МИНИСТЕРСТВО ОБРАЗОВАНИЯ
И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

XII ЭКОНОМИЧЕСКИЕ ЧТЕНИЯ

памяти заслуженного деятеля науки РФ
профессора Александра Петровича Бычкова

Сборник материалов
Международной научно-практической конференции
26–27 октября 2017 г.

Под общей редакцией профессора
Д.М. Хлопцова

Томск
Издательский Дом Томского государственного университета
2018

ОБНАРУЖЕНИЕ СЛУЧАЙНЫХ ВЫБРОСОВ ПУТЕМ ПОСТРОЕНИЯ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

Шерстобитова А.О., магистрант, 2-й курс, **Клемешова А.И., Степанова Е.А.,** бакалавры, 4-й курс НИ ТГУ, г. Томск
Научный руководитель: канд. физ.-мат. наук, доцент Емельянова Т.В.
E-mail: annasherstobitova06@gmail.com, anya-3.4@mail.ru, zhenyutka@mail.ru

Задача построения доверительных интервалов широко востребована при решении широкого класса практических задач. Одной из таких задач является отыскание верхней границы количества людей, находящихся в определенной локальной зоне интереса (например, в зоне накопления аэропорта, в вагоне электропоезда, в некоторой зоне магазина).

Доверительный интервал является показателем точности измерений. Это также показатель того, насколько стабильна полученная величина, т.е. насколько близкая величина к первоначальной величине получится при повторении измерений эксперимента. Главное преимущество интервальной оценки – возможность характеризовать уверенность в вычисленных параметрах распределения.

Каждый час камерами видеонаблюдения фиксируется количество посетителей в каждой зоне магазина. Всего четыре зоны. Задачей исследования является исключение аномального количества посетителей в каждой из четырех зон в каждый час в течение суток.

Теоретические сведения

В работе рассматривается доверительное оценивание скалярного параметра. Для этого находят две такие статистики $T_1 = T_1(X)$ и $T_2 = T_2(X)$, $T_1 < T_2$, для которых при заданном $\gamma \in (0,1)$ выполнено условие

$$P_\theta\big(T_1(X) < \theta < T_2(X)\big) \geq \gamma, \forall\theta \in \Theta. \quad (1)$$

Полученный интервал $(T_1(X), T_2(X))$ называется ***γ- доверительным интервалом*** для параметра θ. Число γ – это ***доверительный уровень*** или ***доверительная вероятность***.

Условие (1) означает, что проводится большое число, не зависящих друг от друга экспериментов, в каждом из которых по n наблюдений над случайной величиной ξ. Далее оценивается параметр θ, используя следующее статистическое правило: если результаты эксперимента оцениваются выборкой X, то неизвестное значение параметра θ лежит внутри интервала $(T_1(X), T_2(X))$ [1]. На практике часто пользуются значениями доверительного уровня γ из небольшого набора заранее выбранных, достаточно близких к 1 значений. Например, $\gamma = 0.9, 0.95, 0.99$.

Иногда рассматриваются односторонние доверительные интервалы:

• верхний вида $\theta < T_2(X)$;

• нижний вида $T_1(X) < \theta$.

Они определяются условиями, которые аналогичны (1), где опускается соответствующая вторая граница.

Асимптотические доверительные интервалы

Как известно, если имеется состоятельная асимптотически нормальная оценка $T_n = T_n(X)$, $X = (X_1, X_2, \dots, X_n)$, для параметра θ, то строится асимптотический (при больших n) доверительный интервал.

В общем случае пусть $n \to \infty$и имеет место соотношение

$$L_\theta\left(\sqrt{n}(T_n - \theta)\right) \to N\left(0, \sigma^2(\theta)\right), \forall\theta \in \Theta,$$

причем асимптотическая дисперсия $\sigma^2(\theta)$ непрерывна по θ. Тогда

$$L_\theta\left(\frac{\sqrt{n}(T_n - \theta)}{\sigma^2(T_n)}\right) \to N(0,1), \forall\theta \in \Theta.$$

Отсюда следует, что при $n \to \infty$ и всех θ

$$P_\theta\left\{\frac{\sqrt{n}|T_n - \theta|}{\sigma(T_n)} < c_\gamma\right\} = 2\Phi\left(c_\gamma\right) - 1 = \gamma,$$

если $c_\gamma = \Phi^{-1}(\frac{1+\gamma}{2})$. Переписав это соотношение в более удобном виде, получим

$$P_\theta\left\{T_n - \frac{c_\gamma\sigma(T_n)}{\sqrt{n}} < \theta < T_n + \frac{c_\gamma\sigma(T_n)}{\sqrt{n}}\right\} \to \gamma.$$

Получается, что $\left(T_n \pm \frac{c_\gamma\sigma(T_n)}{\sqrt{n}}\right)$ – асимптотический – доверительный интервал для θ [2].

Доверительный интервал для параметра бернуллиевской модели

Требуется построить доверительный интервал для параметра θ модели $B(1,\theta)$. Если имеется соответствующая выборка $X = (X_1, X_2, \dots, X_n)$, то оптимальной несмещенной оценкой для θ является выборочное среднее $T = \bar{X}$. Статистика T принимает значения вида $\frac{k}{n}, k = 0,1, \dots, n$, при этом

$$F_T\left(\frac{k}{n}, \theta\right) = P_\theta\left\{T \le \frac{k}{n}\right\} = P_\theta\left\{\sum_{i=1}^{n} X_i \le k\right\} = \sum_{\gamma=0}^{k} C_n^\gamma\, \theta^\gamma(1-\theta)^{n-\gamma}.$$

Здесь

$$\frac{d}{d\theta}F_T\left(\frac{k}{n}, \theta\right) = -nC_{n-1}^k\theta^k(1-\theta)^{n-k-1} < 0, \text{ при } k < n,$$

поэтому при $k < n$ функция $F_T(\frac{k}{n}, \theta)$ монотонно убывает по θ. Следовательно, при $T = \frac{k}{n}$ центральным -доверительным интервалом для θ явля-

ется интервал (θ_1, θ_2), где θ_1 и θ_2определяются в соответствии с уравнениями

$$F_T(t_1, \theta) \le \frac{1-\gamma}{2}, 1 - F_T(t_2 - 0, \theta) \le \frac{1-\gamma}{2},$$

которые в данном случае принимают вид

$$1 - F_T\left(\frac{k-1}{n}, \theta_1\right) = \sum_{\gamma=0}^{k} C_n^\gamma \ \theta_1^\gamma (1-\theta_1)^{n-\gamma} = \frac{1-\gamma}{2},$$

$$F_T\left(\frac{k}{n}, \theta_2\right) = \sum_{\gamma=0}^{k} C_n^\gamma \theta_2^\gamma (1-\theta_2)^{n-\gamma} = \frac{1-\gamma}{2}. \quad (2)$$

При этом, если наблюдалось $T = 1$ (т.е. $k = n$), принимается $\theta_2 = 1$, а при $T = 0$ (т.е.$k = 0$) полагается $\theta_1 = 0$ [3].

Практически границы θ_1и θ_2 вычисляют, пользуясь их связью с квантилями бета-распределения:

$$\sum_{\gamma=0}^{k} C_n^\gamma \theta_2^\gamma (1-\theta_2)^{n-\gamma} = B(\theta_1, k, n-k+1),$$

$$\sum_{\gamma=0}^{k} C_n^\gamma \theta_2^\gamma (1-\theta_2)^{n-\gamma} = 1 - \sum_{\gamma=k+1}^{n} C_n^\gamma \theta_2^\gamma (1-\theta_2)^{n-\gamma} =$$

$$= 1 - B(\theta_2, k+1, n-k).$$

Уравнения (2) принимают вид

$$B(\theta_1, k, n-k+1) = \frac{1-\gamma}{2}, B(\theta_2, k+1, n-k) = \frac{1-\gamma}{2}. \quad (3)$$

Если обозначить $z(p, a, b)$ p-квантиль бета-распределения $Be(a, b)$, т.е. решение уравнения $B(p, a, b) = p, 0 < p < 1$, то из (3) получим

$$\theta_1 = z\left(\frac{1-\gamma}{2}, nT, n - nT + 1\right), \theta_2 = z\left(\frac{1-\gamma}{2}, nT + 1, n - nT\right). \quad (4)$$

Доверительные интервалы (θ_1, θ_2) для широкого диапазона значений при $\gamma = 0,9; 0,95; 0.99$.

Практическая часть. Из постановки задачи видно, что мы имеем дело со схемой Бернулли: проводится n опытов, в каждом из которых может произойти определенное событие – «успех» – с вероятностью p или не произойти – «неудача» – с вероятностью $1 - p$. В нашем случае «успех» – попадание в зону, неудача – непопадание.

Для обработки данных были предоставлены наблюдения почти за 3 месяца.

Проводя первичную обработку данных, мы выявили, что две зоны имеют нулевые показатели. Были построены гистограммы, в которых отображено количество людей в каждый час по двум зонам.

Далее данные были усреднены по часам в течение дня. Отдельно рассматривалось воскресенье, так как в этот день посещаемость существенно мала по сравнению с остальными днями. Данные представлены в виде гистограмм.

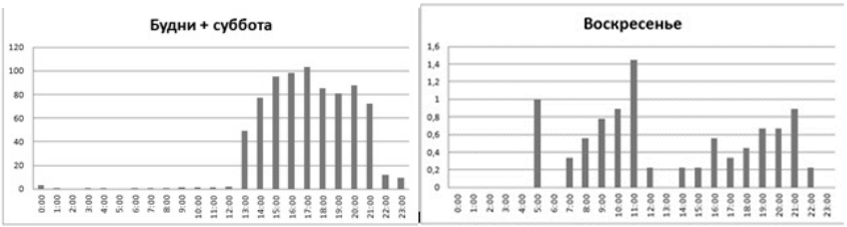

Рис. 1. Распределение посещаемости в зоне 1

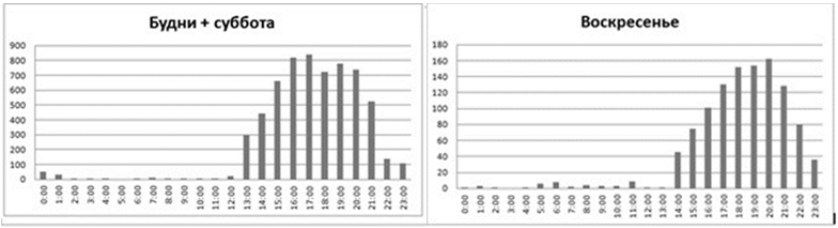

Рис. 2. Распределение посещаемости в зоне 2

Видно, что распределение посетителей близко к нормальному. Следовательно, для построения доверительных интервалов можно применить интегральную теорему Муавра–Лапласа. В явном виде ее применить не можем, так как вероятность p неизвестна. Доверительные интервалы будем строить для вероятности с уровнем доверия $\gamma = 0.95$.

Имеем

$$P\left(\left|\frac{m-np}{\sqrt{np(1-p)}}\right| < \varepsilon\right) \approx \frac{1}{2\pi}\int_{-\varepsilon}^{\varepsilon} e^{-\left(\frac{x^2}{2}\right)}dx.$$

По таблицам распределения по заданному γ находим ε и все сводим к решению неравенства

$$\left|\frac{m-np}{\sqrt{np(1-p)}}\right| < \varepsilon$$

или равносильного ему неравенства $\frac{(m-np)^2}{p(1-p)} < \varepsilon^2$. Разрешаем последнее неравенство относительно p. Получаем

$$\frac{m}{n} - \frac{\varepsilon^2 + \varepsilon\sqrt{4m + \varepsilon^2 - 4\frac{m^2}{n}} - \frac{2m\varepsilon^2}{n}}{2(n+\varepsilon^2)} < p < \frac{m}{n} + \frac{\varepsilon^2 + \varepsilon\sqrt{4m + \varepsilon^2 - 4\frac{m^2}{n}} - \frac{2m\varepsilon^2}{n}}{2(n+\varepsilon^2)}.$$

Если выражение $1-p$ заменить его максимальным значением $\frac{1}{4}$, то полученный интервал $\left(\frac{m}{n} - \frac{\varepsilon}{2\sqrt{n}}; \frac{m}{n} + \frac{\varepsilon}{2\sqrt{n}}\right)$ шире и поэтому имеет больший коэффициент надежности.

Теоретически получили интервалы для частоты попадания людей в зону. Чтобы построить интервалы, нужно полученные границы вероятности умножить на среднее количество людей в двух зонах.

Таким образом, построенные доверительные интервалы позволяют исключить ошибки фиксирования камерами видеонаблюдения аномального количества людей в определенной зоне. Удалось получить распределение параметра с определенной точностью, а значит, имеем хорошее представление об исследуемом объекте.

Литература

1. Крамер Г. Математические методы статистики. М. : Мир, 1975. С. 533.

2. Ивченко Г.И., Медведев Ю.И. Математическая статистика. М. : Высшая школа, 1984. С. 81–84, 89–92.

3. Исаева Н.А., Кривякова Э.Н. Оценивание параметров распределения : метод. пособие. Томск, 1990. С. 9–10.

РОСТ ПРОДОЛЖИТЕЛЬНОСТИ ЖИЗНИ НАСЕЛЕНИЯ КАК ОСНОВА ОБЩЕСТВЕННОГО РАЗВИТИЯ: ПРИОРИТЕТНЫЕ НАПРАВЛЕНИЯ ГОСУДАРСТВЕННОЙ ПОЛИТИКИ

Шибалков И.П., аспирант, НИ ТПУ, г. Томск
Научный руководитель: д-р экон. наук, доцент Недоспасова О.П.
E-mail: shibalkov.ivan@yandex.ru

Одним из важнейших показателей общественного здоровья и социально-экономического развития является средняя ожидаемая продолжи-