Теория 1. Случайные величины – интро

О спикере

Всем привет!
Меня зовут Лидия Храмова и я уже шесть лет люблю и практикую различные методы data science и машинного обучения - от марковских моделей до natural language processing. У меня есть опыт в построении аналитических процессов и production решений в разных отраслях - от стратегии до логистики и финтеха, и сегодня мне бы хотелось поделиться с вами ключевыми навыками анализа для начала вашего пути в data science.

Немного о моем проектном опыте и о том, как от excel дойти до нейронок:
https://youtu.be/tUBrP2o4Jpo?t=8479

История

Не секрет, что для data science – это не только программирование, но в первую очередь особые навыки анализа данных. И если в публикациях рядом со словами data science и машинное обучение упоминаются сложные архитектуры нейронных сетей, то в реальности чем изящнее и легче ваше решение – тем лучше.
И для того, чтобы научиться самостоятельно находить такие решения в море данных вам понадобятся инструменты математической статистики и теории вероятностей.
Если готовы попробовать – давайте начнем!

Итак, несмотря на то что про машинное обучение начали активно говорить сравнительно недавно, статистический анализ данных зародился еще в 18 веке, а если говорить точнее - в 1794 году. Именно в этот год немецкий математик и астроном Карл Гаусс (да, наверняка кто-то из вас слышал про Гауссово или нормальное распределение) уже активно работал над фундаментальными парадигмами статистики, а сам термин «статистика» стал названием новой научной дисциплины.

В это же время активно развивается теория вероятности и появляются первые практические задачи – в основном связанные с анализом ошибок в геодезических и астрономических наблюдений. Однако, уже в 19 веке работы русских математиков Чебышева, Маркова и Колмогорова существенно расширяют границы применимости математической статистики – для задач планирования производства и управления системами массового обслуживания, лингвистики.

А уже в 1936 году благодаря работам сэра Рональда Фишера появляется прообраз современного распознавания образов – тогда известный под менее модным названием «Дискриминантный анализ».
Так что с полной уверенностью можно сказать, что математическая статистика – наука старая и проверенная временем. И больше всего с тех пор добавилось не столько подходов, сколько доступных данных и инструментов для быстрых вычислений.

Понятие вероятности

Так о чем же учит нас математическая статистика? Базовым понятием ее является **вероятность**. Изначально понятие вероятности появилось в связи с существованием азартных игр и попытками подсчитать шансы на выигрыш при игре в кости. Однако в мире существует гораздо больше случаев, где мы сталкиваемся с этим термином – от вопросов о поле будущего ребенка до предсказания будущего дохода от портфеля инвестиций, то есть анализом некоего **случайного эксперимента** с его возможными **событиями** (исходами).

Что же такое вероятность? Существует несколько ее определений.

- **Классическое**

Это упрощенное определение, в этой трактовке вероятностью случайного события A называется отношение числа n несовместных (то есть исключающих друг друга) равновероятных элементарных событий, составляющих событие A, к числу всех возможных элементарных событий N:

$$P(A) = \frac{n}{N}$$

Вспоминая пример про кости, можем посчитать классическую вероятность выпадения шестерки на шестигранном игральном кубике. Подставляя в первую формулу, получим:

$$P(A) = \frac{1}{6}$$

- **Аксиоматическое определение**

Несмотря на удобство классического определения, в реальном мире даже вероятности выпадения той или иной грани на кубике могут отличаться – в следствие неровности кубика, поверхности для броска и других неизвестных нам факторов.

Поэтому в современной математической статистике и анализе данных используется аксиоматический подход, предложенный русским математиком Колмогоровым.

Он предельно просто и логичен – вместо того, чтобы искать некий теоретический закон распределения исходов случайного события, мы начинаем фиксировать реализации этой случайной величины в реальном мире и базировать наши решения уже на этой реальной выборке. Возвращаясь к случаю с игральной костью, мы просто-напросто начнем подкидывать наш кубик и фиксировать номер броска и выпавшее на грани число, набирая таким образом выборку для дальнейшего анализа. И именно такие выборки мы с вами будем изучать дальше.

Распределения случайной величины

Перед тем как приступить к оценке выборок от реализации случайных величин, поговорим о том, какими они бывают.

По характеру всех случайных величин их распределения можно разделить на два типа:

1. **Дискретные** – то есть принимающие ограниченное или счетное число значений.

В качестве примера дискретных величин можно привести, например, такие явления как: значения изменяющейся температуры тела у человека, количество бракованных изделий на производственной линии, количество необработанных звонков в колл-центре, результаты броска игрального кубика или голов в футбольном матче.

Семейство дискретных случайных величин широко применяются в теории массового обслуживания для анализа и моделирования очередей в фастфуд кафе, отделениях банка,

колл центрах, и включает в себя такие распределения как Пуассон, Бернулли, Биномиальное и многие другие. Тем, кому интересна работа с системами массового обслуживания, может пригодиться следующая книга, посвященная теории случайных процессов[1]

2. **Непрерывные** – или принимающие бесконечное число возможных значений на интервале. Примером может служить курс доллара или стоимость финансовых инструментов, убытки при наступлении страхового случая и тд. Семейство непрерывных случайных величин огромно и включает в себя такие законы как Нормальное или Гауссово распределение, Экспоненциальное распределение, Бета распределение, Логнормальное, Логистическое и многие другие.

Ключевым из них можно по праву назвать Нормальное или Гауссово благодаря **центральной предельной теореме статистики**[2], которая гласит, что сумма независимых одинаково распределенных случайных величин имеет распределение близкое к нормальному. Таким образом, нормальное распределение вы встретите практически везде, какими бы данными вы не занимались.

А что же насчет того, как задаются разные распределения? Для формального определения распределения случайной величины существует два способа.

- **Плотность распределения случайной величины** – универсальный вариант, подходящий как для дискретных, так и непрерывных случайных величин. Представляет собой функцию f(x), характеризующая сравнительную вероятность попадания случайной величины на участок x + Δx. Проще всего понять это определение геометрически:

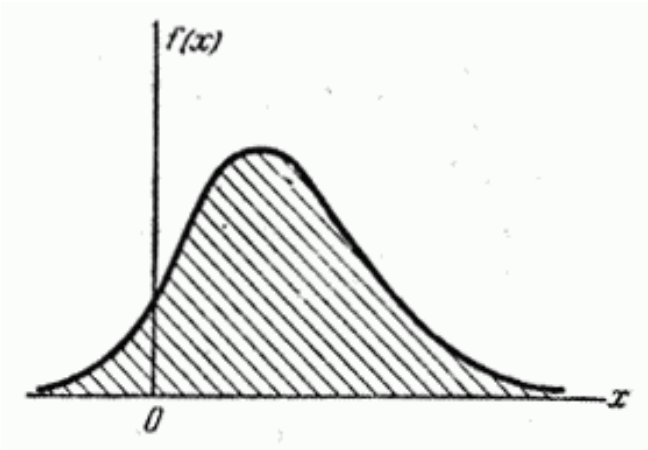

Рисунок 1 Плотность распределения случайной величины

На картинке выше по оси абсцисс представлены возможные значения случайной величины, а по оси ординат – частота реализации этих значений.

Кривая, соединяющая полученные точки, называется **кривой или гистограммой распределения** и очень часто используется для анализа данных, в чем мы убедимся на практике.

Свойства плотности распределения

- всегда принимает неотрицательные значения

[1] Соколов Г.А. Теория случайных процессов для экономистов, Учебное пособие. — М.: Физматлит, 2010. — 208 с.: ил. — ISBN 978-5-9221-1100-3.

[2] https://ru.wikipedia.org/wiki/Центральная_предельная_теорема

- редкие события имеют близкую к нулю вероятность, и их мы можем видеть на левом и правом “хвостах” графика. Формально это записывается следующим образом при x->±∞ f(x)-> 0

Плотность вероятности может выглядеть по-разному и задаваться различным набором параметров, например в случае Нормального распределения N(μ, σ) плотность равна функции Гаусса:

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

А ее кривая распределения имеет вид симметричного колокола:

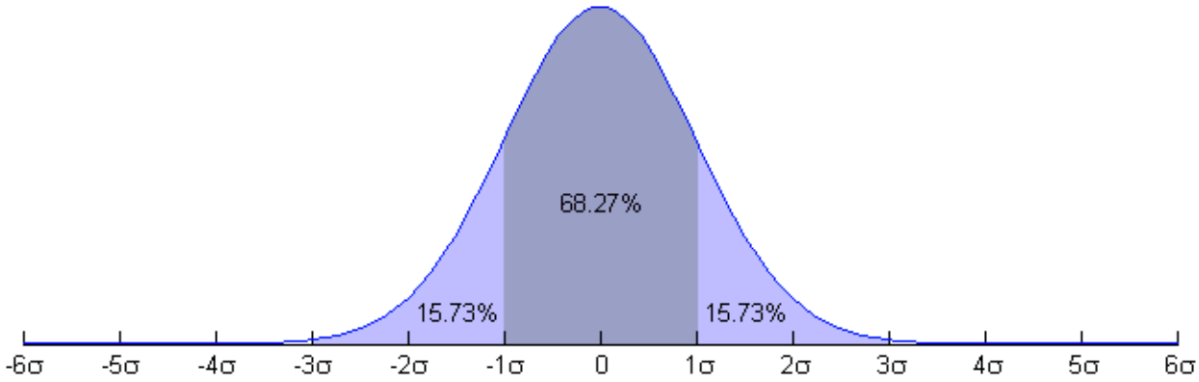

Рисунок 2 Плотность распределения для Нормального закона

- **Функция распределения случайной величины** существует только для непрерывных случайных величин и представляет собой f(x), характеризующую вероятность попадания x в элементарный интервал dx. На картинке это выглядит так:

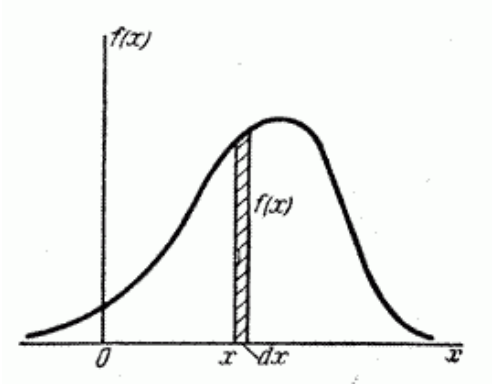

Рисунок 3 Функция распределения случайной величины

Возможно, у вас уже возник вопрос – если у нас уже есть закон распределения, какие полезные характеристики мы можем из него извлечь? μσ

Оценки распределения случайной величины

Глядя на плотность распределения случайной величины у нас уже могут появиться мысли о том, что можно исследовать – например, найти самое частое значение, максимум, минимум. Все это входит в понятие **точечных оценок** случайной величины – то есть оценок, полученных на основе наблюдаемой выборки.

На практике вам понадобится следующий джентельменский набор точечных показателей.

- **Математическое ожидание или среднее** значение случайной величины. Часто обозначается $\bar{x}$, m(x) или EV(x) – от английского expected value (ожидаемое значение). Наиболее распространённый показатель для базовой аналитики – аналитик, отвечая на вопрос, как часто пользователь делает покупки или обычно имеет в виду именно среднее значение.
- **Дисперсия –** физически представляет собой меру рассеяния значений вокруг среднего случайной величины. Полезный показатель для анализа разброса в вашей выборке. Вычисляется по формуле:

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

- **Среднеквадратическое отклонение, S(x) или σ –** еще одна мера рассеяния, корень из дисперсии:

$$S(x) = \sqrt[2]{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- **Мода или Mo(x)** – самое часто встречающееся значение случайной величины. Полезно при базовом анализе самых популярных исходов случайной величины – например, самое популярное время суток для покупки он-лайн игры, самая частая категория покупок по группе клиентов и тд.
- **Квантили или x_α** – семейство оценок, отвечающих на вопрос – какое значение x_α случайная величины не превысит с заданной вероятностью α. Полезно для поиска аномальных значений, выбросов. Чаще всего на практике используются **квартили** – квантили с вероятностями 0.25, 0.5, 0.75, разделяющие всю совокупность возможных исходов на четыре равных промежутка. 0.5 квартиль в статистике еще называют медианой.
- **Интерквартильный размах или IQR**– полезный аналог дисперсии, полезный для оценки разброса случайной величины. Представляет собой разницу между 0.75 и 0.25 квантилем - $x_{0,75} - x_{0,25}$

Для нормального распределения в силу его симметричности медиана будет совпадать с модой и математическим ожиданием, а остальные квантили будут выглядеть следующим образом:

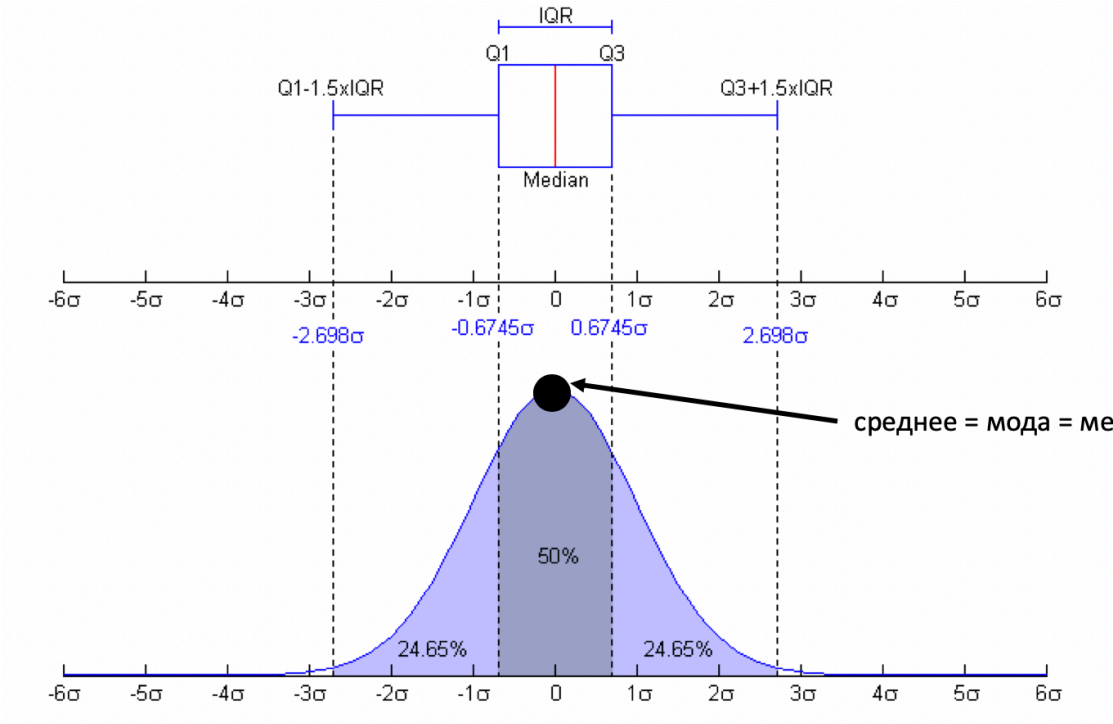

Рисунок 4 Точечные оценки для Нормального закона

Практика 1.1 Практика - анализ распределения случайной величины, симетричные данные

Попробуем применить полученные знания на практике и сделаем это при помощи любимого jupyter notebook.

Проклятие ассиметричности, бимодальность или где чаще всего ошибаются аналитики

Однако далеко не во всех случаях аналитики работают с нормально распределенными данными, что зачастую ведет к серьезным ошибкам. Почему так случается и что с этим делать?

В первую очередь, при работе с распределениями важно помнить, что они могут быть **ассиметричными.** Идеальный колокол Гауссова закона – частая ситуация, особенно при очень больших выборках, где работает центральная предельная теорема, однако не единственная возможная.

Представим, что вы работаете датасаентистом в банке и хотите помочь продакт менеджеру увеличить прибыль по клиентам – так называемую life time value. Вы собрали данные по всем пользователям и думаете с чего начать – в первую очередь приходит в голову вопрос – а сколько в среднем банк зарабатывает на одном клиенте за его жизнь в системе?

В случае ассиметрии данных вы можете получить такую картинку:

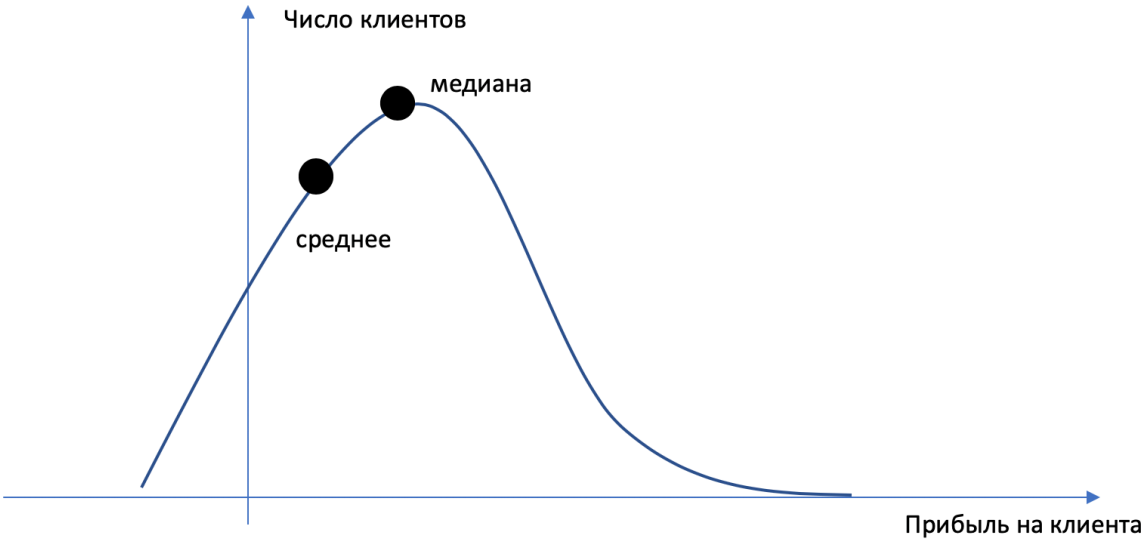

Рисунок 5 Пример ассиметричного распределения прибыли по клиенту в банке

В данном случае средняя оценка оказалась нерелевантной – посчитав только среднее мы не приняли во внимание, что у нашего распределения есть так называемые «хвосты» - наличие выбросов слева или справа, в данном случае – убыточные и очень прибыльные клиенты. Такие «хвосты» приводят к неверным выводам при оценке только по среднему.

Что же делать?
В первую очередь, использовать весь набор точечных оценок, например, квантили и показатели разброса для оценки всего спектра распределения. На примере выше можно видеть, что медиана могла бы быть более релевантной оценкой, чем среднее.
Второй – дополнительно оценить **показатели формы распределения**, например, **коэффициенты ассиметрии и эксцесса**.

- **Коэффициент ассиметрии (skewness)** – характеризует смещенность распределения и вычисляется по формуле

$$y_1 = \frac{m[\,(x - m(x))^3\,]}{S(x)^3}$$

Положителен, если правый хвост распределения длиннее левого, отрицателен – если левый длинные правого. У нормального распределения равен 0.

- **Коэффициент эксцесса** (kurtosis) характеризует остроконечность распределения и вычисляется по формуле:

$$y_2 = \frac{m[\,(x-m(x))^4\,]}{S(x)^4} - 3$$

Положительный при остром пике у среднего значения и отрицателен при гладком.
У нормального распределения эксцесс равен 0.

Помимо ассиметричного распределения, вам может встретиться еще один сложный случай. Представьте, что вы, наученные опытом со средним, сразу нарисовали гистограмму прибыли по клиентам и обнаружили вот такую картину:

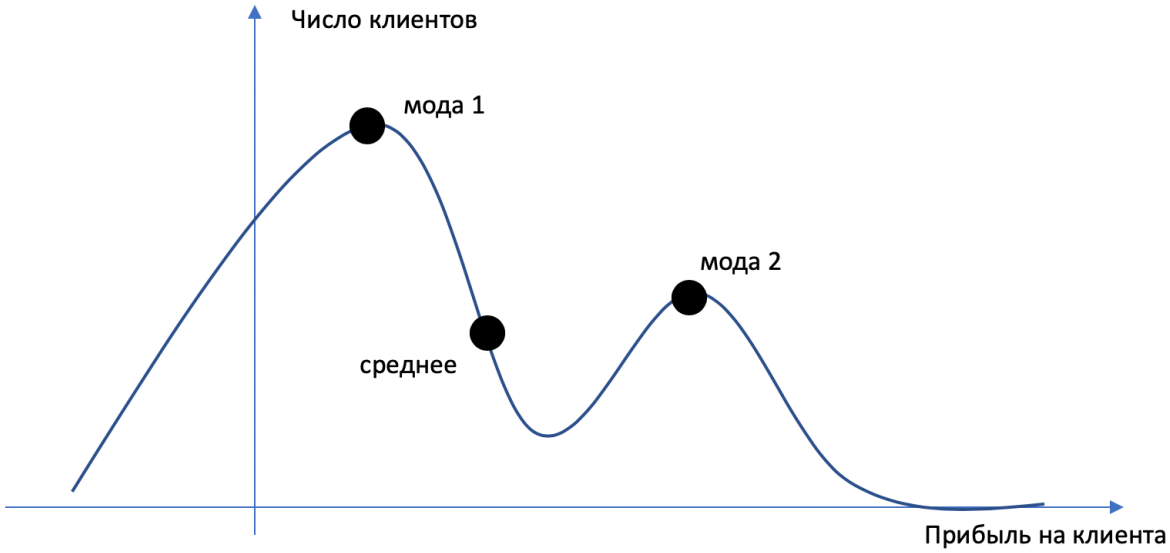

Рисунок 6 Пример бимодального распределения

Наличие подобных горбов на гистограмме указывает на существование нескольких мод – а значит, нескольких кластеров внутри изучаемой метрики. В нашем случае, это явно две группы клиентов с разной прибыльностью, и общая оценка по среднему или любому другому показателю теряет смысл, и возникает необходимость в анализе этих групп по отдельности.
В статистике подобное явление называется **мультимодальность** и говорит о наличии кластеров или структурных изменений в исследуемом процессе.

Что делать, если вы встретили мультимодальное распределение? Однозначно, попробовать отделить группы друг от друга и попытаться выявить причины различий. В нашем примере разница в прибыли по клиентам может говорить об их разном поведении (разные виды покупок и частота) или о разной экономике – кто- то использует более дорогие услуги, а по кому-то банк несет больше расходов, например на кешбек и смс.

Практика 1.2 Практика - анализ распределения случайной величины, ассиметричные данные

Теория 2 – Корреляционный анализ

В мире многие события взаимосвязаны и одна из ключевых задач датасаентиста – обнаруживать скрытые взаимосвязи и находить благодаря им новые точки роста или экономии. Задача выявления связей встретится вам в разных кейсах – от подбора фич для предсказания вероятности оттока клиента, до построения прогноза его покупок в зависимости от множества факторов, включая погоду за окном или время года.
Представьте, что вы уже собрали большой датасет из N фич и теперь думаете, с чего начать отбор лучших признаков. Логично выбрать для начала те фичи, которые максимально связаны с исследуемой целевой функций. В случае анализа оттока, например, вы можете выбрать для начала те признаки, которые больше всего коррелируют с числом дней, кгда клиент был активен.

Однако, как понять, что фичи и целевая функция - случайные величины X1,…,Xn и Y связаны? На этот вопрос отвечает раздел статистики под названием корреллляционно-регрессионный анализ или сокращенно КРА.
Сегодня мы коснемся той его части, которая рассказывает о линейных взаимосвязях между случайными величинами.

Что же такое **линейная корреляция**? Это статистическая взаимосвязь двух и более случайных величин, при которой изменений значений одной сопутствует изменений значений второй.
Корреляция может быть как **прямой**, так и **обратной,** например:

- Прямая связь - изменение числа покупок мороженого в зависимости от температуры воздуха. Чем жарче, тем больше мороженого продается.
- Обратная связь – потребление сахара и здоровье зубов – чем больше сладостей, тем вреднее для эмали.

Оценка силы связи для метрических и ранговых данных

Корреляцию можно измерять несколькими способами – в зависимости, от того, с каким типом данных вы работаете.
Для обычных **количественных** данных подходит **линейный кэффициент корреляции Пирсона**, который вычисляется по следующей формуле:

$$r(x,y) = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt[2]{\sum_{i=1}^{N}(x_i - \overline{x})^2 (y_i - \overline{y})^2}}$$

Коэффициент Пирсона принимает значения от -1 до 1, где -1 – максимальная обратная связь, а 1 – максимальная прямая связь.
Степень силы связи вычисляют обычно по шкале Чеддока, используя для этого значение коэффицента Пирсона по модулю:

Количественная мера тесноты связи	Качественная характеристика силы связи
0,1 - 0,3	Слабая
0,3 - 0,5	Умеренная
0,5 - 0,7	Заметная
0,7 - 0,9	Высокая
0,9 - 0,99	Весьма высокая

Рисунок 7 Шкала силы связи Чеддока

Для работы с количественными и ранговыми данными (например, число покупок и номер региона проживания из справочника) можно использовать **ранговый коэффициент корреляции Кендалла.**
Значения количественного показателя x выставляют в порядке возрастания и присваивают им ранги, значение рангового показателя y- сортируют и рассчитывают следующий коэффициент:

$$\tau = \frac{2S}{n(n-1)}$$

Где S = P – Q, P - суммарное число наблюдений, следующих за текущими наблюдениями с **большим** значением рангов Y, Q - суммарное число наблюдений, следующих за текущими наблюдениями с **меньшим** значением рангов Y.

Для анализа корреляций в большом датасете размерности N * N, где N- число фич, коэффициенты корреляции вычисляются попарно и образуют **матрицу парных корреляций:**

-	y	x_1	x_2	x_3
y	1	r_{yx1}	r_{yx2}	r_{yx3}
x_1	r_{x1y}	1	r_{x1x2}	r_{x1x3}
x_2	r_{x2y}	r_{x2x1}	1	r_{x2x3}
x_3	r_{x3y}	r_{x3x1}	r_{x3x2}	1

Рисунок 8 Матрица парных корреляций

С ее помощью можно удобно визуализировать связи в большом датасете, что мы и проделаем чуть позже в практике.

Мифы про связи - как не выдавать желаемое за действительное

Возможно, вы уже готовы проверить наличие связей в любимом датасете с Kaggle или рабочем проекте, но прежде чем начать, хочется добавить пару слов об ошибках, которые часто встречаются на начальном этапе анализа.

Частенько в реальности между явлениями физического мира существуют нелинейные связи. Такие связи невозможно выявить при помощи коэфициентов линейной корреляции, но можно найти при помощи визуализации **диаграммы рассеяния** – так называется график, более известный под именем scatter plot.

Несколько примеров нелинейных связей:

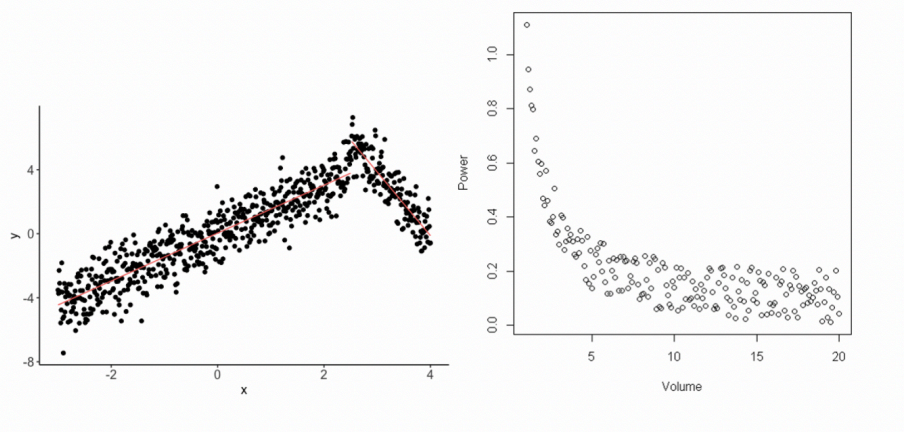

Рисунок 9 Пример нелинейных связей на scatter plot

Однако, в случае очень большого датасета, визуализация может быть слишком затратной для вашего рабочего ноута. Поэтому используется ее в меру)

Практика 2 - ищем взаимосвязи в данных

Теория 3 – доверительные интервалы

Мы с вами разобрались, что такое точечные оценки случайной величины и научились анализировать с их помощью симметричные и ассиметричные данные.
Однако, это не единственный существующий вид статистических оценок.

Вообразим, что мы решаем задачу продуктовой аналитики. В приложении добавили новый дизайн и несколько новых фич – и теперь наша задача выяснить, изменились ли к лучшему продуктовые метрики после релиза.
Конечно, можно было бы просто выбрать базовую метрику – например, число покупок через приложение и посмотреть на график ее изменений во времени (временной ряд).
И увидеть там нечто такое:

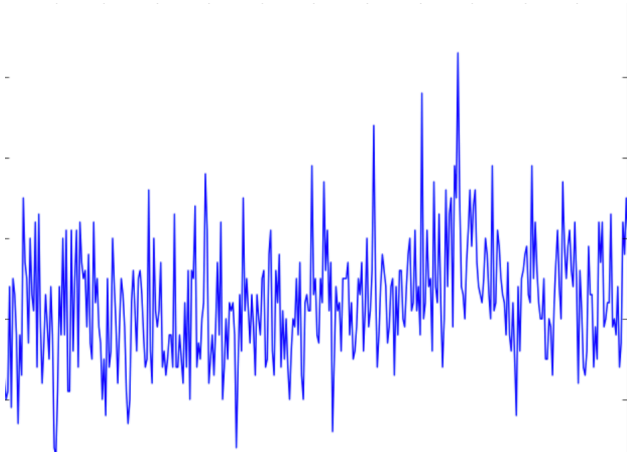

Рисунок 10 Волатильный временной ряд с сезонностью

Появится вопрос – что это за пики и поменялось ли хоть что-нибудь? Несмотря на то, что для анализа временных рядов существует отдельный научный раздел – эконометрика, методы базовой математической статистики тоже смогут нам помочь.
Для поиска ответа воспользуемся понятием **доверительного интервала.**
Доверительный интервал – это особый, **интервальный** вид оценки параметров случайной величины, который позволяет оценить их с заданной **доверительной вероятностью.**

Говоря простым языком, доверительный интервал уровня $\alpha = 95\ \%$ для параметра означает, что при проведении большого числа независимых экспериментов в 95 % случаев истинное значение параметра будет лежать внутри этого интервала.

Как строится доверительный интервал?
Помня о центральной предельной теореме, будем использовать преимущества, которые дает нам нормальное распределение больших датасетов.

Для нормально распределенных выборок существуют два варианта интервалов:

- Для математического ожидания μ с известной дисперсией – вычисляется по формуле

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Где z_α - квантиль нормального распределения уровня α.

- При неизвестной дисперсии

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Левая граница доверительного интервала обычно называется нижним интервалом, а правая – верхним.

Теперь попробуем проверить, как это работает на практике.

Практика 3. Построение доверительных оценок

Теория 4. Аномалии в данных

Мы научились работать с точечными и доверительными оценками и теперь готовы к самому сакраментальному вопросу – вопросу выбросов в данных или аномалий.
Под **аномальным значением** в общем случае понимается наблюдение, отличное от основной совокупности. Чаще всего аналитики задаются вопросом – что делать с этими значениями, удалять или заменять на что-то?

Хорошая практика в данном случае – **идти от задачи**. В случае, если вы хотите делать долгосрочные прогнозы по финансовому инструменту, то, как в нашем примере с акциями Google, повторяющиеся аномальные всплески доходности стоит разметить отдельно и сгладить, дабы не вносить нестабильность в модель за счет разовых событий.
Если же ваша задача – оперативно находить проблемы технического характера – например, поломки в приложении, ведущие к потере клиентов, то задача превращается в кейс отдельной статистической дисциплины – **обнаружения аномалий или anomaly detection.**

Какие техники обнаружения аномалий бывают?

- **Одномерный подход** – выбросы отбираем только по одному признаку. Например, у нас есть котировка акций и для поиска нестандартных значений мы используем только эти данные. В таком случае ваши базовые инструменты – это
  - IQR – все, что выпадает за предел интерквартильного расстояния IQR помечается как выброс (см Рисунок 4)
  - 99 или 25 квантили – все, что больше или меньше соответствующих значений можно рассматривать как необычное значение
  - правило трех сигм – для нормально распределенной выборки все значения содержатся в интервале $m(x) \pm 3\sigma$, поэтому все выходящие за этот промежуток значения можно пометить, как аномальные
  - доверительные интервалы – значения за пределами доверительного интервала можно классифицировать как выбросы с заданной доверительной вероятностью

- **Многомерный подход** – в случае, если вы исследуете события, описываемые N метриками – например, фиксируемые в разные моменты времени показания с датчиков нефтепровода, то могут быть полезны более сложные модели поиска выбросов в многомерном пространстве. Несколько примеров таких алгоритмов:
  - **Isolation forest** - https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf
  - **One Class Support Vector Machine** - https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf

Последние требуют правильного отбора признаков для обучения и больших вычислительных мощностей, поэтому для начала рекомендуется попробовать более простые подходы – они будут эффективны для 80 % задач начинающего датасаентиста.

Практика 4. Поиск аномалий в данных

Вместо заключения

Несмотря на то, что математическая статистика – формальная и строгая наука, в анализе данных, как и в жизни, много открытий и сюрпризов. Помните о базовых принципах, не бойтесь пробовать новое и экспериментировать с миксами методов и моделей. Уверена, что впереди вас ждет много интересного)