УДК 621.382.2:[621.315.592.3+620.192.63]

ЛИНЕЙНОЕ СГЛАЖИВАНИЕ ЭМПИРИЧЕСКОЙ ФУНКЦИИ ВЕРОЯТНОСТИ ДЛЯ МАЛЫХ ВЫБОРОК ДАННЫХ

Н. Н. Мешечек

Брестский государственный технический университет, Брест, Беларусь,
e-mail: cm@bstu.by

The main advantage of nonparametric statistics, including ordinal (rank) statistics, is their independence both from the parameters and from the type of distribution of the random variable.

The proposed two methods for estimating the empirical distribution function, based on order statistics: a reliable estimate of the quantiles of the desired selected level, as well as an estimate of the quantile level for the available points (sample data) have shown their effectiveness for small samples of empirical data. Both methods include setting the required level of confidence level for the result. By specifying the statistical reliability of the interval estimate of the function, it is then possible to approximate the empirical distribution function.

Введение

На начальном этапе статистического (как параметрического, так и непараметрического) анализа эмпирических данных обычно требуется определить эмпирическую функцию распределения как оценку (эмпирическую меру) неизвестной функции распределения генеральной совокупности по имеющейся выборке данных измерений.

Ранее предложенные методы оценивания эмпирической функции распределения, основанные на непараметрических (порядковых) статистиках весьма эффективны для случая выборок данных малого объема [1]. Важнейшим их достоинством является возможность задания достоверности (статистической обеспеченности) одно- или двусторонней интервальной оценки функции. Это позволяет затем аппроксимировать эмпирическую функцию распределения [2].

Способы аппроксимации с последующей экстраполяцией эмпирической функции распределения.

Одним из способов непараметрической статистики является построение эмпирической функции распределения.

Эта функция определяет частотное (эмпирическое) распределение, которое может быть найдено, например, по формуле

$$\hat{F}\left(X\right)=\frac{1}{n}\sum\nolimits_{i=1}^{n} I\left(x_i < X\right) \qquad (1)$$

Известно, что эмпирическая функция распределения является несмещенной состоятельной оценкой неизвестной функции распределения $F\left(X\right)$ для гипотетически бесконечной выборки случайной величины. Однако данное теоретическое утверждение формулируется для случая наращивания числа членов выборки (до 35 и более) [3]. При малых же размерах выборок уже интуитивно понятно, что такая оценка оказывается завышенной на правом конце (возрастающего) вариационного выборочного ряда. Аналогично, равная нулю оценка эмпирической функции распределения левее начала вариационного ряда будет заниженной. Наиболее приближенными к истинным значениям представляются оценки в области среднего, соответствующего величине $F\left(X\right)=0,5$.

Таким образом, получаемые соотношением вида (1) оценка сопротивления элемента (в левой части вариационного ряда) и оценка внешней нагрузки (в правой) являются неточными. Это весьма проблематично с точки зрения безопасности, поскольку может привести к неверным расчетам вероятности отказа, например, к переоценке надежности конструкции.

Как ранее отмечалось, оценивание квантилей распределения случайной величины с известной достоверностью может быть выполнено с применением аппарата непараметрических (порядковых, ранговых) статистик [3]. Квантиль $X_{p,\beta}$ уровня p имеет свою функцию распределения $G(X_p)$.

Преимущества применения ранее рассмотренных методов оценивания эмпирической функции распределения является возможность задания достоверности (статистической обеспеченности) одно- или двусторонней интервальной оценки функции. Это позволяет затем аппроксимировать эмпирическую функцию распределения.

Аппроксимация предусматривает замену одних математических объектов другими, в том или ином смысле близкими к исходным. Другими словами, можно сказать, что это метод приближения, при котором для нахождения дополнительных значений, отличных от исходных данных, приближенная функция проходит не через узлы интерполяции, а между ними.

Применение аппроксимации (в сравнении с интерполяцией) имеет ряд преимуществ:

1. преимущество использования аппроксимации при значительном количестве табличных данных (интерполирующая функция становится громоздкой);

2. интерполирующей функцией невозможно описать данные при повторении эксперимента в одних тех же начальных условиях (требуется статистическая обработка);

3. преимущество использования аппроксимации для сглаживания погрешностей эксперимента. Данные x_i и y_i обычно содержат ошибки, поэтому интерполяционная формула повторяет эти ошибки.

Для аппроксимации функции $G(X_p)$ с экстраполяцией могут быть использованы следующие методы:

1. Линейная аппроксимация;
2. Нелинейная (например, логарифмическая) аппроксимация;
3. Линейная в нелинейной шкале фактора.

В качестве линейной аппроксимации может использоваться:

- линейное трёхточечное (или 5-точечное) сглаживание порядковых статистик как оценок эмпирической функции распределения,
- линейная комбинация порядковых статистик (*L3*-оценка квантилей).

Сглаживание данных эксперимента является процедурой усреднения с помощью интерполяционных полиномов, обеспечивающей получение уточненного значения $\tilde{y}_i$ по заданному значению y_i и ряду близлежащих значений $(...,y_{i-1},y_i,y_{i+1},...)$, известных со случайной погрешностью.

Сглаживание осуществляется по группам точек скользящих вдоль всей таблицы. При линейном сглаживании по трем точкам берут первую группу точек (x_1,y_1) (x_2,y_2) (x_3,y_3) и сглаживают (находят значение аппроксимирующего многочлена) среднюю точку y_2 заменяя ее вычисленным значением $\tilde{y}_2$. Затем берут следующую группу точек (x_2,y_2) (x_3,y_3) (x_4,y_4), вычисляют значение многочлена для средней точки этой группы y_3 и сглаживают ее, заменяя вычисленным значением $\tilde{y}_3$. И так проходят до конца таблицы. После этого производят сглаживание двух первых и двух последних точек по специальным менее точным формулам [4]. На рисунке 1 показан пример сглаживания данных.

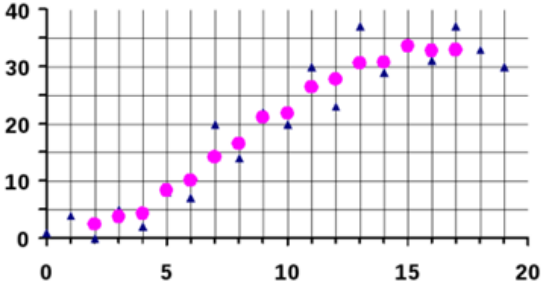

Рисунок 1 — Пример сглаживания экспериментальных данных: Δ - экспериментальные данные, ○ – результаты сглаживания

L-оценки параметров распределений формируются как линейные комбинации порядковых статистик. L-оценки обладают хорошими свойствми робастности [5]. Под робастностью в статистике понимают нечувствительность к малым отклонениям от предположений. Они устойчивы к наличию аномальных ошибок измерений, к малым отклонениям от исходных предположений о виде наблюдаемого закона распределения. Это позволяет использовать L-оценки в процедурах параметрической отбраковки наблюдений [6]. При больших объемах выборок строить L-оценки с исполь-

зованием всего множества порядковых статистик весьма затруднительно и более экономично для вычисления оценок параметров воспользоваться выборочными квантилями.

Оценка квантили необходимого уровня p с заданной обеспеченностью β можно представить как нормированную линейную комбинацию трех первых порядковых статистик эмпирического ряда измерений:

$X_{p,\beta}=aX_{(1)}+bX_{(2)}+cX_{(3)}$, при $p\rightarrow 0$ либо

$X_{p,\beta}=aX_{(N)}+bX_{(N-1)}+cX_{(N-2)}$, при $p\rightarrow 1$.

Параметры a,b,c – коэффициенты линейной комбинации, нормированные условием $a+b+c=1$.

Нормировка коэффициентов a, b и c означает, что корректно учитывается параметр положения случайной величины f и обеспечивается несмещенность оценки искомой квантили. Поскольку оценка включает величины интервалов между порядковыми статистиками, в ней учтен также параметр масштаба, т.е. степень рассеяния случайной величины.

Все коэффициенты оценки зависят от параметров p и β, а также общего числа N результатов единичных испытаний [6].

- Модификация метода наименьших квадратов (интегральное сглаживание на неэквидистантных интервалах) [7].

Типичной задачей обработки данных является установление функциональной зависимости некоторой величины (отклика) от одной или нескольких переменных (факторов). В теории вероятностей функция, приближенно представляющая статистическую зависимость случайных величин, определяется как регрессия, в частности – средняя квадратическая регрессия.

Регрессии (представляющие собой приближенно статистическую зависимость случайных величин), найденные классическим методом наименьших квадратов, равно как и соответствующий им коэффициент детерминации R-квадрат, оказываются недостаточно корректными в случае нерегулярного расположения отсчетов на шкале фактора.

Для оценки качества регрессии предлагается более точный коэффициент интегральной детерминации R_{DD} - квадрат (definite determinative), равный доле полного квадрата отклонения кусочно-гладкого приближения эмпирических данных, которая объяснена регрессионной моделью [7]. Вычисления при этом выполняются на интервале изменения фактора.

Заключение

Как отмечалось ранее [3] в практических задачах анализа надежности строительных конструкций оценивание параметров сопротивления и нагрузок зачастую выполняется на основе сравнительно малых выборках результатов реальных измерений. Статистическая обработка этих результатов обычно связана с построением эмпирических функций распределения. Предложены и рассмотрены способы аппроксимации эмпирическое функции вероятности, которые могут быть эффективно применины для малого объема исходных данных.

Список цитированных источников

[1] Дереченник, С.С., Мешечек, Н.Н. Решение задачи анализа функции состояния на основе приближения хвостовых частей распределений случайных величин нагрузки и сопротивления // Вестник Брестского государственного технического университета. – 2023. – С. 7-9.

[2] Шуленин, В.П. Математическая статистика Ч. 2. Непараметрическая статистика: учебник // Томск: Изд-во НТЛ, 2012. – 388 с.

[3] Дереченник, С.С., Мешечек, Н.Н. Численное решение задачи оценивания эмпирической функции распределения для малых выборок с заданной достоверностью // Вестник Брестского государственного технического университета. – 2024. – С. 67-71.

[4] Румшиский, Л.З. Математическая обработка результатов эксперимента. Справочное руководство // – М: Главная редакция физико-математической литературы издательства “Наука”, 1971. – 192 с.

[5] Шуленин, В.П. Математическая статистика Ч. 3. Робастная статистика: учебник // Томск: Изд-во НТЛ, 2012. – 520 с.

[6] Дереченник, С.С., Тур, В.В. Д 36 Оценивание соответствия прочности бетона: теория и практика : монография. - Брест : Издательство БрГТУ, 2023. - 160 с.

[7] Дереченник. С.С. Интегральная оценка качества регрессионных моделей / С.С.Дереченник, А.В.Дмитриева, С.С.Дереченник-мл. // Вестник Брестского государственного технического университета. - 2009. - № 5 (59): Физика, математика, информатика. - С. 77-80.