НЕТОЛОГИЯ
групп

ОШИБКИ И ПРОВАЛЫ ПРИ ИНТЕРПРЕТАЦИИ АНАЛИТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Максим Чикуров

Data Scientist и руководитель команды аналитики

Работал в компаниях Citibank, BNP Paribas, Barclays Bank, Teradata

maxim.chikurov@gmail.com

О ЧЕМ ПОГОВОРИМ И ЧТО СДЕЛАЕМ

План занятия

1. Виды распределений
2. Проверка статистических гипотез
3. Центральная предельная теорема
4. Причины ошибок в интерпретации данных
5. Дополнительные темы (CSV, Gretl)

1

РАСПРЕДЕЛЕНИЯ

Вероятности и распределение

Распределение вероятностей это функция (закон) которая описывает вероятность получения определённого значения переменной. Другими словами значения переменной различаются на основе **распределение вероятностей**.

Дискретные распределения

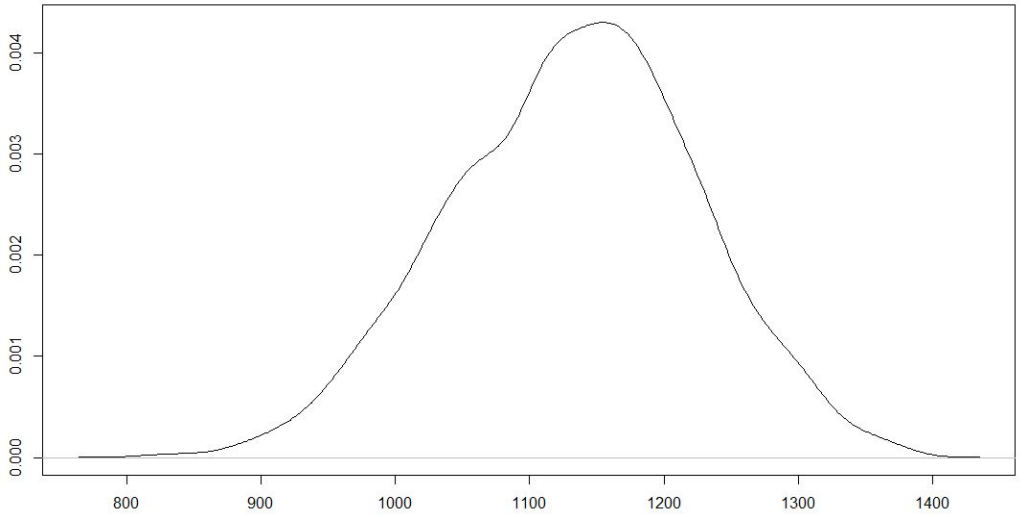
(Кумулятивная) Функция распределения - функция, характеризующая распределение случайной величины или случайного вектора; вероятность того, что случайная величина примет значение, меньшее или равное x.

Функция вероятности - функция, возвращающая вероятность того, что дискретная случайная величина примет определённое значение.

Функция вероятности - это наиболее часто используемый способ охарактеризовать **дискретное распределение.**

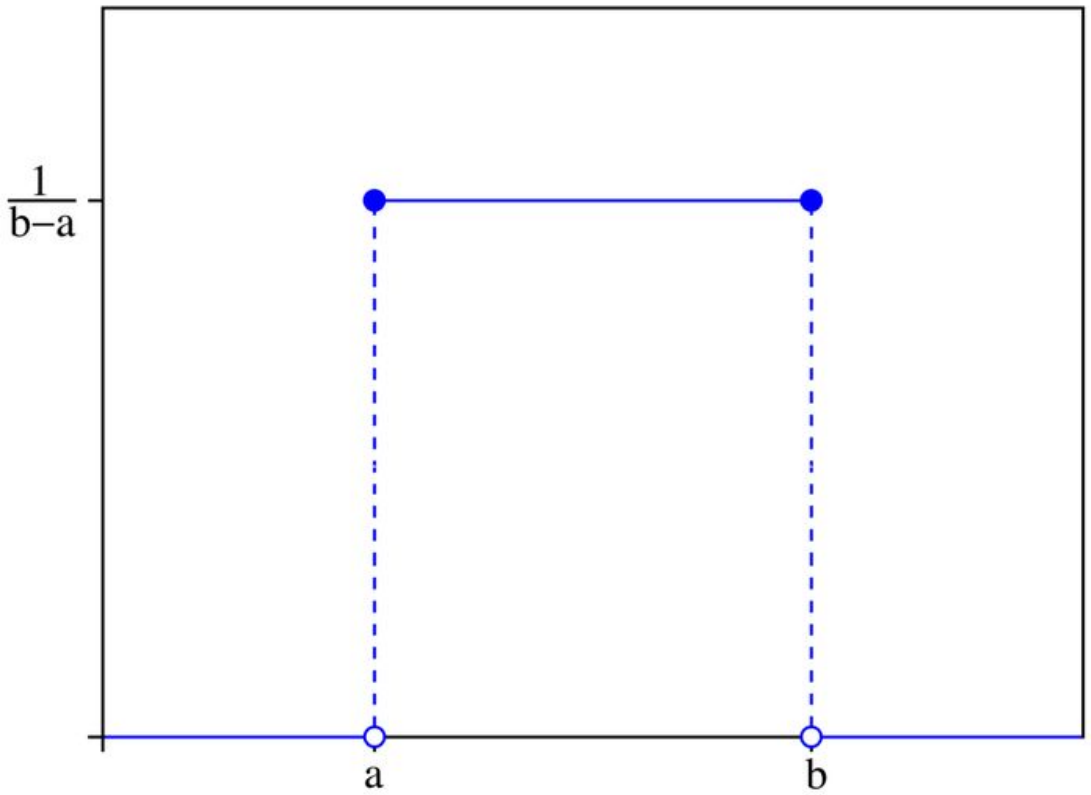
Непрерывные распределения

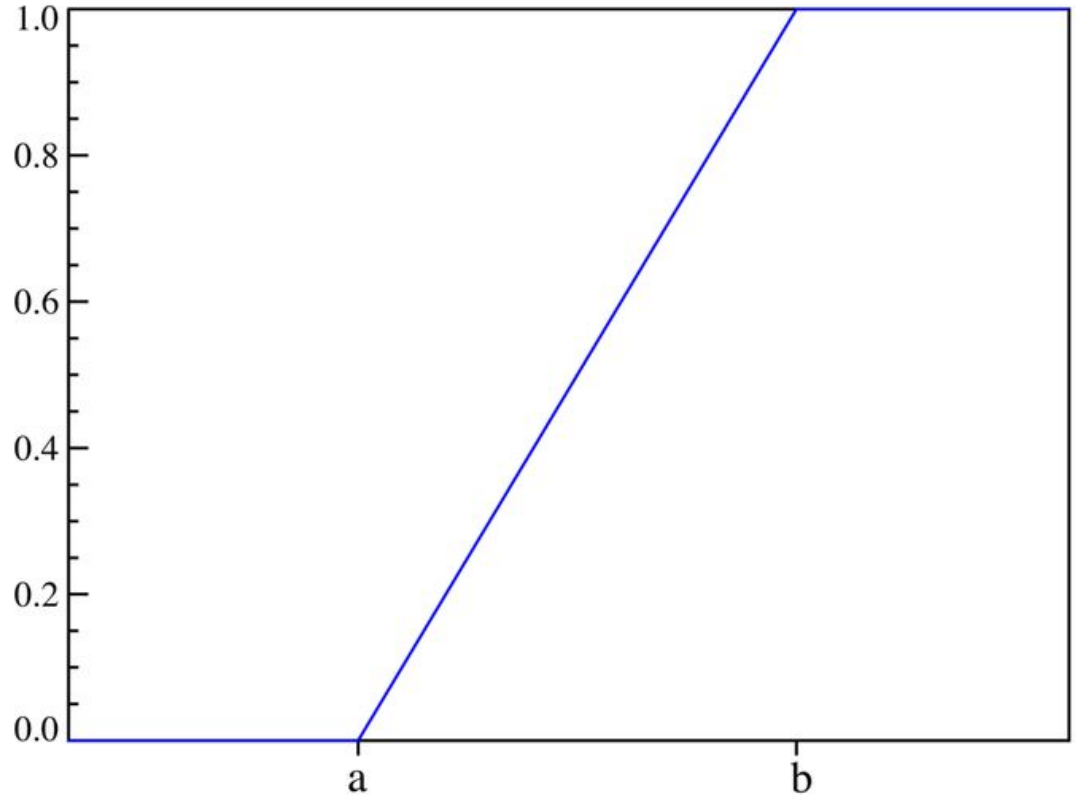
Плотность вероятности используется для вычисления вероятностей в случае непрерывной случайной величины. Значения функции плотности не вычисляются простой подстановкой значений в качестве аргумента, а должны быть проинтегрированы над интервалом значений.



Равномерное распределение

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



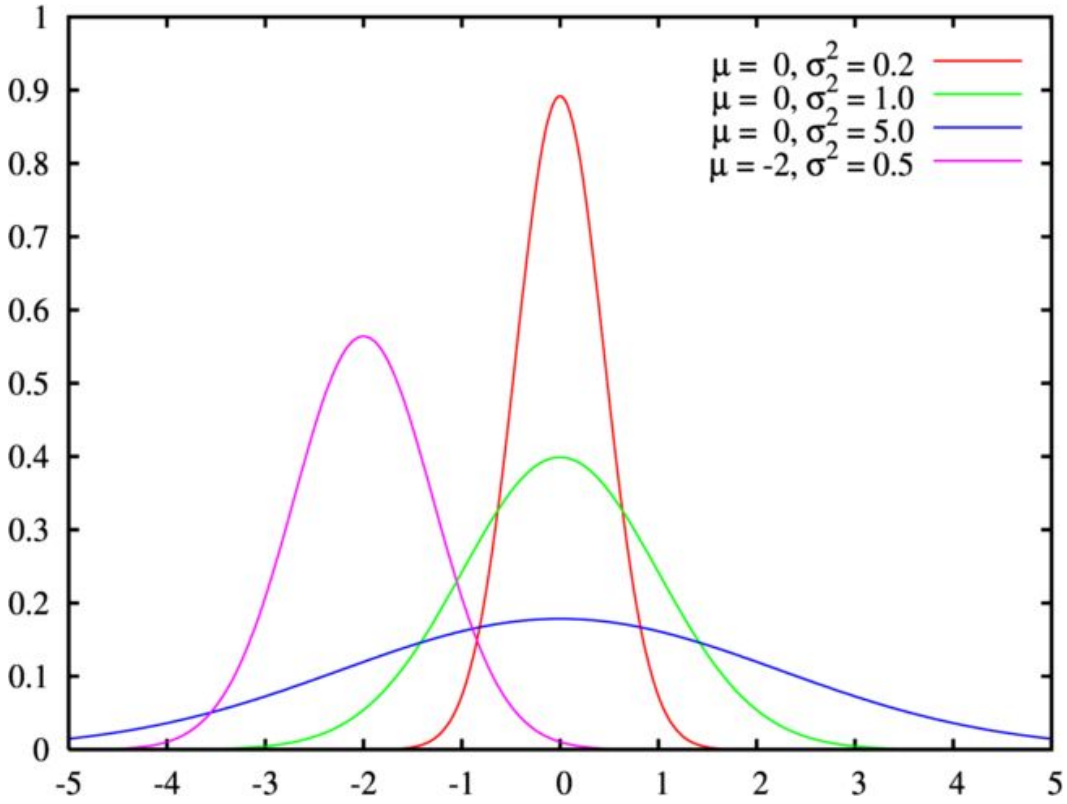


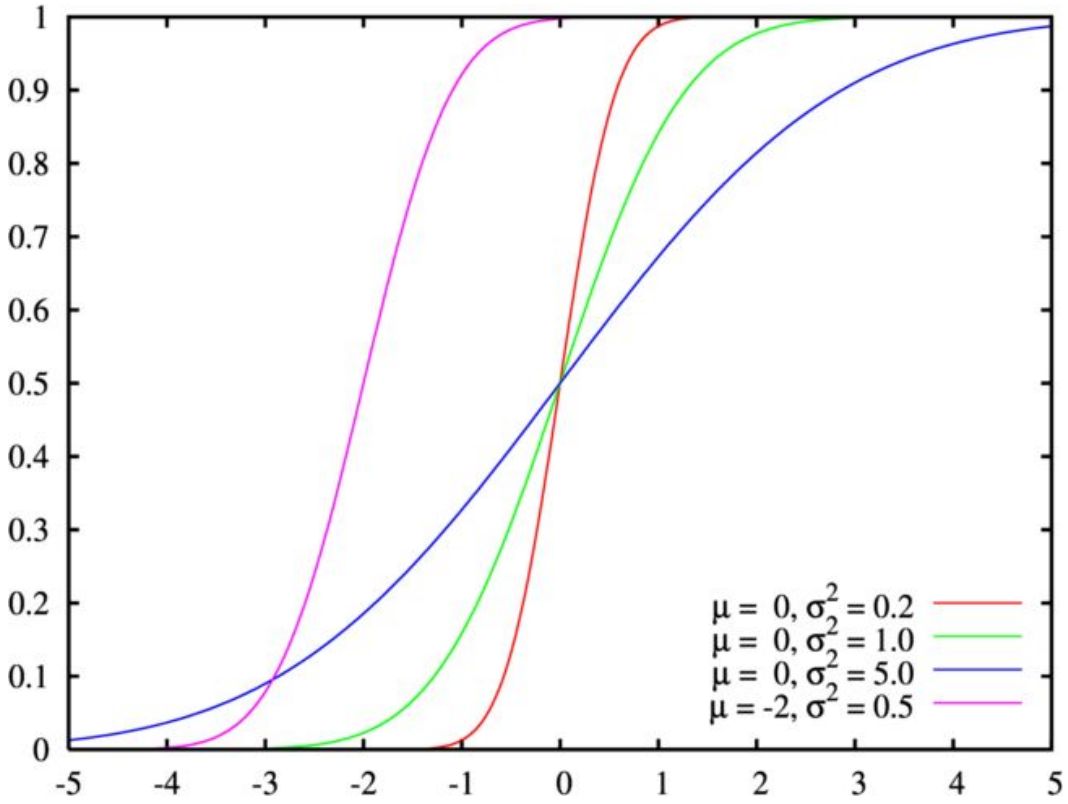
Пример

Вероятность выпадения каждой стороны игрального кубика

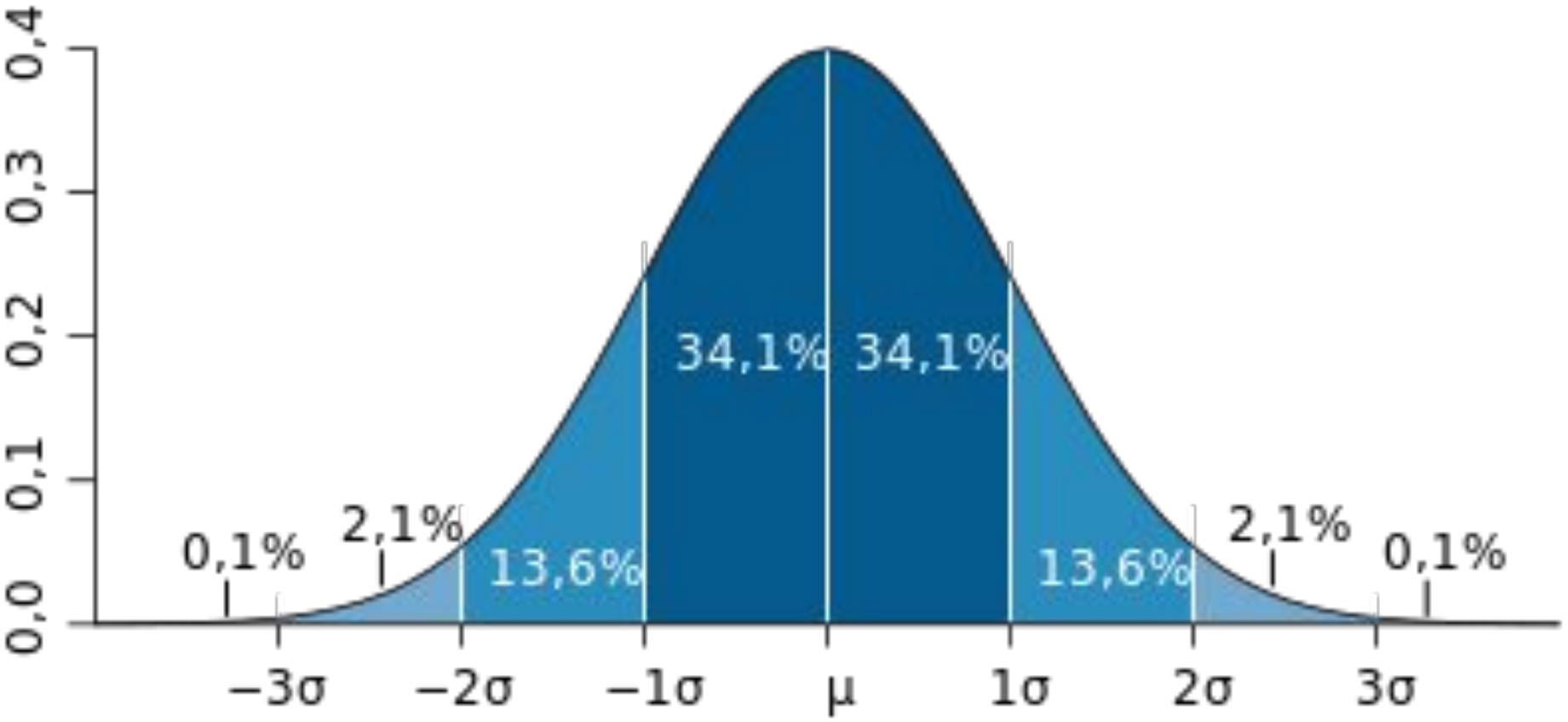
Нормальное распределение

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

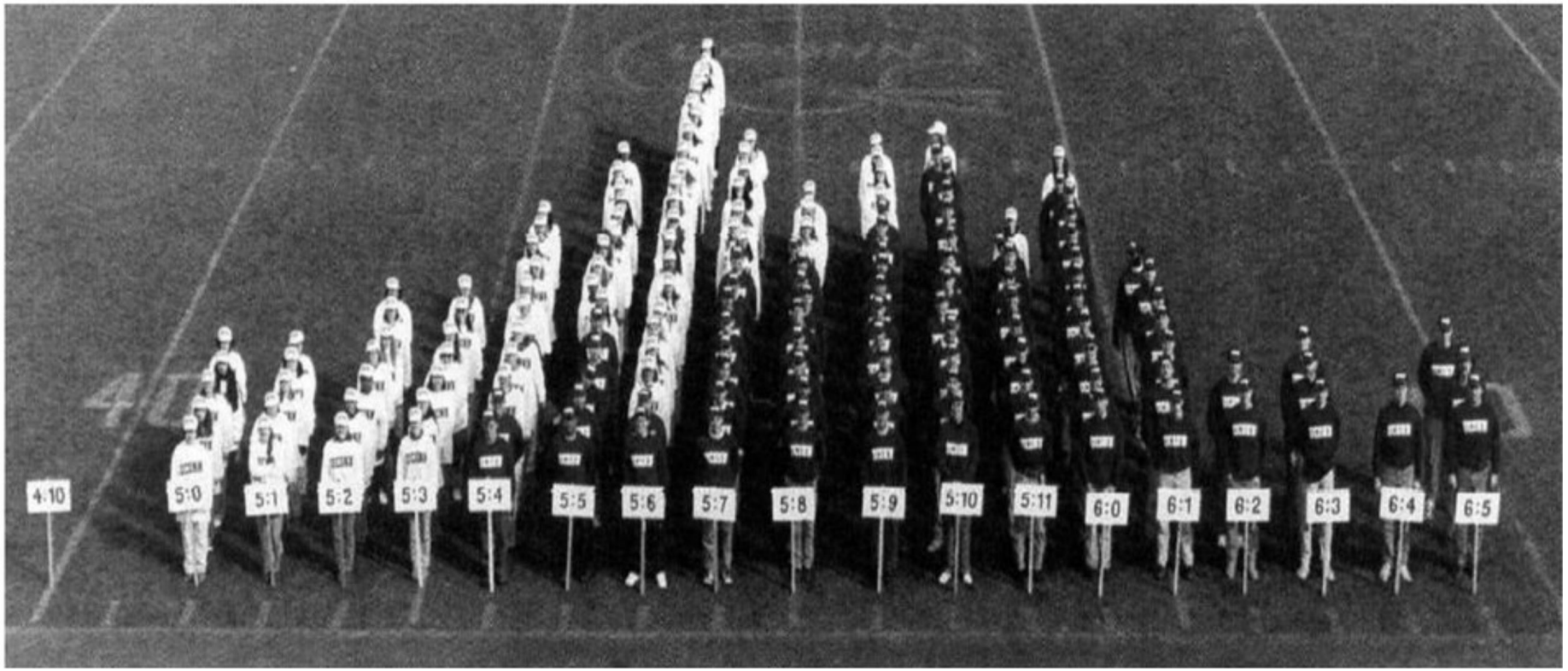




Нормальное распределение



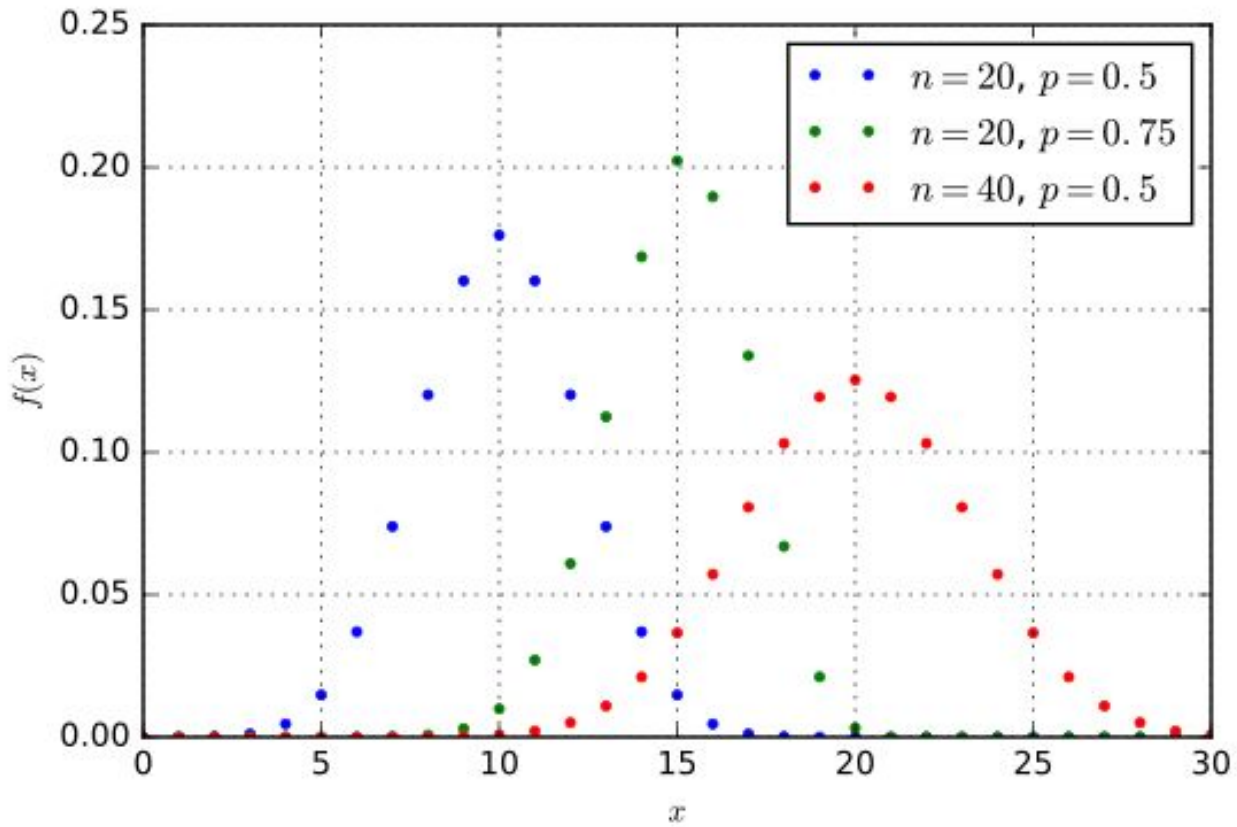
Пример

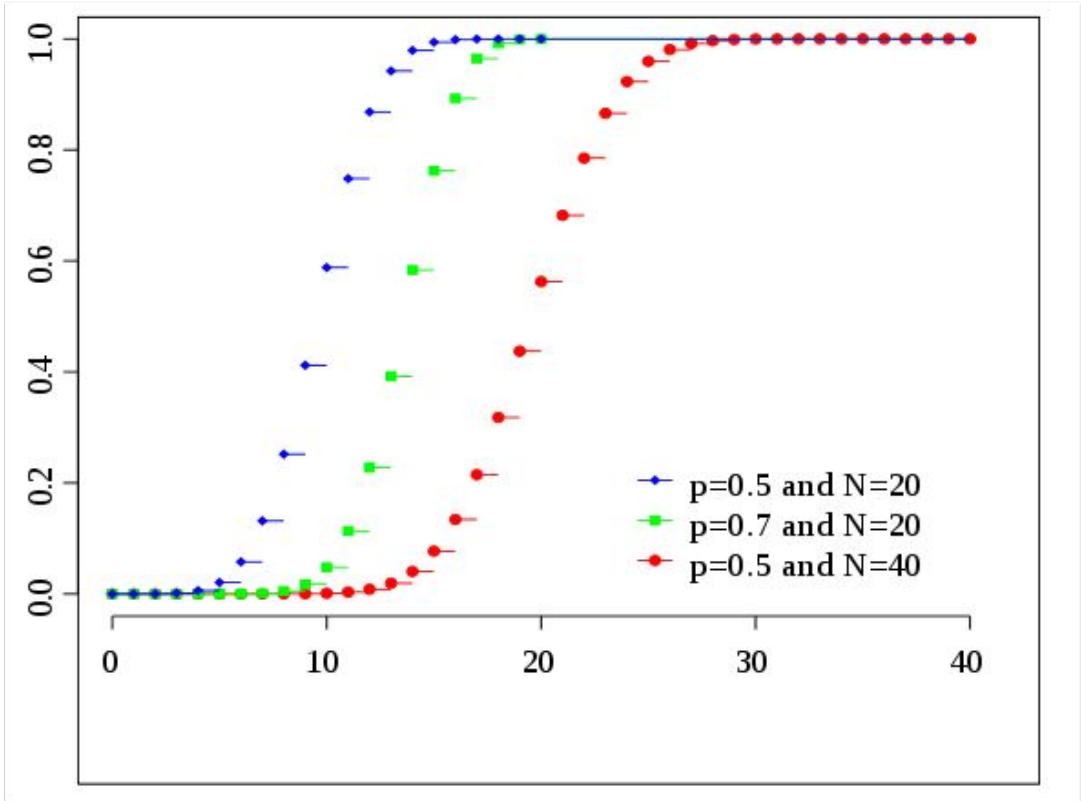


Распределение людей по росту

Биномиальное распределение

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





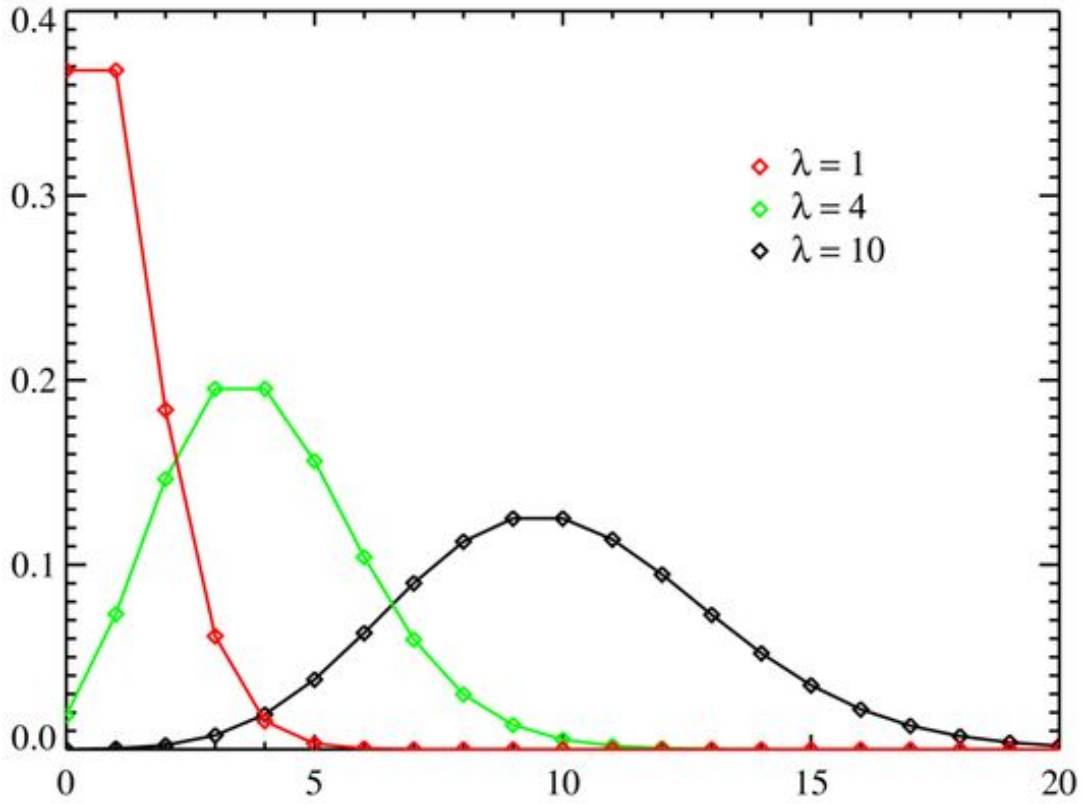
РАСПРЕДЕЛЕНИЯ

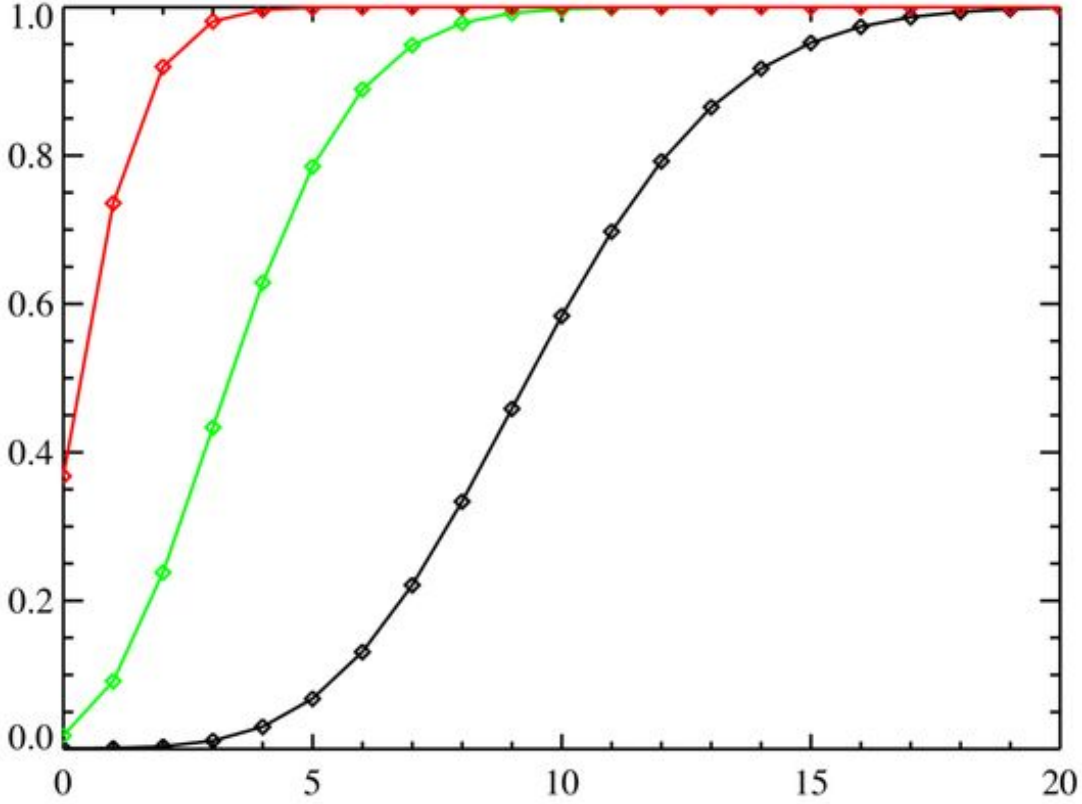
Пример

Количество попаданий по воротам

Распределение Пуассона

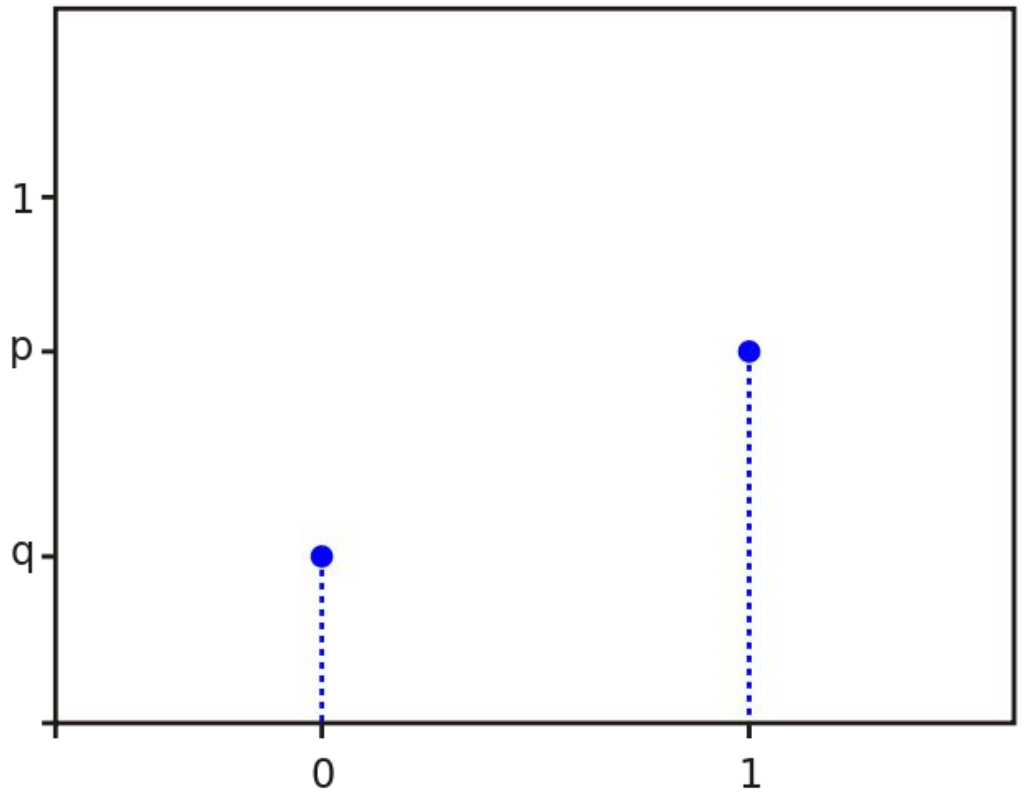
$$p(k) \equiv \mathbb{P}(Y = k) = \frac{\lambda^k}{k!}\, e^{-\lambda}$$

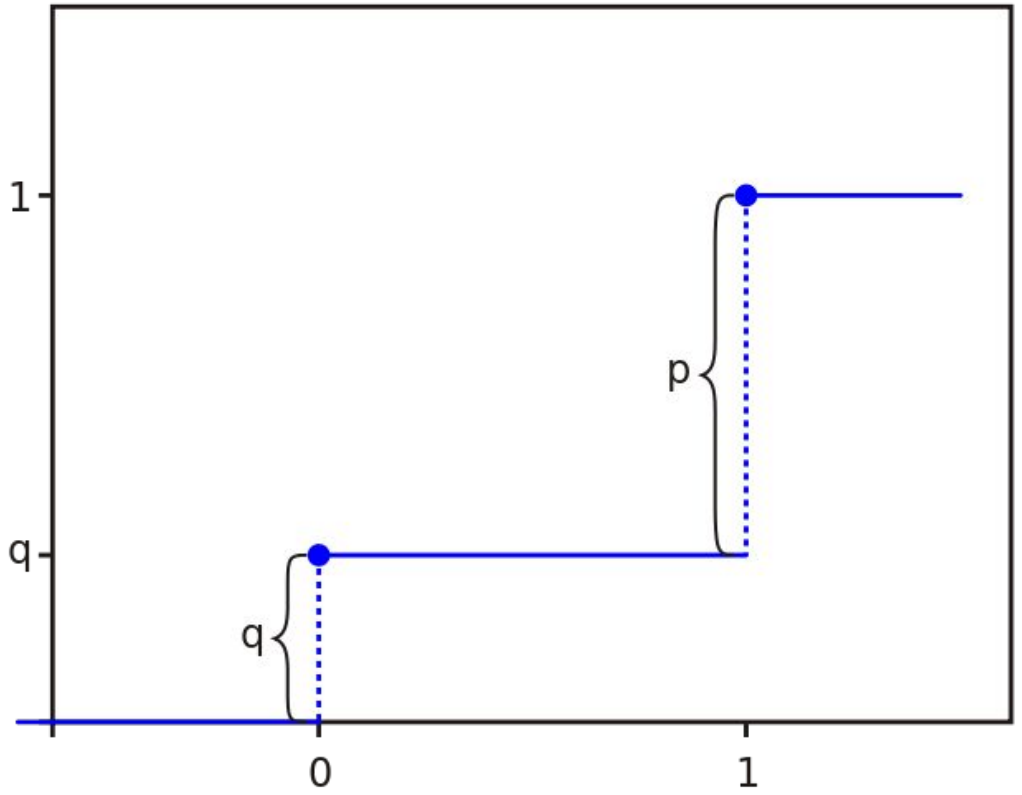




Распределение Бернулли

$$\mathbb{P}(X = 1) = p \qquad q \equiv 1 - p$$
$$\mathbb{P}(X = 0) = q$$





2

ИНСТРУМЕНТЫ ПРОВЕРКИ ГИПОТЕЗ

Ошибки первого и второго рода

Ошибки первого и второго рода - это ключевые понятия задач проверки статистических гипотез.

Данные понятия также часто используются и в других областях, когда речь идет о принятии **«бинарного» решения (да/нет)** на основе некоего критерия (теста, проверки, измерения), который с некоторой вероятностью может давать ложный результат.

Ошибки первого и второго рода

		Верная гипотеза	
		H_0	H_1
Результаты применения критерия	H_0	H_0 верно принята	H_0 Неверно принята ошибка *второго* рода
	H_1	H_0 Неверно отвергнута ошибка *первого* рода	H_0 Верно отвергнута

Вероятность ошибки первого рода при проверке статистических гипотез называют уровнем значимости

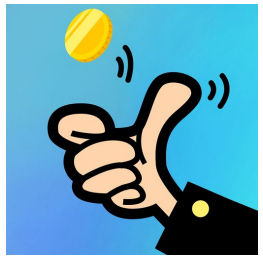
Вероятность ошибки второго рода связана с понятием мощность критерия

Пример

Фильтрация email спама

ИНСТРУМЕНТЫ ПРОВЕРКИ ГИПОТЕЗ

Coin toss / “орёл или решка”



Статистическая мощность

Коротко

Вероятность отклонения основной или нулевой гипотезы при проверке статистических гипотез в случае, когда конкурирующая или альтернативная гипотеза верна.

Чем выше мощность статистического теста, тем меньше вероятность совершить ошибку второго рода.

Величина мощности также используется для вычисления размера выборки, необходимой для подтверждения гипотезы с необходимой силой эффекта.

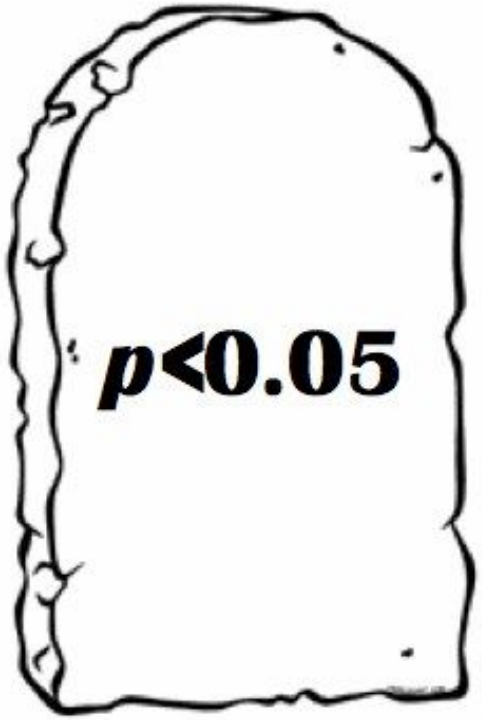
Статистическая значимость

Коротко

В статистике **значение переменной называют статистически значимой,** если мала вероятность случайного возникновения этой или ***ещё более*** крайних величин.

Уровень значимости = вероятность ошибки первого рода.

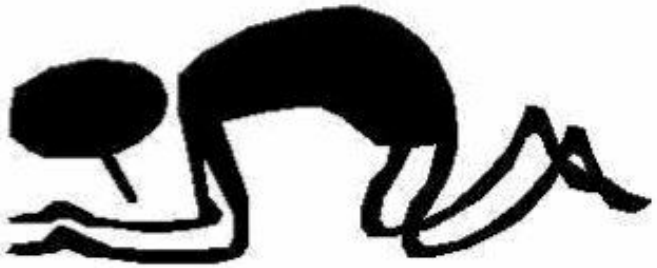
p-value



Величина, используемая при тестировании статистических гипотез.

Фактически это вероятность ошибки при отклонении нулевой гипотезы (ошибки первого рода).

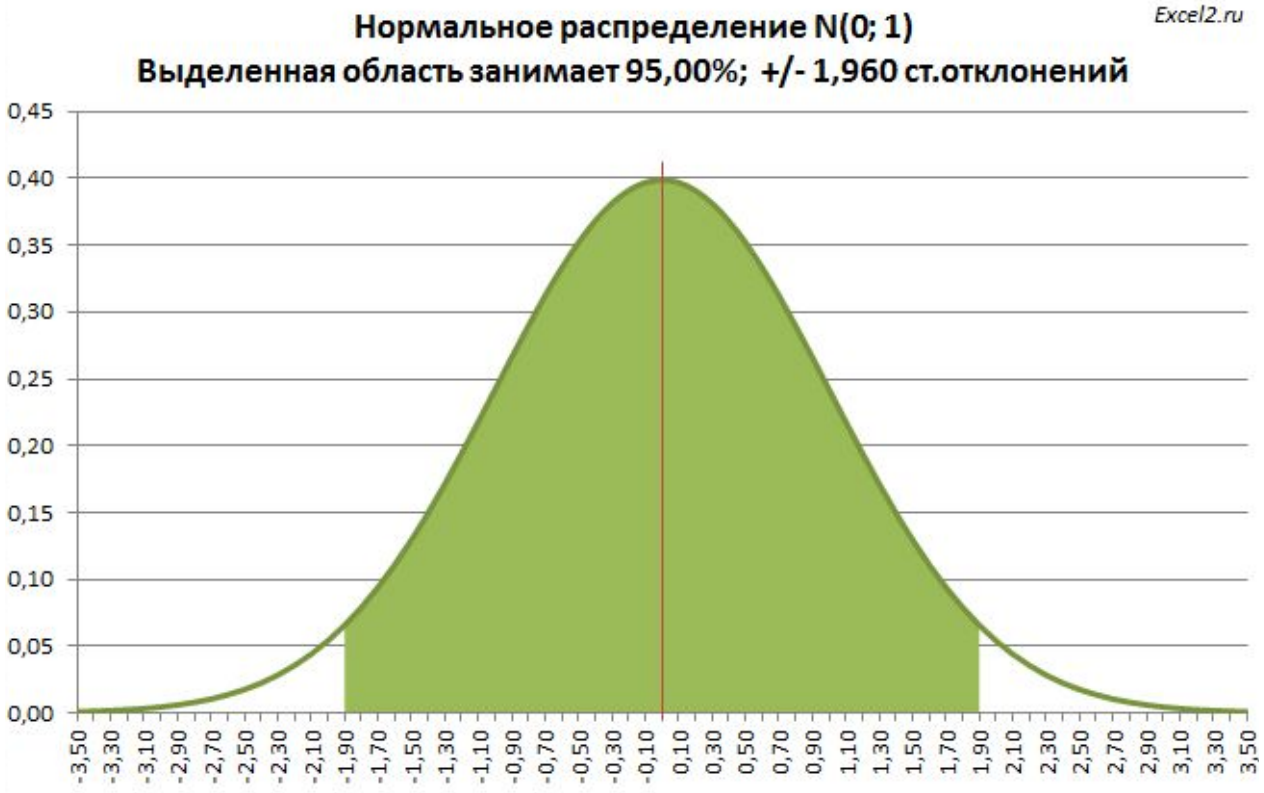
Обычно P-значение равно вероятности того, что случайная величина с данным распределением (распределением тестовой статистики при нулевой гипотезе) примет значение, не меньшее, чем фактическое значение тестовой статистики.



Доверительный интервал

Интервал, содержащий значение случайной величины с заданным уровнем доверия.

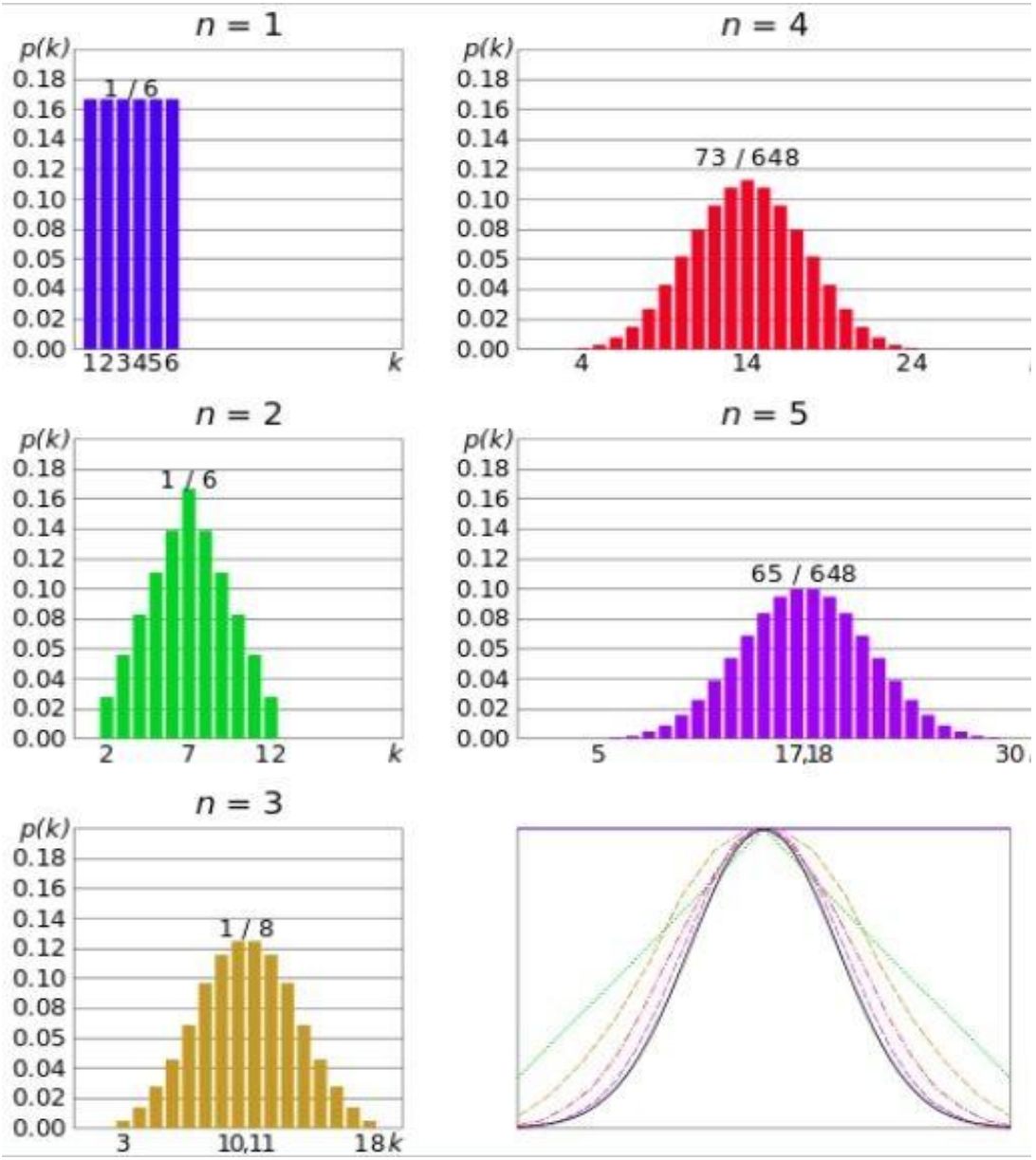
Уровень доверия = 1 - вероятность ошибки первого рода.



Центральная предельная теорема

Коротко

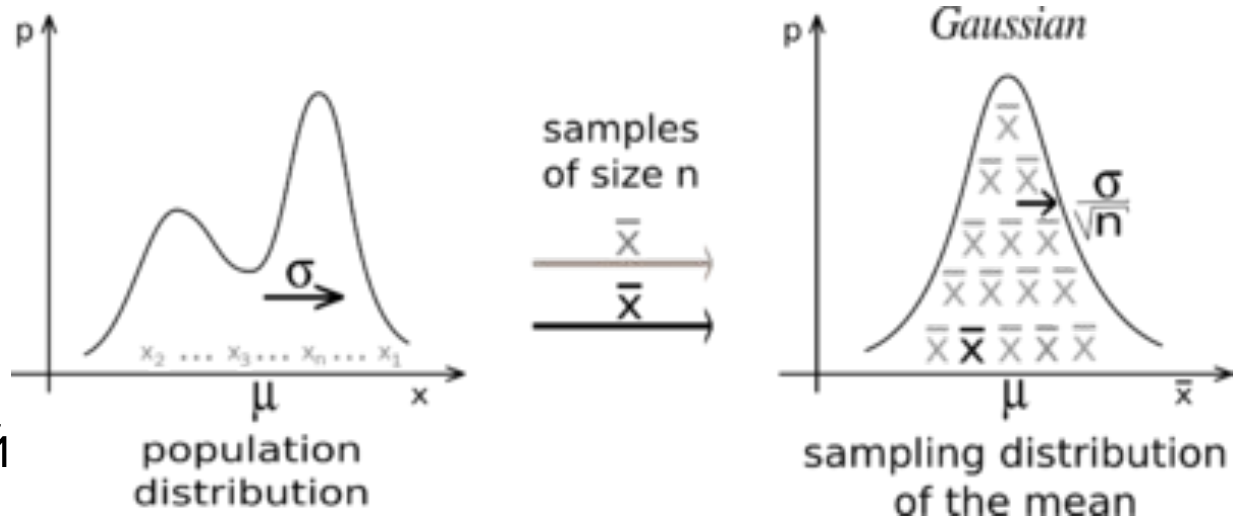
Сумма достаточно большого количества слабо зависимых случайных величин,
имеющих примерно одинаковые масштабы
ни одно из слагаемых
не доминирует, не вносит
в сумму определяющего вклада,
имеет распределение,
близкое к нормальному.



Центральная предельная теорема

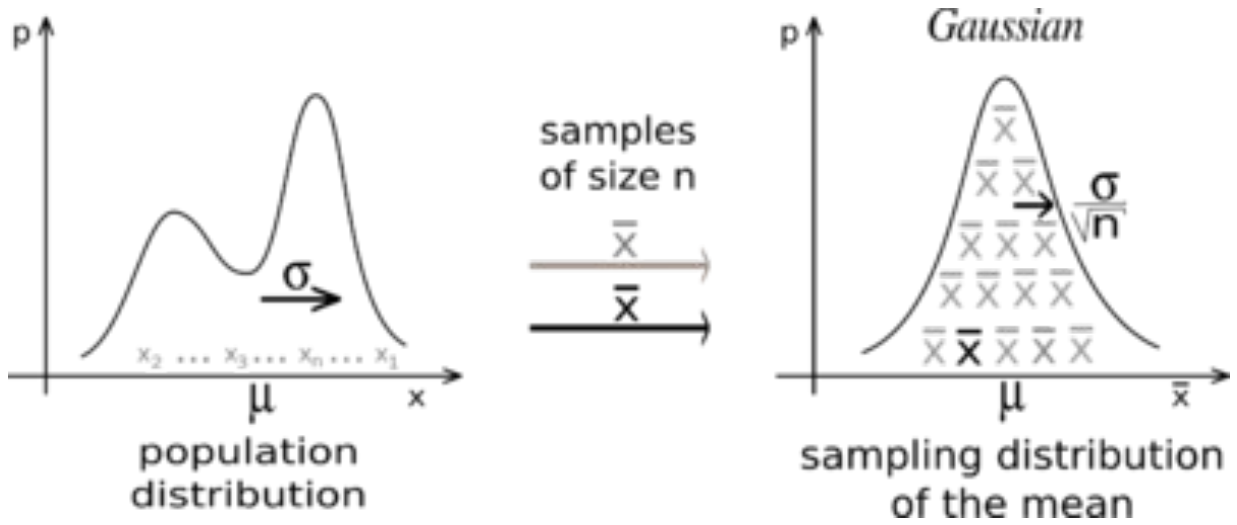
Важно

Какова бы ни была форма распределения генеральной совокупности, выборочное распределение стремится к нормальному, а его дисперсия задается центральной предельной теоремой.



Центральная предельная теорема

Неформально говоря, классическая центральная предельная теорема утверждает, что $\bar{X}_n$ имеет распределение близкое к $N(\mu, \sigma^2/n)$



ПРАКТИКА

ПРАКТИКА

Продолжение кейса грузоперевозок.

Поступил звонок от компании-клиента. Необходима срочная перевозка груза из Таллинна в Санкт-Петербург. Вес груза неизвестен, но известно, что перевезти предстоит 140 коробок. В Таллинне в настоящий момент доступен только один автомобиль грузоподъемностью 1,2 тонны.

Ссылка на датасет: goo.gl/y6JhKG

ПРАКТИКА

Коробки в заказах

	A	B	C
1	Номер коробки	Номер заказа	Вес коробки
2	1	1	5,88
3	2	1	5,69
4	3	1	3,69
5	4	1	5,56
6	5	1	3,45
7	6	1	25,71
8	7	1	4,92
9	8	1	5,56
10	9	1	5,75
11	10	1	3,6
12	11	1	5,32
13	12	1	2,96

Минимум	0,41
Максимум	25,71
Среднее	7,51
Медиана	5,10
Стандартное отклонение	5,79

ПРАКТИКА

КЕЙСЫ

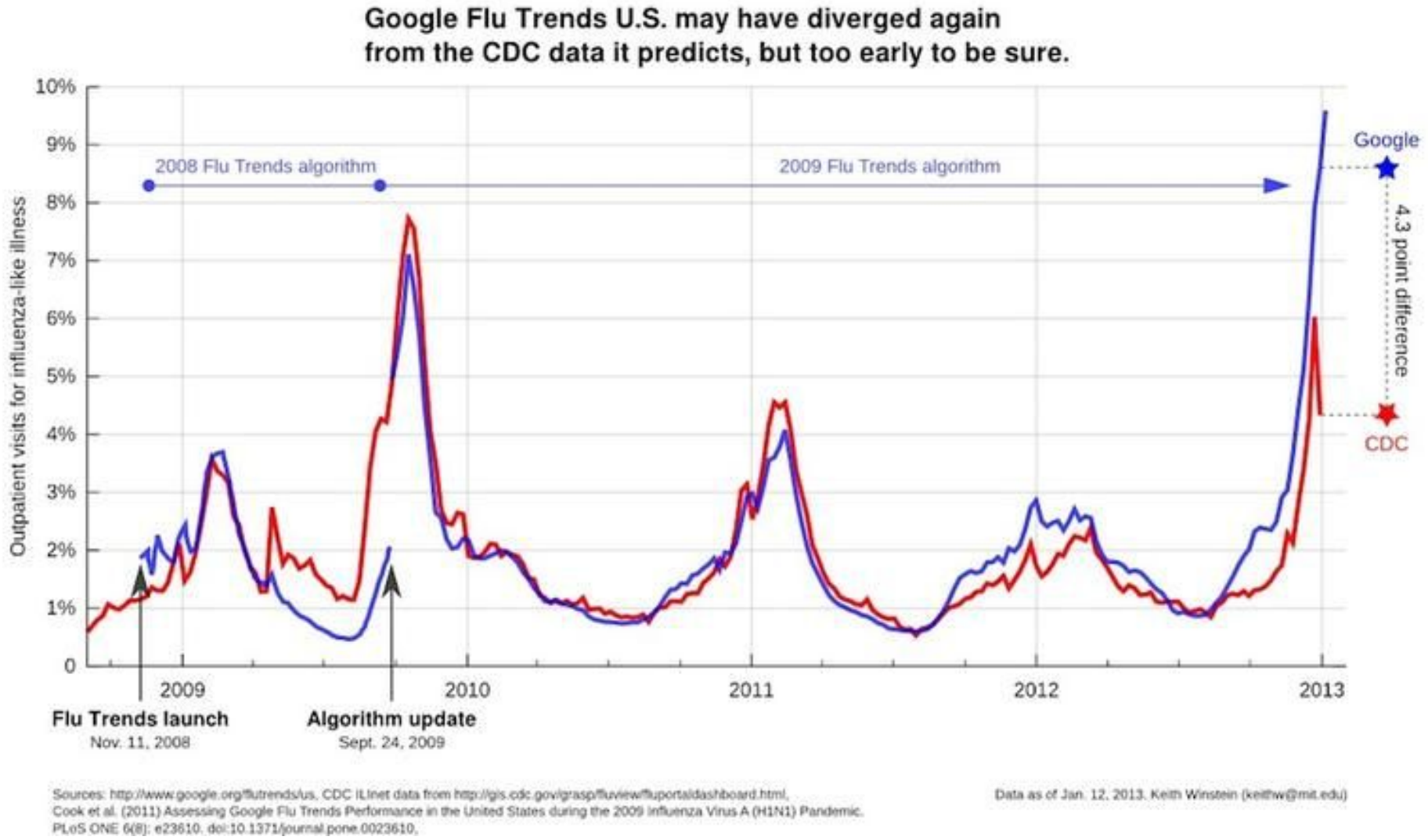
Google Flu Trends

Специалисты Google обратили внимание, что вспышке эпидемий гриппа предшествует всплеск запросов, связанных со здоровьем. Чтобы проверить свои наблюдения, *они взяли 50 миллионов наиболее популярных в США запросов и сопоставили частоту их появления с данными об эпидемиях гриппа, которые наблюдались между 2003 и 2008 годами.*

Им удалось идентифицировать сочетание 45 запросов, частота использования которых коррелирует со вспышками эпидемий.

Особенно интересен тот факт, что всплеск наблюдается по меньшей мере за две недели до того, как медикам удаётся зафиксировать начало эпидемии. В некоторых случаях задержка ещё дольше. Например, первые признаки эпидемии атипичной пневмонии появились в интернете за два с лишним месяца до того, как её заметила ВОЗ.

КЕЙСЫ / Google Flu Trends



Google Flu Trends

Статья в Science указала на существенные неточности в прогнозах Google Flu Trends.

Сервис более чем на 50% преувеличил размах эпидемии гриппа в сезоны 2012–2013 и 2011–2012 годов.

Согласно оценке Google Flu Trends, в разгар прошлогодней эпидемии около 11% жителей США заразились гриппом. Это почти вдвое выше цифр Центра по контролю и профилактике заболеваний США, который не оценивает количество больных по косвенным признакам, а просто пересчитывает их.

Кроме того, алгоритмы Google совершенно прозевали вспышку эпидемии вируса H1N1-A (“свиной грипп”) в 2009-м.

Сеть Target и тесты на беременность

Американская торговая сеть Target узнала о беременности девушки раньше, чем её отец.

«Она ещё в школу ходит,
а вы посылаете ей купоны
на детскую одежду
и памперсы?», —
кричал тогда
рассерженный отец.

Выбор канала привлечения

Представьте предпринимателя, который рекламирует в Google свой интернет-магазин.

Один клик на объявление стоит $1,5, а прибыль от каждой продажи составляет $150. После 500 показов — две продажи. Реклама вырубается как неэффективная. Запускается другое объявление. После 500 показов — пять броней. Вывод: эта кампания в два с половиной раза эффективнее. Зальём туда денег побольше.

Ещё тысяча показов — и всего четыре брони. Конверсия внезапно упала до уровня прошлой кампании. Почему?
Предприниматель принял решение об эффективности рекламы на основе слишком малой выборки.

Summary

1

Для точного анализа не всегда достаточно только тех данных, которые у нас есть

2

Прежде чем делать выводы, надо проверить полученные значения на статистическую значимость

3

При оценке величины по случайной выборке стоит рассчитать доверительный интервал для полученного значения

Дополнительные материалы

1. Оценить статистическую значимость:

   https://ru.wikihow.com/оценить-статистическую-значимость

2. Рассчитать величину P или значение вероятности:

   https://ru.wikihow.com/посчитать-величину-P-или-значение-вероятности

3. Вычислить доверительный интервал

   https://ru.wikihow.com/вычислить-доверительный-интервал

Дополнительные темы

Формат CSV

Применение формата CSV

CSV (от англ. Comma-Separated Values — значения, разделённые запятыми) — текстовый формат, предназначенный для представления табличных данных.

- Каждая строка файла — это одна строка таблицы
- Разделителем (англ. delimiter) значений колонок является символ запятой (,)
- Значения, содержащие зарезервированные символы (двойная кавычка, запятая, точка с запятой, новая строка) обрамляются двойными кавычками (")

Gretl

Инструмент Gretl

GNU Regression, Econometrics and Time-series Library: Библиотека для регрессий, эконометрики и временных рядов) — прикладной программный пакет для эконометрического моделирования.

Эконометрика — наука, изучающая количественные и качественные экономические взаимосвязи с помощью математических и статистических методов и моделей.

http://gretl.sourceforge.net/

СПАСИБО ЗА ВНИМАНИЕ