**Ошибки в данных при статистическом исследовании.**

Всего в статистическом исследовании существует два типа ошибок исходных данных.

**Выбросы** – это такие значения, которые не укладываются в общую картину распределения выборки.

**Пропуски** – это значения элементов выборки, которые не удалось получить по тем или иным причинам.

Первоначально предполагается, что качество полученных для статистического исследования данных является достаточным. Однако в реальности мы имеем дело с методами и устройствами для сбора данных обладающими некоторой, определённо гарантированной долей ошибок. В метрологии такую долю называют погрешностью. Независимо от метода получения данных или типа используемых приборов, сама хаотическая природа вселенной, обусловленная квантовыми эффектами, не позволяет получить данные с абсолютно полной точностью и достоверностью (Принцип неопределенности Гейзенберга). То есть изначально в реальном мире для любых данных существуют ограничения и в точности, и в достоверности.

Погрешность измерений являет собой статистическую величину, выражающую или среднеквадратичное отклонение (погрешность по Пирсону), или доверительный интервал (погрешность по Корнфельду) определяемые в результате измерений одной и той же неизменной величины.

Любой метод, вычисляющий обобщенное значение (статистику) обладает определенной устойчивостью к ошибкам, имеющимся в выборке. Степень этой устойчивости называют **робастностью**. В статистике существуют как робастные, так и не робастные методы, однако, их выбор остаётся исключительно за исследователем.

|  |  |  |
| --- | --- | --- |
| **Что измеряем** | **Не робастный метод** | **Робастный метод** |
| Разброс выборки | Среднее отклонение | Стандартное отклонение |
| Центральную тенденцию | Среднее арифметическое | Медиана |
| Частоту встречаемости | Простой подсчёт | Мода группы |
| Нормальность | Интеграл разности | Асимметрия, эксцесс |

Существует два подхода для снижения влияния ошибок на результаты статистик.

Подходы и методы обработки ошибок.

|  |  |  |
| --- | --- | --- |
| **Подход** | **Поточный** | **Объектный** |
| Выборка рассматривается | Как упорядоченный набор | Как абстрактное множество |
| Лучше подходит для | Малых и нестандартных выборок | Больших выборок со стандартным распределением |
| **Методы удаления выбросов** | | |
|  | Ситуативное исключение | Исключение по квартилям |
| Статистическая подстановка | Исключение по погрешности |
| Компрессионная подстановка | Исключение по Диксону |
| Индикаторный метод | Исключение по Шовене |
| **Методы удаления пустот** | | |
|  | Ситуативная подстановка | Статистическая подстановка |
| Множественная условная подстановка (Multiple imputation) | Регрессионная подстановка |
| Одиночная последовательная подстановка (Single imputation) | Случайная подстановка (Hot-deck) |
| Прогностическая подстановка | Донорская подстановка (Cold-deck) |
| Индикаторный метод | Индикаторный метод |

Вкратце разберём перечисленные методы.

**Индикаторный метод** предполагает исключительно маркирование выбросов или пропусков без изменения самой выборки. Это единственный метод, не снижающий её репрезентативность.

**Ситуативный** или так называемый **«ad-hoc» метод** подразумевает рассмотрение каждой появившейся отдельной ошибки в её собственном контексте со строгим учетом того является ли данное значение выборки ошибочным или напротив отражает какое-либо реальное изменение ситуации. Этот подход является интуитивно понятным, но неудобен при обработке больших выборок, поскольку количество содержащихся в них ошибок может быть слишком велико для анализа вручную. Альтернативой ad-hoc подхода при анализе больших временных рядов являются методы ослабления значимости, предложенные Хольтом и Макридакисом, суть которых сводится к тому, что чем более старые данные мы анализируем, тем менее значимы, как сами данные, так и ошибки, присутствующие в них.

**Метод исключения** - предполагает исключение из выборки всего подмножества неполных и связанных с ними данных. Существует несколько статистически обусловленных методов определения выбросов или элементов, которые необходимо исключить.

**Исключение по квартилям** – удаляются все данные в выборке, значения которых лежат вне диапазона с границей в полтора межквартильных расстояния от внутренних границ нижнего и верхнего квартилей, т.е. лежащих вне промежутка .

**Исключение по Диксону** – это метод основанный на гипотезе о **нормальности распределения** и следующего из нее заключения, что в нормальном распределении каждое 22-е измерение будет выходить из «двух сигм», и каждое 370-е — из трёх. Неподходящие данные исключают, основываясь на таблице вероятности выбросов Диксона (табличный Q-критерий). Используют на малых выборках до 30 элементов. Метод подходит только для нормально распределенной выборки.

**Исключение по Шовене** – также, как и метод Диксона подходит только к нормально распределенным выборкам, однако позволяет работать с объемом не более 20 элементов. По данному критерию отбрасывают единственное значение, сомнительность которого определяется вероятностью получения значения, отклоняющегося от среднего больше, чем сомнительное значение, зависящее от объёма испытаний (критерием Шовене).

**Исключение по погрешности** (по Граббсу) – это экспертный метод. Из выборки исключается определенный процент данных имеющих максимальные отклонения от предполагаемого центра распределения, затем исключают значения, вносящие наибольшие искажения в стандартное отклонение, этот фактор определяют по таблице статистики Граббса. Предполагаемый процент выбросов зависящий от класса точности прибора, метода измерения и измеряемой величины определяется экспертом. Следует отметить, что данный метод значительно снижает репрезентативность данных.

**Статистическая подстановка** – это подстановка статистики, характеризующей центральную тенденцию выборки (моды, медианы, среднего), вместо выброса или отсутствующего элемента данных.

**Компрессионная подстановка** – это подстановка усредненного значения, вычисляемого по значениям соседних элементов выборки с использованием математической или геометрической компрессии.

**Регрессионная подстановка** – это подстановка значения, вычисляемого по уравнению линейной регрессии, определенному для известных значений выборки.

**Случайная подстановка** – это подстановка случайного значения, находящегося в заданном диапазоне, обычно в качестве допустимого диапазона используют 2 и 3 квартили.

**Донорская подстановка** – это подстановка случайно выбранного значения из множества строго определенных значений-доноров, используется в тех случаях, когда связанность известных значений для некоторых пар выборки невозможно идентифицировать.

**Множественная условная подстановка** используется в тех случаях, когда значимость нулевой гипотезы интуитивно сомнительна, и прослеживается чёткая статика или динамика последовательности. Обычно автоматизированные системы обработки данных способны производить такую подстановку без изменения основных статистик выборки. Но в любом случае снижение репрезентативности при данном подходе – неизбежно.

**Одиночная последовательная подстановка** – это метод, когда отсутствующее значение берут равным предыдущему значению в последовательности. Таким методом, например, определяют котировки активов в неторговые (выходные) дни.

**Прогностическая подстановка** - это метод, когда отсутствующее значение определяют по выбранной методике прогнозирования на основании предшествующих данных. Этот метод не стоит путать с регрессионной подстановкой, поскольку он включает в себя понятие последовательности.

К ознакомлению:

<http://arhiuch.ru/lab4a.html>

<http://arhiuch.ru/lab5.html>

<http://arhiuch.ru/lab7.html>