О ПОГРЕШНОСТИ ГАУССОВСКОЙ АППРОКСИМАЦИИ ГИПЕРГЕОМЕТРИЧЕСКОГО РАСПРЕДЕЛЕНИЯ ПРИ ПРОВЕРКЕ ГИПОТЕЗЫ О РАВЕНСТВЕ БИНОМИАЛЬНЫХ ЧАСТОТ

В.В. Поддубный

Томский государственный университет

Известно [1], что при проверке гипотезы о равенстве параметров p_1 и p_2 двух независимых биномиальных распределений

$$P_i(x_i \mid n_i, p_i) = C_{n_i}^{x_i} p_i^{x_i} (1-p_i)^{n_i - x_i}, \quad x_i = \overline{0, n_i}, \; i = 1,2,$$

где $C_n^x = \dfrac{n!}{x!(n-x)!}$ – число сочетаний из n по x, приходится иметь дело со сложной нулевой гипотезой $H_0 : p_1 = p_2$ при неизвестных заранее объемах испытаний n_1 и n_2. Оптимальные оценки этих параметров выражаются эмпирическими частотами $\hat{p}_1 = m_1/n_1$, $\hat{p}_2 = m_2/n_2$, где m_1 и m_2 – наблюдаемые в эксперименте значения x_1 и x_2, а n_1 и n_2, как уже отмечалось, заранее неизвестны и реализуются в ходе эксперимента. Совместная вероятность встретить в эксперименте значения m_1 и m_2 при реализовавшихся значениях n_1 и n_2 и верной нулевой гипотезе H_0, когда $p_1 = p_2$ (обозначим это общее значение через p), выражается произведением биномиальных распределений

$$P(m_1, m_2 \mid n_1, n_2, p) = C_{n_1}^{m_1} C_{n_2}^{m_2} p^{m_1+m_2} (1-p)^{n_1+n_2-(m_1+m_2)},$$

так что величина $s = m_1 + m_2$ является достаточной статистикой для оптимальной оценки p: $\hat{p} = s/(n_1+n_2)$. Условная вероятность получить в эксперименте наблюдения m_1 и m_2 при фиксированном значении $s = m_1 + m_2$ и верной H_0 есть, на самом деле, условная вероятность получить значение m_1, так как $m_2 = s - m_1$. Эта вероятность выражается значением при $x = m_1$ свободного от параметра p гипергеометрического распределения [1, с. 236]

$$h(x \mid s, n_1, n_2) = C_{n_1}^{x} C_{n_2}^{s-x} \Big/ C_{n_1+n_2}^{s}, \quad x = \overline{\max(0, s-n_2), \min(n_1, s)}. \qquad (1)$$

Таким образом, величина m_1 является статистикой, распределение которой при верной нулевой гипотезе не зависит от неизвестного параметра p. Следовательно, при заданном уровне значимости α критерия проверки нулевой гипотезы $H_0 : p_1 = p_2$ против альтернативы $H_1 : p_1 \neq p_2$ решение в пользу альтернативы принимается при $m_1 \le h_{s,n_1,n_2,\alpha/2}$ или при $m_1 \ge h_{s,n_1,n_2,1-\alpha/2}$, где $h_{s,n_1,n_2,\alpha/2}$ и $h_{s,n_1,n_2,1-\alpha/2}$ – квантили соответственно уровней $\alpha/2$ и $1-\alpha/2$ гипергеометрического распределения (1). При $h_{s,n_1,n_2,\alpha/2} < m_1 < h_{s,n_1,n_2,1-\alpha/2}$ оснований отвергнуть нулевую гипотезу нет. Реализовавшемуся в эксперименте значению m_1 соответствует достигнутый уровень значимости

$$p_0 = \sum_{x=\max(0,s-n_2)}^{\min(n_1,s)} \{h(x \mid s, n_1, n_2) \le h(m_1 \mid s, n_1, n_2)\}.$$

Решение в пользу альтернативы принимается при $p_0 \le \alpha$. При $p_0 > \alpha$ оснований отвергнуть нулевую гипотезу нет.

Рассмотренный выше гипергеометрический критерий проверки гипотезы $H_0 : p_1 = p_2$ о равенстве биномиальных частот теоретически работает при любых значениях объемов испытаний n_1 и n_2. На практике, однако, его применимость ограничивается возможностями вычислений факториалов больших чисел, так как уже 170! = 7.2574e+306, т.е. является числом, близким к машинной бесконечности. В связи с этим возникает вопрос об аппроксимации при больших n_1 и n_2 гипергеометрического распределения другим распределением, в каком-то смысле близким к гипергеометрическому, но допускающим вычисления достигнутого уровня значимости при любых как угодно больших значениях n_1 и n_2. Вычисления факториалов больших чисел по формуле Стирлинга не спасают положения, так как разность между $n!$ и «приближением» Стирлинга $S(n) = n^n e^{-n}\sqrt{2\pi n}$ неограниченно возрастает с ростом n, $\lim\limits_{n\to\infty}\left(n! - S(n)\right) = \infty$, хотя их отношение и стремится к единице, $\lim\limits_{n\to\infty}\left(n!/S(n)\right) = 1$ [2, с. 271].

Исследуем теперь вопрос о возможности и точности аппроксимации гипергеометрического распределения (1) нормальным распределением с вероятностями целочисленных значений x

$$P_N(x) = \left(1/\sqrt{2\pi\sigma^2}\right)\exp\left(-(x-\mu)^2/\left(2\sigma^2\right)\right), \qquad (2)$$

где $\mu = s(n_1/n)$, $\sigma^2 = s(n_1/n)(n_2/n)(n-s)/(n-1)$ – математическое ожидание и дисперсия гипергеометрического распределения (1) [3, с. 190], $n = n_1 + n_2$. Результаты сравнительного анализа максимальных значений абсолютного ε и относительного δ расхождений распределений (1) и (2) при $n_1 = \overline{N_1, N_2}$, $n_2 = \overline{N_1, N_2 - n_1 + N_1}$, $N_1 = 2$ и различных N_2 (так что $\max(n_1 + n_2) = 2N_2$) приведены в табл. 1.

Таблица 1

N_2	20	40	60	80	100	120	140	160
ε	0,0955	0,3682	0,6202	0,8426	1,0429	1,2262	1,3962	1,5552
δ	5,457	$1{,}627\cdot10^{2}$	$5{,}309\cdot10^{3}$	$1{,}917\cdot10^{5}$	$7{,}350\cdot10^{6}$	$2{,}929\cdot10^{8}$	$1{,}199\cdot10^{10}$	$5{,}006\cdot10^{11}$

Видно, что относительное расхождение форм распределений (1) и (2) очень быстро растет с ростом максимального значения n. Однако с точки зрения вычисления достигнутого уровня значимости различие этих распределений пренебрежимо мало, и приближение (2) подходит при любом $n > 160$. В табл. 2 приведены значения частоты ν расхождений в принятии решений (доли несовпадающих решений) по распределениям (1) и (2) в тех же условиях.

Таблица 2

N_2	20	40	60	80	100	120	140	160	170
ν	0,01326	0,01014	0,00787	0,00694	0,00598	0,00518	0,00453	0,00405	0,00385

Видно, что с ростом n величина ν убывает, и при $n > 170$ $\nu < 0{,}4\% << \alpha = 5\%$.

Литература

1. Справочник по прикладной статистике /Под ред. Э. Ллойда, У. Ледермана. М.: Финансы и статистика, 1989. Т. 1. 512с.
2. Нейман Ю. Вводный курс теории вероятностей и математической статистики. М.: Наука (Гл. ред. физ.-мат. лит.), 1968. 448 с.
3. Кендэлл М. Дж., Стьюарт А. Теория распределений. М.: Наука (Гл. ред. физ.-мат. лит.), 1968. 588 с.