ТЕХНОСФЕРА

Занятие № 2

Статистические методы анализа данных

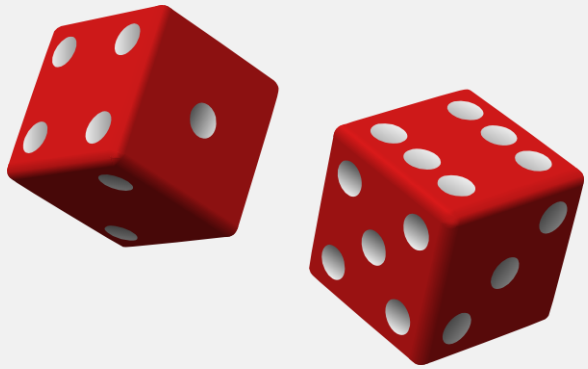
Евгений Некрасов

План занятия

1. Теория вероятностей
2. Применение теории вероятностей к задаче машинного обучения
3. Случайные величины
4. Распределения случайных величин
5. Выборки
6. Статистическое тестирование гипотез
7. Соревновательный анализ данных
8. Объяснение домашнего задания

Теория вероятностей

Теория вероятностей изучает закономерности случайных явлений при многократном повторении опыта.

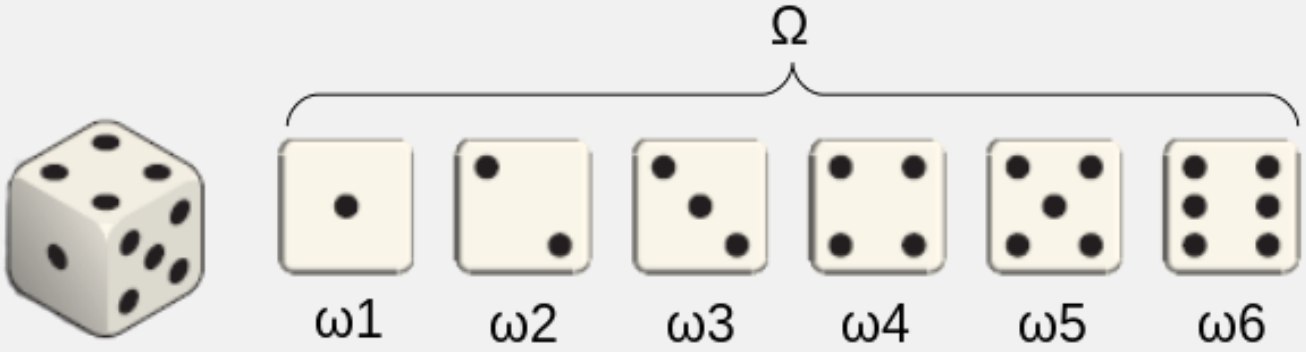


Опыт – комплекс условий.

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

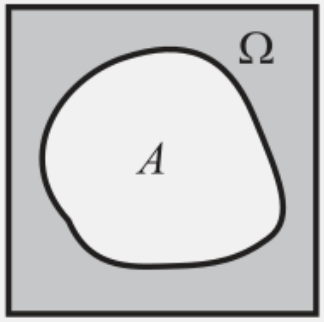
AB – произойдут оба события

A\B – событие A произойдет, событие B нет

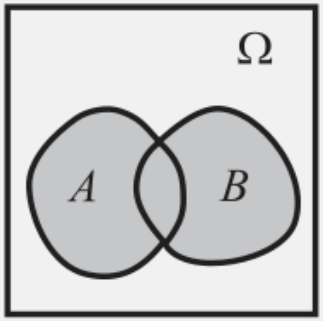
$\overline{A}$ – событие A не произойдет

Свойства событий

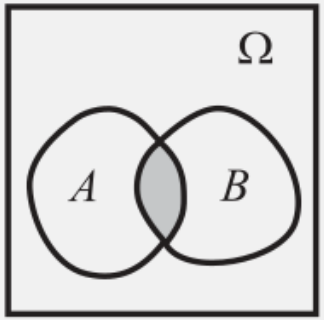
1) $\Omega + A = \Omega$;
2) $\Omega A = A$;
3) $AA = A$ (но не A^2);
4) $A + A = A$ (но не $2A$);
5) $A + \varnothing = A$;
6) $A\varnothing = \varnothing$;
7) $(A \setminus B)(B \setminus A) = \varnothing$;
8) $A + B = B + A$;
9) $AB = BA$;
10) $C(A + B) = CA + CB$;
11) $\overline{A + B} = \overline{A}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



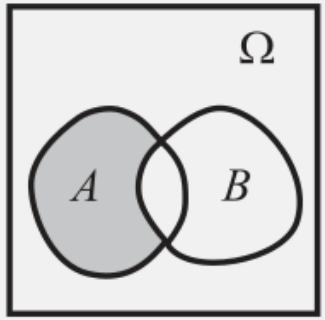
$\overline{A}$



$A+B$



$A\,B$



$A \setminus B$

Вероятностное пространство

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

$(\Omega, \mathcal{F}, P)$ – **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 1

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

$$N = 2^{10}$$

$$P(\omega) = \frac{1}{N} = \frac{1}{1024}$$

$$M = 1$$

$$P(X) = P(\omega) \cdot M = \frac{M}{N} = \frac{1}{1024}$$

Задача 2

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

Задача 2

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

$$N = 2^{10}$$

$$P(\omega) = \frac{1}{N} = \frac{1}{1024}$$

$$M = C_n^k = \frac{n!}{k!(n-k)!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot (8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)} = 45$$

$$P(X) = P(\omega) \cdot M = \frac{M}{N} = \frac{45}{1024} \approx 0.044$$

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 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 ?

$$N = 2^{10}$$

$$M = C_n^k = 45$$

$$P(\omega') = q^8 \cdot (1-q)^2$$

$$P(X) = P(\omega') \cdot M \approx 0.12$$

Задача 4

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

Задача 4

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

$$N = 2^{100}; q = 0.01;$$

$$M = 1$$

$$P(\omega') = q^0 \cdot (1 - q)^{100}$$

$$P(X) = 1 - P(\omega') \cdot M \approx 0.63$$

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:

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

События называются независимыми, если P(AB) = P(A)*P(B) = P(A)P(B|A) = P(B)P(A|B)

Формула сложения вероятностей

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

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

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

$$\mathbf{P}(A_1 + A_2 + A_3) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) - \\ - \mathbf{P}(A_1 A_2) - \mathbf{P}(A_2 A_3) - \mathbf{P}(A_1 A_3) + \mathbf{P}(A_1 A_2 A_3)$$

Формула полной вероятности

Гипотезами называют полную группу попарно не совместных событий.

- $H_1 + \ldots + H_n = \Omega$
- $H_i H_j = \varnothing$

Тогда вероятность события A можно представить в виде:

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

Формула Байеса

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

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

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

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

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

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?

Задача 5

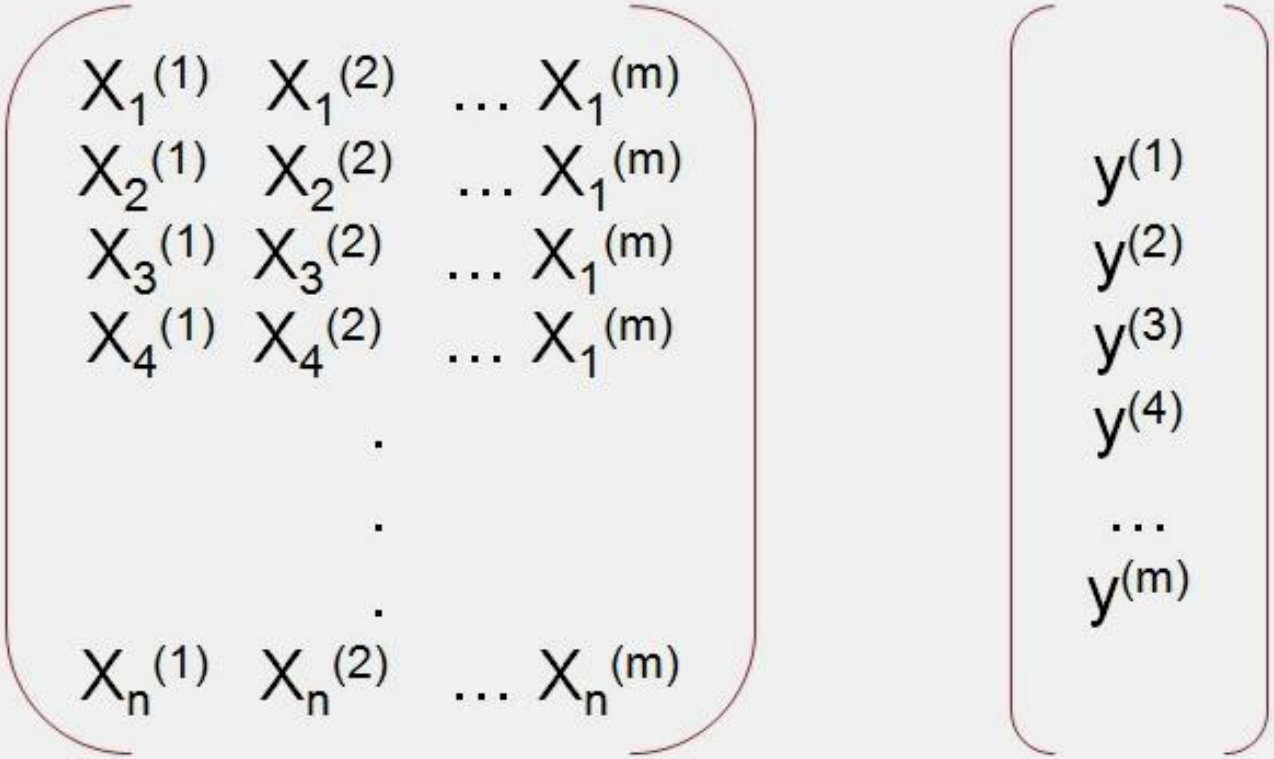
Медицинский прибор диагностирует болезнь у больного с вероятностью 97%, и ошибочно диагностирует болезнь у здоровых людей с вероятностью 5%. Болезнь встречается с частотой 1:600 человек. Какова вероятность что человек действительно болен при условии, что прибор диагностировал болезнь?

$$P(H_1) = \frac{1}{600};\ P(H_2) = \frac{599}{600}$$

$$P(A|H_1) = 0.97;\ P(A|H_2) = 0.05$$

$$P(H_1|A) = \frac{P(H_1)\cdot P(A|H_1)}{P(H_1)\cdot P(A|H_1)+P(H_2)\cdot P(A|H_2)} \approx 0.031$$

Обучение с учителем



X - матрица признаков, y - вектор разметки

Каждая строка представляет объект

Наивный Байесовский классификатор

Наивное предположение: признаки взаимно независимы

$$sample = (feature_1,\ feature_2 \ldots feature_n)$$

$$class = ArgMax(P(class = 0|sample),\ P(class = 1|sample))$$

$$P(class = x|sample) = \frac{P(sample|class{=}x) \cdot P(class{=}x)}{P(sample)},\ x \in \{0, 1\}$$

$$P(sample|class = x) = P(feature_1|class = x) \cdot P(feature_2|class = x) \ldots$$

$$P(sample) = P(feature_1) \cdot P(feature_2) \ldots$$

Задача 6

Озноб	Насморк	Головная боль	Лихорадка	Грипп
1	0	1	1	**0**
1	1	0	0	**1**
1	0	2	1	**1**
0	1	1	1	**1**
0	0	0	0	**0**
0	1	2	1	**1**
0	1	2	0	**0**
1	1	1	1	**1**

Sample = (1,0,1,0) - Грипп или не грипп?

Задача 6

Sample = (1,0,1,0) - Грипп или не грипп?

P(О=1|Грипп=1)=3/5
P(Н=0|Грипп=1)=1/5
P(Г=1|Грипп=1)=2/5
P(Л=0|Грипп=1)=1/5

P(Грипп=1)=5/8

P(О=1|Грипп=0)=1/3
P(Н=0|Грипп=0)=2/3
P(Г=1|Грипп=0)=1/3
P(Л=0|Грипп=0)=2/3

P(Грипп=0)=3/8

P(Грипп=1|Sample)*P(Грипп=1)=3/500
P(Грипп=0|Sample)*P(Грипп=0)=1/54

не грипп!

Наивный Байесовский классификатор

Достоинства:

- простота реализации
- высокая скорость работы
- масштабируемость

Недостатки:

- невысокое качество предсказаний
- не учитывает взаимодействия признаков

Используется для:

- антиспам
- анализ эмоциональной окраски отзывов

Случайная величина

Доход = ставка + чаевые

Ставка – величина обычная

Чаевые – **случайная величина**

Чаевые за конкретный месяц – **реализация случайной величины**

Случайная величина

Пусть задано вероятностное пространство

$(\Omega, \mathcal{F}, P)$.

Случайной величиной (СВ) называется функция $X : \Omega \to R$

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 ≤ x+Δx)/Δx, Δx→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

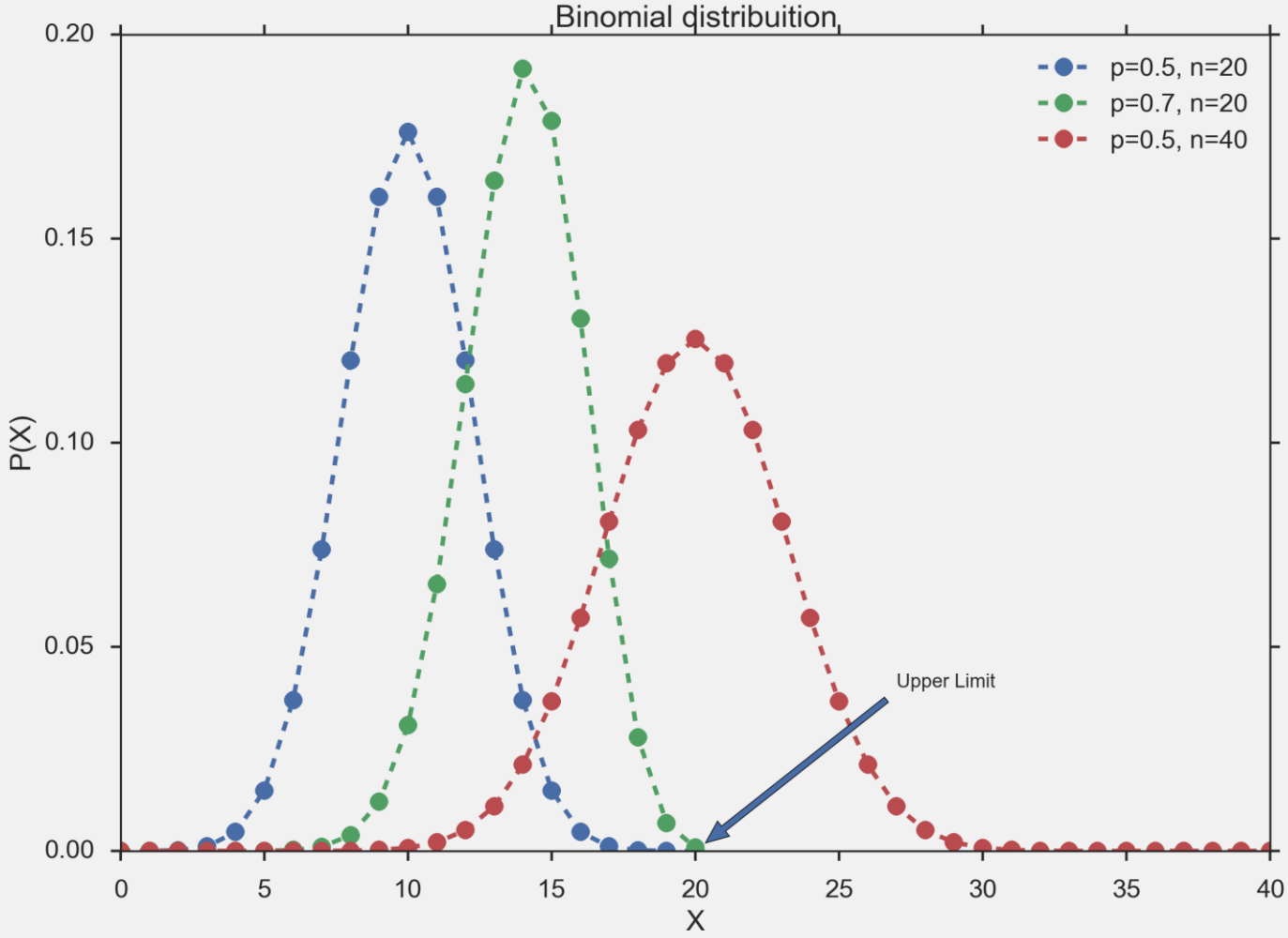
Моделирует количество «успехов» в последовательности из n независимых случайных экспериментов, таких, что вероятность «успеха» в каждом из них постоянна и равна p.

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

$$M[X] = np$$

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

Биномиальное распределение



Распределение Пуассона

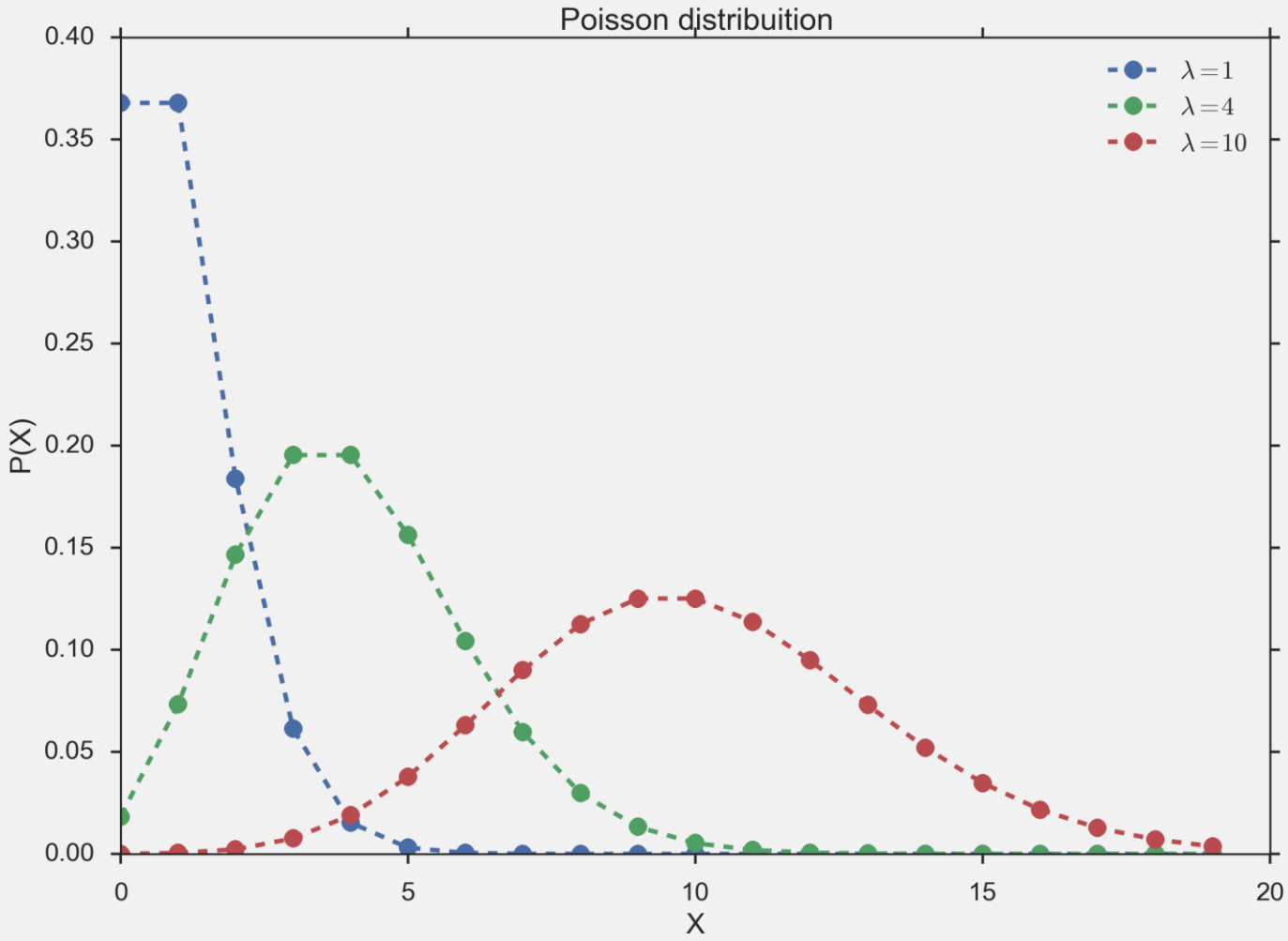
Моделирует число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга.

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

$$M[X] = \lambda$$

$$D[X] = \lambda$$

Распределение Пуассона



Экспоненциальное распределение

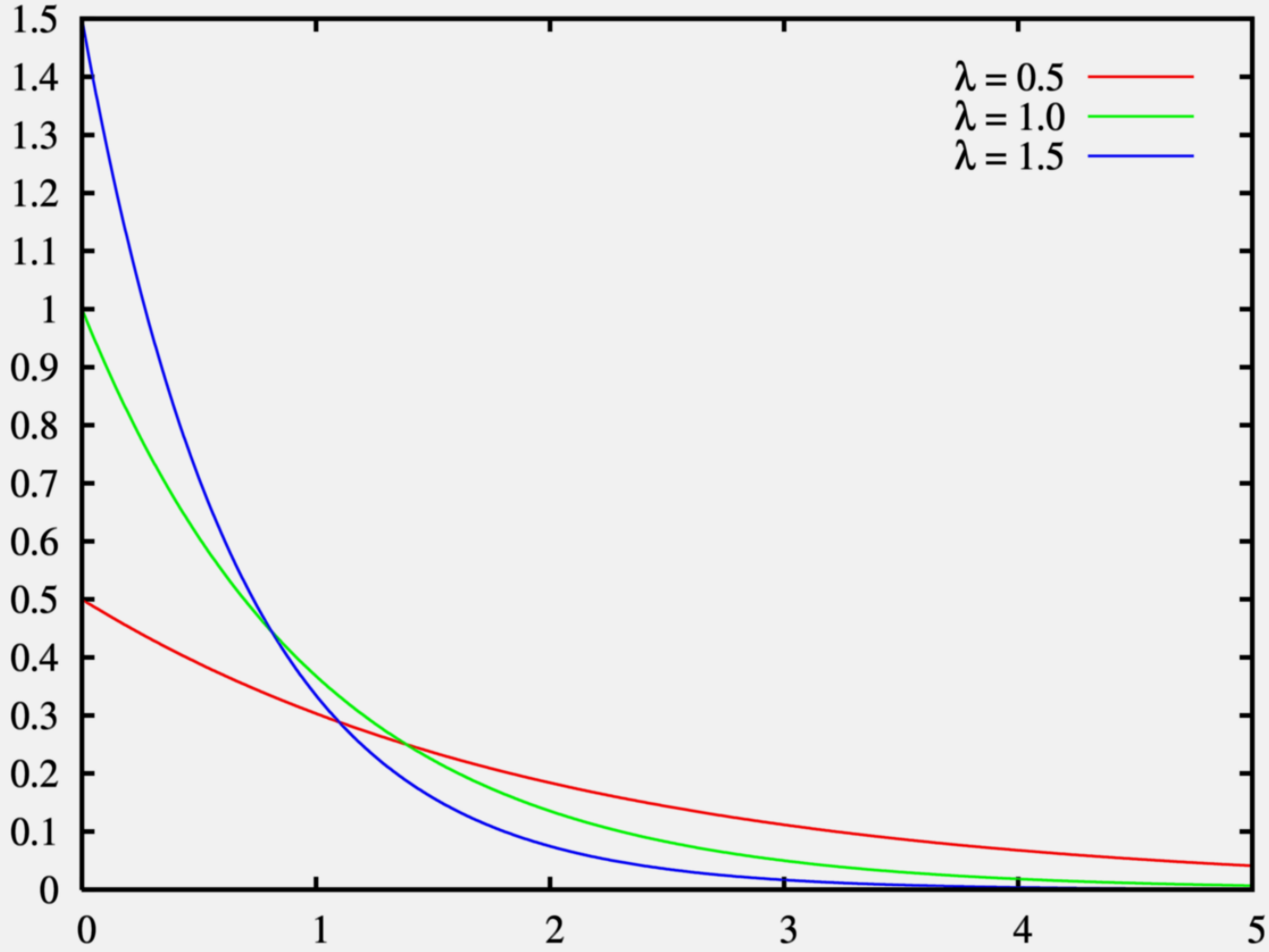
Моделирует время между двумя последовательными свершениями события в Пуассоновском процессе.

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

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

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

Экспоненциальное распределение



Нормальное распределение

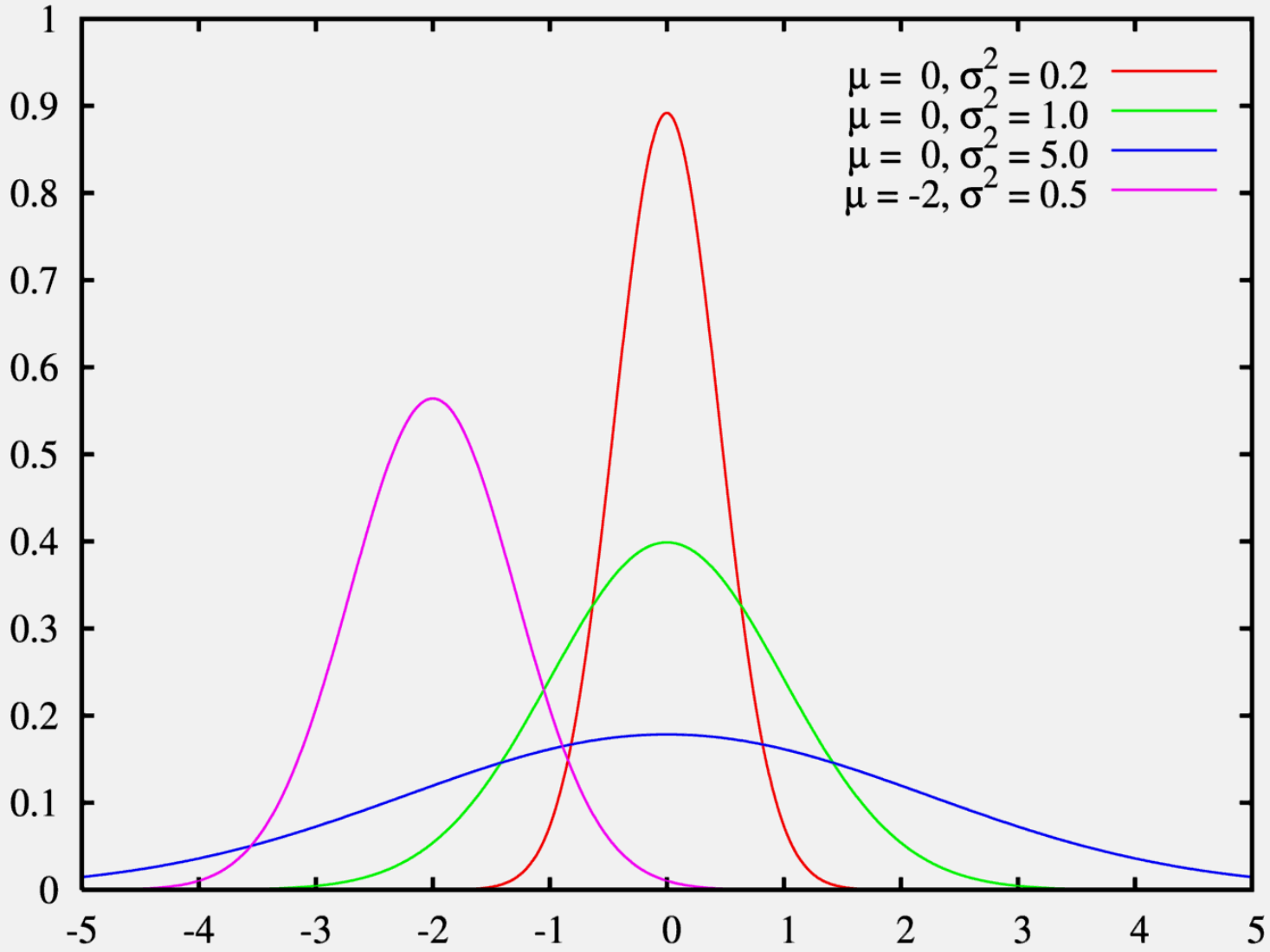
Моделирует результат суммы многих случайных слабо зависимых СВ, каждая из которых вносит малый вклад относительно общей суммы.

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

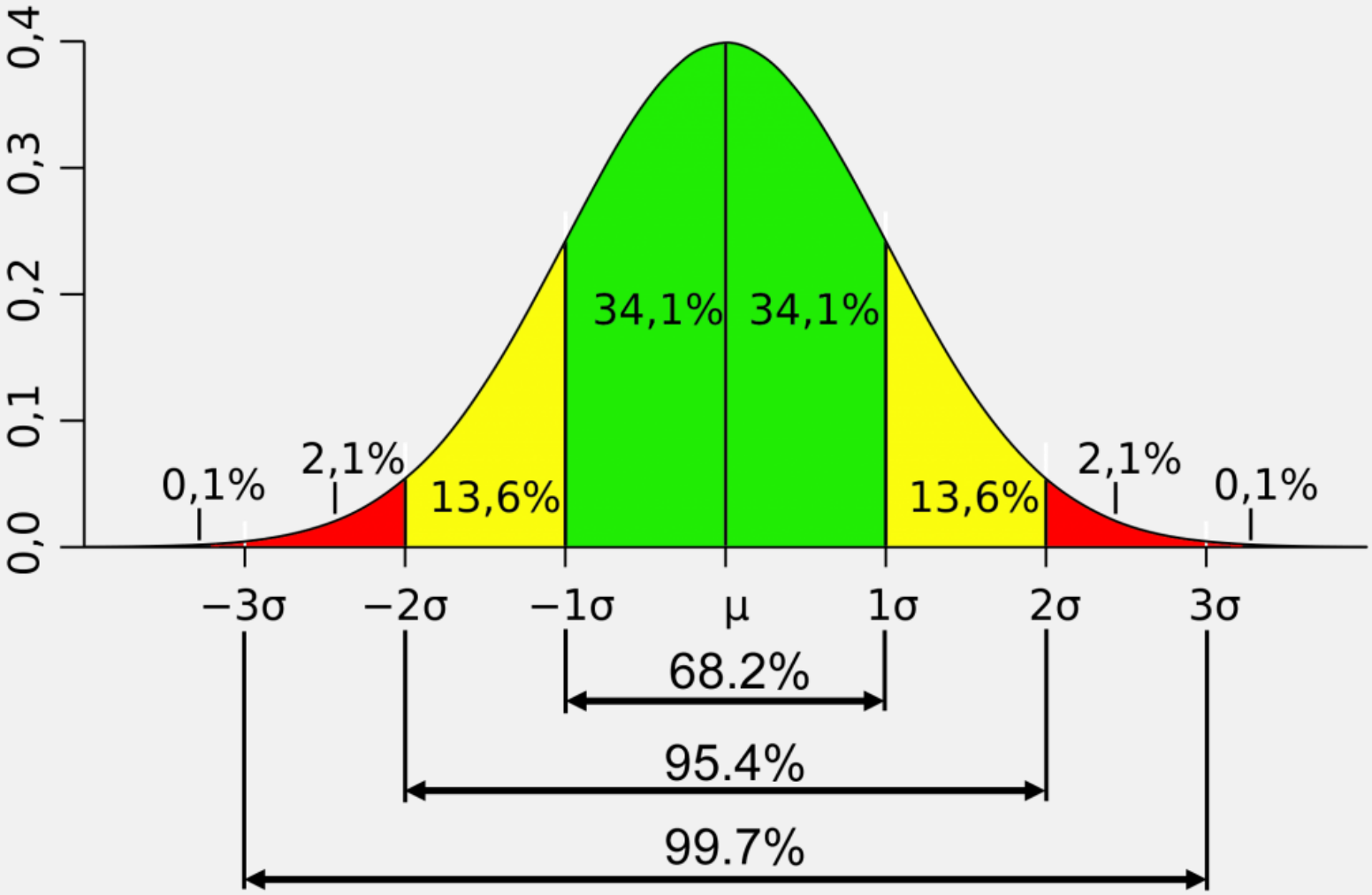
$$M[X] = \mu$$

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

Нормальное распределение



Нормальное распределение

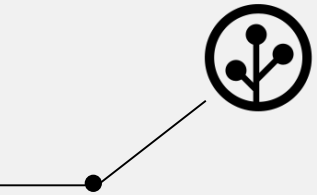


Понятие выборки

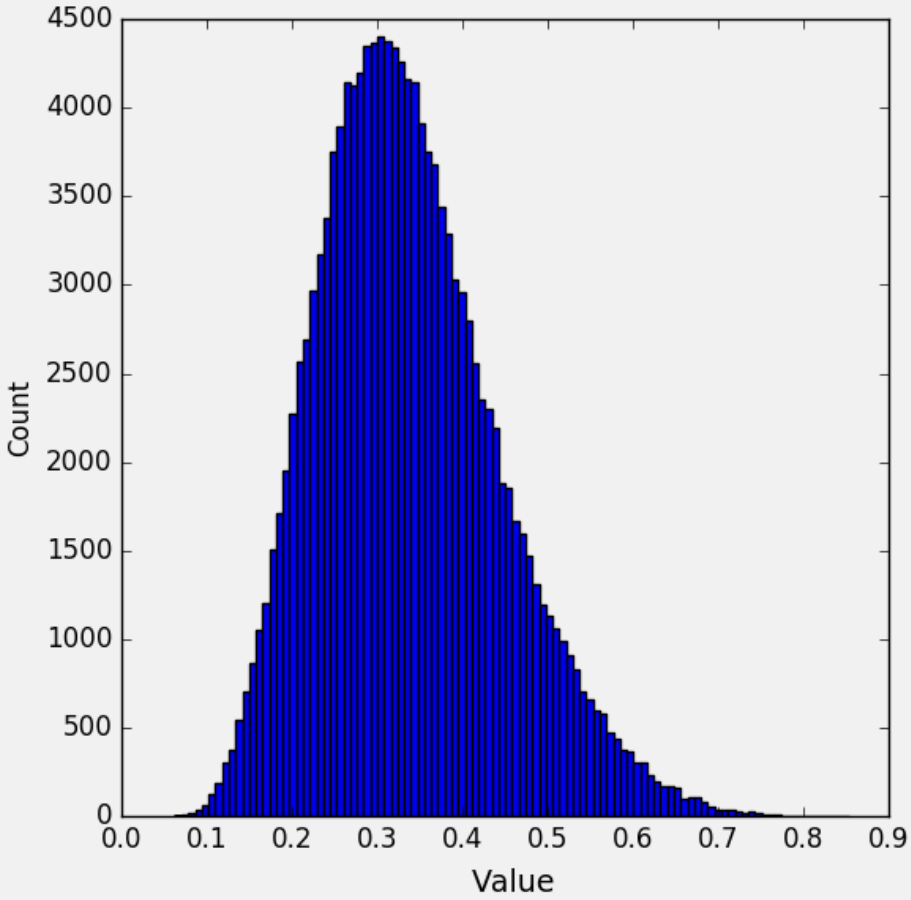
Пусть есть некоторая случайная величина X, вектор из n реализаций X будем называть **выборкой**.

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:

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

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

Квантиль и перцентиль

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

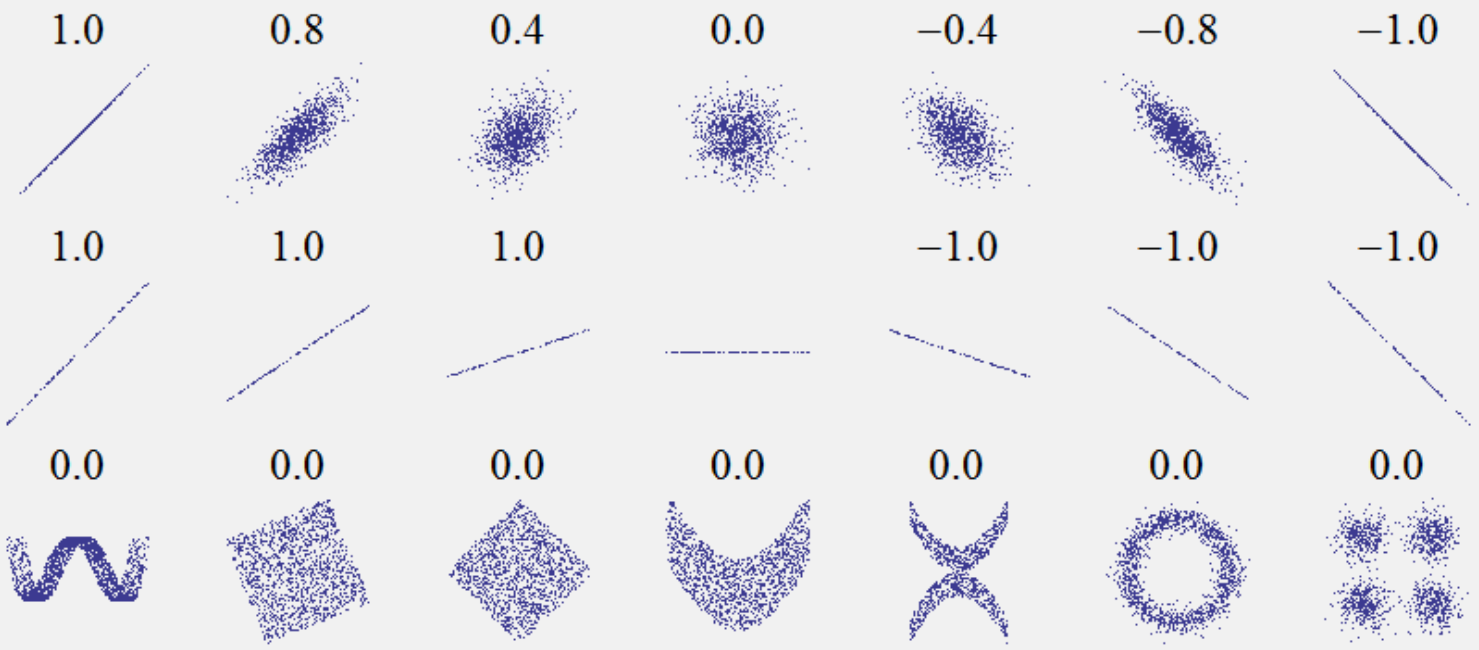
Перцентиль-25 – нижний квартиль

Перцентиль-50 – медиана

Перцентиль-75 – верхний квартиль

Корреляция

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



Статистическое тестирование гипотез

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

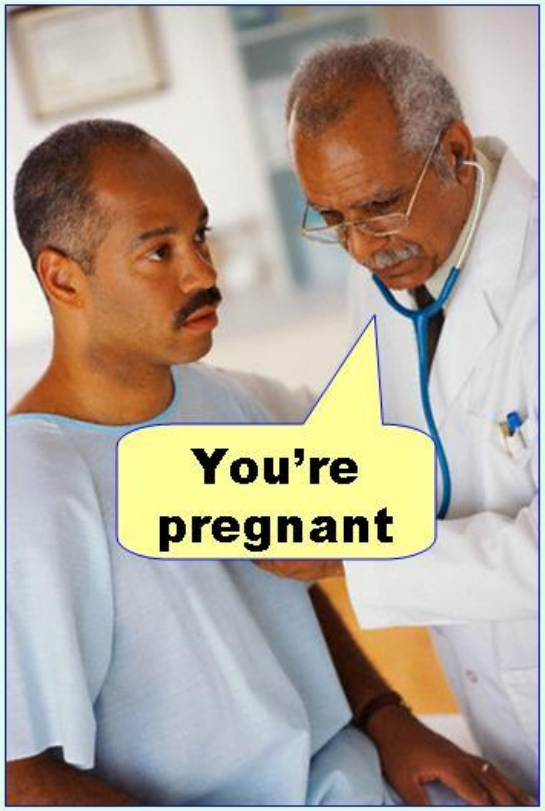
Статистическая ошибка I рода – обнаружение различий или связей, которые на самом деле не существуют.

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

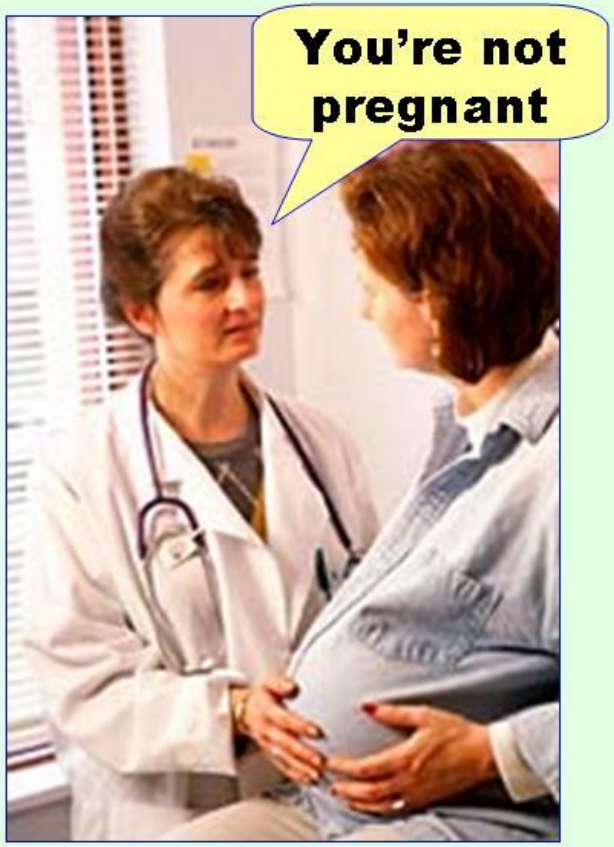
Более критичной ошибкой считается статистическая ошибка первого рода.

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

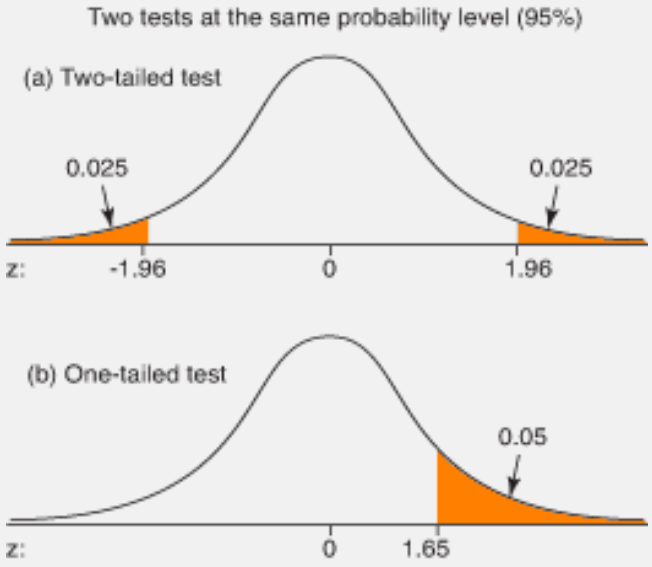
Одновыборочный z-критерий Фишера

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

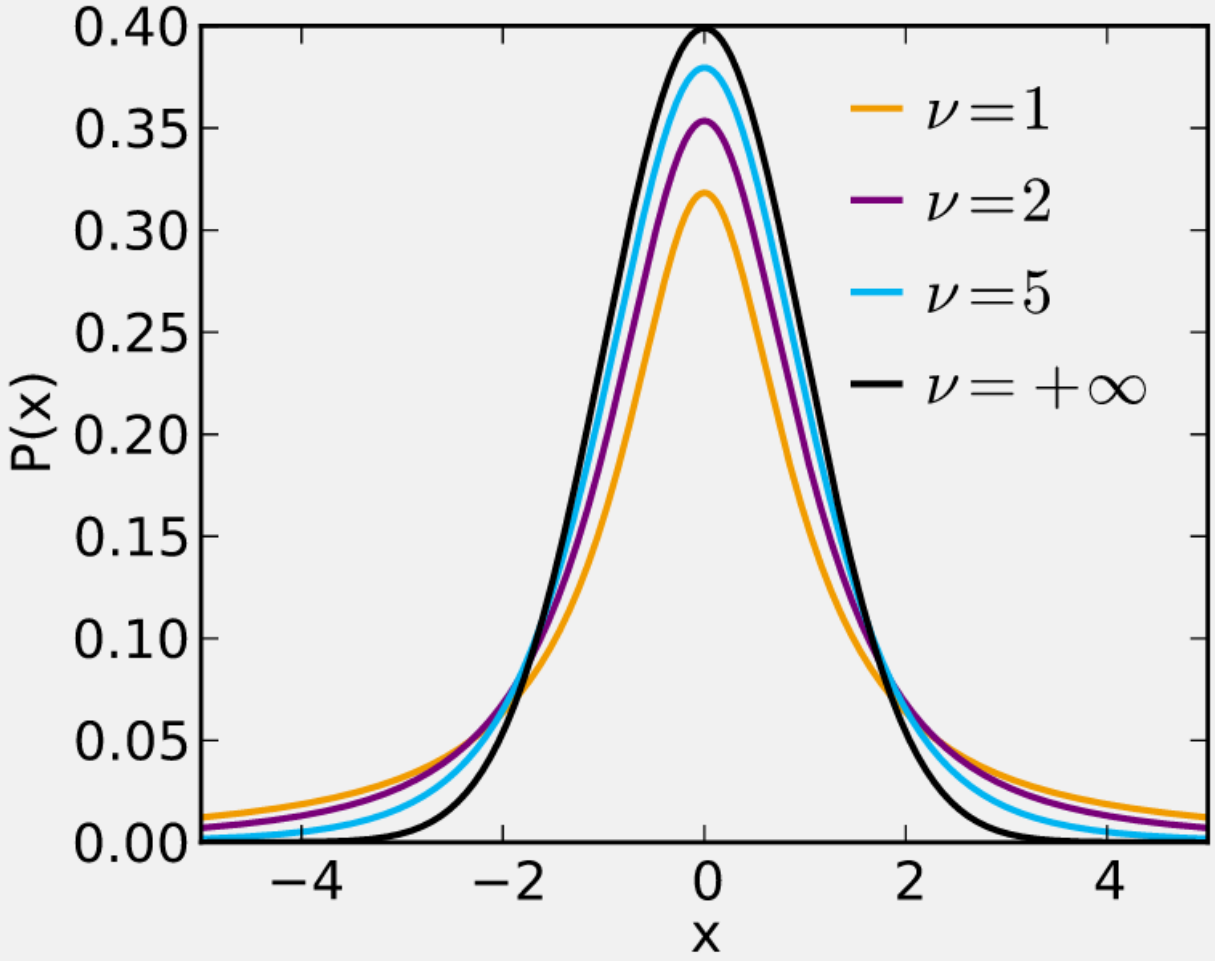
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

$$z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$



Распределение Стъюдента



Госсет Уильям (Стъюдент), 1908

Одновыборочный t-критерий

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

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

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

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

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.

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

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

ТЕХНОСФЕРА

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
- crowdai.org
- drivendata.org
- dataring.ru
- boosters.pro
- mlbootcamp.ru

В среднем соревнование длится несколько месяцев

Соревнование "IntroML2018. 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 125k, Test 125k

- Public/Private split 50/50%

- Метрика accuracy

Данные

train.csv

Id	Sample	Prediction
0	следующий де	следующий день
1	тем бо	тем более
2	так ка	так как
3	меньшей мер	меньшей мере
4	две нед	две недели
5	что эт	что этот

test.csv

Id	Sample
125000	как будт
125001	должна бы
125002	был ег
125003	под ру
125004	вот он
125005	две неде

submit.csv

Id	Prediction
125000	потому что
125001	потому что
125002	потому что
125003	потому что
125004	потому что
125005	потому что

Пример решения задачи

См. код

Домашнее задание №#2

https://www.kaggle.com/c/intromlⅡ2018-1

- Зарегистрироваться на kaggle.com
- Сделать сабмит решения
- Прислать ссылку на код решения
- Прислать ссылку на свой профиль kaggle
- В письме также надо указать фамилию и имя

Срок сдачи

14 октября 2018

ТЕХНОСФЕРА

Спасибо за внимание!

Евгений Некрасов

e.nekrasov@corp.mail.ru