность - вся совокупность изучаемых объектов, интересующая исследователя.

Выборка - часть генеральной совокупности, определенным способом отобранная с целью исследования и получения выводов о свойствах и характеристиках генеральной совокупности.



Виды шкал

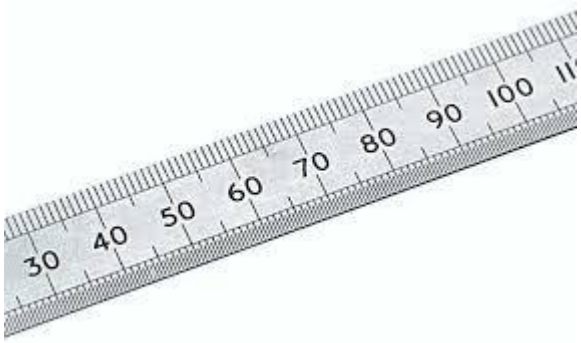
1. Номинальная шкала

2. Порядковая шкала

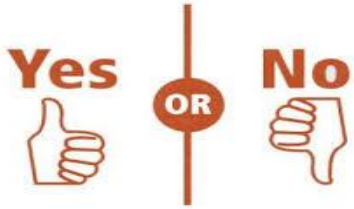
3. Интервальная шкала

Значения показателя ПВ	Условия труда на рабочем месте
0	Допустимые
1 – 2	Вредные
3 – 6	Очень вредные
7 – 14	Неприемлемо вредные
15 – 30	Опасные
более 30	Высокоопасные

4. Относительная шкала



5. Дихотомическая шкала



Основные понятия

Случайной величиной X называется величина, которая в результате опыта (или испытания) принимает какое-либо значение

Пусть в результате независимых испытаний, проведенных в одинаковых условиях, получены числовые значения признака X$\{x^{(1)},x^{(2)}, \ldots,x^{(n)}\}$, где n—объем выборки.

Статистикой называется некоторая произвольная функция от выборки.

Вариационным рядом (статистическим распределением) называется ранжированный в порядке возрастания (или убывания) ряд вариант:x_1, x_2, . . .,x_n($x_1 \le x_2 \le \ldots \le x_n$).

Варианты,x_i	x_1	x_2	. . .	x_k
Частоты,n_i	n_1	n_2	. . .	n_k

Выбросы. Квартили

Выброс — это **нетипичное наблюдение**, то есть такое наблюдение, которое существенно отличается от остальных в выборке.

Что делать с выбросами? Их можно удалить перед подсчетом описательных статистик и отдельно упомянуть в отчёте, что такие наблюдения были.

Квартили — это такие значения, которые делят вариационный ряд на четыре равные части (по 25 % в каждой).

- Q1 Нижний квартиль отделяет 25 % наблюдений с наименьшими значениями от остальных 75 %.
- Q2 Второй квартиль — это медиана (делит вариационный ряд попалам).
- Q3 Верхний квартиль отделяет 25 % наблюдений с наибольшими значениями от остальных 75 %.

Межквартильный размах (IQR) — это **разность между третьим и первым квартилем**.

В этом нам помогут квартили и межквартильный размах. Выбросом считается значение в следующих случаях:

- Если наблюдение меньше, чем значение нижнего квартиля минус межквартильного размаха. Q1-1.5 IQR
- Если наблюдение больше, чем значение верхнего квартиля плюс межквартильного размаха. Q3+1.5 IQR

Пример определения выбросов

Исходные данные	Медиана (Q2)	Квартили (Q1, Q3)			Выбросы
50,5	50,5	50,5	Q1	58,6	50,5
52,6	52,6	52,6	Q2	77,4	52,6
57,5	57,5	57,5	Q3	80,5	57,5
59,7	59,7	59,7			59,7
73,8	73,8	73,8	IQR	21,9	73,8
77,4	77,4	77,4	Q1-1,5 IQR	25,75	77,4
77,9	77,9	77,9	Q3+1,5 IQR	113,35	77,9
80,2	80,2	80,2			80,2
80,8	80,8	80,8			80,8
98	98	98			98
120	120	120			120

Примечание: Для определения квартилей выборка должна быть обязательно упорядочена.

ряд

1. Вычисляют **размах *R* варьирования** признака X как разность между наибольшим x_{max} и наименьшим x_{min} значениями признака:

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

2. Размах R варьирования признака X делится на k равных частей. Число k выбирают, пользуясь одним из следующих правил:

$$k \approx \sqrt{n} \qquad k \approx 1 + \log_2 n \approx 1 + 3{,}221 \cdot \lg n$$

3. Длина h каждого частичного интервала определяется по формуле: $h=R/k$.

4. За начало x_0 первого интервала рекомендуется[6] брать величину $x_0=x_{min}-0{,}5h$.

5. Конец x_k последнего интервала находят по формуле $x_k=x_{max}+0{,}5h$.

Варианты-интервалы, $(x_{i-1}; x_i)$	$[x_0; x_1]$	$(x_1; x_2]$	. . .	$(x_{k-1}; x_k]$
Частоты, n_i	n_1	n_2	. . .	n_k

Пример

Исходные данные

.90	0.79	0.84	0.86	0.88	0.90	0.89	0.85	0.91	0.98	0.91	0.80	0.87
.89	0.88	0.78	0.81	0.85	0.88	0.94	0.86	0.80	0.86	0.91	0.78	0.86
.91	0.95	0.97	0.88	0.79	0.82	0.84	0.90	0.81	0.87	0.91	0.90	0.82
.85	0.90	0.82	0.85	0.90	0.96	0.98	0.89	0.87	0.99	0.85		

Интервальный вариационный ряд

$(x_{i-1}; x_i]$	[0.765-0.795)	[0.795-0.825)	[0.825-0.855)	[0.855-0.885)	[0.885-0.915)	[0.915-0.945)	[0.945-0.975)	[0.975-1.005)
n_i	4	7	7	11	14	1	3	3

Дискретный вариационный ряд

x_i	0.78	0.81	0.84	0.87	0.9	0.93	0.96	0.99
n_i	4	7	7	11	14	1	3	3

Построение интервального вариационного ряда

1. Рассчитаем размах варьирования:

$$R = x_{\max} - x_{\min} = 0.99 - 0.78 = 0.21$$

где x_{min} – наименьшая варианта данной выборочной совокупности;
x_{max} – наибольшая варианта данной выборочной совокупности.

2. Вычислим число равных частей, на которое нужно разделить размах варьирования:

$$k = \sqrt{n} = \sqrt{50} \approx 7$$

3. Вычислим шаг h (длину интервалов):

$$h = \frac{R}{k} = \frac{0.21}{7} = 0.03$$

4. Вычислим начальное значение первого интервала x_0 и конечное значение последнего интервала x_k:

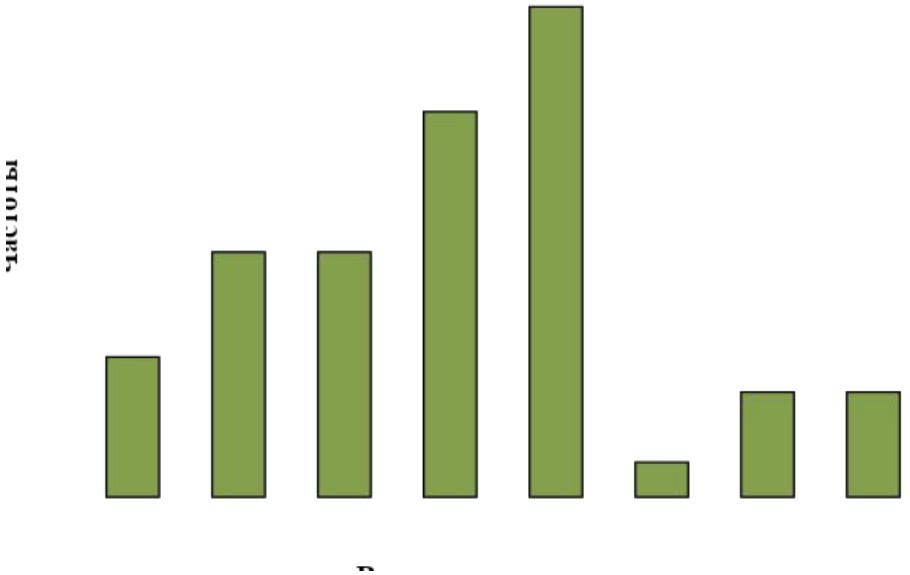
$$x_0 = x_{\min} - 0.5 \cdot h = 0.78 - 0.5 \cdot 0.03 = 0.765$$

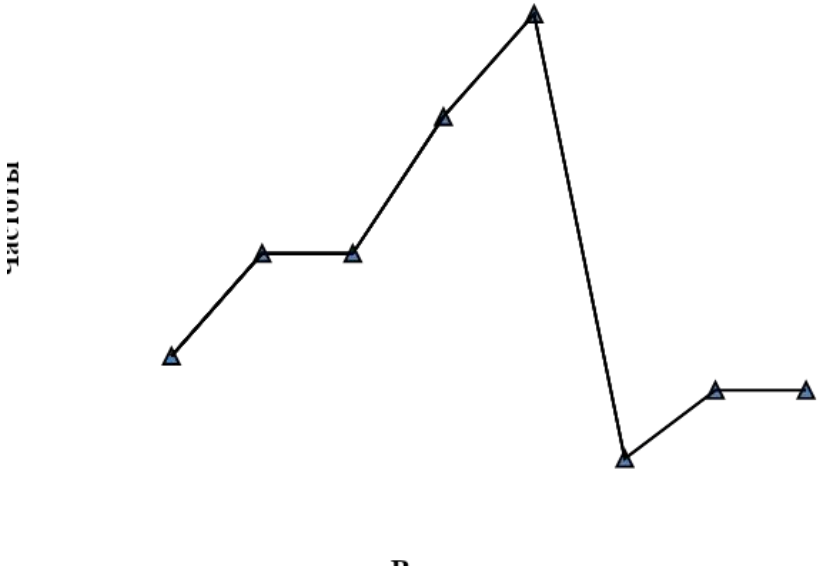
$$x_k = x_{\max} + 0.5 \cdot h = 0.99 + 0.5 \cdot 0.03 = 1.005$$

Гистограмма и полигон

Гистограммой называется ступенчатая фигура, для построения которой по оси абсцисс откладывают отрезки, изображающие частичные интервалы $(x_{i-1};x_i)$ варьирования признака X, и на этих отрезках, как на основаниях, строят прямоугольники с высотами, равными частотам соответствующих интервалов.

Полигоном называется ломанная соединяющая точки с координатами $(x_i;n_i)$.



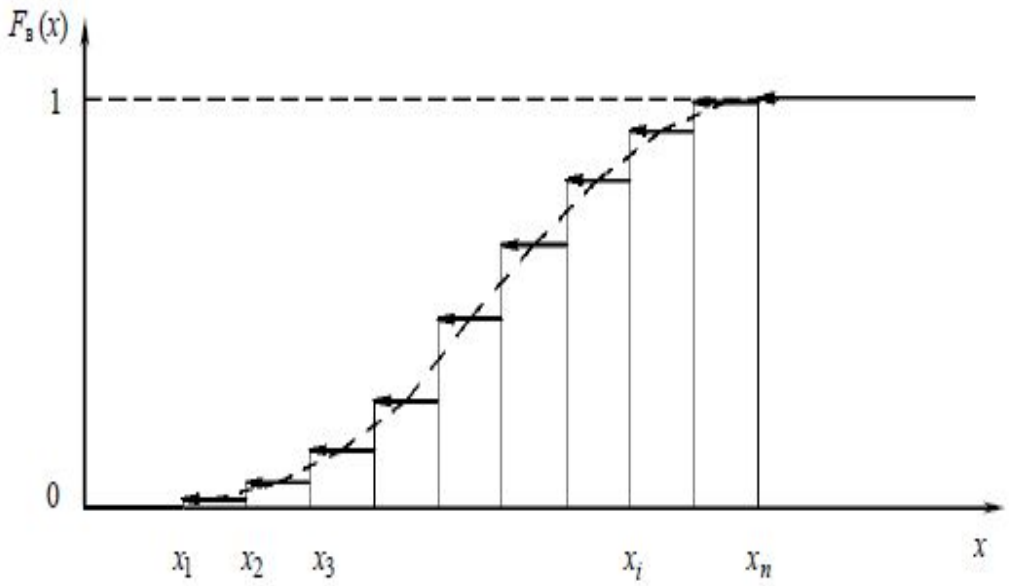


Кумулятивная кривая

Варианты x_i	x_1	x_2	…	x_k
Относительные частоты $w_i = n_i / n$	$w_1 = n_1 / n$	$w_2 = n_2 / n$	…	$w_k = n_k / n$
Накопленные относительные частоты $W_i = W_{i-1} + w_i$	$W_1 = w_1$ ($W_0 = 0$)	$W_2 = W_1 + w_2$	…	$W_k = W_{k-1} + w_k$

Эмпирическая функция распределения

$$F_n^*(x) = \frac{1}{n} \sum_{x_i < x} n_i ,$$



Центральная тенденция

Выборочная средняя

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i, \quad n = \sum_{i=1}^{k} n_i,$$

$$\bar{x} = 0.8724.$$

Свойства среднего

- При расчете *среднего* не допускаются пропущенные значения данных.
- Информативность *среднего* значения переменной высока, если известен ее доверительный интервал.
- С увеличением размера выборки точность оценки *среднего* возрастает.
- С увеличением разброса значений выборки надежность *среднего* падает.

Примечание: При анализе данных средним не следует злоупотреблять, необходимо учитывать его свойства и ограничения. Известны характеристики "средняя температура по больнице" или "средняя высота дома", показывающие некорректность использования этой меры *центральной тенденции* для некоторых случаев.

$S_{M_{e-1}}$

Центральная тенденция

Медианой M_e называют варианту которая делит вариационный ряд на две равные по числу вариант части.

При нечетном объеме выборки $n=2k+1$ $$M_e = x_k$$

При четном объеме выборки $n=2k$ $$M_e \overset{+1}{=} \frac{x_k + x_{k+1}}{2},$$

x_i	0.78	0.81	0.84	0.87	0.9	0.93	0.96	0.99
n_i	4	7	7	11	14	1	3	3

$$M_e = \frac{x_{25} + x_{26}}{2} = 0.87.$$

Примечание: Для определения медианы выборка должна быть обязательно упорядочена.

Центральная тенденция

Модой M_o называют варианту, которая имеет наибольшую частоту.

x_i	0.78	0.81	0.84	0.87	0.9	0.93	0.96	0.99
n_i	4	7	7	11	14	1	3	3

$$M_0 = 0.9.$$

Коэффициент вариации - меры разброса признака относительно его среднего значения.

$$V = \frac{S}{\overline{x}} * 100\%. \qquad V = \frac{S}{\overline{x}} = \frac{0.054279}{0.8724} = 0.062218.$$

Примечание: Если коэффициент вариации превышает 33%, то это говорит о неоднородности информации и необходимости исключения самых больших и самых маленьких значений.

Дисперсия

Формула расчета дисперсии для несгруппированных данных

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

Формула расчета дисперсии для сгруппированных данных

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \frac{1}{n}\sum_{i=1}^{k}x_i^2 n_i - \bar{x}^2 .$$

Если объем выборки $n>50$, то рассчитывают **исправленную дисперсию:**

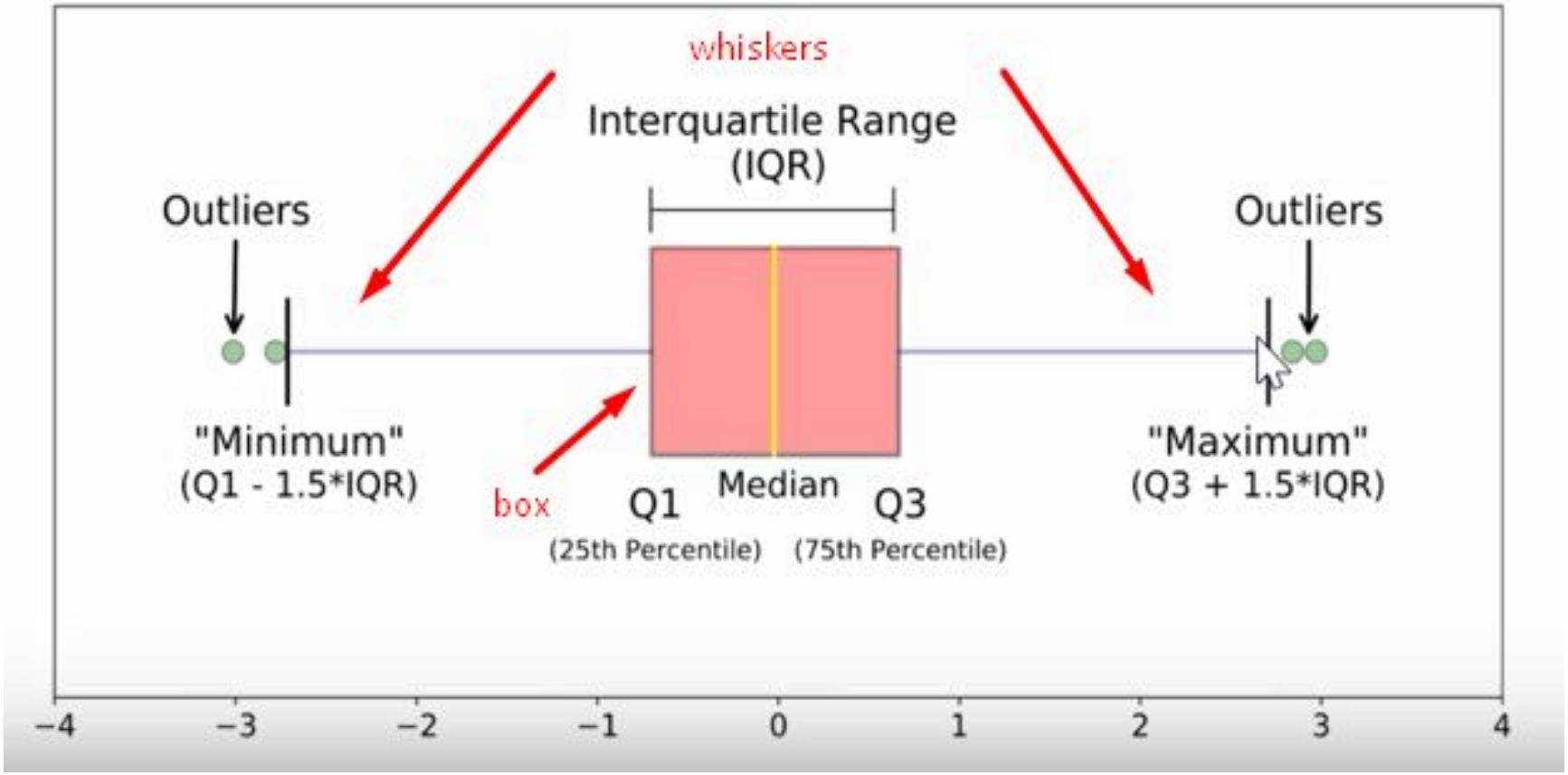
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i,$$

$$S^2 = 0.002946 \qquad s^2 = 0.003006$$

Выборочное среднеквадратическое отклонение

$$S = \sqrt{S^2} = \sqrt{0.002946} = 0.054279, \qquad s = \sqrt{s^2} = \sqrt{0.003006} = 0.05483.$$

Box and whisker plot

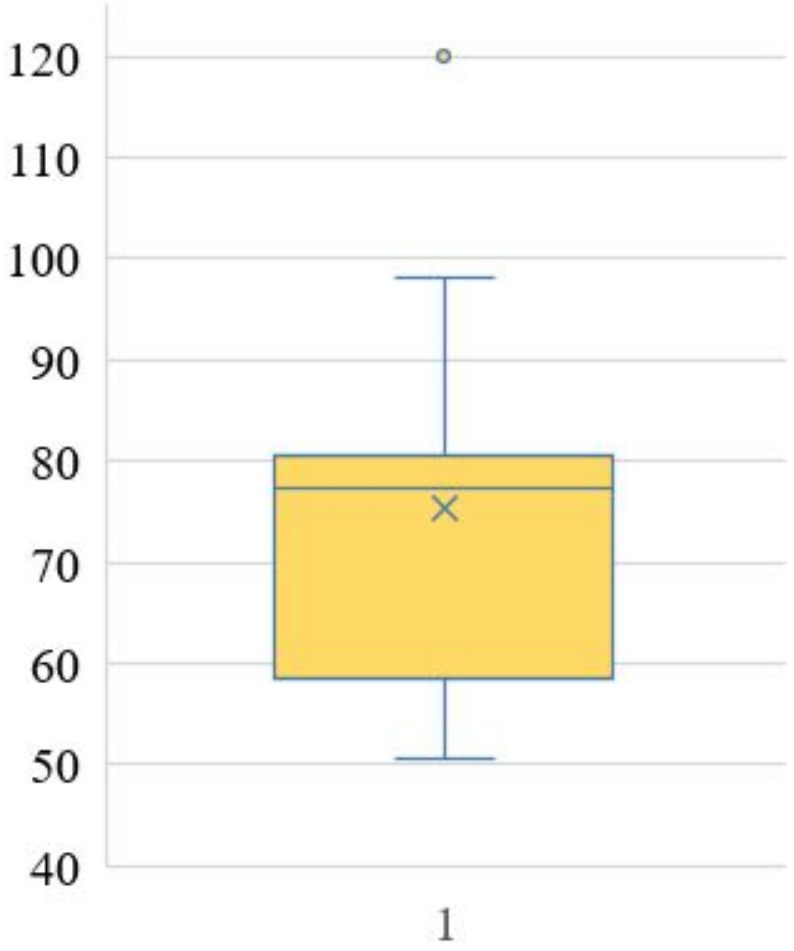


Характеристики графика boxplot

Boxplot показывает **пять важных характеристических значений** для набора данных:

- Минимум .
- Нижний квартиль .
- Медиана .
- Верхний квартиль .
- Максимум .

		Выбросы
Q1	58,6	50,5
Q2	77,4	52,6
Q3	80,5	57,5
		59,7
IQR	21,9	73,8
Q1-1,5 IQR	25,75	77,4
Q3+1,5 IQR	113,35	77,9
Выб.среднее	75,3	80,2
		80,8
		98
		120



Асимметрия

Асимметрия характеризует отклонение распределения признака относительного нормального распределения.

$$A_s = \frac{m_3}{S^3}, \qquad m_3 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i. \qquad m_3 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^3}{n} = \frac{0.00223}{50} = 0.0000446.$$

m_3 - центральный момент третьего порядка $A_s = 0.278894.$

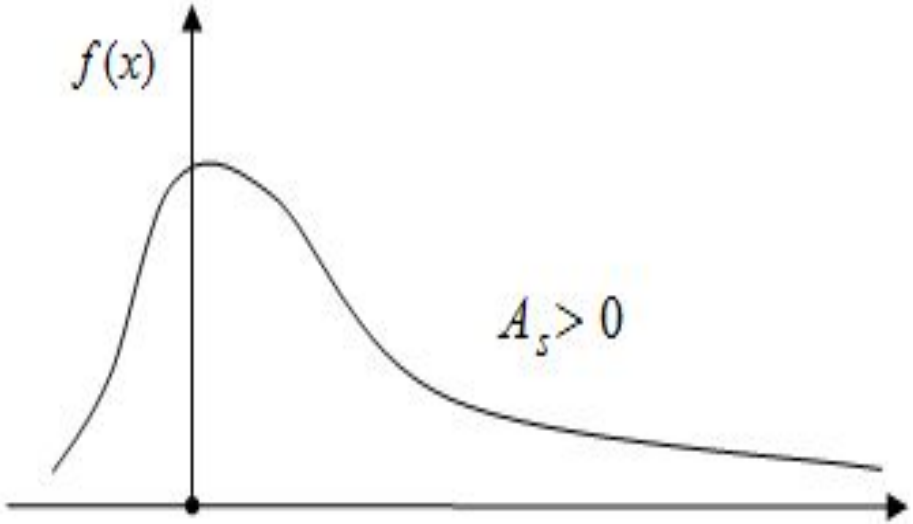


График функции плотности распределения случайной величины с правосторонней асимметрией

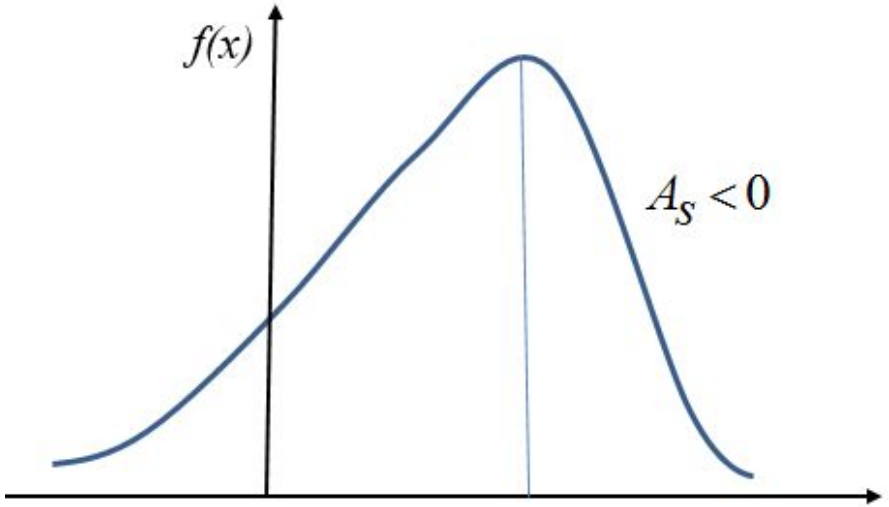


График функции плотности распределения случайной величины с левосторонней асимметрией

Эксцесс

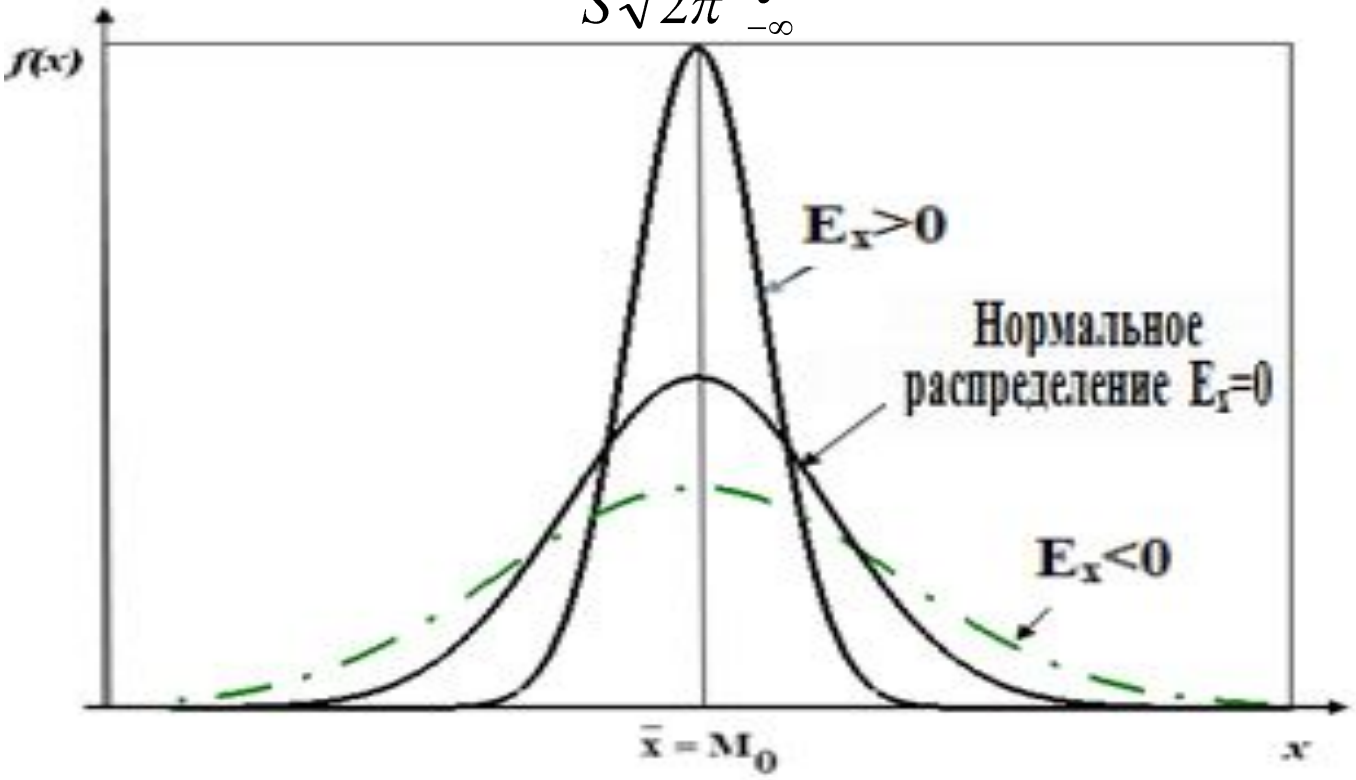
Эксцесс – величина островершинности.

$$E_x = \frac{m_4}{S^4} - 3, \qquad m_4 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i. \qquad m_4 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^4}{n} = \frac{0.001175}{50} = 0.0000235.$$

m_4 - центральный момент четвертого порядка

$E_x = -0.29274057.$

$$F(t) = \frac{1}{S\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{(x-\bar{x})^2}{2S^2}} dx.$$



Доверительные интервалы

Для математического ожидания

$$\bar{x} - t_\gamma \frac{s}{\sqrt{n}} < m < \bar{x} + t_\gamma \frac{s}{\sqrt{n}}$$

$$0{,}856974 < m < 0{,}887826,$$

где, t_γ -γ-квантиль распределения Стьюдента с n-1 степенью свободы.

Примечание: Для поиска t_γ можно воспользоваться функцией Excel =СТЬЮДРАСПОБР(1-γ, n-1).

Для дисперсии

$$\frac{(n-1)s^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_1^2}$$

$$0.045802 < \sigma < 0.068326.$$

$$p_1 = P(\chi^2 \le \chi_1^2) = \frac{\alpha}{2}$$

$$\chi^2_{\frac{1-0.95}{2},50-1} = \text{ХИ2ОБР}((1-0{,}95)/2;50-1) = 70.2241$$

$$p_2 = P(\chi^2 \le \chi_2^2) = 1-\alpha+\frac{\alpha}{2} = 1-\frac{\alpha}{2}.$$

$$\chi^2_{\frac{1+0.95}{2},50-1} = \text{ХИ2ОБР}((1+0{,}95)/2;50-1) = 31.55492$$

Распределение Стьюдента

k	q							
	0,2	0,1	0,05	0,02	0,01	0,005	0,002	0,001
1	3,08	6,31	12,71	31,82	63,66	127,32	318,3	636,61
2	1,89	2,92	4,30	6,96	9,92	14,09	22,33	31,60
3	1,64	2,35	3,18	4,54	5,84	7,45	10,21	12,92
4	1,53	2,13	2,78	3,75	4,60	5,60	7,17	8,61
5	1,48	2,02	2,57	3,36	4,03	4,77	5,89	6,87
6	1,44	1,94	2,45	3,14	3,71	4,32	5,21	5,96
7	1,41	1,89	2,36	3,00	3,50	4,03	4,79	5,41
8	1,40	1,86	2,31	2,90	3,36	3,83	4,50	5,04
9	1,38	1,83	2,26	2,82	3,25	3,69	4,30	4,78
10	1,37	0,81	2,23	2,76	3,17	3,58	4,14	4,59
11	0,36	0,80	2,20	2,72	3,11	3,50	4,02	4,44
12	0,36	0,78	2,18	2,68	3,05	3,43	3,93	4,32
13	0,35	0,77	2,16	2,65	3,01	3,37	3,85	4,22
14	0,34	0,76	2,14	2,62	2,98	3,33	3,79	4,14
15	0,34	0,75	2,13	2,60	2,95	3,29	3,73	4,07
16	0,34	0,75	2,12	2,58	2,92	3,25	3,69	4,02
17	0,33	0,74	2,11	2,57	2,90	3,22	3,65	3,97
18	0,33	0,73	2,10	2,55	2,88	3,2	3,61	3,92
19	0,33	0,73	2,09	2,54	2,86	3,17	3,58	3,88
20	0,33	0,72	2,09	2,53	2,85	3,15	3,55	3,85

Распределение χ^2

K	*p*										
	0,99	**0,95**	**0,90**	**0,50**	**0,25**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**	**0,001**
1	0,0002	0,004	0,02	0,46	1,32	2,71	3,84	5,02	6,63	7,88	10,8
2	0,02	0,1	0,21	1,39	2,77	4,61	5,99	7,38	9,21	10,6	13,8
3	0,12	0,35	0,58	2,37	4,11	6,25	7,81	9,35	11,3	12,8	16,3
4	0,30	0,71	1,06	3,36	5,39	7,78	9,49	11,1	13,3	14,9	18,5
5	0,55	1,15	1,61	4,35	6,63	9,24	11,1	12,8	15,1	16,7	20,5
6	0,87	1,64	2,20	5,35	7,84	10,6	12,6	14,4	16,8	18,5	22,5
7	1,24	2,17	2,83	6,35	9,04	12,0	14,1	16,0	18,5	20,3	24,3
8	1,65	2,73	3,49	7,34	10,2	13,4	15,5	17,5	20,1	22,0	26,1
9	2,09	3,33	4,17	8,34	11,4	14,7	16,9	19,0	21,7	23,6	27,9
10	2,56	3,94	4,87	9,34	12,5	16,0	18,3	20,5	23,2	25,2	29,6
11	3,05	4,57	5,58	10,3	13,7	17,3	19,7	21,9	24,7	26,8	31,3
12	3,57	5,23	6,3	11,3	14,8	18,5	21,0	23,3	26,2	28,3	32,9
13	4,11	5,89	7,04	12,3	16,0	19,8	22,4	24,7	27,7	29,8	34,5
14	4,66	6,57	7,79	13,3	17,1	21,1	23,7	26,1	29,1	31,3	36,1
15	5,23	7,26	8,55	14,3	18,2	22,3	25,0	27,5	30,6	32,8	37,7
16	5,81	7,96	9,31	15,3	19,4	23,5	26,3	28,8	32,0	34,3	39,3
17	6,41	8,67	10,1	16,3	20,5	24,8	27,6	30,2	33,4	35,7	40,8
18	7,01	9,39	10,9	17,3	21,6	26,0	28,9	31,5	34,8	37,2	42,3
19	7,63	10,1	11,7	18,3	22,7	27,2	30,1	32,9	36,2	38,6	43,8
20	8,26	10,9	12,4	19,3	23,8	28,4	31,4	34,2	37,6	40,0	45,3
21	8,90	11,6	13,2	20,3	24,9	29,6	32,7	35,5	38,9	41,4	46,8
22	9,54	12,3	14,0	21,3	26,0	30,8	33,9	36,8	40,3	42,8	48,3