Статистика и типы данных

Когда стоит задача проанализировать данные какого-либо эксперимента очень важно понимать какого типа данных является результат. Если, например, мы измеряли температуру, она может изменяться в непрерывно, а между двумя соседними значениями можно вычислить или определить промежуточное значения. Эквивалент таких данных - числовая прямая - и они называются **интервальными непрерывными**. Тем не менее, интервальные данные не всегда изменяются плавно и непрерывно. Рассмотрим ту же температуру. Она изменяется от абсолютного нуля, то есть, ограничена снизу. Или, к примеру, изменение угла. Его значение может принимать любую величину, но за 360 градусов следует 0. Таким образом, мало того, что данные ограничены сверху и снизу, так они еще изменяются циклично.

Вот другой пример, количество посетителей ресторана, кафе или торгового центра. Допустим, в первый день магазин посетило 154 человек, а в другой день - 105. Легко представить промежуточные значения. Однако, если взять два соседних числа, то промежуточное значение может быть не определено, например 105 и 106. Очевидно, что количество человек - величина целая. У таких величин есть отношение порядка, но вот промежуточные значения есть не всегда. Такие данные называются **дискретными**. Они тоже являются интервальными, но изменяются небольшими «скачками».

Как это понимать и применять?

Статистические тесты делятся на две группы: **параметрические** и **непараметрические**. Параметрические тесты предназначены для обработки параметрических данных, Чтобы данные были параметрическими необходимо выполнение следующих усовий

1. распределение данных близко к нормальному
2. выборка содержит более 30 элементов
3. данные интервальные непрерывные

если хотя бы одно из этих условий не выполняется, данные считаются непараметрическими и обрабатываются непараметрическими тестами.

Основное достоинство непараметрических тестов - это как раз способность работать с непараметричесими (не идеальными) данными. С другой стороны, параметрические тесты обладают большей мощностью, то есть дают большую вероятность увидеть существующую закономерность, нежели при использовании непараметрических тестов. Этому есть простое объяснение: непараметрические данные имеют способность скрывать существующие закономерности именно за счет объединения нескольких значений в группы. На распределение мы повлиять, как правило, не можем, тем не менее иногда за счет преобразований можно привести данные к непрерывному виду и улучшить распределение. Что мы можем сделать, так это иметь достаточно большую выборку, а также работать с непрерывными данными.

В R и Python интервальные данные представлены числами

Есть еще один интересный тип данных, а именно **шкальный**. Например, вы хотите определить качество обслуживания клиентов или удобство мебели. Как удобство, так и степень

удовлетворенности сервис носит субъективный характер, и чтобы их учесть в статистике, мы должны что-то сделать с такими данными, а именно шкалировать. То есть, каждому баллу сопоставить описание. В такой системе баллы можно даже ранжировать.

Число, которорым обозначено значение шкалы более чем условное, зато с ним определены отношения порядка и более того, подобие непрерывности. Например, если безупречный сервис обозначить цифрой 5, а менее удобный - цифрой 4, то в принципе, можно представить сервис с оценкой 4.5. Именно поэтому, к шкальным данным применимо много из того, что используется с интервальными данными, главное не забывать, что значение шкалы весьма условное. Больше всего трудностей возникает, когда данные представлены в разных шкалах и не всегда урвется перевести их в одну шкалу. По-умолчанию, R распознает шкальные данные как обычный числовой вектор, однако иногда используется тип упорядоченный фактор.

Если стоит задача создания шкальных данных из непрерывных, можно воспользоваться командой cut(),

Для статистического анализа шкальных данных всегда требуются непараметрические методы. Если же хочется использовать параметрические методы, необходимо так спланировать эксперимент, чтобы в результате получить интервальные данные. Например, при исследовании размеров улиток не делить их на маленькие. средние или большие, а использовать линейку для измерения длины и ширины. Иногда получение непрерывных данных требует использования специального труднодоступного оборудования. Например, если хотим изучить влияние цвета листьев на эффективность фотосинтеза, то для измерения цвета листьев может потребоваться использовании спектрофотометра (чтобы перевести цвет в длину волны). Здесь можно выйти из положения, если перекодировать данные на этапе обработки. Например, можно перекодировать цвет в шкалу RGB.

Еще один пример перекодировки шкальных данных. Допустим, мы хотим исследовать высоту деревьев в городе с запада на восток. Можно задать направление в котором интересно провести анализ и переименовать деревья от самого левого к самому правому. Тогда мы получим шкальные данные, которые можно обработать непараметрическими методами. Или пометить каждое дерево своими географическими координатами или как расстояние от самого левого дерева. Тогда получим интервальные данные и сможем их обработать параметрическими методами.

Еще один тип данных с которым можно встретиться при обработке статистических экспериментов - **номинальный**. Такие данные часто называют категориальными. Их нельзя упорядочить. В каком-то смысле они еще дальше от числовых данных нежели шкальные. Даже если и удастся представить их числами, то вот упорядочить или представить промежуточные значения точно не получится. (нампример пол; «Мужской» и «Женский» мы можем обозначит 0 и 1, но вот получить промежуточное значение не возможно).

Особый класс номинальных данных это парные данные, то есть имеющие всего два значения. Такие данные иногда можно упорядочить. В них можно перекодировать практически любой тип данных, иногда с частичной потерей информации. Посе этого к ним можно применять специальные методы анализа, например, логистическую регрессию.

Номинальные данные представляются факторами в среде R.

Получилось много букв, но тем не мене это основы и их нужно знать, чтобы не попасть в просак применяя параметрические методы к непараметрическщм данным или анализируя результаты в различных шкалах. А такие ошибки часто встречаются среди новичков.

Использованная литература

А.Б. Шипунов и др. Наглядная статистика. Используем R! *Москва, Издательство «ДМК», 2017 г, 296 с.*