Т. М. Болотская, Б. А. Щукин

МЕТОД ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ ДЛЯ ВЫЯВЛЕНИЯ «ПОДОЗРИТЕЛЬНЫХ» ТЕРМИНАЛОВ

Постановка задачи

Служба безопасности банка, которому принадлежат POS-терминалы в некоторой торговой точке, нуждается в методах, которые позволили бы выявлять «подозрительные» терминалы, осуществляющие мошеннические действия при обработке кредитных карт. Видов мошенничества много, в данной статье анализируется случай, при котором в транзакциях, отправляемых в банк, завышается значение суммы, пробиваемой в чеке. Это завышение может быть значительным, особенно для лиц, совершающих туристические поездки или бизнес-командировки в разные страны: достаточно велика вероятность того, что покупатель не сможет предъявить никаких претензий. Т. е. если чек покупателя содержит сумму y, то в транзакции, отправляемой в банк, проставляется сумма $K_3 y$, где K_3 — коэффициент, больший 1.

Суммы x и $K_3 x$ за операционный день имеют распределения, отличающиеся математическим ожиданием. При этом можно выдвинуть гипотезы. Основную — H_0: $m = m_{H_0}$, где m_{H_0} — математическое ожидание суммы нормальной активности, вычисляемое как среднее арифметическое за все дни l месяцев, и альтернативную — H_1: $m = m_{H_1} = K_3 m_{H_0}$, где m_{H_1} — математическое ожидание суммы «мошеннической» активности, т. е. задачу обнаружения «подозрительных терминалов» можно свести к задаче проверки статистических параметрических гипотез.

Следует отметить, что единственным методом получения достоверных сведений о работе терминала является сравнение сумм чеков, подписанных клиентом, с суммой соответствующих транзакций, отправленных в банк. Эта операция достаточно трудоемка, требует ручного труда и может быть выполнена только ограниченное число раз. На основании этой статистики формируется математическое ожидание нормальной активности и математическое ожидание «мошеннической» активности.

Сведение к методу проверки статистических гипотез позволяет автоматизировать мониторинг проверки работы терминалов и выявление «мошеннических» терминалов с заданными ошибками.

Критерий проверки гипотез

В качестве статистики критерия значимости обычно используется статистика $Z = \frac{(\bar{x} - m)\sqrt{n}}{S}$, где $\bar{x}$ — выборочное среднее, m — математическое ожидание, S — выборочная дисперсия (по выборке из $L = 30l$ дней), n — объем выборки для $\bar{x}$. Для рассматриваемого случая сравнивается сумма x, за 1 прошедший текущий день с m_{H_0}, $n = 1$, и используется статистика критерия: $Z = \frac{x - m}{S}$. Ввиду массовости операций на торговых точках логично предположить нормальное распределение x. По критерию χ^2 проверялась гипотеза: $F(x) = N(m,\sigma)$ [1]. Легко доказать, что закон распределения статистики критерия Z есть распределение Стьюдента с $L - 1$ степенями свободы $T(L - 1)$. Так как $L > 30$, распределение Стьюдента можно заменить нормированным нормальным распределением $N(0,1)$, что дает возможность при расчете ошибок использовать квантили нормированного нормального распределения. С результатами проверки гипотез H_0 и H_1 связаны ошибки: α первого рода и β второго рода.

Расчет параметров мониторинга

Для рассматриваемого случая α представляет собой вероятность «ложного срабатывания» (т. е. вероятность принять нормальную активность за мошенническую), а величина ошибки β— вероятность «пропуска цели» (т. е. вероятность принять мошенническую активность за нормальную).

Для выявления «подозрительного» терминала вычисляем:

$\alpha = P\{z > z_{кр} \,|H_0\}$, … т. е. $\alpha = 1 - F_N\left(\frac{m_{H_0}(K_1 - 1)}{S}\right)$,

где $z_{кр} = \frac{x_{кр} - m_{H_0}}{S}$, F_N — функция распределения нормированного нормального распределения, $x_{кр} = K_1 m_{H_0}$, K_1 — некоторый неизвестный коэффициент, определяющий критическую область (параметр мониторинга). Откуда

$$K_1 = \frac{S}{m_{H_0}} U_{1-\alpha} + 1, \qquad (1)$$

где $U_{1-\alpha} = z_{кр} \,|H_0$ — квантиль нормированного нормального распределения с уровнем значимости $(1 - \alpha)$.

Ошибка 2-го рода β:

$\beta = P\{z < z_{кр} \,|H_1\} = P\{-\infty < z < z_{кр}\}$, $\beta = F_N\left(\frac{m_{H_0}}{S}(K_1 - K_3)\right)$,

где $z_{кр} = \frac{x_{кр} - m_{H_1}}{S}$, $x_{кр} = K_1 m_{H_0}$, $m_{H_1} = K_3 m_{H_0}$, где K_1, K_3 — некоторые коэффициенты (параметры мониторинга, причем K_3 известен из определения $m_{H_1} = K_3 m_{H_0}$), откуда по свойству функции распределения F_N:

$$\frac{m_{H_0}(K_1 - K_3)}{S} = U_\beta \ . \qquad (2)$$

Коэффициент K_3 определяет математическое ожидание значений суммы, характеризующей мошеннические действия. Коэффициент K_1 определяет границу критической области.

Итак, предлагается подход, с помощью которого служба информационной безопасности банка, отвечающая за безопасность операций, может получить конкретные количественные оценки для принятия решений. Из уравнений (1) и (2), связывающих три неизвестных параметра α, β, K_1 (K_3 известен), и задавая один из них можно получить оценки двух других и принять ту или иную гипотезу о работе терминала с известными ошибками 1-го и 2-го рода. Например, если задать ошибку 1-го рода α, из уравнения (1) получим параметр мониторинга K_1, а из уравнения (2) получим ошибку 2-го рода β.

Проверяя отношение выручки x в конце текущего дня определенного терминала к математическому ожиданию нормальной активности этого терминала m_{H_0} получаем коэффициент $K_2 = \frac{x}{m_{H_0}}$. Если $K_2 < K_1$, то может быть принята гипотеза H_0 — терминал работает нормально (эвристика), теория же требует вычисления величины $z_{выб} = \frac{x - m_{H0}}{S}$, при этом если $z_{выб} < U_{1-\alpha}$, то принимаем гипотезу H_0 — терминал работает нормально при вычисляемой ошибке 2-го рода β.

Распределение статистики случайной величины X проверки гипотез при H0 и H1

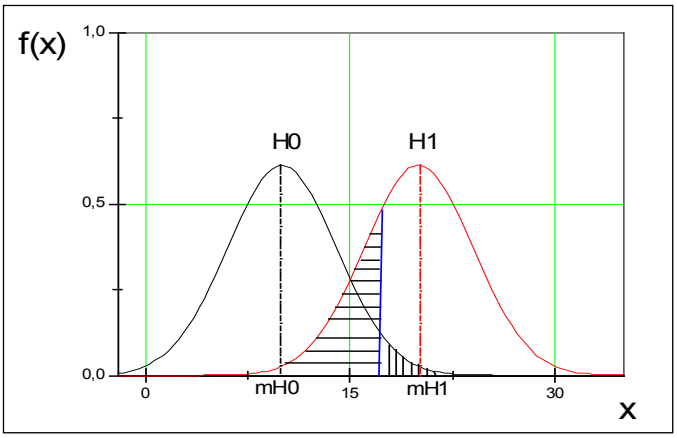

Рис. 1. Распределение статистики случайной величины X

Вычисление предлагаемых коэффициентов (формулы 1, 2) позволяет службе безопасности банка оценивать результаты мониторинга работы терминалов. Если для службы наиболее значим вариант, связанный с пропуском «цели», т. е. «подозрительного» терминала, то следует уменьшать β, хотя при этом неизбежно увеличивается α, что приводит к увеличению числа терминалов, подлежащих проверке. Это видно из приведенного ниже рисунка зависимости между ошибками 1-го и 2-го родов (α и β).

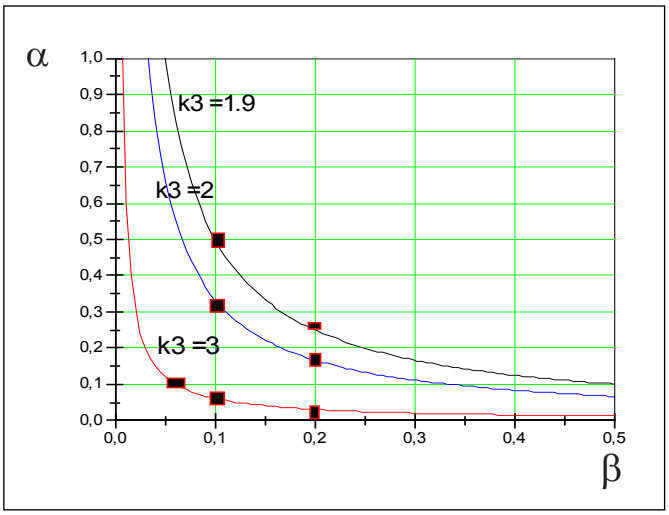

Рис. 2. Зависимость между ошибками α и β

СПИСОК ЛИТЕРАТУРЫ:

1. *Лемешко Б. Ю.* Статистический анализ одномерных наблюдений случайных величин: Программная система. Новосибирск: Изд-во НГТУ, 1995. – 125 с.