

Ошибка репрезентативности (Representativeness error)

Синонимы: Ошибка выборки, Смещение выборки, Sampling error, Sampling bias

Ошибкой репрезентативности в статистике и машинном обучении называют величину расхождения статистических характеристик выборки и генеральной совокупности, из которой она была извлечена. Поэтому ее часто называют ошибкой выборки. Если статистическое исследование производится на выборке, то эта ошибка приводит к тому, что его результаты будут различными для выборки и совокупности.

Более подробно с проблемой репрезентативности в статистике и машинном обучении можно ознакомиться в статье [«Репрезентативность выборочных данных»](#).

Поскольку выборка почти всегда используется для оценки неизвестных параметров совокупности, точное измерение ошибки репрезентативности обычно невозможно по определению. Однако ее можно оценить с помощью общих методов, таких как бутстрап.

Простыми словами, ошибка репрезентативности возникает тогда, когда вместо совокупности наблюдается только выборка из нее. Следовательно, чтобы избежать влияния этой ошибки на результаты статистического исследования, нужно рассматривать всю совокупность, что в большинстве случаев невозможно из-за большого размера или недоступности всех ее наблюдений. Поэтому в реальных задачах используется выборочное исследование, которое неизбежно порождает ошибку репрезентативности.

Выделяют два типа ошибок репрезентативности:

- **случайные** — обусловлены воздействием случайных факторов. Они присутствуют всегда и могут быть статистически оценены и учтены при распространении результатов выборочного исследования на всю генеральную совокупность.
- **систематические** — возникают при нарушении научных принципов формирования выборки.

Расчет ошибок репрезентативности позволяет решить одну из главных проблем организации выборочного наблюдения — оценить репрезентативность выборки. При этом различают среднюю и предельную ошибки.

Средняя ошибка репрезентативности вычисляется по формуле:

$$\mu = \sqrt{\frac{\sigma^2}{n}},$$

где σ — выборочная дисперсия, n — число наблюдений в выборке. Таким образом, средняя ошибка обратно пропорциональна числу наблюдений в выборке. Следует отметить, что данная формула справедлива для случайной повторной выборки, которая рассматривается как результат n повторных независимых испытаний. Результат каждого испытания есть случайная величина, закон распределения которой совпадает с распределением генеральной совокупности.

Предельная ошибка репрезентативности — это максимально возможное расхождение между характеристиками выборки и генеральной совокупности, гарантируемое с определенной вероятностью:

$$\Delta\mu = t\sqrt{\frac{\sigma^2}{n}} = t\mu,$$

где t — коэффициент доверия. Таким образом, предельная ошибка репрезентативности равна t -кратной средней ошибке. При этом каждому t соответствует доверительная вероятность, которую можно определить из статистических таблиц. Фрагмент такой таблицы для нормального закона распределения представлен ниже.

Доверительная вероятность, p	0.683	0.954	0.997
Коэффициент доверия, t	1	2	3

Это означает, что вероятность 2-кратного ($t = 2$) превышения средней ошибки репрезентативности не превысит $p = 0.954$, а 3-кратного — $p = 0.997$. Что же касается доверительной вероятности самой средней ошибки репрезентативности, то она будет $p = 0.683$.

При этом, если объем выборки превышает 30 наблюдений, то доверительную вероятность следует определять по таблицам для нормального закона распределения, а если меньше, то по таблицам для распределения Стьюдента.