Предисловие

Незадолго до настоящего момента газета The New York Times подарила студентам, которые готовились стать выпускниками университетов, статью, озаглавленную следующим образом[1]:

«Сегодняшним выпускникам, только одно слово: статистика»
(For Today's Graduate, Just One Word: Statistics).

В этой статье главный экономист компании Google говорит: «Я не устаю говорить о том, что самой привлекательной работой на следующие 10 лет будет профессия статистика, и я не шучу» (I keep saying that the sexy job in the next 10 years will be statisticians, And I'm not kidding).

И всем известно, что не только Google, но и такие ведущие мировые компании, как Microsoft, IBM, – все они уже сражаются между собой за подобные кадры.

Кроме того, в нашей стране дело с этим обстоит так же. В последнее время меня часто просят читать лекции по статистике для сотрудников фирм, и все мои слушатели говорят с сожалением, что им надо было получше изучать статистику в университетах. И они рассказывают мне о том, как в компаниях нужны кадры, владеющие статистическим анализом, и о том, как их там не хватает.

В этой книге я постарался очень подробно объяснить все, начиная с азов и до практического применения, ориентируясь на читателей, которые чувствуют, что статистический анализ стал нужен в учебных заведениях и вообще в обществе в целом, но не понятно, что и как делать, или на тех читателей, которые, прочитав учебник для начинающих, не понимают, какой метод лучше выбрать, когда дело доходит до практического применения.

Выпущенная в 2011 году книга «Вводный курс в статистику: от критериев до планирования экспериментов» сразу же завоевала огромную популярность. Данная книга написана не с целью поймать второго вьюна[2] (хотя в некоторой степени именно с этой целью...), но, используя содержание и структуру этой книги в качестве основы, мы дополнили ее большим количеством графических изображений, таблиц и иллюстраций. Это позволяет читать ее легко и непринуждённо, листая как книжку с картинками. Кроме того, в книге помещены способы, основанные на бесплатной программе R, для выполнения тех методов анализа, которые невозможно провести с помощью программ табличных вычислений. Конечно, данные, используемые в этой книге, можно скачать с сайта Ohmsha[3].

[1] https://www.nytimes.com/2009/08/06/technology/06stats.html.
[2] Повторить успех. – *Прим. перев.*
[3] http://www.ohmsha.co.jp/data/link/bs01.html.

Далее, на этот раз мне помогал мой коллега по университету Маруяма Ацуси, поэтому я уверен, что доходчивость материала революционно увеличилась (хотя думать так и нескромно).

Итак, давайте вместе откроем дверь в статистику и приступим к статистическому анализу данных!

Август 2017 г. Представитель авторов *Курихара Синъити*

Содержание

Предисловие V

ПРОЛОГ. ЧТО ТАКОЕ СТАТИСТИКА?

Что такое статистика? 2
Возможности статистики 4

ГЛАВА 1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Разнообразные средние 8
Разброс данных ①. Квантиль и дисперсия 10
Разброс данных ②. Коэффициент вариации 12
Связь переменных ①. Коэффициент корреляции 14
Связь переменных ②. Ранговая корреляция 16

ГЛАВА 2. РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ

Вероятность и распределение вероятностей 20
Распределение с равными вероятностями. Равномерное распределение 22
Распределение при подбрасывании монеты. Биномиальное распределение 23
Распределение в виде подвешенного колокола. Нормальное распределение 24
Безразмерное распределение. Стандартное нормальное распределение 26
Узнаём позиции данных. Интервал сигма 29
Форма распределения. Асимметрия и куртозис 30
Распределение редко происходящих событий. Распределение Пуассона 32
Одновременная работа с множеством данных. Распределение хи-квадрат (χ^2) 34
Отношение значений хи-квадрат. Распределение Фишера (F-распределение) 36
Распределение, используемое вместо нормального. Распределение Стьюдента (t-распределение) 37

ГЛАВА 3. СТАТИСТИКА ВЫВОДА

Определяем особенности генеральной совокупности по выборке 42
Статистика вывода 42
Умело угадываем статистический параметр 44
Несмещённая оценка 44
Неограниченное число данных 46
Количество степеней свободы 46
Распределение выборочных статистик ①. Распределение средних 48
Распределение выборочных статистик ②. Распределение доли 50
Распределение выборочных статистик ③. Распределение дисперсии 51
Распределение выборочных статистик ④. Распределение коэффициента корреляции 52
Смещение от истинного значения. Систематическая ошибка и случайная ошибка 54
Две теоремы о выборочном среднем. Закон больших чисел и центральная предельная теорема 56

ГЛАВА 4. ОЦЕНКА ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА

Расширенная оценка ①. Доверительный интервал среднего генеральной совокупности 60
Расширенная оценка ②. Доверительный интервал доли в генеральной совокупности 64
Расширенная оценка ③. Доверительный интервал дисперсии генеральной совокупности 65
Расширенная оценка ④. Доверительный интервал коэффициента корреляции генеральной совокупности 66
Оцениваем статистический параметр путём моделирования. Бутстрэп-метод 68

ГЛАВА 5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Делаем вывод о наличии разницы. Проверка статистической гипотезы 72
Две статистические гипотезы. Нулевая гипотеза и альтернативная гипотеза 74
Порядок проверки статистической гипотезы 76
Проверка гипотезы о равенстве выборочного среднего определённому значению (среднему генеральной совокупности) 78
Два типа ошибок при проверке статистических гипотез. Ошибки 1-го рода и ошибки 2-го рода 84
Проверка гипотезы о равенстве долей признака в выборке определённому значению (доле в генеральной совокупности) 86

Проверка гипотезы о равенстве выборочной дисперсии определённому значению (дисперсии генеральной совокупности) 87
Действительно ли имеется корреляция? Проверка гипотезы о некоррелированности 88
Проверка гипотезы о разности средних ①. Случай двух непарных групп 90
Проверка гипотезы о разности средних ②. Случай двух парных групп 96
Проверка гипотезы о разнице долей. Случай двух непарных групп 98
Доказываем не меньшую эффективность. Испытания не меньшей эффективности 100

ГЛАВА 6. ДИСПЕРСИОННЫЙ АНАЛИЗ И МНОЖЕСТВЕННОЕ СРАВНЕНИЕ

Проверяем эффекты с помощью эксперимента. Однофакторный дисперсионный анализ 104
Проверка однородности дисперсий множества групп. Критерий Бартлетта 110
Учитываем межэкземплярную разницу. Парный однофакторный дисперсионный анализ 112
Находим эффекты взаимодействия. Двухфакторный дисперсионный анализ 114
Повторять проверку нельзя. Множественность сравнений 120
Проверка, допускающая повторение ①. Метод Бонферрони и метод Шеффе 122
Проверка, допускающая повторение ②. Метод Тьюки и метод Тьюки–Крамера 124
Проверка, допускающая повторение ③. Метод Даннетта 128

ГЛАВА 7. НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

Проверка, не зависящая от распределения. Непараметрические методы 132
Проверка качественных данных. Проверка независимости (критерий χ^2 Пирсона) 136
Проверка таблиц сопряжённости 2×2. Точный критерий Фишера 142
Проверка ранговых данных двух непарных групп. U-критерий Манна–Уитни 144
Проверка ранговых данных двух парных групп. Критерий знаков 148
Непараметрическая проверка количественных данных двух парных групп. Критерий знаковых рангов Уилкоксона 150
Проверка ранговых данных множества непарных групп. Критерий Краскела–Уоллиса 152
Проверка ранговых данных множества парных групп. Критерий Фридмана 154

ГЛАВА 8. ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА

Три основных принципа Фишера ①. Повторение 158
Три основных принципа Фишера ②. Рандомизация 160
Три основных принципа Фишера ③. Локальный контроль 162
Различные конфигурации точек эксперимента 164
Прореживаем число проводимых экспериментов. Ортогональный план 166
Ортогональный план на практике ①. Проектирование качества (проектирование параметров) 172
Ортогональный план на практике ②. Совместный анализ 174
Задание размера выборки. Анализ статистической мощности 176

ГЛАВА 9. РЕГРЕССИОННЫЙ АНАЛИЗ

Ищем связи причин и результатов. Регрессионный анализ 186
Применяем к данным числовую формулу. Метод наименьших квадратов 188
Оцениваем точность линии регрессии. Коэффициент смешанной корреляции 191
Проверяем наклон линии регрессии. t-критерий 192
Проверяем правильность анализа. Анализ остатков 195
Регрессионный анализ при наличии нескольких причин. Множественный регрессионный анализ 196
Проблема, возникающая между объясняющими переменными. Мультиколлинеарность 198
Выбираем эффективные объясняющие переменные. Методы выбора переменных 200
Переменные, объясняющие качественные различия ①. Фиктивная переменная сдвига 201
Переменные, объясняющие качественные различия ②. Фиктивная переменная наклона 202
Регрессионный анализ с бинарными переменными. Пробит-анализ 204
Анализируем время до наступления события ①. Кривая выживания 208
Анализируем время до наступления события ②. Сравнение кривых выживания 210
Анализируем время до наступления события ③. Регрессионный анализ пропорциональных рисков Кокса 211

ГЛАВА 10. МНОГОМЕРНАЯ СТАТИСТИКА

Обобщаем данные. Метод главных компонент 216
Открываем скрытые факторы. Факторный анализ 220
Описываем причинную структуру. Моделирование структурными уравнениями (SEM) 226

Классифицируем экземпляры. Кластерный анализ 234
Анализируем связи качественных данных. Анализ соответствий 242

ГЛАВА 11. БАЙЕСОВСКАЯ СТАТИСТИКА И БОЛЬШИЕ ДАННЫЕ

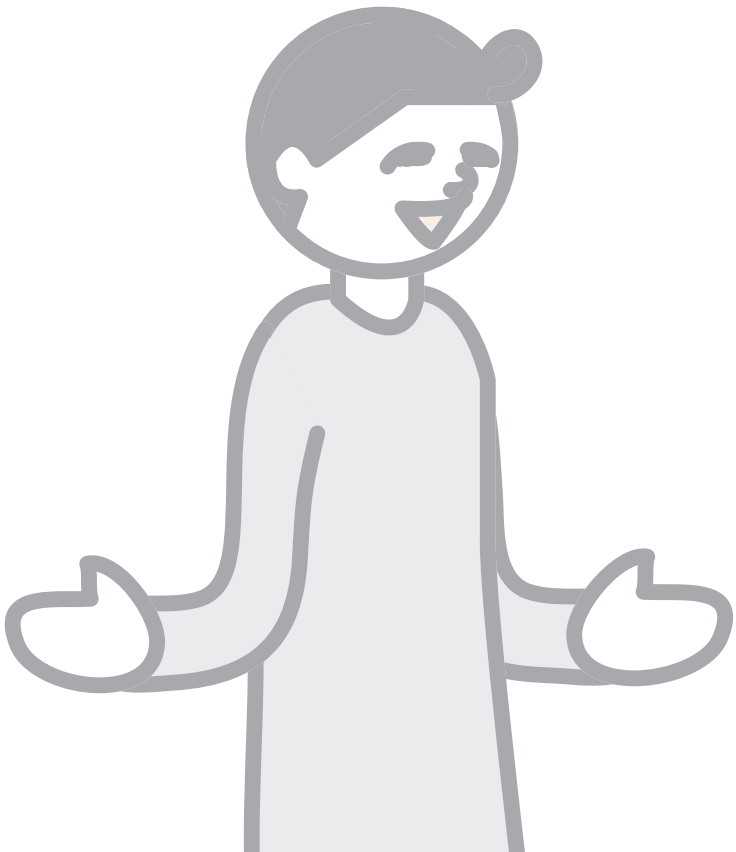
Статистика, использующая знания и опыт. Байесовская статистика 248
Универсальная формула. Теорема Байеса 250
Ищем причину, двигаясь от результата в обратном направлении. Апостериорная вероятность 252
Более точно благодаря новым данным. Байесовское обновление 256
Анализ больших данных ①. Что такое большие данные? 258
Анализ больших данных ②. Ассоциативный анализ 260
Анализ больших данных ③. Прогнозирование трендов и анализ SNS 262

ПРИЛОЖЕНИЕ A. ИНСТАЛЛЯЦИЯ И ИСПОЛЬЗОВАНИЕ R 265

ПРИЛОЖЕНИЕ B. СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ (ТАБЛИЦЫ РАСПРЕДЕЛЕНИЙ), ОРТОГОНАЛЬНЫЕ ТАБЛИЦЫ, БУКВЫ ГРЕЧЕСКОГО АЛФАВИТА 271

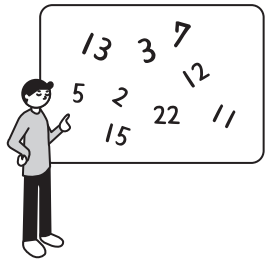
ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ 287

КРАТКАЯ БИОГРАФИЯ АВТОРОВ 291

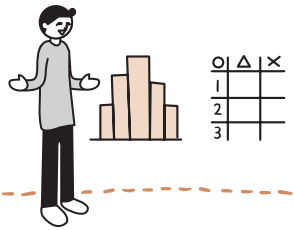


Пролог

Что такое статистика?



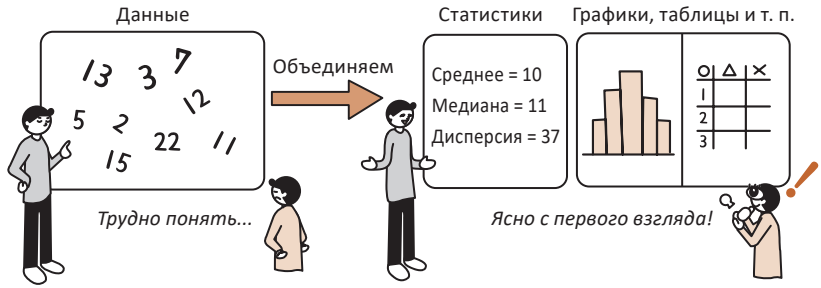
01 Что такое статистика?



Статистика в наши дни стала незаменимой наукой в области не только естественных, но и общественных дисциплин, таких как психология.

Статистика

- Статистика – это наука, в которой данные объединяют в статистики (такие как среднее), графики, таблицы и т. п. и находят особенности этих данных.



Виды статистики

- Статистика как наука включает разделы: описательная статистика, находящая особенности имеющихся данных; статистика вывода, в которой на основе выборки находятся особенности генеральной совокупности, которая стоит за этой выборкой; байесовская статистика, к которой приковано внимание в таких областях, как маркетинг.

* Существует также точка зрения, согласно которой байесовская статистика не включается в статистику вывода.

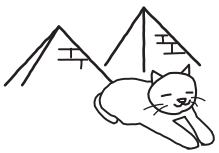
Статистика – набор данных, полученных путём измерения особенностей совокупности, которая является объектом анализа. Слово используется также и в смысле дисциплины.
Статистика – дисциплина, в которой систематизированы методы определения особенностей совокупностей, которые являются объектами исследования, и включающая описательную статистику и статистику вывода.

История статистики

Статистика – это один из разделов науки, поэтому мы говорим об истории ее развития. Не надо думать, что в один день на кого-то нашло озарение, и она появилась. В разделах Colulmn мы познакомим вас с теми из построивших современную статистику великих статистиков, которые внести особенно весомый вклад, но перед этим мы хотели бы кратко ознакомить вас с историей ее развития.

① Зарождение статистики: перепись населения

Хотя слово «статистика» означает научную дисциплину, оно также имеет смысл «набора данных». Английское название дисциплины – statistics – происходит от слова status, которое означает «состояние государства». Истоками статистики были переписи населения, которые государство проводило с целью сбора налогов и наложения повинностей. Например, есть записи о том, что в Древнем Египте переписи проводились для нужд строительства пирамид, а в Японии в период Асука переписи привязывались к площади рисовых полей.



② Самый первый статистический анализ: описательная статистика, начавшаяся с эпидемиологии

В середине XVII века, когда в Лондоне свирепствовала чума, Джон Грант провёл первый статистический анализ: используя статистические данные (записи о смертях), которые хранила церковь. Он установил, что смертность высока среди детей раннего возраста, что она выше в городах, чем в деревнях. Таким образом, он показал, что даже в тех общественных явлениях, которые считались случайно возникающими, путём большого количества наблюдений можно найти определённые закономерности. Подобная описательная статистика впоследствии развилась в полноценную научную дисциплину благодаря трудам Карла Пирсона.

③ Предсказываем общее с помощью теории вероятностей: статистика вывода

В начале XX века Рональд Фишер и Уильям Госсет, известный под псевдонимом Стьюдент, предположили, что на основе малых выборок (малого количества данных) можно судить об особенностях генеральной совокупности (статистических параметров). Кроме того, в последние годы внимание обращено и на байесовскую статистику, в которой считается, что статистические параметры сами по себе имеют вероятностное распределение. Вызывает изумление, что на это потребовалось меньше 100 лет со времени рождения статистики вывода, которая является незаменимой для современной жизни и исследований.

02

Возможности статистики

Статистика превратилась в дисциплину, незаменимую для нашей с вами жизни. Мы попробуем привести конкретные примеры, касающиеся её возможностей.

Описательная статистика

- Нахождение особенностей (среднего, разброса и др.), тенденций и другого для имеющихся данных.
- Эта статистика имеет дело с большим количеством данных (выборками большого размера).

Обработка данных переписей населения	Среднее, дисперсия и др. Глава 1
Корреляционный анализ количества пищи и веса тела	Коэффициент корреляции Глава 1
Расчёт рейтингов успеваемости в (школьных) классах	Нормированные случайные величины Глава 2
Методы закупки товаров для мини-маркетов	Big data (большие данные) Глава 11

Статистика вывода

- Делаем вывод об особенностях генеральной совокупности, используя информацию о выборке.
- Основным содержанием являются: несмещённая оценка, оценка доверительного интервала, проверка статистических гипотез.

Прогноз числа возникновений страховых случаев	Распределение Пуассона Глава 2
Предварительные итоги выборов / Телевизионные рейтинги	Оценка доверительного интервала Глава 4
Проверка эффективности новых лекарственных препаратов	Проверка разности средних двух групп Глава 5
Принятие решений о добавках к кормам	Дисперсионный анализ, метод множественного сравнения Глава 6
Анализ данных дегустации	Непараметрическая статистика Глава 7

Описательная статистика – дисциплина, в которой особенности данных наблюдений находятся с помощью таких статистик, как среднее или дисперсия, с помощью графиков и т. п.
Статистика вывода – дисциплина, в которой на основе данных наблюдений исследуют (оценивают, проверяют и т. п.) особенности генеральной совокупности, которая стоит за этими наблюдениями.

Планирование экспериментов

- «Сборник правил хорошего тона» для успешных экспериментов.
- Существуют также методы, позволяющие экономить время и место.

Порядок проведения и конфигурация точек плана эксперимента	Три основных правила Фишера Глава 8
Контроль качества изделий	Ортогональный экспериментальный план Глава 8
Выбор количества испытуемых (данных)	Анализ статистической мощности Глава 8

Множественный регрессионный анализ и многомерная статистика

- Общее наименование методов обработки множества переменных «за один раз».
- Представление сложной проблемы в виде простой модели для прогнозирования, оценки и т. п.

Оценка подержанных автомобилей скупщиком	Множественный регрессионный анализ Глава 9
Диагностика заболеваний на основе результатов (медицинских) исследований	Пробит-анализ Глава 9
Управленческий консалтинг компаний	Метод главных компонент Глава 10
Испытание на пригодность при приёме на работу	Факторный анализ Глава 10
Позиционирование бренда	Анализ соответствий Глава 10

Байесовская статистика

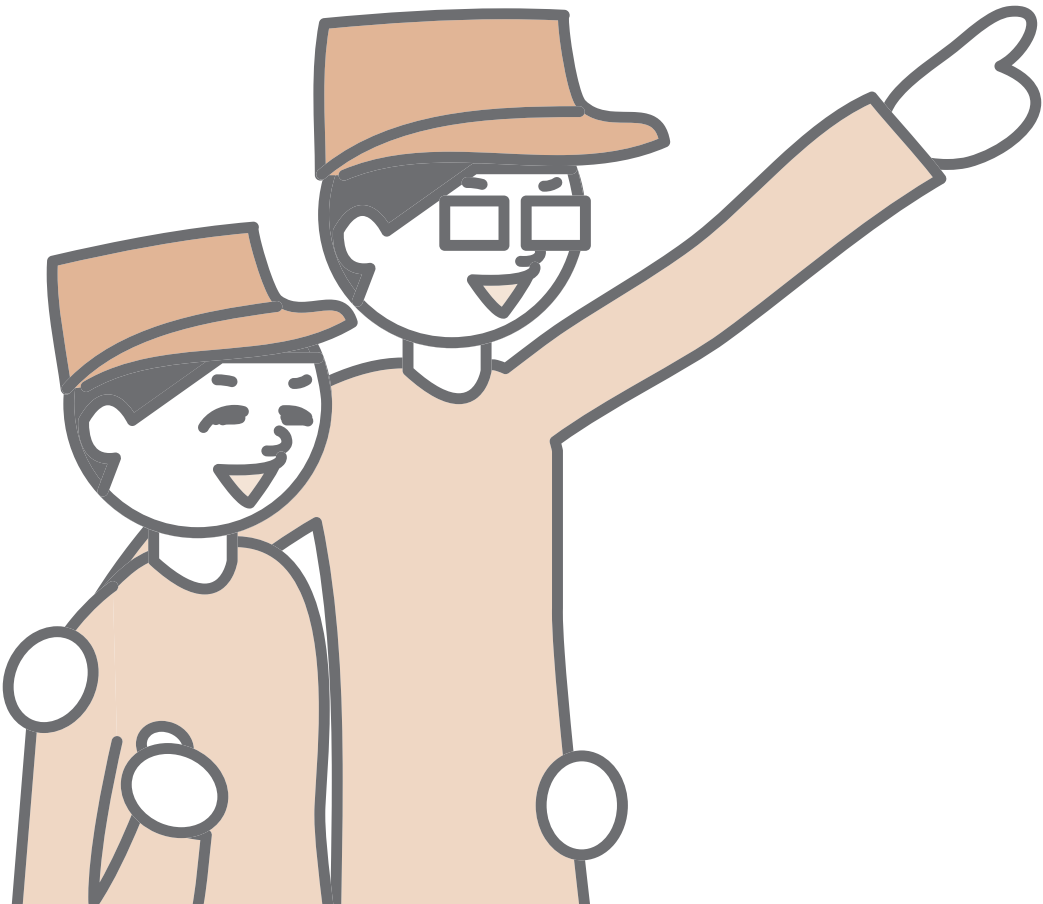
- Возможность гибкого включения знаний, опыта, новых данных.
- Возможность повышения точности путём постепенного обучения.

Отсеивание спама
Машинный перевод
Анализ изображений
Анализ журналов веб-доступа (маркетинг)

Байесовская статистика Глава 11

Планирование экспериментов – методология, касающаяся методов выбора порядка проведения эксперимента в пространстве и во времени, принятия решения о размере выборки, повышения эффективности экспериментов.
Байесовская статистика – статистика, дающая возможность гибкого включения знаний, опыта, новых данных, в основе которой лежит байесовский вывод.

Data is here



Глава 1

Описательная статистика

11 Разнообразные средние

Среднее – это величина, выражающая центральное значение данных.

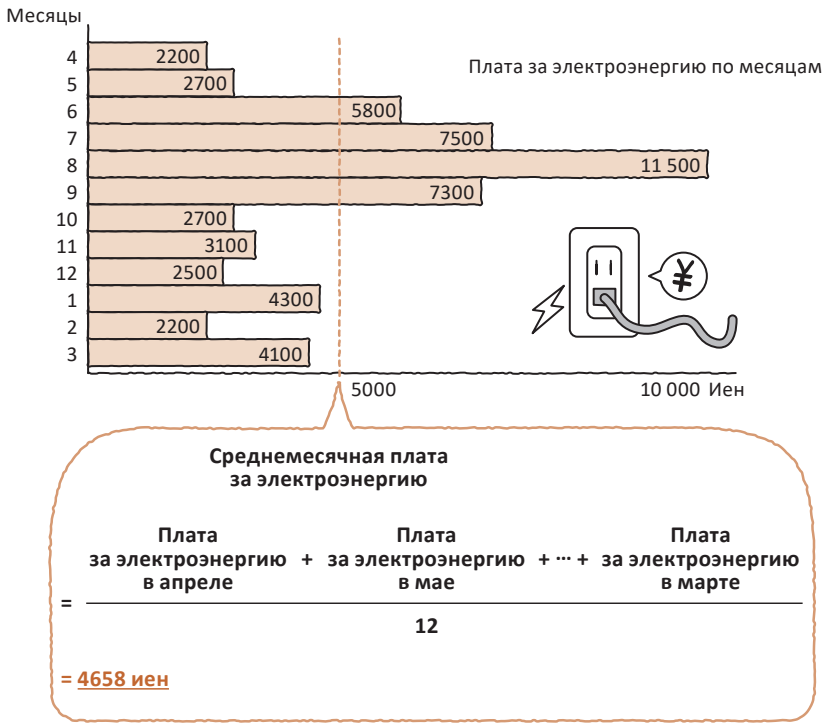
Среднее арифметическое

- Среднее арифметическое x вычисляется следующим образом (здесь x – переменная, n – количество данных):

Среднее арифметическое $\bar{x} = (x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n)/n$.

$\bar{x}$ – это (произносится так) «икс с чертой»

- Пусть имеются данные о плате за электроэнергию по месяцам. Если вы хотите выровнять их и узнать среднюю плату за электроэнергию в месяц, то используйте среднее арифметическое.



Среднее арифметическое – когда говорят «среднее», имеют в виду среднее арифметическое. Это сумма значений, поделённая на количество значений, на неё сильно влияют резко отклоняющиеся значения («выбросы»).

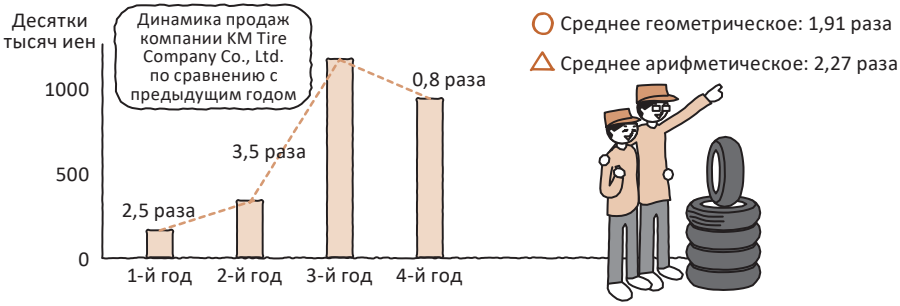
Среднее геометрическое

- Среднее геометрическое x вычисляется следующим образом.

$$\overline{x_G} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \dots \cdot x_{n-1} \cdot x_n}.$$

G означает «Geometric»
$\sqrt[n]{x}$ означает (читается как) корень n-ной степени из x

- Среднее геометрическое подходит для нахождения среднего, например ежегодного прироста какой-либо величины, относительного изменения величины по сравнению с аналогичным периодом прошлого года.



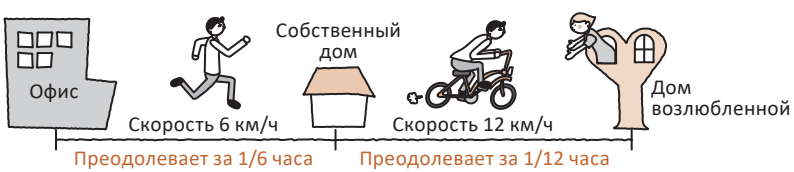
Среднее гармоническое

- Среднее гармоническое x вычисляется следующим образом:

$$\overline{x_H} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_{n-1}} + \frac{1}{x_n}}.$$

H означает «Harmonic»

- Среднее гармоническое используется для нахождения средней скорости при прохождении определённого расстояния.



Мужчина преодолевает 2 км за 1/6 + 1/12 = 1/4 часа, поэтому (средняя скорость) равна 2 ÷ 1/4 = 8 км/ч. Это соответствует гармоническому среднему скорости перемещения до собственного дома и скорости перемещения до дома возлюбленной: $\overline{x_H} = \frac{2}{\frac{1}{6} + \frac{1}{12}} = 8$ км/ч

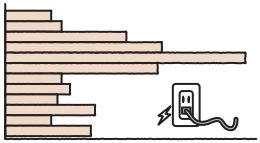
○ Среднее гармоническое: 8 км/ч
△ Среднее арифметическое: 9 км/ч

Среднее геометрическое – применяется, когда мы хотим вычислить среднее значение прироста, процентной ставки.
Среднее гармоническое – применяется для расчёта среднего значения, например скорости движения, электрического сопротивления. Выполняется следующее неравенство: Среднее арифметическое ⩾ Среднее геометрическое ⩾ Среднее гармоническое.

1 2

Разброс данных ①

Квантиль и дисперсия



Одни лишь средние не скажут нам, как разбросаны данные, для этого используются такие показатели, как максимальное значение, минимальное значение, квантиль, межквартильный размах, дисперсия (среднеквадратичное отклонение).

Квантили

- Если выстроить данные, состоящие из n значений, по возрастанию и разбить полученный ряд на k равных частей, то пограничные значения будут называться квантилями.
- Часто используются квартили (k = 4), которые в порядке возрастания значений называются первым (нижним), вторым квартилем и третьим (верхним) квартилями. Второй квартиль называют также медианой, так как он расположен посередине.

Межквартильный размах

- Это разница между 1-м и 3-м квартилями. Чем больше данных сосредоточено в окрестностях медианы, тем меньше будет межквартильный размах.

Отклонение

- Разность между значением и средним значением данных. Если много значений с большим отклонением (по абсолютной величине), можно говорить о том, что этот набор данных имеет большой разброс.

Отклонение (d_i) = Наблюдаемое значение (x_i) – Среднее значение ($\bar{x}$).

Дисперсия

- Отклонения вычисляются для каждого значения данных отдельно, а дисперсия объединяет их в один показатель. Она вычисляется по следующей формуле:

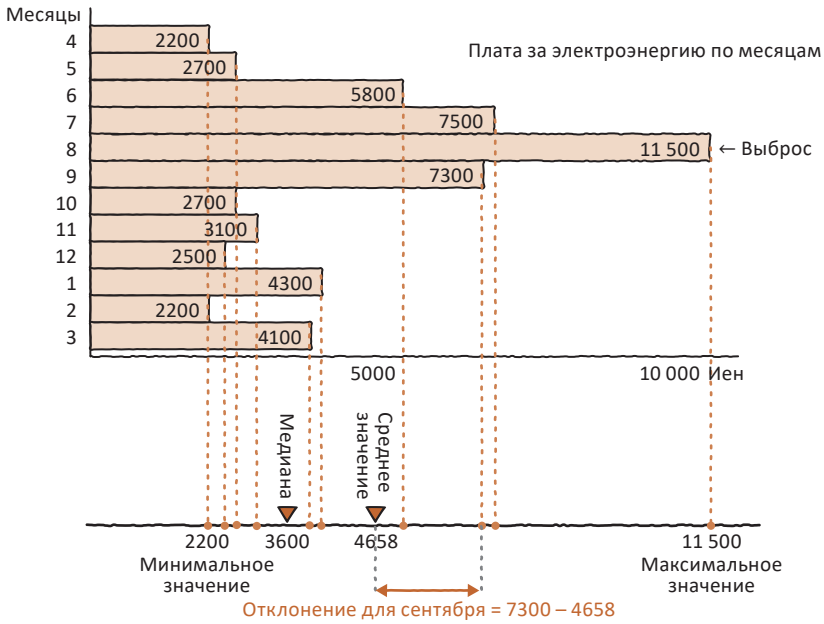
$$S^2 = \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\} \div n$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Квартили – это значения, расположенные на каждой из границ в том случае, когда мы поделим данные, выстроенные в порядке возрастания, на 4 равные части.
Медиана – значение, которое окажется в середине в том случае, когда мы выстроим данные в порядке возрастания. Она мало подвержена влиянию выбросов.

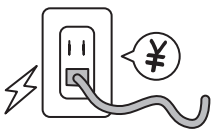
- Первый член в правой части выражения называется суммой квадратов отклонений, а положительный квадратный корень из дисперсии называется среднеквадратичным отклонением (s).

Выброс

- Значение, которое резко отклоняется от среднего значения данных, называется выбросом.



Среднее значение платы за электроэнергию – 4658 иен, максимальное – 11 500 иен
Интересно, в какой сезон мы платим больше?



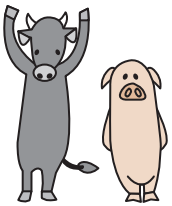
Какой разброс значений платы за электроэнергию?

Дисперсия – показатель того, насколько данные разбросаны вокруг среднего значения. Среднее значение квадратов отклонений.
Среднеквадратичное отклонение – положительный квадратный корень из дисперсии. Удобно тем, что его единица измерения получается такая же, как у данных.

13

Разброс данных ②

Коэффициент вариации

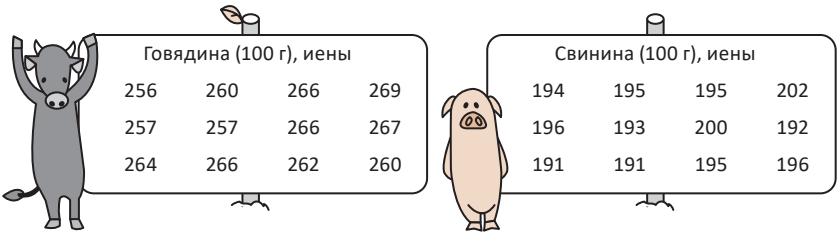


Коэффициент вариации

- Используется в тех случаях, когда требуется сравнить разбросы двух переменных.
- Коэффициент вариации вычисляется по следующей формуле:

Коэффициент вариации (CV) = Среднеквадратичное отклонение (S) ÷ Среднее ($\bar{x}$).

- Где больше вариация цен?



262,5 иены	Среднее арифметическое $\bar{x}$	195,0 иены
4,25 иены	Среднеквадратичное отклонение s	3,19 иены
0,016 иены	Коэффициент вариации CV	0,016 иены

- Хотя среднеквадратичное отклонение цен на говядину больше, чем на свинину, коэффициенты вариации одинаковы. Следовательно, и разбросы одинаковы.

Нахождение среднего и дисперсии с использованием таблицы частотности

Если данные представлены в виде таблицы частотности (см. ниже), можно найти приближённые значения среднего и дисперсии с помощью классификационных индексов (т. е. медиан интервалов).

Интервалы	Классификационные индексы	Частоты
255–259 иен	257 иен	3
260–264 иены	262 иены	4
265–269 иен	267 иен	5

Среднее = (классификационный индекс × частота) ÷ количество данных
= (257 × 3 + 262 × 4 + 267 × 5) ÷ 12 = 262,8
Дисперсия = Среднее квадратов разностей «классификационный индекс – среднее»
= ((257 – 262,8)^2 × 3 + (262 – 262,8)^2 × 4 + (267 – 262,8)^2 × 5) ÷ 12 = 15,97

Коэффициент вариации – значение, полученное делением среднеквадратичного отклонения на среднее значение. Используется для сравнения разбросов в двух группах, имеющих разные единицы измерения.

ЭТО ИНТЕРЕСНО

Легенды о великих ①

ПРИВЕТ, Я...

КАРЛ ПИРСОН

Karl Pearson (1857–1936)

Современная описательная статистика стала солидной наукой благодаря трудам Карла Пирсона: он ввел в статистику такие понятия, как среднеквадратичное отклонение, гистограммы и др. Карл Пирсон родился в 1857 году в Лондоне в семье адвоката, он был настолько слаб здоровьем, что не мог регулярно посещать школу. Несмотря на это, он поступил в университет, где погрузился в математику, а после окончания университета проходил обучение физике, римскому праву в Германии. Там заинтересовался литературой, юриспруденцией, социализмом. Как считается, своё имя Carl он исправил на Karl, потому что находился под большим влиянием знаменитого в то время экономиста Карла Маркса (Karl Marx). Хотя после возвращения на родину в следующем 1880 году он продолжил было изучать юриспруденцию, однако сразу же вернулся в мир математики и последовательно занимал должность профессора прикладной математики в нескольких университетах Лондона.

В мир статистики этого «прикладного математика» Пирсона увлёк его коллега по университету, зоолог Рафаэль Уэлдон, который под влиянием идей Фрэнсиса Гальтона пытался объяснить эволюцию живых организмов с позиций статистики. Уэлдон обратился за помощью к сильному в математике Пирсону. В процессе предпринятых вместе с Уэлдоном попыток найти подход к проблемам наследственности и эволюции с помощью статистических методов Пирсон изобрёл множество понятий, методов, без которых представить себе современную статистику невозможно. Эта его деятельность получила признание, а после смерти Гальтона в 1911 году Пирсон в качестве его преемника становится первым профессором факультета евгеники Университетского колледжа Лондона и основывает первый в мире факультет (прикладной) статистики.

Среди всех многочисленных заслуг Пирсона самой важной можно, наверное, назвать идею метода проверки гипотез с помощью распределения хи-квадрат. Для «критерия согласия», содержание которого почти полностью соответствует «критерию независимости» (глава 7 этой книги), в качестве меры измерения соответствия наблюдаемой и ожидаемой частот он независимо изобрёл статистику, подчиняющуюся распределению хи-квадрат. Правда, само по себе распределение хи-квадрат было уже к тому времени опубликовано германским геодезистом Фридрихом Робертом Гельмертом. Помимо того, он впервые разработал обстоятельные числовые таблицы для обобщенных данных, изобрёл метод оценки статистических параметров под названием метод моментов.

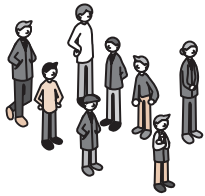
Далее начала стремительно развиваться статистика вывода трудами Фишера и сына Пирсона, Эгона, и тень самого Пирсона стала совсем бледной, однако в последнее время во всём мире усиливается тенденция к переоценке его вклада в науку. Так, например, повторно приковывается внимание к изданной в 1892 году книге «Грамматика науки» (The Grammar of Science). Эта книга является, так сказать, научно-философским трудом, в ней объясняется, что «если наука – это язык, то статистика – это грамматика науки о данных», и считается, что даже Эйнштейн и Нацумэ Сосэки[1] испытали её влияние. Книга в переводе на японский, к сожалению, уже не переиздаётся, однако её английскую версию можно бесплатно посмотреть в интернете.

[1] Нацумэ Сосэки – японский писатель, один из основоположников современной японской литературы. – *Прим. перев.*

14

Связь переменных ①

Коэффициент корреляции



Предполагаемая линейная связь между двумя переменными, например расходами на сбытовую рекламу и торговой выручкой, температурой воздуха и величиной урожая, временем на игры и успеваемостью вида «если одно увеличивается, то другое тоже увеличивается» или «если одно увеличивается, то другое уменьшается», называется корреляцией.

Коэффициент корреляции смешанных моментов Пирсона

- Показатель корреляции, принимающий значения от –1 до 1.
- Коэффициент корреляции переменных x и y вычисляется по следующей формуле:

$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}\sqrt{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}}$$

Потребители	Количество купленных яблок (x)	Количество купленных мандаринов (y)	$x - \bar{x}$	$y - \bar{y}$
1	1	2	–2,5	–0,5
2	2	1	–1,5	–1,5
3	5	4	1,5	1,5
4	6	3	2,5	0,5
Средние	3,5	2,5	0	0

$$r = \frac{(-2,5)(-0,5) + (-1,5)(-1,5) + (1,5)(1,5) + (2,5)(0,5)}{\sqrt{(-2,5)^2 + (-1,5)^2 + (1,5)^2 + (2,5)^2}\sqrt{(-0,5)^2 + (-1,5)^2 + (1,5)^2 + (0,5)^2}} = 0,76$$

- Если r близко к 1, то усиливается положительная корреляция («если одно увеличивается, то и другое увеличивается» или «если одно уменьшается, то и другое уменьшается»), а распределение точек на корреляционной диаграмме направлено вправо и вверх.
- Напротив, в случае близости к –1 усиливается отрицательная корреляция («если одно увеличивается, то другое уменьшается» или «если одно уменьшается, то другое увеличивается»), распределение точек на корреляционной диаграмме направлено вправо и вниз.
- Близость к 0 показывает отсутствие корреляции (некоррелированности), а распределение точек на корреляционной диаграмме напоминает круг.

Коэффициент корреляции – показатель силы взаимосвязи (корреляции) двух переменных. Чем он ближе к 1, тем сильнее положительная корреляция, а чем ближе к –1, тем сильнее отрицательная корреляция. В случае равенства 0 переменные некоррелированы.

ЭТО ИНТЕРЕСНО

Легенды о великих ②

ПРИВЕТ, Я...

Фрэнсис Гальтон

Francis Galton (1822–1911)

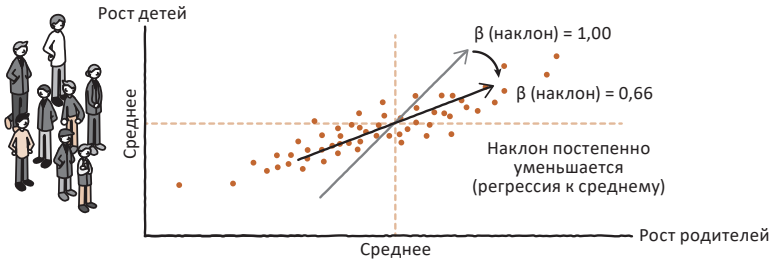
Как следует из его названия, коэффициент корреляции представил в виде формулы Карл Пирсон, однако это понятие изобрёл его учитель евгенист Фрэнсис Гальтон.

Гальтон, родившийся в 1822 году в семье богатого банкира Бирмингема, по воле отца с неохотой поступил на медицинский факультет, но в конце концов стал изучать математику в Кембриджском университете. Его отец умер, когда он заканчивал университет, и он, воспользовавшись представившимся случаем, отправился в экспедицию по Африке, где встретился с людьми разных рас, и, по-видимому, это и привело Гальтона на путь евгеники.

В 1875 году Гальтон в качестве одного из обоснований евгеники попытался доказать, что человеческий рост передаётся по наследству. Сначала он, используя душистый горошек (sweet pea), данные о котором было легко собрать, проверил, не передаётся ли вес семян от родителей к детям. Как он и предсказывал, растения душистого горошка, выросшие из тяжёлых семян, давали тоже тяжёлые семена, но он заметил ещё одно интересное явление: разброс веса у семян от растений-детей был меньше, чем у семян от растений-родителей. Гальтон предположил, что признаки всевозможных организмов не принимают крайние формы и благодаря этому явлению возникает возможность поддерживать существование видов. Другими словами, сила, действующая между поколениями, вызывает постепенный регресс к среднему (к предкам). Он назвал это явление «регрессией». В Англии он измерил рост большого числа родителей и их детей и убедился в том, что это явление имеет место и у человека (рис. ниже).

И в качестве меры, показывающей силу взаимосвязи роста родителей и детей, он ввёл коэффициент корреляции.

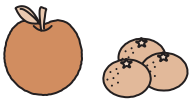
Гальтон был плодотворным и талантливым учёным. За свою жизнь он написал более 340 статей. Межквартильный размах, медиана – это тоже его изобретения. Стремясь предсказывать погоду, он пришёл к идеям, которые легли в основу множественного регрессионного анализа, он внёс вклад в создание методов расследования преступлений с использованием идентификации преступника по отпечаткам пальцев. В поздние годы жизни он по совету Флоренс Найтингейл, которая приходилась ему дальней родственницей, основал факультет статистики в университете, чем внёс огромный вклад в развитие современной статистики. Умер он в возрасте 89 лет.



15

Связь переменных ②

Ранговая корреляция



В том случае, если возможно использовать только ранговые данные, или если между двумя переменными предполагается криволинейная зависимость (корреляционная диаграмма имеет вид кривой), используется коэффициент ранговой корреляции.

Коэффициент ранговой корреляции Спирмена

- Коэффициент корреляции, рассчитанный для ранговых данных, – это коэффициент ранговой корреляции Спирмена.
- Если переменная непрерывна (переменная, принимающая значения из непрерывного диапазона), сначала мы выполняем ранговое преобразование.

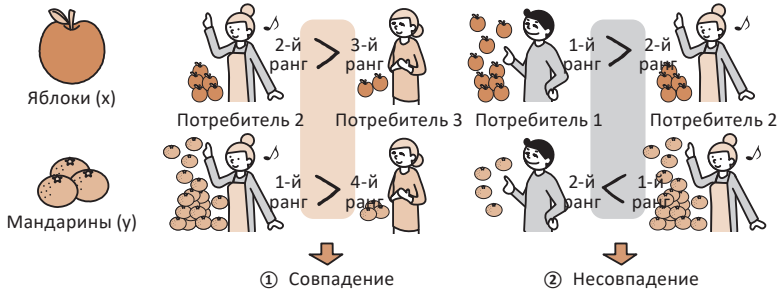
Потребители	Ранги x	Ранги y	$x - \bar{x}$	$y - \bar{y}$
1	1	2	–1,5	–0,5
2	2	1	–0,5	–1,5
3	3	4	0,5	1,5
4	4	3	1,5	0,5
Средние	2,5	2,5	0	0

Коэффициент ранговой корреляции Спирмена:

$$\rho = \frac{(-1,5)(-0,5)+(-0,5)(-1,5)+(0,5)(1,5)+(1,5)(0,5)}{\sqrt{(-1,5)^2+(-0,5)^2+(0,5)^2+(1,5)^2}\sqrt{(-0,5)^2+(-1,5)^2+(1,5)^2+(0,5)^2}} = 0,60.$$

Коэффициент ранговой корреляции Кендалла

- Показатель, применяемый для измерения корреляции, при использовании которого внимание обращено на совпадение рангов по x и по y.
- Если для ранговых данных (x_1, y_1) потребителя 1 и ранговых данных потребителя 2 (x_2, y_2) выполняется:



Коэффициент ранговой корреляции – показатель степени корреляции двух ранговых переменных. Существуют коэффициент ранговой корреляции Спирмена и коэффициент ранговой корреляции Кендалла. Чёткого критерия, какой из этих методов использовать, нет.

① $x_1 < x_2$ и $y_1 < y_2$ или $x_1 > x_2$ и $y_1 > y_2$ → делается вывод: ранги совпадают;
② $x_1 < x_2$ и $y_1 > y_2$ или $x_1 > x_2$ и $y_1 < y_2$ → делается вывод: ранги не совпадают.

- Изучаем данные трёх потребителей, и в случае, если мы видим «совпадение рангов», назначаем им ○, а в случае, если мы видим «несовпадение рангов», назначаем им ×.

Потребители	Ранги x	Ранги y	Потребитель 1	Потребитель 2	Потребитель 3
1	1	2			
2	2	1	×		
3	3	4	○	○	
4	4	3	○	○	×

Потребители	Потребитель 1	Потребитель 2	Потребитель 3	Потребитель 4
Количество ○	2	2	0	4
Количество ×	1	0	1	2

- Если A = количеству ○, B = количеству ×, n = количеству пар данных (в этом примере – 4), то коэффициент ранговой корреляции Кендалла находится по следующей формуле (для случая, если имеются одинаковые ранги, формула будет другой):

Коэффициент ранговой корреляции Кендалла

$$\tau = \frac{(A - B)}{\text{Число комбинаций при извлечении двух штук из четырех}} = \frac{4 - 2}{\frac{1}{2} \cdot 4 \cdot (4 - 1)} = 0{,}33.$$

О числе сочетаний

- Число сочетаний при извлечении двух элементов из A, B, C, D равно 6: (AB), (AC), (AD), (BC), (BD), (CD).
- Число сочетаний при извлечении двух элементов из A, B, C, D, E равно 10: (AB), (AC), (AD), (AE), (BC), (BD), (BE), (CD), (CE), (DE).
- В общем случае, при извлечении 2 элементов из n, число сочетаний находится по формуле ½ n (n – 1).
  При извлечении x элементов из n элементов число сочетаний находится по формуле $C_n^x = \frac{n!}{x!(n-x)!}$ (x! читается как «x факториал» и вычисляется как x! = x×(x – 1)×…×2×1).

Сочетания – метод извлечения x элементов из n отличных друг от друга элементов.