Цензурированные данные (Censored data)

> Синонимы: Неполные данные, Урезанные данные, Trimmed data

В статистике цензурированными называют данные, в которых значения признаков известны только на некоторых интервалах наблюдения. Примером цензурирования может служить взвешивание объектов на весах с ограниченной шкалой, скажем, до 100 кг. Тогда для объектов с весом меньше 100 кг можно узнать точный вес, а для всех прочих можно сказать только, что вес больше 100 кг. В этом случае говорят, что наблюдения в интервале от 100 кг до бесконечности являются цензурированными.

Цензурирование во времени обусловлено ограниченностью интервалов наблюдения. Например, если требуется модель для оценки вероятности ухода клиента в зависимости от интенсивности использования им услуг компании, то к моменту завершения исследования часть клиентов может уйти, а часть — нет. Кроме этого клиент может «выпасть» из поля зрения и по другим причинам — сменить место жительства, заболеть и т.д. Т.е. часть данных о клиенте после определенной точки времени не будет доступна для моделирования, что и приводит к их цензурированию.

Само понятие цензурированных данных было введено в 1949 году датским статистиком Андерсом Халдом в исследованиях в области промышленного контроля качества. Однако наиболее мощным драйвером развития статистических методов моделирования на основе цензурированных данных стал анализ выживаемости — класс статистических методов моделирования, позволяющих оценить вероятность наступления некоторого события.

Изначально основной областью применения анализа выживаемости была медицина, а именно оценка продолжительности жизни пациента при исследовании применения различных методов лечения и лекарств. Позднее методы анализа выживаемости получили распространение в страховом деле, социологии, маркетинге, технической диагностике и т.д.

Целью анализа выживаемости является моделирование процессов наступления терминальных событий для объектов некоторой совокупности. В медицине таким событием является смерть пациента, в маркетинге это может быть уход клиента, завершение жизненного цикла товара, ликвидация компании. В технике — выход из строя оборудования.

Примерами вопросов, на которые дает ответ анализ выживаемости, могут быть «какова будет доля оставшихся клиентов, спустя определенное время после применения маркетинговой стратегии», «какие темпы ухода будут наблюдаться среди оставшихся

клиентов», «какие факторы воздействуют на увеличение или уменьшение шансов ухода клиентов» и т.д.

Для ответа на эти вопросы нужно четко определить «время жизни» объекта — период пребывания объекта в совокупности до наступления терминального события. В случае с выживанием пациента «смерть» однозначна, в других предметных областях наступление терминального события не всегда можно локализовать в отдельном моменте времени.

Анализ выживаемости может выполняться только на основе цензурированных данных, когда целевая переменная — вероятность наступления события, а интервал наблюдений (длительность исследования) ограничены во времени.

Можно выделить следующие механизмы цензурирования данных:

1. **Фиксированное цензурированное**. Имеет место в том случае, когда совокупность из n элементов наблюдается в течение фиксированного промежутка времени. При этом, число объектов m для которых событие наступило, является случайным. Для каждого i-го объекта определен свой период наблюдения ($t_i = 1, \ldots, n$), который может быть разным для разных объектов, но фиксирован заранее. Тогда вероятность того, что для объекта в интервале наблюдения не наступит терминальное событие (пациент останется жив, клиент не уйдет) будет $P(t)$.
2. **Случайное цензурирование**. В этом случае наблюдение совокупности из n элементов производится до тех пор, пока событие не наступит для заданного их числа m. Например, пока из 100 клиентов не уйдет 80. Недостатком метода является то, что продолжительность всего исследования оказывается случайной и заранее неизвестна.

Так же, тип цензурирования данных различают по направлениям:

1. **Правостороннее цензурирование**. Цензурирование справа имеет место, если известно, в какой момент эксперимент был начат и что он закончится в момент времени, расположенный справа от точки начала эксперимента. В этом случае любая точка данных находится выше определенного значение, но неизвестно насколько. Например, если шкала весов проградуирована до 100 кг, то цензурированными оказываются все наблюдения, в которых вес будет превышать 100 кг.
2. **Левостороннее цензурирование**. Цензурирование слева происходит в том случае, если неизвестно, когда эксперимент был начат. Очевидно, что в этом случае любая точка данных находится ниже определенного значение, но неизвестно насколько. Например, если на шкале весов отсутствуют значения меньше 20 и цензурированными окажутся все значения ниже 20.
3. **Интервальное цензурирование**. Точка данных находится где-то на интервале между двумя заданными значениями, но неизвестно где. Например, известно, что вес объекта лежит в диапазоне от 20 до 50 кг.

Левостороннее цензурирование можно рассматривать как частный случай интервального с началом интервала в нуле, а правостороннее — с концом интервала в бесконечности.

Если в исследовании участвует определенное количество объектов, и оно заканчивается в заранее заданное время, после чего любые объекты, не попавшие в исследование, подвергаются правому цензурированию, то говорят, что имеет место цензурирование I-го рода. Например, если мы из 100 клиентов исследуем 80 в течение одного месяца с известной даты, и оставшиеся 20 подвергаются правому цензурированию

Если в исследовании участвует определенное число объектов, и оно останавливается когда события происходят для заданного их числа, а остальные объекты подвергаются правому цензурированию, то имеет место цензурирование II-го рода.

Степенью цензурирования называют вероятность попадания в интервал цензурирования в случае цензурирования I-го рода, или отношение количества цензурированных наблюдений к полному объему совокупности в случае цензурирования II-го рода.