Информационные процессы, Том 9, № 3, 2009, стр. 210–215.

Оценивание скорости убывания экспоненциального хвоста распределения

М.Г. Давиденко

"НТЦ Орион", Москва, Россия

Поступила в редколлегию 22.09.2009

Аннотация—В статье рассматривается оценка скорости убывания хвоста распределения случайных величин в предположении, что вид самого распределения не известен.

1. ВВЕДЕНИЕ

В задачах поверки статистических гипотез обычно предполагается, что все наблюдения – случайные величины с одинаковым или примерно одинаковым законом распределения. Вместе с тем, возникают ситуации, когда главная часть закона распределения наблюдения сосредоточена на некотором множестве M, а остальная, относительно небольшая, на оставшейся части числовой оси; при этом указать общий закон распределения для всех наблюдений со значениями из множества M не представляется возможным. С другой стороны, из общих соображений можно предположить, что хвосты распределений (распределения значений вне множества M) имеют экспоненциальную скорость убывания. Такой вывод может базироваться на теории больших уклонений [1].

На практике возникают ситуации, когда имеются два семейства распределений приведенного выше типа, причем главная часть распределений первого типа находится на множестве M_1, а второго – на множестве M_2, причем множества M_1 и M_2 не пересекаются. Таким образом, возникает задача выбора порога в области между множествами для проверки гипотез о типе наблюдаемых распределений с заданными вероятностями ошибок.

Для решения этой задачи необходимо уметь оценивать хвосты распределений в области малых и сверхмалых вероятностей. Обнаружить такие вероятности статистическими средствами невозможно, но, зная скорость убывания хвоста распределения в области малых вероятностей и экстраполируя ее на область сверхмалых вероятностей, можно такие вероятности вычислить. Решению этой задачи посвящена данная статья.

В разделе 2 приводится постановка задачи. Здесь описывается модель наблюдаемых распределений в частном случае (в остальных случаях задача решается аналогично), когда хвост распределения рассматривается на $+\infty$. Показано, что задача оценивания вероятности хвоста распределения сводится к оценке трех параметров, причем их взаимосвязь такова, что наиболее важным является оценка параметра λ. В разделе 3 рассматриваются методы построения оценки параметра λ. Для частного случая семейства распределений было проведено численное моделирование, результаты которого приведены в разделе 4. Завершается работа выводами.

2. ПОСТАНОВКА ЗАДАЧИ

Предположим, что наблюдаются значения случайных величин со значениями из множества R, R – множество действительных чисел. На интервале $[0; a]$ функция распределения у каждой величины, вообще говоря, своя, а на интервале $(a; +\infty)$ все функции распределения имеют

экспоненциально убывающий хвост

$$P(X_i > x) = 1 - F(x) = \alpha e^{-\lambda(x-a)} \tag{1}$$

(см. рис. 1).

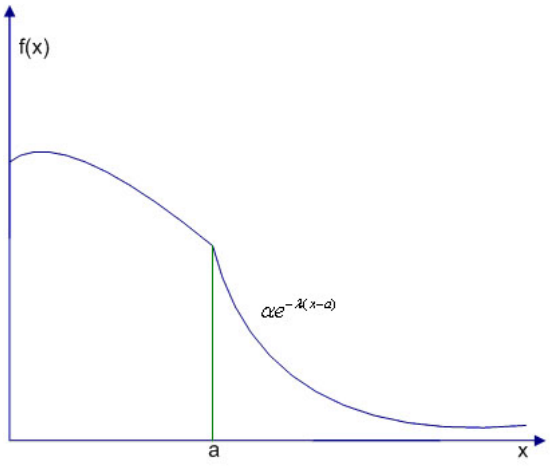

Рис. 1. График функции плотности распределения $f(x)$.

Параметр λ задает скорость убывания хвостов распределений результатов наблюдений. Предполагается, что значение этого параметра у всех наблюдений одинаково. Величина a задает начало области, где выполняется условие (1). Вообще говоря, значение a свое для каждого наблюдения, однако, как будет видно из дальнейшего, в этом случае под значением этого параметра можно понимать его максимальное значение для всех возможных наблюдений. Параметр α определяет вероятность попадания значения случайной величины в экспоненциальный хвост распределения. Как и величина a, его значение различно для различных наблюдений; поэтому в качестве общего для всех наблюдений значения α можно брать его значение, соответствующее максимальному значению a, т.е.

$$\alpha = P(X > a).$$

Как указывалось выше, оценка скорости убывания хвоста распределения необходима для оценки величины $P(X_i > C)$, где $C > a$. Оценить значение этой вероятности по результатам измерений не всегда представляется возможным, в особенности при ее малых значениях. Если параметры α, λ, a известны, то эта вероятность вычисляется по формуле

$$P(X_i > C) = \alpha e^{-\lambda(C-a)}.$$

Понятно, что значения параметров α, λ, a могут быть получены только с некоторой точностью, однако для решения поставленной задачи некоторые ошибки не являются критическими.

1) При неточном вычислении оценки параметра $\hat{\lambda}$, но такой что $\hat{\lambda} < \lambda$, показатель скорости убывания экспоненциального хвоста распределения будет меньше реального, поэтому оценка для $P(x > C)$ будет

$$\alpha e^{-\hat{\lambda}(C-a)} > \alpha e^{-\lambda(C-a)},$$

а с точки зрения проверки гипотез это означает, что оценка вероятности ошибки будет завышена. Если $\hat{\lambda} > \lambda$, то оценка вероятности ошибки будет занижена, что является неприемлемым, поскольку в этом случае решающее правило не будет обеспечивать заданную точность.

Поэтому при вычислении оценки скорости убывания хвоста распределений допустимо занижать ее значение, поскольку это будет приводить только к необходимости проводить дополнительные наблюдения, но нельзя ее завышать.

2) Состоятельная оценка параметра $\hat{a}$ практически невозможна не только из-за того, что ее значение может быть различно у разных наблюдений, но и из-за достаточно смазанной границы начала хвоста экспоненциального распределения, поскольку плотность распределения наблюдений может и не изменяться скачкообразно при выходе из множества M. Однако, как видно из дальнейшего, состоятельность оценки $\hat{a}$ не требуется; необходимо лишь выполнение условия

$$\hat{a} \geq a.$$

3) Оценка параметра $\hat{\alpha}$ определяется как:

$$\hat{\alpha} = \frac{m}{n}, \tag{2}$$

где m - число значений наблюдений, больших $\hat{a}$, а n - общее число наблюдений. Поскольку $\hat{\alpha}$ является состоятельной оценкой $P(X > \hat{a})$, то неточное задание $\hat{a}$, но такое что $\hat{a} > a$, приведет к пересчету параметра α и существенно не повлияет на оценку $P(X > c)$:

$$P(X > C) = \alpha e^{\lambda(\hat{a}-a)} e^{-\lambda(C-\hat{a})} = P(X > \hat{a}) e^{-\lambda(C-\hat{a})}; \tag{3}$$

поэтому из состоятельности оценки (2)

$$P(X > C) \approx \hat{\alpha} e^{-\lambda(C-\hat{a})}.$$

Если же $\hat{a} < a$, то наблюдения могут попасть на отрезок $[0 : a]$, где закон распределения случайной величины не известен, и тогда не будет выполняться свойство (3), что приведет к погрешности в определении $P(X > C)$.

Таким образом, можно сделать следующие выводы.

1. При построении оценки параметра a ее завышенное значение не будет иметь серьезных последствий для определения вероятности $P(X > C)$, поэтому отсутствие возможности состоятельно оценить этот параметр не является препятствием для решения поставленной задачи.

2. Не возникает существенных проблем, если $a \leq \hat{a} \leq C$.

3. Оценка параметра λ должна быть состоятельной или ее значение не должно превосходить значение оцениваемого параметра.

3. ОЦЕНКА ПАРАМЕТРОВ ХВОСТА РАСПРЕДЕЛЕНИЯ

Результаты наблюдений представим в виде вариационного ряда:

$$x_{(1)} \geq x_{(2)} \geq \ldots\ldots \geq x_{(k)} \geq x_{(k+1)} \geq \ldots\ldots \geq x_{(n)}; \tag{4}$$

где значения $x_{(i)}, i \in [1, k]$, приходятся на хвост распределения; значение k неизвестно, но позволяет задать a.

Предположим, что среди наблюдений достаточное их число попадает в хвост распределения. Это означает, что априори известно, что доля значений наблюдений из хвоста распределения не менее α_0.

Пусть $t = [\alpha_0 n]$, $[x]$ - целая часть числа x. В качестве оценки a используется

$$\hat{a} = x_{(t)}.$$

Понятно, что это заниженная оценка a, поскольку использовано минимальное значение этой величины, однако, как указывалось выше, это влияет только на точность окончательной оценки вероятности достижения заданного уровня C. Проблема получения более точной оценки a рассматривается в следующем разделе на основании результатов численного моделирования.

В качестве оценки параметра λ используется

$$\lambda_t = \frac{t}{\sum\limits_{i=1}^{t} (x_{(i)} - \hat{a})}. \tag{5}$$

Вычисленное таким образом значение λ_t является состоятельной оценкой параметра λ при $t \to \infty$ [2].

В качестве оценки вероятности попадания в хвост распределения в этом случае следует взять ее минимальное априорное значение α_0.

Таким образом, предложенный метод оценивания в соответствии с (3) позволяет получить состоятельную оценку для $P(X > C)$.

4. ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ

При компьютерном моделировании, для создания последовательности псевдослучайных чисел (ПСЧ), статистические свойства которых близки к свойствам равномерного распределения на участке $[0, 1]$, использовался линейный конгруэнтный алгоритм [3]. Этот алгоритм использует большой нечетный постоянный множитель и постоянное слагаемое вместе со значением SEED, инициализирующим начальное значение генератора, для итеративного создания случайных чисел и обновления SEED-значения. Получаемые значения пропускались через фильтр (рис. 2) для создания значений случайной величины, равномерно распределенной на интервале $[0, 1]$ и имеющей экспоненциально убывающий хвост распределения на интервале $(1; +\infty)$. Это означает, что при моделировании распределения $a = 1$. Значение параметра $\lambda = 1$, а α выбиралось с учетом варьирования количества значений случайных величин, попавших в хвост распределения, которое необходимо было менять от опыта к опыту.

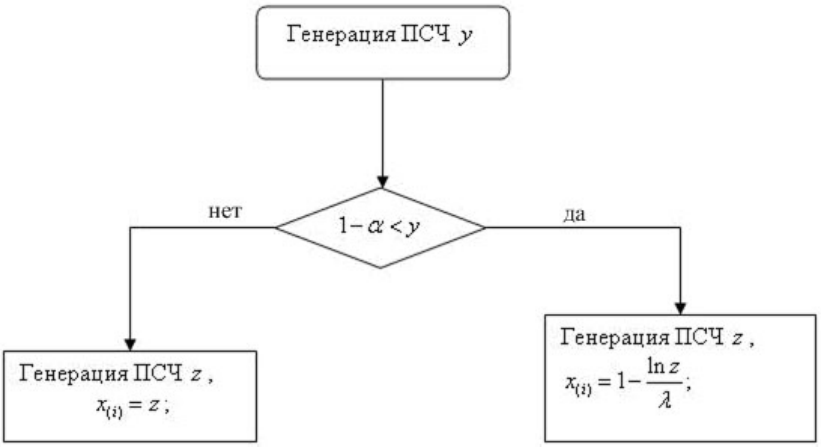

Рис.2. Алгоритм создания выборки значений случайной величины с требуемыми свойствами.

По результатам полученных значений строился вариационный ряд $x_1 \geq x_2 \geq x_3 \geq ..... \geq x_n$ и по формуле (5) вычислялся набор значений λ_t, $t = 10 + 5k$. На рис. 3 представлен график значений λ_t при заданном значении $\alpha = 1$.

На рис. 4 приведены траектории значений λ_t, найденных при заданном значении $\alpha = 0.5$ для различных выборок. Вертикальной прямой изображена граница между хвостом распределения и основной частью.

Из приведенного примера видно, что оценка λ_t при малых значениях t подвержена значительным флуктуациям. Стабилизация значений около правильного значения λ наступала при t порядка 300. Это означает, что в рассматриваемом случае $n = 100$, значение α_0 должно быть в интервале $(0.3; 0.5)$.

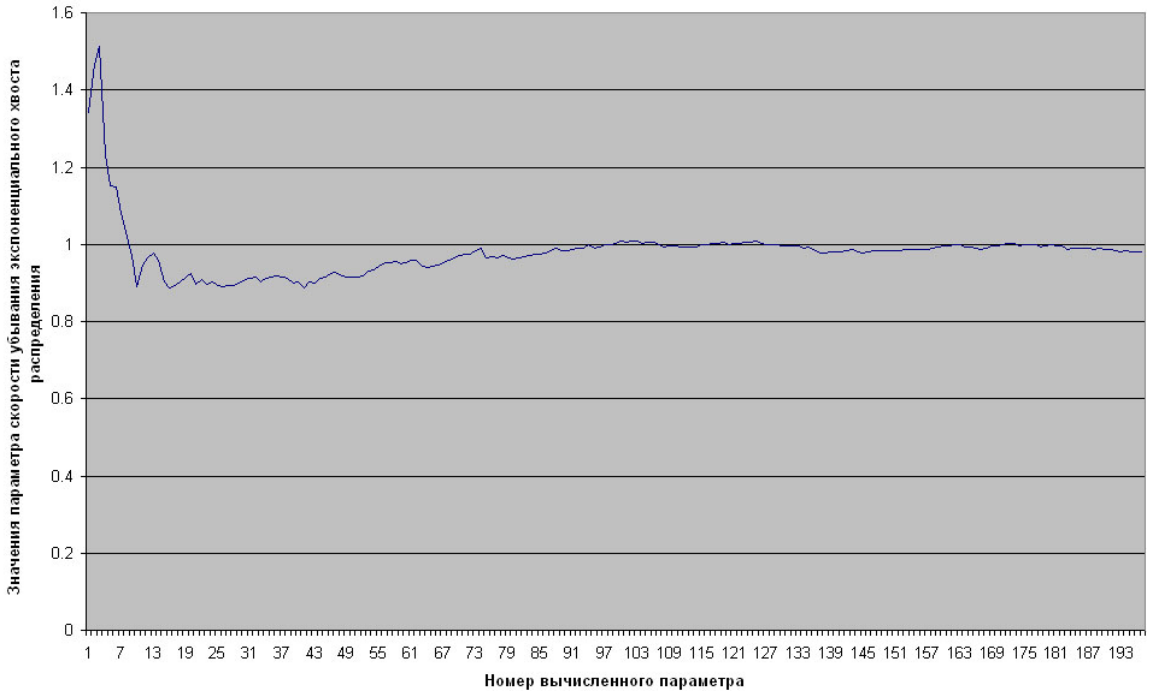

Рис.3. График значений λ_t, когда почти все значения вариационного ряда попали в экспоненциально убывающий хвост моделируемого распределения.

Анализ рисунка показывает, что задача состоятельного оценивания параметра a является сложной. Если учесть, что относительная стабилизация значения оценки λ наступает при относительной погрешности 10%, то отклонение λ_t на 10% вниз после перехода через точку a мы обнаруживаем при t порядка 750, что соответствует значению оценки параметра a 0.5. Как указывалось ранее, такая погрешность в определении значения a может быть неприемлемой.

5. ВЫВОДЫ

Задача оценки хвоста распределения для определения $P(X_i > C)$, сводится к нахождению параметров λ, a, α. Если они известны, то $P(X_i > C) = \alpha e^{-\lambda(C-a)}$,где $C > a$.

Найдены состоятельные оценки параметров λи α, требующие, однако, достаточно большого числа наблюдений. Отыскание оценки параметра a является сложной задачей.

Оценка значения $P(X_i > C)$ по результатам измерений не всегда представляется возможным, однако возможно задавать оценки параметров λ, a, α с учетом следующего:

1) условие $\hat{\lambda} < \lambda$ ведет к заниженной оценке хвоста распределения, а с точки зрения проверки гипотез – к завышению оценки вероятности ошибки, условие $\hat{\lambda} > \lambda$ - неприемлемо, так как приводит к заниженной оценке вероятности ошибки;

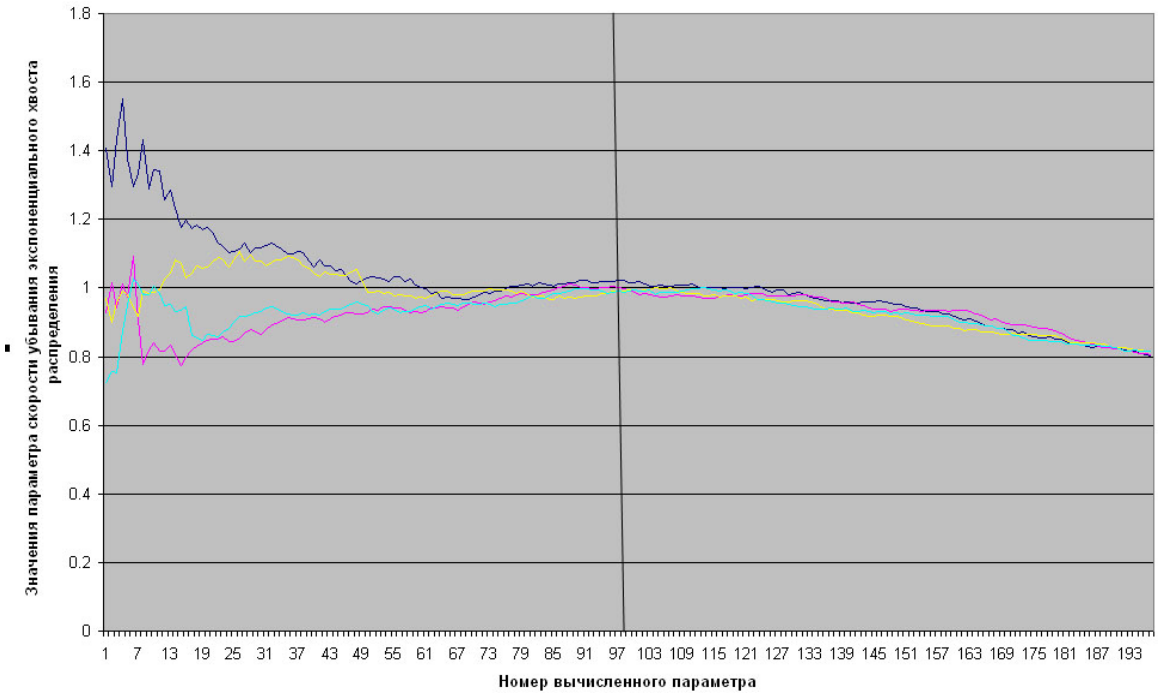

Рис.4. График значений λ_t, когда половина значений вариационного ряда попали в экспоненциально убывающий хвост моделируемого распределения.

2) условие $a \leq \hat{a} \leq C$ ведет к пересчету параметра α и существенно не влияет на значения $P(X_i > C)$, условие $\hat{a} < a$ нежелательно, так как ведет к погрешностям в определении $P(X_i > C)$, из-за повышения вероятности попадания в интервал, где закон распределения случайной величины неизвестен.

В случае, когда наблюдений недостаточно, целесообразно использовать смещенные оценки параметра λ, когда для его оценивания используется часть наблюдений не из хвоста распределения. Но в этом случае нужно в качестве оценки α брать 1, чтобы гарантировать оценку сверху для вероятности $P(X_i > C)$.

СПИСОК ЛИТЕРАТУРЫ

1. Боровков А.А., Боровков К.А. *Асимптотический анализ случайных блужданий. Том 1: Медленно убывающие распределения скачков*. М.: Физматлит, 2008.

2. Калинина В.Н., Панкин В.Ф. *Математическая статистика*. М.: Высш. школа, 1994.

3. Топп У., Форд У. *Структуры данных в C++*. М.: БИНОМ, 2000.