Йенс Брабанд[1], Хендрик Шебе[2]

ОЦЕНКА НАЦИОНАЛЬНЫХ ЭТАЛОННЫХ ЗНАЧЕНИЙ ДЛЯ ЖЕЛЕЗНОДОРОЖНОЙ БЕЗОПАСНОСТИ. СТАТИСТИЧЕСКАЯ ТРАКТОВКА

В настоящей работе обсуждается процедура принятия решения, используемая в Решении Комиссии [1] для определения национальных эталонных значений (NRV – national reference values). В соответствии с директивой в области безопасности, для каждого государства-члена ЕС ежегодно полагается рассчитывать показатели безопасности. В решении для расчетов показателей безопасности и оценки наличия возможной деградации безопасности использована фиксированная процедура. При оценке безопасности решение вместо среднеарифметического значения использует взвешенную сумму.

В этой связи интересно изучить особенности такой процедуры принятия решения, а также преимущества и недостатки данного метода. В данной работе двояким образом рассматривается несколько упрощенная версия процедуры. Во-первых, анализируется взвешенная сумма и выводится ее параметр эффективности. Затем она сравнивается путем разброса с обычным выборочным средним. Теоретические результаты подкрепляются простым имитационным моделированием, проводимым с целью оценки вероятностей ошибок первого и второго рода. В частности конструируются аномальные альтернативные распределения, которые не распознаются процедурой принятия решения.

Ключевые слова: *национальные эталонные значения, показатели железнодорожной безопасности, робастные оценки.*

1. Введение

В Решении Комиссии [1] для каждого государства-члена ЕС определяется набор так называемых значений NRV (национальных эталонных значений), который выводится на основе данных о происшествиях за несколько лет. Статистика происшествий измеряется в показателях FWSI (фатальные случаи и взвешенные серьезные травмы) для разных категорий людей – для пассажиров, персонала, участников дорожного движения на железнодорожных переездах, прочих граждан, лиц с несанкционированным доступом, для общества в целом. Выбор в качестве критерия показателя FWSI пассажиров оправдан тем, что происшествия, связанные с отдельными людьми, имеют относительно высокую вероятность и незначительные последствия в сравнении с проис-

[1] Siemens AG, Брауншвейг, Германия
[2] TÜV Rheinland InterTraffic GmbH, Кёльн, Германия

шествиями, связанными, например, со скоростными поездами, которые происходят редко, но потенциально означают весьма серьезные последствия. Впрочем, в данной работе рассматривается не конкретное значение NRV, а изучается общий подход и его особенности. Значения NRV затем устанавливаются в качестве целевых значений для последующих лет. Таким образом, ежегодно для каждого государства-члена ЕС и для каждого значения NRV должна оцениваться следующая гипотеза: «H_0: Имеются ли свидетельства того, что показатели безопасности государства-члена ЕС с учетом значения NRV ухудшились?»

Данная гипотеза как таковая хорошо известна при решении проблем, связанных с контролем качества. В этом случае оцениваются такие, например, параметры, как длина, ширина и т.д. Метод выборки применяется для того, чтобы выяснить, не появляются ли в процессе производства изделия, которые не соответствуют спецификации. При этом для контролируемого параметра постулируется существование нормального распределения со средним значением μ и разбросом σ, после чего рассчитываются предельные значения, и если определенное число изделий выходит за данные границы, считается, что процесс не контролируется. Подробное описание можно найти в многочисленных изданиях, посвященных статистическим методам контроля качества, например в [5]. В целях мониторинга процесса производства используются карты контроля качества, куда заносятся значения измерений наряду с порядковым номером или временем проведения измерения. Затем, в соответствии с применимыми правилами, на картах чертятся предупредительные границы и пределы, при выходе за которые требуется вмешательство в производственный процесс. В дальнейшем данные карты контроля качества были уточнены и компьютеризированы.

Таким образом, проблема, трактуемая в директиве [1] ERA (Европейского Железнодорожного Агентства), схожа с проблемой хорошо известных карт контроля качества.

Однако ERA [1] предлагается новый алгоритм, который никогда не анализировался со статистической точки зрения. Это как раз и является предметом рассмотрения настоящей статьи.

Во втором разделе будет описан подход, используемый ERA для мониторинга вводимых агентством показателей. В третьем разделе будет дан теоретический анализ основного этапа процедуры принятия решения ERA. После чего, в четвертом разделе, анализ будет подкреплен имитационным моделированием правила с использованием метода Монте-Карло. Последний раздел будет посвящен выводам.

2. Подход, используемый ERA

ERA намерено тестировать гипотезу H_0: «Показатели безопасности государства-члена ЕС не ухудшились с учетом значения NRV» путем использования процедуры, которая на самом деле является комплексным статистическим тестом, состоящим из нескольких этапов. Сначала мы дадим общее определение проблемы, которую хотим обсудить в данной статье:

Пусть X_i – годовые значения. Тогда конкретное скользящее взвешенное среднее (MWA) можно определить как взвешенную сумму:

$$\mathrm{MWA}_{q,p} = \frac{\sum_{i=p}^{q} w_{p,q}(i) X_i}{\sum_{i=p}^{q} w_{p,q}(i)}. \qquad (1)$$

Веса определяются следующим образом:

$$W_{p,q}(i) = 1/\max(\varepsilon;|X_i - \overline{X}_{p,q}|) \quad (2)$$

$$\text{где } \overline{X}_{p,q} = \frac{1}{q-p+1}\sum_{i=p}^{q} X_i \quad (3)$$

как арифметическое среднее. Значение ε полагается равным 0,01 в [1], вероятно, на основе эвристического рассуждения. Отметим, что ERA рассчитывает MWA только на основе пяти последних лет, которые обозначим здесь как $MWA_{n,n-4}$, после чего можем найти выражение правила многоступенчатого принятия решения, которое, по сути, обретает силу закона, будучи директивой ЕС, в следующем алгоритме:

Алгоритм

ЕСЛИ $NRV \geq X_n$ ИЛИ $NRV \geq MWA_{n,n-4}$, ТО ПРИНИМАЕМ H_0 ДРУГОЕ.

ЕСЛИ $NRV \times 1.2 \geq MWA_{n,n-4}$, ТО ПРИНИМАЕМ H_0 ДРУГОЕ.

ЕСЛИ «Первое отклонение (в шагах 1 или 2) за последние три года», ТО ПРИНИМАЕМ H_0 ДРУГОЕ.

ЕСЛИ «Число значительных происшествий не возросло», ТО ПРИНИМАЕМ H_0 ДРУГОЕ.

ОТБРАСЫВАЕМ гипотезу о том, что показатели безопасности не ухудшились.

Этот алгоритм был напрямую получен из описания, приводимого в [1]. При этом в [1] не приводится обоснование алгоритма. Следовательно, объясним, как этот алгоритм работает со статистической точки зрения. В настоящей статье нас особенно интересуют вероятности ошибок первого и второго рода, а именно что H_0 отбрасывается в случае, когда оно справедливо, и что H_0 принимается, будучи ложным. Обозначим совпадающие ошибки α и β, как вероятность ошибки первого и второго рода соответственно.

Начнем с простых наблюдений, сделанных нами во время анализа алгоритма, состоящего из шагов 1-5:

- Шаг 1 дает более или менее случайный результат с вероятностью принятия порядка 0.25 – 0.5 при разумных допущениях; таким образом, он не является важным показателем и не принимается здесь в расчет.
- В шаге 2 постоянное поле допуска кажется совершенно эвристическим.
- Шаг 3 также может считаться малоинформативным, так как данные последующих лет будут иметь зависимости, и вряд ли может дать какой-то ощутимый результат.
- В шаге 4 ERA эксплицитно рекомендует применение статистического теста и предлагает распределение Пуассона для числа серьезных происшествий в качестве нулевой гипотезы.

Основываясь на последнем наблюдении, можно сделать вывод, что ERA предлагает в качестве нулевой гипотезы составной процесс Пуассона, где N_t представляет собой количество происшествий до момента времени t, а S_1, S_2,… – независимую и одинаково распределенную (н.о.р.) последовательность случайных переменных, характеризующую степень серьезности происшествий, как было рассмотрено нами в [6]. Происшествия возникают с частотой λ(t) в моменты времени T_1, T_2 и т.д., где i – индекс происшествия. Для каждого происшествия с индексом i существует некоторая «высота прыжка» S_i, являющаяся степенью тяжести. Степень тяжести измеряется числом проис-

шествий с фатальным исходом (или эквивалентным ущербом). Таким образом, для фиксированного интервала [0,t] процесс X_t описывает сводное количество происшествий с фатальным исходом, а N_t – общее число происшествий.

Данный составной процесс Пуассона может использоваться для расчетов параметров коллективного риска.

Таким образом, исходя из нулевой гипотезы, наблюдаемая величина X_t является случайной переменной, которая, как можно предположить, распределяется следующим образом:

$$\sum_{i=1}^{N_t} S_i \,. \tag{4}$$

при постоянной частоте происшествий $\lambda(t) = \lambda$.

Следовательно,

$$E(X_t) = \lambda t E(S_1) \text{ и } V(X_t) = \lambda t E(S_1^2) \tag{5}$$

остаются верными, когда допускаем, что существуют моменты S_1. При условии достаточного числа происшествий в год, можно применить предельные теоремы для распределения X_t, например, центральную предельную теорему.

В следующем разделе будет проанализирована с теоретической точки зрения оценка (1), после чего будет исследован весь процесс принятия решения с привлечением статистического имитационного моделирования.

3. Теоретический анализ шага 2 алгоритма

При рассмотрении (1) с теоретической точки зрения, видно, что

a) число наблюдений в выборке мало, поскольку используется всего 5 значений,

b) среднее (1) – взвешенное среднее, где сам вес (2) выводится из выборки,

c) вес (2) усечен снизу, с тем чтобы отграничить его от ноля, а выбор константы усечения ε несколько произволен,

d) вес (2) снижает влияние больших наблюдений, так как обратно пропорционален разнице между наблюдением и выборочным средним.

Форма оценки (1) хорошо известна в робастной статистике. Ее можно встретить в [2], Раздел 2.3d, в виде так называемой W-оценки Тьюки, введенной Тьюки [3] с весовой функцией (2). Тем самым результаты из [2] можно напрямую использовать для специальной формы (1) данной W-оценки. Было продемонстрировано в [2], что статистическое поведение (1) точно такое же, как и поведение M-оценки Хьюбера с функцией $\psi(u)=u\,w(u)$. Следует напомнить, что M-оценка Хьюбера определяется как оценка T_n, удовлетворяющая равенству

$$\sum_{i=1}^{n} \psi(X_i; T_n) = 0 \,, \tag{6}$$

Рассмотрим, например, [2]. Теперь теорию робастных оценок можно просто применить к (1), чтобы посмотреть, каковы свойства. Двумя основными свойствами являются робастность и эффективность.

а) Робастность

Оценка называется робастной, если она устойчива к выбросам, присутствующим в выборке. Выбросы – это значения, выпадающие из нормального пространства значений выборки, которое состоит из одинаково распределенных случайных переменных.

Могут применяться два основных параметра робастности оценки. Первый – так называемая точка срыва. Это тот порог выбросов в выборке, до которого оценка все еще дает непротиворечивые результаты. То есть, несмотря на то что какая-то часть данных в выборке выпадает из совокупности, параметры совокупности оцениваются корректно. После некоторых расчетов можно обнаружить, что точка срыва определятся на уровне 50%. Это справедливо для выборки любого размера. Любое наблюдение, если оно достаточно крупное, разделяется своим абсолютным отличием от выборочного среднего. Выборочное среднее подвергается искажению только в том случае, если более половины наблюдений X_i являются выбросами.

Второй важной характеристикой является функция влияния, введенная Хампелем. Для статического значения T, построенного на основе независимых и одинаково распределенных случайных переменных с общей функцией распределения F, и выброса y она определяется как

$$ЕСЛИ(y;T;F)=\lim_{t\to 0}\frac{T(1-t)F+t\Delta_y)-T(F)}{t} \tag{7}$$

См. [2, Раздел 2.1b]. Δ_y означает распределение вероятности приведения массы 1 к точке y. Функция влияния показывает, насколько сильно единичное наблюдение может повлиять на саму оценку. Это можно легко посчитать, если дать одному наблюдению стремиться к бесконечности и посмотреть, как при этом меняется оценка. Если y – значение выброса, а m – выборочное среднее значений X_i, можно увидеть, что функция влияния среднеарифметического (3)

$$ЕСЛИ(y)=\lim_{t\to 0}\frac{(1-t)m+ty-m}{t}=y-m. \tag{8}$$

Как правило, робастная оценка характеризуется ограниченной функцией влияния. Можно заметить, что функция влияния нормального выборочного среднего не ограничена и возрастает вместе со значением выброса y. Вот почему выборочное среднее не является робастной оценкой среднего совокупности.

Оценка (1) может выражаться в виде

$$T(F)=\frac{\int w(x)xdF(x)}{\int w(x)dF(x)},$$

где

$$w(x)=\frac{1}{\left|x-\int xdF(x)\right|}.$$

Здесь и далее целые располагаются на всей вещественной оси.

Для оценки (1) функция влияния выглядит как

$$ЕСЛИ(y) = y\frac{\int wdF(x)\int \beta^2 x\,|x-y|\,dF(x) - \int wxdF(x)\int \beta^2\,|x-y|\,dF}{\beta^2} + O(1), \quad (9)$$

где m – среднее F, w – взвешенная функция и β = |x-m|. Кроме того, O(1) означает член порядка 1 по сравнению с y, когда y стремится к бесконечности.

Функция влияния (9) не ограничена, если y возрастает. Следовательно, одно единичное наблюдение X_i может повлиять на оценку среднего таким образом, что если единичное наблюдение стремится к бесконечности, то и оценка среднего тоже.

Вывод этого небольшого исследования состоит в том, что оценка (1) не является робастной.

b) Эффективность

Для того чтобы получить дисперсию оценки (1), результаты [2] просто используются для оценок типа местоположения в Разделе 2.2b. Дисперсия (variance) выводится следующим образом

$$Var(\psi, F) = \frac{\int \psi^2 dF}{\left(\int \psi' dF\right)^2}. \quad (10)$$

Теперь, памятуя о том, что ψ(u)= u w(u), при w(u) = max(εm;|u-m|), видим, что

$$\psi'(u) = w(u) + u\,w'(u) = \begin{cases} \max(\varepsilon m, |u-m|) + u \text{ for } \varepsilon m < |u-m| \\ \max(\varepsilon m, |u-m|) \quad \text{иначе} \end{cases} \quad (11)$$

Если пренебречь влиянием ε, которое в любом случае мало, и попросту считать его равным нолю, получаем

$$Var(\psi, F) = \frac{\int x^2 (x-m)^2 dF}{\left(\int (u+|u-m|)dF\right)^2}. \quad (12)$$

Можно пренебречь ε, поскольку оно, похоже, введено для того, чтобы избежать деления на ноль и сделать вычисление возможным.

Отметим, что данное значение должно быть поделено на n, чтобы получить дисперсию оценки, построенной на выборке размера n. Эту величину необходимо сравнить с дисперсией простого среднеарифметического, то есть

$$Var(F) = \int (x-m)^2 dF = \sigma^2. \quad (13)$$

Из равенства (12) получаем

$$Var(\psi, F) = \frac{m_4 - 2m_3 m + m^4}{\left(m + \int |x-m|\,dF\right)^2}, \quad (14)$$

где m_4 и m_3 – обычные четвертые и третьи моменты распределения. Так как значения X суть суммы других случайных переменных, нормальное распределение может считаться хорошим приближением. Не забудьте, что мы допустили существование моментов значений S, так что гарантируется слабая сходимость к нормальному распределению. Правда, приближение будет хорошим только тогда, когда число событий достаточно велико. Так что для малого числа событий приближение будет грубым. Теперь для нормального распределения (14) получаем

$$Var(\psi;F)=(2m^4 + 6m^2\sigma^2 + 3\sigma^4)/(m+2\sigma)^2. \quad (15)$$

Делим (15) на (13) и получаем

$$Eff = (2+6v^2+3v^4)/(v^2+4v^3+4v^4), \quad (16)$$

при $v = \sigma/m$ в качестве коэффициента вариации. Eff представляет собой относительную эффективность робастной оценки, связанную с нормальным выборочным средним. Можно заметить, что она стремится к бесконечности, как v стремится к нолю, т.е. оценка (1) имеет гораздо больший разброс, чем обычное выборочное среднее. Даже при v=1 можно видеть, что Eff = 11/7, т.е. дисперсия робастной оценки примерно на 57% больше, чем дисперсия обычной оценки.

Следовательно, чем больше число наблюдений за единичный год, тем меньше значение v и, таким образом, тем больше потеря эффективности оценки (1).

Примечания:

1. Мы пренебрегли тем фактом, что оценка (1) не является итерированной оценкой, т.е. среднее вычисляется для оценки (1) один раз, и вычисление оценки не повторяется при получении нового среднего, после того как (1) рассчитано для улучшения оценки. Правда, введенная при этом погрешность имеет порядок 1/n, где n – размер выборки.
2. При вычислении эффективности мы заменили выборочное среднее средним совокупности в весовой функции, тем самым введя погрешность порядка 1/n.
3. Целые числа без указания пределов располагаются на всем пространстве определения функции распределения F.

В заключение нашего теоретического исследования, мы видим, что оценка (1) утратила эффективность, применяет меньшие веса к большим наблюдениям, но не является робастной. Что это означает на практике? Внезапное изменение статистики за один год или несколько лет не окажет влияния на статистику так, как это произошло бы в случае с невзвешенной оценкой, т.е. выборочным средним. Кроме того, разброс оценки гораздо больше, чем разброс обычного выборочного среднего. Это означает, что мы наблюдали бы гораздо больше «случайного шума» в нашем скользящем взвешенном среднем MWA, т.е. в MWA могли бы присутствовать ненужные случайные колебания, в отличие от использования выборочного среднего.

4. Результаты имитационного моделирования

Для подкрепления теоретических изысканий Раздела 2 был проведен ряд имитационных тестов на разных моделях функций распределения.

Моделируются ежегодные н.о.р. наблюдения заданного распределения со средним NRV, чтобы проверить нулевую гипотезу σ. Некоторые эксперименты оставили впечатление, что для такого симметричного распределения с таким регулярным поведением не было заметно какой-то большой разницы между решениями, основанными на нормальном усреднении (3) и взвешенном

скользящем среднем (1). Так что в качестве верхнего предела α можем постулировать нормальное приближение

$$P(N > \frac{NRV}{\sigma\sqrt{5}}), \tag{17}$$

где N – стандартная нормальная случайная переменная. Данная аппроксимация выводится из того факта, что мы усредняем 5 значений NRV согласно (1) и допускаем нормальное приближение. Наши имитационные тесты подтверждают, что распределение MWA, действительно, более плотно центрируется вокруг NRV, при этом погрешность на удивление мала. Это вполне логично, ведь MWA складывается из нескольких значений NRV.

Для второго эксперимента берем гамма-распределение в качестве несимметрического решения и сравниваем результаты перекоса распределения для равного σ.

Результаты в таблице 1 показывают большую зависимость вероятностей погрешности для шага 2 от σ. При малом σ, α практически равно нолю, но при изменениях в среднем значении менее чем на 20%, β близко к 1, в то время как при большем σ, α возрастает, а β убывает. Заметим, что среднее было выбрано равным единице, так что σ также является коэффициентом дисперсии.

Таблица 1. Результаты имитационного моделирования вероятности ошибки первого рода α

Эксперимент	σ=20%	σ=30%	σ=50%
Нормальное приближение	0,013	0,068	0,19
Нормальное MWA	0,01	0,074	0,18
Нормальные шаги с 1 по 3	0,0043	0,032	0,08
Гамма-MWA	0,029	0,084	0,15
Гамма-шаги 1-3	0,011	0,032	0,06

Диапазон значений σ, представленный в таблице 1, может считаться вполне репрезентативным, как демонстрирует подробная оценка швейцарской статистики происшествий. К сожалению, детали результатов являются конфиденциальными. Результаты также показывают адекватность нормального приближения, в особенности для больших значений σ. Они также показывают, что шаги 1 и 3 при комбинировании только обеспечивают фактор между 2 и 3, что указывает на высокую степень зависимости. Отметим, что мы не моделировали шаг 4, поскольку для этого случая значение α определено ERA как равное 0,05, но также с высокой степенью зависимости. Кроме того, нам пришлось бы сделать гораздо больше допущений в отношении составного процесса Пуассона.

Также были проведены и другие качественные эксперименты, чтобы выявить некую «плохую» альтернативную гипотезу, которую вряд ли можно распознать во время процедуры принятия решения. В этих экспериментах н.о.р. наблюдения были искажены фиксированными или рандомизированными линеарными трендами либо периодически возникающими выбросами, репрезентирующими крупномасштабные происшествия. Общее наблюдение таково: процедура принятия решения довольно медленно обнаруживает линеарный тренд, и требуется, как правило, от 4 до 5 лет, чтобы нулевая гипотеза была отвергнута процедурой, после того как тенденция уже стала заметной в данных. Этот эффект можно объяснить комбинированным действием подавления выбросов скользящим усреднением в шаге 2 и «однократным» исключением в шаге 3. Как и ожидалось, выбросы также довольно хорошо подавляются процедурой принятия решения, например, при низком σ, скажем 10% от NRV, процедура принятия решения допускает крупные выбросы, случающиеся не реже, чем раз в три года, что на практике не могло бы допускаться.

5. Заключение

В настоящей статье рассмотрена, как с точки зрения теоретической трактовки, так и путем имитационного моделирования с использованием метода Монте-Карло, процедура принятия решения, представленная в [1]. Выяснилось, что взвешенная сумма (MWA), используемая в процедуре, имеет больший разброс, чем обычное выборочное среднее. Заметим, что оценка максимального правдоподобия всегда асимптотически эффективна, при условии, что выполняются условия регулярности. В нашем случае оценка максимального правдоподобия – выборочное среднее, и оно оказывается эффективной оценкой. Оценка применяет меньшие веса к более крупным наблюдениям, но не является робастной. Метод MWA дает хорошее усредненное значение, но теряет эффективность, в отличие от выборочного среднего. Также оказалось, что если вероятность ошибки первого рода довольно мала для типичных параметров, то вероятность ошибки второго рода может быть весьма высокой, так как, в частности, большие значения при этом не принимались бы в расчет малыми весами, а следовательно, тенденции в альтернативных гипотезах не распознавались бы на ранней стадии.

Для статистических процедур принятия решений также не свойственно учитывать в процедуре выборочную вариацию, а значит, в самой процедуре не было предусмотрено какого-либо фиксированного целевого значения для вероятности погрешности первого рода.

Если бы нужна была оценка с ограниченной робастностью, то можно было бы использовать так называемое α-усеченное среднее. Данная оценка пренебрегает наименьшими γ/*100% и наибольшими γ/*100% наблюдениями в выборке и рассчитывает только среднее значение по отношению к оставшимся элементам выборки. Такая оценка гораздо проще, а ее точку срыва (2γ) можно изменить в зависимости от усечения. В этом случае не надо было бы применять взвешивание, а потеря эффективности была бы гораздо меньшей, чем в случае с настоящей оценкой.

Благодарность

Авторы выражают благодарность двум анонимным рецензентам за помощь в редактировании статьи.

Литература

1. Решение Комиссии от 5 июня 2009 года о принятии единого метода безопасности при оценке соответствия целевым показателям безопасности, в соответствии со Статьей 6 Директивы 2004/49/ ЕС Европейского Парламента и Совета; 2009/460/ЕС.

2. **F.R. Hampel, E.M. Ronchetti, P.J. Rosseeuw, W.A.Stahel**, Robust Statistics – The Approach based on Influence Functions, Wiley 1986 [**Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. Робастность** в статистике. Подход на основе функций влияния].

3. **J.W. Tukey**, Exploratory Data Analysis, Addison Wesley 1977 [Тьюки Дж.Н. Анализ результатов наблюдений].

4. Директива 2004/49/ЕС Европейского Парламента и Совета от 29 апреля 2004 года о безопасности железных дорог.

5. **A.J. Duncan,** Quality Control and Industrial Statistics, Homewood / Illinois 1965 [Дункан А.Дж. Контроль качества и промышленная статистика].

6. **J. Braband, H. Schäbe**, The collective risk, the individual risk and their dependence on exposition time, ESREL 2011, Proceedings , Advances in Safety, Reliability and Risk Management, pp. 1783-1787 [**Брабанд Й. Шебе Х.** Коллективный риск, индивидуальный риск и их зависимость от времени их изложения].