

Занятие № 2

Статистические методы анализа данных

Евгений Некрасов

План занятия

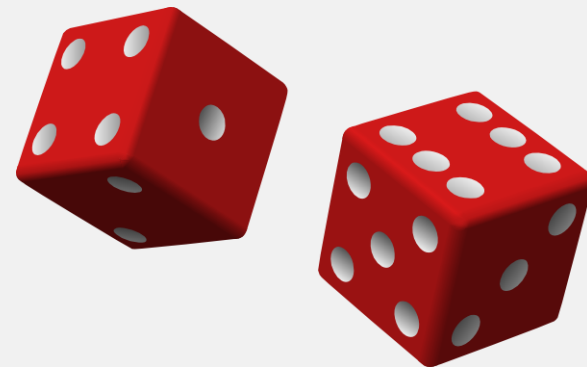


1. Теория вероятностей
2. Случайные величины
3. Распределения случайных величин
4. Выборки
5. Статистическое тестирование гипотез
6. Разбор домашнего задания



Теория вероятностей изучает закономерности случайных явлений при многократном повторении опыта.

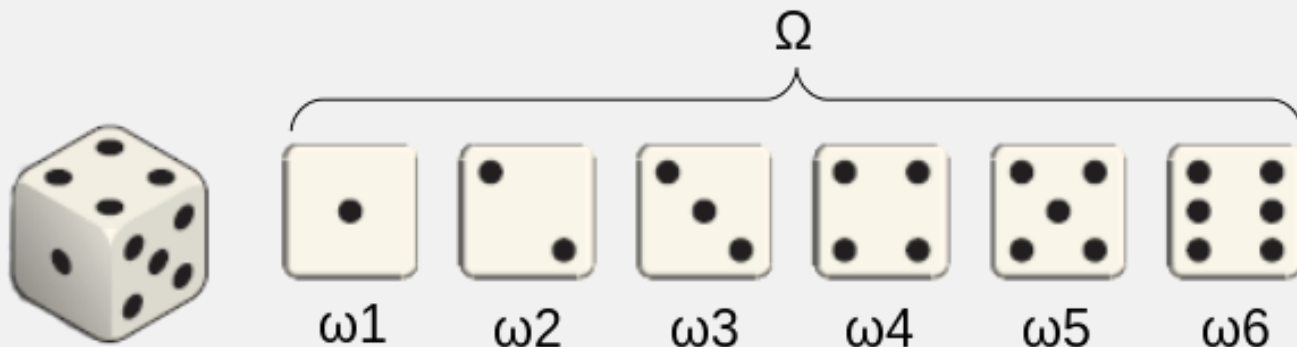
Опыт — комплекс условий.



Событие **случайно**, если при многократном воспроизведении опыта оно иногда происходит, а иногда нет.

Частота случайного события **устойчива**.

Пространство элементарных событий



ω – элементарные исходы, исходы случайного эксперимента, из которых в эксперименте происходит ровно один

Ω – пространство элементарных исходов

$A+B$ – произойдет хотя бы одно из событий

AB – произойдут оба события

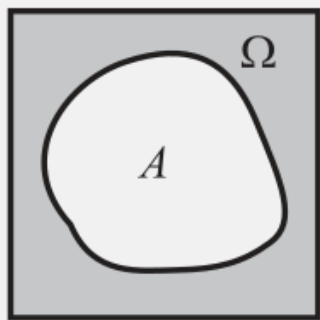
$A \setminus B$ – событие A произойдет, событие B нет

\overline{A} – событие A не произойдет

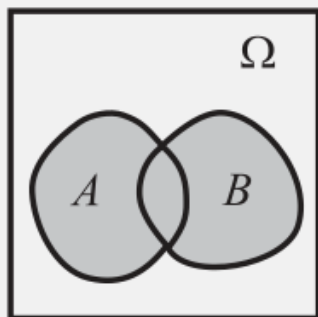
Свойства событий



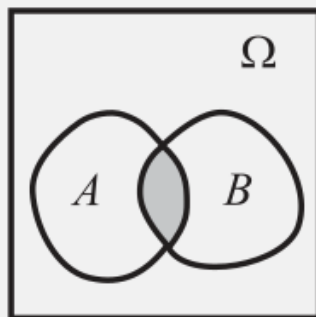
- 1) $\Omega + A = \Omega$;
- 2) $\Omega A = A$;
- 3) $AA = A$ (но не A^2);
- 4) $A + A = A$ (но не $2A$);
- 5) $A + \emptyset = A$;
- 6) $A\emptyset = \emptyset$;
- 7) $(A \setminus B)(B \setminus A) = \emptyset$;
- 8) $A + B = B + A$;
- 9) $AB = BA$;
- 10) $C(A + B) = CA + CB$;
- 11) $\overline{A + B} = \overline{A} \cdot \overline{B}$, $\overline{AB} = \overline{A} + \overline{B}$;
- 12) $A + \overline{A} = \Omega$, $\overline{\overline{A}} = A$.



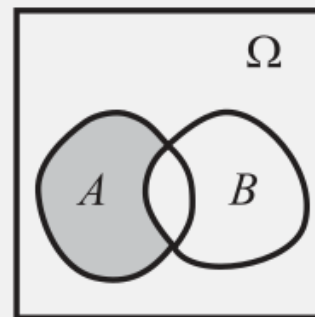
\overline{A}



$A+B$



AB



$A \setminus B$



Случайные события – элементы сигма-алгебры событий \mathcal{F} – множества подмножеств Ω .

Аксиомы:

1. \mathcal{F} является алгеброй событий
2. Каждому событию A из \mathcal{F} поставлено в соответствие $P(A) \geq 0$
3. $P(\Omega) = 1$
4. $P(A + B) = P(A) + P(B)$ – для несовместных событий

(Ω, \mathcal{F}, P) – вероятностное пространство

Свойства вероятности



- $P(\emptyset) = 0$
- $P(A) \leq 1$
- Если $A \subset B$, то $P(A) \leq P(B)$
- $P(A + B) = P(A) + P(B) - P(AB)$

Задача 1



Мы подбрасываем монетку 10 раз, какова вероятность, что все 10 раз выпадет орел?

Задача 2



Мы подбрасываем монетку 10 раз, какова вероятность, что орел выпадет ровно 8 раз?

Задача 3



Мы подбрасываем монетку 10 раз, какова вероятность, что орел выпадет ровно 8 раз, при этом монетка кривая и вероятность выпадения орла $q = 0.6$?

Задача 4



Если 1 раз не помыть руки перед едой вероятность заразиться 1%, какова вероятность заразиться, если не помыть руки перед едой 100 раз?

Условная вероятность



Условная вероятность $P(A|B)$ – вероятность события A при условии, что событие B уже произошло.

$$P(AB) = P(A)P(B|A) = P(B)P(A|B)$$

Формула умножения вероятностей



Вероятность одновременного появления нескольких событий можно вычислить по формуле умножения вероятностей:

$$\mathbf{P}(A_1 A_2 \cdot \dots \cdot A_n) = \mathbf{P}(A_1) \mathbf{P}(A_2 | A_1) \cdot \dots \cdot \mathbf{P}(A_n | A_1 \cdot \dots \cdot A_{n-1})$$

Формула сложения вероятностей



Вероятность появления хотя бы одного из нескольких событий вычисляют по формуле сложения вероятностей:

$$\mathbf{P}(A_1 + \dots + A_n) = \sum_{i=1}^n (-1)^{i-1} p_i,$$

$$\text{где } p_1 = \sum_{i=1}^n \mathbf{P}(A_i), \quad p_2 = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbf{P}(A_i A_j), \dots, \quad p_n = \mathbf{P}(A_1 \cdot \dots \cdot A_n)$$

$$\begin{aligned} \mathbf{P}(A_1 + A_2 + A_3) = & \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) - \\ & - \mathbf{P}(A_1 A_2) - \mathbf{P}(A_2 A_3) - \mathbf{P}(A_1 A_3) + \mathbf{P}(A_1 A_2 A_3) \end{aligned}$$

Формула полной вероятности



Гипотезами называют полную группу попарно не совместных событий.

$$- H_1 + \dots + H_n = \Omega$$

$$- H_i H_j = \emptyset$$

Тогда вероятность события A можно представить в виде:

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

Формула Байеса



Пусть известны вероятности гипотез $P(H_i)$ и условные вероятности $P(A|H_i)$

$$P(H_i)P(A|H_i) = P(A)P(H_i|A)$$

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{P(A)}$$

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

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{\sum_{k=1}^n P(H_k)P(A|H_k)}$$

Задача 5



Медицинский прибор диагностирует болезнь у больного с вероятностью 97%, и ошибочно диагностирует болезнь у здоровых людей с вероятностью 5%. Болезнь встречается с частотой 1:600 человек. Какова вероятность что человек действительно болен при условии, что прибор диагностировал болезнь?

Случайная величина



Доход = ставка + чаевые

Ставка – величина обычная

Чаевые – **случайная величина**

Чаевые за конкретный месяц –
реализация случайной величины



Случайная величина



Пусть задано вероятностное пространство (Ω, \mathcal{F}, P) .

Случайной величиной (СВ) называется функция $X : \Omega \rightarrow \mathbb{R}$

Реализацию СВ будем обозначать x

Поведение независимой СВ полностью определяется законом распределения

Закон распределения



$F(x) = P(x < X)$ – **функция распределения СВ**

$f(x) = P(x = X)$ – **функция вероятности для дискретных СВ**

$f(x) = P(x < X \leq x + \Delta x) / \Delta x, \Delta x \rightarrow 0$ – **функция плотности вероятности для непрерывных СВ**

$$f(x) = F'(x)$$

$$F(x) = \int_{-\infty}^x f(t) dt$$

Математическое ожидание СВ



$$M[X] = \sum_{i=1}^{\infty} x_i p_i$$

$$M[X] = \int_{-\infty}^{\infty} x f(x) dx$$

$$M[cX] = cM[X]$$

$$M[X + c] = M[X] + c$$

Дисперсия СВ



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

$$D[X] = \int_{-\infty}^{\infty} f(x) (x - M[X])^2 dx$$

$$D[X] = M[X^2] - (M[X])^2$$

$$D[cX] = c^2 D[X]$$

$$D[-X] = D[X]$$

$$D[X + c] = D[X]$$

Биномиальное распределение



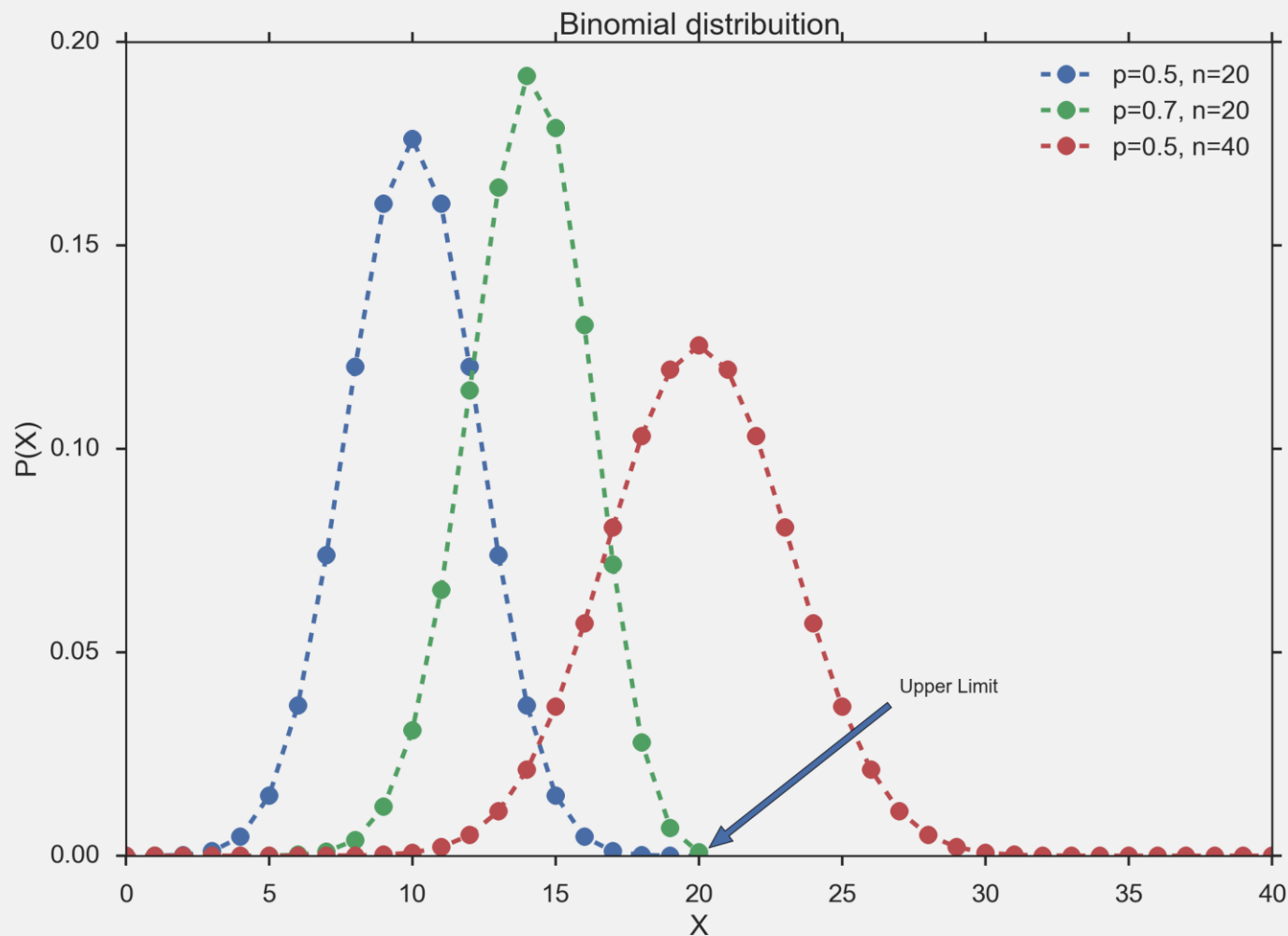
Моделирует количество «успехов» в последовательности из n независимых случайных экспериментов, таких, что вероятность «успеха» в каждом из них постоянна и равна p .

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

$$M[X] = np$$

$$D[X] = np(1 - p)$$

Биномиальное распределение



Распределение Пуассона



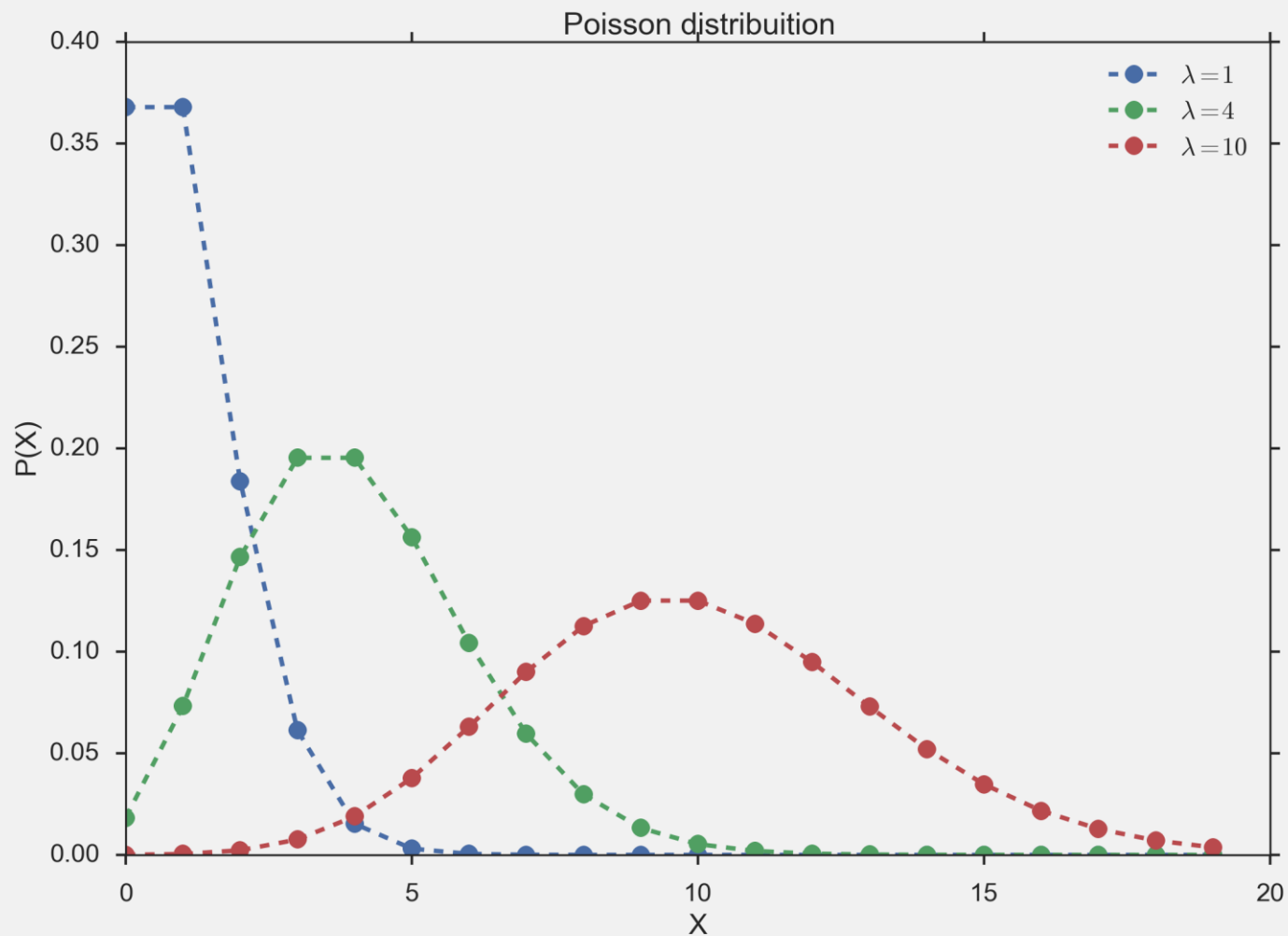
Моделирует число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга.

$$p(k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

$$M[X] = \lambda$$

$$D[X] = \lambda$$

Распределение Пуассона



Экспоненциальное распределение



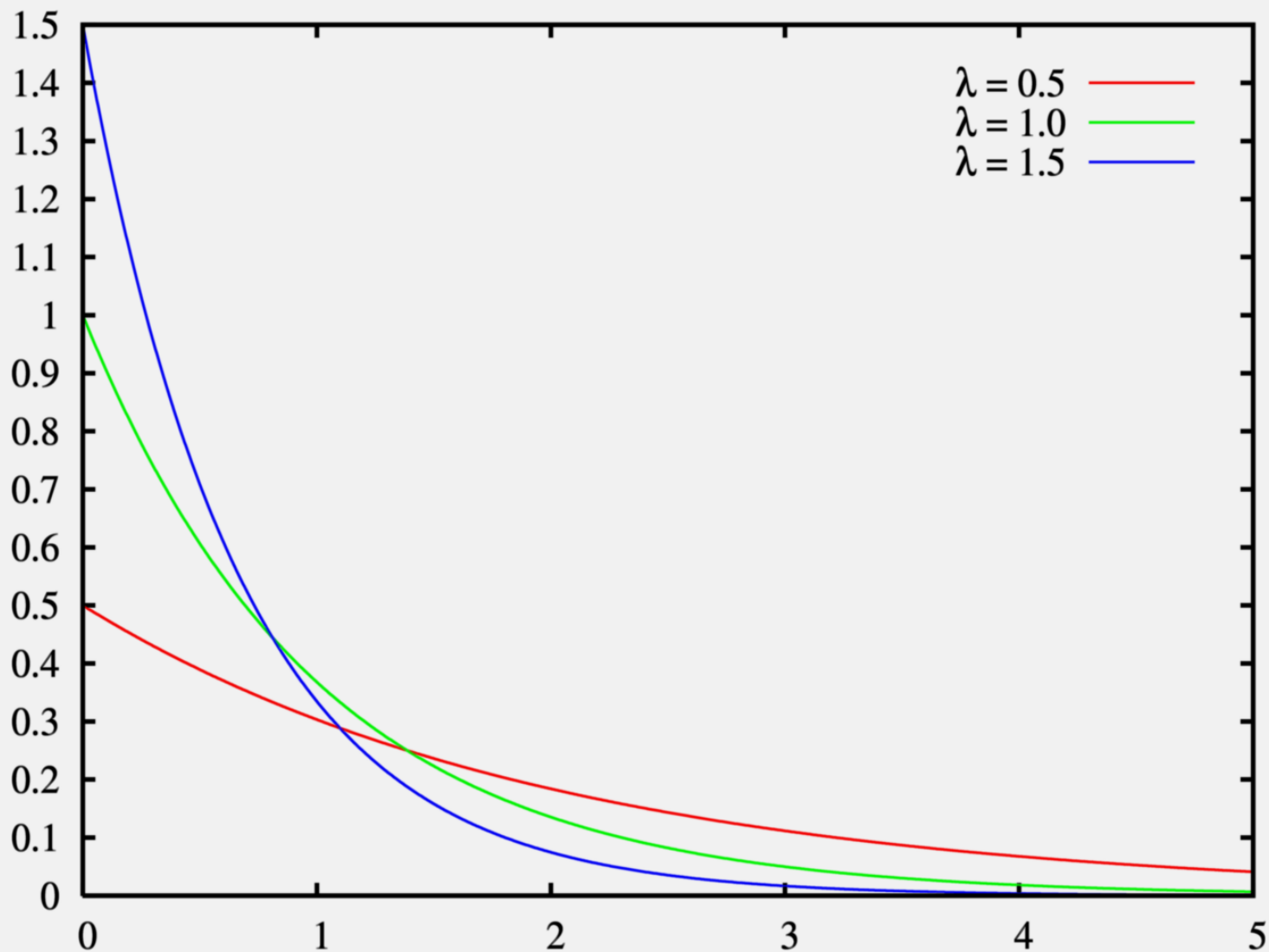
Моделирует время между двумя последовательными свершениями события в Пуассоновском процессе.

$$f(x) = \lambda e^{-\lambda x}$$

$$M[X] = \frac{1}{\lambda}$$

$$D[X] = \frac{1}{\lambda^2}$$

Экспоненциальное распределение



Нормальное распределение



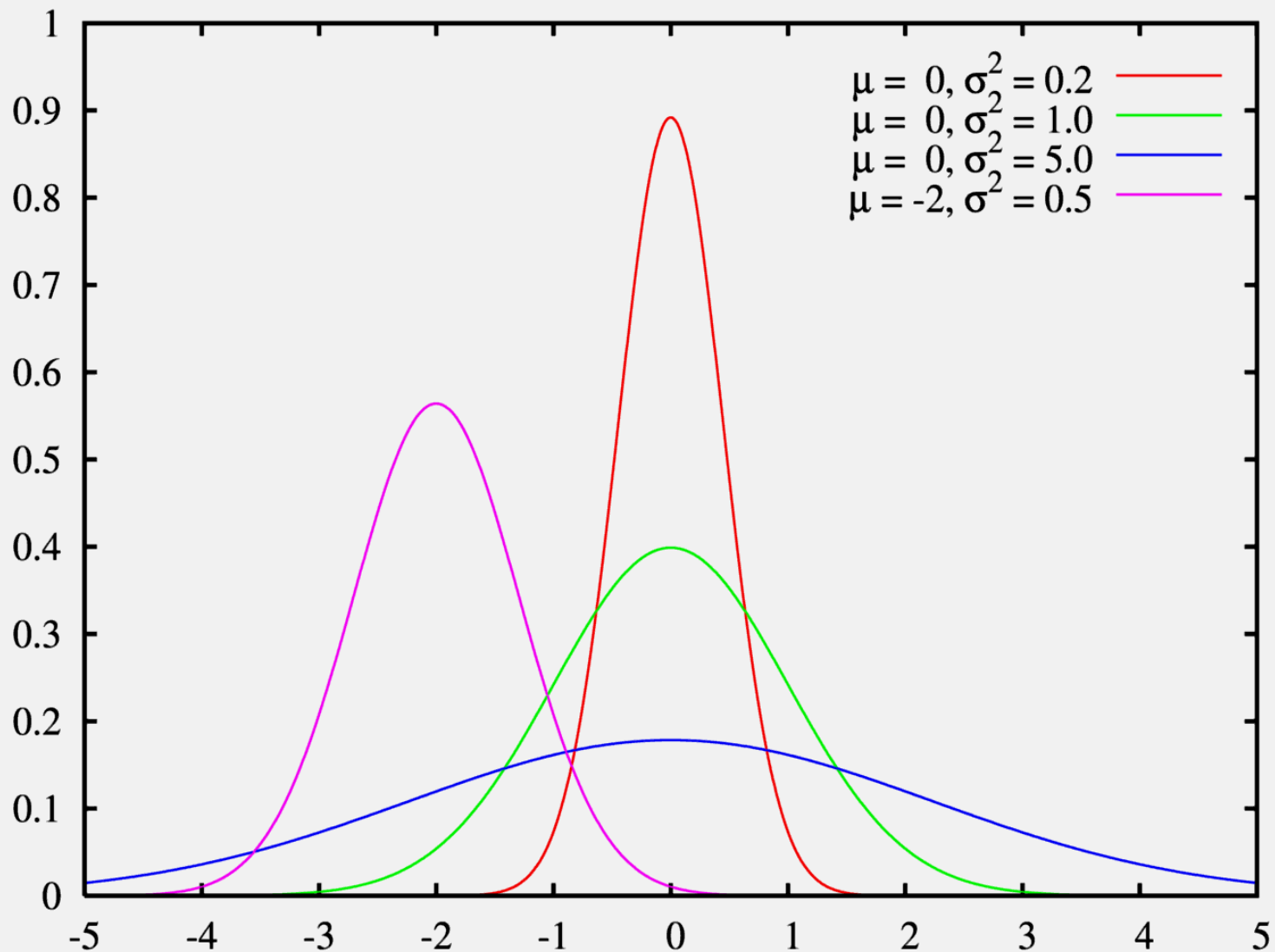
Моделирует результат суммы многих случайных слабо зависимых СВ, каждая из которых вносит малый вклад относительно общей суммы.

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

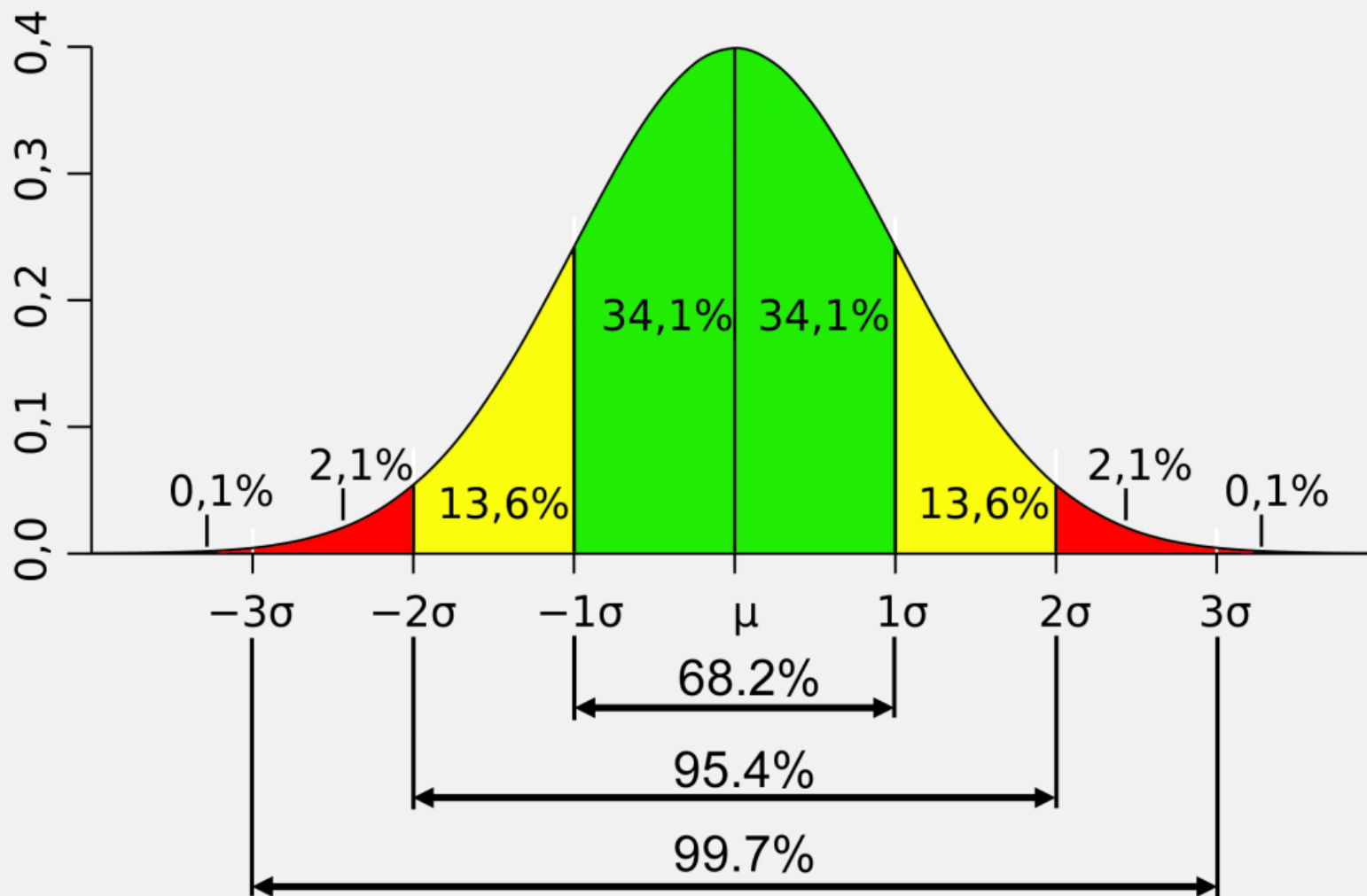
$$M[X] = \mu$$

$$D[X] = \sigma^2$$

Нормальное распределение



Нормальное распределение



Понятие выборки



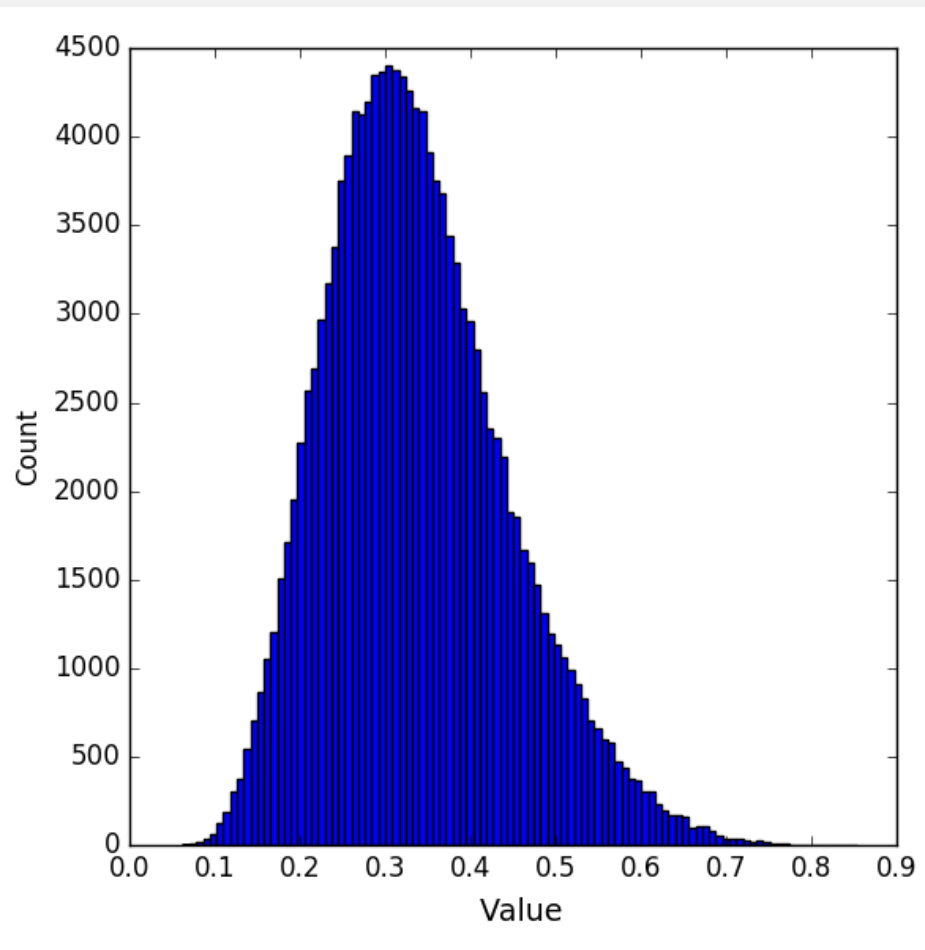
Пусть есть некоторая случайная величина X , вектор из n реализаций X будем называть **выборкой**.

Выборку можно также рассматривать как единственную реализацию случайного вектора из n одинаковых независимых случайных величин X .

Гистограмма



Гистограмма – геометрическое изображение эмпирической функции плотности вероятности некоторой случайной величины, построенное по выборке.



Оценка СВ по выборке



По выборке можно оценить математическое ожидание и дисперсию СВ следующим образом:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

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

Квантиль и перцентиль



Квантиль – значение, которое заданная СВ не превышает с фиксированной вероятностью.

Перцентиль-25 – нижний квартиль

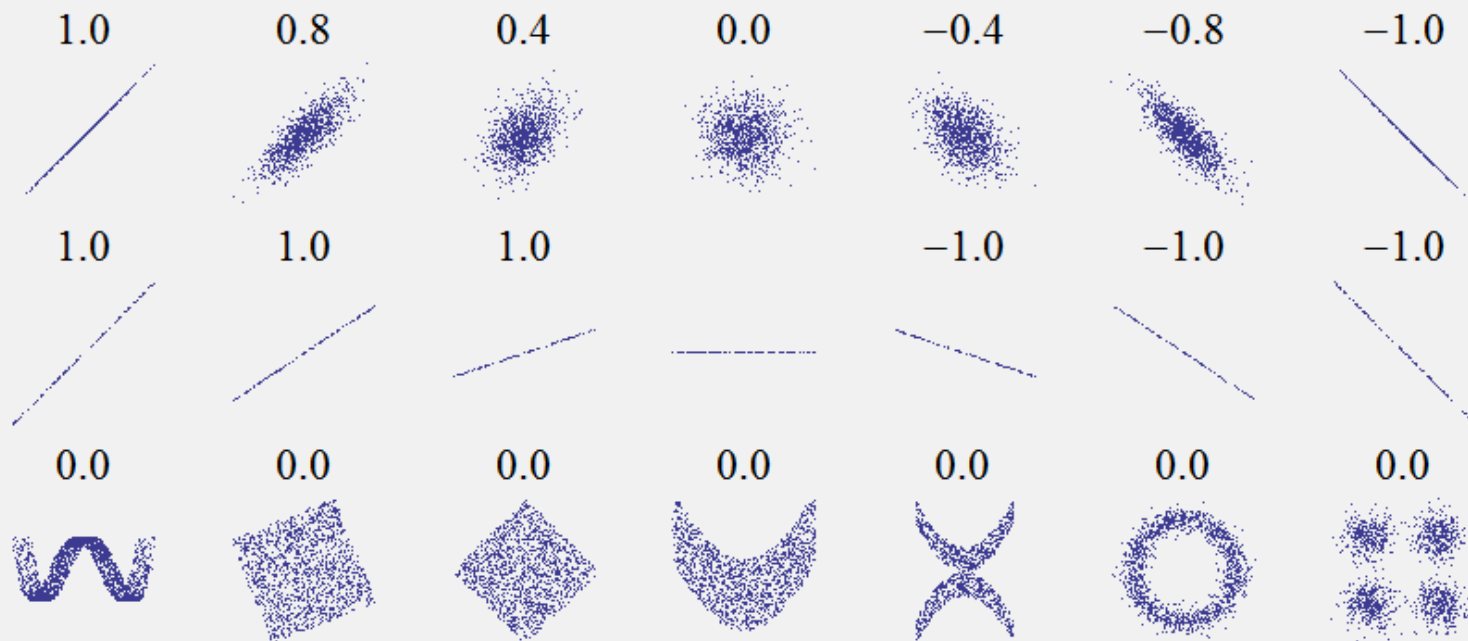
Перцентиль-50 – медиана

Перцентиль-75 – верхний квартиль

Корреляция



$$r_{xy} = \frac{M[XY] - M[X]M[Y]}{\sqrt{D[X]D[Y]}} = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m (x_i - \bar{x})^2 \sum_{i=1}^m (y_i - \bar{y})^2}}$$





Статистической гипотезой H называется любое предположение относительно параметров или закона распределения СВ, проверяемое по выборке.

Проверяемая гипотеза называется **нулевой** и обозначается H_0 .

Гипотеза, конкурирующая с H_0 называется **альтернативной** и обозначается H_1 .

Нулевая и альтернативная гипотезы представляют полную группу несовместных событий отклонение одной влечет принятие другой.

Какие бывают гипотезы?



- О значении параметра
- О виде закона распределения
- О независимости двух СВ
- Об однородности наблюдений

Ошибки I и II рода



Статистическая ошибка I рода – обнаружение различий или связей, которые на самом деле не существуют.

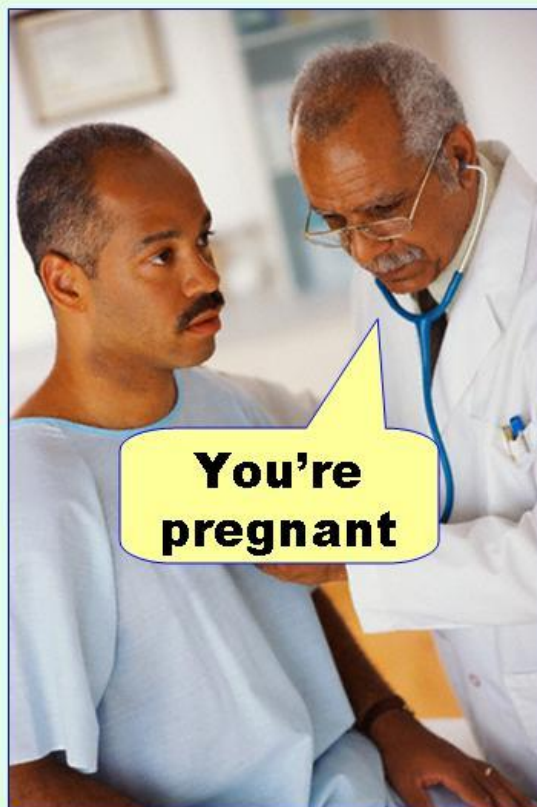
Статистическая ошибка II рода – не обнаружение различий или связей, которые на самом деле существуют.

Более критичной ошибкой считается статистическая ошибка первого рода.

Ошибки I и II рода



Type I error
(false positive)



Type II error
(false negative)



Подход к проверке статистической гипотезы



1. Формулировка основной гипотезы H_0 и альтернативной гипотезы H_1 .
2. Выбор подходящего статистического теста.
3. Задание уровня значимости α , вероятности допустить ошибку первого рода.
4. Вычисление эмпирического значения критерия по тесту
5. Сравнение с критическим значением критерия по тесту и принятие решения

Одновыборочный t-критерий



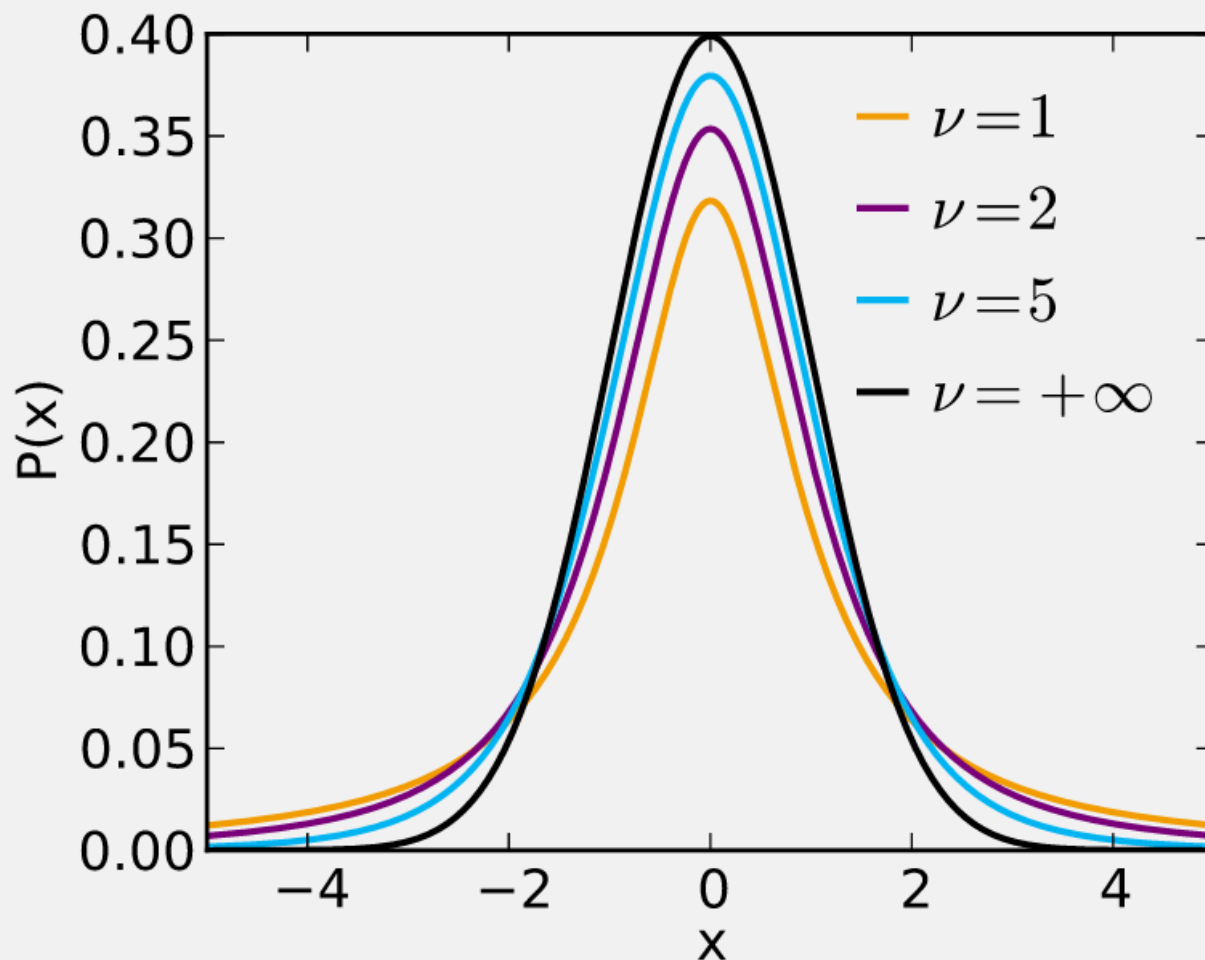
Применяется для проверки нулевой гипотезы H_0 о равенстве математического ожидания некоторому известному значению μ .

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

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

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}; \quad df = n - 1$$

Распределение Стьюдента



Двухвыборочный t-критерий для выборок с одинаковой дисперсией



Применяется для проверки нулевой гипотезы H_0 о равенстве математических ожиданий двух независимых выборок с одинаковой дисперсией.

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}; \quad df = n_1 + n_2 - 2$$

Семинар

Соревновательный анализ данных



Участники получают 2 выборки:

Train – выборка с разметкой

Test – выборка без разметки

Цель:

Максимально точно относительно метрики соревнования выполнить предсказание разметки для **Test**

Решение – это файл с предсказанной разметкой

Public и Private leaderboard



Выборку **Test** делят на 2 части:

Public – используется для оценки качества решения в процессе соревнования, доступна сразу после отправки решения

Private – используется для финальной оценки решений и определения победителей, доступна только после завершения соревнования

Существует лимит на количество отправленных решений

Соревновательный анализ данных



Платформы:

- kaggle.com
- topcoder.com
- drivendata.org
- dataring.ru

В среднем соревнование длится несколько месяцев

Соревнование "Introduction to machine learning. Autocompletion."



- предоставлен набор словосочетаний из двух слов, которые были набраны из русскоязычных текстов
- на основе предоставленного набора тренировочных словосочетаний надо построить алгоритм дополнения второго слова в словосочетании, когда известно первое слово и не менее двух первых букв второго слова.
- Train 0.5M, Test 0.5M
- Public/Private split 50/50%
- Метрика accuracy

Данные



train.csv

Id	Sample	Prediction
0	очень тру	очень трудно
1	гусеница вступ	гусеница вступила
2	мне не	мне нельзя
3	спасении гр	спасении грешной
4	почуяла чужо	почуяла чужого
5	обучением сол	обучением солдат

test.csv

Id	Sample
500000	пересекли необъятн
500001	были вм
500002	просто си
500003	прядку грив
500004	пока недоступн
500005	который каже

submit.csv

Id	Prediction
500000	пересекли необъятные
500001	были вмятины
500002	просто синий
500003	прядку гривы
500004	пока недоступного
500005	который кажется

Пример решения задачи



См. код

Домашнее задание №#2



- Зарегистрироваться на kaggle.com
- Сделать сабмит решения
- Выложить решение на github.com
- Прислать ссылку на код решения
- Прислать ссылку на свой профиль [kaggle](https://kaggle.com)

Срок сдачи

15 октября 2017



Спасибо за внимание!

Евгений Некрасов

e.nekrasov@corp.mail.ru