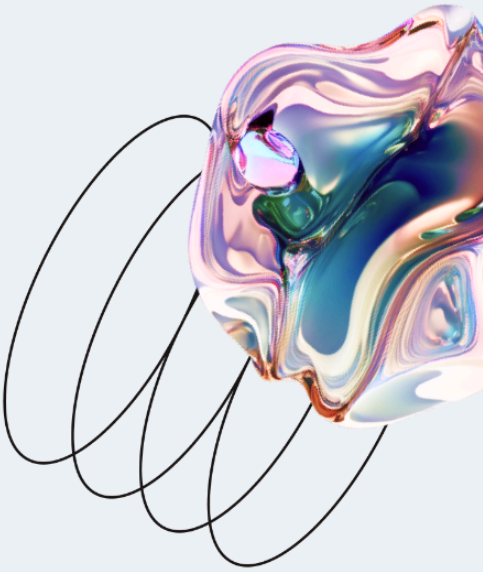
GeekBrains



Основы статистики

Data Science

Оглавление

Введение 4

Словарь терминов 4

Почему статистика? 5

Этапы процесса анализа данных 5

Исследуйте свои данные: наблюдения, переменные, типы переменных 6

Матрица данных и таблица частот 10

Графики и формы распределения 12

Для категориальных переменных 13

Для количественных переменных 14

Среднее значение, медиана и мода 17

Среднее значение 19

Диапазон, межквартильный диапазон и прямоугольная диаграмма 20

Диапазон 20

Пример диапазона 20

Межквартильные диапазоны и выбросы 21

Межквартильный диапазон (IQR) 22

Как рассчитать межквартильный размах (IQR): 22

Пример расчёта IQR 23

Преимущество IQR 23

Что такое блочные диаграммы? 24

Дисперсия и стандартное отклонение 25

Почему дисперсия и стандартное отклонение являются хорошими показателями изменчивости? 25

Дисперсия 26

Как рассчитать дисперсию шаг за шагом: 26

Интуиция 27

Среднеквадратичное отклонение 28

Свойства стандартного отклонения 28

Генеральная совокупность против выборочной дисперсии и стандартного отклонения 28

Может возникнуть вопрос, почему во время расчёта дисперсии, мы возводим разницу в квадрат? 29

Нормальное распределение, биномиальное распределение и распределение Пуассона 29

Нормальное распределение или распределение Гаусса, кривая нормального распределения 30

Свойства нормального распределения: 30

Расчёт вероятности нормального распределения 31

Распределение в форме колокола и эмпирическое правило 32

Испытание Бернулли и биномиальное распределение 33

Испытание Бернулли или биномиальное испытание 34

Примеры испытания Бернулли или биномиального испытания 34

Биномиальное распределение 34

Формула биномиального распределения 35

Пример биномиального распределения 35

Распределение Пуассона 36

Пример распределения Пуассона 38

Заключение 39

Введение

Всем привет! Добро пожаловать на лекцию по статистике в контексте передовых технологий Data Science! В течение этой лекции мы будем рассматривать роль статистики в анализе данных и узнаем, как она помогает нам извлекать ценную информацию из больших объёмов данных.

1. **Что вы узнаете?**

В ходе этой лекции вы узнаете основные понятия и методы статистики, которые используются в Data Science. Мы рассмотрим такие важные темы, как описательная статистика, вероятность, распределения, статистические тесты и многое другое. Вы также узнаете, как применять эти методы для анализа данных и принятия обоснованных решений на основе статистических выводов.

2. **Чему вы научитесь?**

В результате изучения этой лекции вы научитесь:

- Понимать основные понятия и термины статистики, используемые в Data Science.
- Применять методы описательной статистики для анализа данных и получения сводной информации о них.
- Оценивать вероятность событий и использовать её для принятия решений на основе данных.
- Разбираться в различных типах распределений и применять их для моделирования данных.
- Использовать статистические тесты для проверки гипотез, а также для того, чтобы делать статистически обоснованные выводы.

Словарь терминов

Искусственный интеллект — программа, которая демонстрирует когнитивные способности, подобные человеческим.

Глубокое обучение — это более специализированная версия машинного обучения, в которой используются более сложные методы для решения сложных задач.

Обучение с учителем — один из самых основных типов машинного обучения. В этом типе алгоритм машинного обучения обучается на размеченных данных.

Почему статистика?

Статистические методы в основном полезны для обеспечения правильной интерпретации ваших данных. Эти кажущиеся отношения действительно «значительны» или значимы, и это не происходит случайно. На самом деле, статистический анализ помогает найти смысл в бессмысленных числах.

Таким образом, статистика — это не что иное, как некоторое числовое значение, которое может описать определённое свойство вашего набора данных. Есть несколько хорошо известных статистических данных, таких как медиана (или «среднее значение»), стандартное отклонение и т. д. Стандартное отклонение — это изменчивость в наборе данных вокруг среднего значения. Дисперсия — это квадрат стандартного отклонения. Линейный тренд — ещё один пример «статистики» данных.

Этапы процесса анализа данных

Прежде чем приступить к работе с конвейером анализа данных, вы должны знать, что он состоит из пяти основных этапов.

Шаг 1. Определитесь с целями или задайте вопрос

Первым шагом конвейера анализа данных является определение целей. Эти цели обычно могут потребовать сбора и анализа значительного объёма данных.

Шаг 2. Что измерять и как измерять

Измерение обычно относится к присвоению чисел для обозначения различных значений переменных. Предположим, в ходе своего исследования вы пытаетесь выяснить, существует ли связь между ростом и весом человека, имеет ли смысл измерять рост и вес собак с помощью весов.

Шаг 3. Сбор данных

Как только вы узнаете, какие типы данных вам нужны для вашего статистического исследования, вы сможете определить, могут ли ваши данные быть собраны из существующих источников / баз данных, или нет. Если данных недостаточно, вам необходимо собрать новые данные. Даже если у вас есть существующие данные, очень важно знать, как они были собраны. Это поможет вам понять, что вы можете определить ограничения обобщаемости результатов и провести надлежащий анализ.

Чем больше данных у вас есть, тем лучше корреляции, создание лучших моделей и поиск более действенных идей становится для вас проще. Особенно полезно

собирать данные из как можно более разнообразных источников: это помогает значительно упростить дальнейшую работу.

Шаг 4. Очистка данных

Ещё один важный шаг в конвейере анализа данных — улучшить качество существующих данных. Слишком часто специалисты по данным исправляют орфографические ошибки, обрабатывают пропущенные значения и удаляют бесполезную информацию. Это самый важный шаг, поскольку ненужные данные могут привести к неверным результатам и ввести бизнес в заблуждение.

Шаг 5. Обобщение и визуализация данных

Исследовательский анализ данных помогает лучше понять данные. Потому что картинка действительно стоит тысячи слов, так как многие люди понимают картинки лучше, чем лекцию. Точно так же меры дисперсии указывают на распределение данных вокруг центра. Корреляция относится к степени, в которой две переменные движутся синхронно друг с другом.

Шаг 6. Моделирование данных

Теперь создавайте модели, которые сопоставляют данные с вашими бизнес-результатами и дают рекомендации. Именно здесь уникальный опыт специалистов по данным становится важным для успеха бизнеса. Сопоставление данных позволяет строить модели, прогнозирующие результаты бизнеса.

Шаг 7. Оптимизируйте и повторите

Анализ данных — это повторяемый процесс, который иногда приводит к постоянным улучшениям как в бизнесе, так и в самой цепочке создания стоимости данных.

Теперь вы знаете шаги, связанные с конвейером анализа данных. Прежде чем перейти к более сложным методам, я предлагаю начать своё путешествие по анализу данных.

Исследуйте свои данные: наблюдения, переменные, типы переменных

Набор данных содержит информацию об образце. Набор данных состоит из наблюдений. **Наблюдение (кейс)** — экспериментальная единица. Это могут быть лица, от которых собираются данные. Когда данные собираются от людей, мы иногда называем их **участниками**. Когда данные собираются с животных, часто

используется термин **«субъекты»**. Другой синоним — **экспериментальная установка**. Итак, кейсы — это не что иное, как объекты в коллекции.

Каждый кейс имеет один или несколько атрибутов или качеств, называемых переменными, которые являются характеристиками кейсов. **Переменная** — это измеряемая характеристика, которая может принимать различные значения. Другими словами, что-то, что варьируется между различными наблюдениями.

Этим переменная отличается от константы, которая одинакова для всех наблюдений в исследовании.

Наблюдение (кейс)

Экспериментальный блок, из которого собираются данные.

Переменная

Характеристики наблюдений, которые могут принимать разные значения (другими словами, что-то, что может варьироваться).

Постоянная

Характеристика, одинаковая для всех случаев в исследовании.

Пример:

Предположим, вы собираете информацию о больных раком молочной железы. Теперь для каждого больного раком вы хотите знать следующую информацию.

Код образца: идентификационный номер

Толщина комка: 1—10

Однородность размера ячейки: 1—10

Однородность формы клеток: 1—10

Маргинальная адгезия: 1—10

Размер одной эпителиальной клетки: 1—10

Голые ядра: 1—10

Мягкий хроматин: 1—10

Нормальные ядрышки: 1—10

Митозы: 1—10

Класс: 2 для доброкачественных, 4 для злокачественных

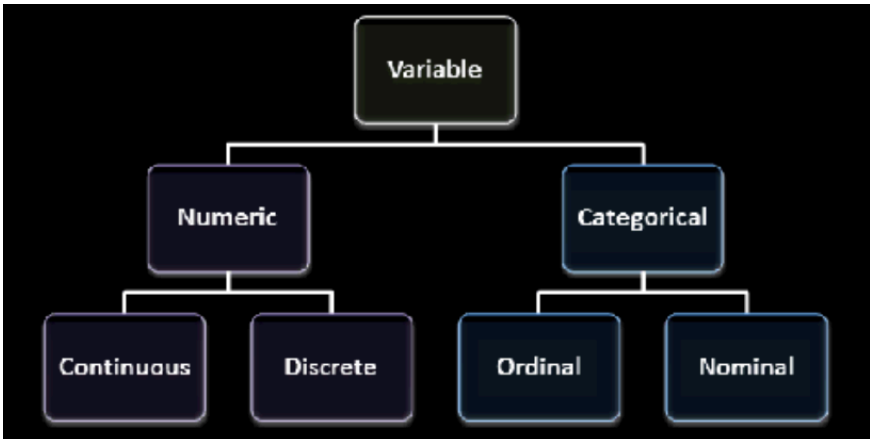
В этом примере пациенты с раком молочной железы сами являются кейсами, а все эти характеристики пациентов являются переменными.

В исследовании кейсами могут быть самые разные вещи. Это могут быть отдельные пациенты и группы пациентов. Но они также могут быть, например, компаниями, школами или странами и т. д.

У нас может быть много разных типов переменных, представляющих разные характеристики. По этой причине существуют различные уровни измерений или различные типы переменных.

Переменная — это характеристика, которую можно измерить и которая может принимать различные значения. Рост, возраст, доход, провинция или страна рождения, оценки, полученные в школе, и тип жилья — всё это примеры переменных. Переменные можно разделить на две основные категории: категориальные и числовые.

Затем каждая категория подразделяется на две подкатегории: номинальная или порядковая для категориальных переменных, дискретная или непрерывная для числовых переменных. Кратко рассмотрим эти типы.



Категориальные переменные

Категориальная переменная (также называемая качественной переменной) относится к характеристике, которую нельзя измерить количественно. Категориальные переменные могут быть как номинальными, так и порядковыми. И номинальные, и порядковые переменные можно назвать категориальными переменными.

Номинальная переменная

Номинальная переменная состоит из различных категорий, которые не имеют порядка. Номинальная переменная — описывает имя, метку или категорию без

естественного порядка. Пол человека или тип жилья являются примерами номинальных переменных.

Пример:

Пол пациента может быть мужским, женским. Или различным может быть город, в котором он проживает. Здесь каждая категория отличается друг от друга, но нет порядка ранжирования. Точно так же в приведённом ниже примере переменная «вид транспорта для поездки на работу» является номинальной.

Mode of transportation for travel to work	Number of people
Car, truck, van as driver	9,929,470
Car, truck, van as passenger	923,975
Public transit	1,406,585
Walked	881,085
Bicycle	162,910
Other methods	146,835

Порядковая переменная

Второй уровень измерения — это порядковый уровень. Порядковая переменная — это переменная, значения которой определяются отношением порядка между различными категориями. Существуют не только различия между категориями переменной, но и различия в порядке значений. Примером может быть «Самый высокооплачиваемый», «Среднеоплачиваемый» и «Самый низкооплачиваемый» сотрудник.

В приведённом ниже примере переменная «поведение» является порядковой, поскольку категория «Отлично» лучше, чем категория «Очень хорошо», что, в свою очередь, лучше, чем категория «Хорошо» и т. д. Существует некоторое естественное упорядочивание, но оно ограничено, поскольку мы не знаем, насколько «отличное» поведение лучше, чем «очень хорошее».

Student behaviour ranking	
Behaviour	**Number of students**
Excellent	5
Very good	12
Good	10
Bad	2
Very bad	1

Количественные/числовые переменные

Числовая переменная (также называемая количественной переменной) — это количественная характеристика, значениями которой являются числа (за исключением чисел, которые представляют собой коды, обозначающие категории). Числовые переменные могут быть как непрерывными, так и дискретными.

Непрерывная переменная

Переменная непрерывна, если возможные значения переменной образуют интервал. Примером может служить, опять же, рост пациента. Кто-то может быть ростом 172 сантиметра, а кто-то — 174 сантиметра. Но также, например, рост может составлять 170.2461 сантиметра. У нас не набор отдельных чисел, а бесконечная область значений.

Дискретная переменная

Переменная дискретна, если её возможные категории образуют набор отдельных чисел.

Для приведённых выше данных о раке молочной железы однородность размера ячейки (1—10) является примером дискретной переменной.

Матрица данных и таблица частот

Матрица данных

Матрица данных — это прямоугольная таблица или матрица, в которой строки представляют наблюдения или случаи, а столбцы — переменные или атрибуты. Каждая ячейка матрицы содержит значение, соответствующее переменной для данного наблюдения. Матрица данных может использоваться для организации и хранения данных для лёгкого анализа и интерпретации.

Таблица частот

С другой стороны, таблица частот представляет собой табличное представление частотного распределения категориальной переменной. Она показывает количество или частоту наблюдений, которые попадают в каждую категорию переменной. Каждая строка в таблице частот представляет категорию переменной, а соответствующий столбец показывает количество наблюдений, попадающих в эту категорию. Таблицы частот можно использовать для суммирования и визуализации категорийных данных, а также для расчёта различных сводных статистических данных, таких как мода и процент наблюдений в каждой категории.

Если вы проводите исследование, вы должны думать о своих данных с точки зрения наблюдений и переменных.

Наблюдения — это объекты, которыми в вашем исследовании могут быть люди, животные или предметы, а *переменные* — это интересующие характеристики. Сейчас я расскажу, как можно упорядочить и представить свои наблюдения и переменные. Возьмём, к примеру, и предположим, что вас интересует «Primera División», главное футбольное соревнование в Испании. Здесь интересующие вас случаи — это отдельные футболисты в лиге, а переменные, на которые вы обращаете внимание, — это возраст, масса тела, забитые голы, членство в команде и цвет волос. Лучший способ упорядочить всю эту информацию — с помощью матрицы данных.

Таким образом, *матрица данных* — это табличное представление случаев и переменных вашего статистического исследования. Каждая строка матрицы данных представляет случай, а каждый столбец представляет собой переменную.

Полная матрица данных может содержать тысячи, тысячи или даже больше наблюдений.

Переменные

Наблюдения

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
6.5	3.2	5.1	2	Iris-virginica
6.4	2.7	5.3	1.9	Iris-virginica
6.8	3	5.5	2.1	Iris-virginica
6.7	3.1	4.4	1.4	Iris-versicolor
5.6	3	4.5	1.5	Iris-versicolor
5.8	2.7	4.1	1	Iris-versicolor

Чтобы получить больше информации, очень полезно обобщать данные. Хороший способ сделать это — составить таблицу частот. Таблица частот показывает, как значения переменной распределяются по наблюдениям. Рассмотрим следующий пример, чтобы понять это. Мы можем получить частоту элементов, а затем процент, или даже вычислить совокупный процент.

Class	Frequency	Percentage	Cumulative Percentage
Iris-setosa	2	25%	25%
Iris-virginica	3	38%	63%
Iris-versicolor	3	38%	100%
Total	8	100%	

Здесь мы имеем всего 8 случаев, и среди 8 случаев 2 случая (25%) принадлежат Iris-Setosa.

3 случая, что означает, что 38% случаев принадлежат Iris-Virginia, и аналогично ещё 38% относятся к Iris Versicolor.

Вышеприведённый пример относится к категориальной переменной с именем class. Но подумайте, если ваша переменная является количественной, то вычисление процента для каждого конкретного значения не имеет смысла. В этом случае сначала приведите свои данные к некоторым порядковым категориям, используя интервалы. Затем займитесь остальными делами.

Графики и формы распределения

Формы распределений: определения, примеры

Форма распределения или способ распространения данных определяется сочетанием его центральной тенденции, изменчивости и асимметрии.

Главная тенденция

Центральная тенденция распределения относится к тому, где собираются данные. Наиболее распространёнными мерами центральной тенденции являются среднее значение, медиана и мода. Выбор меры может повлиять на форму распределения. Например, распределение с длинным хвостом в одну сторону может иметь среднее значение, отличное от медианы, которая является мерой центрального значения.

Изменчивость

Изменчивость распределения относится к тому, насколько разбросаны данные. Одним из способов измерения изменчивости является вычисление диапазона, представляющего собой разницу между самым высоким и самым низким значениями. Другой способ — вычислить стандартное отклонение, которое измеряет разброс данных вокруг среднего значения.

Асимметрия

Асимметрия относится к степени асимметрии в распределении. Распределение, скошенное влево, имеет хвост, простирающийся влево, а распределение, скошенное вправо, имеет хвост, простирающийся вправо. Если распределение симметрично, то оно имеет нулевую асимметрию.

По мере построения набора данных он может генерировать различные формы среди десятков возможностей, каждая из которых представляет своё распределение. Изучение формы распределения может обеспечить визуальное представление, иллюстрирующее распределение данных.

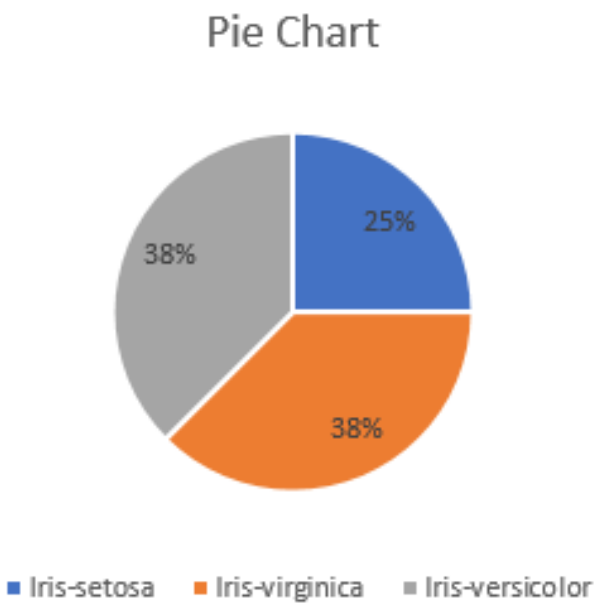
Для категориальных переменных

Если интересующая переменная является категориальной, то, как правило, **круговая** диаграмма или **гистограмма** являются лучшим представлением.

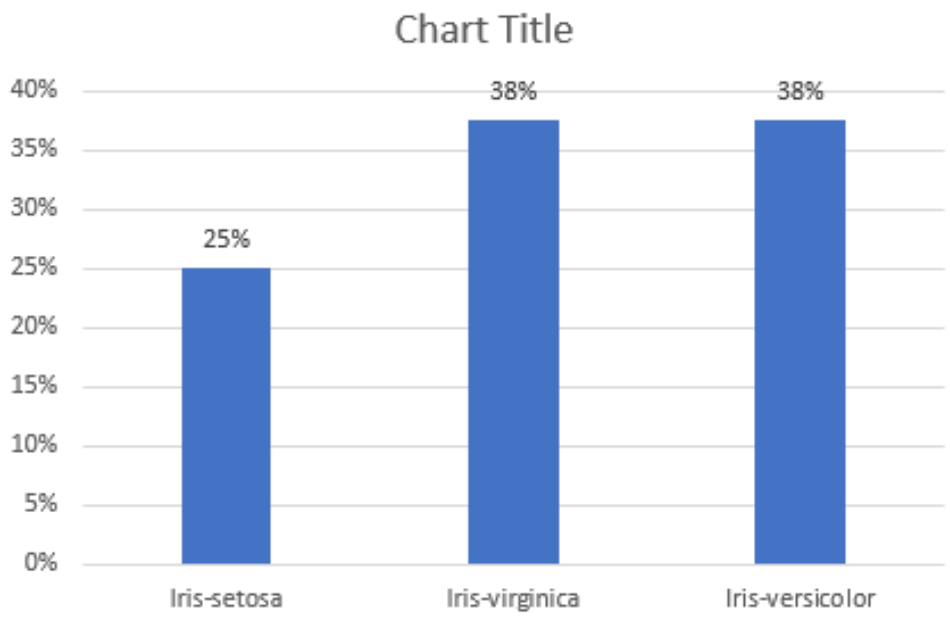
Если категорий слишком много, то круговая диаграмма может быть запутанной, но столбчатая диаграмма даст чёткое представление**. Таким образом, гистограмма имеет преимущество перед круговой диаграммой, когда ни одна из переменных не слишком высока.**

Class	Frequency	Percentage	Cumulative Percentage
Iris-setosa	2	25%	25%
Iris-virginica	3	38%	63%
Iris-versicolor	3	38%	100%
Total	8	100%	

Если мы нарисуем круговую диаграмму для процентного столбца приведённой выше таблицы, то она будет выглядеть следующим образом.



Для той же таблицы, если мы нарисуем гистограмму для процентного столбца, она будет выглядеть так.



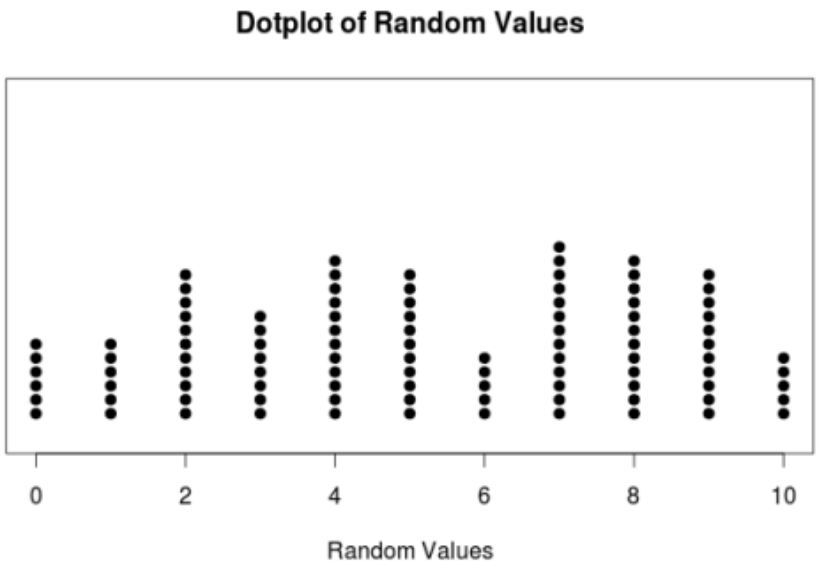
По сути, для категориальной переменной вы можете сделать множество представлений, перечисленных ниже:

- Столбчатая диаграмма
- Круговая диаграмма
- Таблица частот
- Таблица сопряжённости
- Сегментированный столбчатый график
- Относительная частота
- Мозаичный график

Для количественных переменных

Точечный график

Если вы работаете с количественными переменными или числовыми переменными, то точечный график — это один из видов представления, которое можно использовать. Точечный график выглядит так. Нанесите каждую точку на график после проведения горизонтальной линии и отметьте на ней возможные значения через равные промежутки времени.

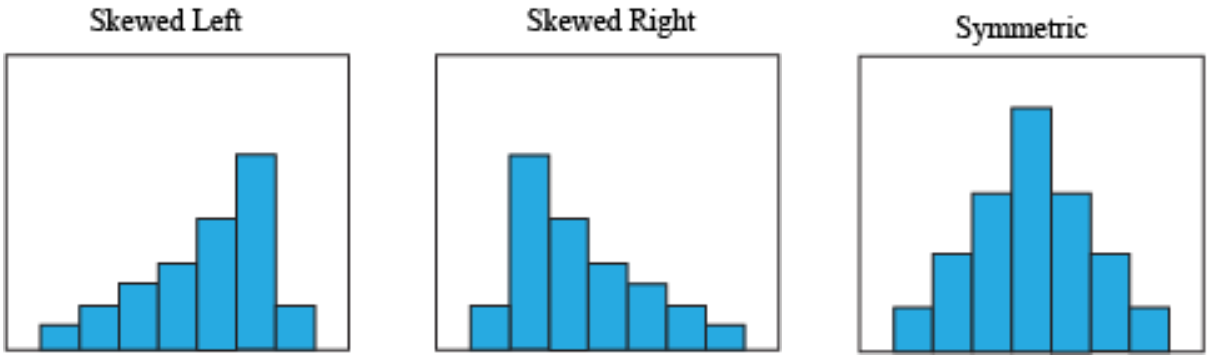


Но если у вас очень большая выборка, то точечный график может выглядеть беспорядочно.

В таком случае может быть полезен другой вид представления, называемый гистограммой.

Гистограмма

Гистограмма похожа на столбчатую диаграмму в том смысле, что она использует столбцы для отображения частот или относительных частот возможных значений переменной. Однако есть одно важное отличие. Разница в том, что столбцы гистограммы соприкасаются друг с другом. Это касание означает, что значения переменной интервала/отношения представляют лежащую в основе непрерывную шкалу. Ниже приведён пример гистограммы.

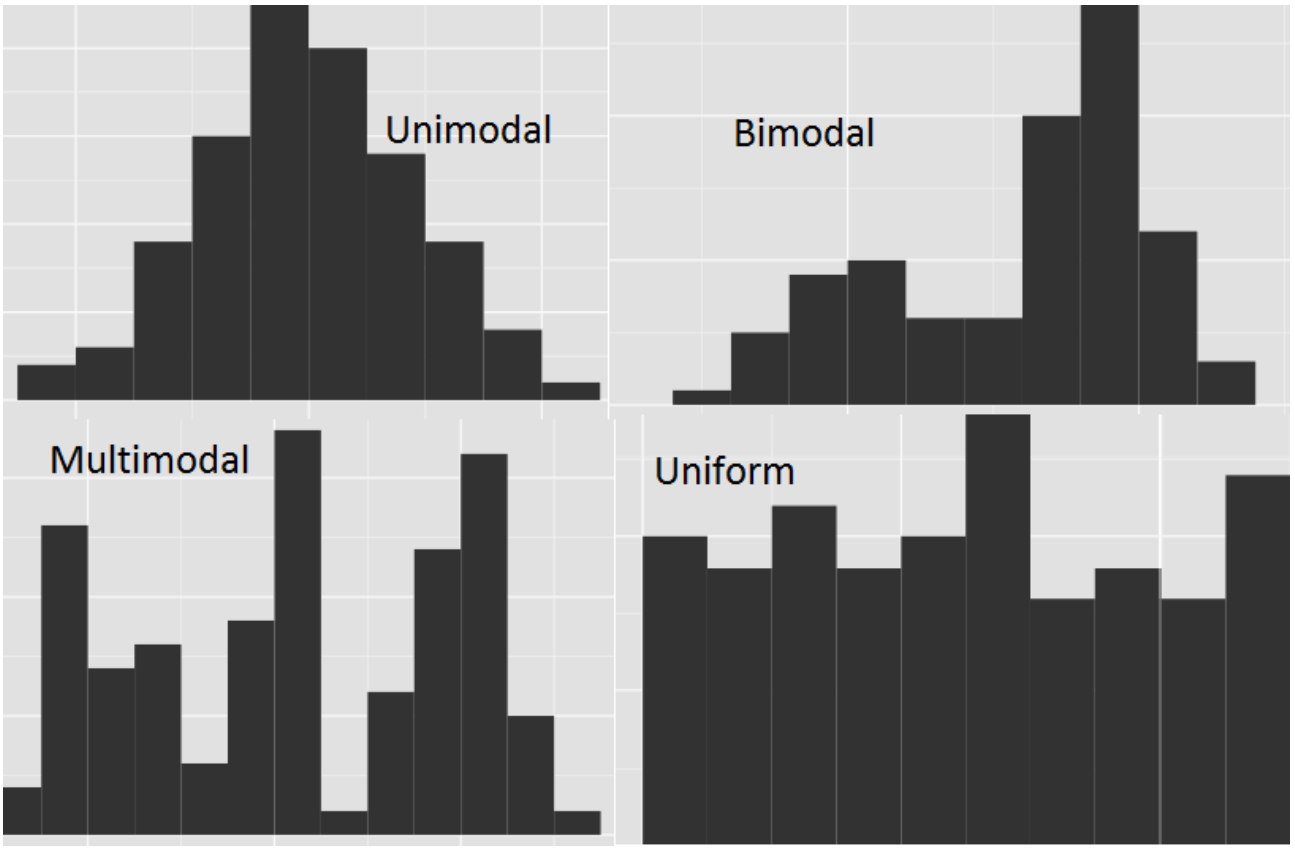


Обратите внимание на приведённую выше гистограмму и посмотрите на распределение. Есть три вида форм.

1. Правая имеет форму колоколообразной кривой, имеет одну вершину и приблизительно симметрична.
2. Левый график — перекошенный и унимодальный.
3. Средняя перекошен вправо и также унимодальный.

Есть четыре вида модальностей:

- **Унимодальный:** имеет только один пик
- **Бимодальный:** имеет два пика
- **Мультимодальный:** у него много пиков
- **Равномерный:** всё распределено равномерно



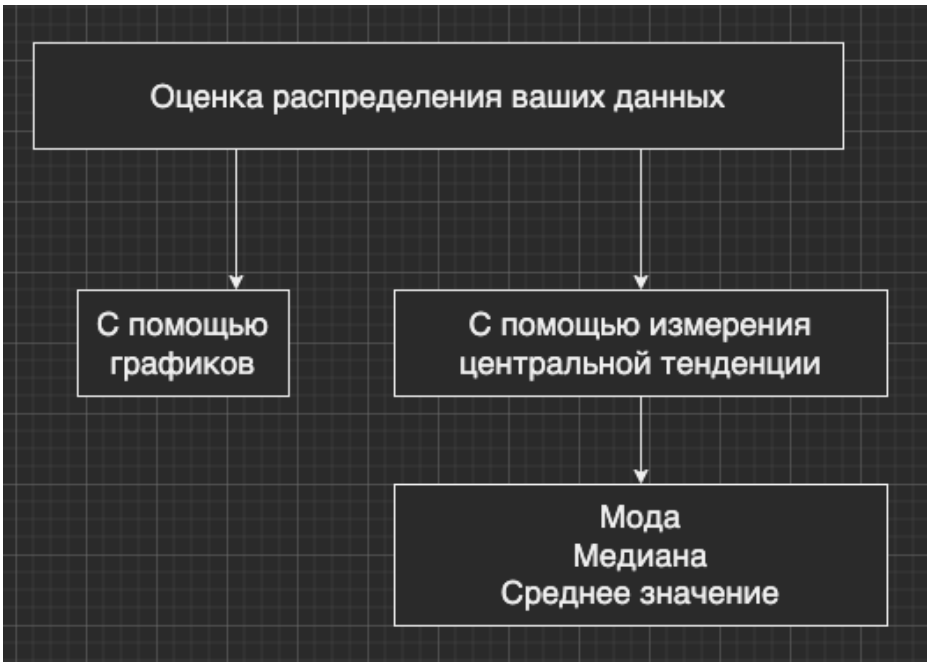
При работе с любыми данными не забывайте следить за формой распределения. Это имеет существенное значение, поскольку может повлиять на статистические методы, которые вы собираетесь использовать позже.

Таким образом, форма распределения определяется сочетанием его центральной тенденции, изменчивости и асимметрии. Различные меры центральной тенденции и изменчивости, а также различные степени асимметрии могут давать широкий диапазон форм распределения.

Среднее значение, медиана и мода

Среднее значение, медиана и мода — это три меры центральной тенденции, используемые в статистике для описания центрального или типичного значения набора данных.

Чуть ранее мы рассмотрели, как суммировать распределение ваших данных с точки зрения графиков. Теперь пришло время измерить центр вашего распределения. Как только речь заходит об измерении центральной тенденции переменной, на «сцену» выходят мода, медиана и среднее значение.



Мода

Мода — это наиболее распространённое значение в наборе данных. Набор данных может иметь несколько мод или вообще не иметь моды, если нет значения, которое встречается более одного раза.

Если интересующая вас переменная измеряется на номинальном или порядковом (категориальном) уровне, то нахождение моды является наиболее часто используемым методом для измерения центральной тенденции ваших данных.

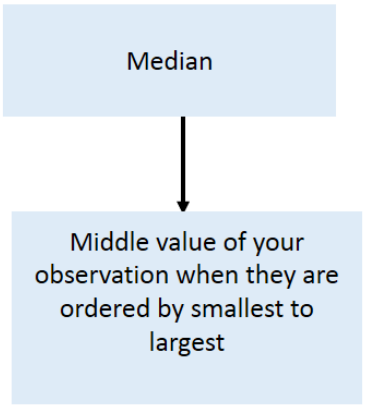
Найти моду несложно. По сути, это значение, которое встречается чаще всего. Другими словами, мода является наиболее распространённым результатом. Мода — это название категории, которое встречается чаще.

Существует вероятность наличия более одной моды в вашей переменной.

Медиана

Второй мерой центральной тенденции является медиана. Медиана — это не что иное, как среднее значение ваших наблюдений, когда они расположены в порядке от наименьшего к наибольшему.

Итак, медиана — это среднее значение в отсортированном наборе данных. Это значение, которое отделяет самые высокие 50% данных от самых низких 50%. Чтобы найти медиану, нужно отсортировать данные по порядку и найти значение, которое находится ровно посередине. Если имеется чётное количество значений, медиана является средним значением двух серединных значений.



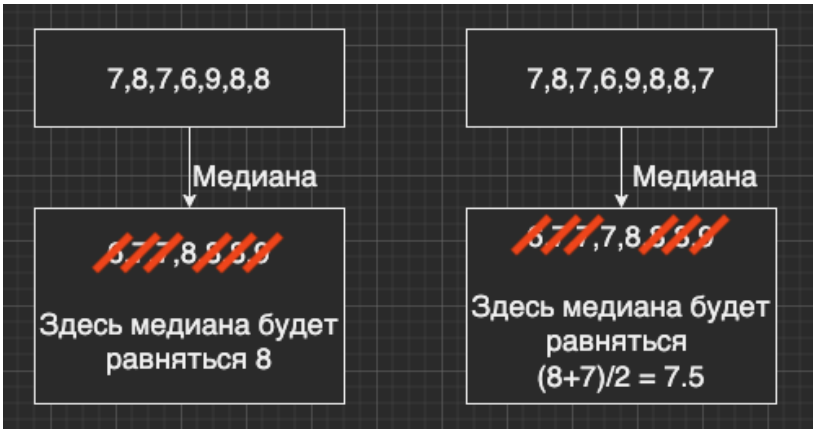
Процесс нахождения медианы включает в себя два шага:

Шаг 1. Упорядочивайте измерения от меньшего к большему

Шаг 2. Найдите среднее значение

Если у вас нечётное количество наблюдений, найти среднее значение несложно. Предположим, у вас есть 5 измерений. Итак, после упорядочивания всегда 3-я позиция является средним значением.

В случае если у вас чётное число измерений (допустим, 6 измерений), единого среднего значения нет. Тогда как вычислить медиану? Необходимо просто взять среднее из двух серединных значений.



Среднее значение

Среднее значение, также известное как среднее арифметическое или среднее, представляет собой сумму всех значений в наборе данных, разделённую на количество значений. Это мера центральной тенденции, которая отражает центр масс данных. Формула среднего значения:

Среднее значение = (сумма значений) / (количество значений)

$$\overline{X} = \frac{\Sigma X}{n}$$

Среднее значение представляет собой сумму всех значений, делённую на количество наблюдений. Это не что иное, как среднее значение.

Теперь возникает вопрос: когда и какое измерение центральной тенденции использовать?

- Если данные являются категориальными (номинальными или порядковыми), невозможно рассчитать среднее значение или медиану. Тогда необходимо выбрать моду.
- Если ваши данные количественные, используйте среднее значение или медиану. По сути, если ваши данные имеют некоторые значительные выбросы или данные сильно искажены, то медиана является лучшим показателем для определения центральной тенденции. В противном случае выберите среднее значение.

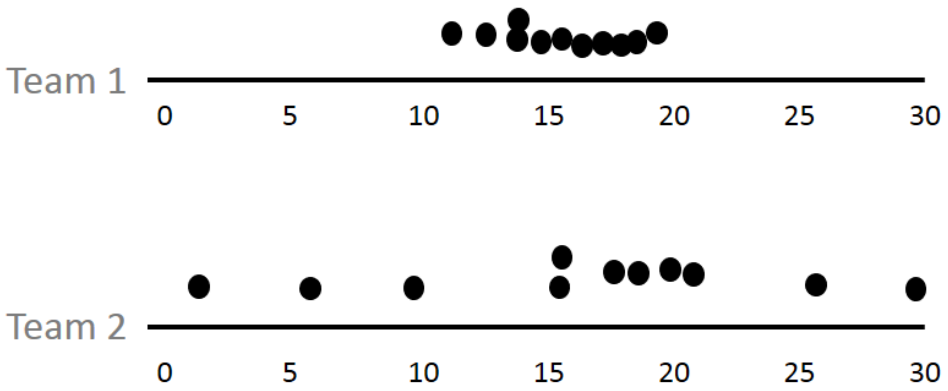
Диапазон, межквартильный диапазон и прямоугольная диаграмма

Диапазон

Диапазон — это статистическая мера, которая рассчитывается путём вычитания минимального значения набора данных из максимального значения. Это простая мера изменчивости, но она чувствительна к выбросам, поскольку даже одно экстремальное значение может сильно повлиять на диапазон.

Пример **диапазона**

Возьмём приведённый ниже пример:



Если вы считаете обе команды, их мода = 14,1, медиана = 15 и среднее = 15.

Это указывает на то, что если вы когда-нибудь адекватно описываете распределение, ему может потребоваться **больше информации, чем меры центральной тенденции.**

В этой ситуации на первый план выходят **меры изменчивости.** Какие же существуют меры изменчивости?

- Диапазон
- Межквартильный диапазон
- Блочная диаграмма, чтобы получить хорошее представление о том, как распределяются значения в распределении

Наиболее простой мерой изменчивости является диапазон. Это разница между самым высоким и самым низким значением.

Для приведённого выше примера диапазон будет следующим:

Диапазон (Команда 1) = 19,3 — 10,8 = 8,5

Диапазон (Команда 2) = 27,7 — 0 = 27,7

Поскольку диапазоны учитывают только количество экстремальных значений, иногда это может не оказать должного влияния на изменчивость. В этом случае вы можете использовать другую меру изменчивости, называемую межквартильным размахом (IQR).

Межквартильные диапазоны и выбросы

Межквартильный размах (IQR) — это мера статистической дисперсии, основанная на делении набора данных на квартили. В частности, это разница между верхним квартилем (Q3) и нижним квартилем (Q1) набора данных.

Чтобы рассчитать IQR, нужно сначала расположить данные в порядке от наименьшего к наибольшему. Затем находится медиана (Q2) набора данных и нижний квартиль (Q1), который является медианой нижней половины набора данных (т. е. точек данных ниже медианы), а верхний квартиль (Q3) является медианой верхней половины набора данных (т. е. точек данных выше медианы). Наконец, IQR рассчитывается как разница между Q3 и Q1.

IQR часто используется в качестве меры изменчивости или разброса в наборе данных и считается надёжным статистическим показателем, поскольку он менее чувствителен к выбросам или экстремальным значениям, чем диапазон или стандартное отклонение. Он также обычно используется в столбчатых диаграммах для визуализации распределения набора данных.

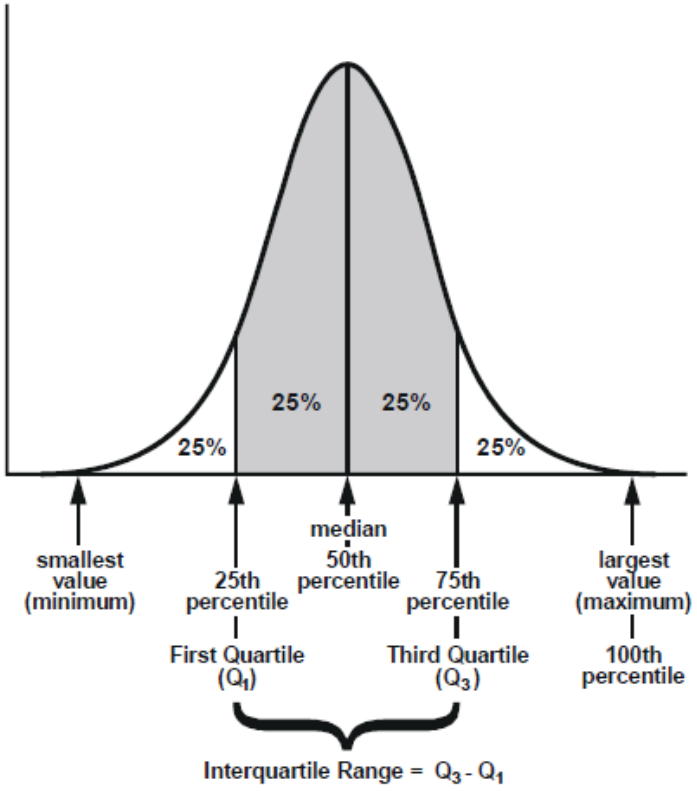
Предположим, в некоторых случаях вы сравниваете две группы. Вы уже рассчитали центральную тенденцию ваших данных, то есть среднее значение, медиану и моду

для обеих групп. Иногда может случиться так, что среднее, медиана и мода одинаковы для обеих групп.

Межквартильный диапазон (IQR)

Межквартильный размах даёт ещё одну меру изменчивости. Это лучшая мера дисперсии, чем **диапазон**, потому что **она не учитывает экстремальные значения**. Он поровну делит распределение на четыре равные части, называемые квартилями. Первые 25% — это 1-й квартиль (Q1), последние — 3-й квартиль (Q3), а средние — 2-й квартиль (Q2).

2-й квартиль (Q2) делит распределение на две равные части по 50%. Итак, в основном это то же самое, что и **медиана.**



Межквартильный размах — это расстояние между третьим и первым квартилем, или, другими словами:

IQR = Q3 - Q1

Как рассчитать межквартильный размах (IQR):

Шаг 1. Упорядочить датасет от низкого к высокому

Шаг 2. Найдите медиану или, другими словами, Q2

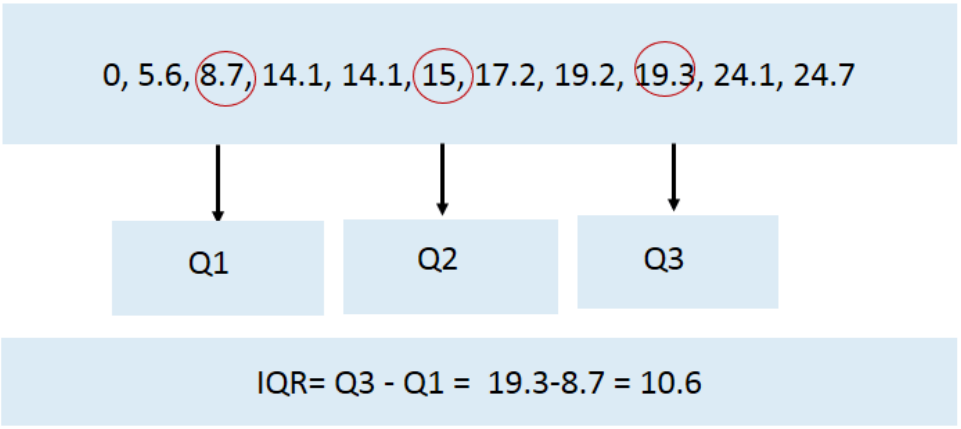
Шаг 3. Затем найдите Q1, взглянув на медиану левой части Q2

Шаг 4. Аналогичным образом найдите Q3, взглянув на медиану справа от Q2

Шаг 5. Теперь вычтите Q1 из Q3, чтобы получить IQR

Пример расчёта IQR

Рассмотрим приведённый ниже пример, чтобы получить чёткое представление.



Рассмотрим другой пример, чтобы лучше понять.

Рассмотрим следующие числа: 1, 3, 4, 5, 5, 6, 7, 11. Q1 — среднее значение в первой половине набора данных. Поскольку в первой половине набора данных имеется чётное количество точек данных, среднее значение является средним из двух средних значений; то есть Q1 = (3 + 4)/2 или Q1 = 3,5. Q3 — среднее значение во второй половине набора данных. Опять же, поскольку вторая половина набора данных имеет чётное количество наблюдений, среднее значение является средним из двух средних значений; то есть Q3 = (6 + 7)/2 или Q3 = 6,5. Межквартильный диапазон равен Q3 минус Q1, поэтому IQR = 6,5 — 3,5 = 3.

Преимущество IQR

- Основное преимущество IQR заключается в том, что на него не влияют выбросы, поскольку он не принимает во внимание наблюдения ниже Q1 или выше Q3.
- Тем не менее может быть полезно искать возможные выбросы в вашем исследовании.
- Как правило, наблюдения можно квалифицировать как выбросы, если они лежат более чем на 1,5 IQR ниже первого квартиля или на 1,5 IQR выше третьего квартиля.

Выбросы = Q1 — 1,5 * IQR или

= Q3 + 1,5 * IQR

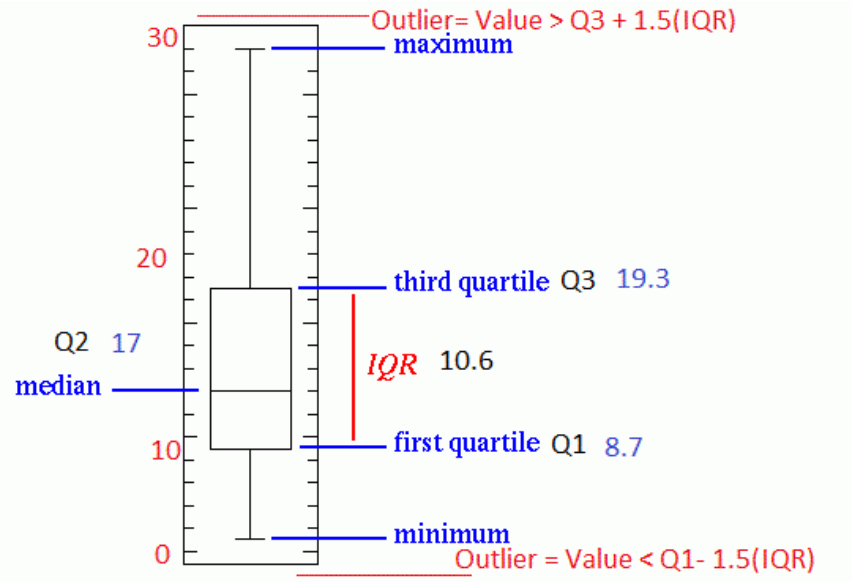
Что такое блочные диаграммы?

Блочные диаграммы — это графические представления, которые обычно используются для отображения распределения набора данных и его сводной статистики. Блочные диаграммы отображают медиану, квартили, диапазон и выбросы набора данных. Центральный прямоугольник представляет IQR, а медиана показана линией внутри прямоугольника. Нижние и верхние усы представляют минимальное и максимальное значения набора данных, которые не считаются выбросами, а любые точки за пределами усов отображаются как отдельные точки, представляющие выбросы. Блочные диаграммы полезны для быстрой визуализации центральной тенденции и изменчивости набора данных и выявления любых экстремальных значений.

Итак, блочная диаграмма — это график, который в основном используется, когда вы описываете центр и изменчивость ваших данных.

Это также полезно для обнаружения выбросов в данных.

Внимательно наблюдайте за приведённым выше первым примером IQR, когда он нанесён на блочную диаграмму.



Диаграмму выше также можно описать так:

- Границами коробки служат первый и третий квартили (25-й и 75-й процентили соответственно).
- Линия в середине ящика — медиана (50-й процентиль).
- Концы усов — края статистически значимой выборки (без выбросов).

Дисперсия и стандартное отклонение

Дисперсия и стандартное отклонение — это статистические меры, которые используются для описания степени изменчивости или разброса в наборе данных.

Ранее мы обсудили, что **размах, межквартильный размах (IQR) и прямоугольная диаграмма** очень полезны для измерения **изменчивости данных**.

Есть два других вида изменчивости, которые в статистике очень часто используются для исследований.

1. **Дисперсия**
2. **Среднеквадратичное отклонение**

Дисперсия измеряет, насколько далеко набор чисел разбросан от их медианного или среднего значения. Он рассчитывается путём получения среднего значения квадратов разностей между каждым числом и средним значением набора данных. Более высокая дисперсия указывает на то, что числа более разбросаны по сравнению со своим средним значением.

Стандартное отклонение представляет собой квадратный корень из дисперсии и обеспечивает меру отклонения данных от среднего значения. Он выражается в тех же единицах, что и данные, и является более интуитивной мерой разброса данных, поскольку имеет тот же масштаб.

И дисперсия, и стандартное отклонение являются важными показателями во многих областях статистики, включая проверку гипотез, контроль качества и анализ данных.

Почему дисперсия и стандартное отклонение являются хорошими показателями изменчивости?

Поскольку дисперсия и стандартное отклонение учитывают **все значения переменной** для расчёта изменчивости ваших данных.

Существует два типа дисперсии и стандартного отклонения с точки зрения выборки и совокупности.

Дисперсия

Вот формула для расчёта дисперсии выборки и генеральной совокупности и стандартного отклонения. Есть небольшая разница, внимательно наблюдайте за ними.

Для выборки

$$variance = s^2 = \Sigma \frac{(x - \bar{x})^2}{n} - 1$$

$$standard\ deviation = s = \sqrt{s^2}$$

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n - 1}$$

Для генеральной совокупности

$$variance = \sigma^2 = \Sigma \frac{(x - \bar{x})^2}{n} - 1$$

$$standard\ deviation = \sigma = \sqrt{\sigma^2}$$

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n}$$

Где:

- X — индивидуальное значение
- N — количество значений генеральной совокупности
- x̄ — среднее значение генеральной совокупности

Как рассчитать дисперсию шаг за шагом:

1. Вычислите среднее значение: x̄
2. Вычтите среднее из каждого наблюдения: X - x̄
3. Возведите в квадрат каждое из полученных наблюдений: (X - x̄)^2
4. Сложите эти квадраты результатов вместе
5. Разделите эту сумму на количество наблюдений n (в случае совокупности), чтобы получить дисперсию **S2**. Если вы рассчитываете выборочную дисперсию, разделите на n - 1
6. Используйте положительный квадратный корень, чтобы получить стандартное отклонение **S**

X	$X - \bar{X}$	$(X - \bar{X})^2$
0	-15	225
24.1	9.1	82.81
5.6	-9.4	88.36
14.1	-0.9	0.81
17.2	2.2	4.84
8.7	-6.3	39.69
19.2	4.2	17.64
14.1	-0.9	0.81
27.7	12.7	161.29
15	0	0
19.3	4.3	18.49
		639.74

Здесь:

N = 11

N - 1 = 10

Среднее ($\bar{x}$) = 15

Выборочная дисперсия (s^2) = 639,74/10 = 63,97

Генеральная совокупность (σ^2) = 639,74/11 = 58,16

S = 8,00

σ = 7,6

Интуиция

1. Если дисперсия высока, это означает, что у вас большая изменчивость в вашем наборе данных. Другими словами, мы можем сказать, что вокруг вашего среднего значения разбросано больше значений.
2. Стандартное отклонение представляет собой среднее расстояние наблюдения от среднего значения.
3. Чем больше стандартное отклонение, тем больше изменчивость данных.

Свойства дисперсии

- Она всегда неотрицательная, поскольку каждый член суммы дисперсии возводится в квадрат, и поэтому результат либо положительный, либо нулевой.
- Дисперсия всегда имеет квадратные единицы. Например, дисперсия набора гирь, оценённая в килограммах, будет дана в килограммах в квадрате. Поскольку дисперсия генеральной совокупности возводится в квадрат, мы не можем напрямую сравнивать её со средним значением или самими данными.

Среднеквадратичное отклонение

Стандартное отклонение — это мера того, насколько разбросаны числа. Его символ — σ (греческая буква «сигма») для стандартного отклонения совокупности и S для стандартного отклонения выборки. Это квадратный корень из дисперсии.

Свойства стандартного отклонения

- Оно описывает квадратный корень из среднего значения квадратов всех значений в наборе данных и также называется среднеквадратичным отклонением.
- Наименьшее значение стандартного отклонения равно 0, поскольку оно не может быть отрицательным.
- Когда значения данных группы схожи, стандартное отклонение будет очень низким или близким к нулю. Но когда значения данных меняются друг относительно друга, стандартное отклонение будет высоким или далеко от нуля.

Генеральная совокупность против выборочной дисперсии и стандартного отклонения

Основная задача логической статистики (или оценки и прогнозирования) состоит в том, чтобы составить мнение о чём-либо, используя только неполную выборку данных.

В статистике очень важно различать совокупность и выборку. Совокупность определяется как все члены (например, частота, цена, годовой доход) определённой группы. Генеральная совокупность — вся группа.

Выборка — это часть совокупности, которая используется для описания характеристик (например, среднего значения или стандартного отклонения) всей совокупности. Размер выборки может быть меньше 1%, 10% или 60% генеральной совокупности, но это никогда не вся совокупность целиком. Поскольку и выборка, и совокупность — не одно и то же, поэтому в их формуле есть небольшая разница.

Может возникнуть вопрос, почему во время расчёта дисперсии, мы возводим разницу в квадрат?

Чтобы избавиться от негативов, чтобы негатив и позитив не отменяли друг друга при сложении.

+5 -5 = 0

Нормальное распределение, биномиальное распределение и распределение Пуассона

Нормальное распределение, биномиальное распределение и распределение Пуассона — три важных распределения вероятностей, используемых в статистике и анализе данных.

Нормальное распределение, также известное как распределение Гаусса, представляет собой непрерывное распределение вероятностей, которое часто используется для описания природных явлений, таких как рост и вес. Для него характерна колоколообразная кривая, симметричная и центрированная вокруг среднего значения. Стандартное отклонение определяет ширину кривой и описывает изменчивость данных.

Биномиальное распределение — это дискретное распределение вероятностей, которое используется для моделирования количества успешных результатов в фиксированном числе независимых испытаний. Оно характеризуется двумя параметрами — вероятностью успеха в одном испытании и количеством испытаний. Биномиальное распределение широко используется в таких областях, как финансы, биология и контроль качества.

Распределение Пуассона — это дискретное распределение вероятностей, которое используется для моделирования количества событий, происходящих за фиксированный интервал времени. Он характеризуется одним параметром — средним количеством событий в единицу времени. Распределение Пуассона часто используется в таких областях, как эпидемиология, финансы и телекоммуникации.

Нормальное распределение или распределение Гаусса, кривая нормального распределения

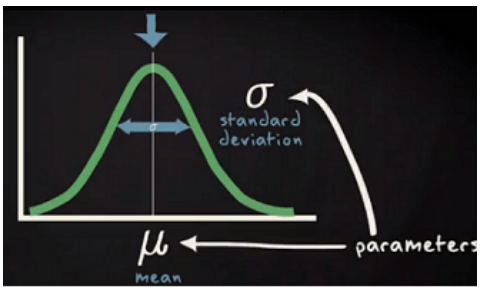
В теории вероятностей нормальное распределение или распределение Гаусса является очень распространённым непрерывным распределением вероятностей. Нормальное распределение иногда неофициально называют кривой нормального распределения.

Плотность вероятности нормального распределения:

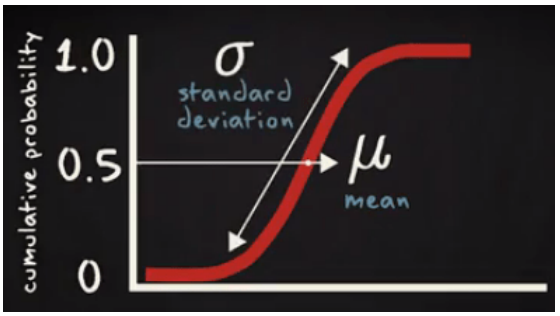
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

μ — среднее значение или ожидание распределения

σ^2 — это дисперсия



Кумулятивное нормальное распределение вероятностей будет выглядеть так, как показано на диаграмме ниже.



Свойства нормального распределения:

- Среднее значение, мода и медиана равны
- Кривая симметрична в центре (т.е. вокруг среднего, μ)

- Ровно половина значений находится слева от центра и ровно половина значений — справа
- Общая площадь под кривой равна 1

Расчёт вероятности нормального распределения

Функция плотности вероятности или p.d.f указывает вероятность на единицу случайной величины. Вот пример p.d.f ежедневного времени ожидания водителем такси компании Uber. По оси X показано ежедневное время ожидания и вероятность этого ожидания по оси Y в час.

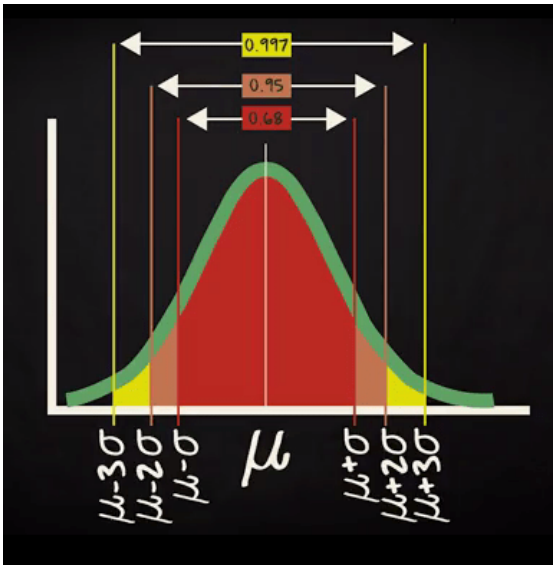
Предположим, что один водитель такси Uber хочет знать вероятность того, что он будет ждать более 7 часов в день.

Тогда его заинтересует жёлтая поверхность, показанная выше. На основе этого графика можно оценить площадь. То же самое вы можете получить ниже кумулятивной кривой вероятности.

Вероятность ожидания более 7 часов будет рассчитываться с использованием дополнительного правила 1 - P. Поскольку нас интересует вероятность ожидания более 7 часов, а сумма всех вероятностей равна 1, то вероятность ожидания более 7 часов это 1 - P в точке графика от X = 7. Таким образом, P следует вычесть из 1, чтобы получить желаемый результат.

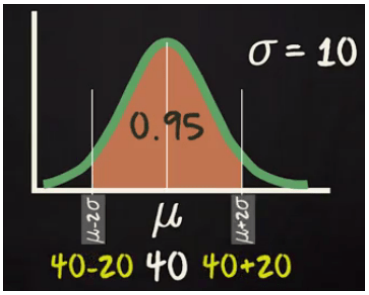
Распределение в форме колокола и эмпирическое правило

Если распределение имеет форму колокола, то предполагается, что около 68% элементов имеют z-показатель от -1 до 1; около 95% имеют z-показатель от -2 до 2; и около 99% имеют z-показатель от -3 до 3.



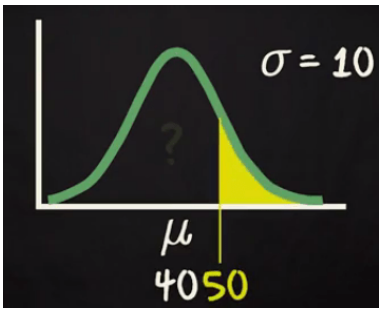
Предположим, что время, которое вы тратите в будние дни на поездки, определяется нормальным распределением со средним значением = 40 минут и SD = 10 минут.

Каков будет ваш диапазон времени в пути в течение 95% ваших дней в неделю?



Как вы знаете, 95% будут находиться в пределах 2 стандартных отклонений от вашего среднего значения. Таким образом, диапазон будет от (40 - 20) = 20 до (40 + 20) = 60 минут.

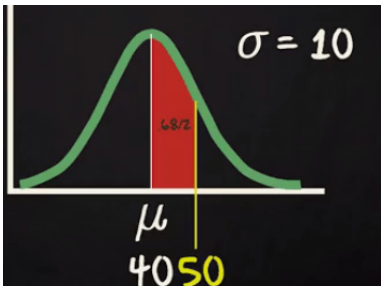
Теперь ещё один вопрос, на который вы хотите ответить: какова вероятность того, что вы будете путешествовать более 50 минут?



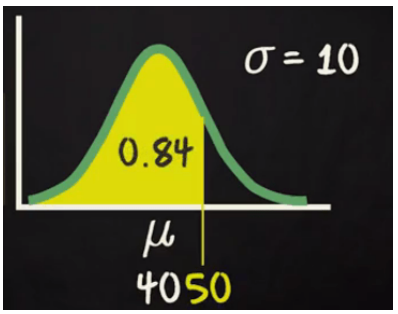
На самом деле вас интересует жёлтая поверхность, представленная на диаграмме выше. Вы знаете, что нормальное распределение симметрично. Таким образом, половина вероятности приходится на одну сторону от среднего, а другая половина — на другую сторону от среднего.

Поскольку SD = 10. Таким образом, одно стандартное отклонение будет составлять диапазон от 30 до 50.

Вы уже знаете, что для левой стороны выше 40 вероятность равна 0,5. Теперь, если вы рассчитаете вероятность от 40 до 50, она будет равна половине 1 стандартного отклонения, т. е. 0,68/2 = 0,34.



Таким образом, вероятность путешествовать менее 50 минут = 0,5 + 0,34 = 0,84



Но вас интересует время в пути более 50 минут, поэтому оно будет 1- 0,84 = 0,16.

Испытание Бернулли и биномиальное распределение

Каждая случайная величина имеет соответствующее распределение вероятностей. Распределение вероятностей применяет теорию вероятности для

описания поведения случайной величины. Дискретная случайная величина X имеет конечное число возможных целочисленных значений. Распределение вероятностей X перечисляет значения и их вероятности в таблице.

Value of X	x_1	x_2	x_3	...	x_k
Probability	p_1	p_2	p_3	...	p_k

- Каждая вероятность p представляет собой число от 0 до 1.
- Сумма вероятностей должна быть равна 1.

Это свойство мы уже изучали ранее. Теперь мы обсудим наиболее важную вероятность для дискретной случайной величины — биномиальное распределение. Перед этим необходимо знать об испытании Бернулли.

Испытание Бернулли или биномиальное испытание

Испытание Бернулли (или биномиальное испытание) — это случайный эксперимент с ровно двумя возможными исходами, «успехом» и «неудачей», в котором вероятность успеха одинакова при каждом проведении эксперимента**.**

- Событие (или испытание) приводит только к одному из двух взаимоисключающих исходов — успех/неудача.
- Вероятность успеха известна, P(success) = π

Примеры испытания Бернулли или биномиального испытания

- Одно подбрасывание монеты (орёл или решка), P(орёл) = π = 0,5

- Выживаемость человека после операции АКШ, P(выживаемость) = π = 0,98

- Выберите человека из населения Индии, P (ожирение) = π = 0,31.

Биномиальное распределение

Распределение называется биномиальным, если выполняются следующие условия:

1. Каждое испытание имеет бинарный результат (один из двух результатов помечен как «успех»).
2. Вероятность успеха известна и постоянна для всех испытаний.

3. Количество испытаний указано.
4. Испытания независимы. То есть результат одного испытания не влияет на результат последующих испытаний.

Если все вышеперечисленные условия соблюдены, то биномиальное распределение описывает вероятность X успехов в n испытаниях. Классический пример биномиального распределения — количество орлов (X) при n подбрасываниях монеты.

Обозначение для биномиального распределения X ~ B (n, л), который читается как «X распределён биномиально с n испытаниями и вероятностью успеха в одном испытании, равной π».

Формула биномиального распределения

Используя эту формулу, можно рассчитать распределение вероятностей биномиальной случайной величины X, если известны n и π.

$$P(X) = \frac{n!}{X!(n-X)!}\pi^{X}(1-\pi)^{n-X}$$

n! называется «n факториалом» = n(n-1)(n-2) . . .(1)

P(X) = количество сценариев * один сценарий

Первый факторный член даёт количество сценариев, а второй член описывает вероятность успеха в степени числа успехов и вероятность неудачи в степени числа неудач.

Пример **биномиального распределения**

Какова вероятность того, что при 6 подбрасываниях монеты выпадет 2 орла?

- Успех = «орёл»
- n = 6 испытаний
- π = 0,5
- X = количество орлов в 6 бросках, здесь 2.

- X имеет биномиальное распределение с n = 6 и π = 0,5.
- X ~ B (6, 0,5)

$$P(X=2)=\frac{6!}{2!(6-2)!}0.5^2(1-0.5)^{6-2}=15*0.5^6=0.234$$

Таким образом, вероятность выпадения двух орлов равна 0,234.

Рассмотрим другой пример:

Какова вероятность того, что в выборке из 8 больных с сердечным приступом умрут 2 пациента, если вероятность смерти от сердечного приступа = 0,03?

Предположим, что вероятность смерти одинакова для всех больных.

- Смерть от сердечного приступа — бинарная переменная (Да или Нет)
- «Успех» в данном случае определяется как смерть от сердечного приступа
- n = количество «испытаний» = 8 пациентов
- π = 0,03 = вероятность успеха
- X = количество смертей. Здесь X = 2
- X ~ B (8, 0,03)

Если вы будете следовать той же формуле, вы получите P(x=2) = 0,021.

Распределение Пуассона

Другим распределением вероятностей для дискретных переменных является распределение Пуассона. Распределение Пуассона используется для определения вероятности того, что количество событий произойдёт за определённое время или в определённом пространстве. Оно было назван в честь Симеона Д. Пуассона (1781—1840), французского математика.

Примеры событий в пространстве или во времени:

- количество клеток в заданном объёме жидкости;
- количество звонков в час на линию помощи;

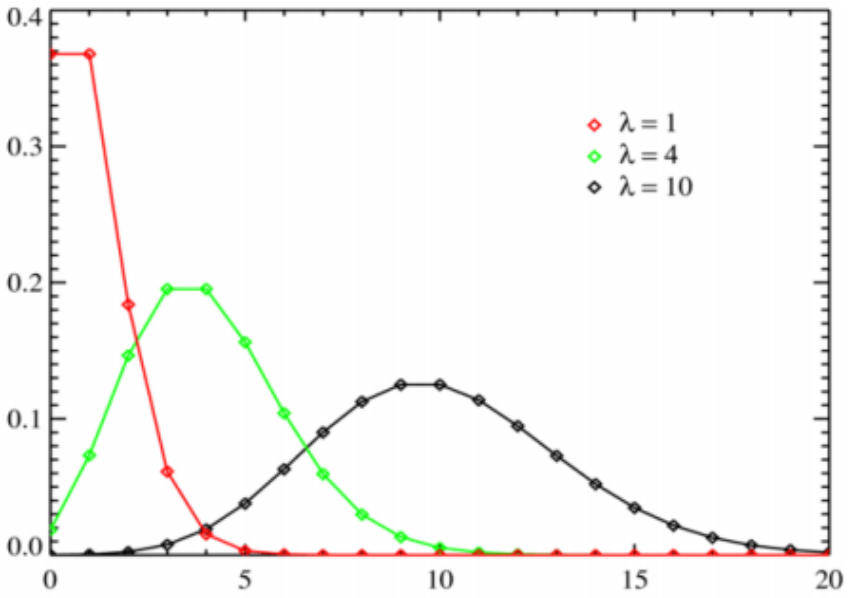
- количество коек в отделении неотложной помощи заполнено / 24 часа.

Подобно биномиальному распределению и нормальному распределению существует множество распределений Пуассона.

- Каждое распределение Пуассона определяется средней скоростью, с которой происходит событие.
- Скорость обозначается λ.
- λ = «лямбда», греческая буква «L» — для распределения Пуассона есть только один параметр.

Вероятность того, что в указанном пространстве или времени имеется ровно X вхождений, равна:

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!}$$



Горизонтальная ось представляет собой индекс X. Функция определена только при целых значениях X. Соединительные линии являются только направляющими для глаза и не указывают на непрерывность. Обратите внимание, что по мере увеличения λ распределение начинает напоминать нормальное распределение.

- Если λ равно 10 или больше, нормальное распределение является разумным приближением к распределению Пуассона.

- Среднее значение и дисперсия для распределения Пуассона одинаковы и равны λ.
- Стандартное отклонение распределения Пуассона равно квадратному корню из λ.

Пример **распределения Пуассона**

В крупную городскую больницу каждый понедельник поступает в среднем 80 пациентов в отделение неотложной помощи. Какова вероятность того, что их будет больше 100?

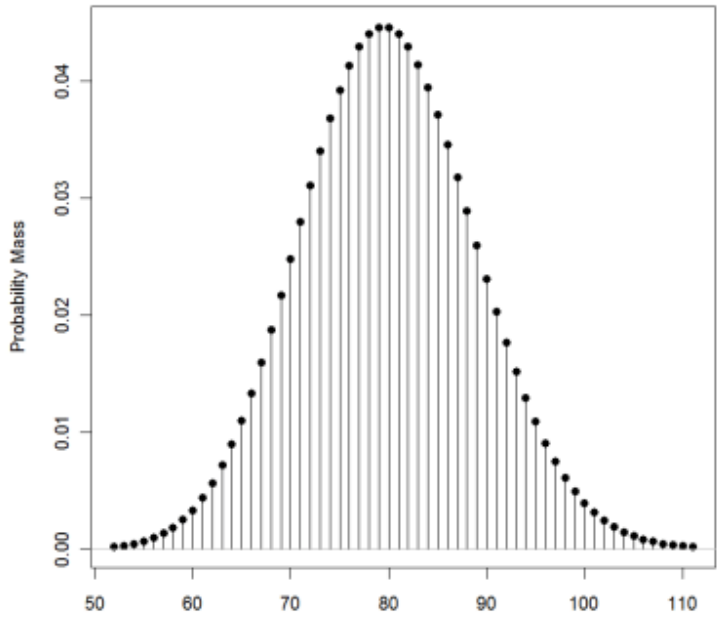
Если мы предположим, что λ = 80 и x = 100, то мы получим значение вероятности, равное 0,01316885.

Чтобы получить тот же результат, мы можем использовать нормальное приближение, а затем получить значение вероятности.

Отделение неотложной помощи принимает в понедельник?

- λ — скорость допуска / день в понедельник = 80
- мы можем использовать нормальное приближение, так как λ > 10

Нормальное приближение имеет среднее значение = 80 и SD = 8,94 (квадратный корень из 80 = 8,94).



Теперь мы можем использовать тот же способ, которым вычисляем p-значение для нормального распределения. Если вы сделаете это, вы получите значение 0,01263871, что очень близко к 0,01316885, которое мы получаем непосредственно из формулы Пуассона. Основная цель здесь — показать вам, как нормальное приближение работает для распределения Пуассона.

Понимание этих распределений и их свойств необходимо для многих приложений в таких областях, как финансы, инженерия и наука. Анализируя данные с использованием этих распределений, мы можем делать прогнозы, а также выводы о реальных явлениях.

Заключение

Сегодня мы поговорили о том, почему важно знать основы статистического анализа данных и как применять его на практике. Мы рассмотрели основные понятия статистики и то, как они характеризуют наш набор данных. Это лишь вершина айсберга, и если вы захотите углубиться статистический анализ, вас ждёт долгий, но интересный путь.
Я желаю вам успехов. До новых встреч!