2.1 Что такое качественные и количественные данные?

Если кратко, то количественные данные обеспечивают числа для общего представления о той или иной ситуации, связанной с продукцией, услугой, уровнем обслуживания и т. д. А качественные данные эти общие моменты уточняют — делают более развернутыми и подробными.

Категориальные. Это данные с ограниченным числом уникальных значений или категорий (например, пол или религия). Категориальные переменные могут быть текстовыми или числовыми, в которых категории закодированы числовыми кодами (например, 0 = Женский, а 1 = Мужской). Также эти данные называются качественными данными.

Категориальные переменные могут быть либо номинальные, либо порядковые

Номинальная. Переменную можно рассматривать как номинальную, когда ее значения представляют категории без естественного упорядочения, например, подразделение компании, где работает наемный сотрудник. Примеры номинальных переменных включают регион, почтовый индекс или религию

Порядковая. Переменную можно рассматривать как порядковую, когда ее значения представляют категории с некоторым естественным для них упорядочением, например, уровни удовлетворенности обслуживанием от крайней неудовлетворенности до крайней удовлетворенности. Примеры порядковых переменных включают баллы, представляющие степень удовлетворенности или уверенности, или баллы, оценивающие предпочтение.

Категориальные переменные определяют категории в диаграмме, обычно, чтобы показать на графике отдельные графические элементы или группы элементов. Количественные переменные часто подытоживаются внутри категорий категориальных переменных.

К примеру, на диаграмме с переменной дохода по категориям пола будет представлено среднее значение дохода для мужчин и среднее значение дохода для женщин. Исходные значения количественных переменных могут быть представлены при помощи диаграммы рассеяния. Например, диаграмма рассеяния может для каждого наблюдения показывать зарплату в настоящее время и зарплату при приеме на работу. Категориальную переменную можно использовать для группировки наблюдений по полу.

2.2

Шкала измерения в статистике — это способ представления переменных (признаков, атрибутов) и их группировки в различные категории. Она определяет характер значений, присвоенных переменным в наборе данных.

Шкала измерения используется для определения и описания переменных в наборах данных. Она определяет методы, которые могут быть использованы для их анализа. В зависимости от типа анализируемых данных определяется тип шкалы измерения. Выделяют 4 основных вида шкал: номинальная, порядковая, интервальная и шкала отношений

Шкалы измерения используются для представления как качественных, так и количественных данных. Номинальная и порядковая шкалы используются для измерения качественных данных, в то время как интервальная и шкала отношений используются для измерения количественных.

Основными свойствами шкал измерений являются:

Идентифицируемость — возможность присвоения числовых значений каждой переменной в наборе данных. Например, в анкете запрашивается пол респондента — «Мужчина» и «Женщина». Для этих двух значений могут быть определены идентифицирующие значения — 1 и 2 соответственно. К таким значениям не могут быть применены арифметические операции, потому что они служат только для идентификации, а не описания.

Величина (магнитуда) — это размерность шкалы измерения, где значения могут быть упорядочены от наименьшего к наибольшему. Например, место в соревновании распределяется от 1-го, 2-го, 3-го до наименьшего.

Равенство интервалов — означают, что шкала имеет стандартизированный порядок, т.е. разность между двумя любыми соседними уровнями шкалы одинакова. Упорядоченность шкалы не гарантирует равенство интервалов. Например, в примере с местами в соревновании, каждая позиция имеет одинаковую разницу интервалов равную 1, но при этом 2-й участник может финишировать на 20 секунд позже, чем первый, а третий на 40 секунд позже, чем второй.

Абсолютный ноль — естественное и однозначное присутствие нулевой точки, изменение которой невозможно. Данная точка характеризует отсутствие измеряемого признака. Например, 0 градусов по Кельвину является абсолютным нулем на шкале, а 0 градусов по Цельсию — нет, т.к. за него принято одно из произвольно взятых физических явлений — температура плавления льда.

Номинальная шкала (категориальная, наименований) — это шкала измерения, которая используется для идентификации. Она является самой «слабой» из четырех видов шкал в смысле возможности обработки данных. Она присваивает номера атрибутам для удобства идентификации, но может использоваться только как метка. Единственный вид статистического анализа, который можно выполнить с использованием номинальной шкалы, это вычисление процентных долей и частот. Данные в номинальной шкале можно проанализировать графически с помощью гистограммы и круговой диаграммы. Например, если измерить атрибут «Товар» в номинальной шкале, то она будет выглядеть так: 1 — мороженное; 2 — соки; 4 — выпечка. При этом значения шкалы не определяют какого-либо приоритета между товарами, а просто идентифицируют их. Очевидно, что такая шкала может использоваться только для самого просто анализа.

Порядковая шкала (ординальная, ранговая) — предполагает ранжирование (упорядочивание) значений переменной в зависимости от масштабирования. Атрибуты в порядковой шкале обычно располагаются в порядке возрастания или убывания. Порядковая шкала может быть использована в исследованиях рынка, рекламы и опросов удовлетворенности клиентов. Она использует квалификаторы, такие как «очень», «высоко», «больше», «меньше» и т. д. В порядковой шкале можно использовать для статистического анализа такие статистики как медиана, но не среднее значение. Существуют и другие виды анализа, которые могут быть проведены с использованием порядковой шкалы. Например, компания-разработчик ПО может провести опрос пользователей для оценки нового приложения в шкале: «Отлично», «Очень хорошо», «Хорошо», «Плохо», «Очень плохо». Атрибуты в этом примере перечислены в порядке убывания.

Интервальная шкала (разностей) — это шкала, в которой уровни упорядочены, а интервалы между ними равны. Её можно рассматривать как расширение порядковой шкалы. Основным отличием является свойство равных интервалов. Интервальная шкала не только позволяет однозначно определить, какое значение больше (меньше), но и на сколько. Кроме того, в отличие от порядковой и номинальной шкал, в интервальной могут выполняться арифметические операции. Типичным примером является измерение температуры по шкале Фаренгейта. Интервальную шкалу можно использовать при расчете среднего значения, медианы, моды, стандартного отклонения и других статистик.

Шкала отношений (абсолютная) является «наивысшим» уровнем представления данных. Она может рассматриваться как расширение интервальной шкалы, и, следовательно, удовлетворяет четырем свойствам шкалы измерения: идентифицируемостью, величиной, равноинтервальностью и наличием абсолютного нуля. Примерами шкал отношения являются длина, вес, время и т. д. В исследованиях рынка примерами шкалы отношений являются цена, количество клиентов, суммы продаж и т. д. Она широко используется в маркетинге и рекламе. Шкала отношений совместима со всеми методами статистического анализа и может использовать как показатели центральной тенденции (среднее значение, медиана, мода и т. д.), так и разброса значения (дисперсии, размаха, стандартного отклонения и т. д.).

3.4

Нормальное распределение - это когда значения случайной величины будут сгруппированы вокруг среднего значения, и чем дальше от среднего значения, тем меньше вероятность того, что такое значение появится.

Биномиальное распределение — это распределение вероятностей в последовательности экспериментов, где эксперимент даёт двоичный результат. При этом результаты независимы друг от друга

Равномерное распределение характеризуется наличием **постоянной прямой вероятности** значения на определенном интервале, а также оно связано с событиями, которые в равной степени вероятны.

распределение Пуассона – это распределение количества: количества раз того, как что-то произойдёт. Оно параметризуется не вероятностью p и количеством испытаний n, но средней интенсивностью λ, что, в аналогии с биномиальным, просто постоянное значение np. Распределение Пуассона – то, о чём надо вспоминать, когда идёт речь о подсчёте событий за определённое время при постоянной заданной интенсивности.

Когда есть что-то, типа прихода пакетов на роутер или появления покупателей в магазине или что-то, ожидающее в очереди – думайте «Пуассон»

t-Распределение Стьюдента – это основа t-теста, который многие нестатистики изучают в других областях. Оно используется для предположений о среднем нормального распределения и так же стремится к нормальному распределению с увеличением своего параметра. Отличительная особенность t-распределения – его хвосты, которые толще, чем у нормального распределения.

распределение хи-квадрат – распределение сумм квадратов нормально-распределенных величин. На этом распределении построен тест хи-квадрат, который сам основан на сумме квадратов разниц, которые должны быть нормально распределены.

Распределение χ 2 (хи-квадрат) используется для оценки следующих проблем:

Подходит ли набор данных определенному виду распределения

Одинаковы ли распределения двух совокупностей

Независимы ли друг от друга два события

Есть ли в совокупности другая изменчивость

Кривая стремится вправо.