УДК 519.24

М. Г. Лапина
Санкт-Петербургский Государственный Морской Технический Университет
г. Санкт-Петербург, Россия

ОЦЕНКА ПАРАМЕТРОВ НОРМАЛЬНО РАСПРЕДЕЛЕННОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ПО ЕЕ ЦЕНЗУРИРОВАННОЙ ВЫБОРКЕ

В статье рассмотрены существующие методы оценки параметров нормально распределенной случайной величины по ее цензурированной выборке. Сформулирован закон распределения цензурированной выборки, на базе чего предложен новый метод оценки параметров. Продемонстрировано сравнение численных результатов методов, в ходе чего выявлено превосходство предложенного нового метода над уже существующими.

Точный прогноз пассажирского спроса — один из важнейших факторов, влияющих на успешную работу любой авиакомпании. Прогноз всегда базируется на исторических данных, позволяющих получить статистическую выборку, оценить параметры ее распределения. Однако эти данные отражают лишь количество проданных авиабилетов, но не общий спрос.

Авиакомпании заранее устанавливают пределы бронирования билетов. Если спрос на продукт оказался ниже предела, то количество проданных билетов соответствует спросу. В противном случае, будет зафиксировано количество проданных билетов, равное пределу бронирования. Таким образом, истинное значение спроса будет усечено пределом бронирования или, как еще говорят, «цензурировано».

В данной статье рассмотрены различные методы оценки параметров распределения общего/ неограниченного спроса и проведены сравнительные расчеты для исходных данных с различной долей цензурирования.

Пусть X и $Y-$ независимые случайные величины, распределенные по нормальному закону: $X \in N(\mu_X, \sigma_X)$, $Y \in N(\mu_Y, \sigma_Y)$. $x_i -$ наблюдения над случайной величиной X (элементы нецензурированной выборки). $y_i -$ наблюдения над случайной величиной Y (элементы выборки ограничений). Выборка с элементами $z_i \in Z -$ цензурированная выборка, причем

$$z_i = min\{x_i, y_i\}.$$

Применительно к оценке параметров общего спроса в авиаиндустрии: X отражает общий спрос, $Y -$ ограничения/пределы бронирования, $Z -$ наблюдаемый спрос.

Рассмотрим существующие методы оценки параметров распределения нецензурированной выборки x_i по имеющейся цензурированной выборке z_i.

Метод Naïve 1 (N1): заменяет все цензурированные наблюдения на математическое ожидание всех наблюдений рассматриваемой выборки.

Метод Naïve 2 (N2): заменяет все цензурированные наблюдения на математическое ожидание всех нецензурированных наблюдений.

Метод Naïve 3 (N3): заменяет все цензурированные наблюдения на большее из рассматриваемого цензурированного наблюдения и математического ожидания всех нецензурированных наблюдений [1].

Метод Expectation Maximization (EM): является итерационным методом, состоящим из двух шагов:

- восстанавливающего шага, E −шага, на котором производится замена наблюдаемых цензурированных данных их «восстановленными» значениями;
- максимизационного, M −шага, на котором за счет максимизации функции правдоподобия производится оценка параметров распределения нецензурированной выборки [2].

В методе EM предполагается, что цензурированная случайная величина, распределена по усеченному слева нормальному закону [3]. Для

восстановления элементов нецензурированной выборки для наблюдений, подвергшихся цензурированию, воспользуемся следующими соотношениями:

$$\zeta_i^{(k-1)} = M[x \,|\, x \geq z_i, X \sim N(\hat{\mu}_X^{(k-1)}, \hat{\sigma}_X^{(k-1)})],$$

$$(\zeta_i^2)^{(k-1)} = M[x^2 \,|\, x \geq z_i, X \sim N(\hat{\mu}_X^{(k-1)}, \hat{\sigma}_X^{(k-1)})],$$

где $\zeta_i -$ восстановленные значения случайных величин выборки x_i.

$M[x \,|\, x \geq z_i, X \sim N\left(\hat{\mu}_X^{(k-1)}, \hat{\sigma}_X^{(k-1)}\right] -$ математическое ожидание случайной величины x усеченного слева нормального распределения.

$M[x^2 \,|\, x \geq z_i, X \sim N\left(\hat{\mu}_X^{(k-1)}, \hat{\sigma}_X^{(k-1)}\right] -$ математическое ожидание квадрата случайной величины x усеченного слева нормального распределения [2].

Метод Projection-Detruncation (PD): также является итерационным методом, состоящим из двух шагов – восстанавливающего и максимизационного. Отличие метода от предыдущего состоит в способе расчета восстановленных наблюдений.

Основная идея метода Projection-Detruncation состоит в том, что известна вероятность τ недооценки параметров распределения нецензурированной выборки по цензурированным наблюдениям [4]. Параметр τ рассчитывается как отношение вероятности превышения прогнозируемого значения ζ_i к вероятности превышения наблюдаемого цензурированного элемента z_i. Тогда восстановленные значения вычисляются из уравнения:

$$\int_{\zeta_i^{(k-1)}}^{\infty} f\left(x \,|\, \left(\hat{\mu}_X^{(k-1)}, \hat{\sigma}_X^{(k-1)}\right)\right) dx = \tau \int_{z_i}^{\infty} f\left(x \,|\, \left(\hat{\mu}_X^{(k-1)}, \hat{\sigma}_X^{(k-1)}\right)\right) dx.$$

Рассмотрим **новый метод**, идея которого состоит в формулировании закона распределения для цензурированной выборки и последующего применения метода максимального правдоподобия.

Предположим, что кроме самой выборки с элементами z_i, известно также, какие элементы были цензурированы, а какие нет. Тогда, если некоторый

элемент $z_i \in Z$ цензурирован, то наблюдаем на самом деле элемент $y_i \in Y$, где $Y \in N(\mu_Y, \sigma_Y)$, а если нет, то элемент $x_i \in X$, где $X \in N(\mu_X, \sigma_X)$.

В результате, поскольку будем наблюдать x_i, при условии, что $x_i < y_i$, вероятность осуществления которого равна $P(y_i > x_i)$, то элементы $z_i = x_i$ распределены по закону с плотностью вероятности:

$$\frac{1}{\sigma_X}\varphi\left(\frac{z-\mu_X}{\sigma_X}\right)\left(1-\Phi\left(\frac{z-\mu_Y}{\sigma_Y}\right)\right),$$

$\Phi(x) -$ функция распределения для нормированной нормальной случайной величины, $\varphi(x) -$ плотность для нормированной нормальной случайной величины), а элементы $z_i = y_i$, которые будем наблюдать с вероятностью $P(x_i > y_i)$, распределены по закону с плотностью вероятности:

$$\frac{1}{\sigma_Y}\varphi\left(\frac{z-\mu_Y}{\sigma_Y}\right)\left(1-\Phi\left(\frac{z-\mu_X}{\sigma_X}\right)\right).$$

Функция правдоподобия составляется следующим образом:

$$L(\mu_X, \sigma_X, \mu_Y, \sigma_Y) =$$
$$= \prod_{i=1}^{n}\frac{1}{\sigma_X}\varphi\left(\frac{x_i-\mu_X}{\sigma_X}\right)\left(1-\Phi\left(\frac{x_i-\mu_Y}{\sigma_Y}\right)\right)$$
$$\times \prod_{i=1}^{m}\frac{1}{\sigma_Y}\varphi\left(\frac{y_i-\mu_Y}{\sigma_Y}\right)\left(1-\Phi\left(\frac{y_i-\mu_X}{\sigma_X}\right)\right),$$

где $m -$ количество цензурированных элементов в выборке Z, $n -$ количество нецензурированных элементов. Из условия максимизации функции правдоподобия определяются оценки параметров μ_X и σ_X.

Было проведено тестирование всех описанных методов с учетом размера выборки, процента цензурирования, параметров распределения. Анализ результатов методов продемонстрировал превосходство методов Expectation Maximization и нового метода над тремя упрощенными методами и методом Projection-Detruncation при $\tau = 0.5$. При увеличении размера выборки и процента цензурирования было выявлено превосходство нового метода над методом Expectation Maximization.

Примеры числовых расчетов приведены в табл. 1 – 3:

Таблица 1. Сравнительная таблица $(\hat{\mu}_X, \hat{\sigma}_X)$ *при выборке размером 10 элементов и параметрах* $\mu_X = 20.8, \sigma_X = 6.0$

	2.8%	**33.1%**	**56.7%**	**77.2%**	**89.3%**
Новый метод	20.8, 6.0	20.4, 5.6	19.5, 4.5	17.6, 2.6	15.3, 1.4
EM	20.8, 6.0	20.4, 5.6	19.5, 4.5	17.6, 2.6	15.3, 1.4
PD	20.8, 6.0	19.9, 4.9	18.5, 3.1	16.5, 1.2	14.9, 0.4

Таблица 2. Сравнительная таблица $(\hat{\mu}_X, \hat{\sigma}_X)$ *при выборке размером 50 элементов и параметрах* $\mu_X = 19.7, \sigma_X = 7.0$

	2.58%	**29.24%**	**51.42%**	**71.92%**	**95.04%**
Новый метод	19.7, 7.0	19.5, 6.9	19.0, 6.4	18.3, 5.8	14.7, 3.6
EM	19.7, 7.0	19.5, 6.9	19.0, 6.4	18.3, 5.8	13.9, 3.2
PD	19.7, 7.0	19.1, 6.1	17.9, 4.8	15.9, 3.2	10.7, 0.7

Таблица 3. Сравнительная таблица $(\hat{\mu}_X, \hat{\sigma}_X)$ *при выборке размером 100 элементов и параметрах* $\mu_X = 20.4, \sigma_X = 5.5$

	2.3%	**27.73%**	**56.31%**	**80.85%**	**98.44%**
Новый метод	20.4, 5.5	20.1, 5.1	19.7, 4.7	19.2, 4.4	16.9, 3.6
EM	20.4, 5.5	20.1, 5.1	19.7, 4.7	19.2, 4.4	14.7, 2.5
PD	20.3, 5.5	19.8, 4.6	18.7, 3.4	16.9, 2.1	11.7, 0.4

Все вычисления были выполнены с помощью программы *Mathematica.*

Список использованных источников

1. L. R. Weatherford, S. Polt. Better unconstraining of airline demand data in revenue management systems for improved forecast accuracy and greater revenues // Journal of revenue and pricing management. 2002. Vol. 1. No 3. P. 234-254;

2. Kalyan T. Talluri, Garrett J. van Ryzin. The theory and practice of Revenue Management. Kluwer Academic Publishers. Boston. 2004. P.474-478, 485-486;

3. Вадзинский Р.Н. Справочник по вероятностным распределениям - СПб: Наука,2001. - С. 200-202;

4. Richard H. Zeni. Improved forecast accuracy in revenue Management by unconstraining demand estimates from censored data/ PhD thesis. Rutgers University. Newark. New Jersey. 2001. P. 78-101.