Московский политехнический университет

Математические методы анализа данных

Лежнина Юлия Аркадьевна

Москва 2022

План лекции 1

- **Основы теории вероятностей.**
- **Свойства вероятности.**
- **Условная вероятность.**
- **Случайные величины.**
- **Примеры.**
- **Характеристики случайных величин.**
- **Графический анализ.**
- **Эмпирическая функция.**
- **Гистограммы.**
- **Ящик с усами.**
- **Диаграмма рассеивания.**

Основные понятия теории вероятностей

- **Случа́йный эксперименте́нт** (случайное испытание, случайный опыт) — математическая модель соответствующего реального эксперимента, результат которого невозможно точно предсказать.
- **Пространство элементарных событий** — множество Ω всех различных исходов случайного эксперимента. Элемент этого множества называется элементарным событием или исходом.
- **Случа́йное собы́тие** — подмножество множества исходов случайного эксперимента.
- **Случайной величиной** называется функция, определенная на пространстве элементарных исходов, и принимающая свои значения на некотором множестве.

Вероятность

- *Вероятностью случайного события* A *называется отношение числа* n *несовместимых равновероятных элементарных событий, составляющих событие* A*, к числу всех возможных элементарных событий* N*:*

$$\Pr(A) = \frac{n}{N}$$

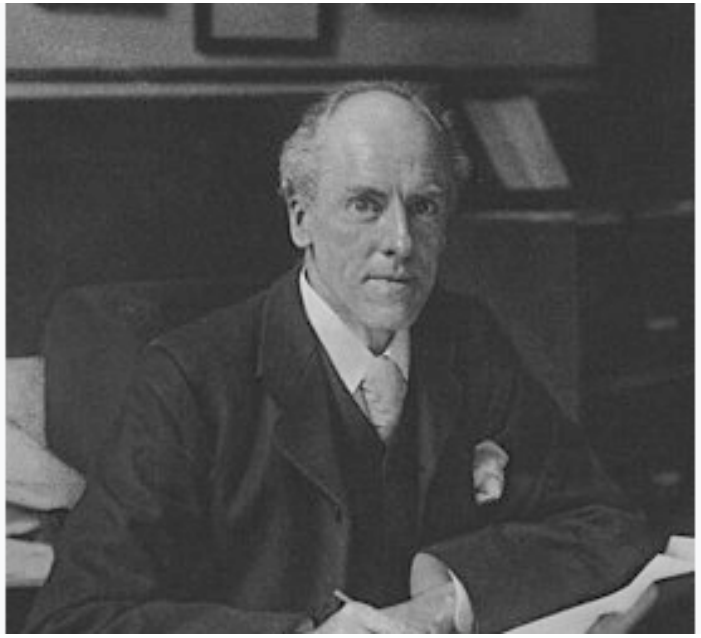
Жорж-Луи Леклéрк, граф де Бюффóн

N=4040

Орел – 2048 выпадений
Решка – 1992 выпадения

P(Орел)=0,507

Карл Пи́рсон



N=24000

Орел – 12012 выпадений
Решка – 11988 выпадения

P(Орел)=0,5005

Статистическая вероятность

- Статистической вероятностью случайного события называется отношение m, числа испытаний, в которых это событие появилось, к общему числу n, проведённых испытаний, и обозначается:

$$W(A) = \frac{m}{n}$$

Свойства вероятности

$$\mathbf{P}\{\varnothing\} = 0;$$

$$0 \leqslant \mathbf{P}\{A\} \leqslant 1; \qquad \mathbf{P}\{\bar{A}\} = 1 - \mathbf{P}\{A\};$$

$$\mathbf{P}\{A\} \leqslant \mathbf{P}\{B\}; \qquad \mathbf{P}\{B \setminus A\} = \mathbf{P}\{B\} - \mathbf{P}\{A\};$$

$$\mathbf{P}\{A + B\} = \mathbf{P}\{A\} + \mathbf{P}\{B\} - \mathbf{P}\{AB\}.$$

Условная вероятность

$$P(A/B) = \frac{P(AB)}{P(B)}$$

Формула полной вероятности

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A \mid H_i)$$

Формула Бейеса

$$P(Hi|A) = \frac{P(H_i) \cdot P(A/H_i)}{\sum_{i=1}^{n} P(H_i)P(A/H_i)} = \frac{P(H_i) \cdot P(A/H_i)}{P(A)}$$

Законы распределения случайных величин

Случайные величины

дискретные

непрерывные

Функция распределения. Плотность

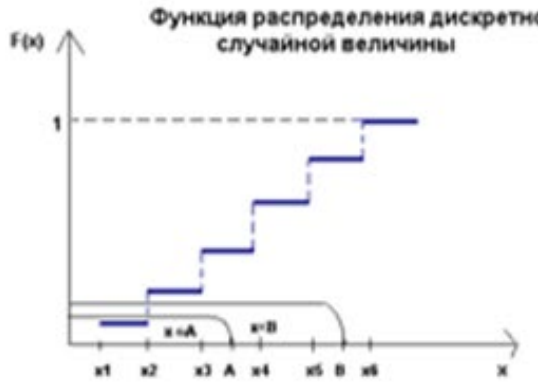
Функция распределения

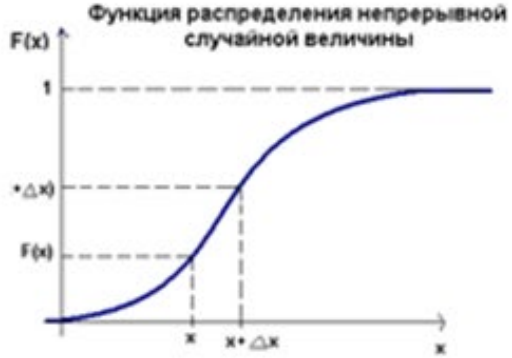
Функцией распределения вероятностей F(x) случайной величины X в точке x называется вероятность того, что в результате опыта случайная величина примет значение, меньше, чем x, т.е. $F(x)=P\{X < x\}$.

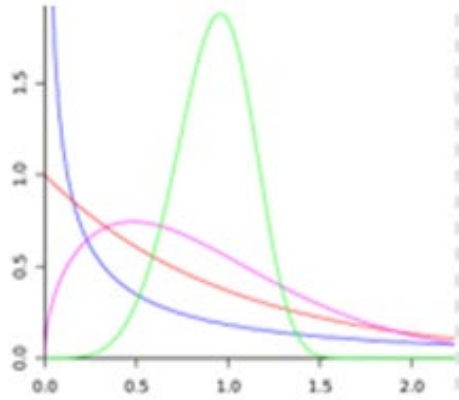
Плотность и функция вероятности

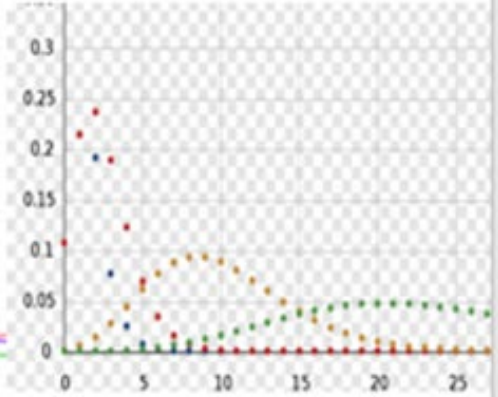
дифференциальная функция распределения. Она представляет собой производную функции распределения.

для дискретной случайной величины рассматриваем функцию вероятности



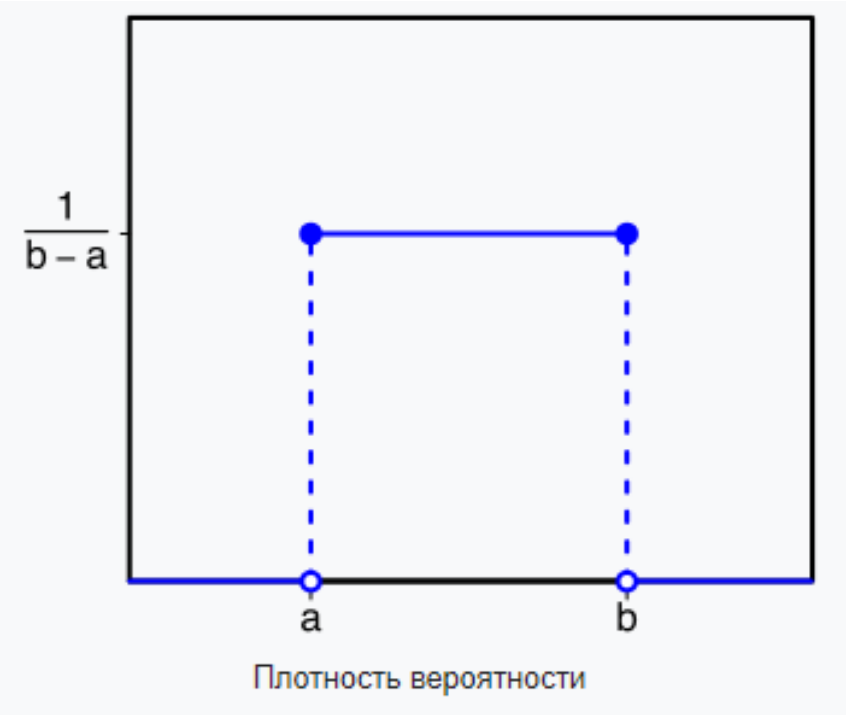






Равномерное распределение

Говорят, что случайная величина имеет непрерывное равномерное распределение на отрезке [a,b], где a и b действительные числа, если её плотность fX(x) имеет вид:



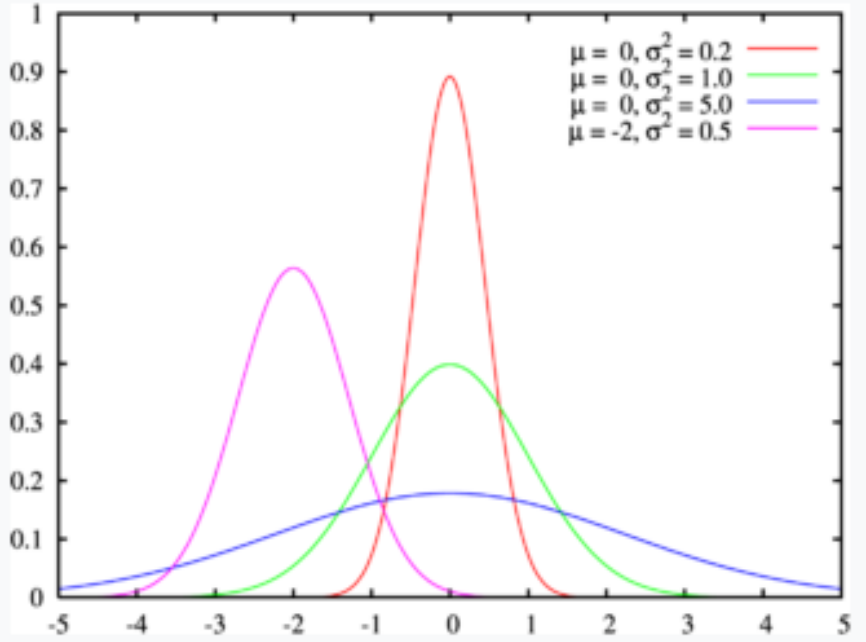
Плотность вероятности

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}.$$

Математическое ожидание	$\dfrac{a+b}{2}$
Медиана	$\dfrac{a+b}{2}$
Мода	любое число из отрезка $[a,b]$
Дисперсия	$\dfrac{(b-a)^2}{12}$

Нормальное распределение

Норма́льное распределе́ние, также называемое распределением Гаусса или Гаусса — Лапласа — распределение вероятностей, которое в одномерном случае задаётся функцией плотности вероятности, совпадающей с функцией Гаусса (здесь параметр μ — математическое ожидание (среднее значение), медиана и мода распределения, а параметр σ — среднеквадратическое отклонение, σ ^{2} — дисперсия распределения :



Зеленая линия соответствует стандартному нормальному распределению

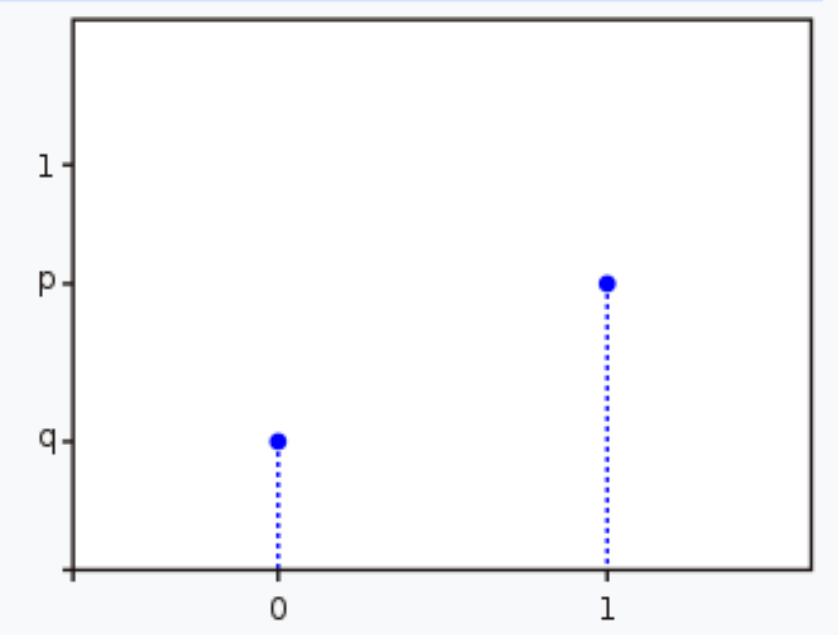
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

Математическое ожидание	μ
Медиана	μ
Мода	μ
Дисперсия	σ^2

Распределение Бернулли

Случайная величина X имеет распределение Бернулли, если она принимает всего два значения: 1 и 0 с вероятностями p и q= 1-p соответственно. Таким образом:

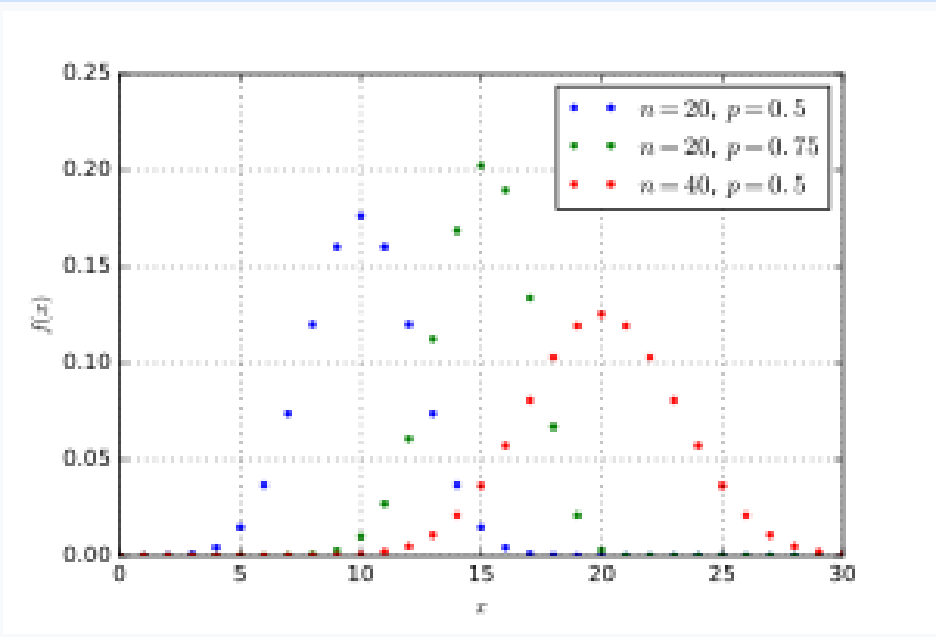
$$\mathbb{P}(X = 1) = p,$$
$$\mathbb{P}(X = 0) = q.$$



Функция вероятности	$\begin{matrix} q & k=0 \\ p & k=1 \end{matrix}$
Функция распределения	$\begin{matrix} 0 & k<0 \\ q & 0\le k<1 \\ 1 & k\ge 1 \end{matrix}$
Математическое ожидание	p
Мода	$\begin{cases} 0, & q>p \\ 0,1, & q=p \\ 1, & q<p \end{cases}$
Дисперсия	pq

Биномиальное распределение

Пусть X_1,…,X_n — конечная последовательность независимых случайных величин, имеющих одинаковое распределение Бернулли с параметром p, то есть при каждом n величина X_i принимает значения 1 («успех») и 0 («неудача») с вероятностями p и q=1-p соответственно. Тогда случайная величина Y=X_1+…+X_n имеет биномиальное распределение с параметрами n и p.

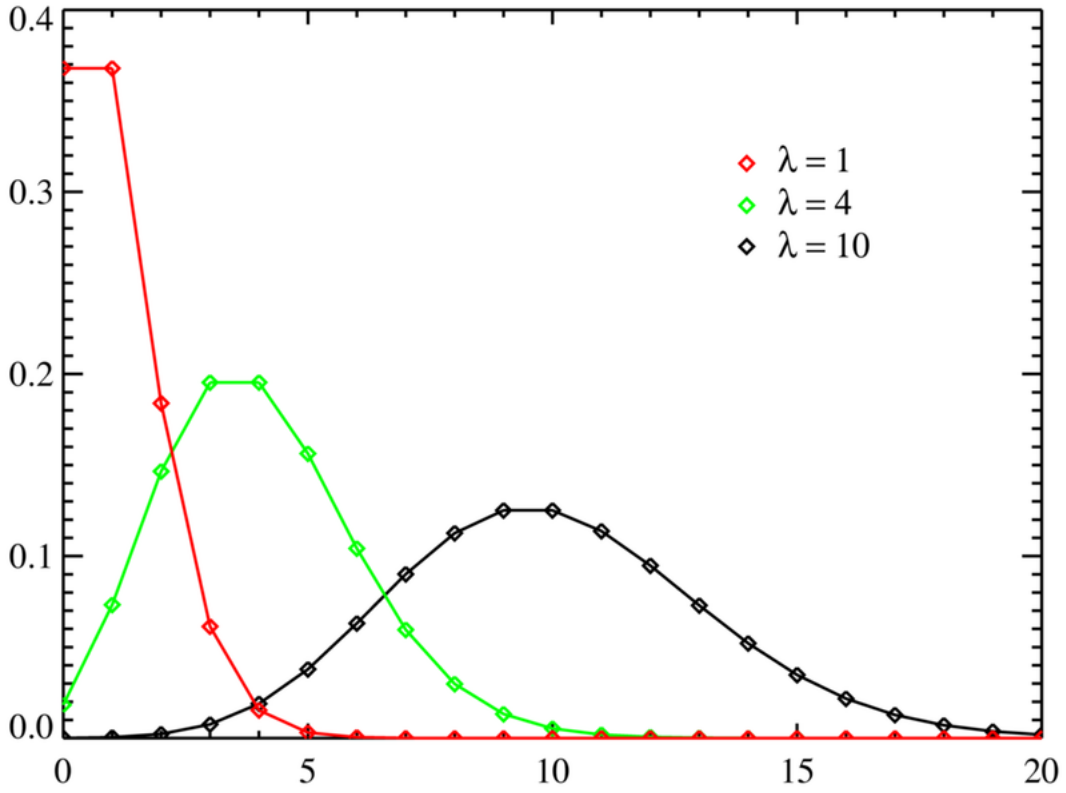


$$p_Y(k) \equiv \mathbb{P}(Y = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, \ldots, n,$$

Функция распределения	$I_{1-p}(n - \lfloor k \rfloor, 1 + \lfloor k \rfloor)$
Математическое ожидание	np
Медиана	одно из $\{\lfloor np \rfloor - 1, \lfloor np \rfloor, \lfloor np \rfloor + 1\}$
Мода	$\lfloor (n+1)\,p \rfloor$
Дисперсия	npq

Распределение Пуассона

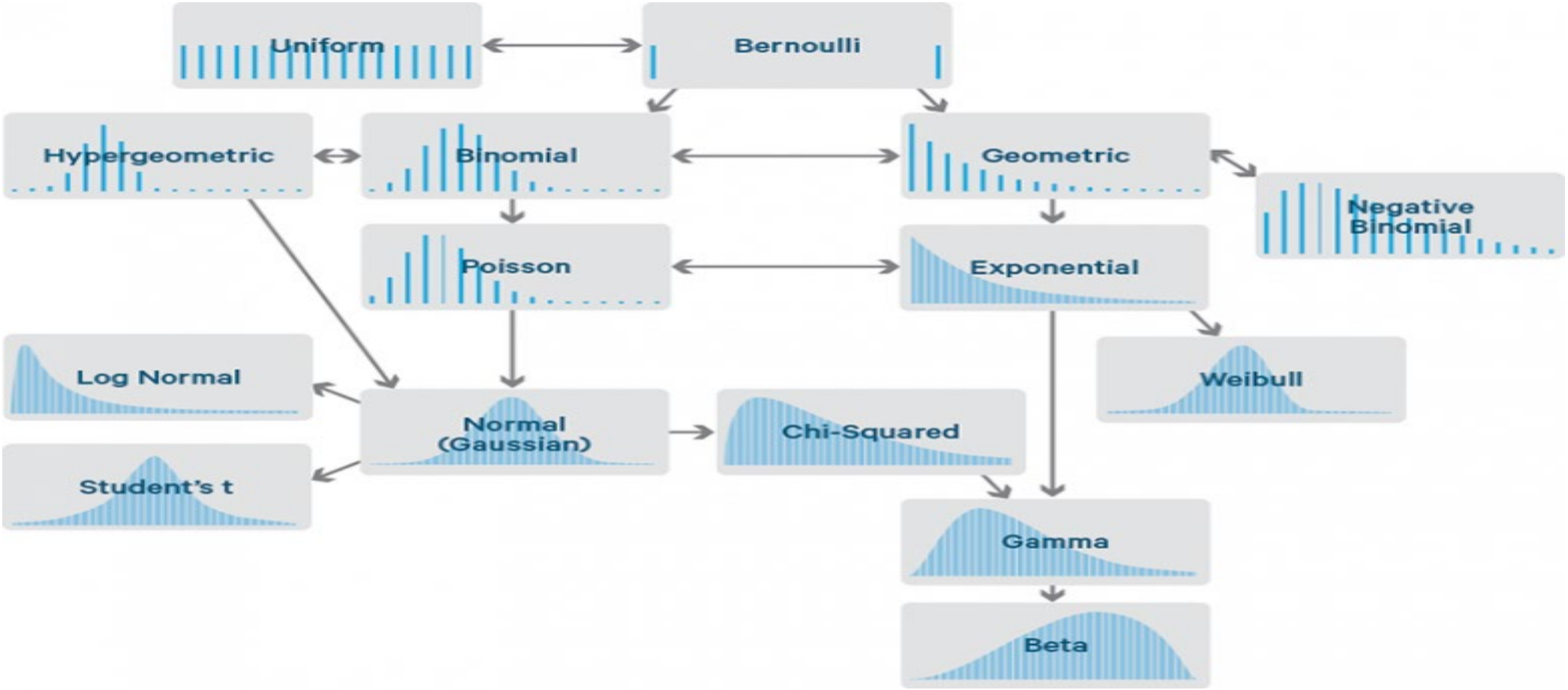
Распределе́ние Пуассо́на — распределение дискретного типа случайной величины, представляющей собой число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга.



$$p(k) \equiv \mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

Математическое ожидание	λ
Медиана	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
Мода	$\lfloor \lambda \rfloor$
Дисперсия	λ

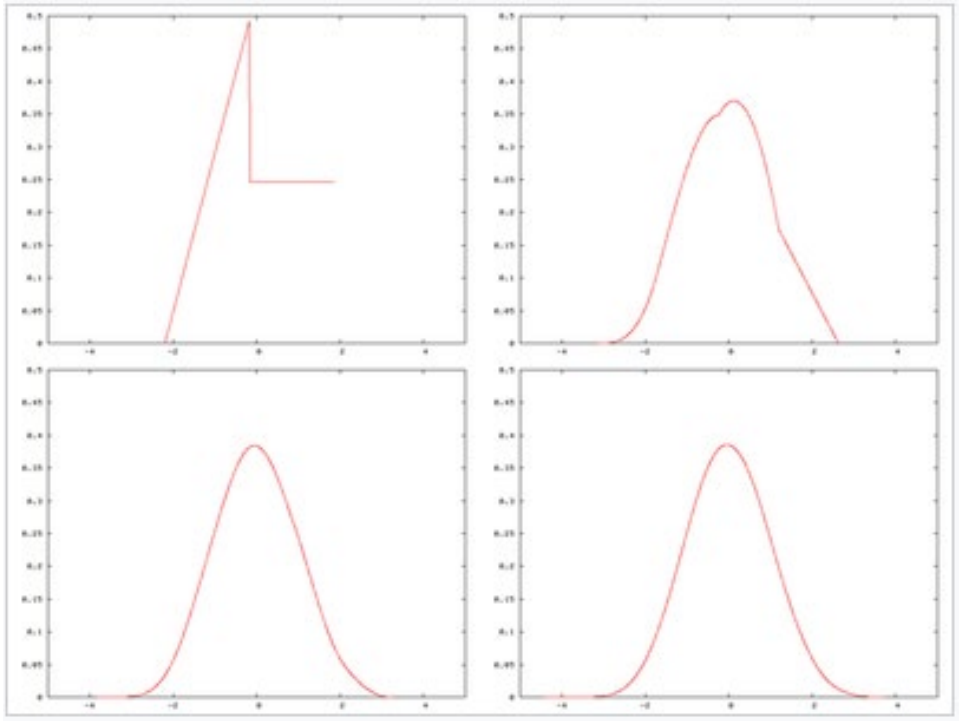
Карта связей распределений



https://habr.com/ru/post/331060/

Центральная предельная теорема (Ляпунова)

- Сумма большого числа как угодно распределенных независимых случайных величин распределена асимптотически нормально, если только слагаемые вносят равномерно малый вклад в сумму.



«Сглаживание» распределения суммированием. Показана функция плотности вероятности одной случайной величины, а также распределения суммы двух, трёх и четырёх случайных величин с такой же функцией распределения.

Характеристики случайных величин

Математическое ожидание

Дискретная

Непрерывная

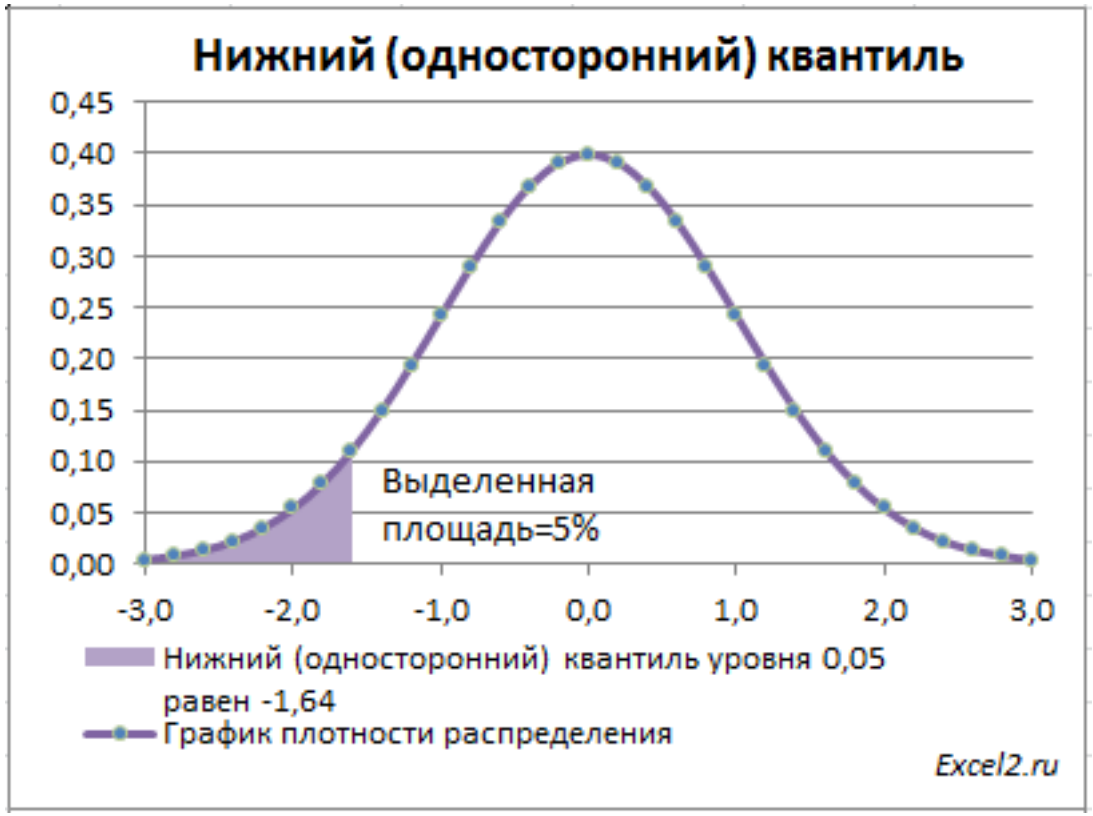
$$\mathbb{P}(X = x_i) = p_i \ , \sum_{i=1}^{\infty} p_i = 1,$$

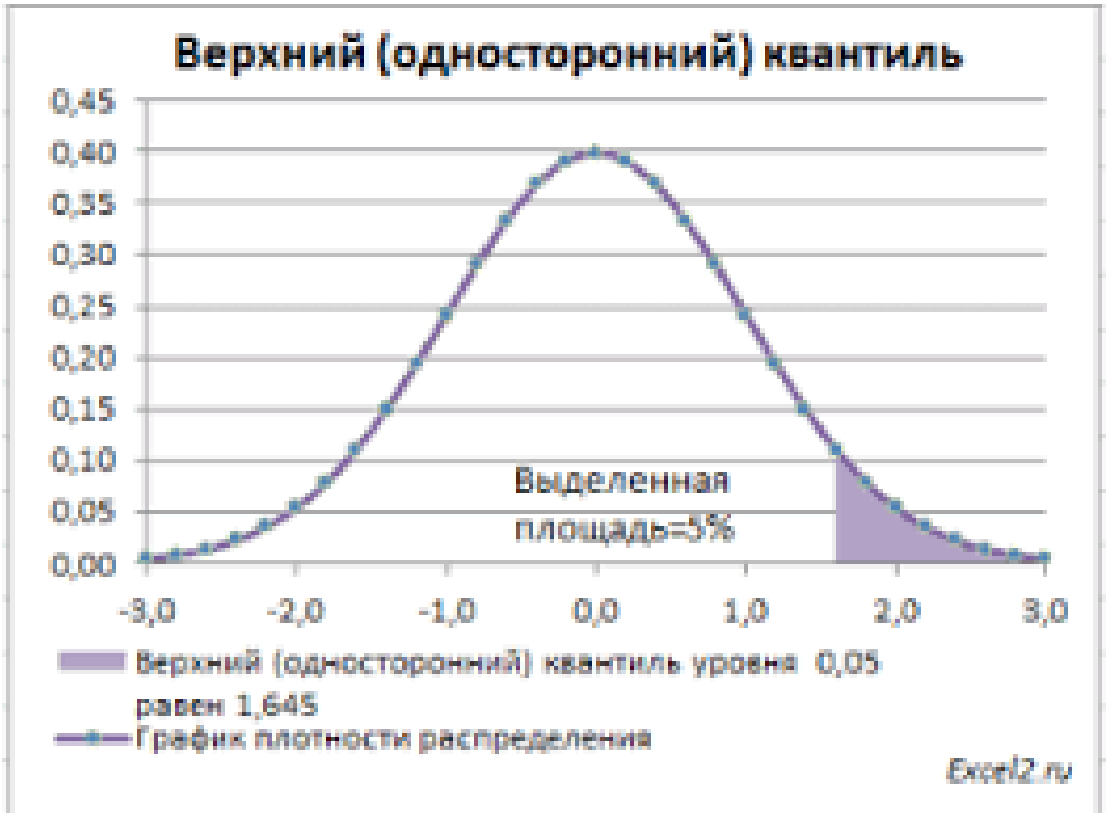
$$f_X(x),$$

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \, p_i \, .$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, f_X(x) \, dx.$$

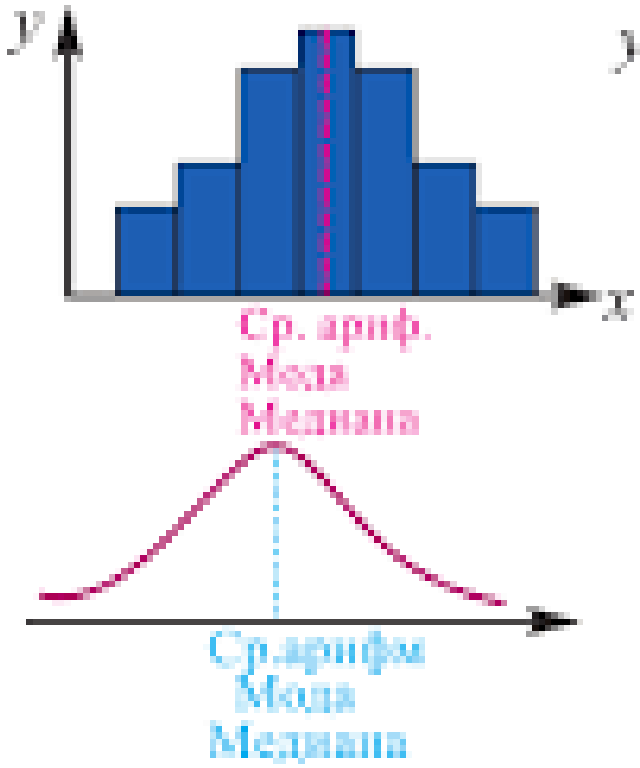
Квантили, мода, медиана



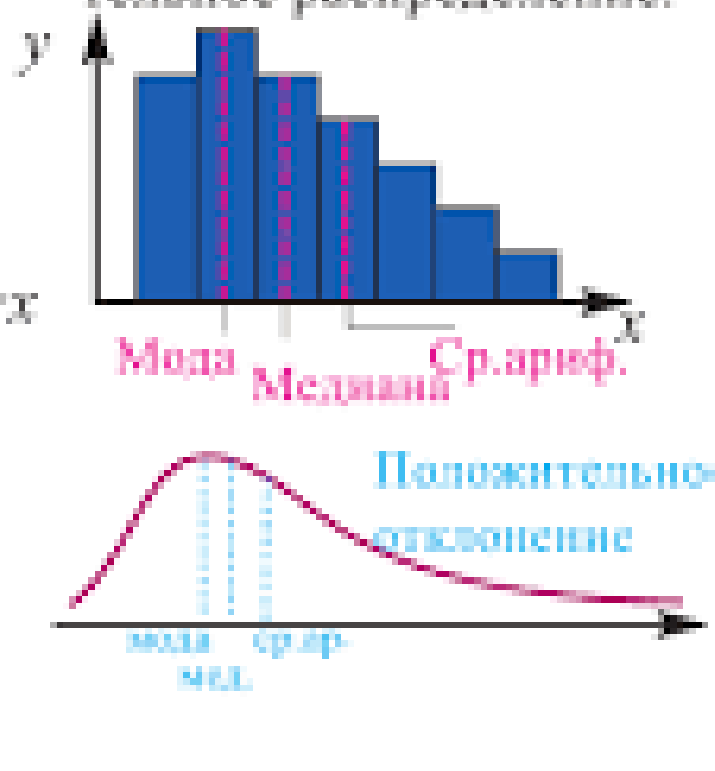


Квантили, мода, медиана

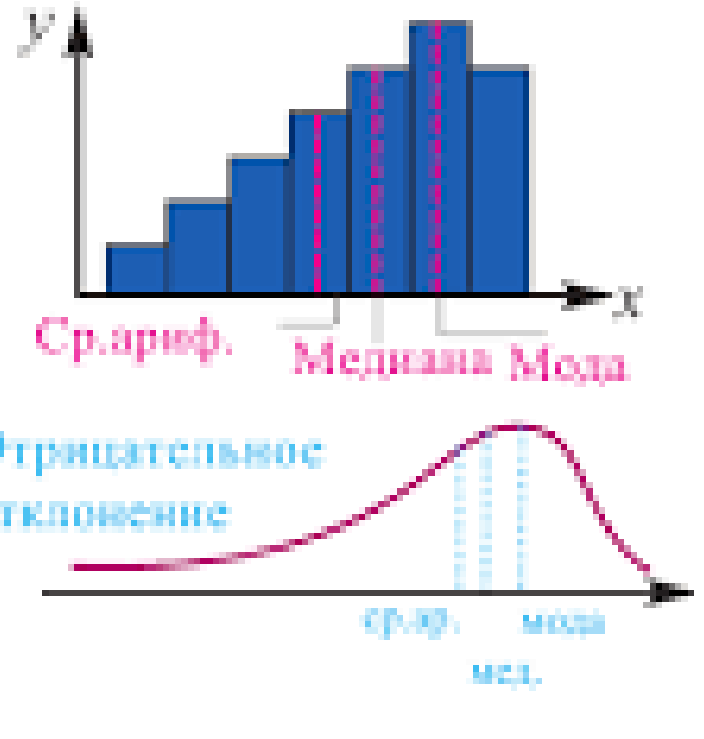
1.Нормальное распределение



2. Распределение с отклонением вправо. Положительное распределение.



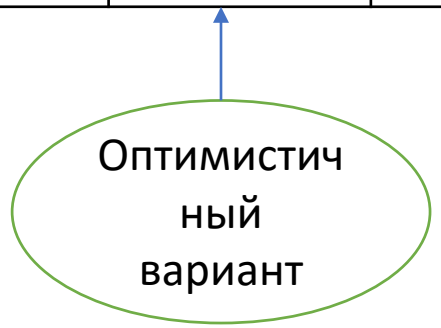
3. Распределение с отклонением влево. Отрицательное распределение.

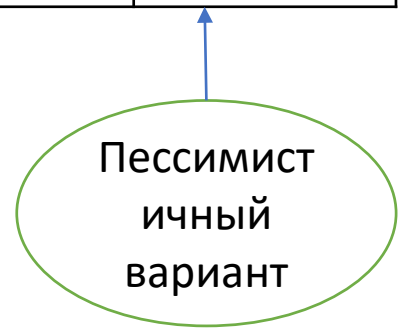


пример

- N=25

Кол-во человек	**1**	**1**	**2**	**1**	**3**	**4**	**1**	**12**
Зарплата	45000	15000	1000	**5700**	5500	3700	3000	**2000**
примечание				матожидание			медиана	мода





Вариативность случайной величины

Дисперсия

$$D[X] = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right],$$

Дискретная

$$D[X] = \sum_{i=1}^{n} p_i (x_i - \mathbb{E}[X])^2,$$

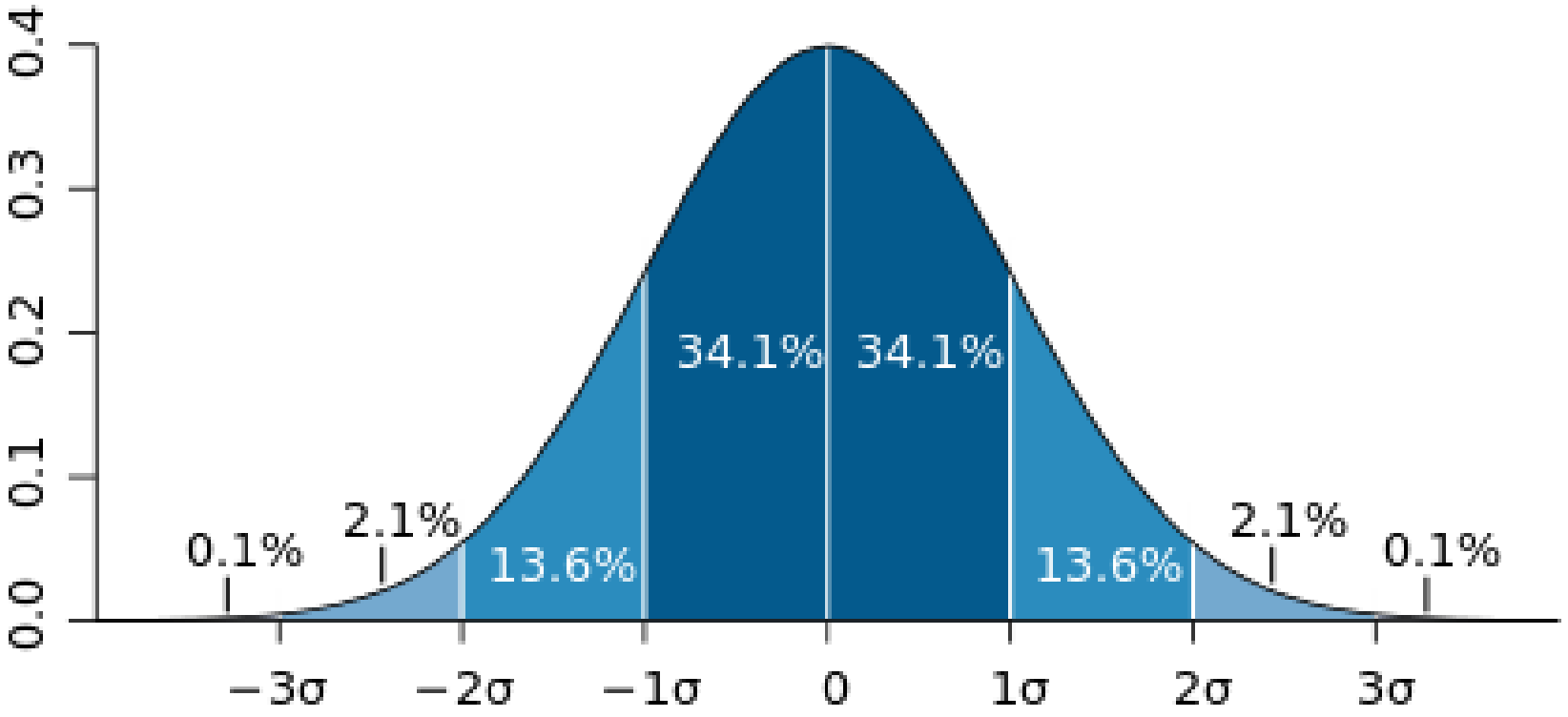
$$D[X] = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (x_i - x_j)^2 = \sum_{i=1}^{n}\sum_{j=i+1}^{n} p_i p_j (x_i - x_j)^2,$$

Непрерывная

$$D[X] = \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^2 f(x)\,dx$$

$$D[X] = \frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x_2 - x_1)^2 f(x_1) f(x_2)\,dx_1\,dx_2.$$

Квартили, правило двух σ



a=900

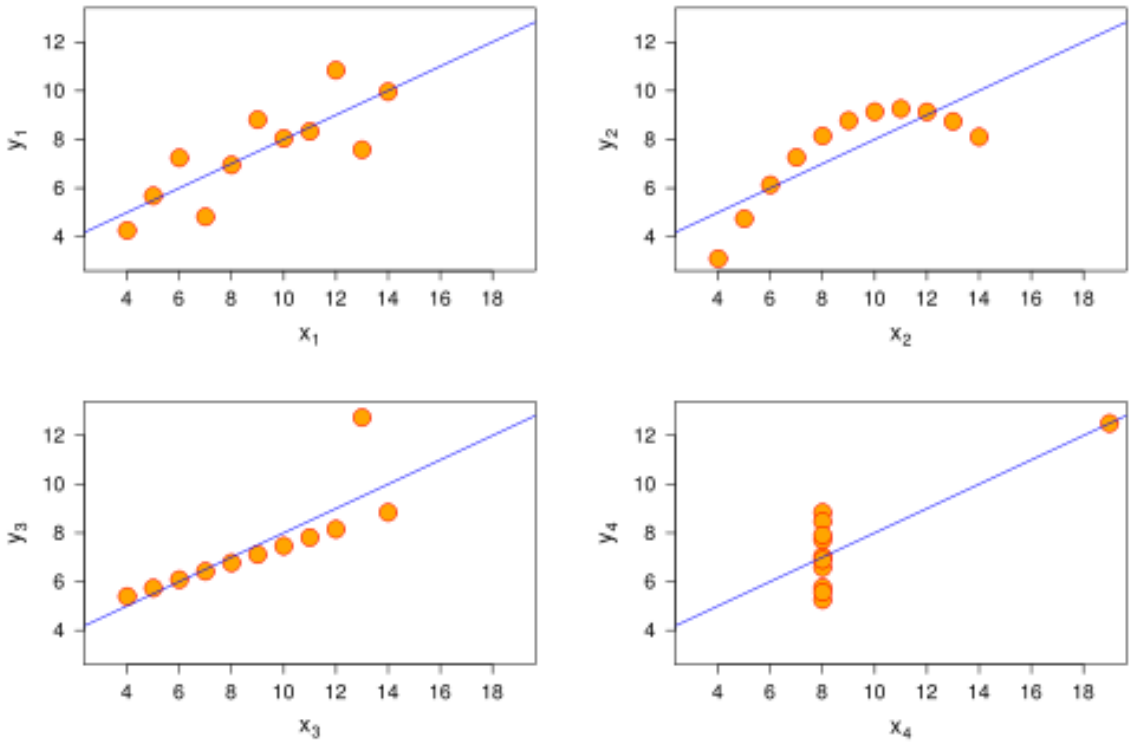
σ^2=400

Сколько кликов ожидать с вероятностью 95%?

Квартет Энскомба

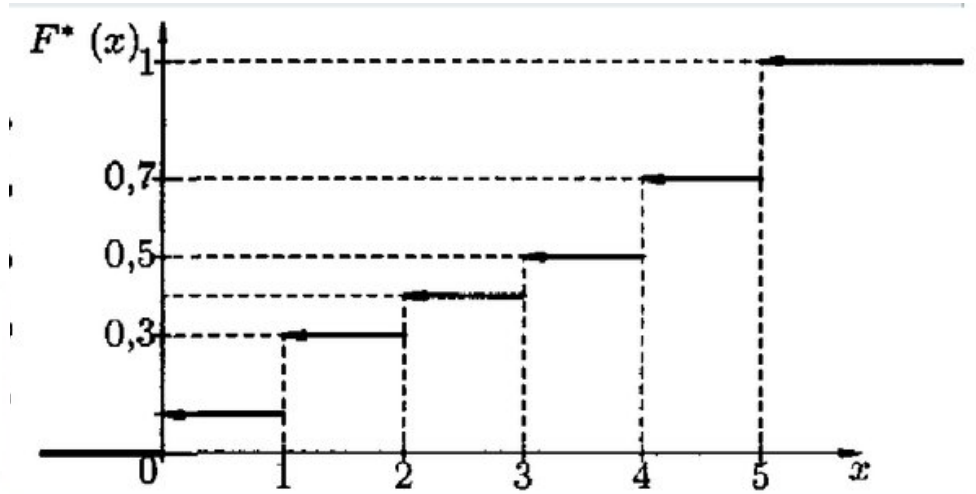
I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89

Характеристика	Значение
Среднее значение переменной x	9.0
Дисперсия переменной x	10,0
Среднее значение переменной y	7,5
Дисперсия переменной y	3,75
Корреляция между переменными x и y	0,816
Прямая линейной регрессии	$y = 3 + 0{,}5x$
Коэффициент детерминации линейной регрессии	0,67

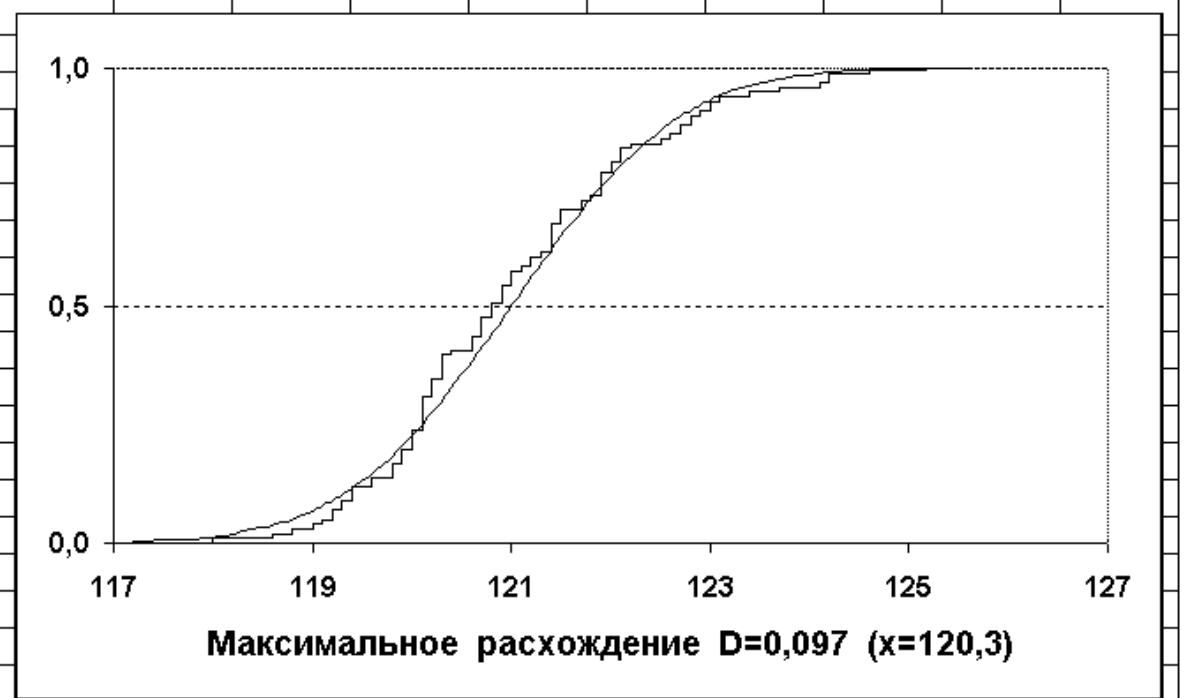


Эмпирическая функция

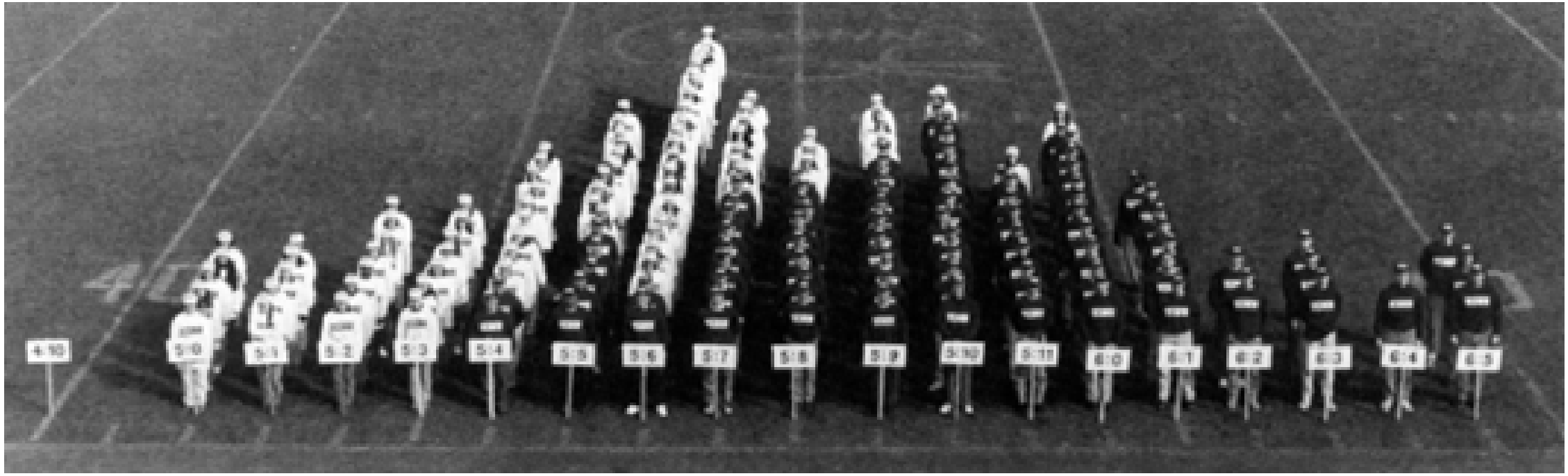
x_i	0	1	2	3	4	5
n_i	1	2	1	1	2	3



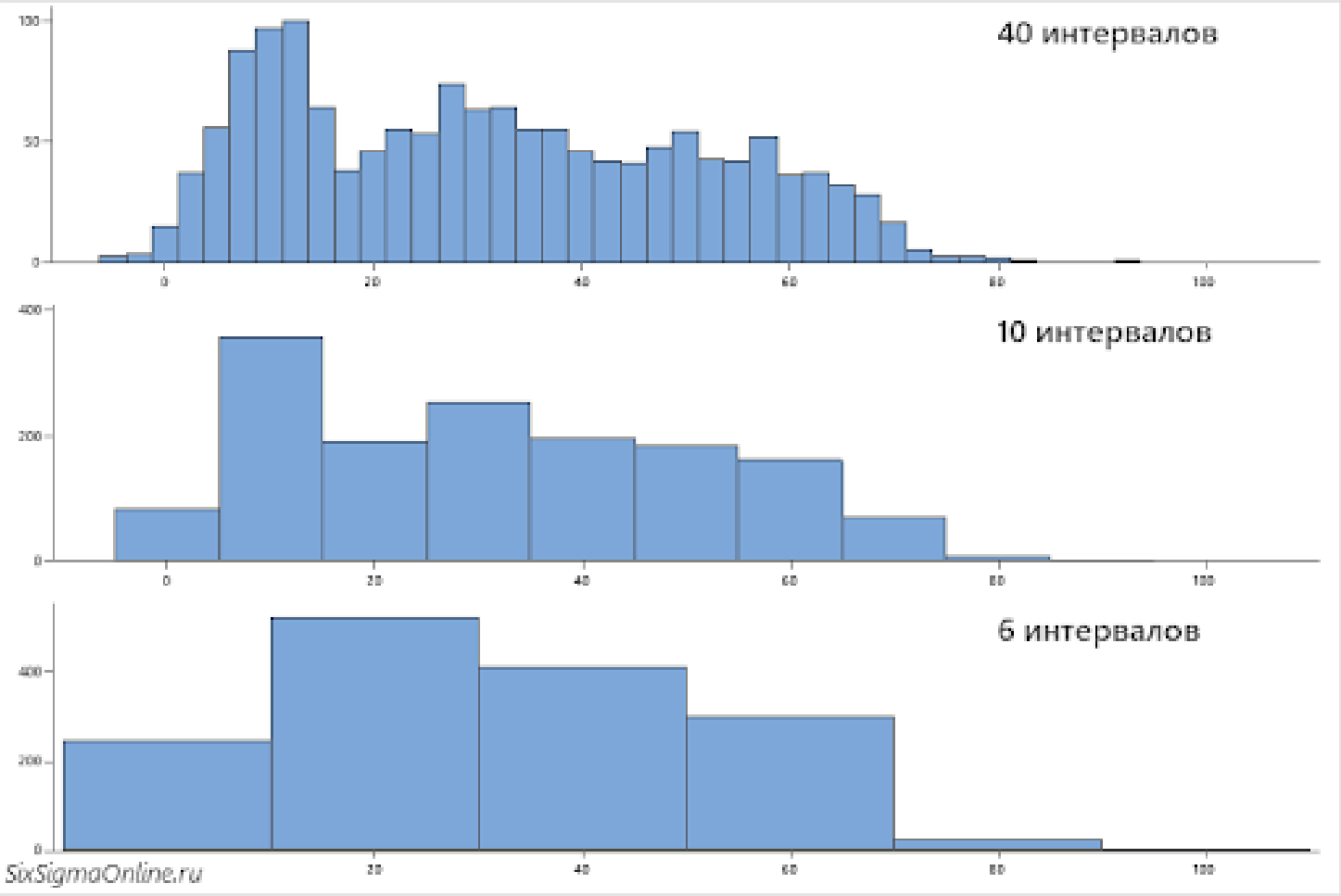
	A	B	C	D
2	117,0	0,00	0,001	0,001
3	118,0	0,00	0,012	0,012
4	118,0	0,01	0,012	0,002
5	118,6	0,01	0,036	0,026
6	118,6	0,02	0,036	0,016
7	118,8	0,02	0,049	0,029
8	118,8	0,03	0,049	0,019
9	119,0	0,03	0,066	0,037
10	119,0	0,04	0,066	0,027
11	119,1	0,04	0,077	0,037
12	119,1	0,05	0,077	0,027
13	119,2	0,05	0,088	0,038
14	119,2	0,06	0,088	0,029
15	119,2	0,06	0,088	0,029
16	119,2	0,07	0,088	0,019
17	119,3	0,07	0,101	0,031
18	119,3	0,08	0,101	0,021
19	119,3	0,08	0,101	0,021
20	119,3	0,09	0,101	0,011
21	119,4	0,09	0,114	0,025



Гистограммы



Гистограммы



Формулы для определения количества интервалов

метод Стёрджеса

$$n = 1 + \lfloor \log_2 N \rfloor,$$

$$k = \lceil \sqrt{n} \rceil$$

Формулы для определения ширины интервалов

метод Скотта

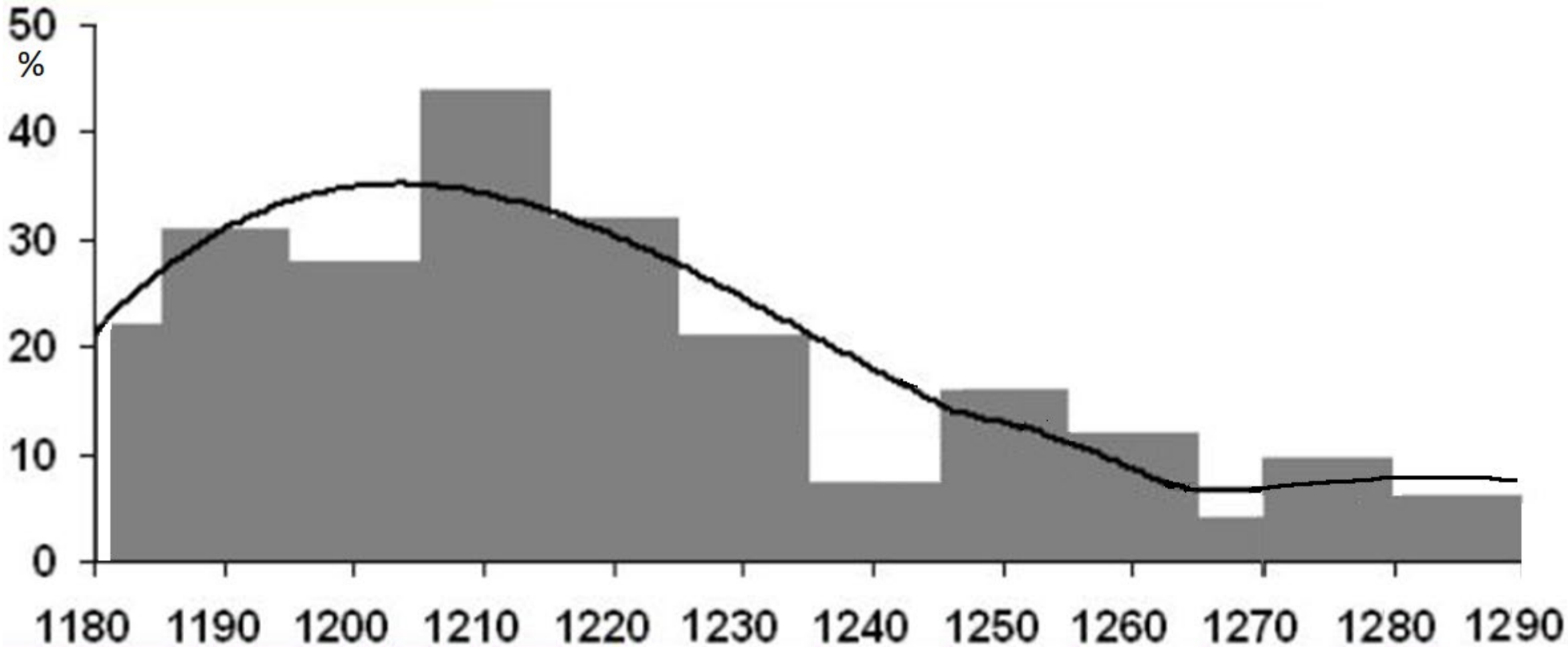
$$n = 3.5 \cdot \hat{\sigma} \cdot N^{-1/3}$$

метод Фридмана- Диакониса

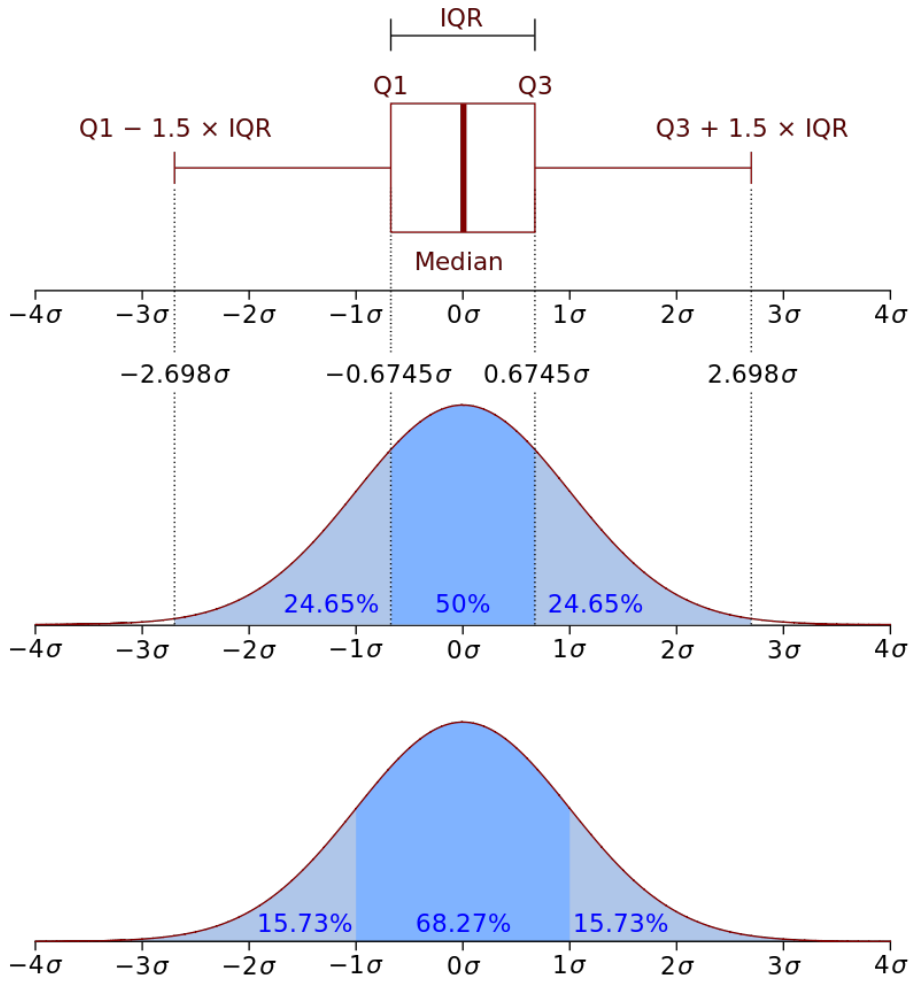
$$n = 2 \cdot IQR \cdot N^{-1/3}$$

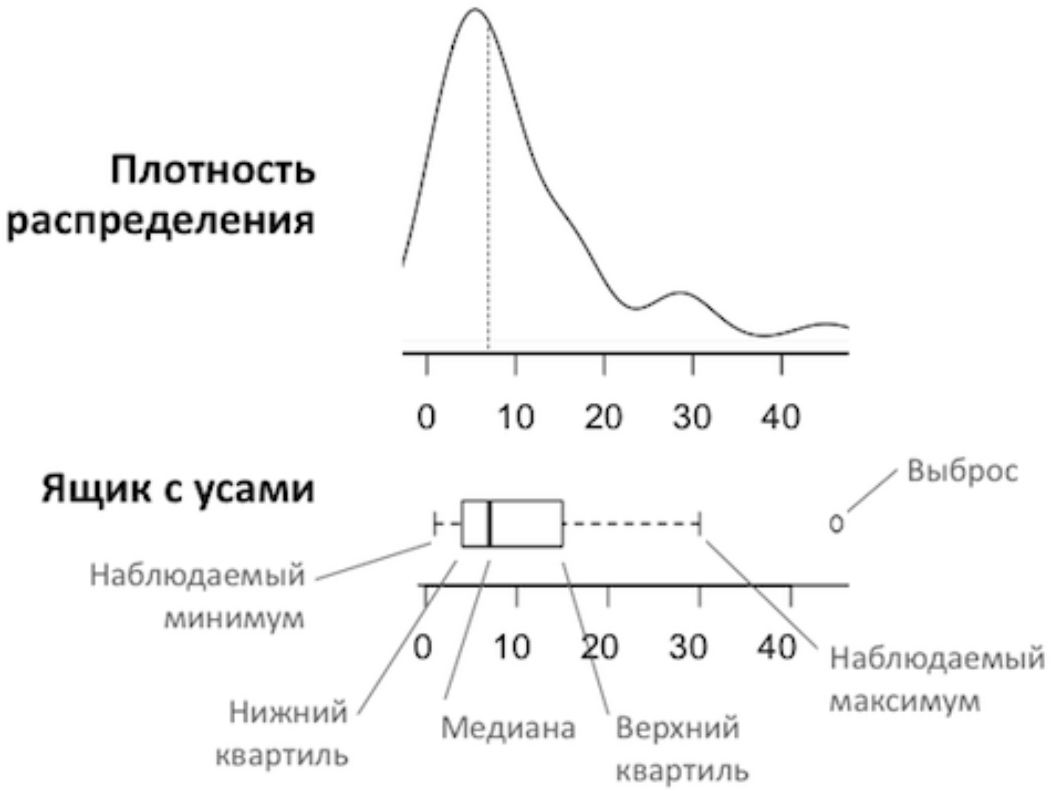
Гистограммы для квартета Энскомба

Смеси распределений времени отклика



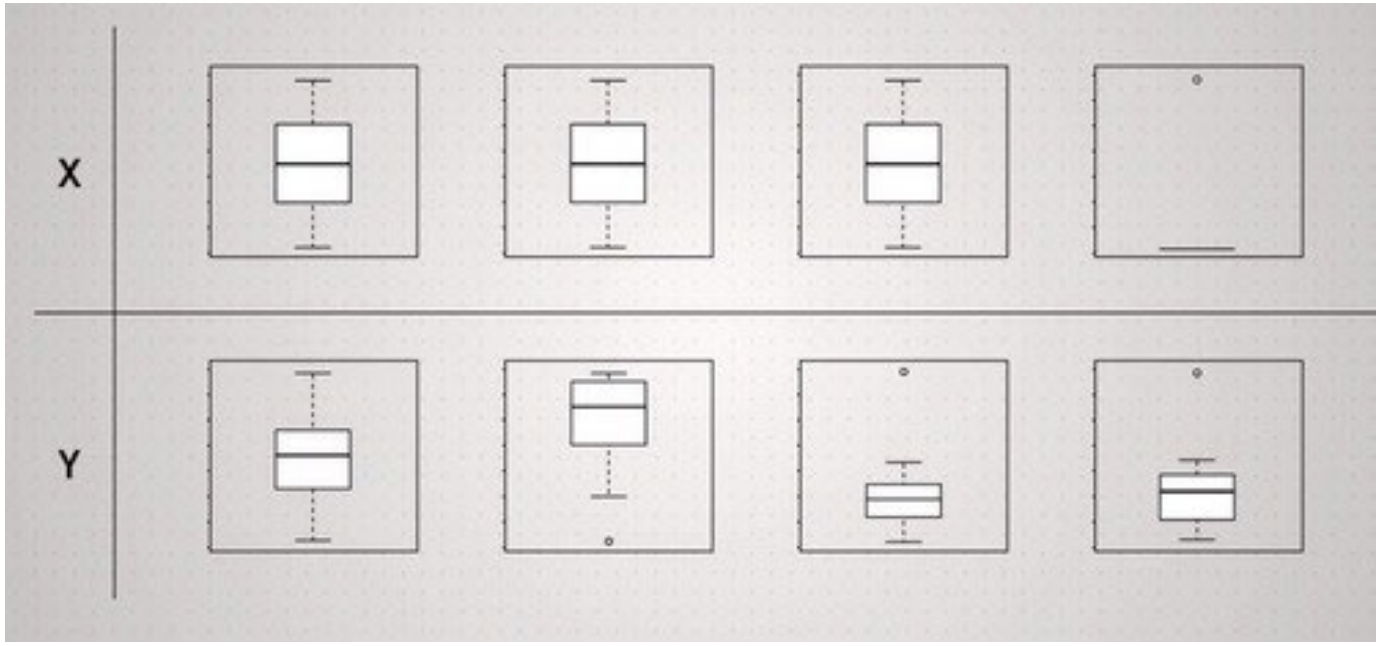
Ящик с усами


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%


Плотность распределения
0 10 20 30 40
Ящик с усами
Выброс
Наблюдаемый минимум
0 10 20 30 40
Наблюдаемый максимум
Нижний квартиль
Медиана
Верхний квартиль

Ящики с усами

Для квартета Энскомба



Для времени отклика

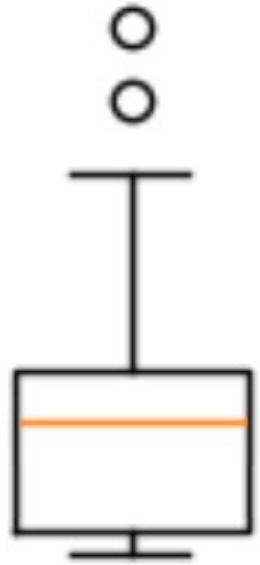


Диаграмма рассеивания