О близости биномиального распределения к нормальному для ограниченного числа наблюдений

А.С. Кондрик
Вычислительный центр ДВО РАН
email: sky_home@mail.ru

К.В. Михайлов
email: mikv.regs@gmail.com

С.В. Нагаев
Институт математики им. С. Л. Соболева СО РАН
email: nagaev@math.nsc.ru

В.И. Чеботарёв
Вычислительный центр ДВО РАН
email: cheb_8@mail.ru

15 апреля 2011 г.

Исследуется вопрос погрешности нормальной аппроксимации для биномиальных распределений с фиксированным числом испытаний n. На специально выбранной конечной сетке значений вероятности успеха p производятся компьютерные вычисления такой погрешности. При помощи аналитических методов оценивается ошибка, возникающая при замене погрешности аппроксимации в произвольной точке вычисленной погрешностью в ближайшей сеточной точке. Показывается, что величина этой ошибки зависит в том числе и от выбора шага сетки. Предлагается способ построения неравномерной сетки по p, имеющей меньшее число точек по сравнению с равномерной, но при этом обеспечивающей такую же точность вычисления погрешности. В результате выводится верхняя оценка константы в неравенстве Берри–Эссеена для двухточечных распределений при условии $n = 1..200$.

Пусть X, X_1, X_2, ... , X_n – независимые случайные величины с одним и тем же распределением:

$$\mathbf{P}(X = 1) = p, \quad \mathbf{P}(X = 0) = q = 1 - p.$$

Введем следующие обозначения: $F_{n,p}(x)$ – функция распределения случайной величины $S_n = \sum_{i=1}^{n} X_i$, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$, $G_{n,p}(x) = \Phi\big(\frac{x-np}{\sqrt{npq}}\big)$,

$$\Delta_n(p) = \sup_{x\in\mathbb{R}} \Big| F_{n,p}(x) - G_{n,p}(x) \Big|, \quad \varrho(p) = \frac{\mathbf{E}|X|^3}{(\mathbf{E}X^2)^{3/2}},$$

$$K_n(p) = \frac{\sqrt{n}}{\varrho(p)} \Delta_n(p), \quad \mathcal{K}_N = \sup_{n\geq N} \max_{p\in(0,0.5]} K_n(p).$$

Не теряя общности выводов, мы будем рассматривать только случай

$$0 < p \leq 0.5.$$

Задача получения наиболее точной верхней оценки для величины $\mathcal{K}_1$ связана с задачей нахождения абсолютной константы в неравенстве Берри – Эссеена. Историю вопроса и библиографические ссылки см., например, в [1] и [2].

В [1] найдена функция $E_0(p,n)$, удовлетворяющая следующим трем условиям:

a) $K_n(p) \le E_0(p,n), \quad n \ge 200,$

b) для каждого $p \in (0, 0.5]$ последовательность $E_0(p,n)$ убывает при $n \ge 200$,

c) $\lim\limits_{n\to\infty} E_0(p,n) = \mathcal{E}(p) \equiv \dfrac{2-p}{3\sqrt{2\pi}\,[p^2+(1-p)^2]}.$

Из работы Эссеена [3] следует, что, во-первых, для каждого $0 < p \le 0.5$ существует предел $\lim\limits_{n\to\infty} K_n(p)$, и, во-вторых, этот предел не превосходит

$$K \equiv \frac{3+\sqrt{10}}{6\sqrt{2\pi}} = 0.409732\ldots\ .$$

Кроме того, $\lim\limits_{n\to\infty} K_n(p_0) = K$, где $p_0 = \frac{4-\sqrt{10}}{2}$. Заметим, что $\max\limits_{p\in(0,0.5]} \mathcal{E}(p) = \mathcal{E}(p_0) = K$.

В [1] также показано, что $\mathcal{K}_{200} < 0.4215$. Для того, чтобы мы имели право считать верным неравенство $\mathcal{K}_1 < 0.4215$, остается убедиться, что

$$\max_{1\le n<200} \max_{p\in(0,0.5]} K_n(p) < 0.4215.$$

Мы получим более точное неравенство.

Теорема 1. *При* $1 \le n \le 200$ *справедлива оценка*

$$\max_{p\in(0,0.5]} K_n(p) < K. \tag{1}$$

Доказательство будет построено, во-первых, из численных расчетов величин $K_n(p)$, $n = 1, 2, \ldots, 200$, по специально выбранным сеткам значений $p \in (0, 0.5]$ и, во-вторых, из теоретической оценки погрешности, возникающей при замене промежуточных значений $K_n(p)$ ближайшими сеточными.

Очевидно, что для любой пары (n,p) величина $\sup\limits_{x\in\mathbb{R}} \left|F_{n,p}(x) - G_{n,p}(x)\right|$ достигается в одной из $n+1$ точек разрыва функции $F_{n,p}(x)$. Заметим, что мы рассматриваем функции распределения, непрерывные слева. Поэтому

$$\Delta_n(p) = \max_{k=0,1,\ldots,n} \Delta_{n,k}(p),$$

где

$$\Delta_{n,k}(p) = \max\Big\{|F_{n,p}(k) - G_{n,p}(k)|,\ |F_{n,p}(k+1) - G_{n,p}(k)|\Big\}.$$

Обозначим

$$\omega(p) = p^2 + q^2, \quad q_i = 1 - p_i, \quad \varrho_i = \varrho(p_i) \equiv \frac{\omega(p_i)}{\sqrt{p_i q_i}}.$$

Лемма 1. *Пусть* $h > 0$, $0 < p_1 < p < p_2 = p_1 + h \le 0.5$. *Тогда при любых* $n \ge 1$ *и* $0 \le k \le n$

$$\Big|\frac{1}{\varrho}\,\Delta_{n,k}(p) - \frac{1}{\varrho_i}\,\Delta_{n,k}(p_i)\Big| \le L(p_1,h)\,\min_{j=1,2}|p-p_j|, \quad i=1,2,$$

где

$$L(p_1,h) = \Big(\frac{1.332}{p_1^2 q_1} + \frac{0.121}{p_1 q_1}\Big)\,\frac{1}{\varrho_2} + \frac{1-2p_1}{2\varrho_2\omega(p_2)}\,\Big(\frac{\varrho_1^2}{\omega^2(p_1)} + 2\Big).$$

Лемма, приведённая выше, может быть получена из оценки погрешности в локальной предельной теореме для схемы Бернулли и ее применения для оценки производной $\frac{d}{dp}\,\Delta_{n,k}(p)$.

Теорема 2. *Пусть* $0 < p^* < 0.5$ *некоторое фиксированное число,* $p^* \le p \le 0.5$, p_i – *узел сетки на* $[p^*, 0.5]$ *с шагом* h, *ближайший к* p. *Тогда при любых* $n \ge 1$ *и* $0 \le k \le n$

$$\Big|\frac{\sqrt{n}}{\varrho}\,\Delta_{n,k}(p) - \frac{\sqrt{n}}{\varrho_i}\,\Delta_{n,k}(p_i)\Big| \le \frac{h}{2}\sqrt{n}\,L(p^*,h).$$

Утверждение теоремы 2 следует из леммы 1 и свойств функции $L(p,h)$: она убывает по первому аргументу и возрастает по второму.

Доказательство теоремы 1. В [2] получено неравенство из которого следует (1) при условии $p \le 0.109$. Поэтому для доказательство теоремы необходимо оценить $K_n(p)$ при $0.109 \le p \le 0.5$, $\quad n = \overline{1,200}$.

Рассмотрим равномерную сетку $\mathcal{Q}$ на $[0.109, 0.5]$ с шагом h, который будет выбран позднее. Зафиксируем произвольное $n \in [1,200]$ и $p \in [p^*, p^{**}]$, причем $0.109 \le p^* < p^{**} \le 0.5$. Обозначим $\mathcal{Q}' = \mathcal{Q} \cap [p^*, p^{**}]$.

Теорема 2 даёт неравенство

$$\frac{\sqrt{n}}{\varrho(p)}\,\Delta_{n,k}(p) \le \frac{\sqrt{n}}{\varrho(p_i)}\,\Delta_{n,k}(p_i) + \frac{h}{2}\sqrt{n}\,L(p^*,h),$$

где p_i – узел сетки, ближайший к p. Считая $h \le 10^{-4}$, мы сохраним это неравенство, если заменим $L(p^*,h)$ на $L(p^*) \equiv L(p^*, 10^{-4})$, так как по второму аргументу функция $L(p,h)$ возрастает.

В результате получаем, что

$$K_n(p) = \max_{0\le k\le n}\frac{\sqrt{n}}{\varrho(p)}\,\Delta_{n,k}(p) \le \max_{0\le k\le n}\max_{p_i\in\mathcal{Q}'}\frac{\sqrt{n}}{\varrho(p_i)}\,\Delta_{n,k}(p_i) + \frac{h}{2}\sqrt{n}\,L(p^*). \tag{2}$$

Таким образом для доказательства того, что $K_n(p) < K$ нам нужно выбрать шаг сетки h, исходя из неравенства

$$\frac{h}{2}\sqrt{n}\,L(p^*) < \varepsilon = 10^{-6}, \tag{3}$$

а также провести вычисления и убедиться в справедливости неравенства

$$\max_{0\le k\le n}\max_{p_i\in\mathcal{Q}'}\frac{\sqrt{n}}{\varrho(p_i)}\,\Delta_{n,k}(p_i) < K - \varepsilon.$$

Заметим, что в силу убывания $L(p)$ шаг h можно выбирать тем больше, чем правее находится левый край отрезка $[p^*, p^{**}]$. Разобьем отрезок $[0.109, 0.5]$, например, на такие подотрезки: $[0.109, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.4]$ и $(0.4, 0.5]$. На каждом из них определим h согласно (3). Получим сетку $\mathcal{Q}_0$ на $[0.109, 0.5]$ (с переменным шагом). Затем, воспользовавшись компьютером, проверим, что выполнено неравенство

$$\max_{1\le n\le 200}\max_{0\le k\le n}\max_{p_i\in\mathcal{Q}_0}\frac{\sqrt{n}}{\varrho(p_i)}\,\Delta_{n,k}(p_i) < 0.4096. \tag{4}$$

Теперь из (2)-(4) находим, что для каждого $1 \le n \le 200$

$$\max_{p\in[0.109,0.5]} K_n(p) \le \max_{0\le k\le n}\max_{p_i\in\mathcal{Q}_0}\frac{\sqrt{n}}{\varrho(p_i)}\,\Delta_{n,k}(p_i) \le \max_{1\le n\le 200}\max_{0\le k\le n}\max_{p_i\in\mathcal{Q}_0}\frac{\sqrt{n}}{\varrho(p_i)}\,\Delta_{n,k}(p_i) + \varepsilon < K.$$

□

На самом деле, стремясь минимизировать время вычислений, мы выбирали другие подотрезки, при помощи последовательного вызова рекурсивной функции Найти подотрезок. Нами учитывался тот факт, что из-за убывания функции $L(p)$ на правой половине любого подтрезка требуется менее густая сетка, чем на левой половине, при одинаковой точности расчётов.

В формализованном виде процедура поиска первого подходящего подотрезка выглядит следующим образом.

Найти подотрезок(p_0, p_1).

1 $h_{\min}$ принимает значение 10^{-12}
2 p_a принимает значение p_0, p_b берётся равным $(p_0 + p_1)/2$
3 h_a принимает значение, найденное из (3) для отрезка $[p_a, p_b]$
4 h_b принимает значение, найденное из (3) для $[p_b, p_1]$
5 **если** $h_a/h_b > 1.5$ и $h_a > h_{\min}$ **то**
вернуть Найти подотрезок(p_a, p_b)
иначе
вернуть $[p_0, p_1]$

Заметим, что для уменьшения числа узлов получаемой сетки каждый раз выполнялись вызовы указанной функции для отрезков $[p_0, p_1]$, $[p_0, (p_0+p_1)/2]$ и $[p_0, p_0+(p_1-p_0)/4]$, а в качестве первого подотрезка выбирался наименьший из полученных.

Это обусловлено тем, что возможна ситуация, когда на отрезке $[p_0, p_0 + (p_1 - p_0)/4]$ для достижения выбранной точности требуется шаг h_m, а на отрезке $[p_0+(p_1-p_0)/4, p_1]$ потребуется шаг $2h_m$, но аналогичные вычисления для $[p_0, (p_1 + p_0)/2]$ и $[(p_1 + p_0)/2, p_1]$ дадут, что отношение шагов не превысит 1.5.

Счёт на построенных таким образом сетках при n близких к 200 требовал примерно на 2 порядка меньше времени, чем счёт с использованием равномерной сетки.

Список литературы

[1] Нагаев С.В., Чеботарёв В.И. // Теория вероятн. и ее примен. 2011. (в печати).

[2] Королев В.Ю., Шевцова И.Г. О верхней оценке абсолютной постоянной в неравенстве Берри – Эссеена // Теория вероятн. и ее примен. 2009. Т. 53. Вып. 4. С. 671–695.

[3] Esseen C.-G. A moment inequality with an application to the central limit theorem // Scand. Aktuarietidskr. J. 1956. Vol. 3–4. P. 160–170.