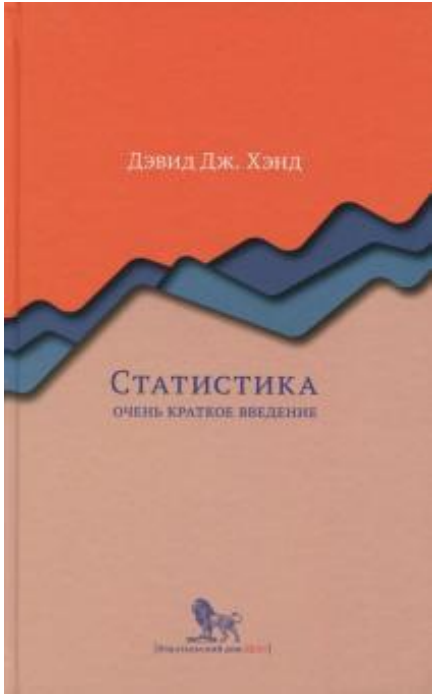
Видео 0

Полезные книжки



Видео 1

Виды данных

- Качественные

- Количественные

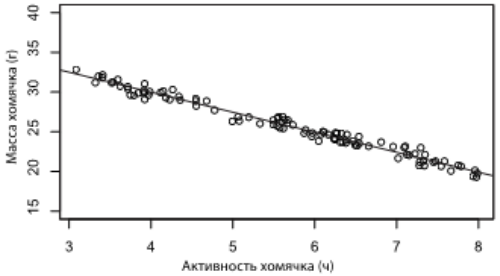
Виды данных

- Качественные
  - Номинальная шкала (номера машин, телефонов)
  - Порядковая шкала (оценки в школе)
- Количественные
  - Интервальная шкала (температура, время)
  - Абсолютная шкала (цены, количества)

Задача (Всерос 2018)

Начинающий исследователь Василий однажды читал научную статью, где изучалась связь активности хомячков (в часах в среднем в течение суток) и их массы. Там приводился следующий график:

Точками здесь обозначены отдельные хомячки, а прямая построена так, что она лежит как можно ближе к точкам. Прямая имеет уравнение $y = 40 - 2{,}5x$, из чего авторы исследования сделали вывод, что увеличение активности на час в среднем уменьшает массу хомячка на 2,5 грамма.



Василий решил использовать эту технику, чтобы оценить влияние образования людей на на их доход. Для этого он опросил 1000 человек, спрашивая у каждого, сколько тот зарабатывает (переменная Income) и каков его последний на данный момент уровень образования (переменная Education). Переменная Education для каждого человека принимает одно из 5 значений:

Education	Уровень образования
1	не окончил школу
2	окончил только школу
3	окончил техникум/колледж
4	окончил университет
5	получил ученую степень

Построив рядом с полученными точками прямую так же, как было сделано в исследовании, кото- рое он читал, Василий обнаружил, что ее уравнение имеет вид $\text{Income} = -10 + 20 \cdot \text{Education}$ (это уравнение прямой, самой близкой к точкам на графике), то есть каждая ступень образования в среднем увеличивает доход на 20 тысяч рублей в месяц.

Исследование, проведенное Василием, не свободно от недостатков. Вам нужно высказать содержательную критику по следующим пунктам:

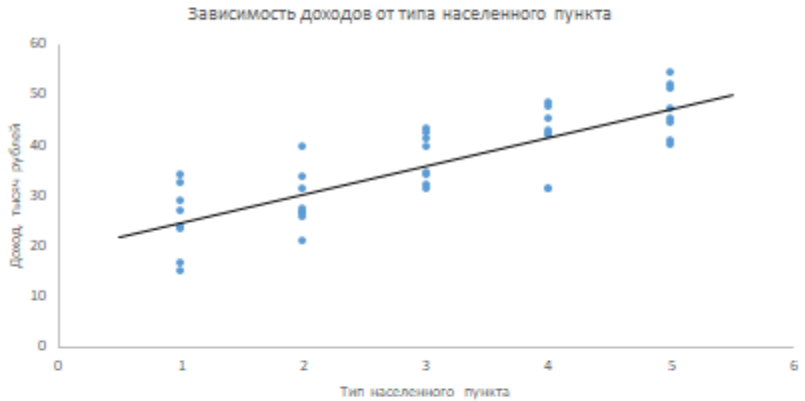
- а) (3 балла) Кодируя уровень образования цифрами от 1 до 5, Василий неявно предполагает очень специфическую форму зависимости дохода от образования. Объясните, какую форму зависимости предполагает Василий, почему так лучше не делать и как ему стоило бы правильно учесть образование в своей модели?
- б) (3 балла) Во-вторых, Василий не учел все факторы, которые могут влиять на доход. Какие? Предложите, как надо было организовать исследование, чтобы корректно измерить влияние образования на доход.

Почему важно понимать тип используемой шкалы?

- Простейший пример: исследование зависимости дохода человека от типа населенного пункта
- Тип населенного пункта закодируем так:
  - 1 – Село
  - 2 – ПГТ
  - 3 – Город
  - 4 – Областной центр
  - 5 – Москва
- И опросим 40 человек, по 8 из каждого типа

Почему важно понимать тип используемой шкалы?

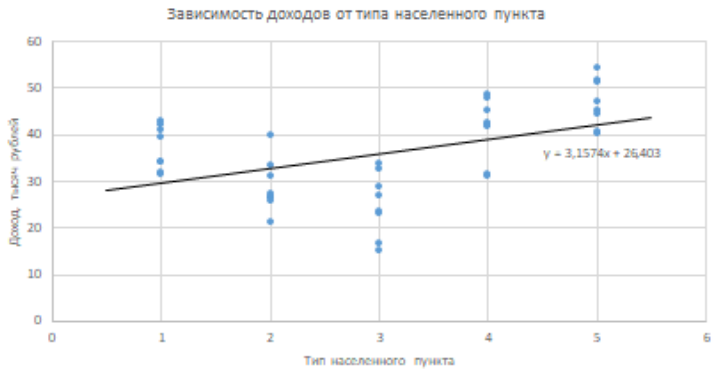
- Получили:



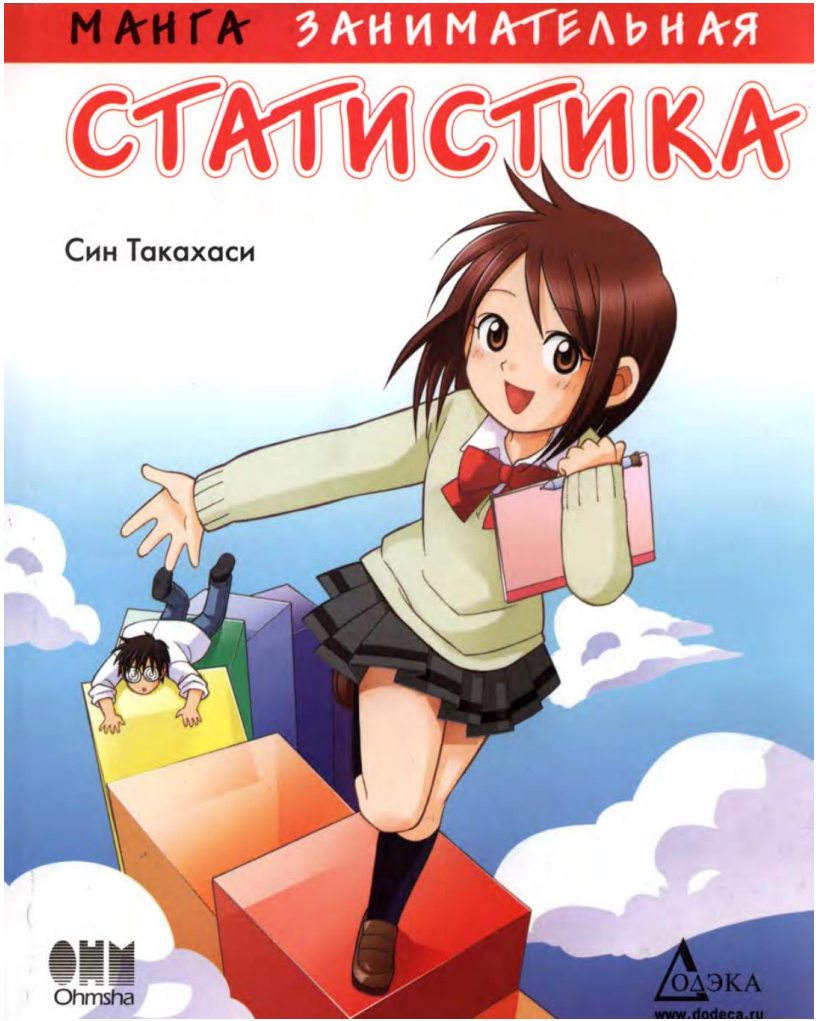
- Доход = 19 + 5.6*Тип населенного пункта

Почему важно понимать тип используемой шкалы?

- Давайте перекодируем:
  - 3 – Село
  - 2 – ПГТ
  - 1 – Город
  - 4 – Областной центр
  - 5 – Москва



- Доход = 26.4 + 3.16*Тип населенного пункта



Дополнительное чтение:
манга “Занимательная статистика”
глава 1

Видео 2

Видео 3

Показатели центра распределения

- Среднее – среднее арифметическое

$$\frac{7+7+5+8+9+10+9+10+7}{9} = \frac{72}{9} = 8$$

Показатели центра распределения

- Мода – самое частое значение в выборке

Значение	Частота
5	1
7	3
8	1
9	2
10	2

Мода!

Показатели центра распределения

- **Медиана – такое число, что половина элементов выборки меньше него, а половина – больше**

Упорядочим наблюдения по возрастанию

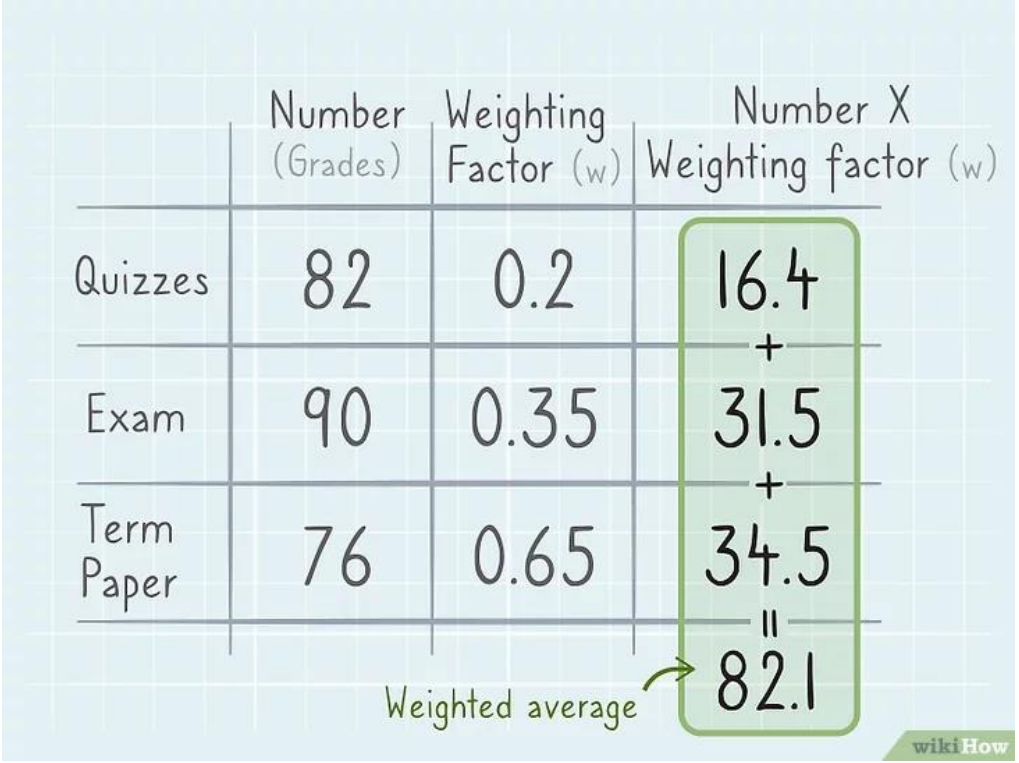
5 гр.
7 гр.
7 гр.
7 гр.
8 гр. **Медиана!**
9 гр.
9 гр.
10 гр.
10 гр.

Показатели центра распределения

- При **большой неоднородности данных, медиана и среднее** могут дать очень сильно **различающиеся результаты**
- К примеру: **100 человек** с доходом около **10 т.р.** и **10 человек** с доходом около **100 т.р.**
- **Среднее** около **20** т.р.
- **Медиана** около **10** т.р.

Про взвешенное среднее

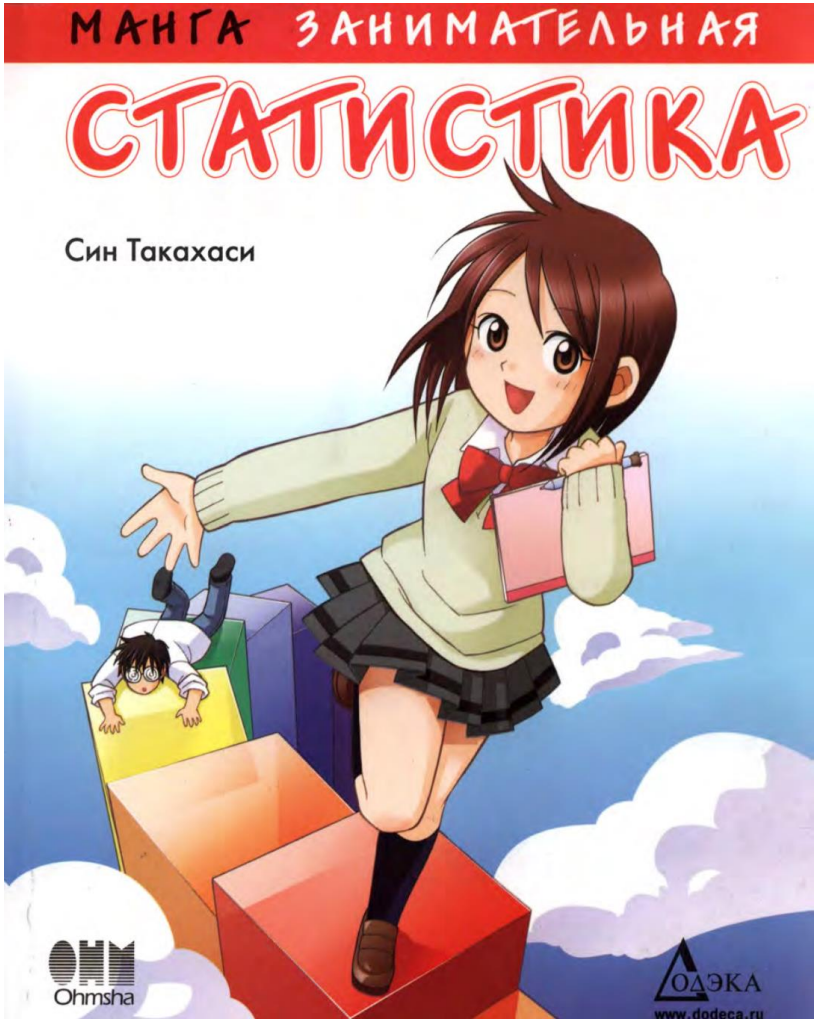
$$\overline{x} = \frac{\sum x_i \cdot f_i}{\sum f_i}$$



Задача (Высшая Проба 2016)

Вася, начинающий экономист и большой любитель сладостей, получил от одной кондитерской фабрики заказ: исследовать, сколько средств жители его родного города N-ска тратят на пирожные. В N-ске есть **три района: Центр, застроенный малоэтажными домами** ещё в царское время, **Спальный район, застроенный типовыми многоэтажками**, и **Частный сектор, застроенный частными домами**. В Центре проживает 10 тысяч человек, в Спальном районе — 100 тысяч, в Частном секторе — 2 тысячи. Вася должен опросить 500 жителей N-ска. Определиться с тем, кого конкретно нужно опрашивать, Вася не может. Он рассматривает несколько альтернатив:

- воспользоваться имеющейся у него базой данных со списком адресов и домашних телефонов всех жителей города: **случайным образом выбирать людей** и звонить им, пока не наберется 500 ответивших человек;
- **опросить по телефону отдельно 300 жителей Спального района, 150 жителей Центра и 50 жителей Частного сектора**, выбрав их случайным образом из имеющейся у Васи базы данных со списком адресов и домашних телефонов всех жителей города;
- **нанять 10 студентов, отправить их в 10 самых населенных домов N-ска** и дать задание каждому студенту опросить по 50 жителей своего дома.



Дополнительное чтение:
манга “Занимательная статистика”
глава 2

Видео 4

Показатели неоднородности

- **Размах – разность между максимальным и минимальным значением в выборке**

10 гр. - максимальное

5 гр. - минимальное

Размах = 10 – 5 = 5

Показатели неоднородности

- Стандартное отклонение $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$$\sqrt{\frac{(7-8)^2+(7-8)^2+(6-8)^2+(8-8)^2+(9-8)^2+(11-8)^2+(8-8)^2+(9-8)^2+(7-8)^2}{9}}=$$

$$=\sqrt{\frac{(1+1+4+0+1+9+0+1+1)}{9}}=\sqrt{\frac{18}{9}}=\sqrt{2}\approx 1.4$$

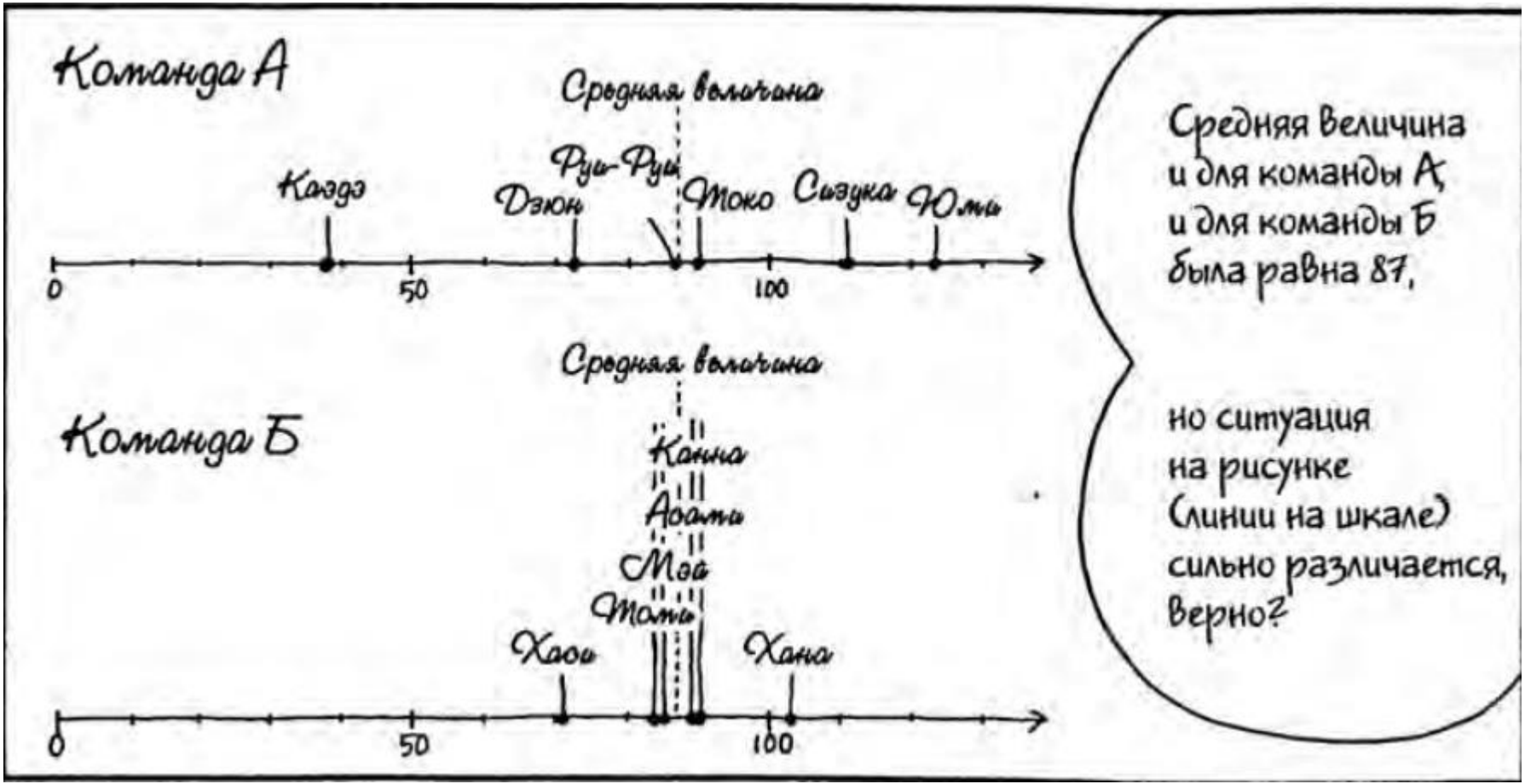
Показатели неоднородности

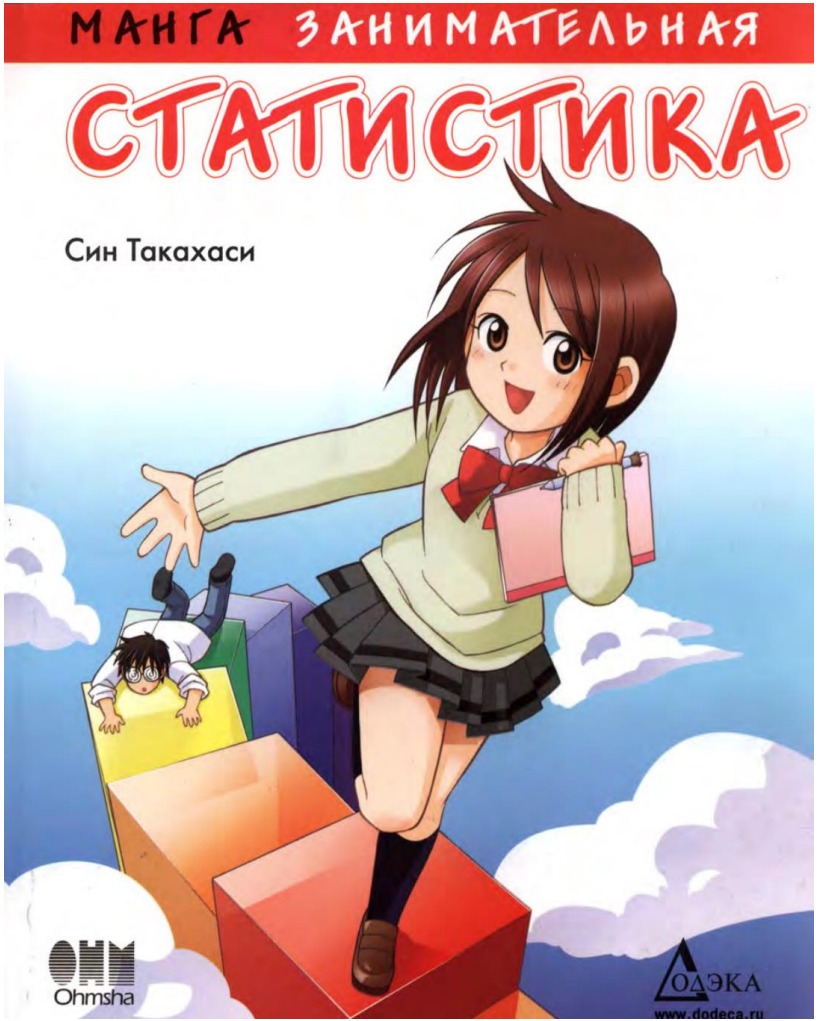
- Коэффициент вариации – стандартное отклонение, деленное на среднее

1.4 – стандартное отклонение

8 среднее

Коэф. Вариации = 1.4/8 = 0.175

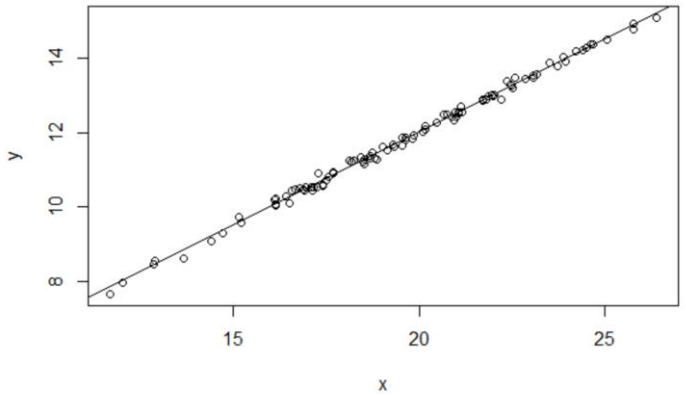

Команда А
Средняя величина
Каэдэ
Дэюн
Руи-Руи
Токо
Сидзука
Юмы
0
50
100
Средняя величина
Команда Б
Канна
Аоаты
Моа
Тоты
Хаои
Хана
0
50
100
Средняя величина и для команды А, и для команды Б была равна 87,
но ситуация на рисунке (линии на шкале) сильно различается, верно?

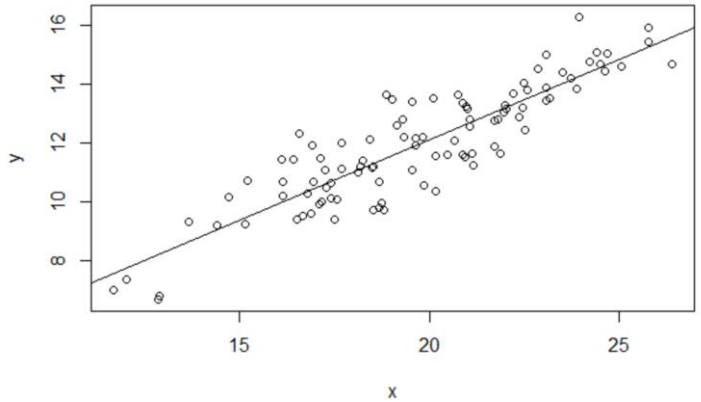


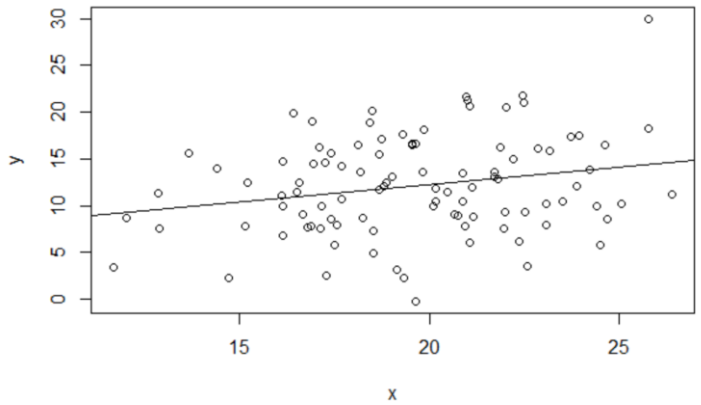
Дополнительное чтение:
манга “Занимательная статистика”
глава 2

Видео 5

Связь: на графике







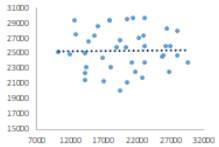
Корреляция

- Мера **линейной** связи между **двумя** показателями
- Рассчитывается по формуле:

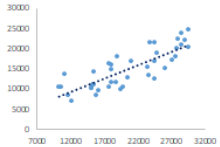
$$r_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}}$$

- Изменяется в пределах **от -1 до +1**
- Корреляция, равная **-1**, говорит об **отрицательной линейной связи** между переменными, **+1 – положительной**.
- **Нулевая** корреляция говорит об **отсутствии связи**.

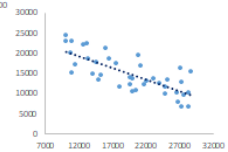
Корреляция



Корреляция около 0

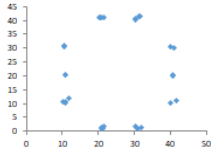


Корреляция +0.82

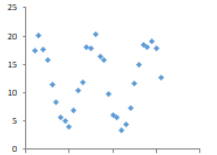


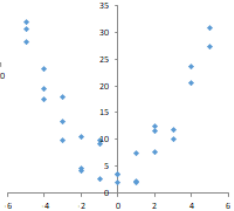
Корреляция -0.76

Корреляция - мера линейной связи



Корреляция везде около 0





Связь - таблицы сопряженности

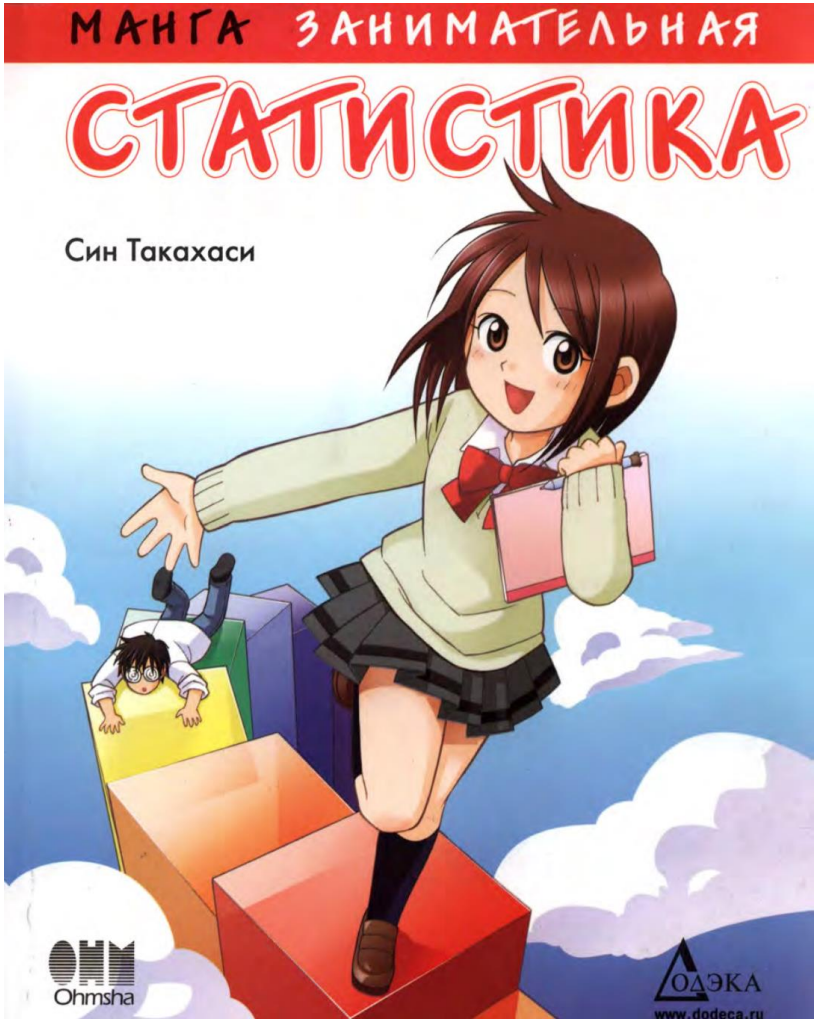
Цвет глаз / Цвет волос	Светлые	Тёмные
Серые	50 (a)	32 (b)
Карие	15 (c)	80 (d)

$$K_a = \frac{ad - bc}{ad + bc}$$

Задача (ВП 2018)

Экономист Вася прогнозирует курс криптовалюты smthCoin. Добыча этой валюты проходит достаточно остроумным путём: чтобы получить одну монету, нужно пройти уровень в компьютерной игре. С каждой добытой монетой сложность уровней повышается, а общий запас smthCoin в мире ограничен 100000 монет, когда все они будут получены, добыча монет прекратится.

1) Вася обнаружил, что сложность добычи монет в течение последних пяти лет неуклонно растёт, как и цена монет. Означает ли это, что через несколько лет монеты точно будут стоить гораздо больше, чем сейчас (сложность-то вырастет!)? Обоснуйте свой ответ. Что будет происходить с ценой smthCoin, когда будут добыты все монеты?



Дополнительное чтение:
манга “Занимательная статистика”
глава 6