УДК 004.652.44 https://doi.org/10.26160/2309-8864-2023-15-39-41

ВИРТУАЛЬНОЕ РАСШИРЕНИЕ ТАБЛИЦ СТАТИЧЕСКИХ ПАРАМЕТРОВ

Моисеев А.А.
Научно-производственное предприятие «Технос – РМ», Мытищи

Ключевые слова: реляционные базы данных, статический параметр, случайные величины, статистические выводы, представительные данные, бутстрэп, эмпирическое распределение, функция Хэвисайда, рекуррентное соотношение, мультипликативное приращение, генеральная совокупность.
Аннотация. Предлагается метод виртуального расширения таблиц статических параметров, базирующийся на бутстрэпном подходе, т.е. на предположении, что истинное распределения значений численного параметра может быть приближено эмпирическим, построенным на основе наличных данных. Формирование выборок бинарных значений предлагается осуществлять методом статистической имитации, используя расчетную долю единиц в исходной выборке. Предложен также метод приведения формируемой выборки к генеральной совокупности, продемонстрированный на примере введения гендерной поправки.

VIRTUAL ENLARGEMENT OF STATIC PARAMETERS TABLES

Moiseev A.A.
Scientific and industrial enterprise “Technos – RM”, Mytischi

Keywords: relational databases, static parameter, random value, statistical inference, sampling representative data, bootstrap, empirical distribution, Heaviside function, recurrence relation, multiplicative increment, parent population.
Abstract. Virtual enlargement of static parameters tables based on bootstrap application, i.e. on proposition that true distribution can be approximated with empirical one, which is formed using on hand data. For binary values forming proposed to apply statistical simulation using calculated part of units in initial sample. Proposed also the method of sample reduction to parent population, which demonstrated by the example of gender amendment introducing.

Основу реляционных баз данных составляют двумерные таблицы вида «объект (субъект) – статический параметр», значения параметров в которых могут быть численного или бинарного типа [1]. В предположении о независимости объектов они могут интерпретироваться как выборки независимых случайных величин соответствующего типа, на основе которых могут строиться статистические выводы. Обычной проблемой при этом является непредставительность этих выборок, затрудняющая обоснование построенных выводов.

Возможным выходом в этой ситуации является использование виртуального расширения этих таблиц с использованием бутстрэпного подхода [2]. В основе последнего лежит предположение, что истинное распределение численных значений может быть приближено эмпирическим, построенным на основе наличных данных – исходной выборки значений x_1, …, x_n численного параметра. Данное распределение имеет вид [3]:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x - x_i), \quad (1)$$

где I – функция Хэвисайда [4]; n – объем исходной выборки (число объектов в исходной таблице).

Для выборки значений численного параметра строится массив мультипликативных приращений вида:

$$\begin{cases} y_1 = 1, \\ y_i = \dfrac{x_i}{x_{i-1}}, i = 2, ..., n. \end{cases} \quad (2)$$

Построение дополнительных численных значений в ходе расширения осуществляется с помощью рекуррентного соотношения:

$$x_i = x_{i-1} y,$$

где $i = n+1, \ldots, N$; N – объем расширенной выборки; y – результат случайной выборки из массива приращений $y_1, \ldots, y_n$, считающихся равновероятными.

В случае множественности численных переменных используется случайный выбор номер значения параметра в выборке (номер объекта). Этому номеру сопоставляется значение приращения, а также значения приращений прочих численных параметров, рассчитанные по исходным выборкам их значений.

Для расчета выборок дополнительных значений бинарных параметров можно использовать следующий подход. По исходной бинарной выборке рассчитывается доля единиц E в ее составе. При формировании дополнительного значения в ходе расширения всякий раз разыгрывается случайное число ξ, равномерно распределенное в интервале (0, 1). Если $\xi > E$, то дополнительное значение считается равным 1 (true). В противном случае оно считается равным 0 (false). Таким образом формируется строка значений в таблице, соответствующая дополнительному объекту.

В ходе расширения таблиц параметров можно, также, осуществить ее коррекцию к генеральной совокупности. Продемонстрируем это на примере введения гендерной поправки для значений численных параметров [5].

Пусть: z – доля мужчин в исходной выборке, %;

$(100 – z)$ – доля мужчин в исходной выборке, %;

z_0 – доля мужчин в генеральной совокупности, %;

$(100 – z_0)$ – доля женщин в генеральной совокупности, %.

Коэффициенты поправки для мужчин составляет при этом:

$$m = \frac{z_0}{z}.$$

а для женщин:

$$w = \frac{100 - z_0}{1000 - z}.$$

Правка численных параметров осуществляется в соответствии с выражениями:

$x \to mx$ для мужчин; $x \to mx$ для женщин.

Значения бинарных параметров при этом сохраняются неизменными, как и методика их имитации при расширении выборки. Предложенный подход, таким образом, позволяет расширять таблицы статических параметров и тем самым обеспечивает повышение представительности и как следствие – обоснованности формируемых на этой основе статистических выводов.

Список литературы

1. Мейер Д. Теория реляционных баз данных. – М.: Мир, 1987. – 608 с.
2. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – М.: Финансы и статистика, 1988. – 263 с.
3. Корн Г., Корн Т. Справочник по математике. – М.: Наука, 1973. – 832 с.
4. Зорич В.А. Математический анализ, часть II. – М.: МЦНМО, 2012. – 818 с.
5. Елисеева И.И., Юзбашев М.М. Общая теория статистики. – М.: Финансы и статистика, 2009. – 656 с.

Сведения об авторе:
Моисеев Александр Александрович – к.т.н., старший научный сотрудник.