

Введение в анализ данных

Лекция 13

Статистика и визуализация

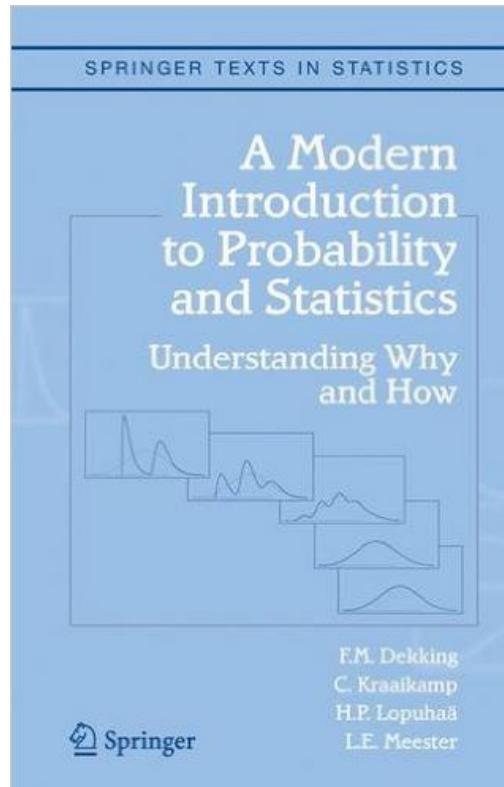
Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2017

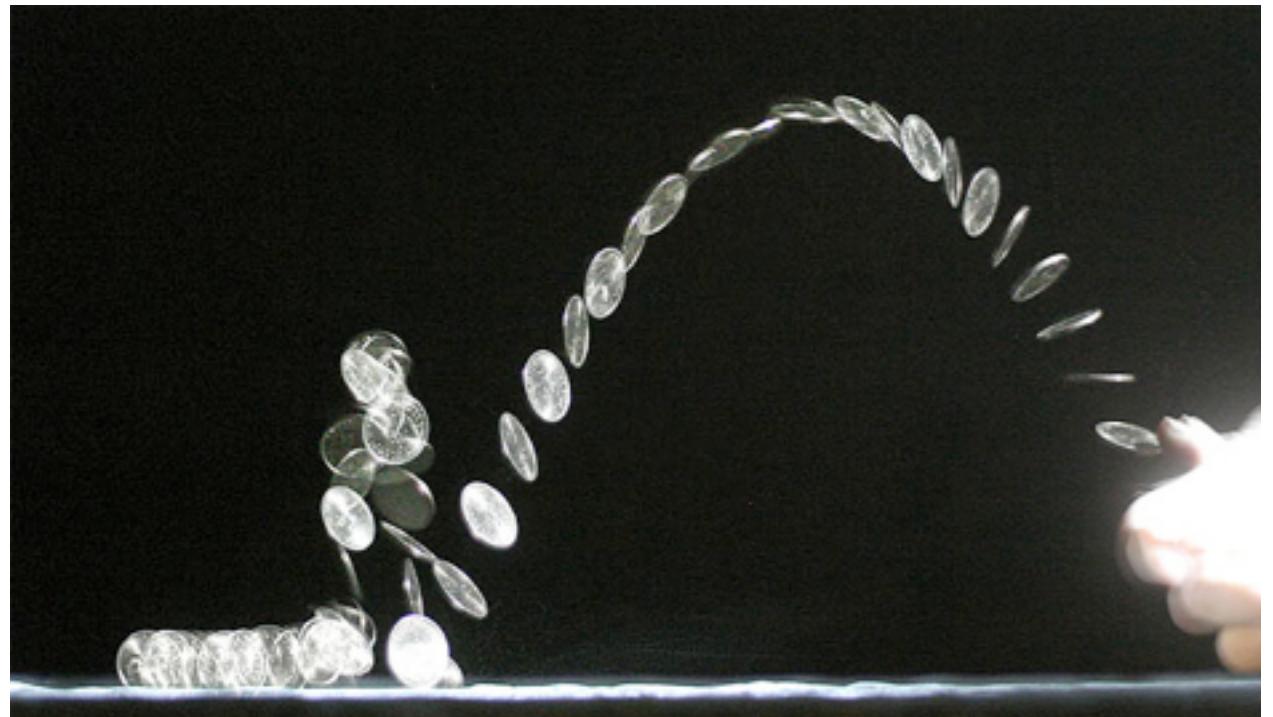
Литература

- Dekking F.M., Kraaikamp C., Lopuhaa H.P., Meester L.E. **A Modern Introduction to Probability and Statistics**. Springer, 2005.



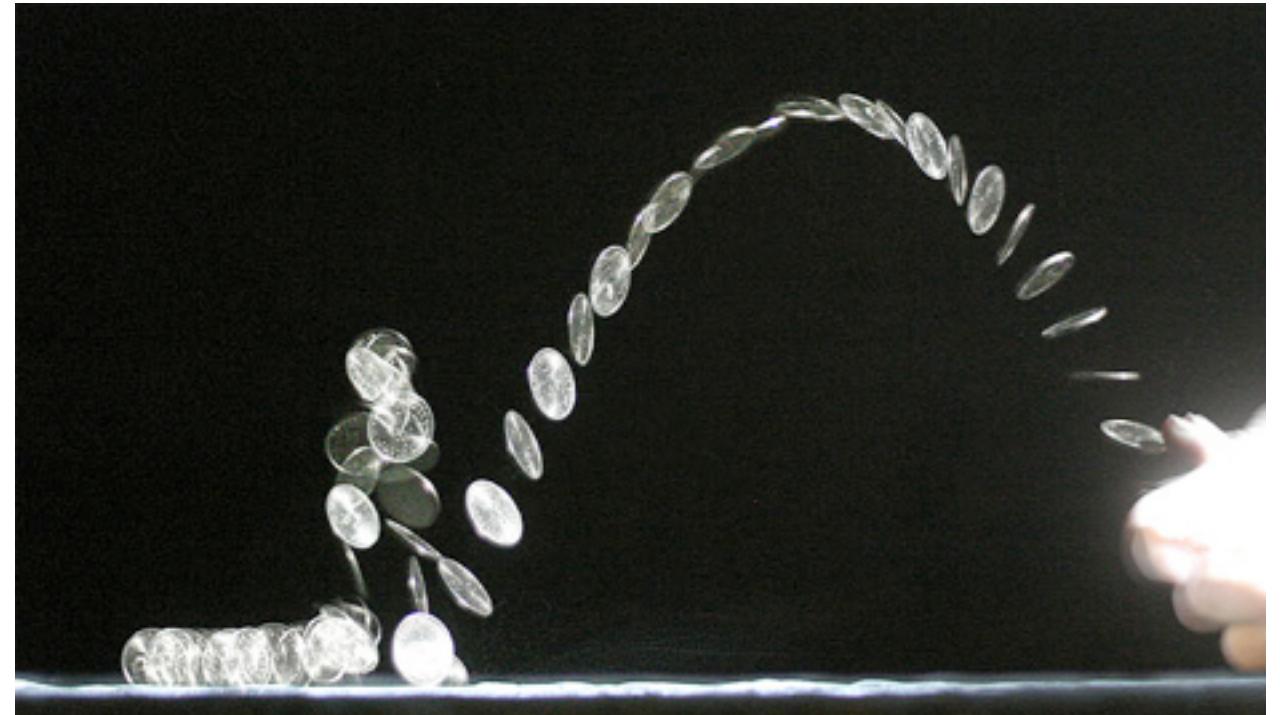
Подбрасывание монетки

- Случайный ли результат?



Подбрасывание монетки

- Случайный ли результат?
- Нет!
- Зависит от:
 - параметров броска
 - свойств монетки
 - свойств воздуха
 - ...
- Очень сложно описать с помощью формул



Случайность

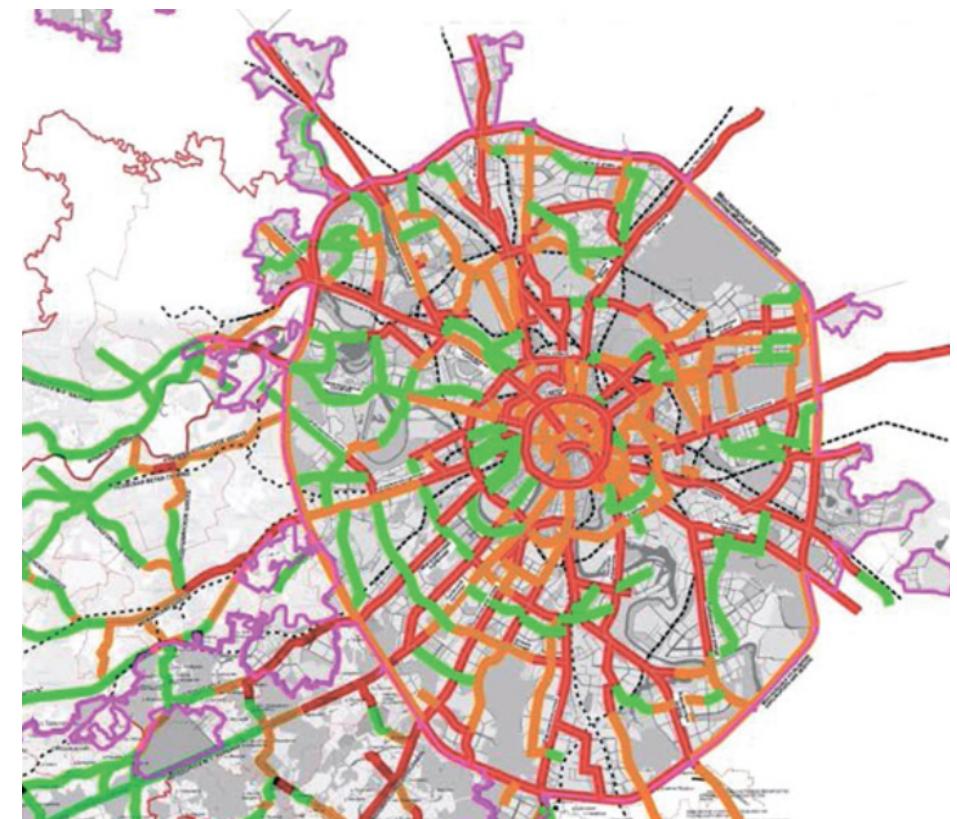
- Большинство процессов в мире можно описать уравнениями
- (кроме квантовых)
- Это может быть слишком сложно
- Проще считать исход случайным

Статистическая физика

- В комнате очень много молекул
 - $\sim 10^{30}$ штук
- Поведение каждой легко описать
- Поведение воздуха в комнате — 10^{30} уравнений
- Проще объявить состояние воздуха случайным

Транспортные потоки

- Сотни тысяч машин в Москве
- Можно пытаться проследить траекторию каждой машины
- Проще считать положение машины случайным
- Распределение машин
- Много зависимостей



Случайность

- Случайная величина — функция, выдающая случайные значения
- Изучаем **вероятности событий**
- Какова вероятность того, что случайная величина выдаст конкретное значение?
- В какой доле случаев она примет это значение?

Дискретные случайные величины

Дискретная случайная величина

- Принимает конечное или счетное число значений
- Возможные значения: $\{a_1, a_2, a_3, \dots\}$
- Вероятности: p_1, p_2, p_3, \dots
- Из свойств вероятностей: $\sum_{i=1}^{\infty} p_i = 1$
- $P(X = a_i) = p_i$ — функция вероятности

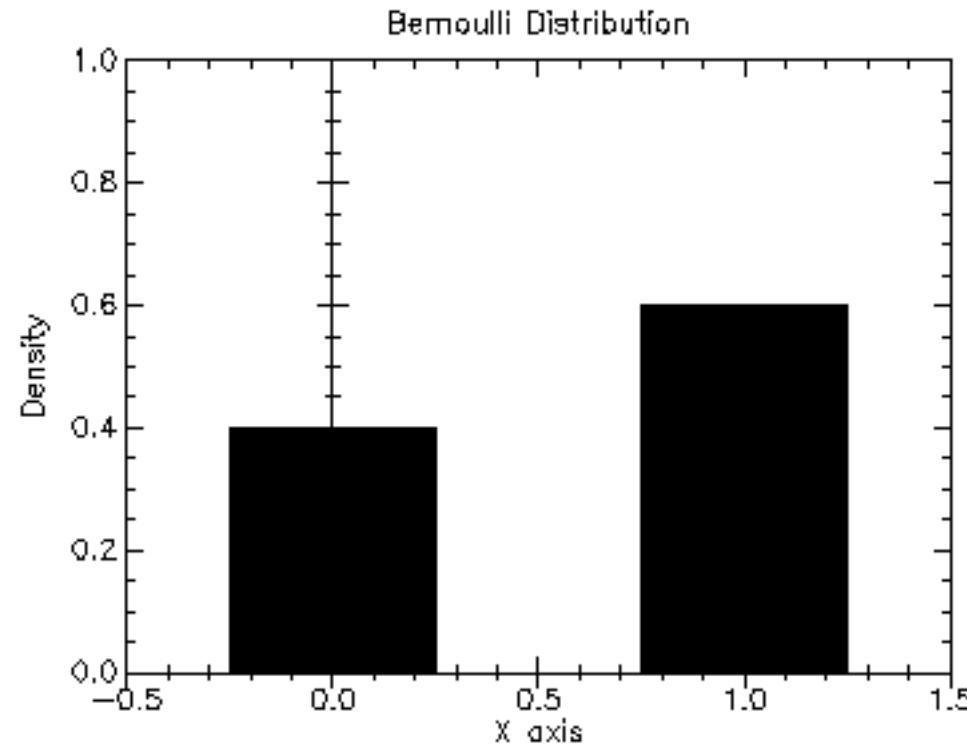
Распределение Бернулли

- Студент проходит тест без подготовки (одна задача)
- $\xi = 1$, если ответ правильный
- $\xi = 0$, если ответ неправильный
- $\xi \sim \text{Ber}(p)$
- $P(\xi = 1) = p$
- $P(\xi = 0) = 1 - p$
- Если 10 вариантов ответа, то $p = 0.1$



Распределение Бернуlli

- Гистограмма:



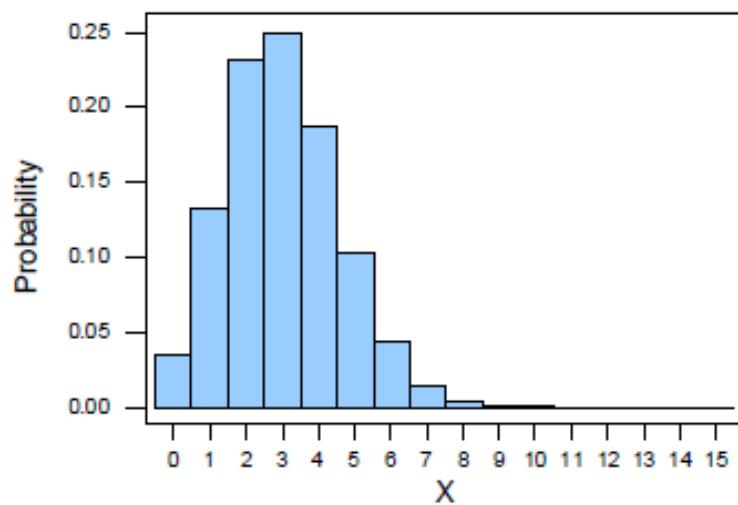
Биномиальное распределение

- Студент проходит тест без подготовки (n задач)
- Сколько задач он решил?
- $\xi_i \sim \text{Ber}(p)$ — решил ли i -ю задачу
- $\eta = \sum_{i=1}^n \xi_i \sim \text{Bin}(n, p)$

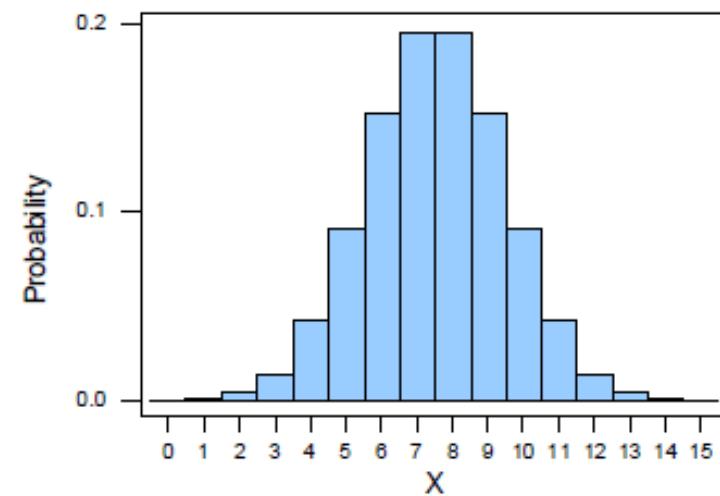
- $P(\eta = k) = C_n^k p^k (1 - p)^{n-k}$

Биномиальное распределение

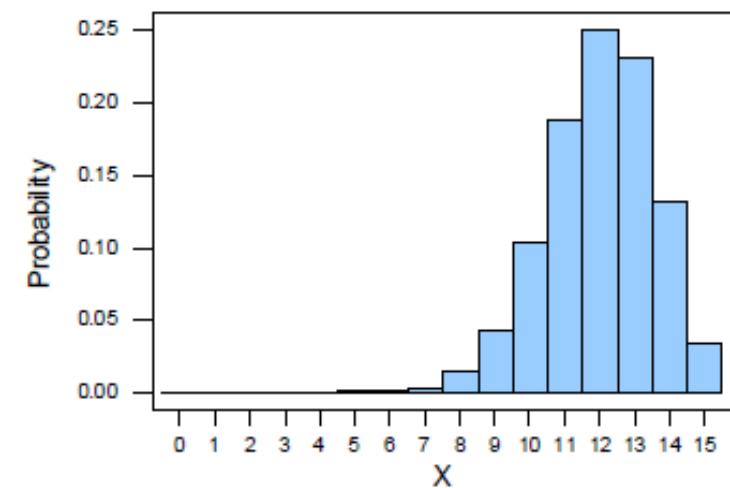
Binomial distribution with $n = 15$ and $p = 0.2$



Binomial distribution with $n = 15$ and $p = 0.5$



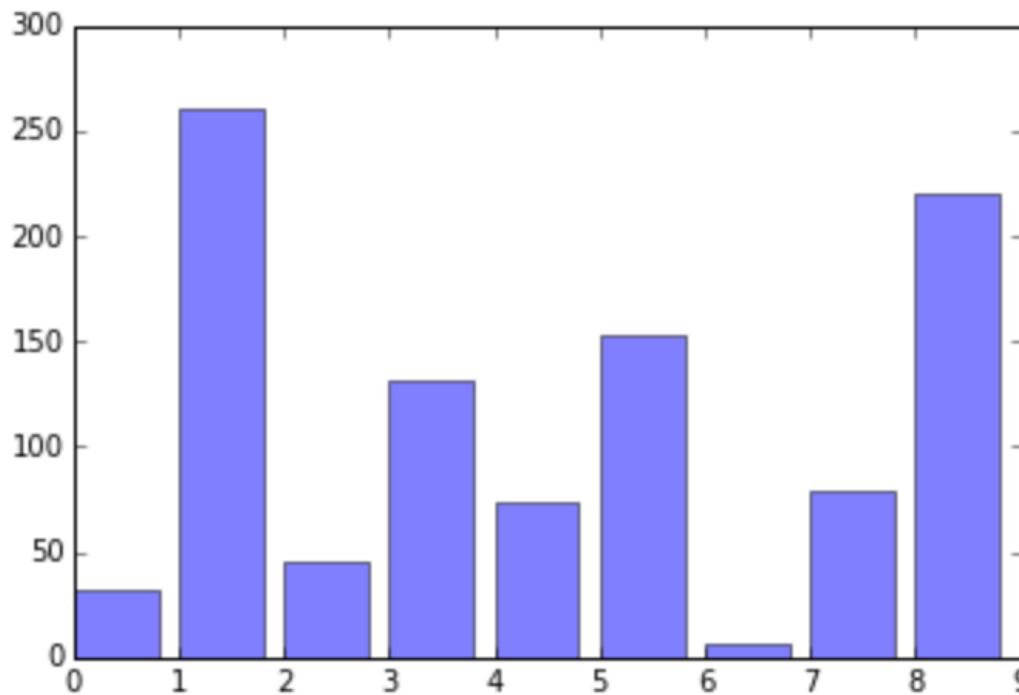
Binomial distribution with $n = 15$ and $p = 0.8$



Мультиномиальное распределение

- Студенты выбирают майнор (n вариантов)
- $\xi_i = (1, 0, \dots, 0)$, если i -й студент выбрал первый майнор
- $P(\xi_{ij} = 1) = p_j$ — вероятность выбрать j -й майнор
- $\eta = \sum_{i=1}^n \xi_i \sim \text{Mult}(n, p)$

Мультиномиальное распределение



Распределение Пуассона

то есть спустя три года, скромницей, с чемоданчиком балерины в руке, затем — шестнадцати лет, в пачках, с газовыми крыльцами за спиной, вольно сидящей на столе, с поднятым бокалом, среди бледных гуляк, затем — лет восемнадцати, в фатальном трауре, у перил над каскадом, затем... ах, во многих еще видах и позах, вплоть до самой последней — лежачей.

При помощи ретушировки и других фотофокусов как будто достигалось последовательное изменение лица Эммочки (искусник, между прочим, пользовался фотографиями ее матери), но стоило взглянуть ближе, и становилась безобразно ясной аляповатость этой пародии на работу времени. У Эммочки, выходившей из театра в мехах с цветами, прижатыми к плечу, были ноги, никогда не плясавшие; а на следующем снимке, изображавшем ее уже в венчальной дымке, стоял рядом с ней жених, стройный и высокий, но с кругленькой физиономией м-сье Пьера. В тридцать лет у нее появились условные морщины, проведенные без смысла, без жизни, без знания их истинного значения, — но знатоку говорящие совсем странное, как бывает, что случайное движение ветвей совпадает с жестом, понятным для глухонемого. А в сорок лет

Распределение Пуассона

то есть спустя три года, скромницей, с чемоданчиком балерины в руке, затем — шестнадцати лет, в пачках, с газовыми крыльцами за спиной, вольно сидящей на столе, с поднятым бокалом, среди бледных гуляк, затем — лет восемнадцати, в фатальном трауре, у перил над каскадом, затем... ах, во многих еще видах и позах, вплоть до самой последней — лежачей.

При помощи ретушировки и других фотофокусов как будто достигалось последовательное изменение лица Эммочки (искусник, между прочим, пользовался фотографиями ее матери), но стоило взглянуть ближе, и становилось 1 раз бразно ясной аляповатость этой пародии на работу времени. У Эммочки, выходившей из театра в мехах с цветами, прижатыми к плечу, были ноги, никогда не плясавшие; а на следующем снимке, изображавшем ее уже в венчальной дымке, стоял рядом с ней жених, стройный и высокий, но с кругленькой физиономией м-сье Пьера. В тридцать лет у нее появились условные морщины, проведенные без смысла, без жизни, без знания их истинного значения, — но знатоку говорящие совсем странное, как бывает, что случайное движение ветвей совпадает с жестом, понятным для глухонемого. А в сорок лет

Распределение Пуассона

то есть спустя три **года**, скромницей, с чемоданчиком балерины в руке, затем — шестнадцати **лет**, в пачках, с газовыми крыльцами за спиной, вольно сидящей на столе, с поднятым бокалом, среди бледных гуляк, затем — **лет** восемнадцати, в фатальном трауре, у перил над каскадом, затем... ах, во многих еще видах и позах, вплоть до самой последней — лежачей.

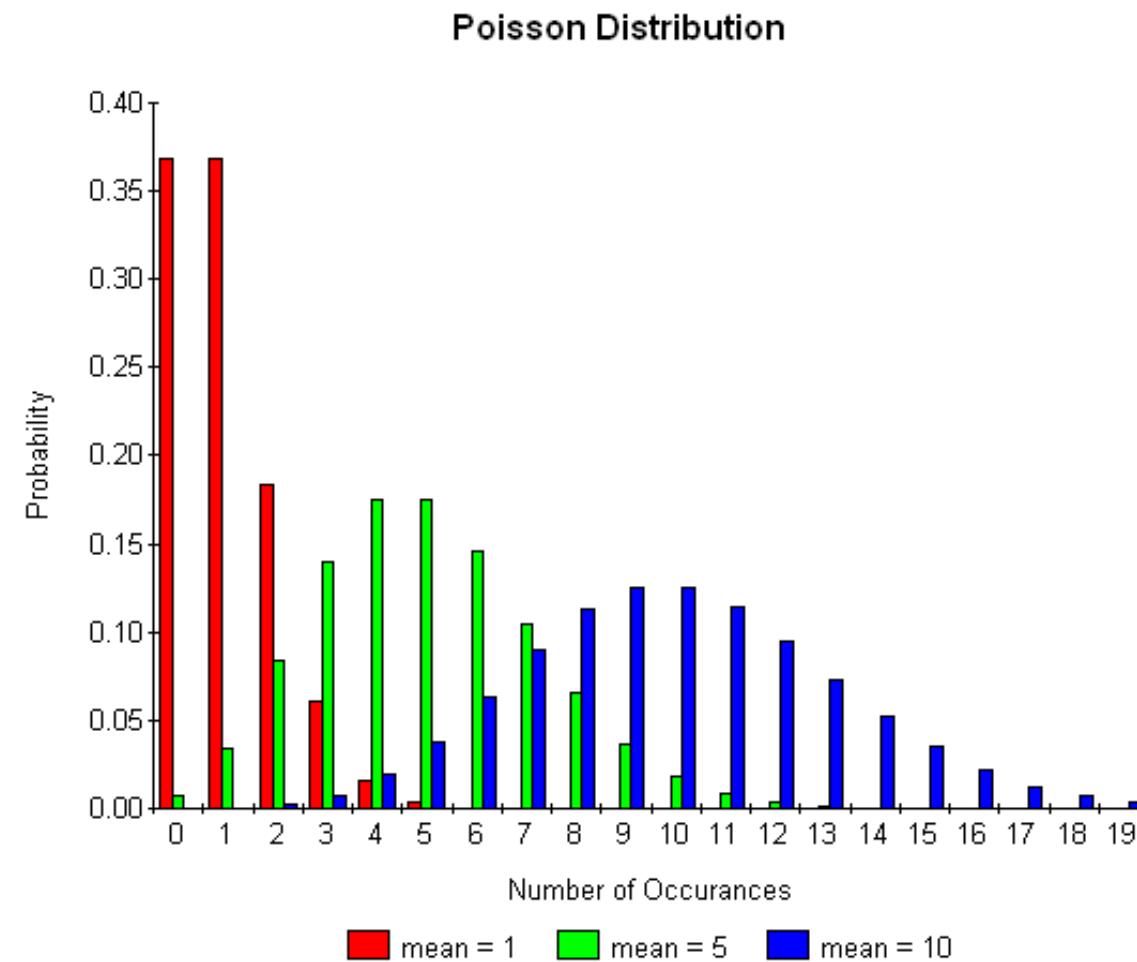
При помощи ретушировки и других фотофокусов как будто достигалось последовательное изменение лица Эммочки (искусник, между прочим, пользовался фотографиями ее матери), но стоило взглянуть ближе, и становилось **21 раз** бразно ясной аляповатость этой пародии на работу времени. У Эммочки, выходившей из театра в мехах с цветами, прижатыми к плечу, были ноги, никогда не плясавшие; а на следующем снимке, изображавшем ее уже в венчальной дымке, стоял рядом с ней жених, стройный и высокий, но с кругленькой физиономией м-сье Пьера. В тридцать **лет** у нее появились условные морщины, проведенные без смысла, без жизни, без знания их истинного значения, — но знатоку говорящие совсем странное, как бывает, что случайное движение ветвей совпадает с жестом, понятным для глухонемого. А в сорок **лет**

Распределение Пуассона

- ξ_w — число использований слова w в тексте
- $P(\xi_w = k)$ — вероятность того, что слово w встретится k раз

$$P(\xi_w = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda > 0$$

Распределение Пуассона



Распределение Пуассона

- Подходит для моделирования редких событий
- Пример: количество покупателей в магазине каждую минуту
- Свойство отсутствия памяти
- Для количества изюминок в булочках с изюмом тоже подходит

Непрерывные распределения

Непрерывная случайная величина

- Принимает континуум или больше значений
- Пример: отклонение времени начала лекции от 10:30 (в минутах)
- $P(\xi = 2) = ?$
- $P(\xi = 2.1) = ?$
- $P(\xi = 2.18) = ?$
- $P(\xi = 2.187) = ?$

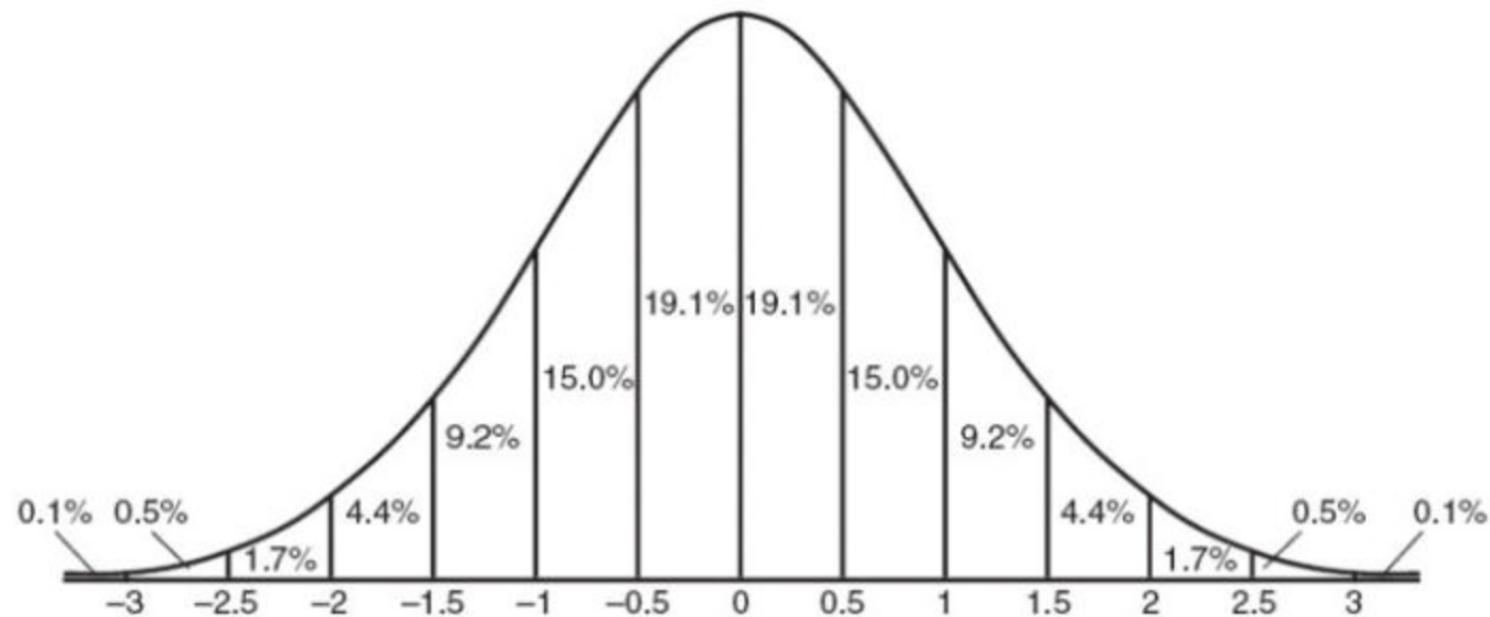
Непрерывная случайная величина

- Принимает континуум или больше значений
- Пример: отклонение времени начала лекции от 10:30 (в минутах)
- $P(\xi = 2) = 0$
- $P(\xi = 2.1) = 0$
- $P(\xi = 2.18) = 0$
- $P(\xi = 2.187) = 0$
- Вероятность каждого элементарного исхода равна нулю!

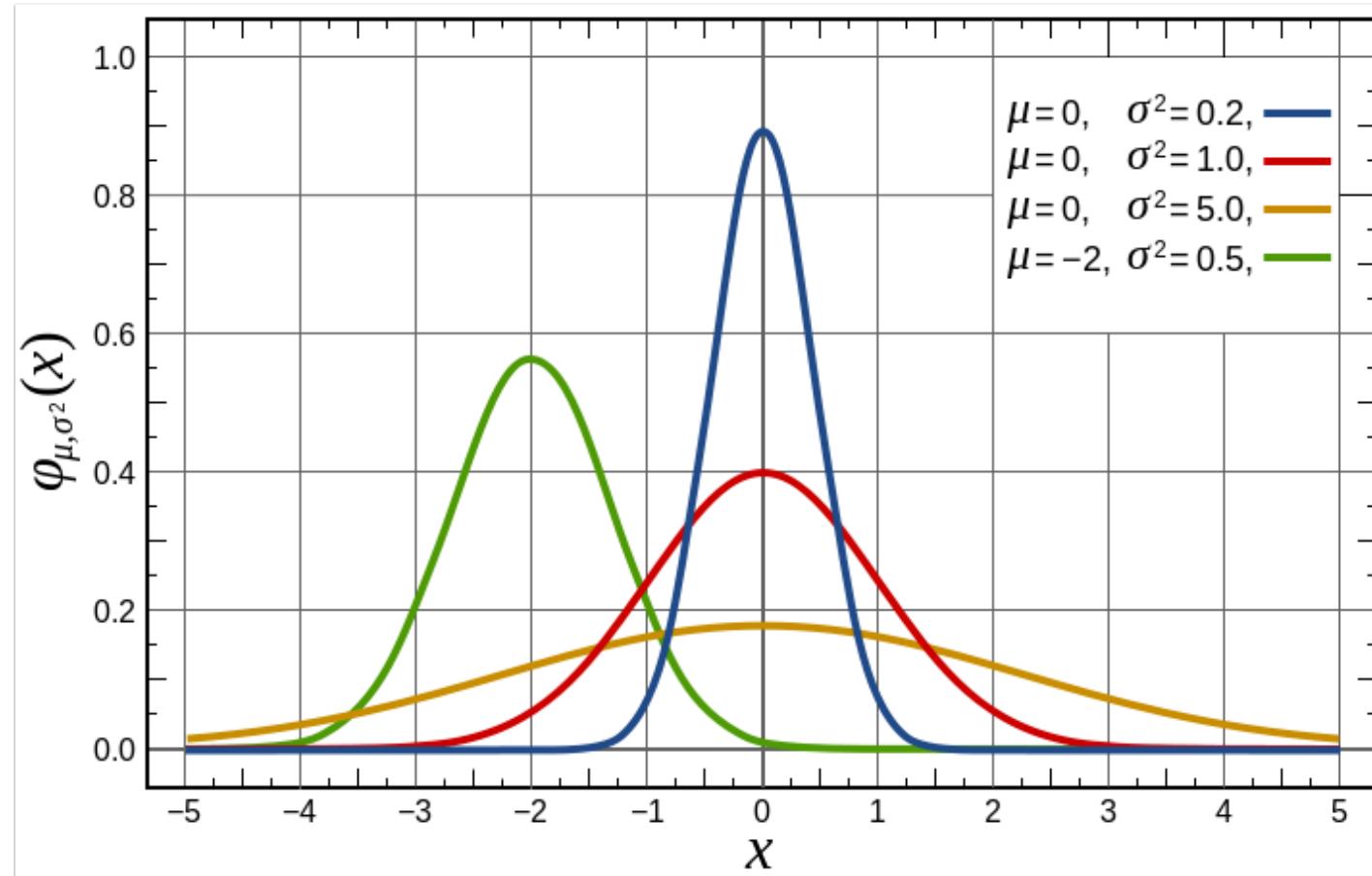
Плотность распределения

Плотность

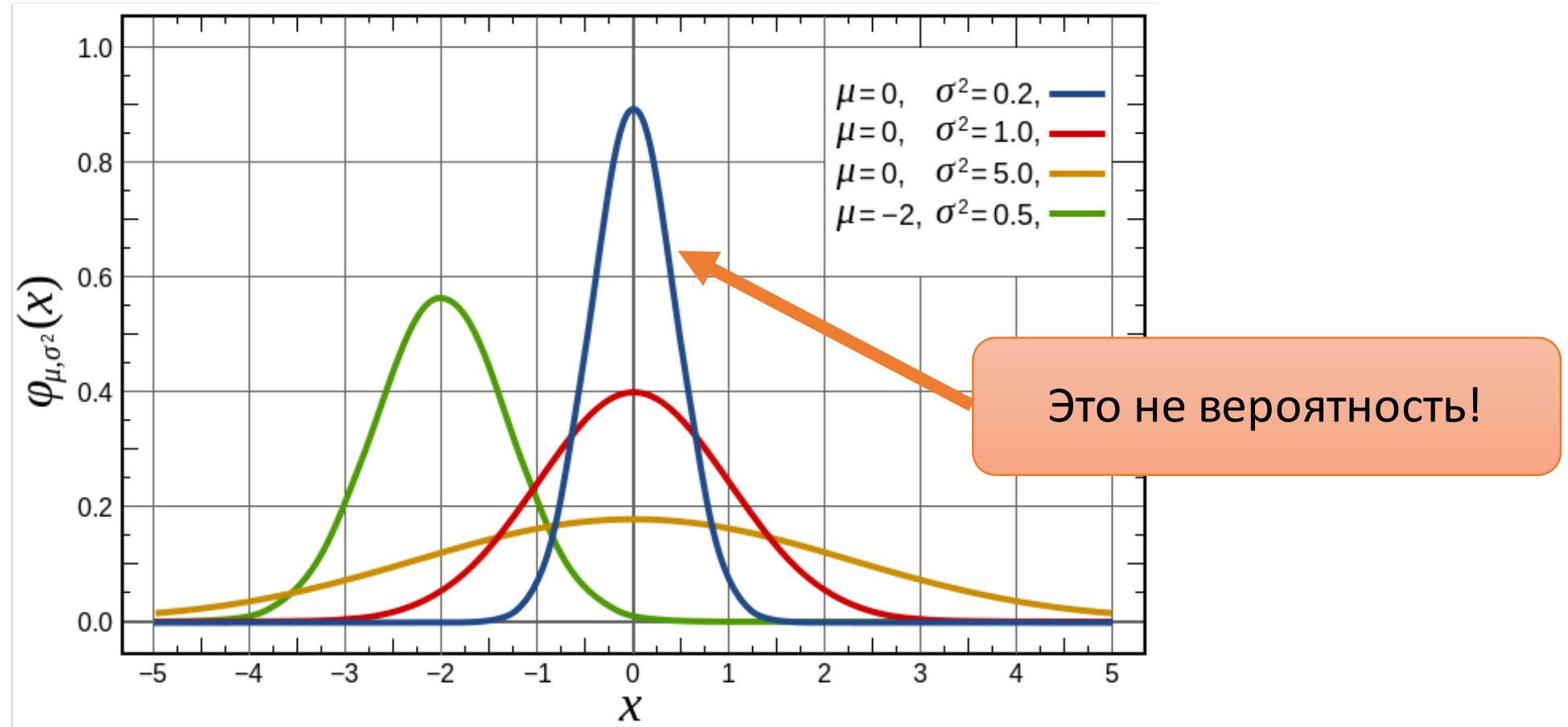
$$P(a \leq \xi \leq b) = \int_a^b p(x)dx$$



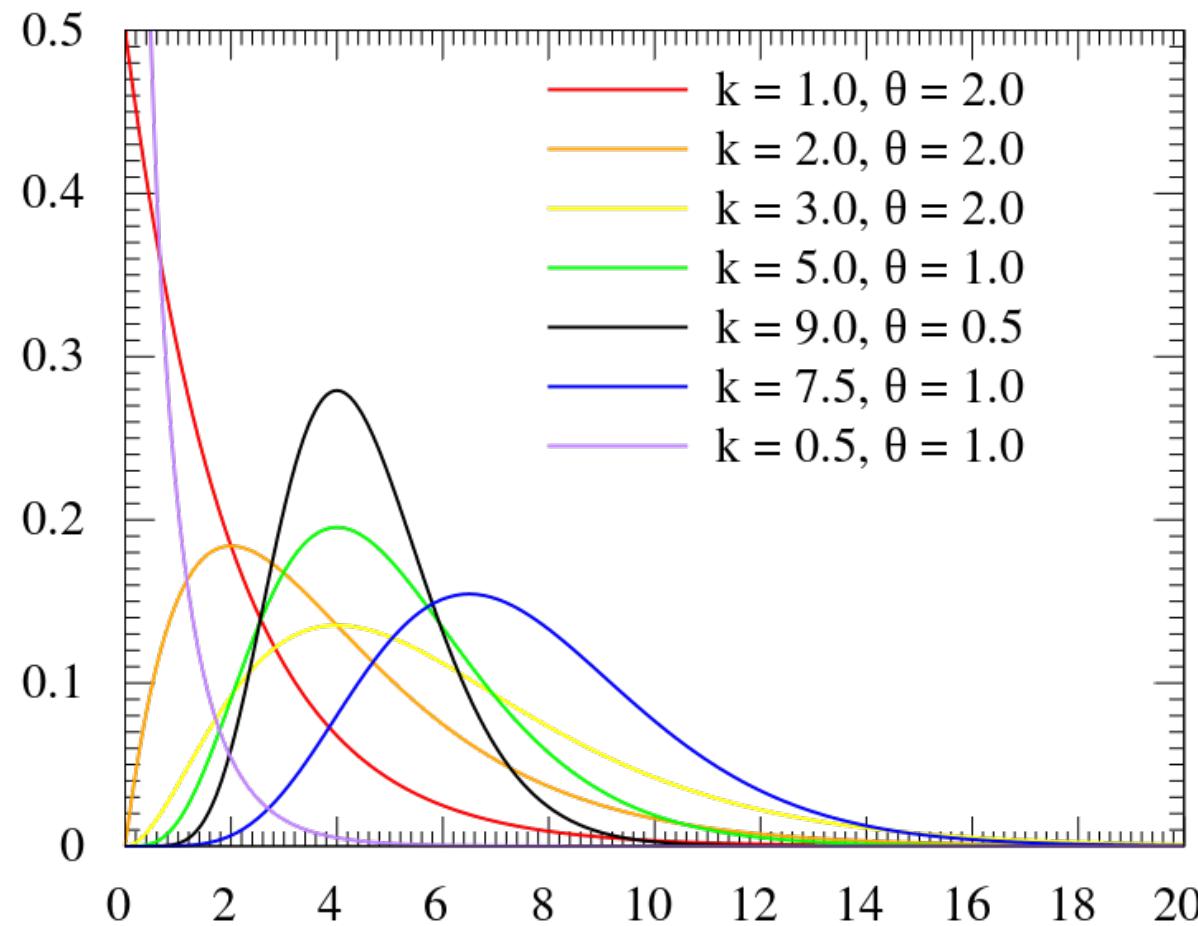
Плотность распределений



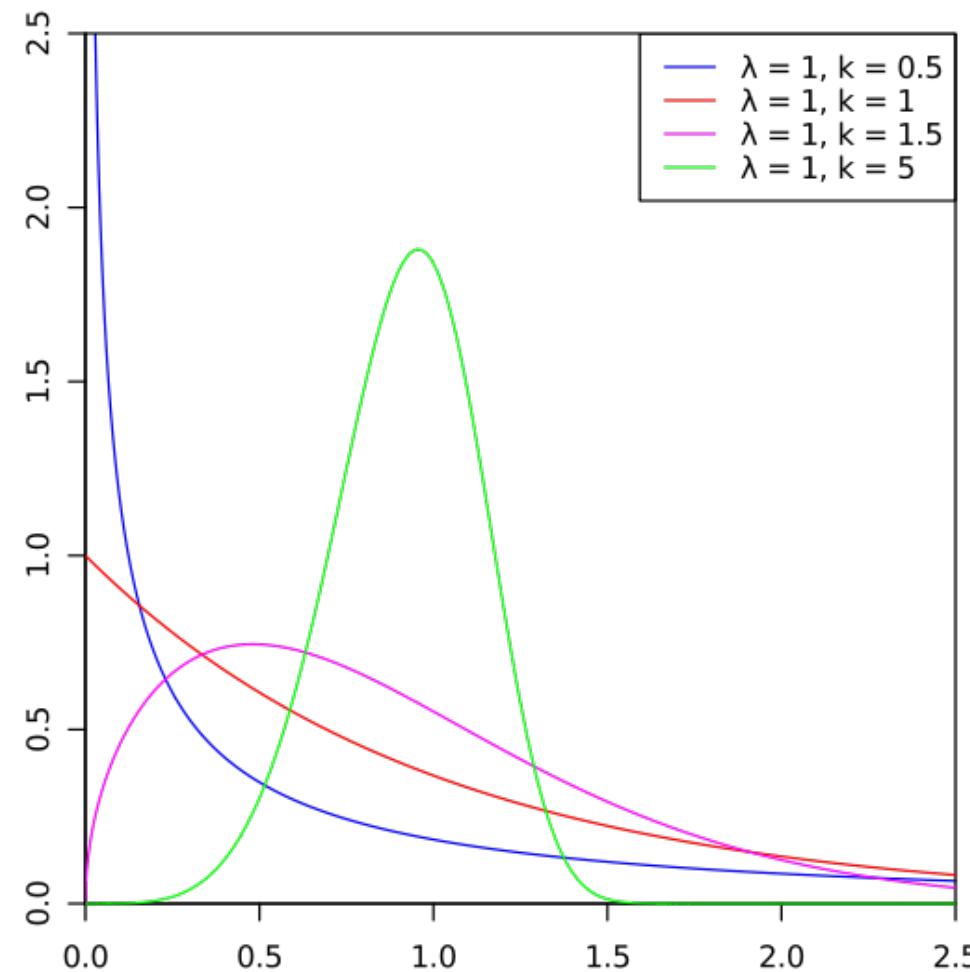
Плотность распределений



Плотность распределений



Плотность распределений

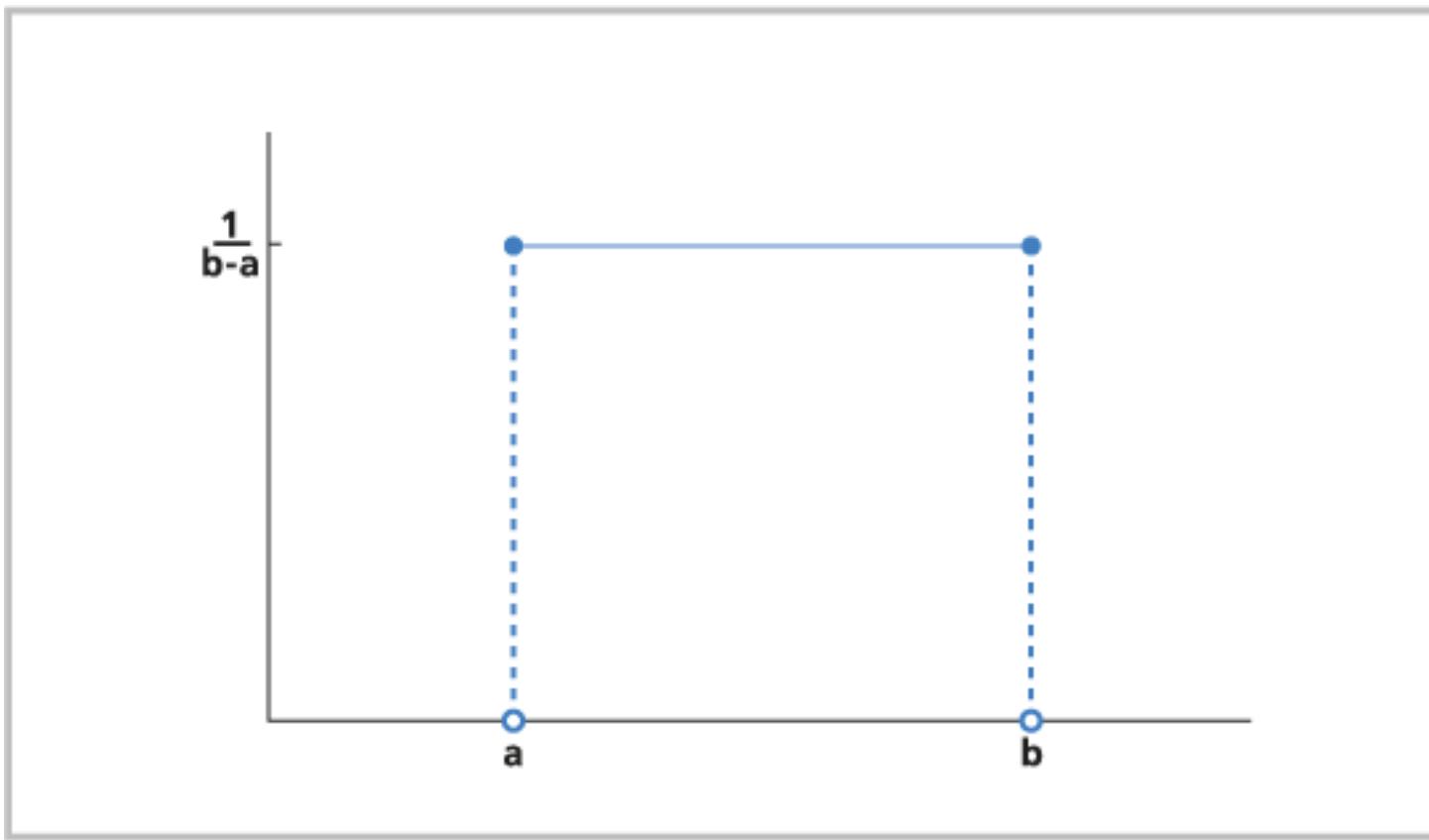


Равномерное распределение

- Носитель (множество с ненулевой плотностью): $[a, b]$
- $\xi \sim R[a, b]$

$$p(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{иначе} \end{cases}$$

Равномерое распределение



Равномерное распределение

- Автобус приходит каждые 5 минут
- Человек приходит в случайный момент на остановку
- Сколько ему придется ждать?
- $\xi \sim R[0, 5]$
- $P(\xi \geq 3) = \frac{2}{5}$

Равномерное распределение

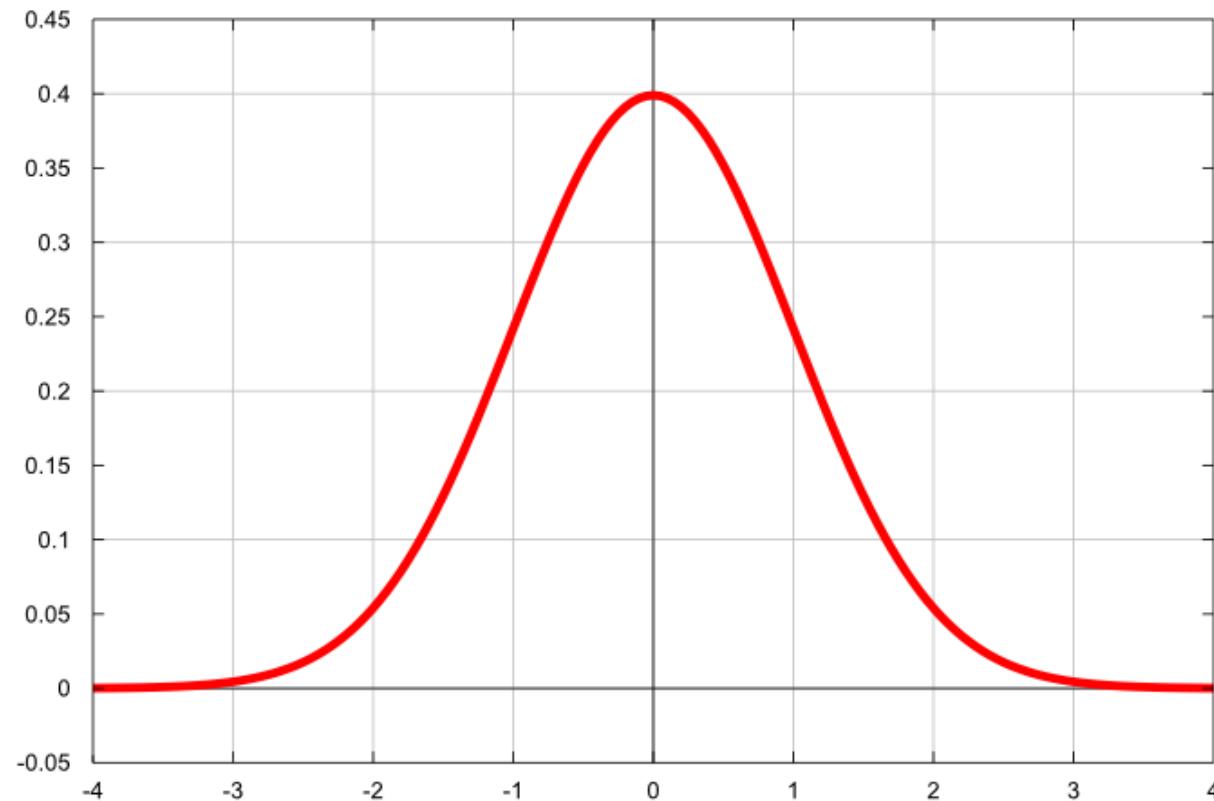
- Не очень распространено
- Легко эмулировать на компьютере
- Позволяет генерировать числа из любого распределения

Нормальное распределение

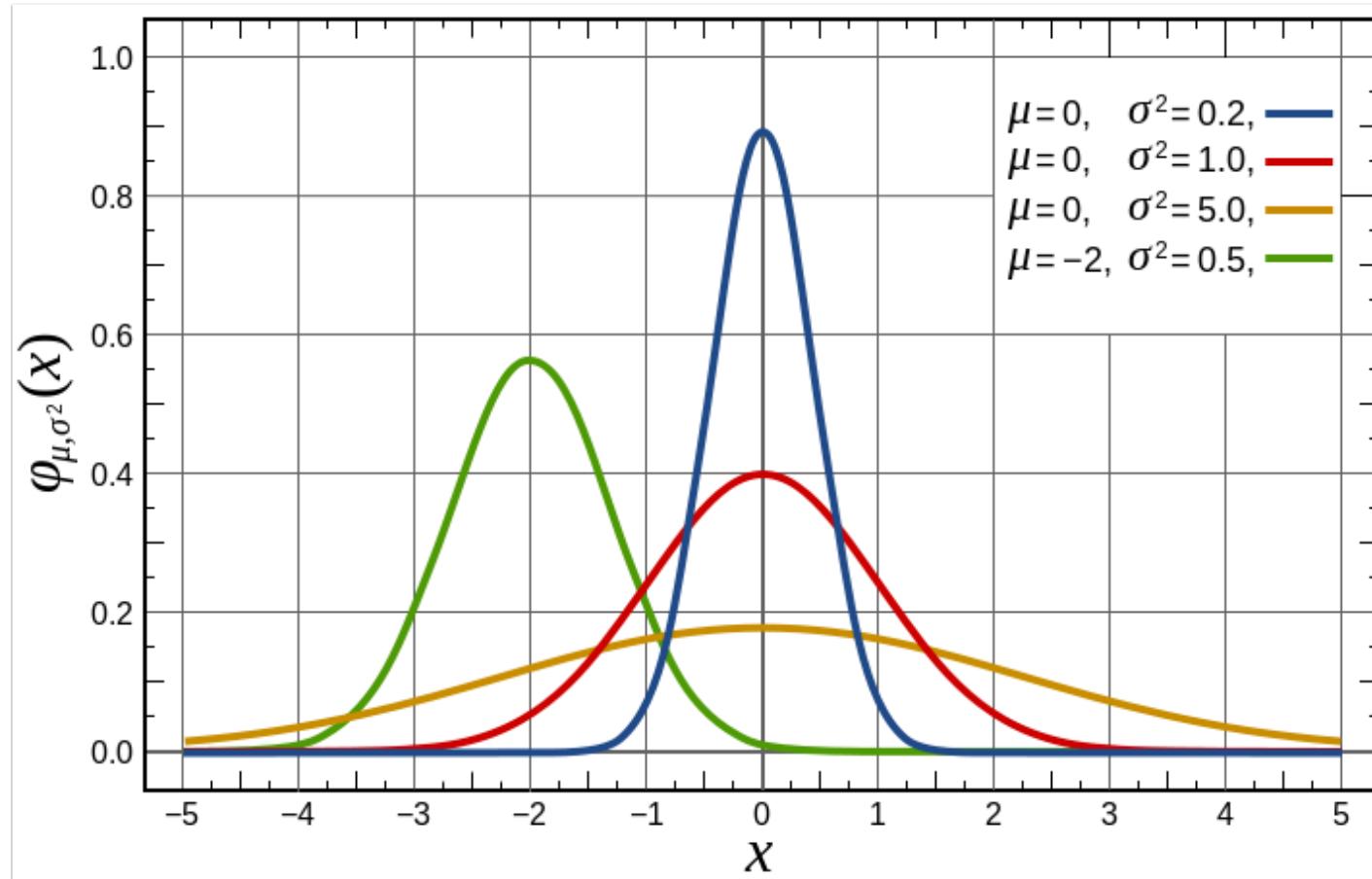
- Носитель: \mathbb{R}
- $\xi \sim N(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Нормальное распределение



Нормальное распределение



Нормальное распределение

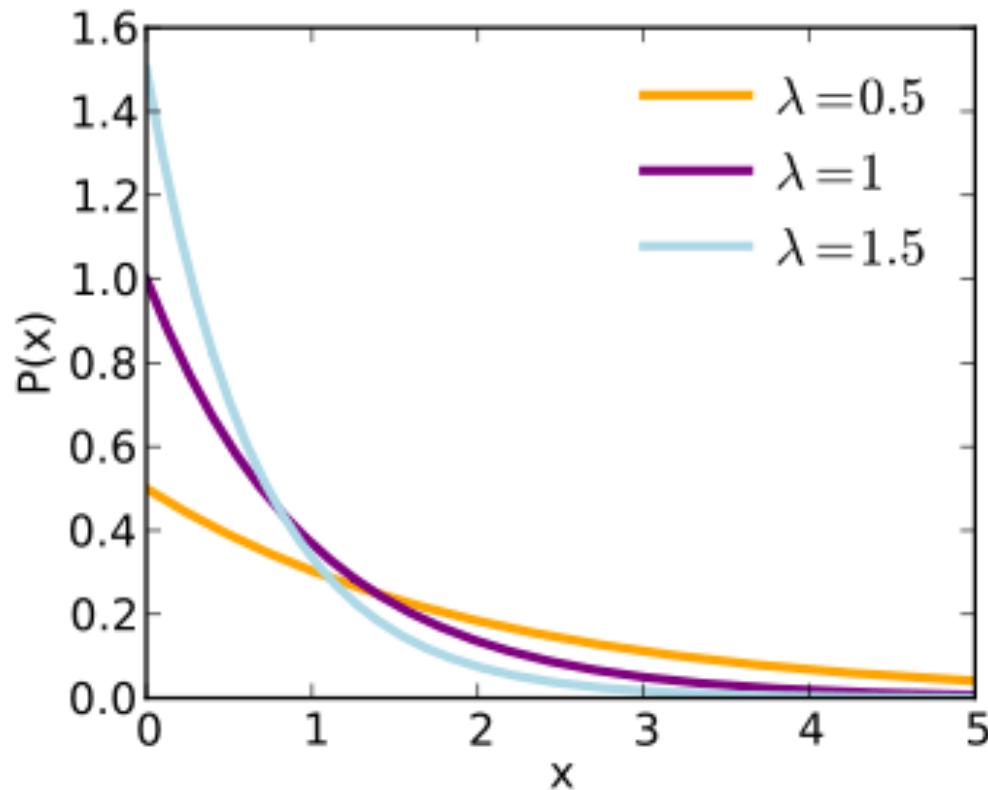


Экспоненциальное распределение

- Носитель: $[0, +\infty]$
- $\xi \sim \exp(\lambda)$

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Экспоненциальное распределение



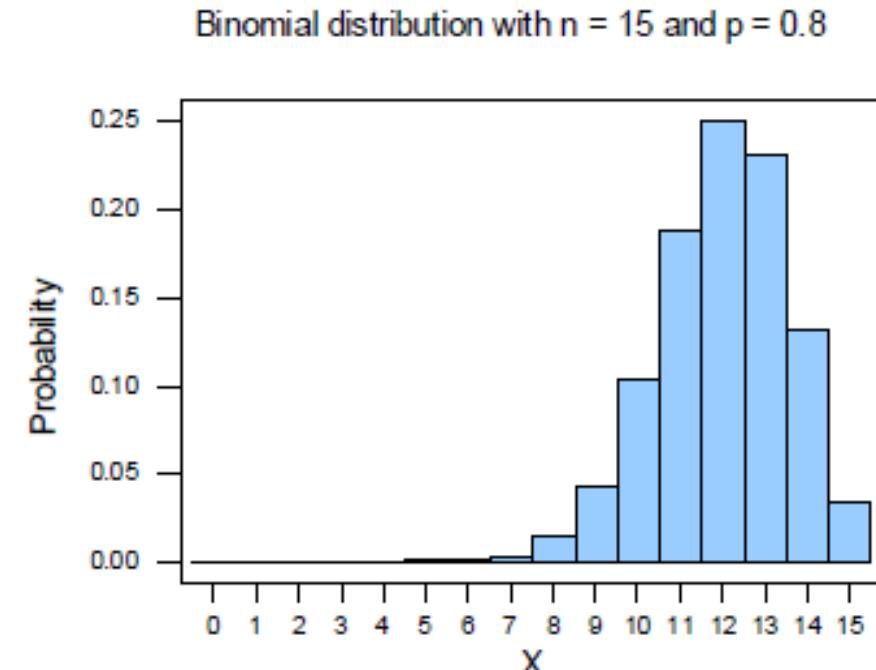
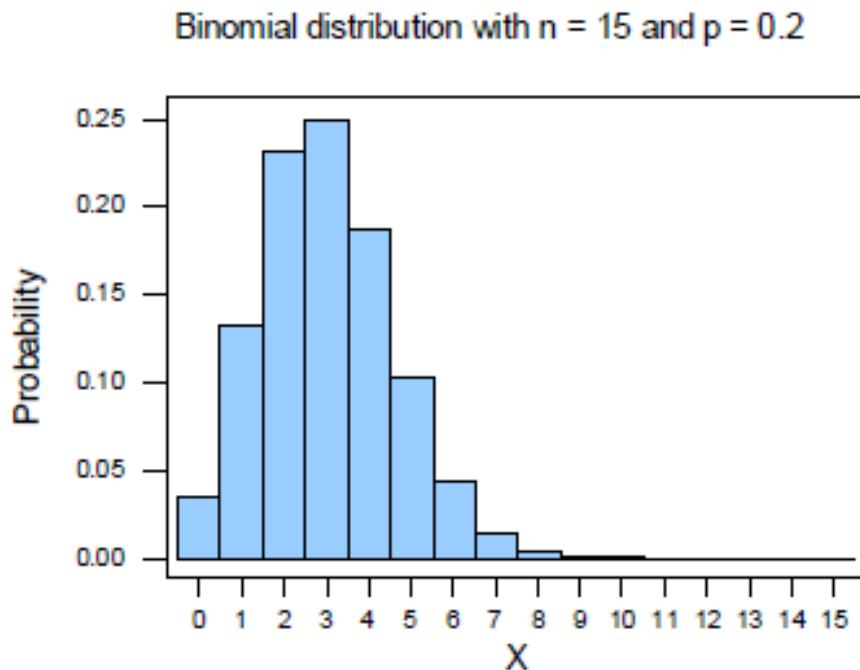
Экспоненциальное распределение

- Моделирует расстояние между редкими событиями
- Время до следующего звонка в колл-центр
- Время до следующего вопроса студента на лекции
- Расстояние между двумя соседними мутациями в ДНК

Характеристики случайных величин

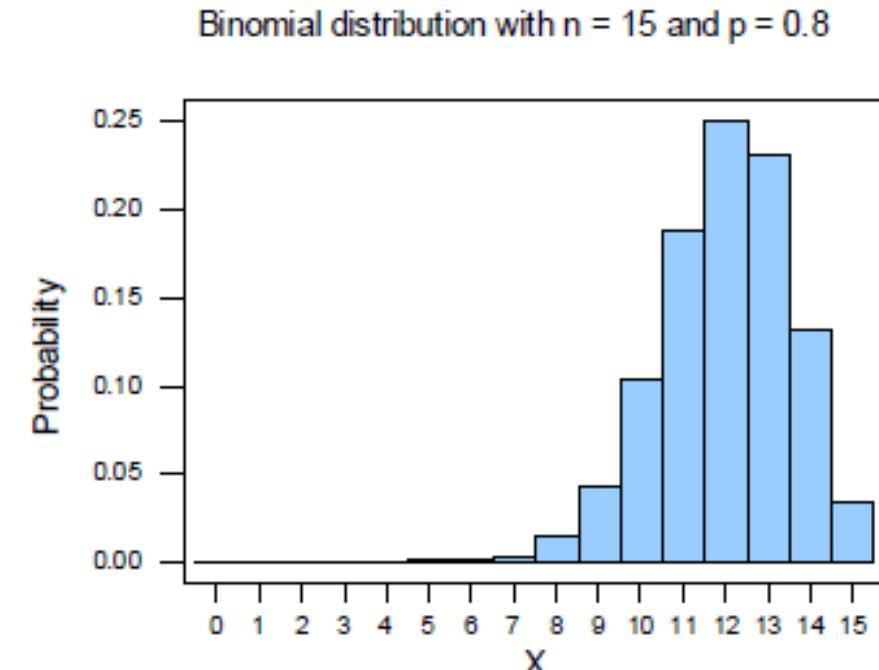
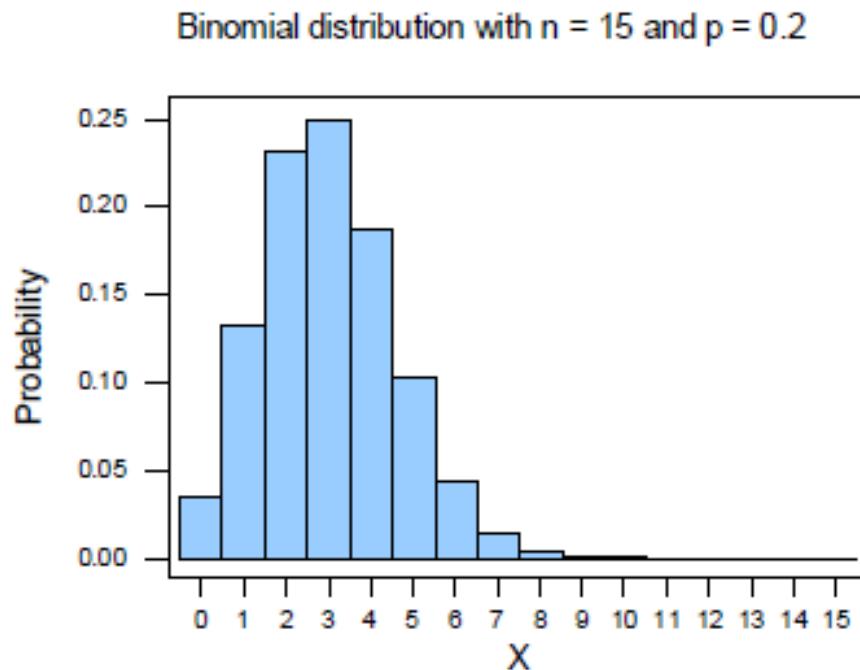
Среднее значение

- Студент правильно отвечает на один вопрос с вероятностью p
- На сколько вопросов он ответит, если всего их n ?
- $\xi \sim \text{Bin}(n, p)$



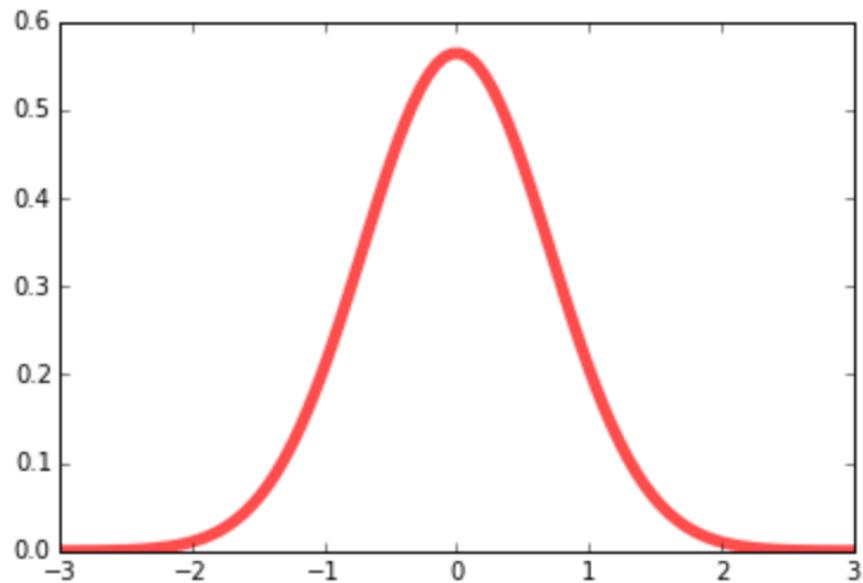
Среднее значение

- На сколько вопросов в среднем будут отвечать такие студенты?
- Ответ: 3 и 12

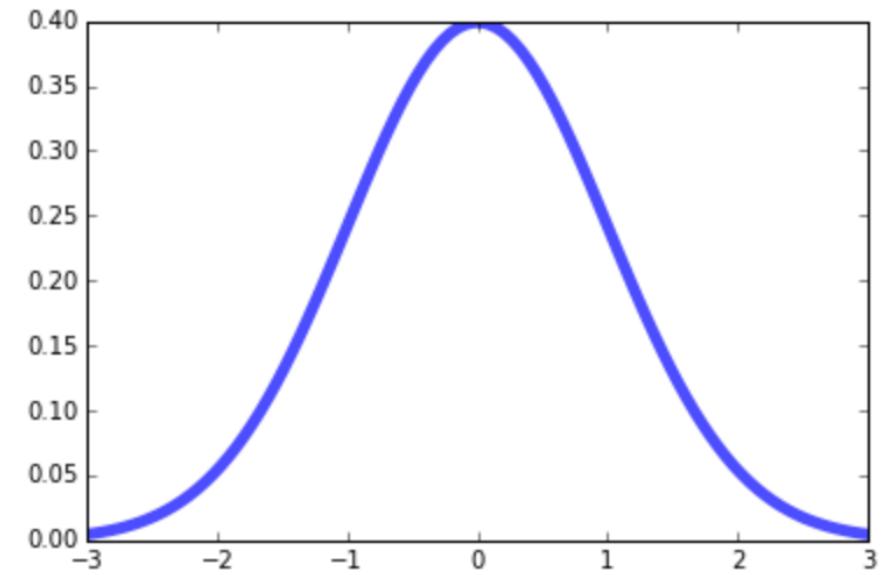


Разброс

- Отклонение времени начала лекции от официального начала
- $\xi \sim N(\mu, \sigma^2)$



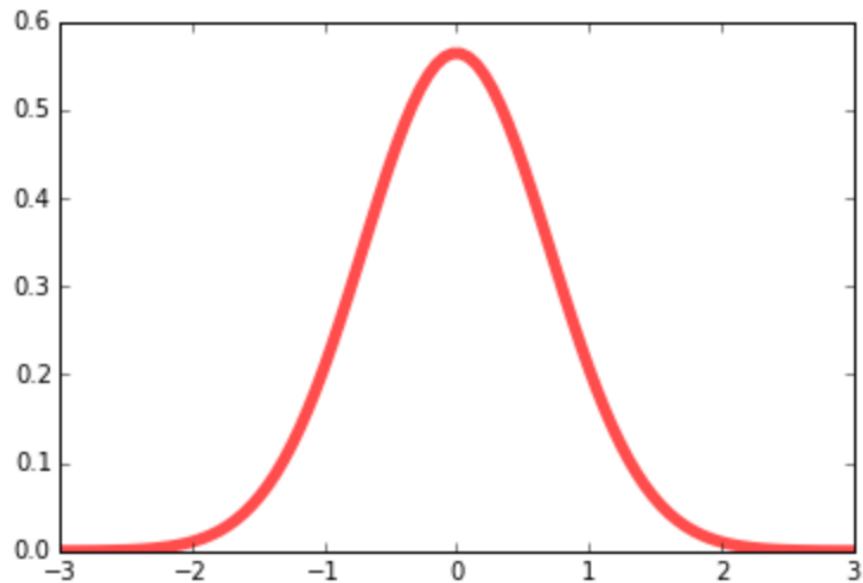
Лекции 1-й парой



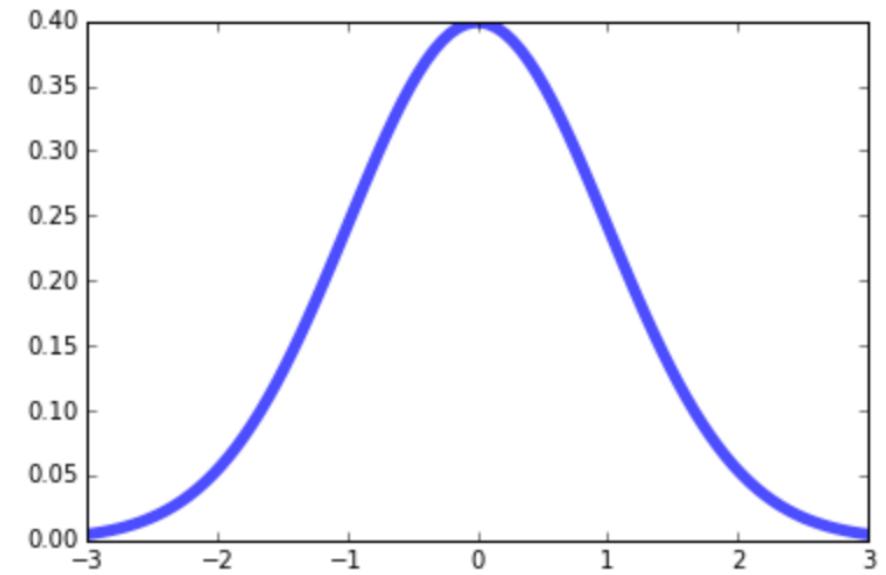
Лекции 2-й парой

Разброс

- Разброс на второй паре выше!



Лекции 1-й парой



Лекции 2-й парой

Математическое ожидание

- Характеризует среднее значение случайной величины

$$\mathbb{E}\xi = \begin{cases} \sum_{i=1}^n x_i p_i, & \text{для дискретных величин} \\ \int_{-\infty}^{+\infty} x p(x) dx, & \text{для непрерывных величин} \end{cases}$$

Математическое ожидание

- Для $\text{Pois}(\lambda)$: $\mathbb{E}\xi = \lambda$
- Для $\text{Bin}(n, p)$: $\mathbb{E}\xi = np$
- Для $R[a, b]$: $\mathbb{E}\xi = (a + b)/2$
- Для $N(\mu, \sigma^2)$: $\mathbb{E}\xi = \mu$

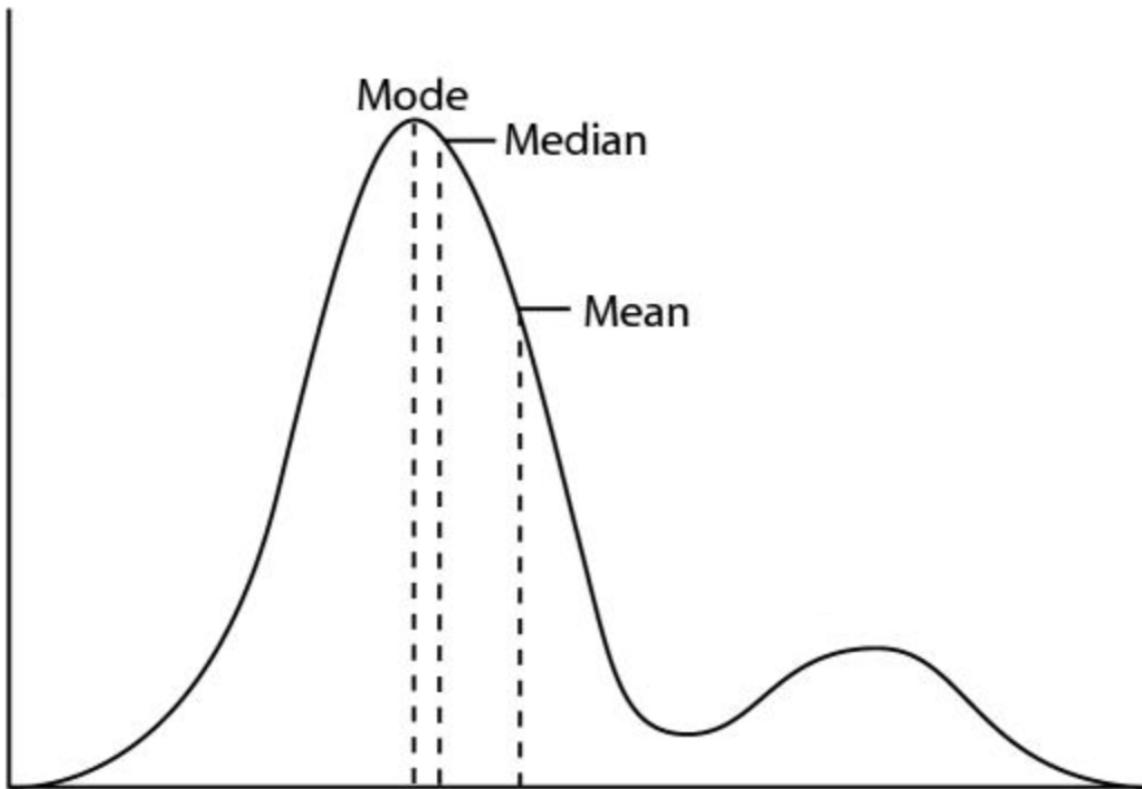
Медиана

- Такое число m , что попасть левее и правее — равновероятно
- $P(\xi \leq m) \geq 0.5$ и $P(\xi \geq m) \geq 0.5$

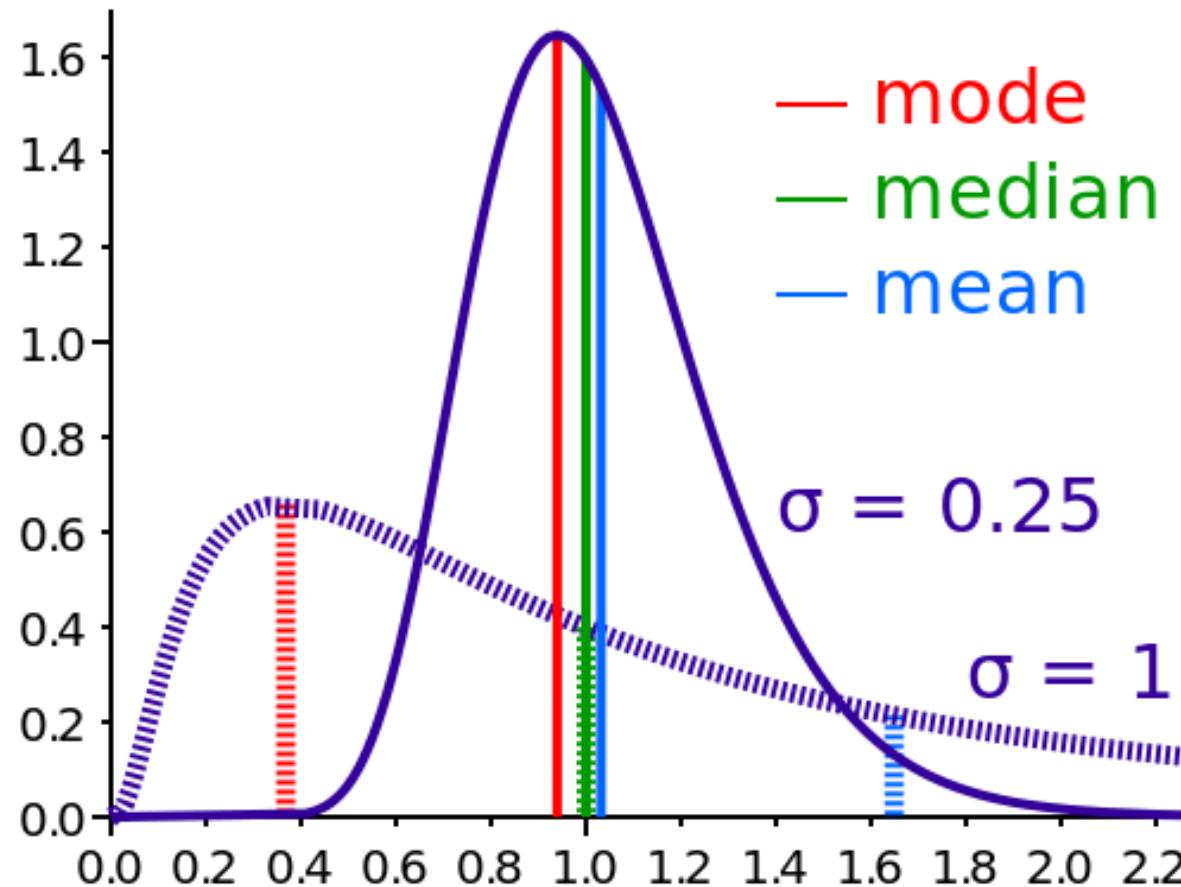
Мода

- Для дискретных величин: точка с максимальной вероятностью
- Для непрерывных величин: точка максимума плотности

Средняя величина



Средняя величина

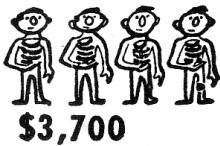


В чем разница?

- Опросили 100 человек
- 99 имеют доход 10.000 рублей
- 1 имеет доход 1.000.000 рублей
- Среднее: $\frac{99 \cdot 10000 + 1000000}{100} = 19900$
- Медиана: 10000
- Мода: 10000



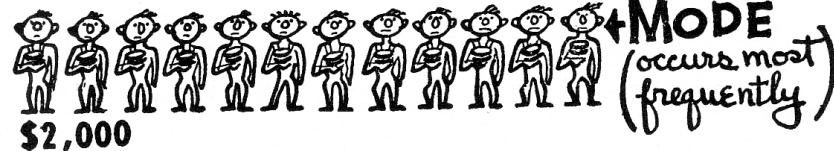
← ARITHMETICAL AVERAGE



← MEDIAN (the one in the middle)
12 above him, 12 below

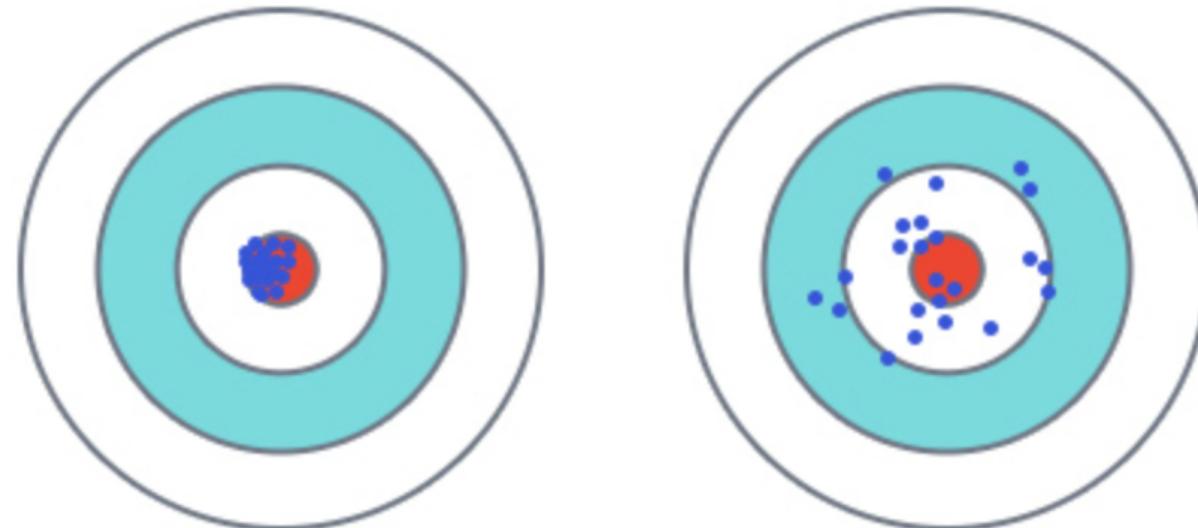


← MODE
(occurs most frequently)



Дисперсия

- Мера разброса случайной величины
- $D\xi = E(\xi - E\xi)^2$
- Стандартное отклонение: $\sqrt{D\xi}$



Дисперсия

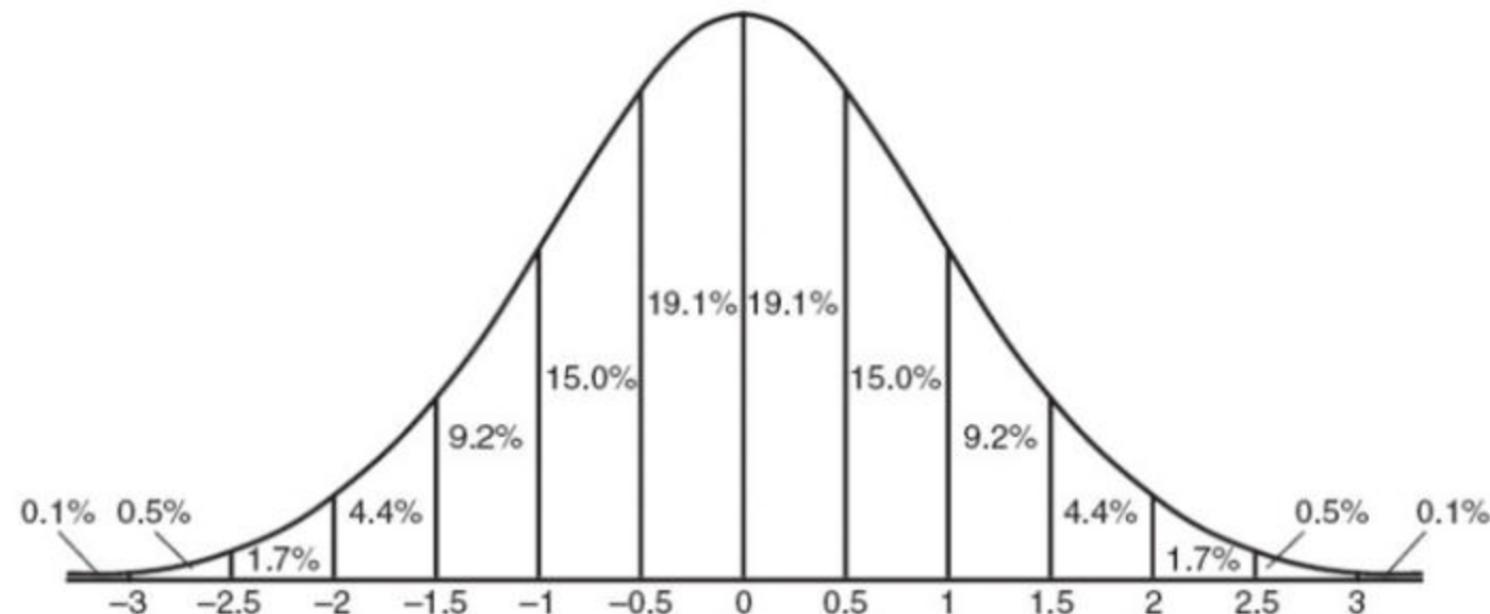
- Для $\text{Pois}(\lambda)$: $\mathbb{D}\xi = \lambda$
- Для $\text{Bin}(n, p)$: $\mathbb{D}\xi = np(1 - p)$
- Для $R[a, b]$: $\mathbb{D}\xi = (b - a)^2/12$
- Для $N(\mu, \sigma^2)$: $\mathbb{D}\xi = \sigma^2$

Дисперсия

- Опросили 100 человек
- 99 имеют доход 10.000 рублей
- 1 имеет доход 1.000.000 рублей
- Дисперсия: 9702990000
- Стандартное отклонение: ~98503
- Что-нибудь более устойчивое?

Квантиль

- Q_p — p -квантиль
- Такое число t , что вероятность попасть левее равна p
- Медиана — 0.5-квантиль



Квантиль

- $Q_{0.25}, Q_{0.75}$ — квартили
- $Q_{0.01}, \dots, Q_{0.99}$ — перцентили

Интерквартильный размах

- Устойчивая к выбросам мера разброса:

$$IQR = Q_{0.75} - Q_{0.25}$$

- В нашем примере: $IQR = 0$

ЗБЧ и ЦПТ

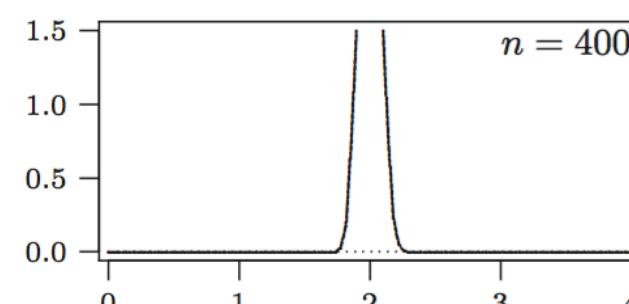
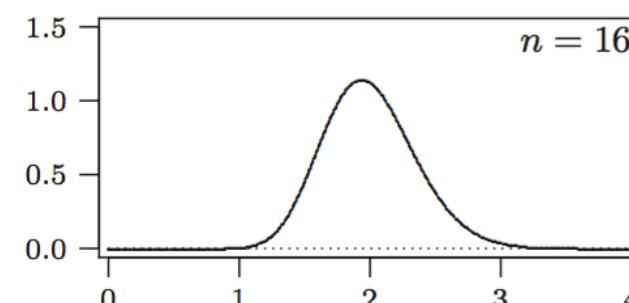
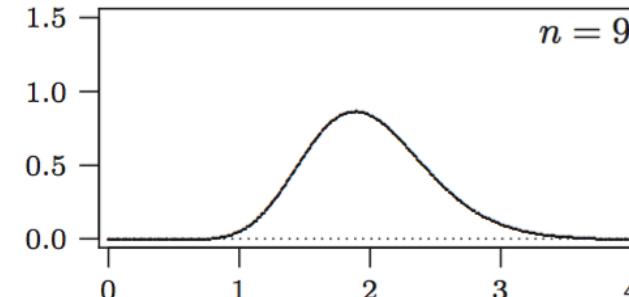
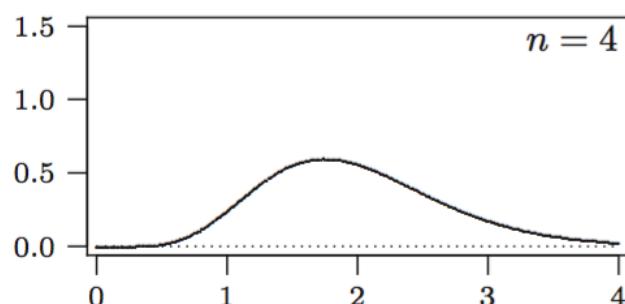
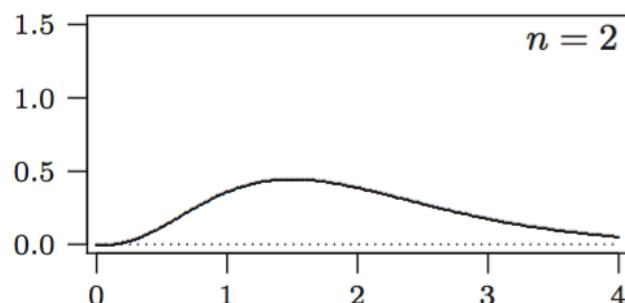
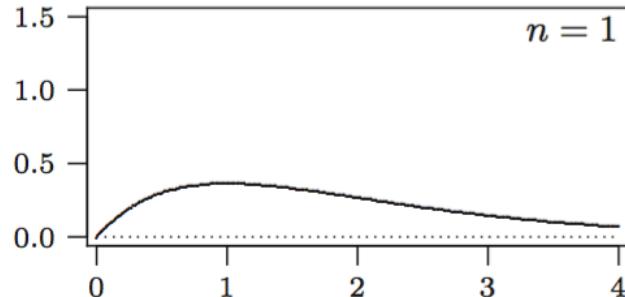
Усреднение наблюдений

- Наблюдение: усреднение результатов повышает их точность
- Измерение артериального давления
- Измерение скорости света
- Усреднение соседних пикселей изображения

Усреднение наблюдений

- $\xi_1, \xi_2, \dots, \xi_n$ — независимые одинаково распределенные случайные величины (наблюдения)
- $\mathbb{E}\xi_i = \mu, \mathbb{D}\xi_i = \sigma^2$
- $\overline{\xi_n} = \frac{1}{n}(\xi_1 + \dots + \xi_n)$
- $\mathbb{E}\overline{\xi_n} = \mu, \mathbb{D}\overline{\xi_n} = \frac{\sigma^2}{n}$
- Усреднение уменьшает дисперсию!

Усреднение наблюдений



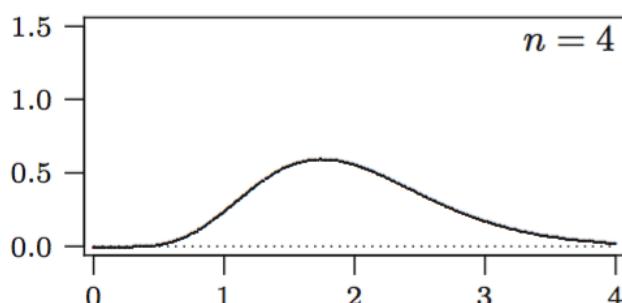
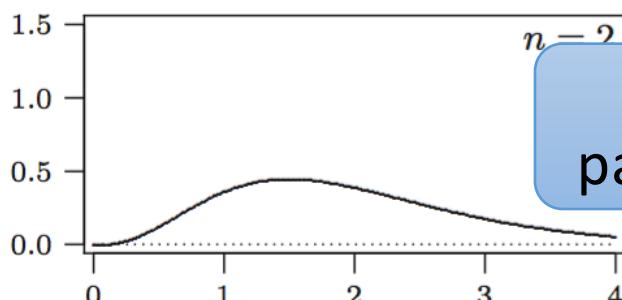
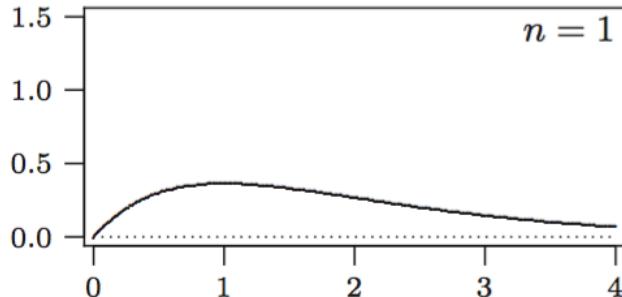
Закон больших чисел

- Среднее по наблюдениям стремится к матожиданию:

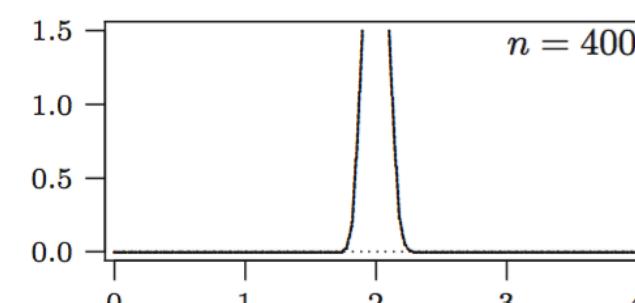
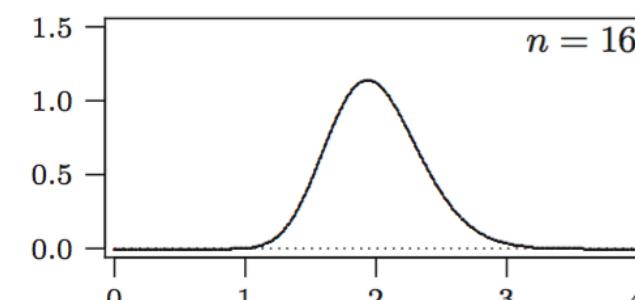
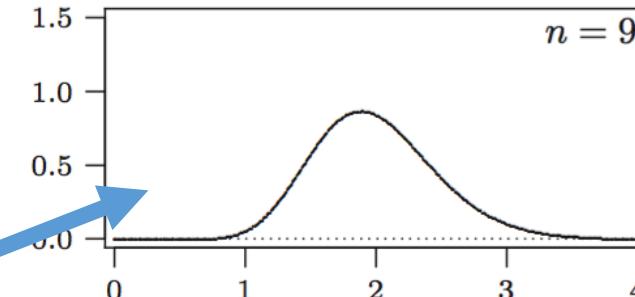
$$\lim_{n \rightarrow \infty} P(|\bar{\xi}_n - \mathbb{E}\xi_1| > \varepsilon) = 0$$

- Обосновывает утверждение «Вероятность события равна его доле в бесконечном числе экспериментов»

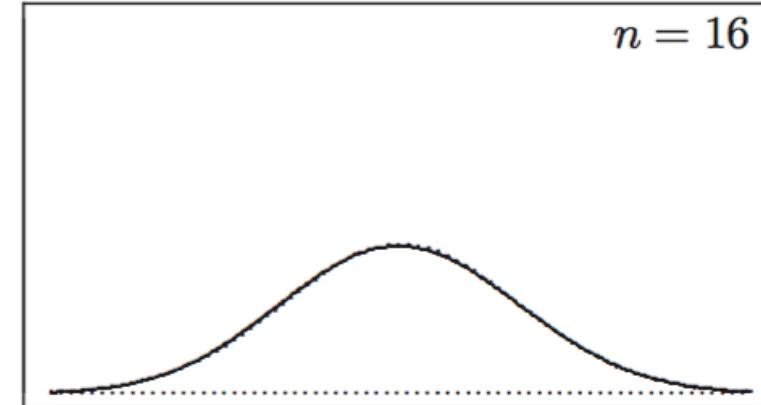
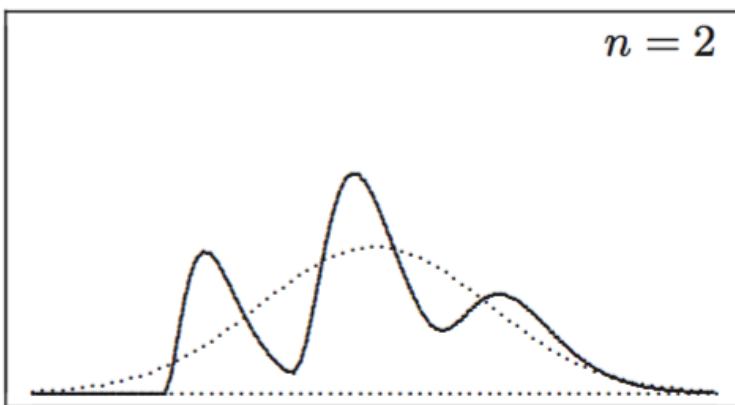
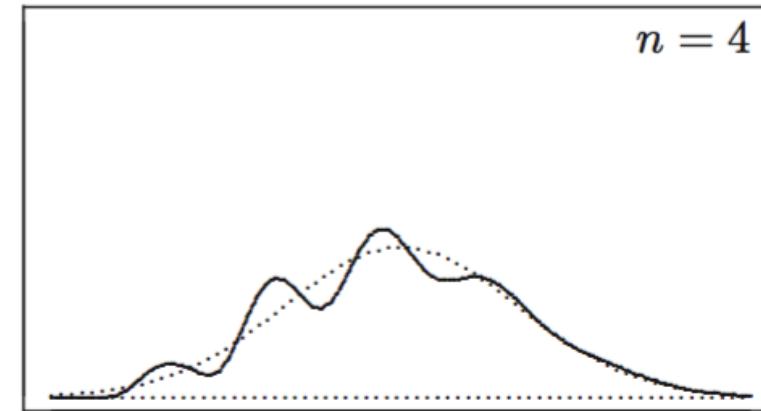
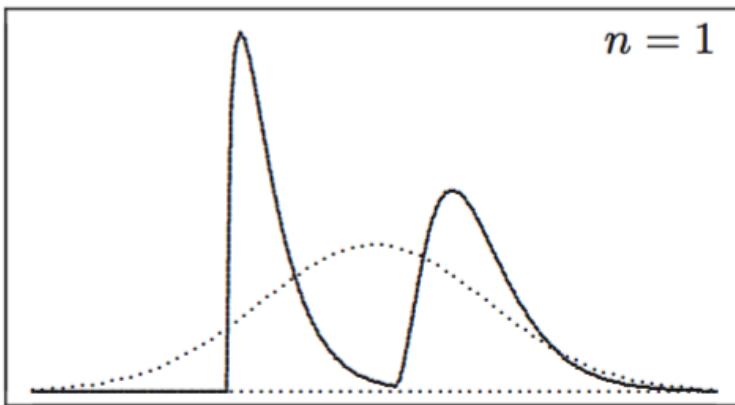
Усреднение случайных величин



Нормальное
распределение!



Усреднение случайных величин



Центральная предельная теорема

- Распределение среднего нормированных величин стремится к нормальному:

$$\sqrt{n} \frac{\bar{\xi}_n - \mu}{\sigma} \rightarrow N(0, 1)$$

Пример

- Бухгалтер решил округлять все числа до целых
- $\$99.53 \rightarrow \100
- $\$100.42 \rightarrow \100
- Какая ошибка накопится после округления 100 чисел?
- $\xi_i \sim R[-0.5, 0.5]$
- $P(|\xi_1 + \dots + \xi_{100}| > 10) = ?$

Пример

$$P(\xi_1 + \dots + \xi_{100} > 10) =$$

$$= P\left(\sqrt{100} \frac{\frac{\xi_1 + \dots + \xi_{100}}{100} - 0}{\sqrt{1/12}} > \sqrt{100} \frac{\frac{10}{100} - 0}{\sqrt{1/12}}\right)$$


$$\sqrt{n} \frac{\overline{\xi}_n - \mu}{\sigma}$$

Пример

$$P(\xi_1 + \dots + \xi_{100} > 10) =$$

$$= P\left(\sqrt{100} \frac{\frac{\xi_1 + \dots + \xi_{100}}{100} - 0}{\sqrt{1/12}} > \sqrt{100} \frac{\frac{10}{100} - 0}{\sqrt{1/12}}\right) \approx \{\text{ЦПТ}\}$$

$$\approx P(N(0, 1) > 3.46) = 0.0003$$

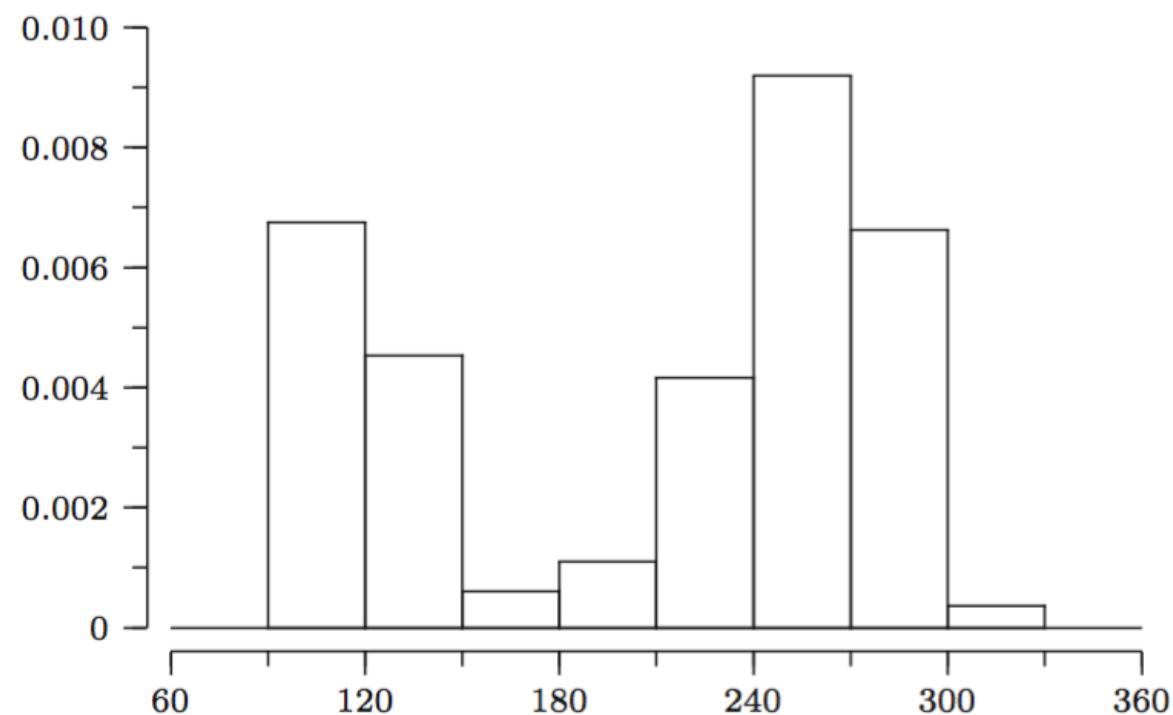
Ответ: 0.0006

Визуализация

Одномерная выборка

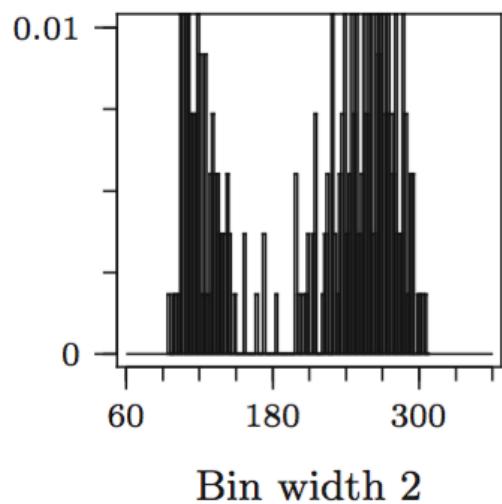
- Old Faithful
 - Длительность извержения гейзера

Гистограмма

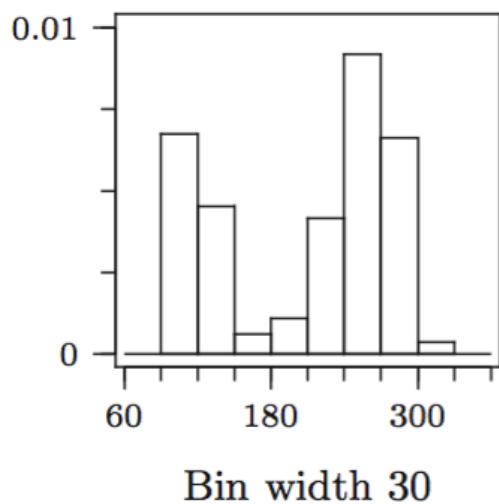


Гистограмма

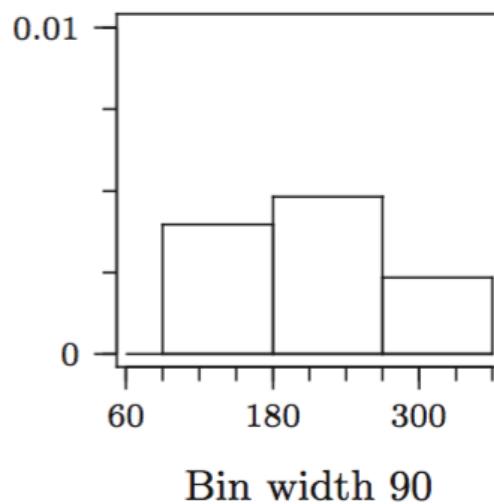
- Выбор ширины столбца:



Bin width 2



Bin width 30

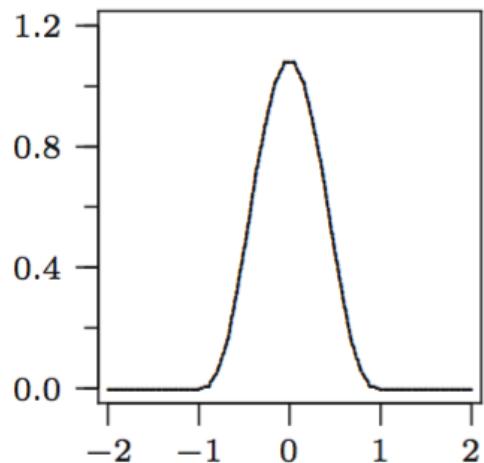


Bin width 90

Ядерное сглаживание

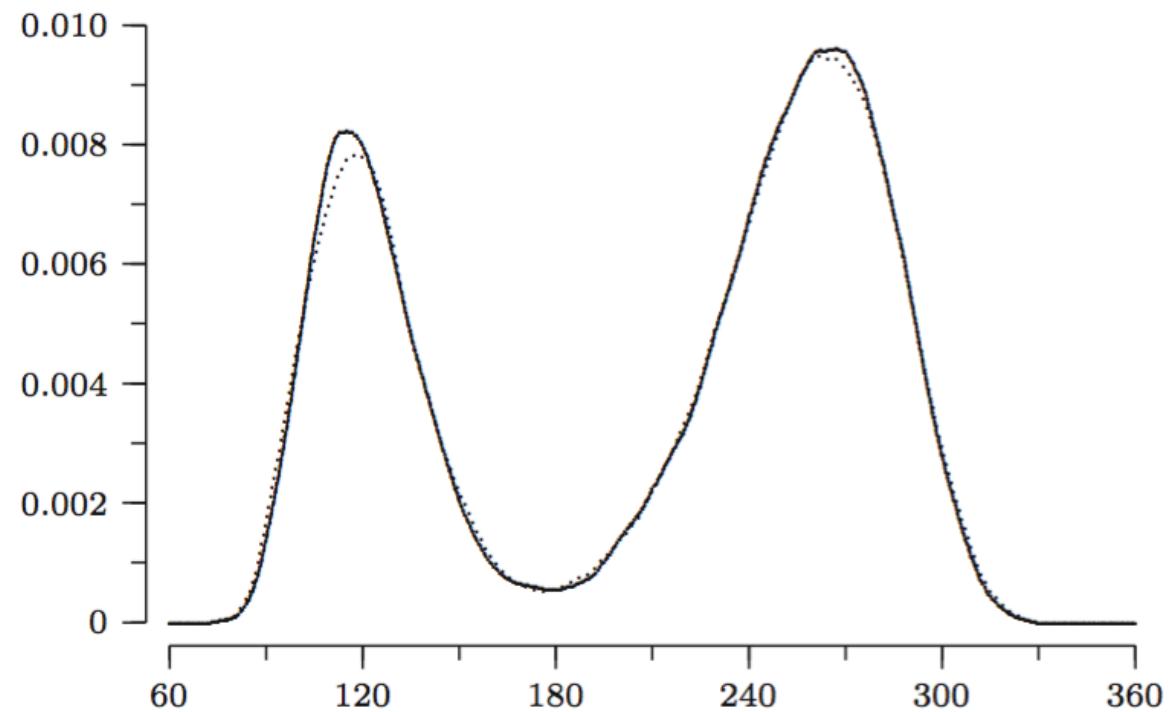
- Значение графика в точке — сумма взвешенных расстояний до наблюдений:

$$f(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

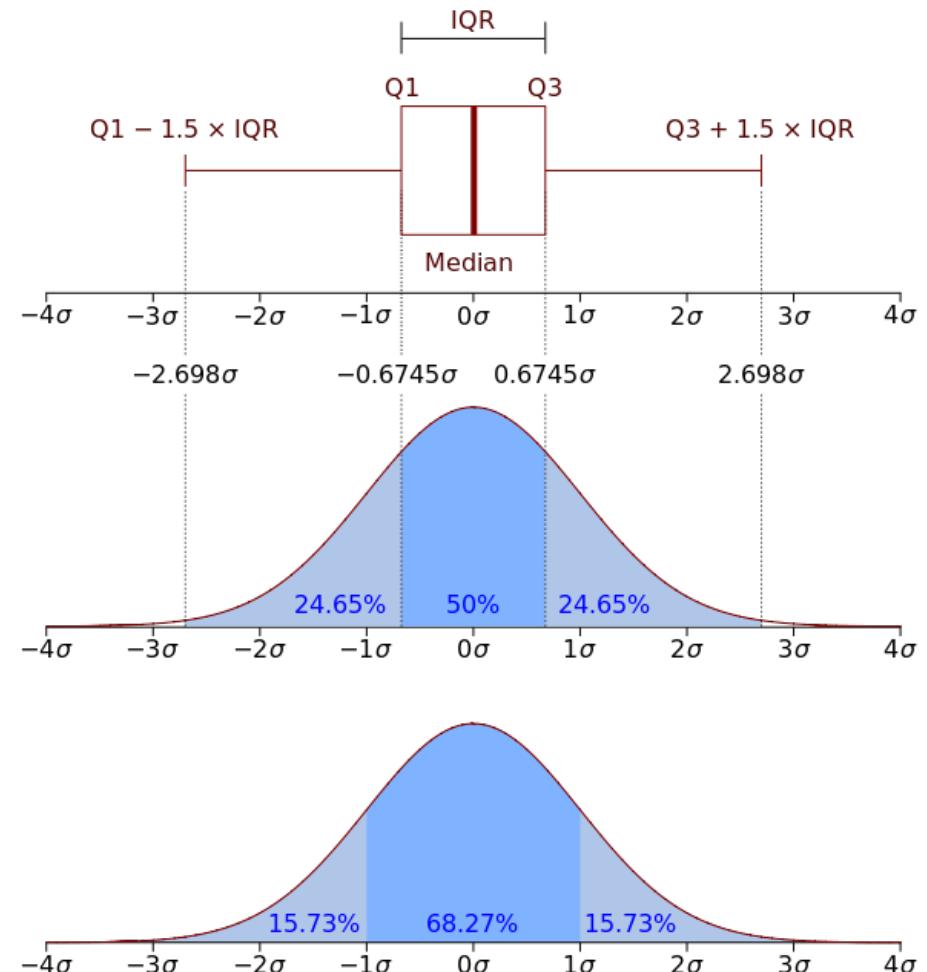


Triweight kernel

Ядерное сглаживание



Boxplot (ящик с усами)



Пример реальной задачи

- <https://www.kaggle.com/c/prudential-life-insurance-assessment>
- Страховые компании нередко просят много документов
 - Получение может занять до 30 дней
- Хотим автоматизировать процессы и оценивать риск на основе простых факторов
- 8 уровней риска

Данные

- 128 столбцов
- Есть вещественные, целочисленные и категориальные признаки

	Id	Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	...	Med
0	2	1	D3	10	0.076923	2	1	1	0.641791	0.581818	...	0
1	5	1	A1	26	0.076923	2	3	1	0.059701	0.600000	...	0
2	6	1	E1	26	0.076923	2	3	1	0.029851	0.745455	...	0
3	7	1	D4	10	0.487179	2	3	1	0.164179	0.672727	...	0
4	8	1	D2	26	0.230769	2	3	1	0.417910	0.654545	...	0

Вещественные признаки

Вещественные признаки

	Insurance_History_5	Family_Hist_2	Family_Hist_3	Family_Hist_4	Family_Hist_5
count	33985.000000	30725.000000	25140.000000	40197.000000	17570.000000
mean	0.001733	0.474550	0.497737	0.444890	0.484635
std	0.007338	0.154959	0.140187	0.163012	0.129200
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000400	0.362319	0.401961	0.323944	0.401786
50%	0.000973	0.463768	0.519608	0.422535	0.508929
75%	0.002000	0.579710	0.598039	0.563380	0.580357
max	1.000000	1.000000	1.000000	0.943662	1.000000

Вещественные признаки

Insurance_History_5	Family_Hist_2	Family_Hist_3	Family_Hist_4	Family_Hist_5
count 33985.000000	30725.000000	25140.000000	40197.000000	17570.000000
mean 0.001733	0.474550	0.497737	0.444890	0.484635
std 0.007338	0.154959	0.140187	0.163012	0.129200
min 0.000000	0.000000	0.000000	0.000000	0.000000
25% 0.000400	0.362319	0.401961	0.326944	0.401786
50% 0.000973	0.463768	0.519608	0.422535	0.508929
75% 0.002000	0.579710	0.598039	0.563380	0.580357
max 1.000000	1.000000	1.000000	0.943662	1.000000

Много пропусков!

Целочисленные признаки

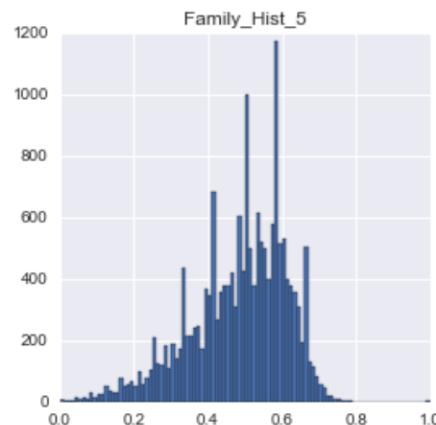
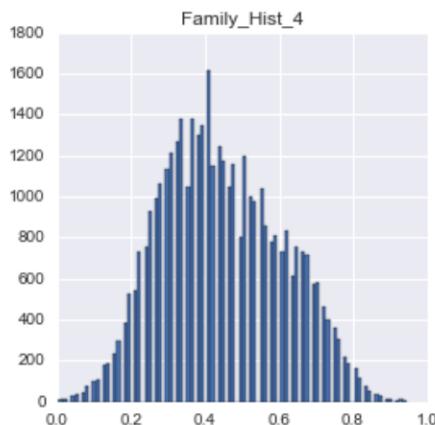
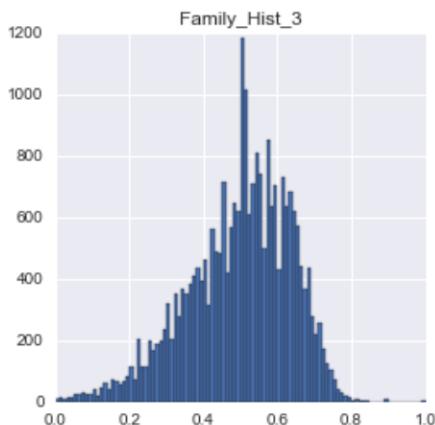
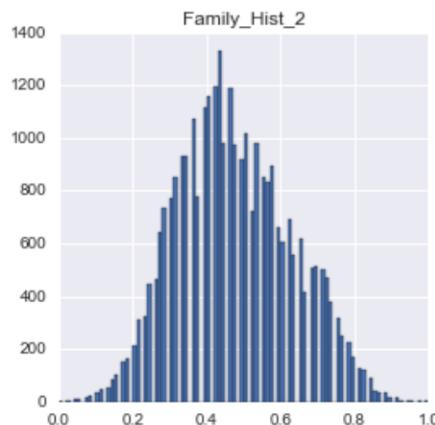
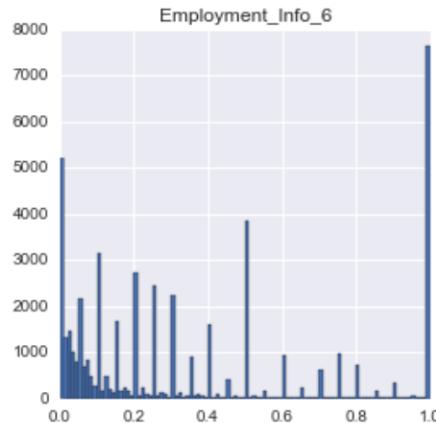
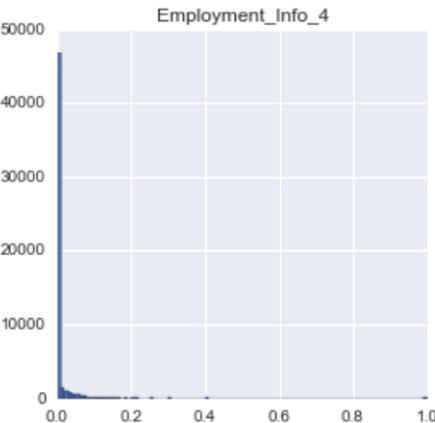
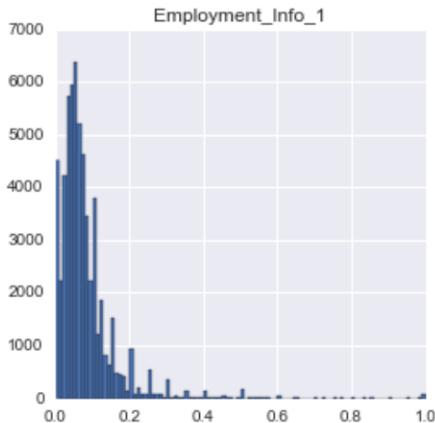
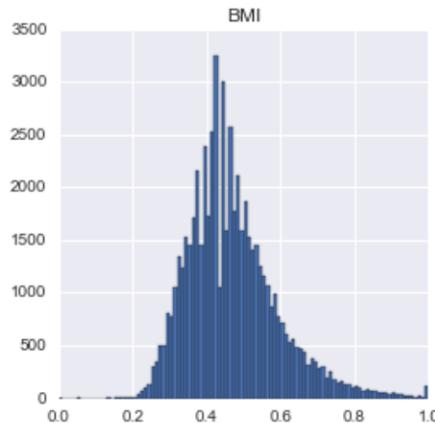
	Medical_History_1	Medical_History_10	Medical_History_15	Medical_History_24	Medical_History_32
count	50492.000000	557.000000	14785.000000	3801.000000	1107.000000
mean	7.962172	141.118492	123.760974	50.635622	11.965673
std	13.027697	107.759559	98.516206	78.149069	38.718774
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	8.000000	17.000000	1.000000	0.000000
50%	4.000000	229.000000	117.000000	8.000000	0.000000
75%	9.000000	240.000000	240.000000	64.000000	2.000000
max	240.000000	240.000000	240.000000	240.000000	240.000000

Целочисленные признаки

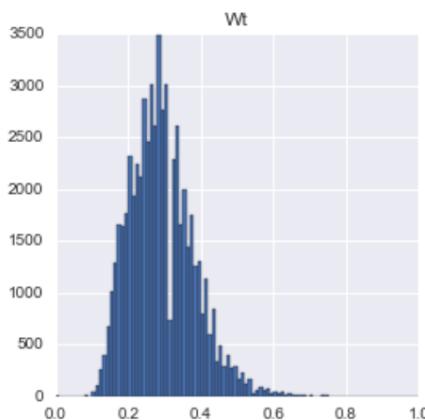
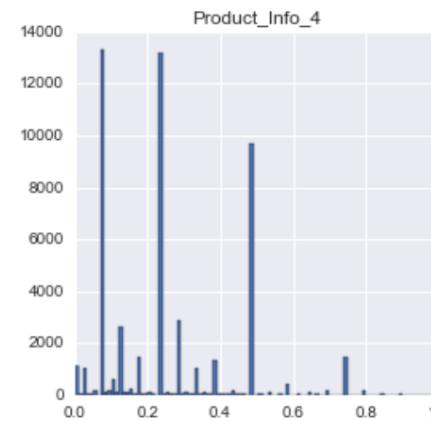
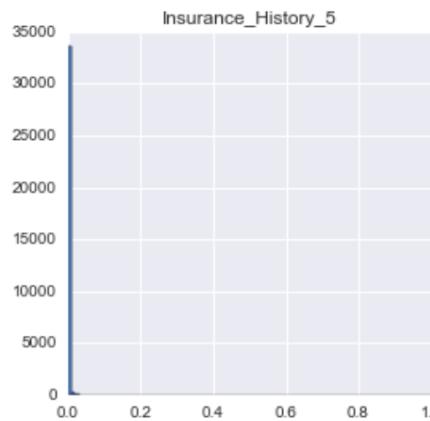
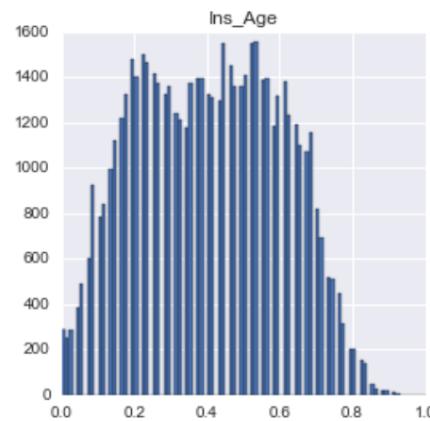
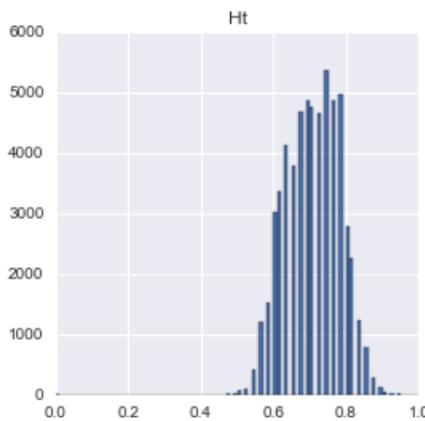
	Medical_History_1	Medical_History_10	Medical_History_15	Medical_History_24	Medical_History_32
count	50492.000000	557.000000	14785.000000	3801.000000	1107.000000
mean	7.962172	141.118492	123.760974	50.635622	11.965673
std	13.027697	107.759559	98.516206	78.149069	38.718774
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	8.000000	17.000000	1.000000	0.000000
50%	4.000000	229.000000	117.000000	8.000000	0.000000
75%	9.000000	240.000000	240.000000	64.000000	2.000000
max	240.000000	240.000000	240.000000	240.000000	240.000000

Практически не заполнены,
можно удалить

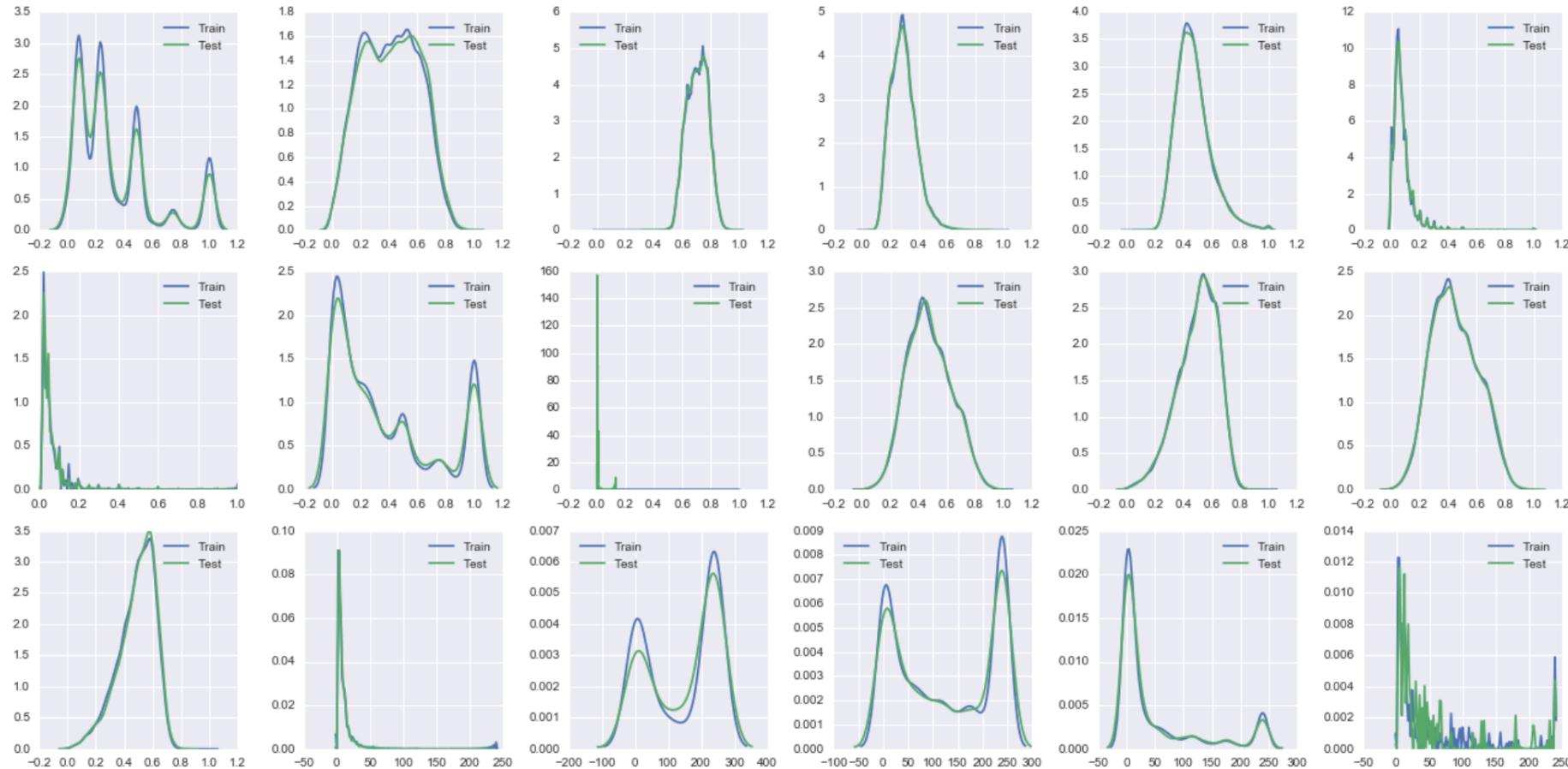
Распределения



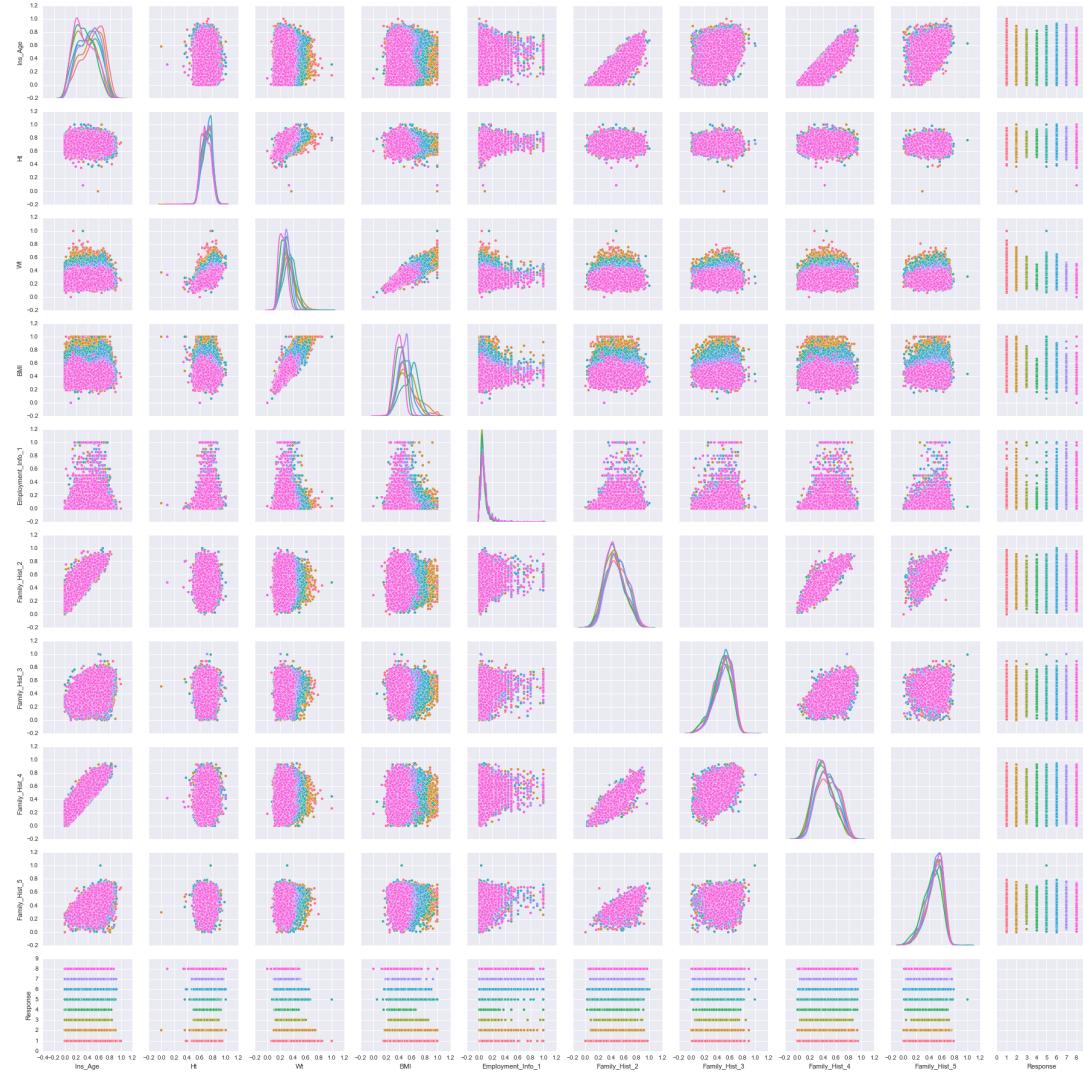
Распределения



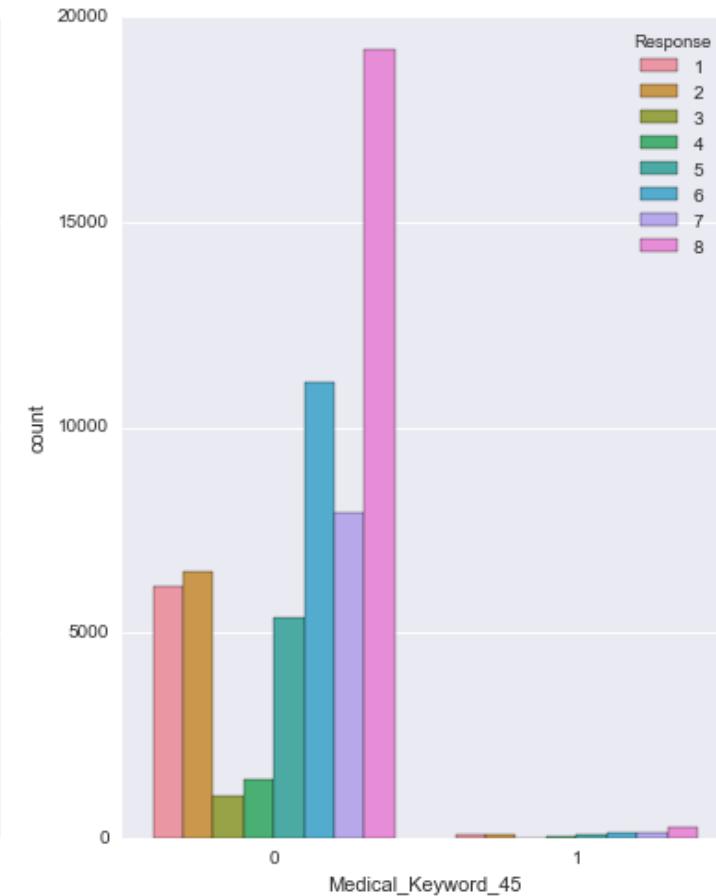
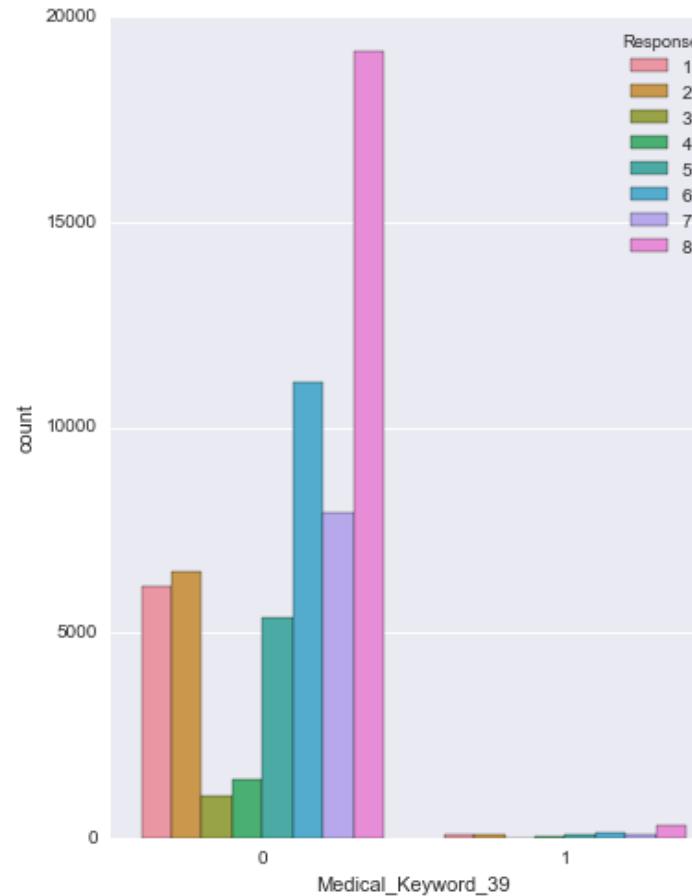
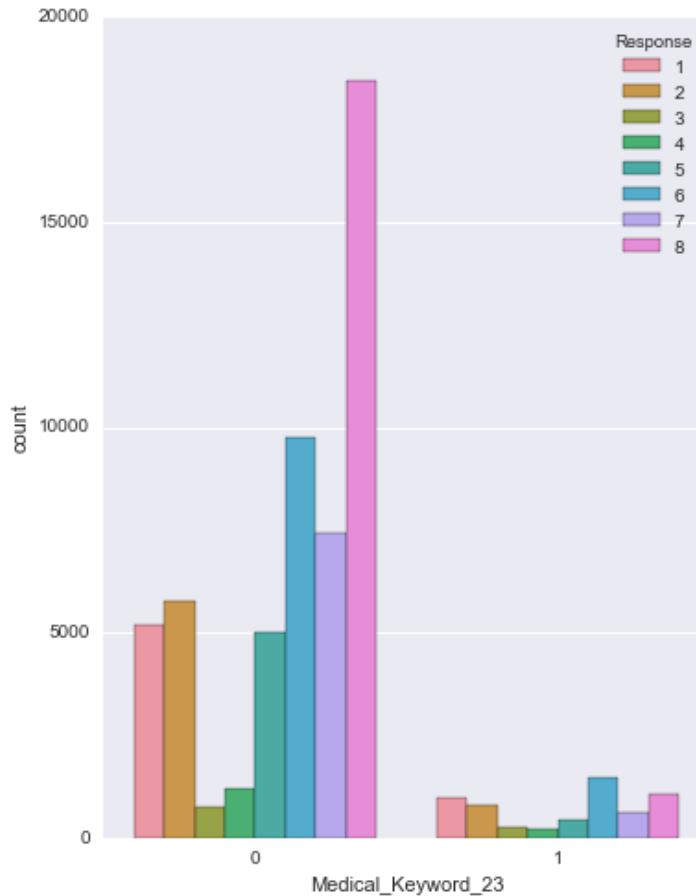
Распределения на обучении и контроле



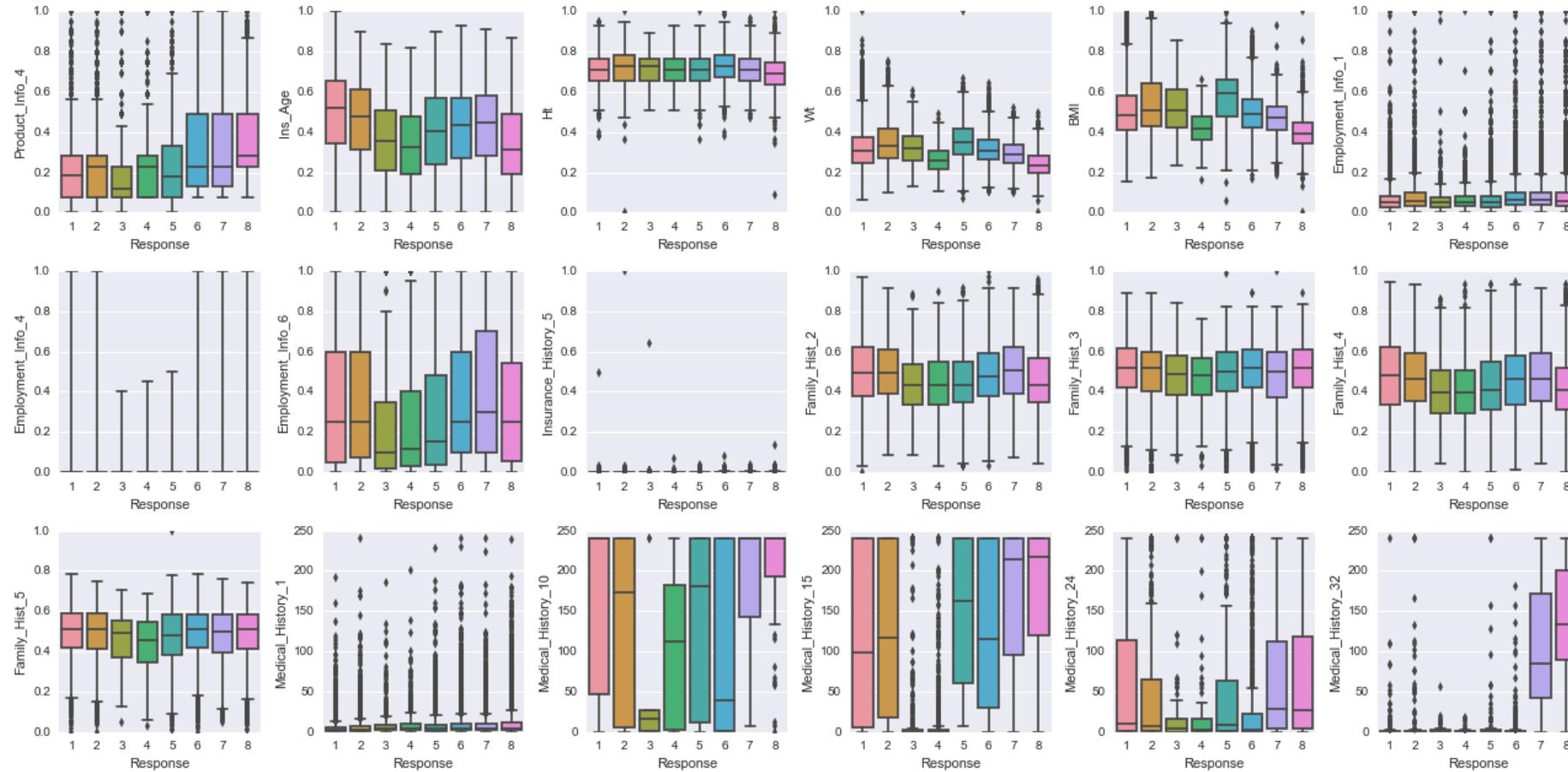
Пары признаков



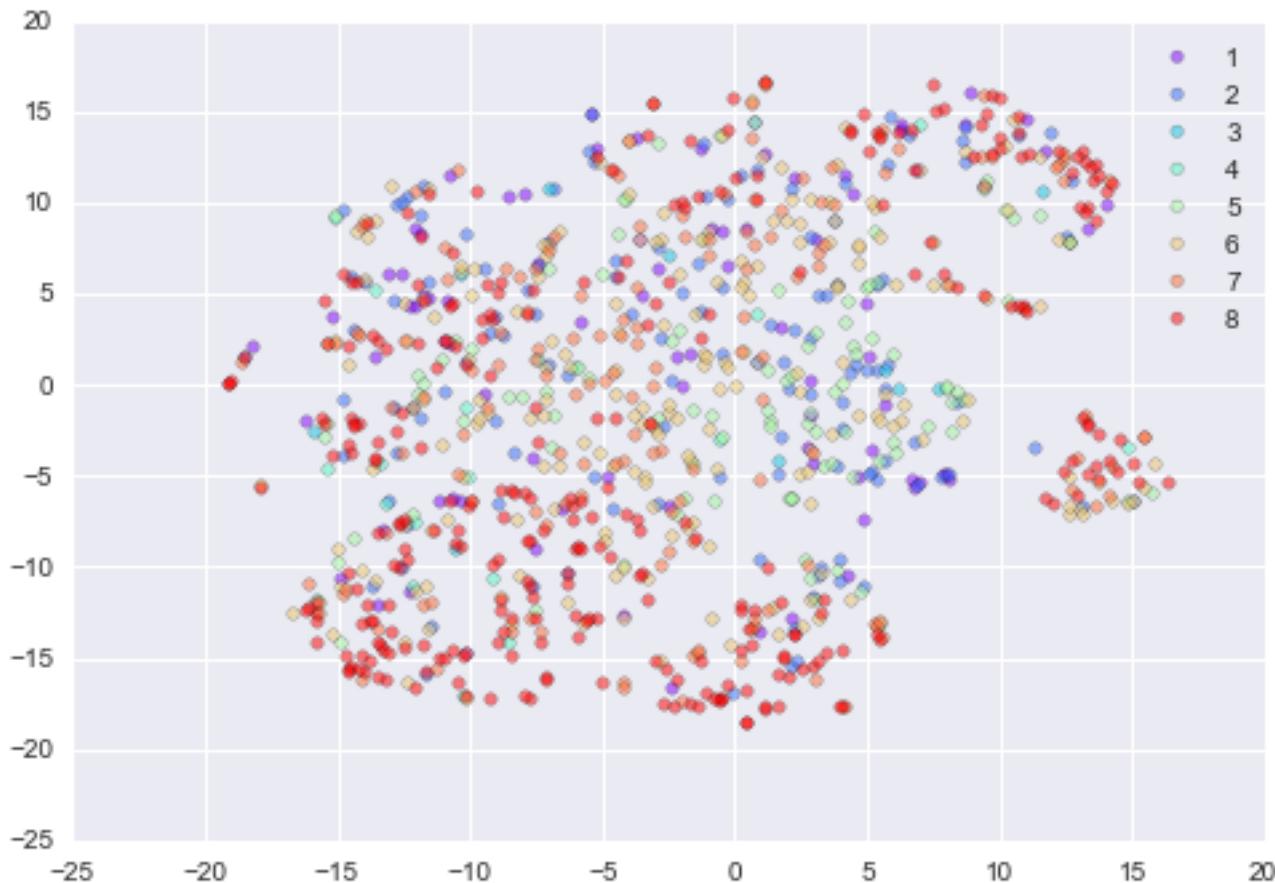
Бинарные признаки и целевая переменная



Вещественные признаки и целевая переменная



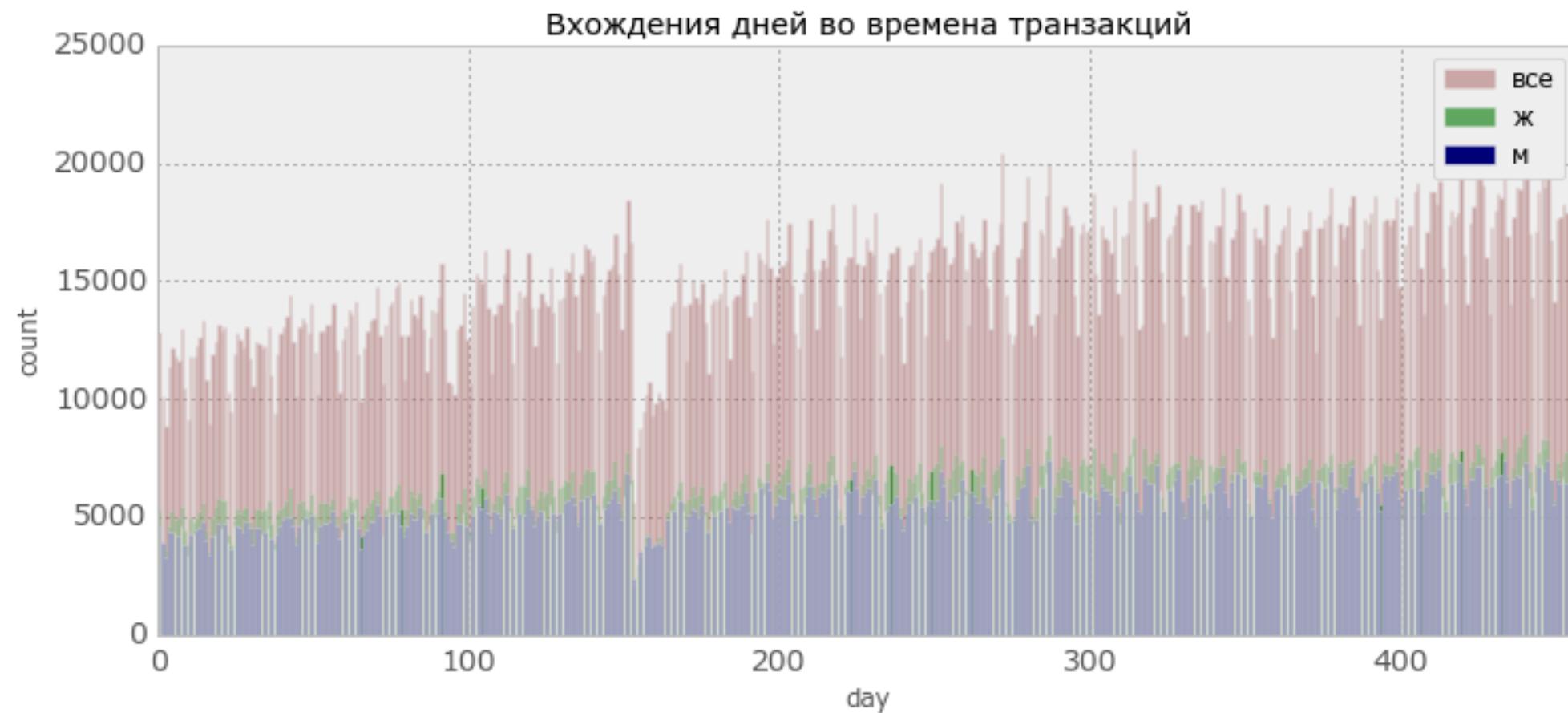
t-SNE



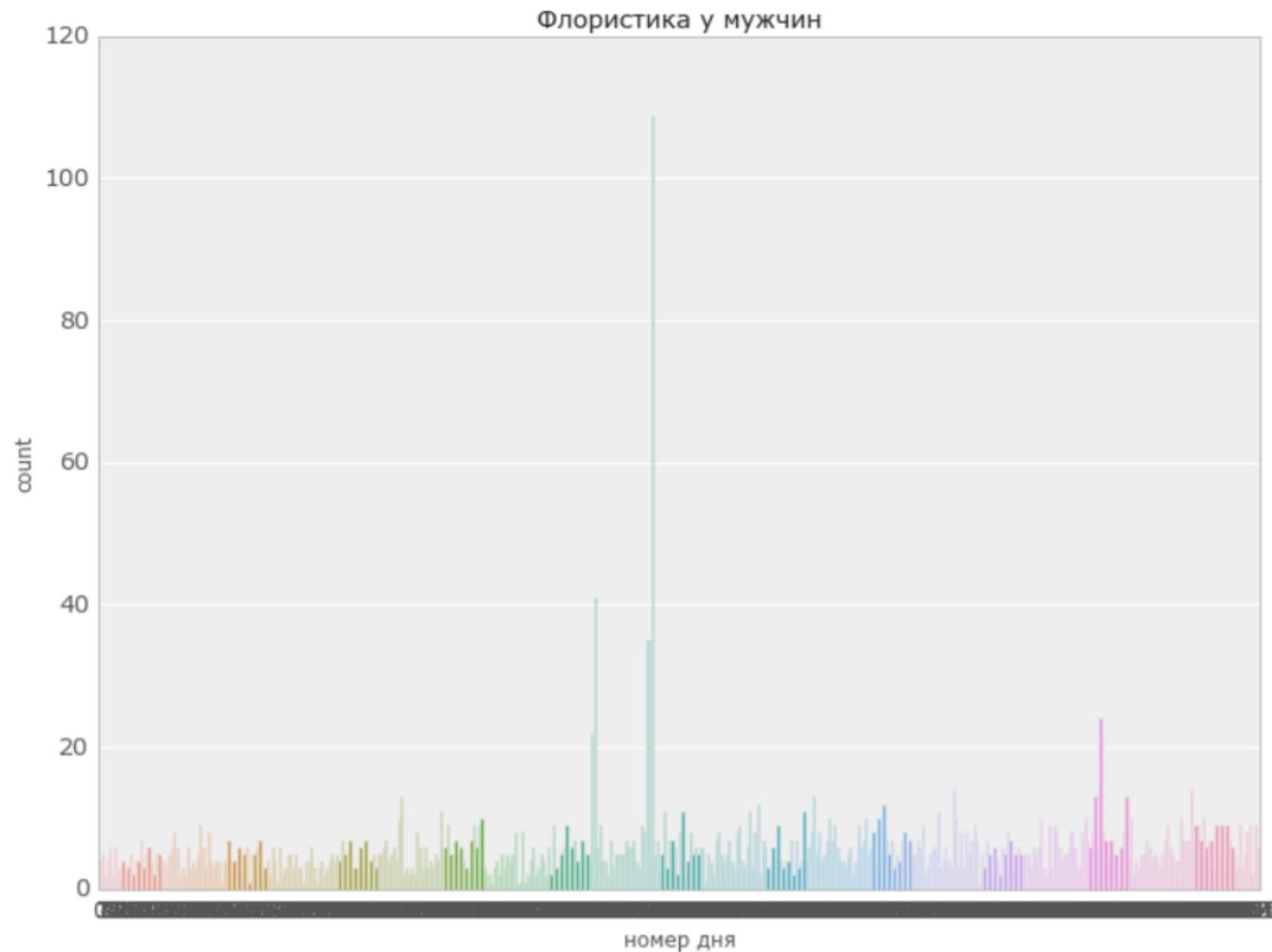
Деанонимизация данных

- Соревнование Сбербанка: предсказание трат по транзакциям
- Вместо дат — число дней, прошедших с некоторого момента

Деанонимизация



Деанонимизация



Резюме

- Распределения бывают дискретными и непрерывными
- Характеристики случайных величин: матожидание, медиана, дисперсия, квантили и т.д.
- Центральная предельная теорема — среднее большого числа значений имеет нормальное распределение
- Грамотная визуализация позволяет увидеть много интересного