Зачем знать распределение случайной величины?

Тестирование данных на нормальность часто является первым этапом их анализа, так как большое количество статистических методов исходит из предположения нормальности распределения изучаемых данных.

Например, пусть необходимо проверить гипотезу о равенстве средних значений в двух независимых выборках. Для этой цели подходит критерий Стьюдента. Но применение критерия Стьюдента обосновано, только если данные подчиняются нормальному распределению. Поэтому перед применением критерия необходимо проверить гипотезу о нормальности исходных данных. Или проверка остатков линейной регрессии на нормальность – позволяет проверить, соответствует ли применяемая модель регрессии исходным данным. Нормальное распределение естественным образом возникает практически везде, где речь идёт об измерении с ошибками.

Проверку выборки на нормальность можно производить несколькими путями. Можно, например, построить гистограмму или можно воспользоваться критериями нормальности. Если гистограмма имеет колоколообразный симметричный вид, можно сделать заключение о том, что анализируемая переменная имеет примерно нормальное распределение. Однако при интерпретации гистограмм следует соблюдать осторожность, поскольку их внешний вид может сильно зависеть как от числа наблюдений, так и от шага, выбранного для разбиения данных на классы.

Критерии нормальности – это группа статистических критериев, предназначенных для проверки нормальности распределения. Существует целый ряд статистических критериев, специально разработанных для проверки нормальности распределения данных. В общем виде проверяемую при помощи этих тестов нулевую гипотезу можно сформулировать так: «Анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение». Если получаемая при помощи того или иного теста вероятность ошибки p оказывается меньше некоторого заранее принятого уровня значимости (например, 0.05), нулевая гипотеза отклоняется. К таким критериям относят критерии Шапиро-Уилка, Андерсона-Дарлинга, Хи-квадрат и др. Подробное изучение критериев выходит за рамки данного курса. Однако, рекомендуем вам самостоятельно углубиться в эту тему. Критерии

проверки нормальности реализованы в библиотеке SciPy в модуле stats. Например, для применения критерия Шапиро-Уилка, следует использовать программный код:

```
from scipy.stats import shapiro
data1 = # your data
stat, p = shapiro(data)
```

Проверка статистических гипотез

Для чего нужны все эти сведения из статистики и теории вероятностей? Наука о данных часто предусматривает формулировку и проверку статистических гипотез о данных и процессах, которые их порождают.

Статистическая гипотеза – это некоторое предположение о свойствах и характеристиках исследуемых генеральных совокупностей. Для наших целей под статистическими гипотезами будем понимать утверждения типа «исследователи данных больше предпочитают Python, чем R». Классическая трактовка подразумевает наличие главной, или нулевой, гипотезы H_0, которой исследователь придерживается изначально, и альтернативной гипотезы H_1, относительно которой нужно ее сопоставить. Для того чтобы принять решение, можно ли отклонить H_0 как ложную или принять ее как истинную, используют специальные статистические критерии – это специальные случайные величины, которые принимают различные действительные значения. В разных задачах критерии разные.

Ошибка первого рода состоит в том, что гипотеза H_0 будет отвергнута, хотя на самом деле она правильная. Вероятность допустить такую ошибку называют *уровнем значимости* (p-value) и обозначают буквой α («альфа») или p. Мы будем обозначать как α, чтобы не путать с вероятностью.

Ошибка второго рода состоит в том, что гипотеза H_0 будет принята, но на самом деле она неправильная. Вероятность совершить эту ошибку обозначают буквой β («бета»). Значение 1-β называют *мощностью критерия* – это вероятность отвержения неправильной гипотезы.

Уровень значимости задается исследователем самостоятельно, наиболее часто выбирают значения 0.1, 0.05, 0.01. И тут возникает мысль, что чем меньше «альфа», тем вроде бы лучше. Но это только вроде: при уменьшении вероятности α – *отвергнуть правильную гипотезу* растет вероятность β – *принять неверную гипотезу* (при прочих равных условиях). Поэтому перед исследователем стоит задача грамотно

подобрать соотношение вероятностей α и β, при этом учитывается тяжесть последствий, которые повлекут за собой та и другая ошибки.

Выделяют 5 шагов при проверке гипотез:

- Определение нулевой (H_0) и альтернативной гипотезы (H_1) при исследовании. Определение уровня значимости критерия.
- Отбор необходимых данных из выборки.
- Вычисление значения статистического критерия K, отвечающего H_0.
- Вычисление критической области, проверка статистики критерия на предмет попадания в критическую область. Критической областью называется область значений критерия, при которых отвергается H_0. А критические значения – это граница критической области.
- Интерпретация достигнутого уровня значимости α и результатов.

t-критерий Стьюдента. t-критерий Стьюдента используется для определения статистической значимости различий средних величин. Может применяться как в случаях сравнения независимых выборок (например, группы больных гриппом и группы здоровых), так и при сравнении связанных совокупностей (например, средняя частота пульса у одних и тех же пациентов до и после приема лекарства).

Для применения данного критерия необходимо, чтобы исходные данные имели нормальное распределение. В случае применения двухвыборочного критерия для независимых выборок также необходимо соблюдение условия равенства дисперсий.

Одновыборочный t-критерий применяется для проверки нулевой гипотезы о равенстве математического ожидания (среднего значения) случайной величины некоторому известному значению m.

Пример. Мы хотим узнать, отличается ли средняя масса землероек в Заповеднике от массы землероек = 90 г.

В Python эту задачу можно решить с помощью scipy.stats.ttest_1samp. В качестве значений параметров мы передаем выборку и значение m (параметр popmean):

```
from scipy import stats
stats.ttest_1samp(data, popmean=5.0)
```

Метод возвращает t-статистику и уровень значимости (p-value), по которому можно сделать вывод отвергать гипотезу или нет.

Двухвыборочный t-критерий. Пусть имеются две независимые выборки нормально распределенных случайных величин. Необходимо проверить по выборочным данным нулевую гипотезу равенства математических ожиданий (средних значений) этих случайных величин.

Пример. Различаются ли по массе тигры-самцы и тигры-самки в зоопарке?

Для решения подобных задач с помощью Python используют scipy.stats.ttest_ind:

```
from scipy import stats
stats.ttest_ind(data_1, data_2)
```

Метод возвращает t-статистику и уровень значимости (*p*-value), по которому можно сделать вывод отвергать гипотезу или нет.

Описательная статистика

На сегодняшний день статистический анализ данных является важной частью науки о данных. Статистика – это набор математических методов и инструментов, позволяющих преобразовывать необработанные данные в значимую информацию, которую легко интерпретировать и обобщать.

Данные – это просто «сырые» наблюдения. Чтобы трансформировать эти наблюдения в имеющие смысл идеи, применяется описательная статистика.

Описательная статистика – это раздел статистической науки, в рамках которого изучаются методы описания и представления основных свойств данных. Описательную статистику можно применять к одному или нескольким наборам данных. Когда описывают и вычисляют характеристики одной переменной, то выполняют одномерный анализ. Когда исследуют статистические связи между несколькими переменными, то выполняют многомерный анализ данных.

Описательная статистика использует три основных метода агрегирования данных:

- Табличное представление
- Графическое изображение
- Расчет статистических показателей

Меры центральной тенденции. Мера центральной тенденции в статистике – число, служащее для описания множества значений одним-единственным числом (для краткости). Например, вместо перечисления величин зарплат всех сотрудников

организации говорят о средней зарплате. Существует множество мер центральной тенденции, мы рассмотрим наиболее популярные.

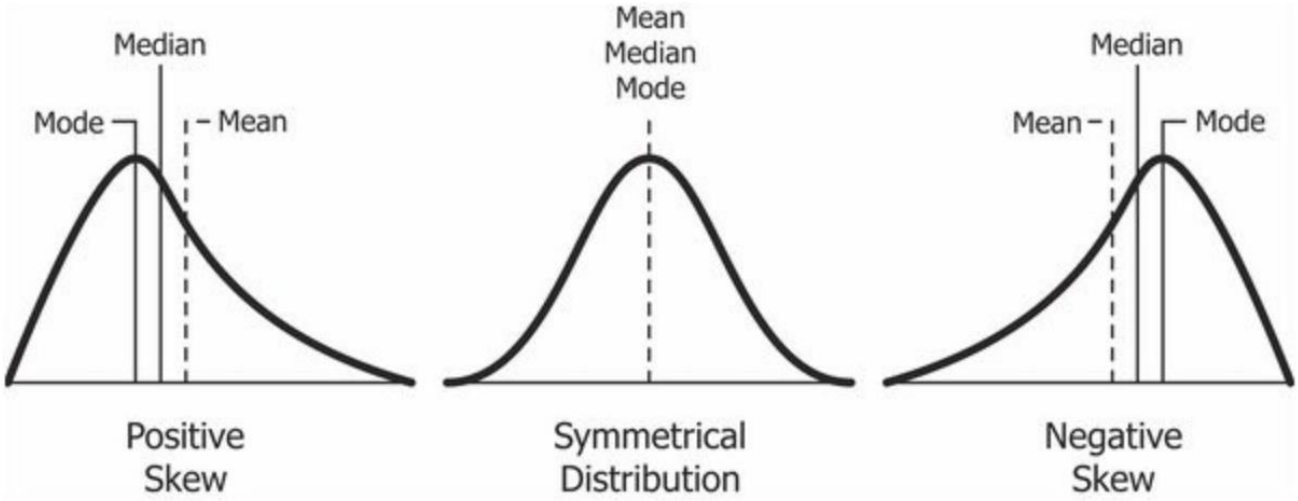
Чаще всего используются среднее (или среднее арифметическое) значение, медиана и мода.

Среднее значение берется как сумма значений, деленная на их количество.

Медиана – значение, которое делит упорядоченные по возрастанию (убыванию) наблюдения пополам. Медиана является ближайшим к центру значением (если число точек данных нечетное) либо средним арифметическим, взятым как полусумма двух ближайших к середине значений (если число точек четное).

Обратите внимание, что медиана – в отличие от среднего – не зависит от каждого значения в наборе данных. Например, если сделать наибольшую точку еще больше (или наименьшую точку еще меньше), то среднее значение изменится в большую (или меньшую) сторону, а срединные точки останутся неизменными, следовательно, и медиана тоже не изменится. Таким образом, среднее значение очень чувствительно к выбросам (экстремальным значения в данных, которые находятся далеко за пределами других наблюдений) в данных. Так как выбросы являются, скорее всего, плохими данными (или иначе - нерепрезентативными для ситуации, которую мы пытаемся понять), то среднее может иногда давать искажение.

Мода – это наиболее часто встречающееся значение.



Вариация. Вариация служит мерой разброса данных. Как правило, меры вариации – это статистические показатели, у которых значения, близкие к нулю, означают полное отсутствие разброса, а большие значения (что бы это ни означало) – очень большой разброс. Например, самым простым показателем является размах,

который определяется как разница между максимальным и минимальным значениями данных.

Размах равен нулю, когда максимально и минимальное значения эквивалентны, что происходит только тогда, когда все элементы равны между собой, и значит, разбросанность в данных отсутствует. И наоборот, когда размах широкий, то максимум намного больше минимума, и разбросанность в данных высокая. Как и медиана, размах не особо зависит от всего набора данных. Набор данных, все точки которого равны 0 или 100, имеет тот же размах, что и набор данных, чьи значения представлены числами 0, 100 и большого количества чисел 50, хотя кажется, что первый набор должен быть разбросан больше.

Более точным показателем вариации является дисперсия. Простыми словами дисперсия – это средний квадрат отклонений. То есть вначале рассчитывается среднее значение, затем берется разница между каждым исходным и средним значением, возводится в квадрат, складывается и затем делится на количество значений в данной выборке. Формулы вычисления дисперсии:

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$$, где

x – отдельные значения,

$\bar{x}$– среднее арифметическое по выборке,

n – общее количество наблюдений.

Дисперсия измеряется в единицах, которые представляют собой квадрат исходных единиц. Поскольку такие единицы измерения трудно интерпретировать, то вместо дисперсии используют стандартное отклонение (также его называют средним квадратическим отклонением), которое вычисляется как корень из дисперсии.

Размах и стандартное отклонение имеют ту же проблему с выбросами, что и среднее. Более надежной альтернативой является вычисление интерквартильного размаха. Прежде чем мы дадим определение этому понятию, необходимо определить, что такое квартиль.

Как мы уже определили ранее, медиана – это значение, которое находится в выборке ровно посередине. Но, кроме медианы, существует еще много чисел, которые

делят ряд наблюдений на равное число частей. Фактически можно на любое количество равных частей разделить выборку и получить какую-то свою меру.

Квантиль – это значение, ниже которого лежит определенное количество наблюдений, соответствующих выбранной частоте. С понятием квантиля плотно связано понятия процентиль. Процентиль показывает процент наблюдений, лежащих ниже выбранного значения. Квартили распределения – это квантили, кратные 25%, то есть соответствующие 25%, 50% и 75%. Их еще иногда называют соответственно «первый», «второй» и «третий» либо «нижний», «средний» и «верхний».

- 0.25-квантиль называется первым (или нижним) квартилем, т.е. меньше этого значения 25% значений в выборке.
- 0.5-квантиль называется медианой или вторым квартилем, т.е. меньше и больше медианы 50% наблюдений.
- 0.75-квантиль называется третьим (или верхним) квартилем, т.е. меньше этого значения 75% значений в выборке и больше 25% значений.

Интерквартильным размахом – называется разность между третьим и первым квартилями. Интерквартильный размах является характеристикой разброса распределения величины и может применяться как аналог дисперсии. Медиана и интерквартильный размах могут быть использованы вместо математического ожидания и дисперсии в случае, если данные содержат выбросы.

Теперь рассмотрим, как можно вычислить изученные показатели с помощью языка программирования Python. Библиотека pandas содержит метод .describe(), который позволяет вывести сразу несколько статистик.

```
import pandas as pd
df = pd.read_excel('mobile.xlsx')
df
```

	Количество SMS за месяц	Количество звонков	Среднемесячный расход
0	56	82	121.54
1	1	221	287.51
2	36	68	113.70
3	23	96	410.23
4	29	139	537.60
...	...	...	...
4487	30	66	186.20
4488	23	112	500.68
4489	5	189	470.42
4490	69	124	858.99
4491	24	136	151.92

```
df.describe()
```

	Количество SMS за месяц	Количество звонков	Среднемесячный расход
count	4492.000000	4492.000000	4492.000000
mean	21.243321	140.480632	506.155512
std	27.911864	91.742992	646.252023
min	0.000000	2.000000	3.180000
25%	3.000000	94.000000	152.880000
50%	6.000000	129.000000	316.960000
75%	32.000000	168.000000	600.032500
max	179.000000	635.000000	5142.760000

Метод .describe() возвращает для каждой переменной следующие характеристики:

- Количество наблюдений (count),
- Среднее значение (mean),
- Стандартное отклонение (std),
- Минимальное значение (min),
- 25-й процентиль (25%) – значение, меньше которого 25% наблюдений,
- Медиана набора чисел (50-й процентиль, 50%),
- 75-й процентиль (75%) – значение, больше которого 25% наблюдений,
- Максимальное значение (max).

Для категориальных переменных .describe() возвращает другие статистики:

- top –наиболее распространенное значение,
- unique – количество уникальных значений,
- freq – частота.

Библиотека pandas содержит методы, предназначенные для вычисления отдельных статистик. Некоторые из методов приведены в таблице.

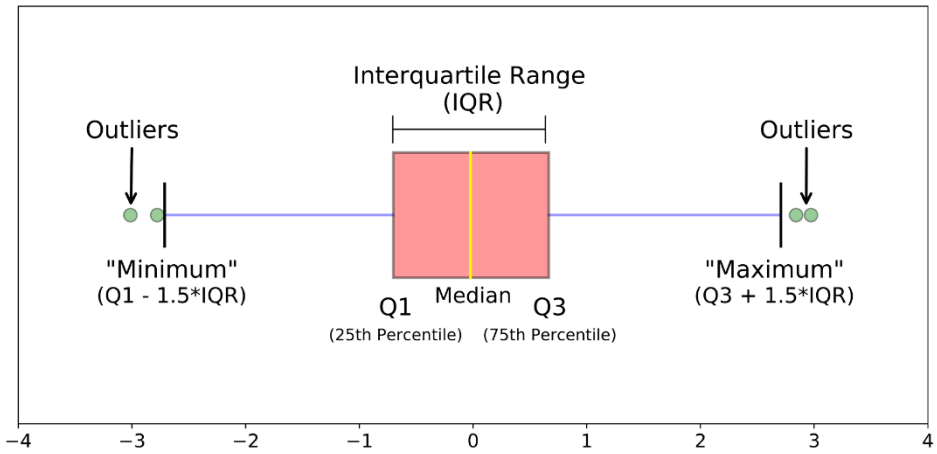
Метод	**Описание**
.count()	Количество значений, исключая отсутствующие
.sum()	Сумма значений
.mean()	Среднее значение
.median()	Медиана
min()	Минимальное значение
.max()	Максимальное значение
.quantile()	Выборочный квантиль в диапазоне от 0 до 1
.std()	Стандартное отклонение
.var()	Дисперсия

Диаграмма размаха. Диаграммы размаха («ящик с усами») (Box and Whisker Plot или Box Plot) – это удобный способ визуального представления групп числовых данных через квартили. Диаграммы размаха используются в описательной статистике и позволяют быстро исследовать один или более наборов данных в графическом виде.

Диаграмма размаха строится следующим образом:

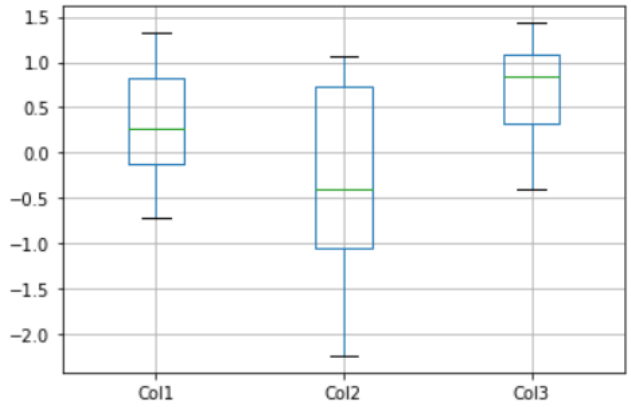
- Ящик отображает диапазон данных, находящийся между первым и третьим квартилем, а медиана делит этот ящик на две части.

- Прямые линии, исходящие из ящика, называются «усами» и используются для обозначения степени разброса (дисперсии) за пределами верхнего и нижнего квартилей.
- Выбросы иногда отображаются в виде отдельных точек, находящихся на одной линии с усами и за пределами их границ.
- Диаграммы размаха могут располагаться как горизонтально, так и вертикально.



Построить диаграмму размаха в Python можно несколькими способами. Например, можно использовать метод .boxplot() из pandas:

```
np.random.seed(1234)
df = pd.DataFrame(np.random.randn(10, 4),
                  columns=['Col1', 'Col2', 'Col3', 'Col4'])
boxplot = df.boxplot(column=['Col1', 'Col2', 'Col3'])
```



Также можно воспользоваться методами из библиотек matplotlib(https://matplotlib.org/stable/gallery/pyplots/boxplot_demo_pyplot.html) или seaborn(https://seaborn.pydata.org/generated/seaborn.boxplot.html).

Корреляционный анализ

Корреляционный анализ – метод изучения взаимосвязи между двумя и более случайными величинами.

Между переменными (случайными величинами) может существовать функциональная связь, т.е. одна из переменных может быть определена как функция от другой (например, $y = kx$). Но между переменными может существовать и связь другого рода, проявляющаяся в том, что одна из них реагирует на изменение другой изменением своего закона распределения. Такую связь называют *стохастической*. Она появляется в том случае, когда имеются общие случайные факторы, влияющие на обе переменные. В качестве меры зависимости между переменными используется коэффициент корреляции.

Определим понятие ковариации. Ковариацию называют парным аналогом дисперсии.

$$cov(x;y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{\text{ср}})(y_i - y_{\text{ср}})$$

В отличие от дисперсии, которая измеряет отклонение одной единственной переменной от ее среднего, ковариация измеряет отклонение двух переменных в тандеме от своих средних.

Коэффициентом корреляции между случайными величинами ξ и η называют число

$$\rho(\xi,\eta) = \frac{\mathbf{cov}(\xi,\eta)}{\sqrt{\mathbf{D}\xi\mathbf{D}\eta}},$$

где $\mathrm{cov}(\xi,\eta)$ – ковариация между ξ и η, $\mathrm{D}\xi$ и $\mathrm{D}\eta$ – дисперсии ξ и η.

Коэффициент корреляции изменяется в пределах от –1 до +1. Если коэффициент корреляции отрицательный, это означает, что с увеличением значений одной переменной значения другой убывают. Если переменные независимы, то коэффициент корреляции равен 0 (обратное утверждение верно только для переменных, имеющих нормальное распределение). Но если коэффициент корреляции не равен 0 (переменные называются некоррелированными), то это значит, что между переменными существует зависимость. Крайние значения +1 и -1 коэффициента корреляции соответствуют наличию линейной зависимости, которая является самой сильной из всех возможных форм зависимости.

Корреляционный анализ позволяет установить силу и направление стохастической взаимосвязи между переменными. Если переменные измеряются в количественной шкале и имеют нормальное распределение, то корреляционный анализ осуществляется посредством вычисления коэффициента корреляции Пирсона. Для оценки силы и направления связи между переменными, измеренными в порядковой шкале, используются непараметрические ранговые коэффициенты корреляции: коэффициент ранговой корреляции Кендалла и коэффициент корреляции Спирмена.

Корреляционный анализ используется в экономике, социологии и психологии, медицине, управления качеством, биометрии и других сферах. Популярность корреляционного анализа объясняется тем, что коэффициенты корреляции относительно просты в расчете, и их применение не требует специальной математической подготовки. С другой стороны – коэффициенты корреляции легко интерпретировать.

Значительная корреляция между двумя случайными величинами всегда является свидетельством существования некоторой статистической связи в данной выборке, но эта связь не обязательно должна наблюдаться для другой выборки и иметь причинно-следственный характер. Часто заманчивая простота корреляционного исследования подталкивает исследователя делать ложные интуитивные выводы о наличии причинно-следственной связи между парами признаков, в то время как коэффициенты корреляции устанавливают лишь статистические взаимосвязи. Например, рассматривая пожары в конкретном городе, можно выявить весьма высокую корреляцию между ущербом, который нанес пожар, и количеством пожарных, участвовавших в ликвидации пожара, причем эта корреляция будет положительной. Из этого, однако, не следует вывод «увеличение количества пожарных приводит к увеличению причиненного ущерба», и тем более не будет успешной попытка минимизировать ущерб от пожаров путем ликвидации пожарных бригад. Корреляция двух величин может свидетельствовать о существовании общей причины, хотя сами явления напрямую не взаимодействуют. Например, обледенение становится причиной как роста травматизма из-за падений, так и увеличения аварийности среди автотранспорта. В этом случае две величины (травматизм из-за падений пешеходов и

аварийность автотранспорта) будут коррелировать, хотя они не связаны причинно-следственно друг с другом, а лишь имеют стороннюю общую причину – гололедицу.

Как правило, данные для исследования удобнее всего обрабатывать и анализировать представляя их в виде DataFrame. Поэтому, мы рассмотрим, как вычислить коэффициенты корреляции с помощью pandas. В pandas есть удобный метод .corr(). Метод имеет всего 2 параметра:

- method, который может принимать значения ‘pearson’ (корреляция Пирсона), ‘kendall’ (корреляция Кенделла), ‘spearman’ (корреляция Спирмена),
- min_periods – минимальное количество наблюдений, необходимых для каждой пары столбцов (задавать не обязательно).

Парадокс Симпсона. Нередким явлением при анализе данных является парадокс Симпсона (или парадокс объединения) – статистический эффект, когда при наличии двух групп данных, в каждой из которых наблюдается одинаково направленная зависимость, при объединении этих групп направление зависимости меняется на противоположное.

Например, представим, что вы можете отождествить всех членов сети как исследователей данных с Восточного побережья либо как исследователей данных с Западного побережья. Вы решаете выяснить, с какого побережья исследователи данных дружелюбнее.

Побережье	Количество пользователей	Среднее число друзей
Западное	101	8.2
Восточное	103	6.5

Очевидно, что исследователи данных с Западного побережья дружелюбнее, чем исследователи данных с Восточного побережья. Ваши коллеги выдвигают разного рода предположения, почему так происходит: может быть, дело в солнце, или в кофе, или в органических продуктах, или в безмятежной атмосфере Тихого океана? Продолжая изучать данные, вы обнаруживаете что-то очень необычное. Если учитывать только членов с ученой степенью, то у исследователей с Восточного побережья друзей в среднем больше, а если рассматривать только членов без ученой степени, то у исследователей с Восточного побережья друзей снова оказывается в среднем больше.

Побережье	Степень	Количество пользователей	Среднее число друзей
Западное	Есть	35	3.1
Восточное	Есть	70	3.2
Западное	Нет	66	10.9
Восточное	Нет	33	13.4

Как только вы начинаете учитывать ученые степени, корреляция движется в обратную сторону. Группировка данных по признаку «восток- запад» скрыла тот факт, что среди исследователей с Восточного побережья имеется сильная асимметрия в сторону ученых степеней. Подобный феномен возникает в реальном мире с определенной регулярностью. Главным моментом является то, что корреляция измеряет связь между двумя переменными при прочих равных условиях. Если данным назначать классы случайным образом, как и должно быть в хорошо поставленном эксперименте, то допущение «при прочих равных условиях» может быть и не плохим. Единственный реальный способ избежать таких неприятностей – это знать свои данные и обеспечивать проверку всех возможных спутывающих факторов. Очевидно, это не всегда возможно. Если бы у вас не было данных об образовании 200 исследователей данных, то вы бы просто решили, что есть нечто, присущее людям с Западного побережья, делающее их общительнее.