

Основы работы с данными: сбор, анализ, визуализация

Лекция 5. Основы статистики: МЦТ, нормальное распределение, ген сов-сть, выборка. Работа с табличными данными в pandas. Постановка исследовательской задачи.



Максим Карпов
@buntar29
mekarpov@hse.ru

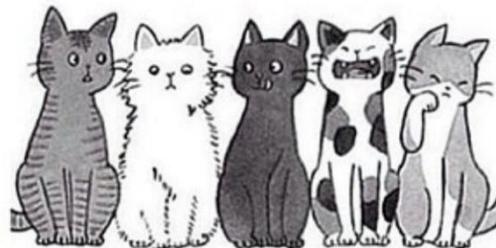


HSE
University

Генеральная
совокупность



Выборка



Наблюдение
(измерение)



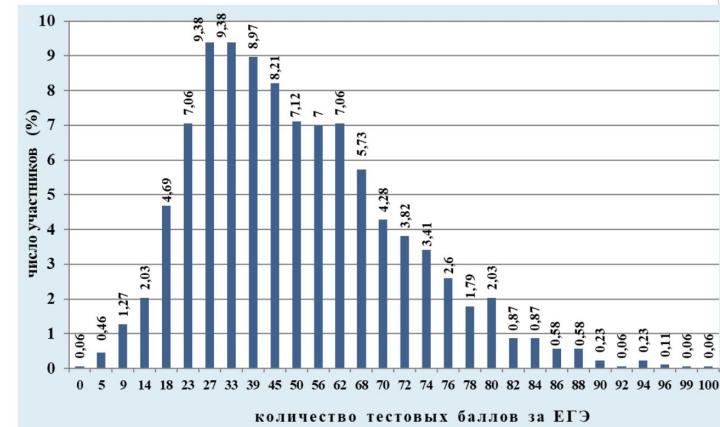
Гистограмма распределения

Чтобы получить полную информацию об исследуемых данных, необходимо получить закон распределения этих данных. Сейчас мы не будем давать определение понятия *распределение*. Но обсудим, что такое *гистограмма распределения*.

Гистограмма - это способ представления статистических данных в графическом виде - в виде столбчатой диаграммы. Она отображает распределение отдельных измерений параметров изделия или процесса. Иногда ее называют частотным распределением, так как гистограмма показывает частоту появления измеренных значений параметров объекта.

Высота каждого столбца указывает на частоту появления значений параметров в выбранном диапазоне, а количество столбцов - на число выбранных диапазонов.

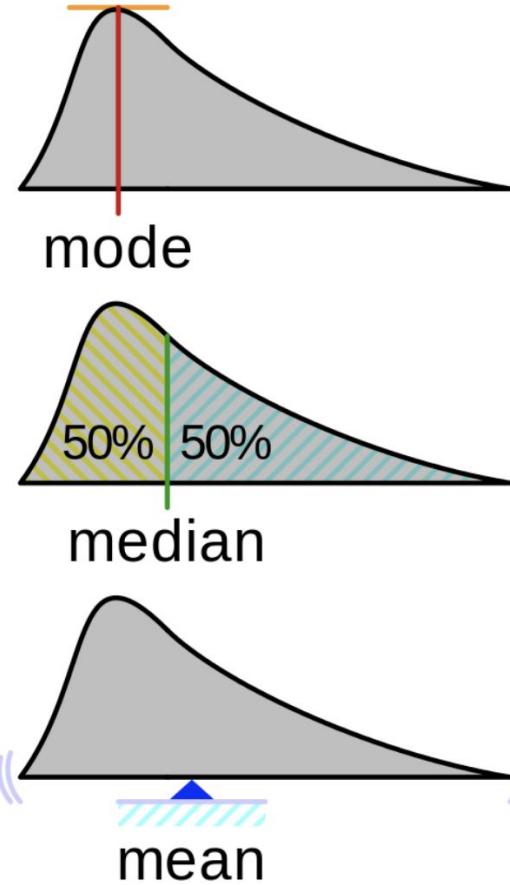
Гистограмма распределения (пример)



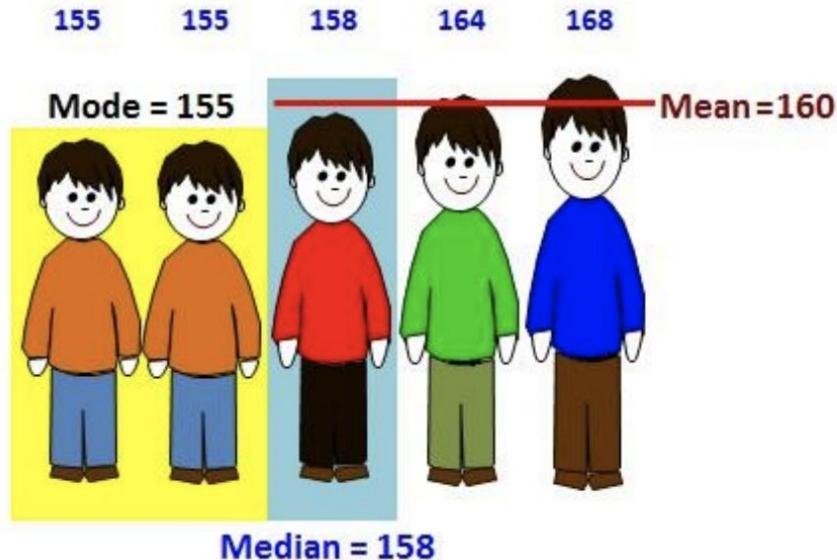
Мода (mode) -наиболее часто встречающееся значение, локальный максимум.

Медиана(median)-значение, которое делит распределение пополам(его площадь в т.ч.): половина значений больше медианы, половина -не меньше.

Среднее значение (mean) -сумма всех значений переменной, делённая на количество значений.



Среднее значение, медиана и мода - это различные способы измерить “центральное” число в наборе данных. Каждый из этих способов пытается выразить информацию о данных с помощью одного “усредненного” числа.



Умение интерпретировать описательные статистики и особенности распределения переменных.

A-4

У нас есть данные о численности зайцев и рысей в отдельно взятом лесу за 20 лет. Изучите описательные статистики в таблице ниже и выберите верное утверждение:

	Зайцы	Рыси
Среднее за 20 лет	34080	20166
Медиана за 20 лет	25400	12300
Стандартное отклонение за 20 лет	21413	16655

1. В популяции рысей наблюдается большая вариативность показателей
2. Разница между медианой и средним значением для численности зайцев может свидетельствовать о том, что в некоторые годы численность популяций зайцев была экстремально низкой
3. Разница между медианой и средним значением для численности рысей может свидетельствовать о том, что в некоторые годы численность популяции рысей была экстремально высокой
4. Гарантированно, что в один из годов в популяции рысей было 20166 особей

Умение определить типы переменных.

A-5

Соотнесите переменную и ее тип:

Количество детей в семье	количественная дискретная
Рост человека	количественная непрерывная
Наличие профиля в социальной сети Facebook	категориальная бинарная
Сорт винограда	категориальная номинальная
Уровень дохода семьи (низкий, средний, высокий)	категориальная порядковая

Умение определить типы переменных и выбрать подходящие метрики для их описания.

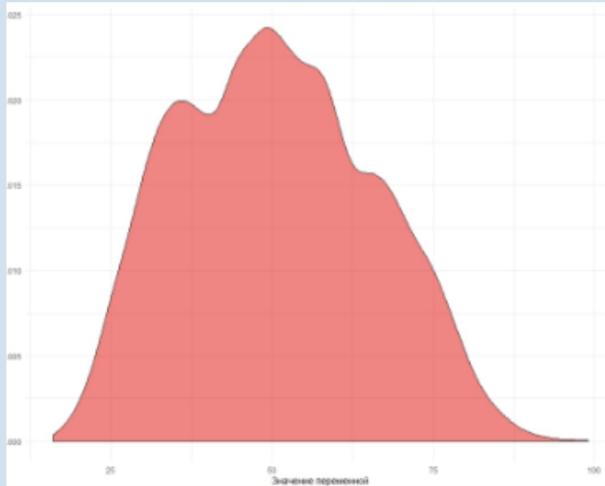
A-6

Какую меры измерения центральной тенденции корректно использовать для переменной с информацией об уровне образования, закодированной числами? (1 – средняя школа, 2 – бакалавриат, 3 – магистратура, 4 – аспирантура).

1. Медиана
2. Мода
3. Арифметическое среднее
4. Гармоническое среднее

A-7

Какой переменной может соответствовать график распределения, приведенный ниже



1. Количество детей в российских семьях
2. Рост взрослого человека в сантиметрах
3. Возраст владельцев домохозяйств
4. Количество воды, выпиваемое человеком в сутки в литрах

Умение
определить
распределение
переменной.



Умение оценить смещение выборки

A-9

Мы хотим узнать, какой средний уровень образования у совершеннолетних женщин в России. Из вариантов ниже выберите тот, в котором не происходит гарантированного смещения выборки:

- а) Были опрошены только работающие женщины.
- б) Было опрошено такое количество женщин из каждого региона, которое соответствует доле женщин, проживающих в этом регионе согласно переписи населения.
- в) Опрос был проведен посредством социальных сетей для их пользователей.
- г) Были опрошены жительницы всех городов с населением более 300 000 человек.

Меры разброса

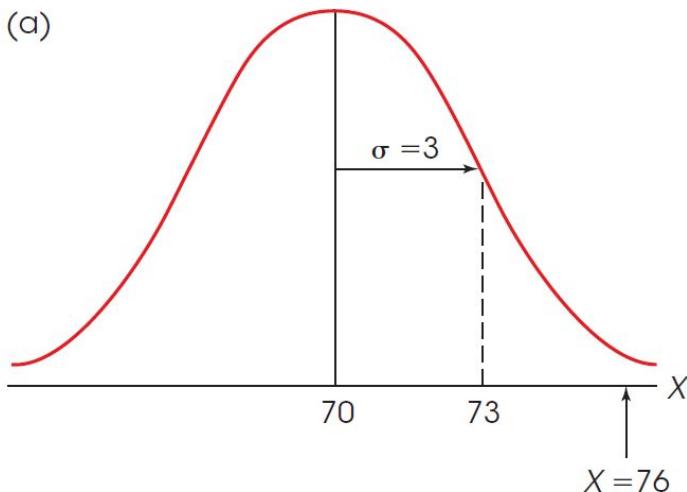
Меры разброса Variability

Говорят о том, насколько сильно рассеяны значения в данных.

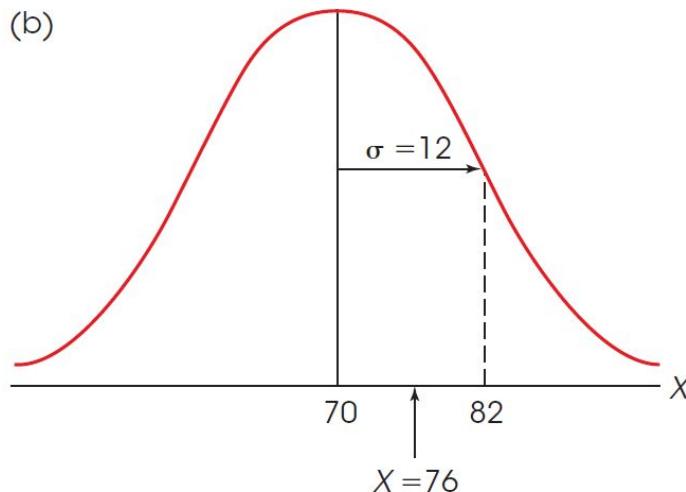
Размах, дисперсия, стандартное отклонение.

Зачем нужны меры разброса?

(a)

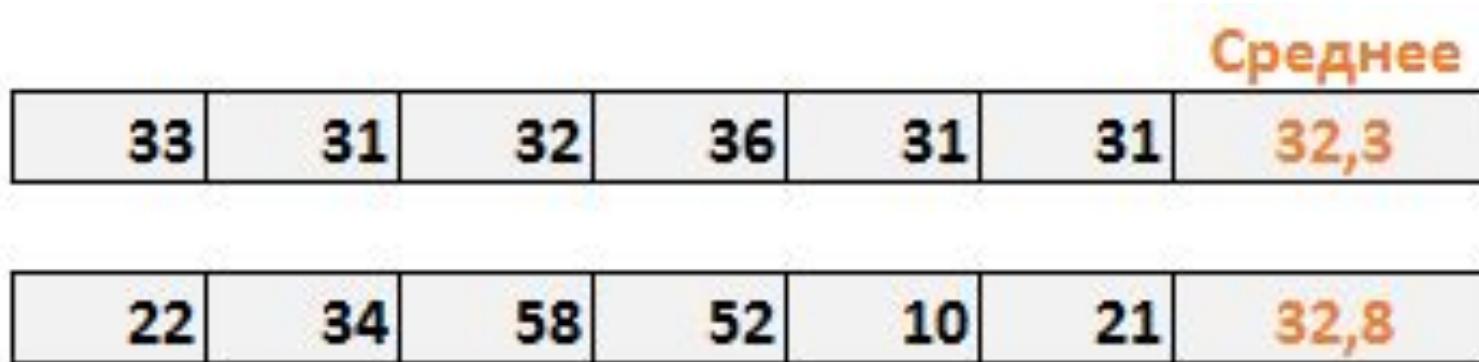


(b)



Чтобы описывать распределения более полно. Так мера разброса говорит нам, например, насколько плотно значения сконцентрированы вокруг среднего.

Зачем нужны меры разброса



Представьте, что вы хотите сделать некий вывод про людей. И в том и в другом случае вы скажите, что ваши выводы корректны для людей 32 лет.

Но разве это корректно?

Размах Range

Разность между самым большим и самым маленьким значениями в выборке.

- Плохо подходит, когда в данных есть выбросы.
=> необычно большой размах или крайне экстремальные минимальное или максимальное значения могут быть поводом для дальнейшего исследования.

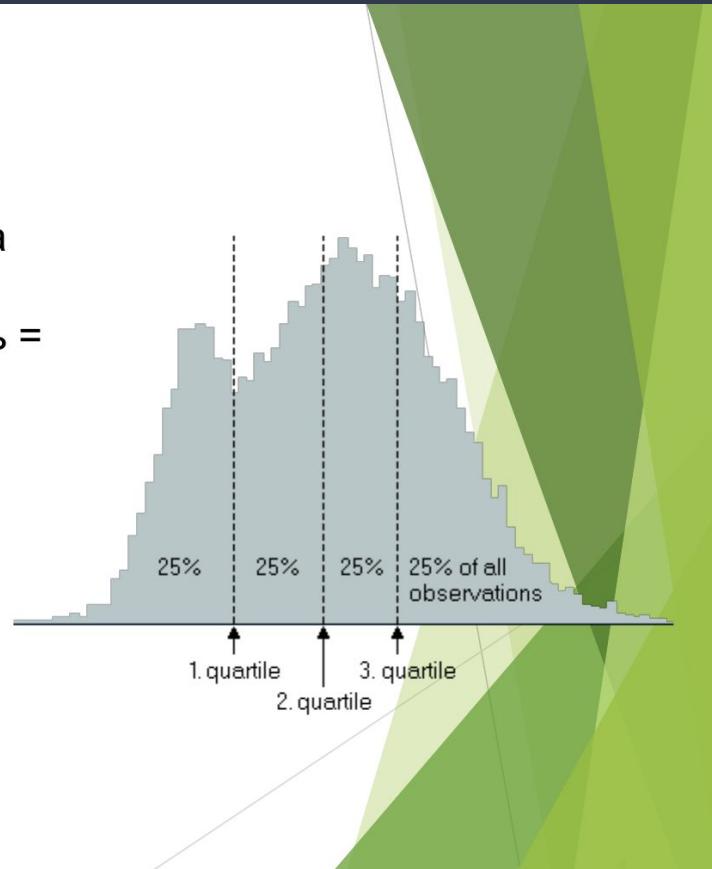
$$\text{range} = X_{\max} - X_{\min}$$

Квартили

Квартили(quartiles) делят распределение на четыре части так, что в каждой из них оказывается поровну значений (2-я квартиль = медиана).

1-я квартиль = 25% процентиль
3-я квартиль = 75% процентиль

Интерквартильный размах –
разница между третьей и первой
квартилями.



Интер- квартильный размах

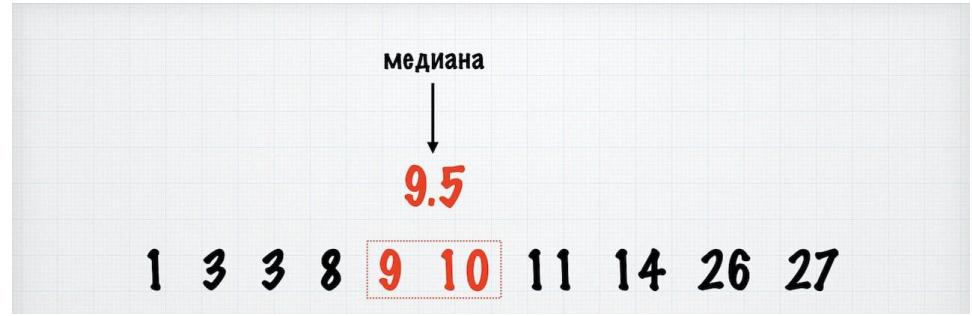
Interquartile range

Разность 75-й и 25-й перцентилей
(медианы половинок данных до
медианы и после медианы (которая
сама по себе 50-я перцентиль)).

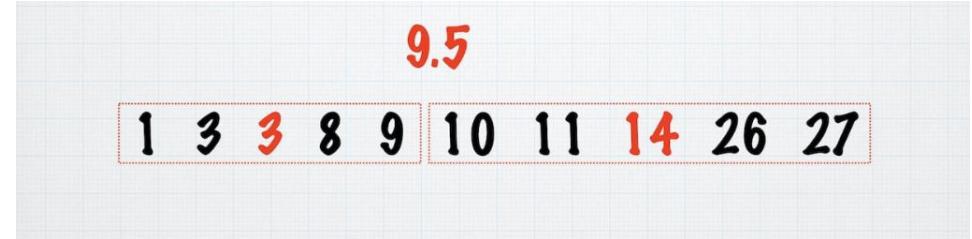
Применяется в паре с медианой!

Интерквартильный размах

Interquartile range



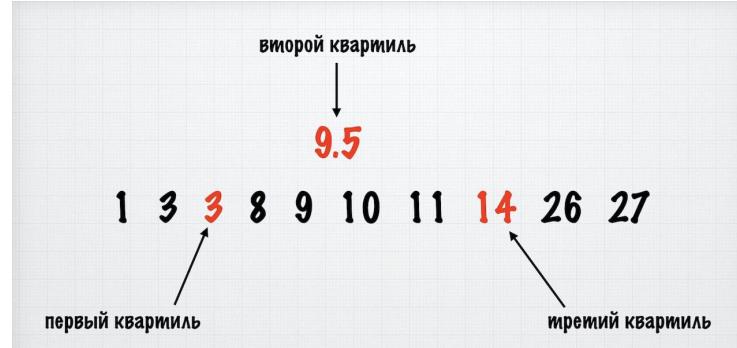
- Находим медиану (срединное значение выборки, она же 50-ю перцентиль (справа и слева от нее по 50% выборки)



- Теперь, чтобы найти 25-ю и 75-ю перцентиль - находим “медианы” левой и правой половин выборок (слева и справа от медианы).

Интерквартильный размах

Interquartile range



- Медиана всей выборки – это второй квартиль, медианы левой и правой половин – это, соответственно первый (или нижний) и третий (или верхний) квартили. Они же 25-я и 75-я перцентили.



- Разница между 75-й и 25-й перцентилями и будет интерквартильным размахом.

Интерквартильный размах

Interquartile range



- В данном случае, интерквартильных размах равен **$14 - 3 = 11$**
- Это значит, что 50% наших данных с медианой 9.5 находятся в интервале между 25 и 75 перцентилем (2 и 3 квартилью) и имеют интерквартильный размах 11 (от 3 до 14).

Интерквартильный размах

Interquartile range



- Представьте, что мы опросили 10 человек, о том, сколько они тратят в месяц на ребенка и спросили о возрасте их детей. Например, у нас получилось, что они тратят, в среднем, 30 000 рублей в месяц. Медиана нашей выборки 9.5 и мы бы хотели сделать вывод, что “опрошенные в среднем тратят 30 000 рублей в месяц на детей 9.5 лет”.
- Но если мы учтем интерквартильный размах (11, 50% детей наших опрошенных находятся в интервале от 3 до 14 лет), то увидим, что делать вывод, что наша выборка отвечает на вопрос про детей 9.5 лет - неверно. Корректно указать и медиану, и интерквартильный размах.

Интерквартильный размах Interquartile range



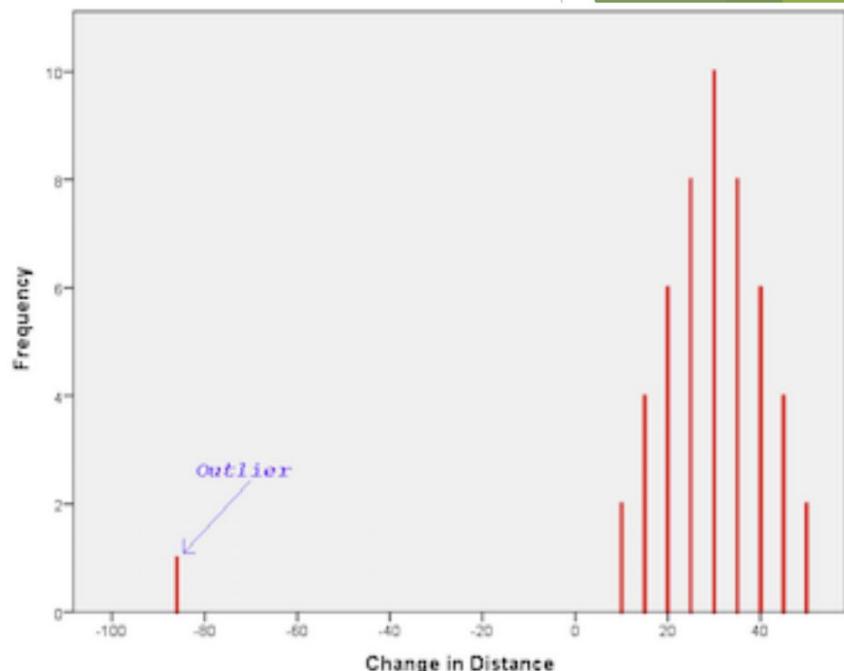
- Если в выборке нечетное количество элементов, то половинки выборок будут содержать четное количество элементов. Действуем так же как в случае с медианой (находим серединки между двумя элементами, складываем их и делим на два).

Выбросы

Выбросы - это наблюдения в анализируемых данных, значения которых сильно отличаются от других значений из этой выборки, размах,

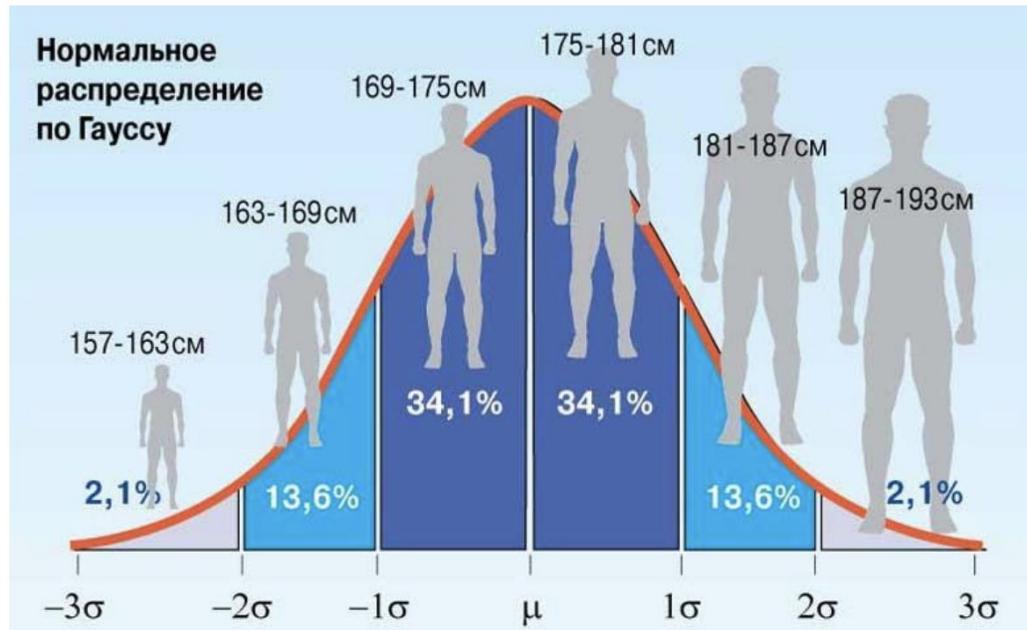
Выбросы могут происходить из-за:

1. Ошибки эксперимента.
2. Специфических условий в данной точке.
3. Случайности.



Распределение случайной величины

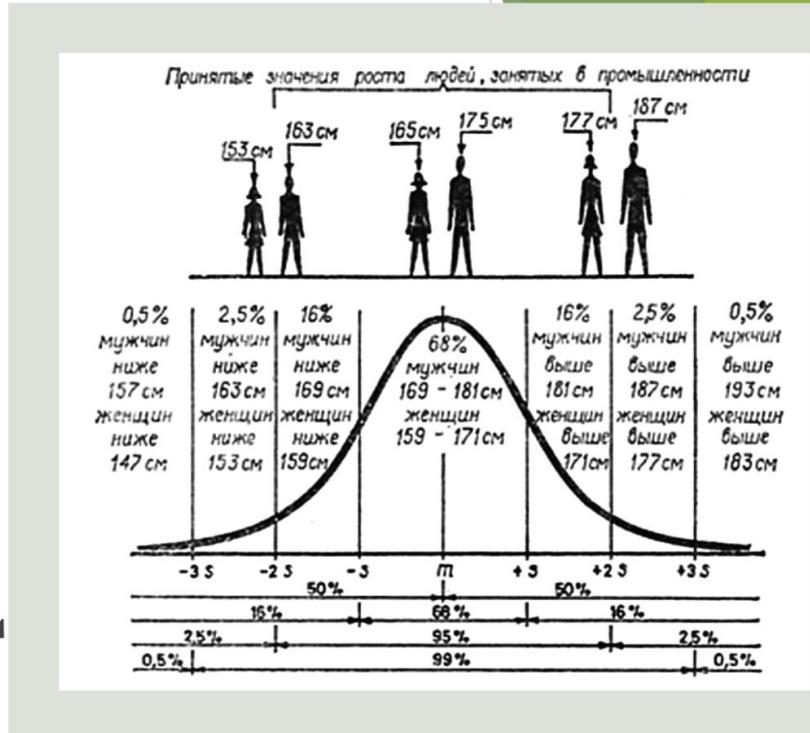
- ▶ Пример: если взять 1000 случайных жителей мира и измерить их рост, то окажется, что рост большинства людей будет составлять около 165-175 см. Но будут и люди значительно ниже, и значительно выше.



Плотность (вероятности) распределения

На рисунке изображена плотность распределения роста людей, занятых в промышленности.

- ▶ Мы видим, что значение роста большинства людей колеблется около среднего $t \approx 165$ см для женщин и $t \approx 175$ см для мужчин.
- ▶ При этом *вероятность того, что рост случайно взятой женщины попадёт в интервал $(165 - 6; 165 + 6) = (159; 171)$ см равна 68%.*



КАК ВЫБРАТЬ ИНТЕРЕСНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ВОПРОС?

«Обыкновенный» вопрос

- Можно просто загуглить
- С ними мы чаще сталкиваемся
- Ответ на него не предполагает открытия чего-то нового

Исследовательский вопрос

- Более дискуссионен
- На него «сложнее» ответить
- Информация из разных источников



«Обыкновенный» вопрос



- Сколько человек записались на олимпиаду по анализу данных?
- Сколько в Москве троллейбусов?
- Где находится штаб-квартира корпорации Apple?
- Что такое регрессия?

Исследовательский вопрос



Обязательно:

- на исследовательский вопрос можно ответить (есть данные)
- если мы на него ответим, то наше понимание мира вокруг улучшится

Примеры:

- Как 16-ти часовые перелёты влияют на скорость реакции пилотов коммерческих авиалайнеров?
- Правда ли, что наличие телевизора в доме мешает развитию ребенка?
- Как связано место рождения ребёнка и вероятность развития у него астмы?

Исследовательский вопрос



Какая из серий сериала “Рик и Морти” самая лучшая?

Это плохо поставленный исследовательский вопрос

Какая из серий Рика и Морти была просмотрена больше всего раз?

С этим вопросом всё в порядке



Исследовательский вопрос



- Может возникнуть как решение проблемы «неполноты» существующей теории
- Ответ может дополнять существующую теорию
- Может «породить» новую
- Помогает «проверить на прочность» существующую теорию



Исследовательский вопрос – насколько он у вас хороший?



- Заранее представляйте, каким может оказаться ответ и как он впишется в теорию
- Какие данные и источники планируется привлечь?
- Даже если ответ представляется возможным найти, то какой ценой?
- Как и любому проекту исследованию нужен план
- Наблюдение или эксперимент? Какой?
- Упрощайте!

Исследовательский вопрос



- **Каковы детерминанты социальной мобильности?**

Исследование, в полной мере отвечающее на этот вопрос,
потребовало бы огромного количества ресурсов

- **Действительно ли место рождения является детерминантом
социальной мобильности?**

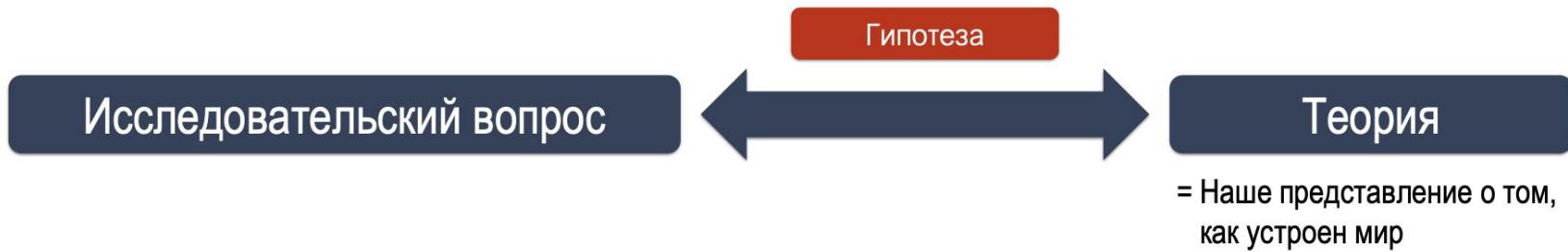
Уже лучше!

ЗАЧЕМ НУЖНА ГИПОТЕЗА?



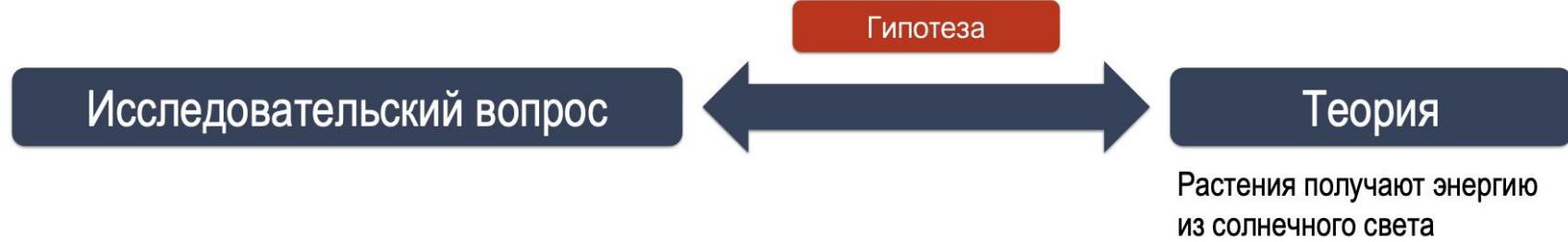
Как нам проверить, что наше представление о том, как устроен мир, верно?

ЗАЧЕМ НУЖНА ГИПОТЕЗА?



Гипотеза – это утверждение, следующее из теории, которое при этом можно проверить на данных

ПРИМЕР С ФОТОСИНТЕЗОМ



Как нам сформулировать гипотезу, чтобы проверить эту теорию?

ПРИМЕР С ФОТОСИНТЕЗОМ

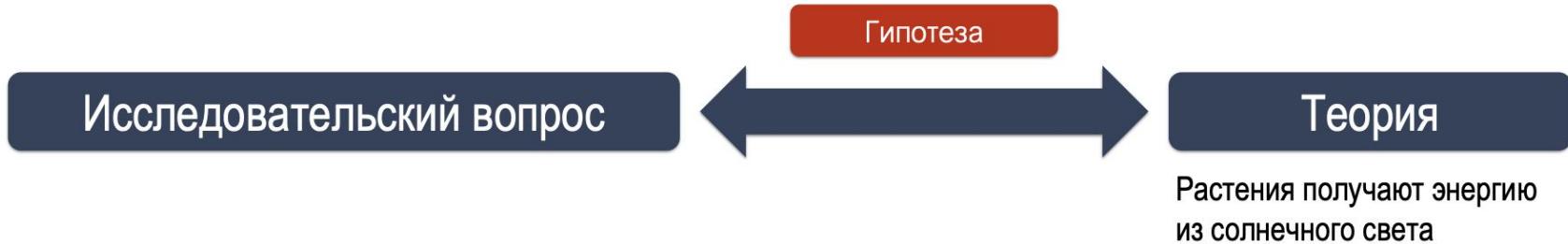


Как нам сформулировать гипотезу, чтобы проверить эту теорию?

- Растения получающие больше света, лучше растут

Плохая гипотеза

ПРИМЕР С ФОТОСИНТЕЗОМ



Как нам сформулировать гипотезу, чтобы проверить эту теорию?

- Когда мы поместим растение в темную комнату, без доступа солнечного света, растение погибнет

Хорошая гипотеза