Введение в анализ данных

Лекция 6

Математическая статистика и анализ данных

Евгений Соколов

sokolov.evg@gmail.com

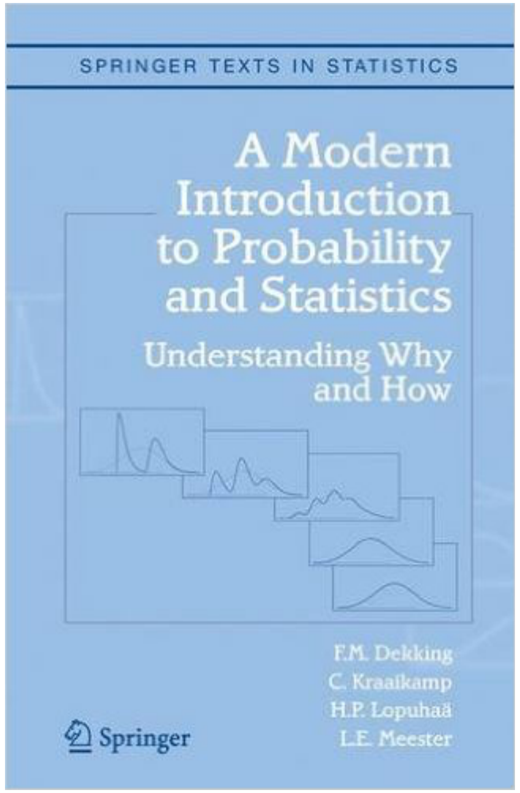
НИУ ВШЭ, 2016

План на сегодня

- Непрерывные распределения
- Матожидание, дисперсия и другие моменты
- Закон больших чисел и центральная предельная теорема
- Визуализация
- Метод максимального правдоподобия

Литература

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

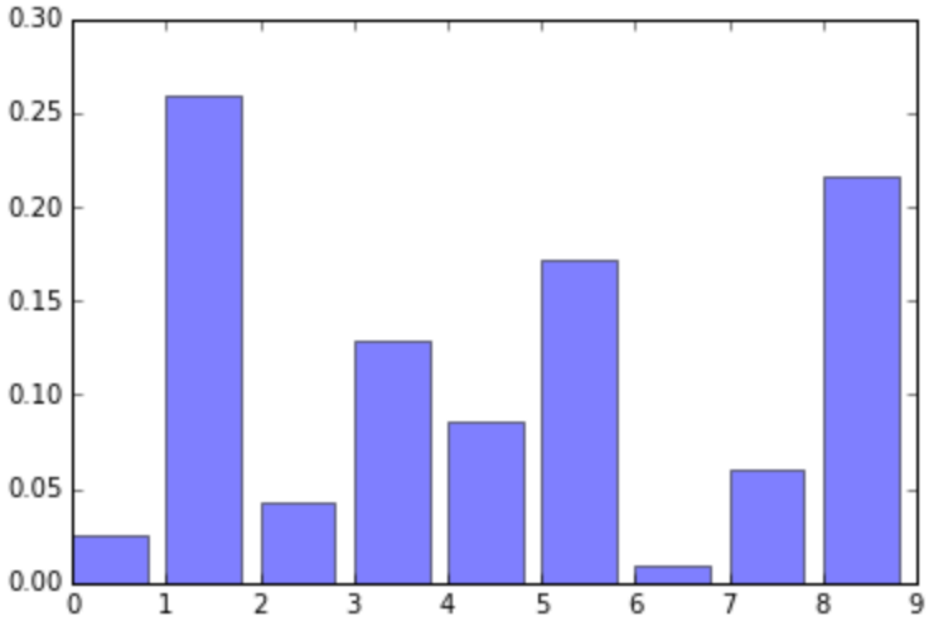


Непрерывные распределения

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

Вероятность записи на конкретный майнор



Непрерывная случайная величина

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

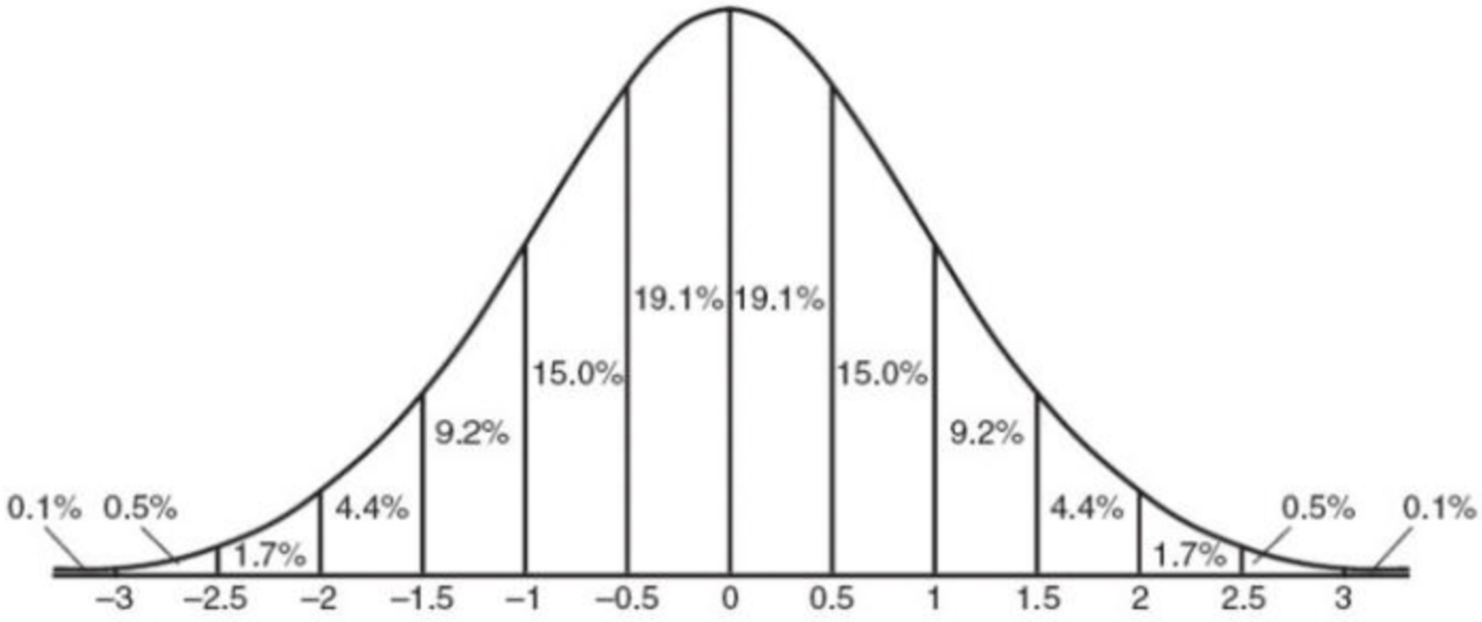
Непрерывная случайная величина

- 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!

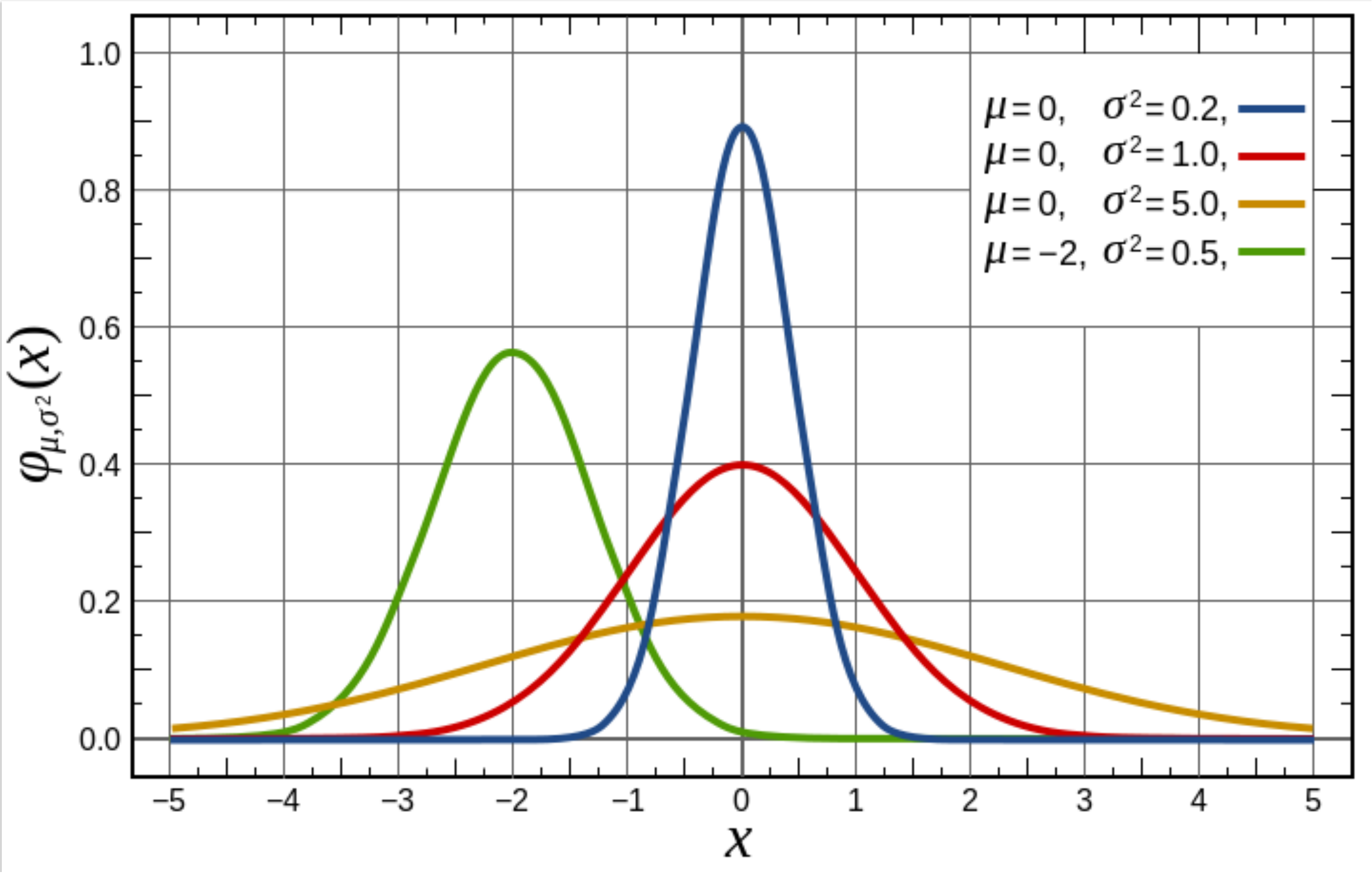
Плотность распределения

Плотность

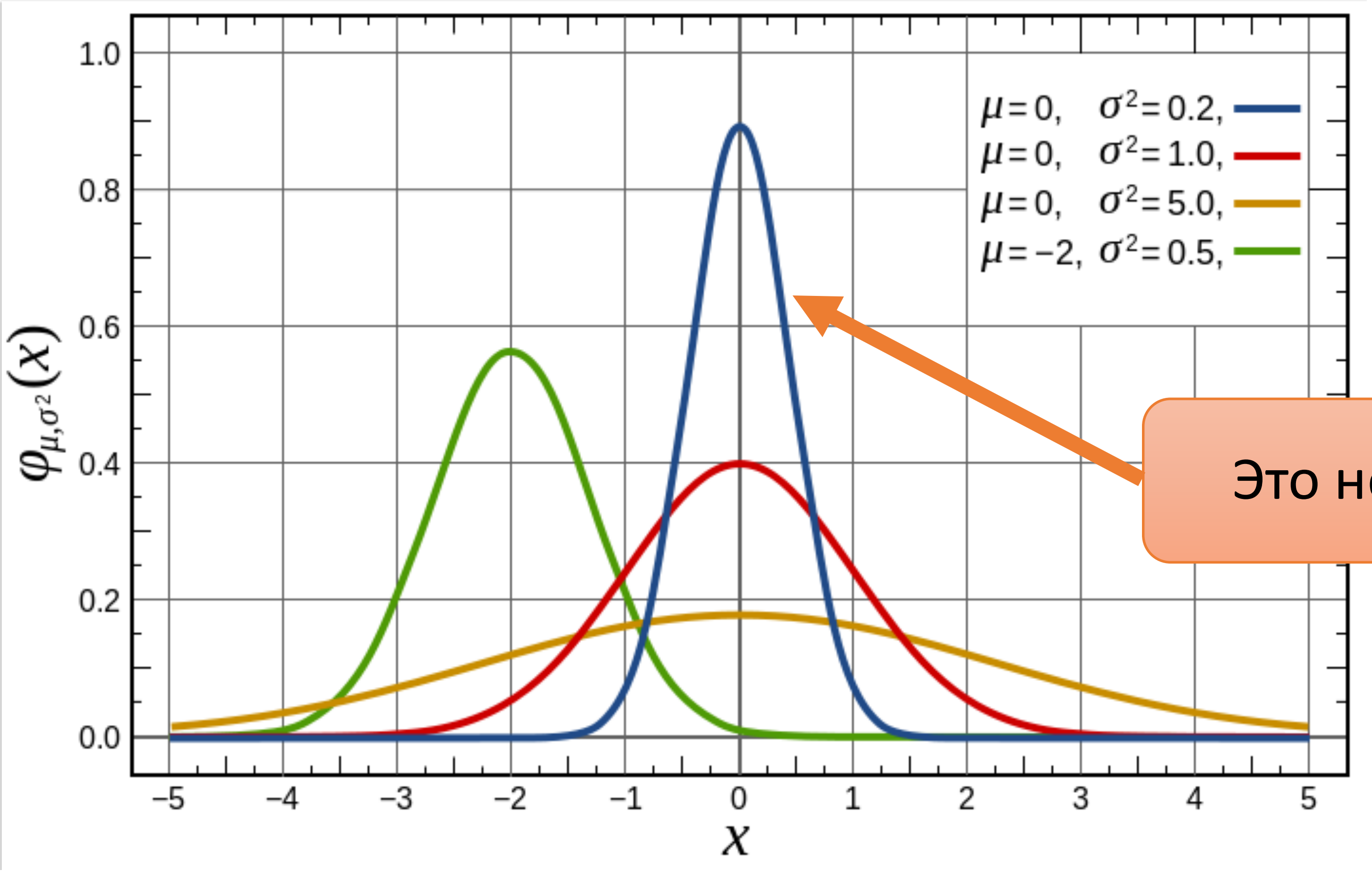
$$P(a \le \xi \le b) = \int_a^b p(x)dx$$



Плотность распределений

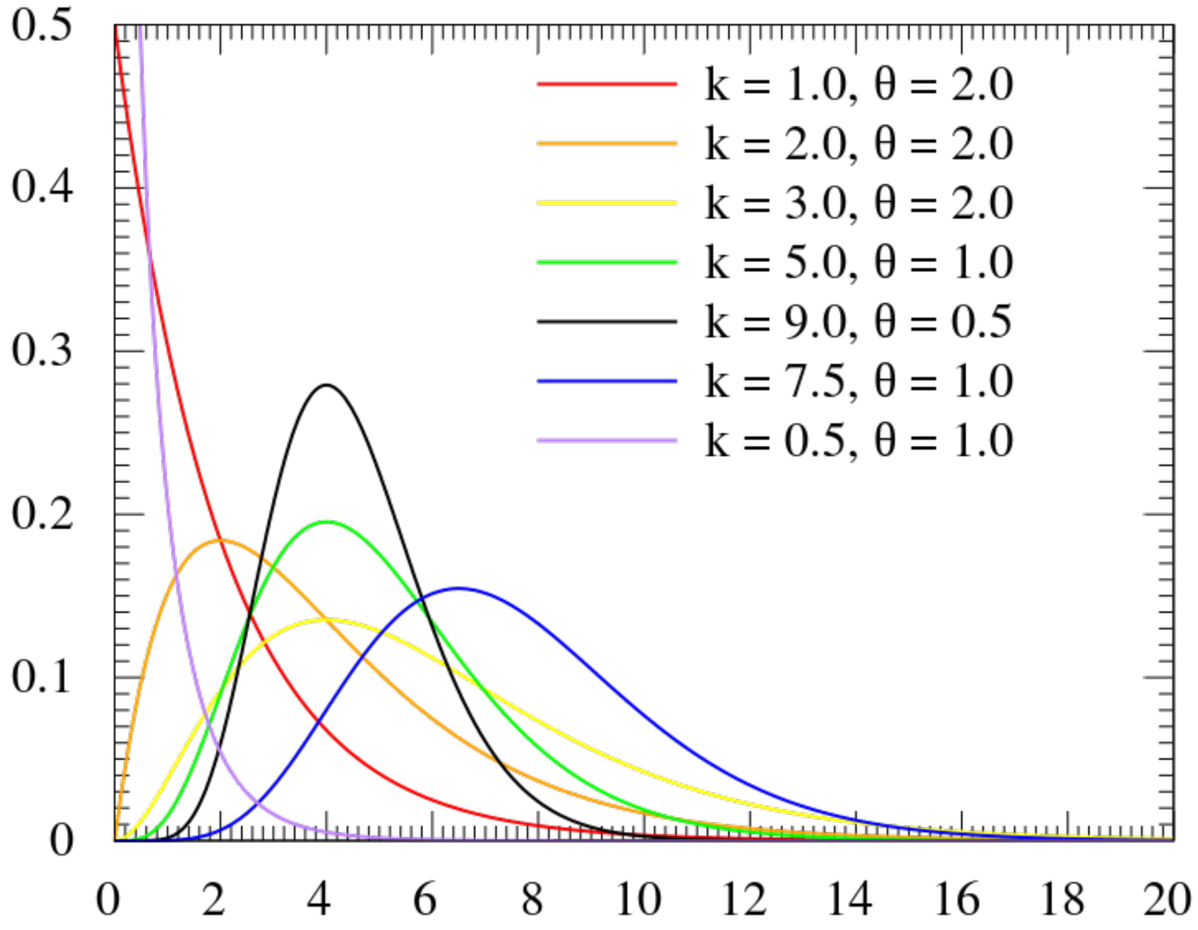


Плотность распределений

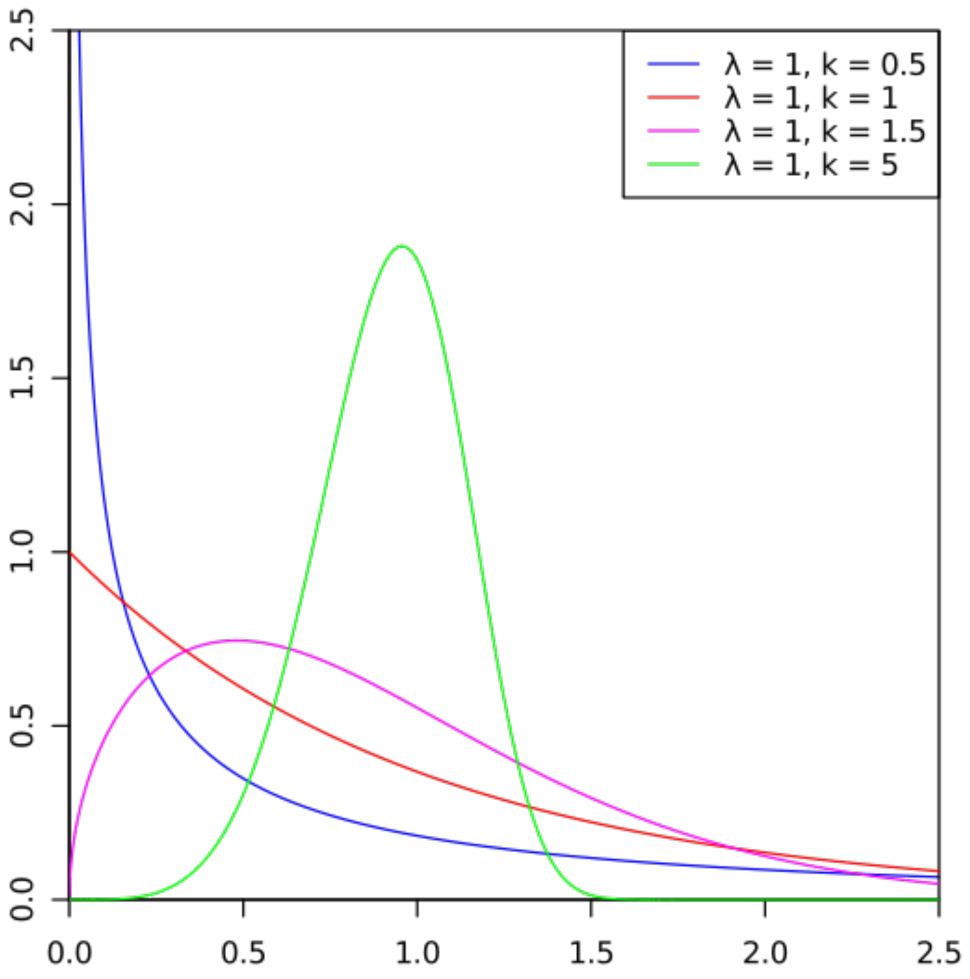


Это не вероятность!

Плотность распределений



Плотность распределений

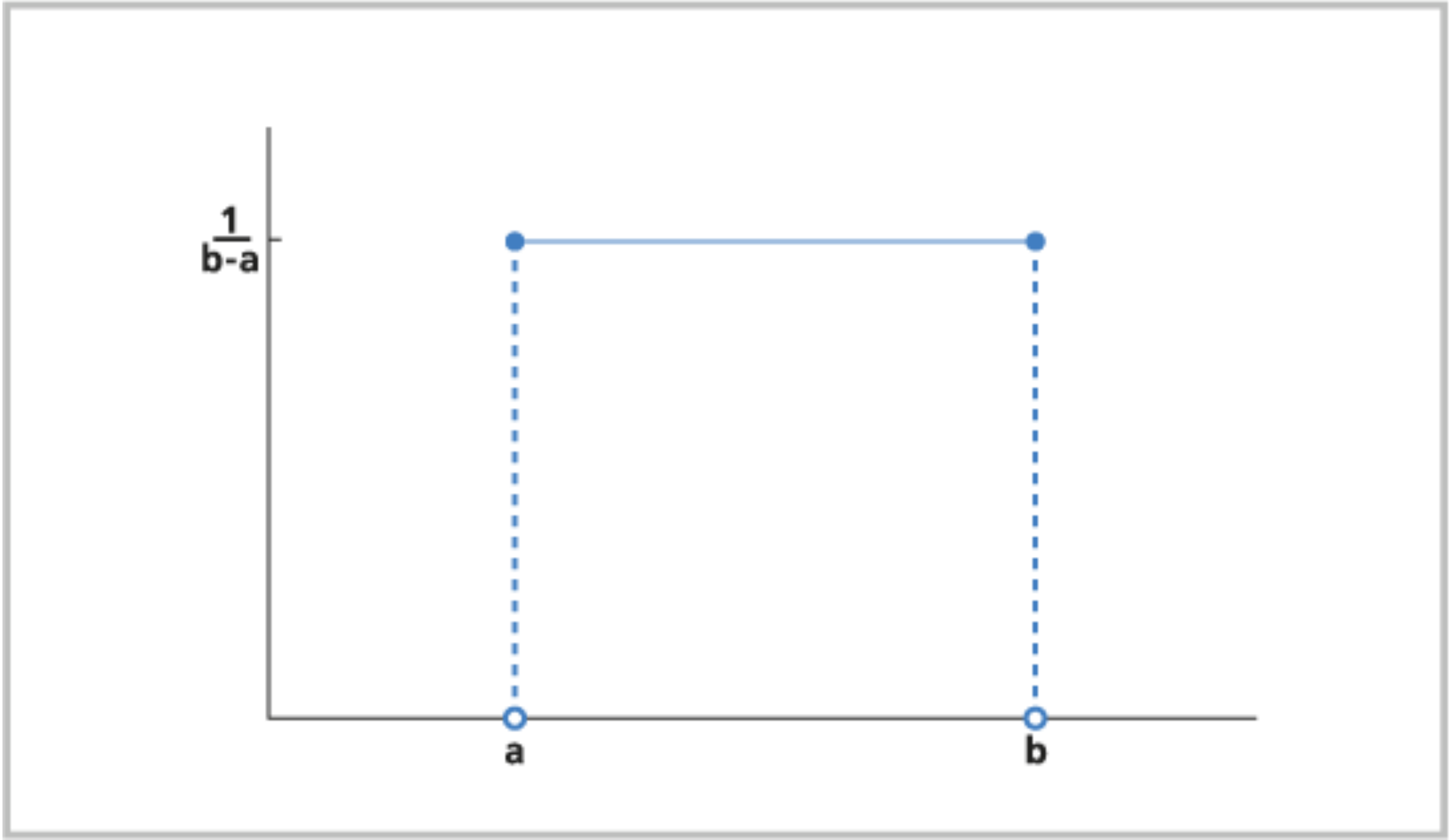


Равномерное распределение

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

$$p(x) = \begin{cases} \dfrac{1}{b-a}, a \le x \le b \\ 0, \qquad\qquad \text{иначе} \end{cases}$$

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

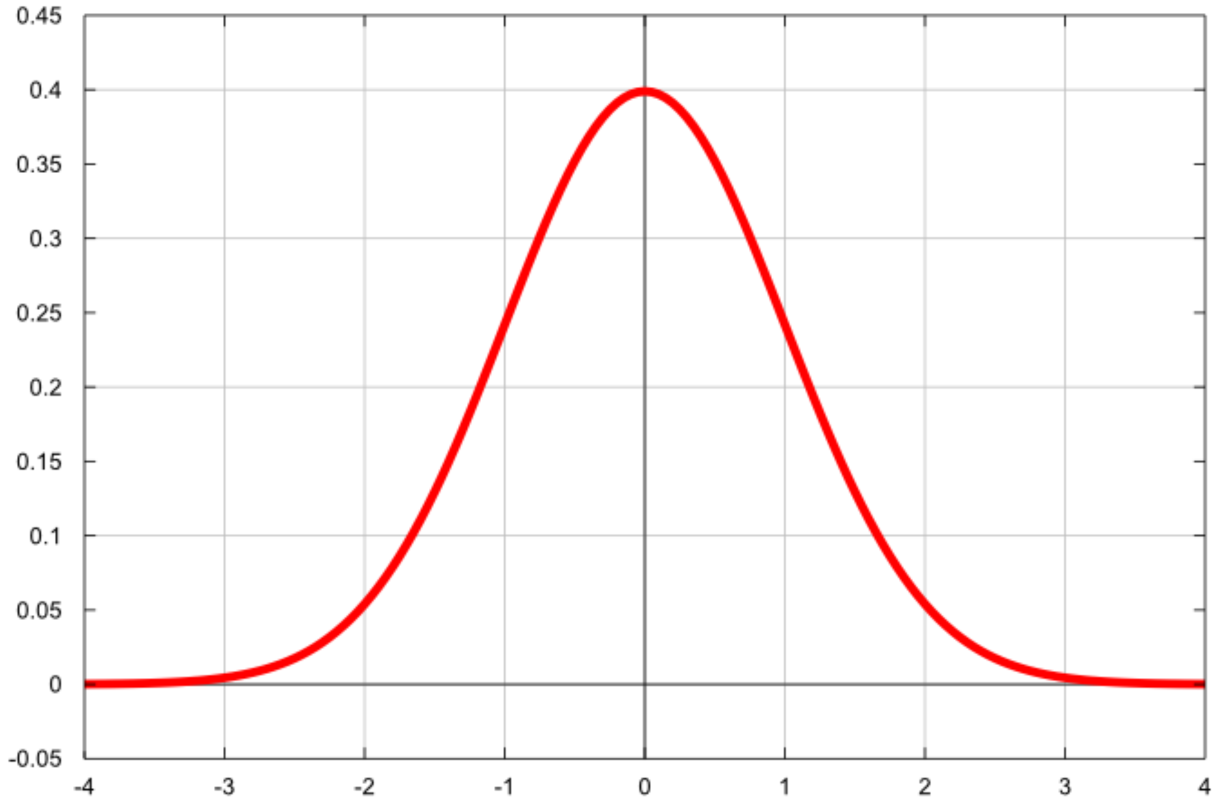
- Не очень распространено
- Легко эмулировать на компьютере
- Позволяет генерировать числа из любого распределения

Нормальное распределение

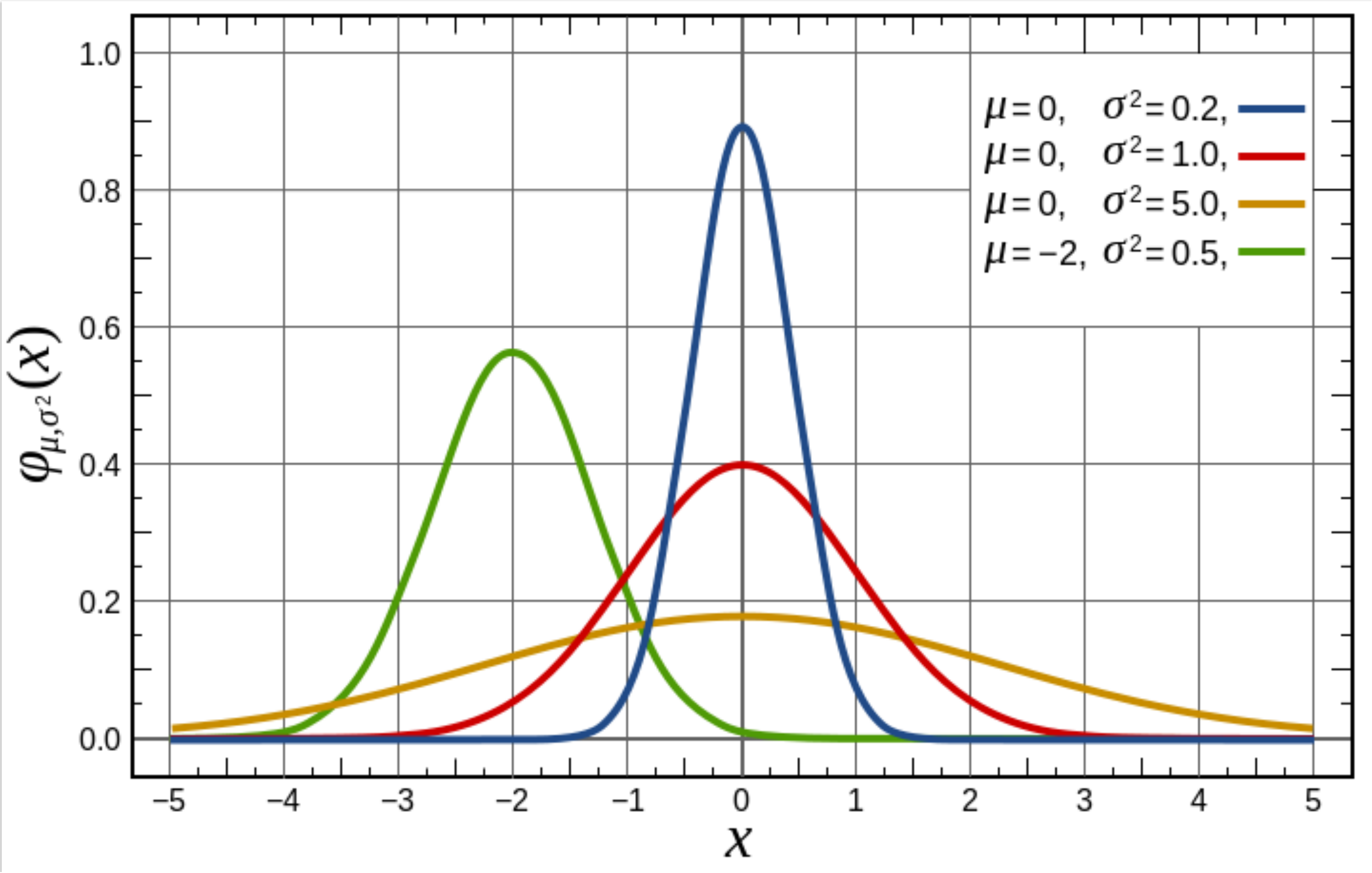
- Носитель: $\mathbb{R}$
- $\xi \sim N(\mu, \sigma^2)$

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

Нормальное распределение



Нормальное распределение



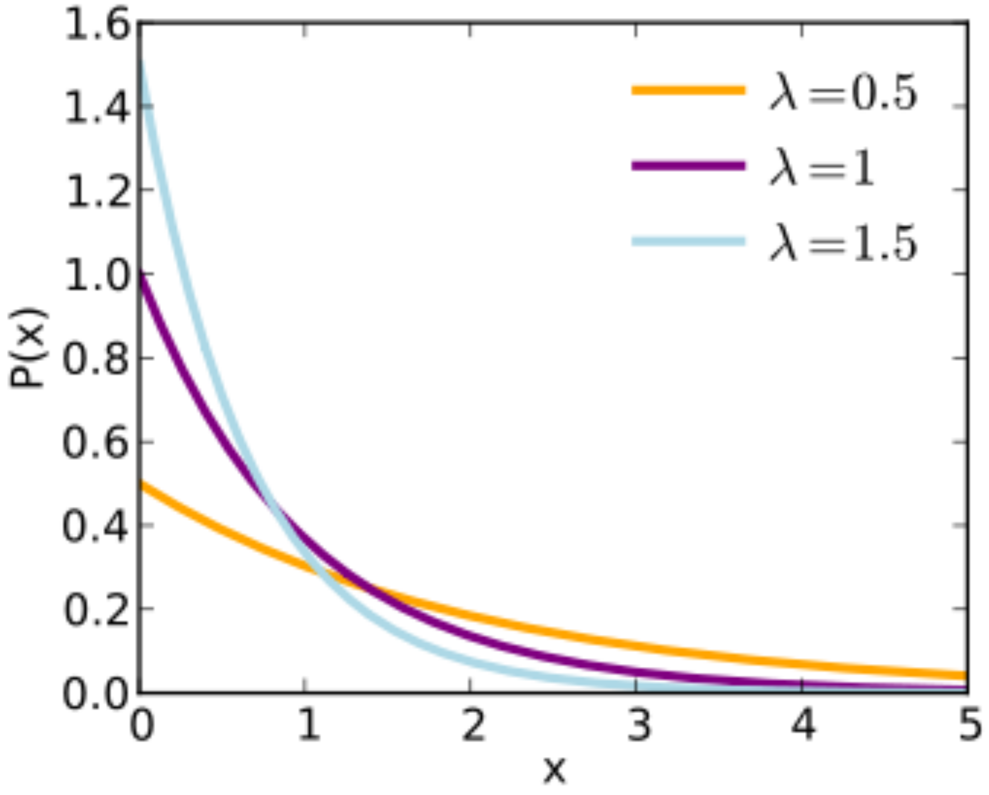
Нормальное распределение

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



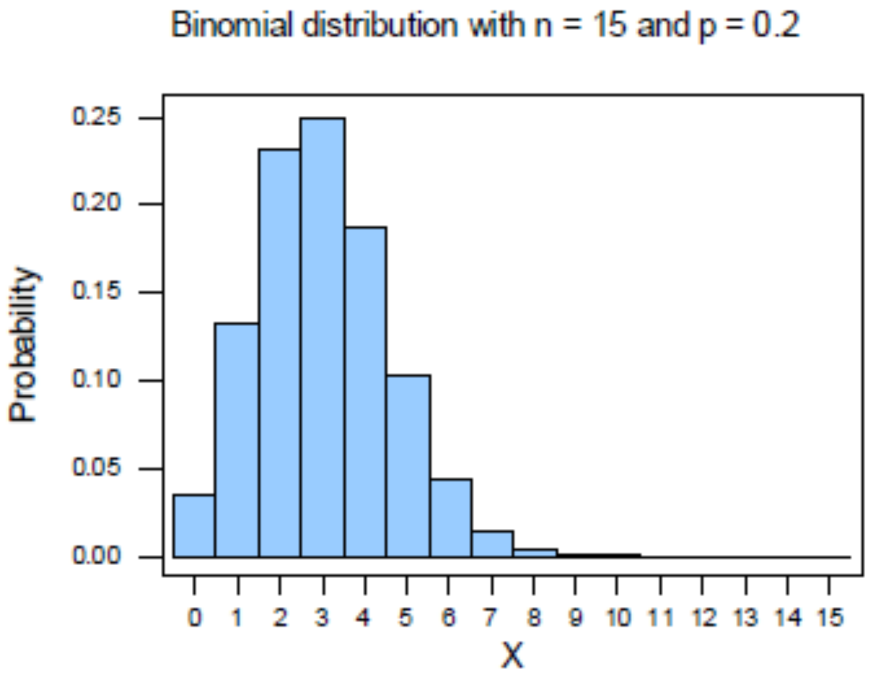
Экспоненциальное распределение

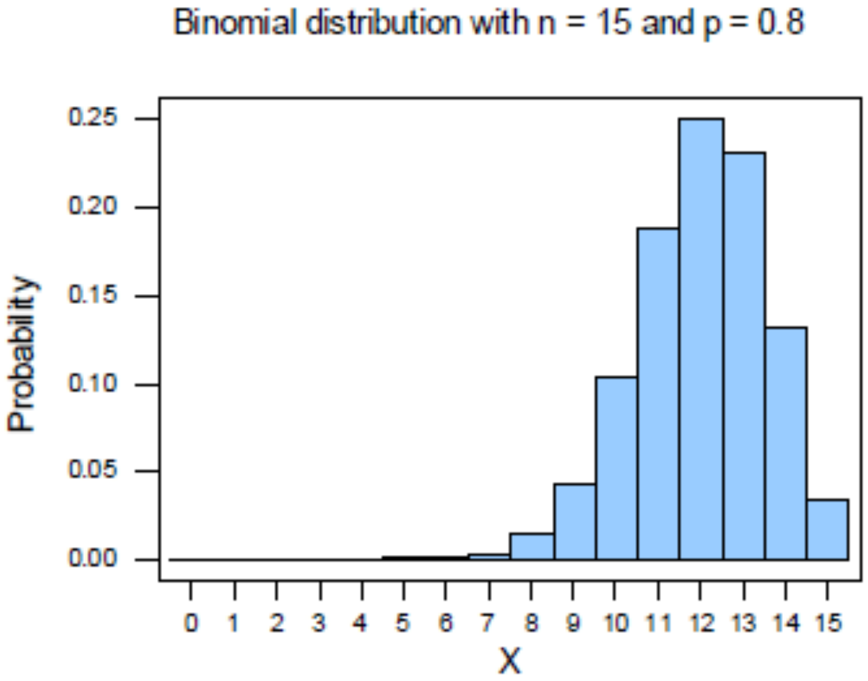
- Моделирует расстояние между редкими событиями
- Время до следующего звонка в колл-центр
- Время до следующего вопроса студента на лекции
- Расстояние между двумя соседними мутациями в ДНК

Характеристики случайных величин

Среднее значение

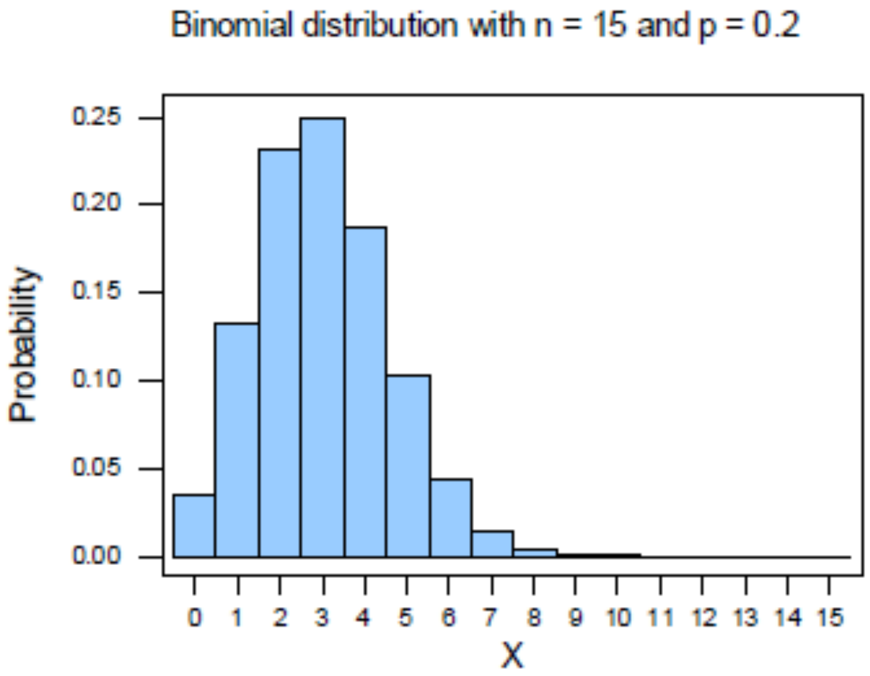
- Студент правильно отвечает на один вопрос с вероятностью p
- На сколько вопросов он ответит, если всего их n?
- $\xi \sim \mathrm{Bin}(n, p)$

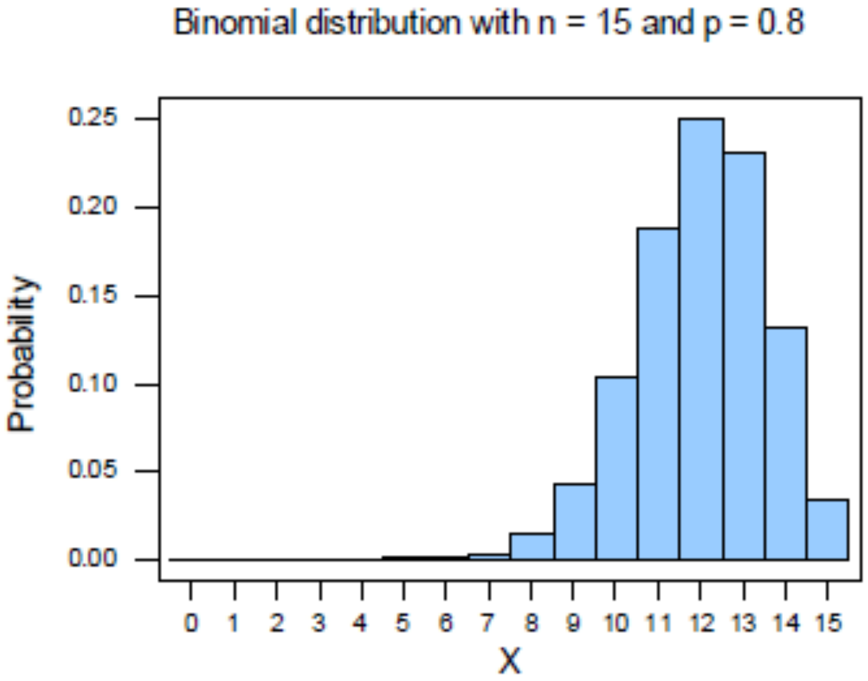




Среднее значение

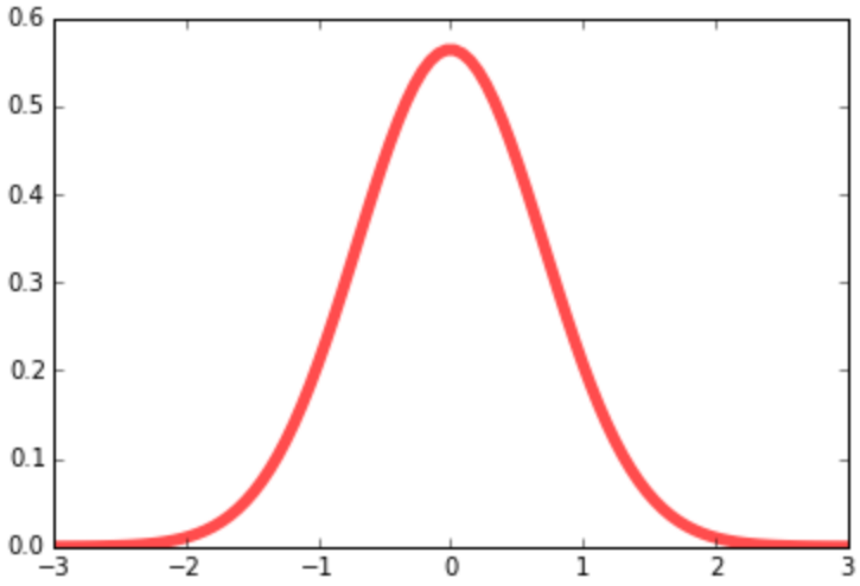
- На сколько вопросов в среднем будут отвечать такие студенты?
- Ответ: 3 и 12



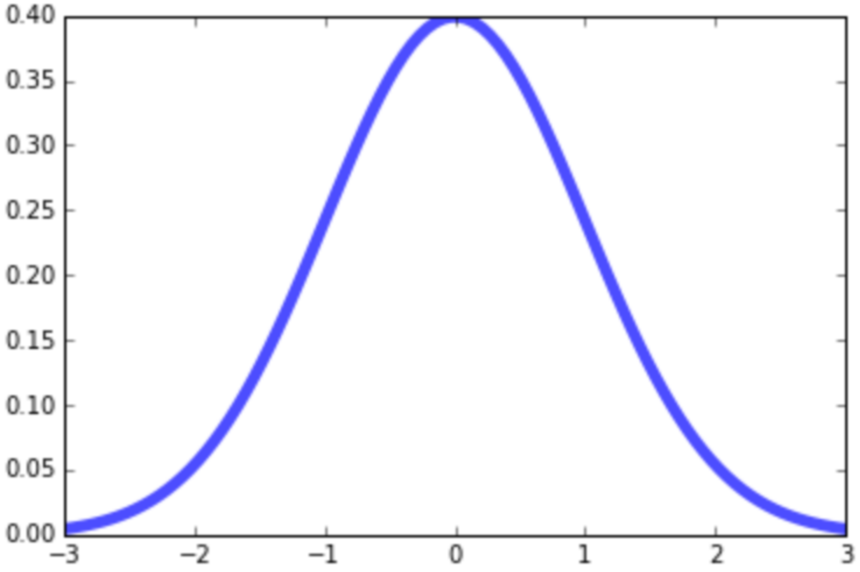


Разброс

- Отклонение времени начала лекции от 10:30
- $\xi \sim N(\mu, \sigma^2)$



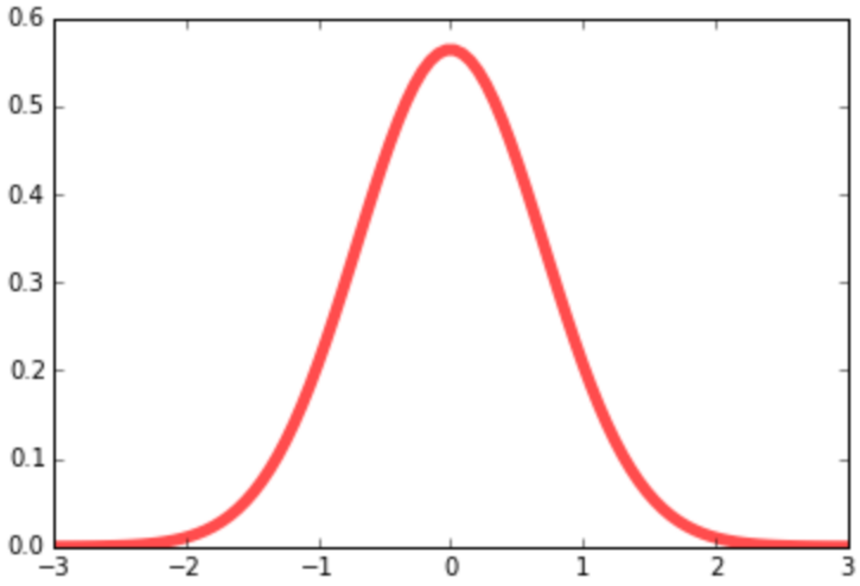
Лектор первого потока



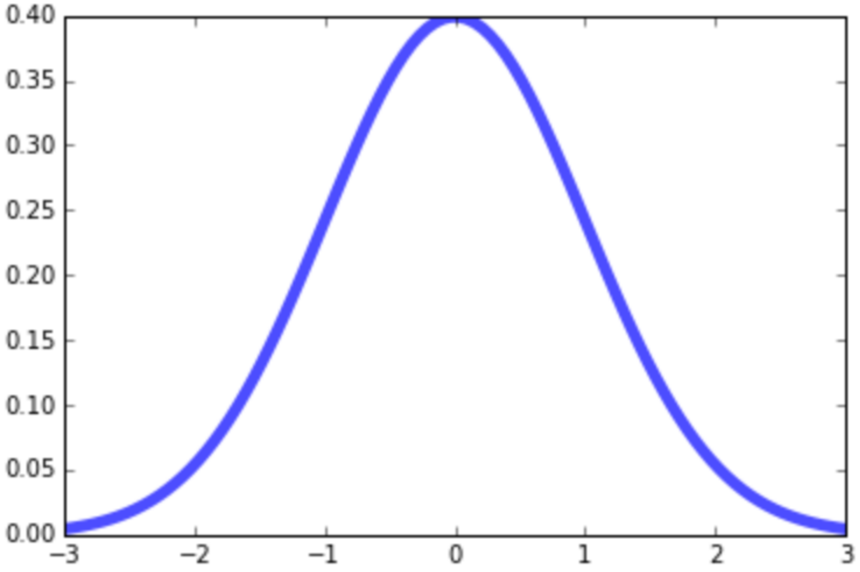
Лектор второго потока

Разброс

- Разброс на втором потоке выше!



Лектор первого потока



Лектор второго потока

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

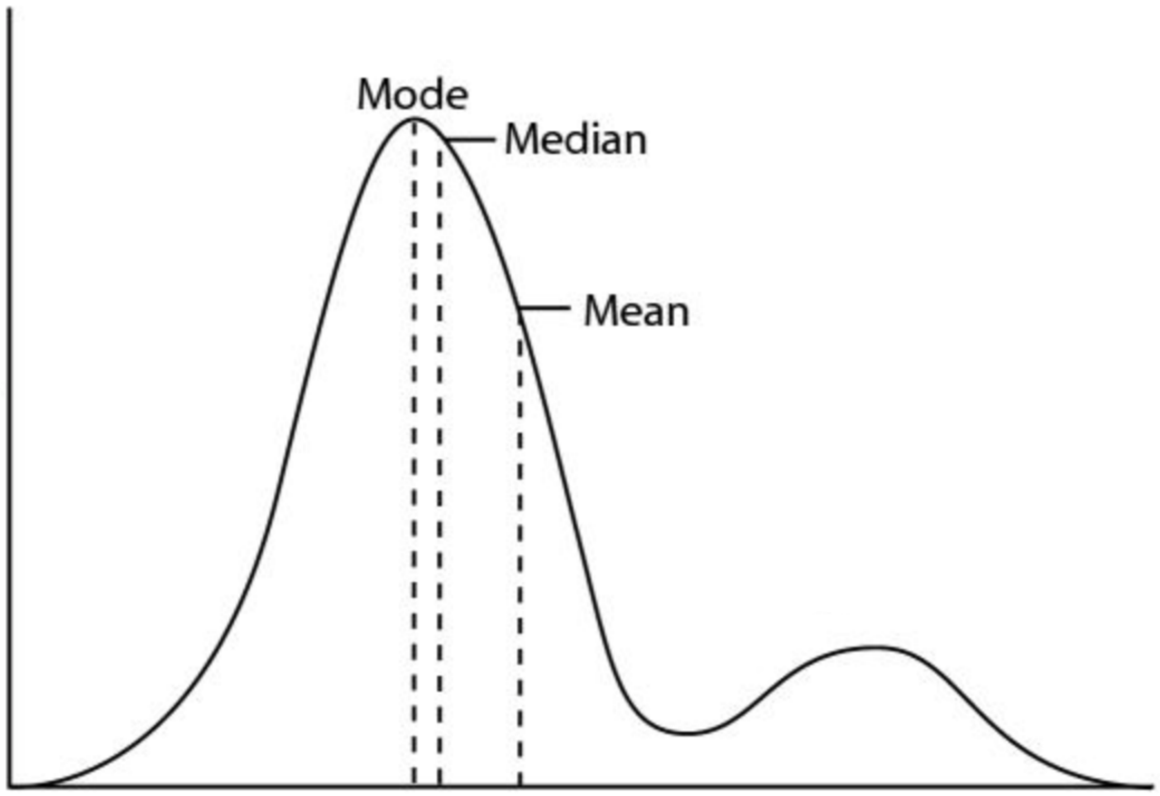
Медиана

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

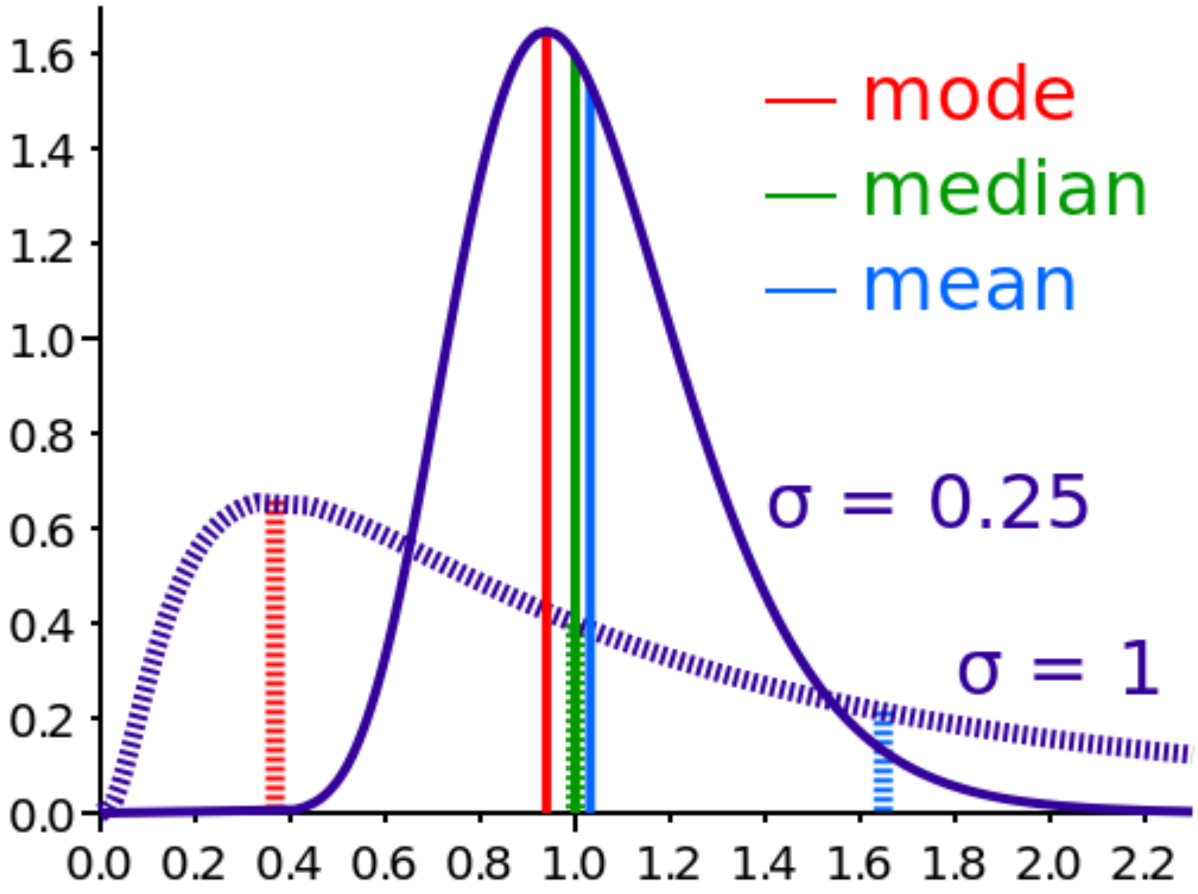
Мода

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

Средняя величина

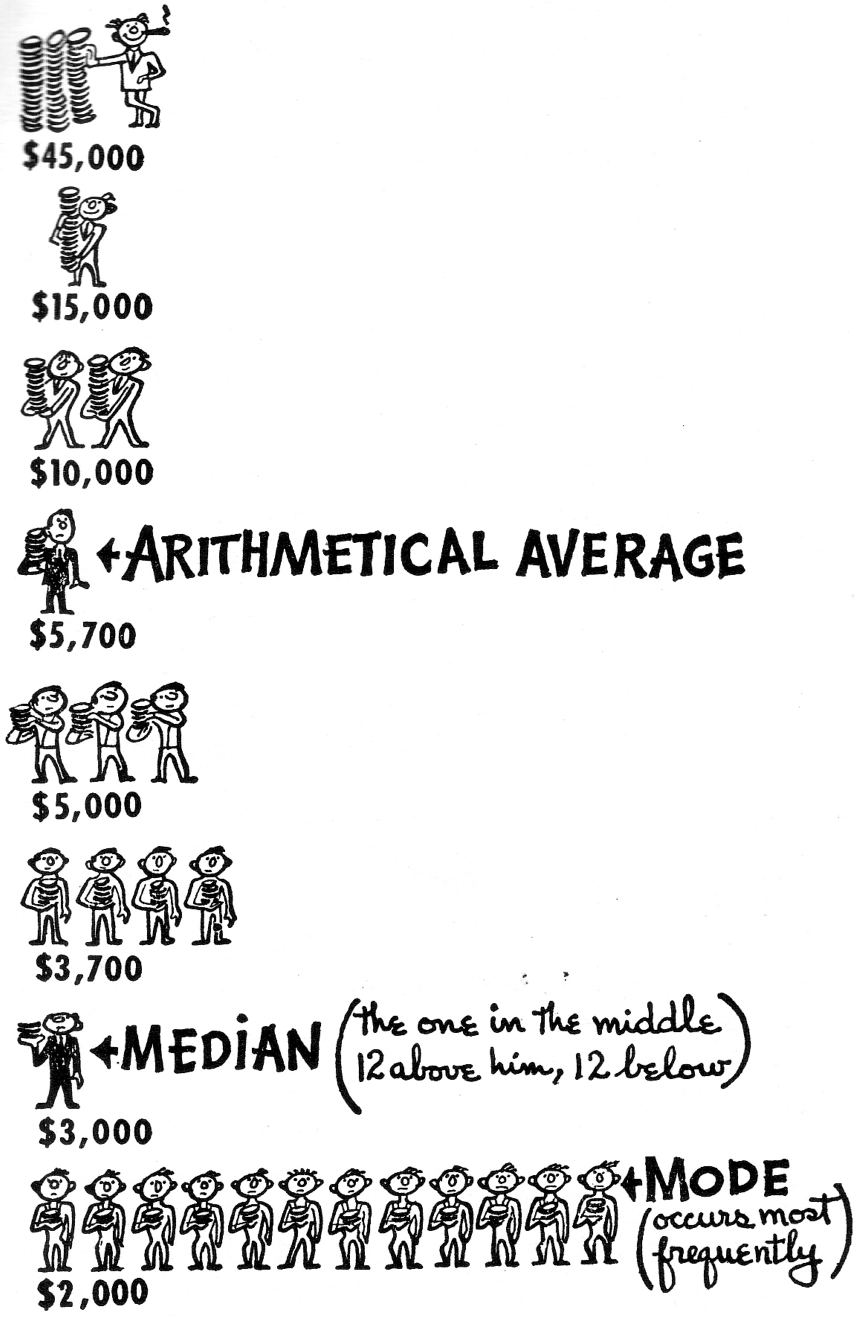


Средняя величина



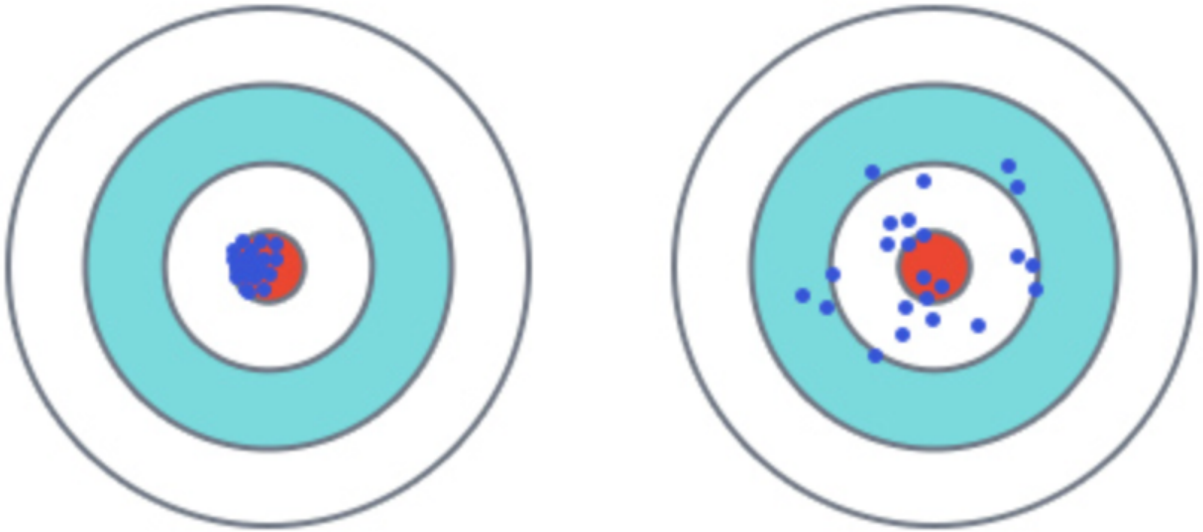
В чем разница?

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


$45,000
$15,000
$10,000
ARITHMETICAL AVERAGE
$5,700
$5,000
$3,700
MEDIAN
(the one in the middle
12 above him, 12 below)
$3,000
MODE
(occurs most
frequently)
$2,000

Дисперсия

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



Дисперсия

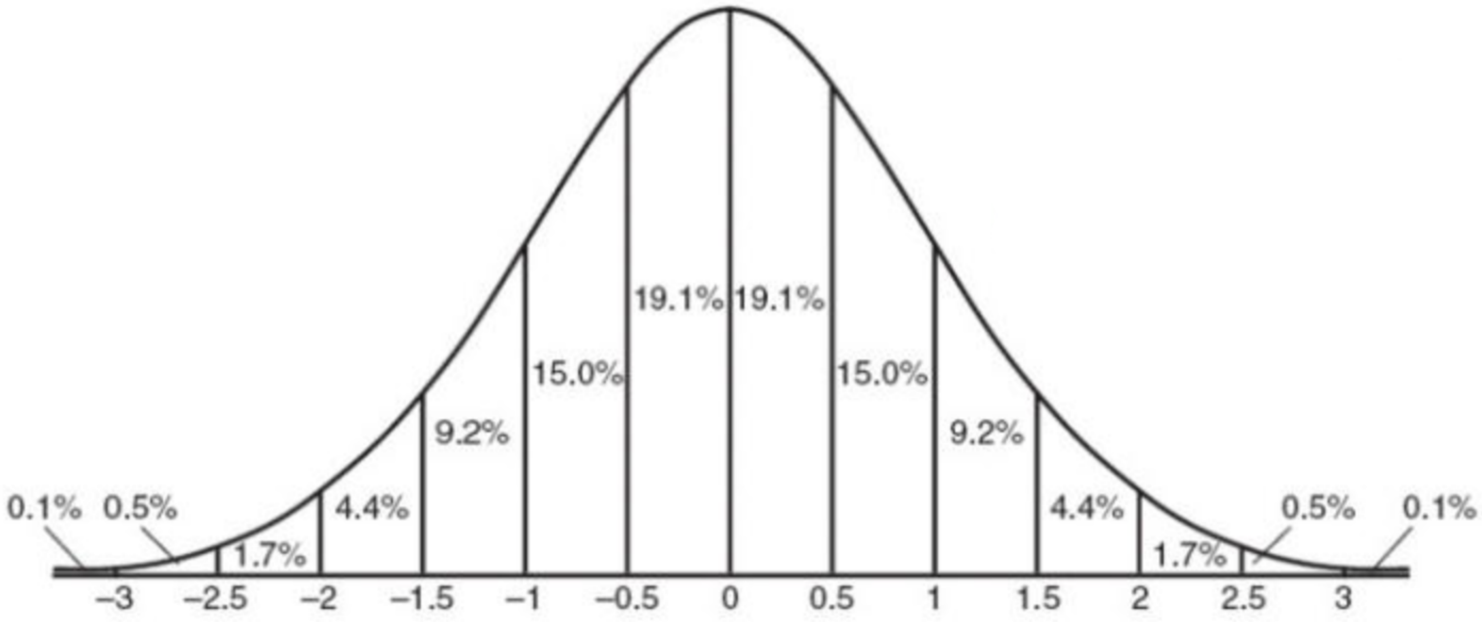
- Для $\mathrm{Pois}(\lambda)$: $\mathbb{D}\xi = \lambda$
- Для $\mathrm{Bin}(n, p)$: $\mathbb{D}\xi = np(1 - p)$
- Для $R[a, b]$: $\mathbb{D}\xi = (b - a)^2/12$
- Для $N(\mu, \sigma^2)$: $\mathbb{D}\xi = \sigma^2$

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 m, что вероятность попасть левее равна p
- Медиана — 0.5-квантиль



Квантиль

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

Интерквартильный размах

- Устойчивая к выбросам мера разброса:

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

- В нашем примере: $IQR = 0$

ЗБЧ и ЦПТ

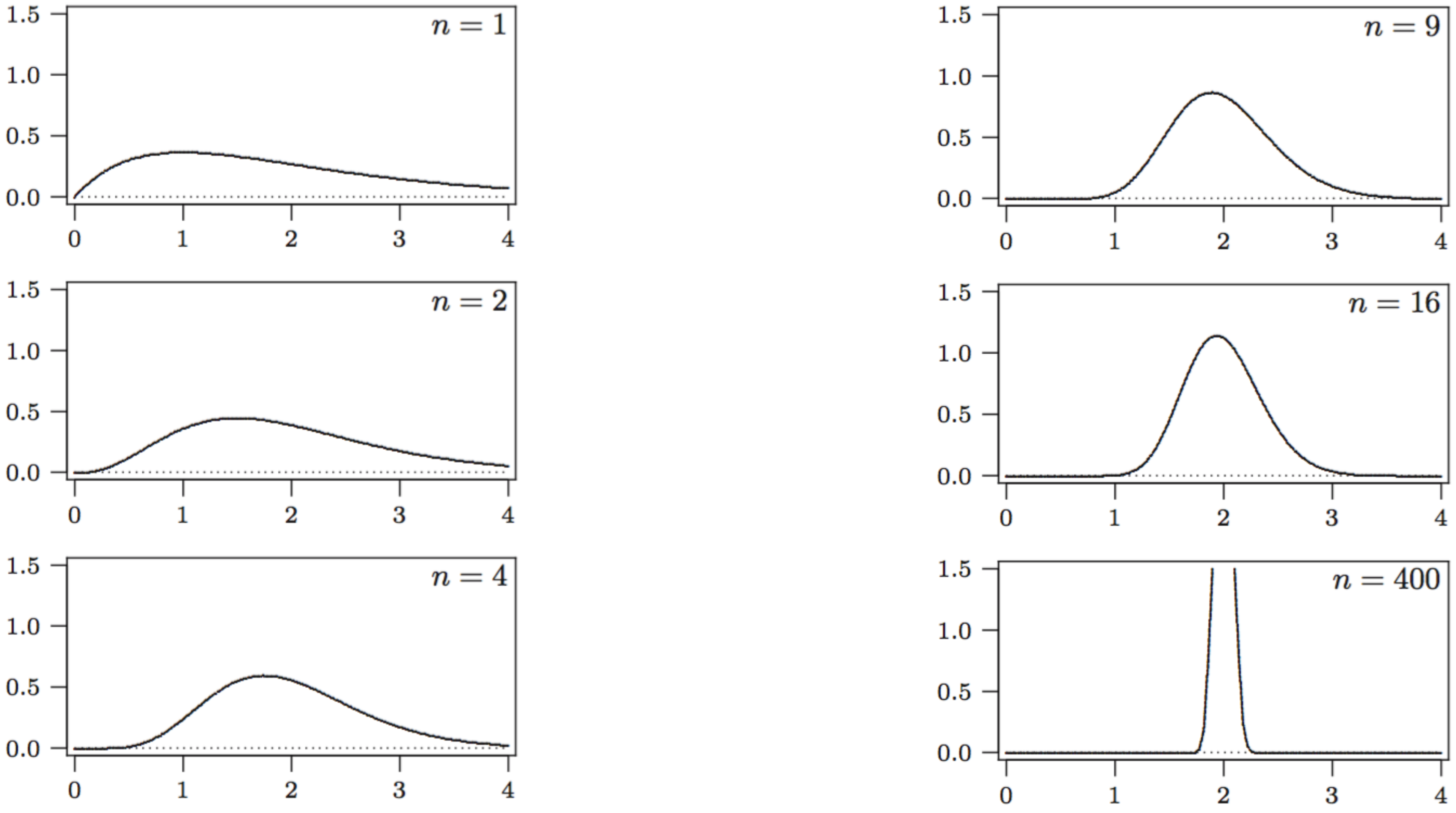
Усреднение наблюдений

- 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)
- $\mathbb{E}\xi_i = \mu, \mathbb{D}\xi_i = \sigma^2$
- $\overline{\xi_n} = \frac{1}{\mathrm{n}}(\xi_1 + \dots + \xi_n)$
- $\mathbb{E}\overline{\xi_n} = \mu, \mathbb{D}\overline{\xi_n} = \frac{\sigma^2}{n}$
- Усреднение уменьшает дисперсию!

Усреднение наблюдений



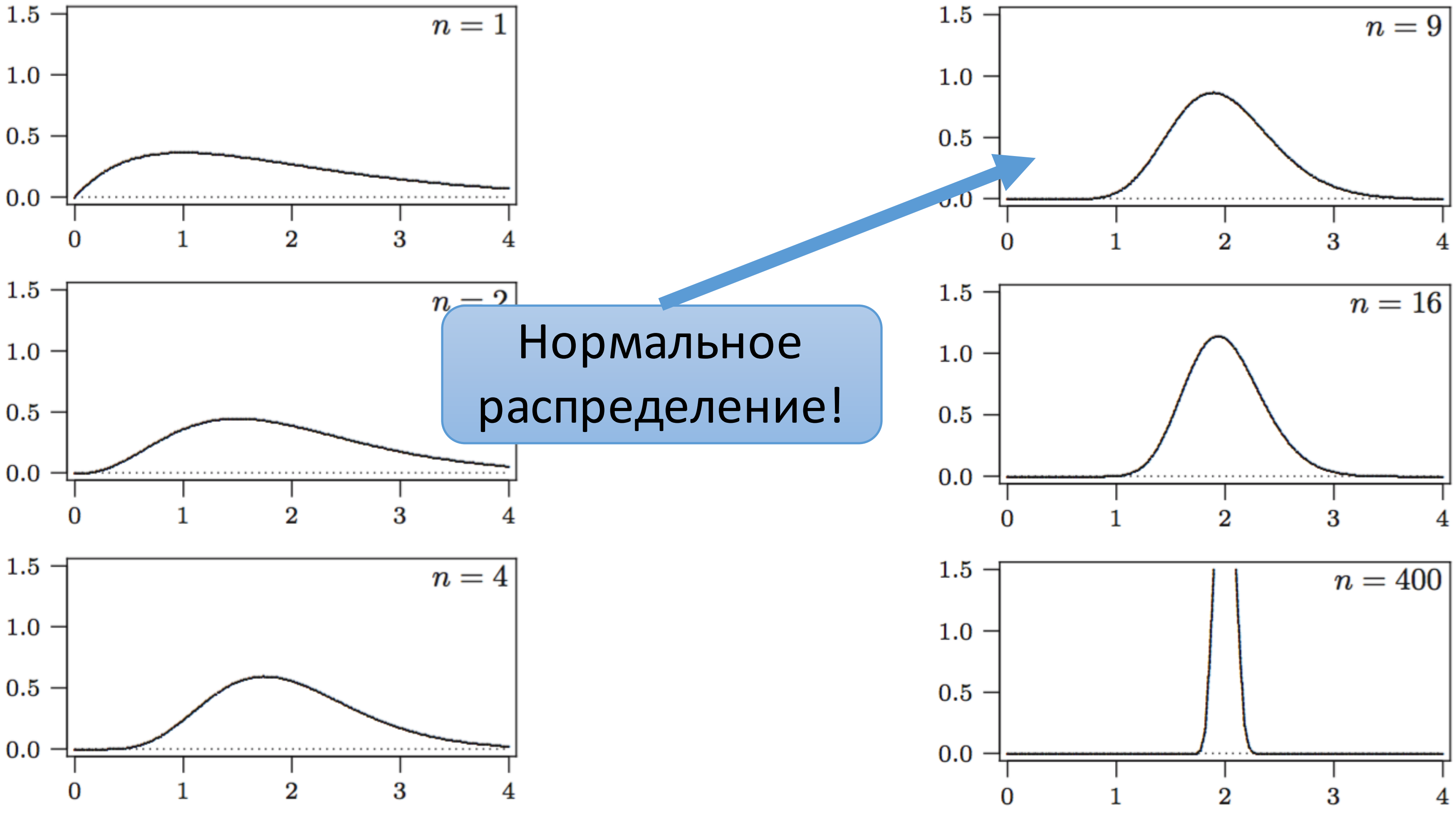
Закон больших чисел

- Среднее по наблюдениям стремится к матожиданию:

$$\lim_{n\to\infty} P\left(\left|\overline{\xi_n} - \mathbb{E}\xi_1\right| > \varepsilon\right) = 0$$

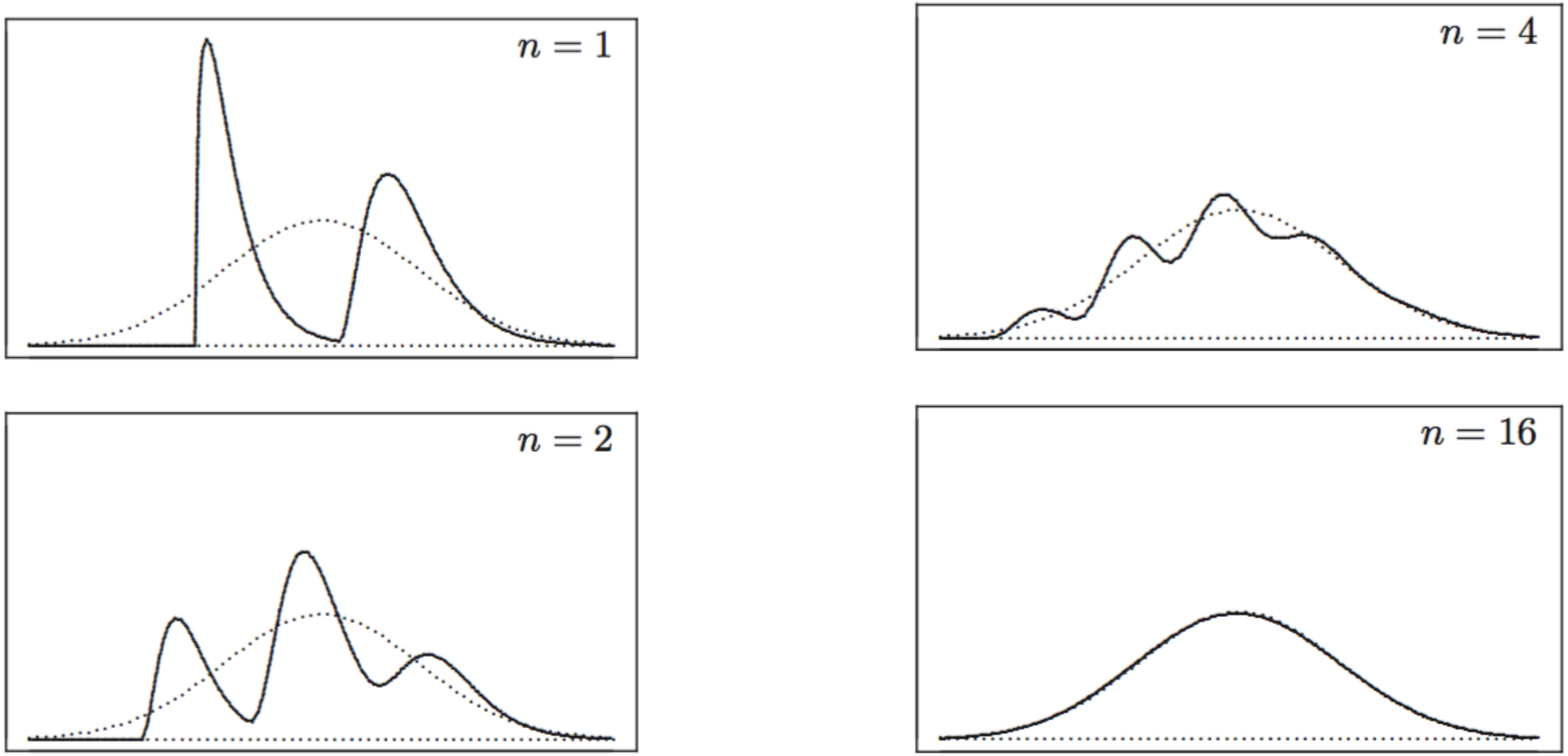
- Обосновывает утверждение «Вероятность события равна его доле в бесконечном числе экспериментов»

Усреднение случайных величин



Нормальное распределение!

Усреднение случайных величин



Центральная предельная теорема

- Распределение среднего нормированных величин стремится к нормальному:

$$\sqrt{n}\,\frac{\overline{\xi_n} - \mu}{\sigma} \to N(0,1)$$

Пример

- Бухгалтер решил округлять все числа до целых
- $\$99.53 \to \$100$
- $\$100.42 \to \$100$
- 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

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

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

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

Ответ: 0.0006

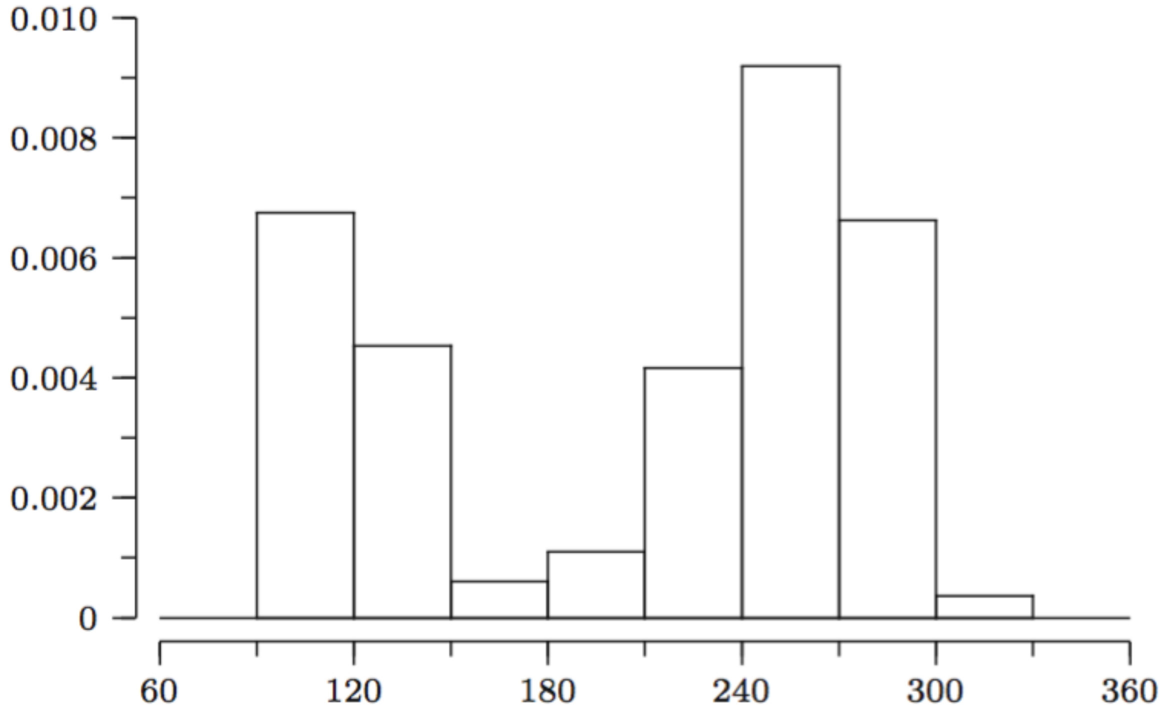
Визуализация

Одномерная выборка

- Old Faithful
- Длительность извержения гейзера

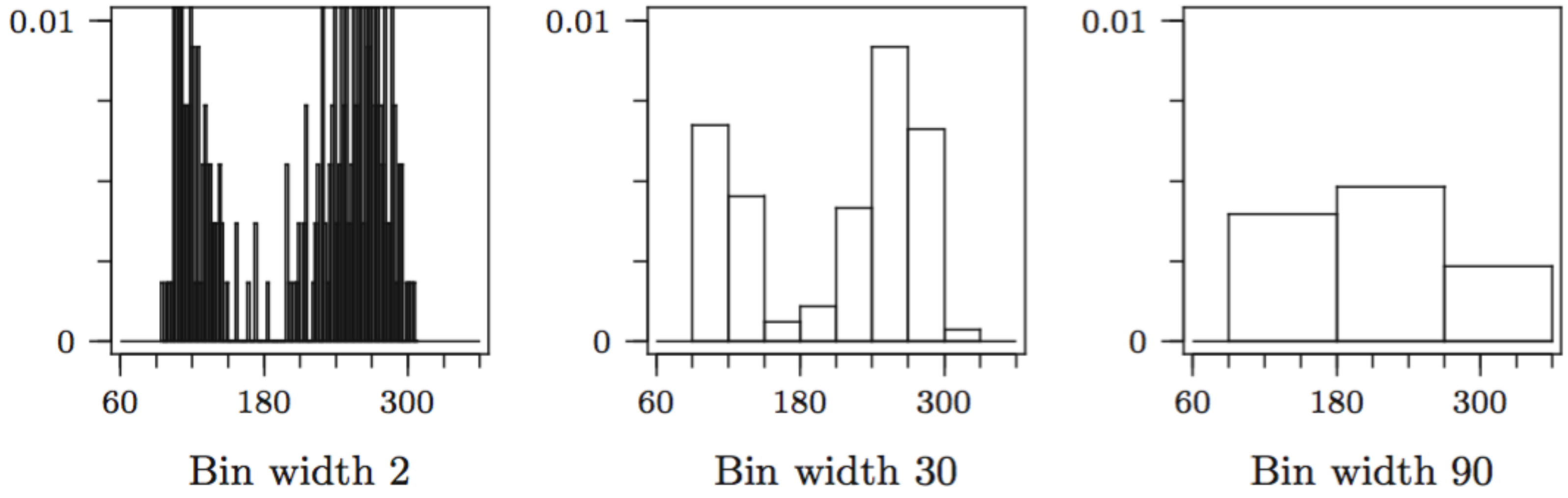
216	108	200	137	272	173	282	216	117	261
110	235	252	105	282	130	105	288	96	255
108	105	207	184	272	216	118	245	231	266
258	268	202	242	230	121	112	290	110	287
261	113	274	105	272	199	230	126	278	120
288	283	110	290	104	293	223	100	274	259
134	270	105	288	109	264	250	282	124	282
242	118	270	240	119	304	121	274	233	216
248	260	246	158	244	296	237	271	130	240
132	260	112	289	110	258	280	225	112	294
149	262	126	270	243	112	282	107	291	221
284	138	294	265	102	278	139	276	109	265
157	244	255	118	276	226	115	270	136	279
112	250	168	260	110	263	113	296	122	224
254	134	272	289	260	119	278	121	306	108
302	240	144	276	214	240	270	245	108	238
132	249	120	230	210	275	142	300	116	277
115	125	275	200	250	260	270	145	240	250
113	275	255	226	122	266	245	110	265	131
288	110	288	246	238	254	210	262	135	280
126	261	248	112	276	107	262	231	116	270
143	282	112	230	205	254	144	288	120	249
112	256	105	269	240	247	245	256	235	273
245	145	251	133	267	113	111	257	237	140
249	141	296	174	275	230	125	262	128	261
132	267	214	270	249	229	235	267	120	257
286	272	111	255	119	135	285	247	129	265
109	268								

Гистограмма



Гистограмма

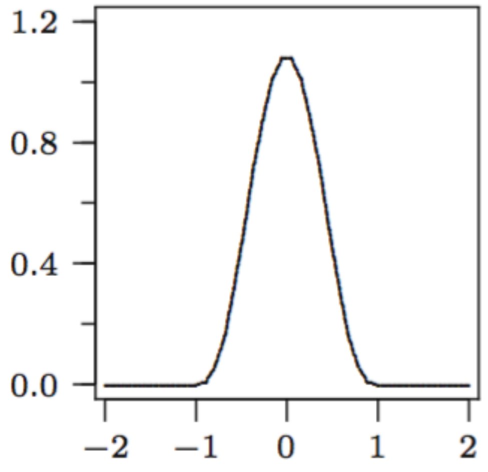
- Выбор ширины столбца:



Ядерное сглаживание

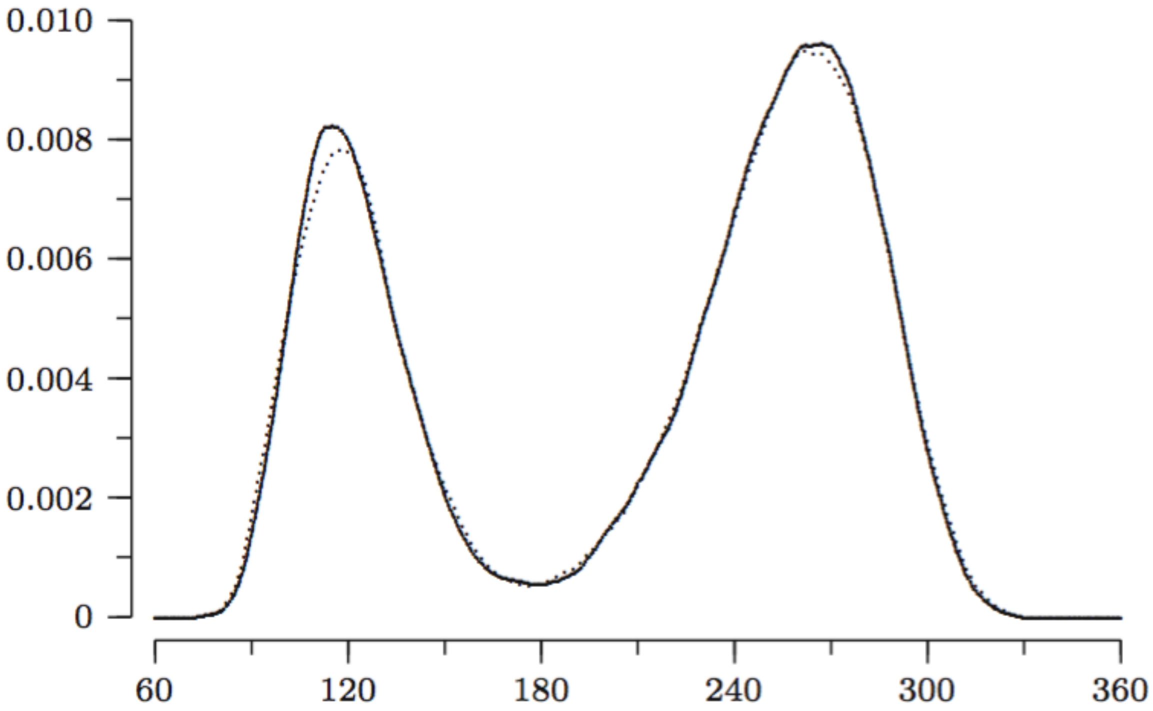
- Значение графика в точке — сумма взвешенных расстояний до наблюдений:

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

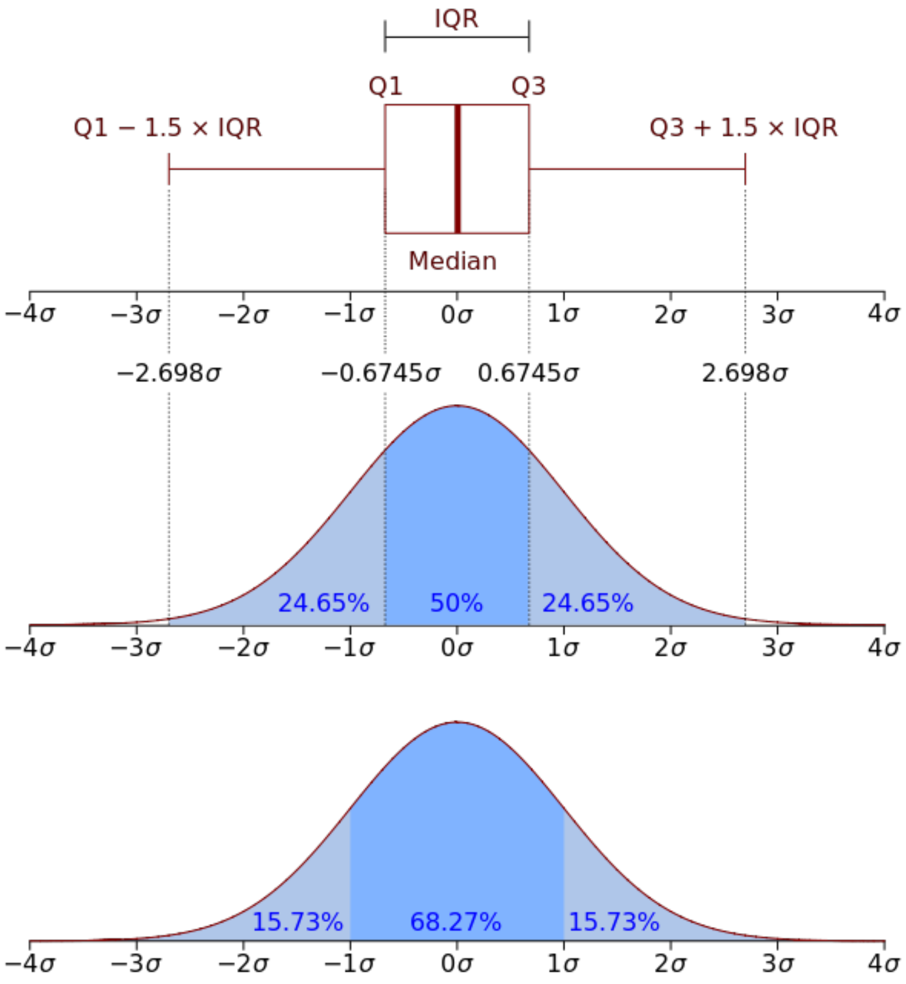


Triweight kernel

Ядерное сглаживание



Boxplot (ящик с усами)



Метод максимального правдоподобия

Правдоподобие

- Выборка: $X_1, X_2, \dots, X_n$
- Предполагаемое распределение: $X_i \sim P(\theta)$
- Задача: оценить параметр θ

Пример

- Выборка: $0, 1, 1, 1, 1, 0, 1, 0, 1, 1$ (подбрасывания монетки)
- Предполагаемое распределение: $X_i \sim \mathrm{Ber}(p)$
- Чему равен p?
- $p \approx \frac{7}{10}$
- Почему такая оценка имеет смысл?

Правдоподобие

- Выборка: $X_1, X_2, \dots, X_n$
- Предполагаемое распределение: $X_i \sim P(\theta)$
- Задача: оценить параметр θ
- Правдоподобие X_i: $P(X_i \mid \theta)$
- Правдоподобие выборки: $L(\theta) = P(X_1 \mid \theta)P(X_2 \mid \theta) \dots P(X_n \mid \theta)$
- Правдоподобие — вероятность получить нашу выборку при параметре θ

Максимизация правдоподобия

- Выбираем параметр, дающий наибольшую вероятность получения выборки:

$$L(\theta) \to \max_{\theta}$$

Пример

- Предполагаемое распределение: $X_i \sim \mathrm{Ber}(p)$
- $P(X_i = 1 \mid p) = p$
- $P(X_i = 0 \mid p) = 1 - p$
- $P(X_i \mid p) = p^{X_i}(1-p)^{1-X_i}$

$$L(p) = p^{\sum X_i}(1-p)^{n-\sum X_i}$$

$$\arg\max_p L(p) = \frac{\sum_{i=1}^{n} X_i}{n}$$

Классификация текстов

- Наивный байесовский классификатор
- Нужно оценить $p\left(x^j = 1 |\, y\right)$
- Оценка максимального правдоподобия:

$$p_{jy} = p\left(x^j = 1 |\, y\right) = \frac{\sum_{i=1}^{\ell}\left[x_i^j = 1\right][y_i = y]}{\sum_{i=1}^{\ell}[y_i = y]}$$

- Доля текстов с данным словом среди всех текстов класса

Резюме

- Непрерывные случайные величины
- Если распределение надо описать одним числом:
  - Матожидание, медиана, мода
  - Дисперсия, интерквартильный размах
- Уменьшение дисперсии: закон больших чисел
- Работа с суммой независимых факторов: ЦПТ
- Визуализация признаков: гистограммы и ядерное сглаживание
- Метод максимального правдоподобия

На следующей лекции

- Линейная регрессия
- Стохастический градиентный спуск
- Вещественные и категориальные признаки в линейных моделях
- Переобучение и регуляризация