Гомогенность (Homogeneous)

> Синонимы: Однородность

В статистике, гомогенность — понятие, связанное с предположением о том, что наблюдения в выборке данных являются однородными по некоторому набору параметров. При этом однородность может иметь место по одним признакам, но не иметь по другим. Однородность выборки означает, что все ее элементы имеют схожие характеристики в соответствии с целями исследования.

Гомогенность данных является важным аспектом любого исследования, поскольку от нее зависят точность и надежность получаемых результатов. Если выборка неоднородна, то результаты анализа могут быть сильно искажены, что повлечет за собой ошибки в принятии решений на основе данных.

Чтобы получить однородную выборку, необходимо применять специальные подходы к ее формированию. Например, проводить стратификацию — разделение исходной совокупности на группы (страты) по заданным критериям, и использовать случайный отбор элементов внутри каждой группы.

Предположение о гомогенности имеет важную роль в машинном обучении с точки зрения репрезентативности обучающей выборки. Если исходный набор данных является однородным, то любые извлеченные из него подмножества будут иметь одно и то же распределение, следовательно, отражать одни и те же закономерности и зависимости, которые должны быть обнаружены ML-моделью в процессе обучения.

Таким образом однородная выборка дает точные и достоверные результаты исследования, которые можно обобщить на всю генеральную совокупность. А если в выборку попали элементы из групп, имеющих разные распределения, то результаты могут оказаться некорректными и не отразить реальную картину исследуемого бизнес-процесса.

Для проверки гомогенности используются различные статистические критерии, называемые **критериями однородности**. Они, в отличие от критериев согласия, проверяют не соответствие данных какому-то конкретному распределению, а гипотезу, что у выборок одинаковое распределение.

Например, клиенты, которые выбирают товар A, — образуют первую выборку, а клиенты, предпочитающие товар B, — вторую. Значения в них отражают некоторые потребительские свойства товара. Требуется выяснить, имеется ли значимое различие между потребительскими свойствами товаров A и B. В анализе данных и маркетинге такие критерии часто называют A/B-тест.

Критерии однородности делятся на параметрические и непараметрические. Первые основаны на предположении, что распределение признака в совокупности подчиняется некоторому известному закону. К таким относятся критерии Уилкоксона-Манна-Уитни (U-критерий Манна-Уитни), Критерий Ван дер Вардена, медианный критерий и т.д.

Непараметрическими называют критерии, использование которых не требует предварительного вычисления оценок неизвестных параметров закона распределения признака. Примерами являются критерий однородности Смирнова, критерий Фишера и др.