Бутстрап (Bootstrap)

Разделы: Алгоритмы

В статистике и анализе данных бутстрапом называют статистическую процедуру, основанную на выборке с замещением для определения точности (смещения) выборочных оценок дисперсии, среднего, стандартного отклонения, доверительных интервалов и других структурных характеристик совокупности.

Метод разработан и впервые опубликован в 1972 году Бредли Эфроном.

В основе идеи бутстрапа лежит оценка структурных характеристик генеральной совокупности на основе перевыборки (resampling) из выборки. Иными словами, перевыборка по отношению к выборке рассматривается как выборка по отношению к генеральной совокупности.

Алгоритм работы метода следующий:

1. Из генеральной совокупности формируется случайная выборка из $N(t)$ наблюдений (например, если требуется определить среднюю сумму чека посетителя супермаркета, будем оценивать ее на основе выборки из 1 000 клиентов).
2. К выборке применяется случайная перевыборка с возвратом (псевдовыборка) того же объема, но в которую некоторые наблюдения могут попасть несколько раз, а другие не попасть совсем. Например, если выборка содержала 5 значений (1, 2, 3, 4, 5), то результатом перевыборки может быть (2, 2, 4, 5, 5). Затем вычисляется ее среднее.
3. Процедура перевыборки повторяется достаточно много раз (несколько десятков, сотен или даже тысяч), и для каждого случая вычисляется среднее.
4. Из полученного набора средних значений вычисляется среднее и рассматривается как среднее всей генеральной совокупности.

Важнейшим преимуществом бутстрапа являются:

- простота реализации;
- отсутствие необходимости гипотез о параметрах распределения данных;
- возможность оценивания многих статистических характеристик (среднего, дисперсии, стандартного отклонения, доверительных интервалов, квантилей, коэффициентов корреляции и др.).

К недостатку метода можно отнести использование малореалистичного предположения о независимости перевыборок и значительные вычислительные затраты при их многократном построении.

Метод оказывается особенно полезным, когда теоретическое распределение данных неизвестно или объем выборки мал для прямой статистической оценки.

В анализе данных бутстрап используется для оценки точности аналитических моделей.