Доклады Академии Наук СССР
1939. Том XXII, № 6

МАТЕМАТИКА

Р. О. КУЗЬМИН

О ЗАКОНЕ РАСПРЕДЕЛЕНИЯ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ В ВЫБОРКАХ ИЗ НОРМАЛЬНОЙ СОВОКУПНОСТИ

(Представлено академиком С. Н. Бернштейном 6 I 1939)

§ 1. Две случайные величины x и y находятся в нормальной корреляции, если вероятность p_s того, что точка (x, y) принадлежит площади s, выражается формулой:

$$p_s = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \iint\limits_s e^{-\frac{1}{2(1-r^2)}\left[\frac{(x-x_0)^2}{\sigma_x^2} - \frac{2r(x-x_0)(y-y_0)}{\sigma_x\sigma_y} + \frac{(y-y_0)^2}{\sigma_y^2}\right]} dxdy.$$

При этом параметры $x_0, y_0, \sigma_x, \sigma_y$ и r имеют следующие выражения через математические ожидания соответствующих переменных:

$$x_0 = Ex,\ y_0 = Ey;\ \sigma_x^2 = E(x-x_0)^2;\ \sigma_y^2 = E(y-y_0)^2;\ r\sigma_x\sigma_y = E(x-x_0)(y-y_0).$$

Иными словами, x_0 и y_0 — средние значения переменных x и y, σ_x и σ_y — их стандарты, r — коэффициент корреляции. При этом всегда $\sigma_x > 0$, $\sigma_y > 0$, $-1 < r < 1$. Обозначения взяты из книги В. Романовского (1).

Значительную важность представляет вопрос о распределении эмпирического коэффициента корреляции $\bar{r}$ в выборках из данной нормальной совокупности. Дифференциальный закон распределения выборок объема n обозначается в дальнейшем через $D_n(\bar{r})$. Он был изучен Р. Фишером (2) в 1915 г. и может быть представлен формулой:

$$D_n(\bar{r}) = \frac{(1-r^2)^{\frac{n-1}{2}}(1-\bar{r}^2)^{\frac{n-4}{2}}}{\pi(n-3)!} \frac{d^{n-2}}{d\tau^{n-2}} \frac{\arccos(-\tau)}{\sqrt{1-\tau^2}}, \tag{1}$$

в которой после дифференцирования следует положить $\tau = r\bar{r}$.

К. Пирсон и его сотрудники (3) в обширном мемуаре исследовали свойства функции $D_n(\bar{r})$. Однако их метод приводил к сложным вычислениям. Кроме того им не удалось полностью доказать одно из важных свойств $D_n(\bar{r})$, в силу которого при фиксированном r и $n \to \infty$ распределение $D_n(\bar{r})$ стремится к нормальному распределению со средней r и с дисперсией $\frac{1-r^2}{\sqrt{n-1}}$. По словам В. Романовского до сих пор нет полного и строгого доказательства этой теоремы.

Настоящая работа имеет целью восполнить этот пробел. Предлагаемая мной формула для $D_n(\bar{r})$ дает простой путь для удобного изучения этой величины.

§ 2. Непосредственным интегрированием без труда проверяется равенство:

$$\frac{\arccos(-\tau)}{\sqrt{1-\tau^2}} = \int_1^\infty \frac{dx}{(x-\tau)\sqrt{x^2-1}}. \tag{2}$$

Отсюда легко следует, что

$$\frac{d^{n-2}}{d\tau^{n-2}} \frac{\arccos(-\tau)}{\sqrt{1-\tau^2}} = (n-2)! \int_1^\infty \frac{dx}{(x-\tau)^{n-1}\sqrt{x^2-1}}. \tag{3}$$

Подставляя в (1), имеем:

$$D_n(\bar{r}) = \frac{n-2}{\pi} \cdot (1-r^2)^{\frac{n-1}{2}} (1-\bar{r}^2)^{\frac{n-4}{2}} \int_1^\infty \frac{dx}{(x-\tau)^{n-1}\sqrt{x^2-1}}. \tag{4}$$

Обозначая последний интеграл через J и полагая $x = \frac{1+\sigma\tau}{1-\sigma}$, находим:

$$J = \frac{1}{(1-\tau)^{n-\frac{3}{2}}\sqrt{2}} \int_0^1 \frac{(1-\sigma)^{n-2}\sigma^{-\frac{1}{2}}d\sigma}{\sqrt{1-\frac{1-\tau}{2}\sigma}} =$$

$$= \frac{1}{(1-\tau)^{n-\frac{3}{2}}\sqrt{2}} \left[\int_0^1 (1-\sigma)^{n-2}\sigma^{-\frac{1}{2}}d\sigma + 0(1)\int_0^1 (1-\sigma)^{n-2}\sigma^{\frac{1}{2}}d\sigma\right]. \tag{5}$$

Полученные интегралы представляют бэта-функции. Выражая их через гамма-функцию и применяя формулу Стирлинга, имеем:

$$J = \frac{1}{(1-\tau)^{n-\frac{3}{2}}} \left[\sqrt{\frac{\pi}{2n}} + 0\left(\frac{1}{n\sqrt{n}}\right)\right]. \tag{6}$$

Из (4), (5) и (6) после простых вычислений получаем:

$$D_n(\bar{r}) = \sqrt{\frac{n-1}{2\pi}} \left[\frac{(1-r^2)(1-\bar{r}^2)}{(1-r\bar{r})^2}\right]^{\frac{n-4}{2}} \frac{(1-r^2)^{\frac{3}{2}}}{(1-r\bar{r})^{\frac{5}{2}}} \left[1+0\left(\frac{1}{n}\right)\right]. \tag{7}$$

§ 3. Из равенства (7) легко получить асимптотическую формулу для $D_n(\bar{r})$. Для этого вводим новую переменную, полагая

$$\bar{r} = r + \frac{u}{\sqrt{n-1}}.$$

После этого легко получаются формулы:

$$\frac{(1-r^2)^{\frac{3}{2}}}{(1-r\bar{r})^{\frac{5}{2}}} = \frac{1}{(1-r^2)}\left[1+0\left(\frac{u}{\sqrt{n}}\right)\right],$$

$$\frac{n-4}{2}\log\frac{(1-r^2)(1-\bar{r}^2)}{(1-r\bar{r})^2} = -\frac{u^2}{2(1-r^2)^2} + 0\left(\frac{u^2+u^3}{\sqrt{n}}\right). \tag{8}$$

Будем считать временно, что $|u| < n^{\frac{1}{12}}$. В таком случае из (8) получается:

$$\left[\frac{(1-r^2)(1-\bar{r}^2)}{(1-r\bar{r})^2}\right]^{\frac{n-4}{2}} \frac{(1-r^2)^{\frac{3}{2}}}{(1-r\bar{r})^{\frac{5}{2}}} = \frac{1}{1-r^2} e^{-\frac{u^2}{2(1-r^2)^2}} \left[1 + 0\left(\frac{1}{n^{\frac{1}{4}}}\right)\right]. \tag{9}$$

После этого (7) можно переписать в таком виде:

$$D_n(\bar{r}) = \sqrt{\frac{n-1}{2\pi}}\, \frac{e^{-\frac{u^2}{2(1-r^2)^2}}}{1-r^2} \left[1 + 0\left(\frac{1}{n^{\frac{1}{4}}}\right)\right]. \tag{10}$$

Отсюда при $|u_1| < n^{\frac{1}{12}}$ и $|u_2| < n^{\frac{1}{12}}$ получается:

$$\int_{\bar{r}_1}^{\bar{r}_2} D_n(\bar{r})\, d\bar{r} = \frac{1}{\sqrt{2\pi(1-r^2)}} \int_{u_1}^{u_2} e^{-\frac{u^2}{2(1-r^2)^2}} du + 0\left(\frac{1}{n^{\frac{1}{4}}}\right). \tag{11}$$

Значение интеграла в правой части, распространенного на интервалы, в которых $|u| > n^{\frac{1}{12}}$, представляет величину меньшую, чем $n^{-\frac{1}{4}}$. Поэтому ограничение, в силу которого $|u_1| < n^{\frac{1}{12}}$, $|u_2| < n^{\frac{1}{12}}$, можно отбросить. Формула (11) верна при любых u_1 и u_2. Ее можно переписать в таком виде:

$$\int_{\bar{r}_1}^{\bar{r}_2} D_n(\bar{r})\, d\bar{r} = \frac{\sqrt{n-1}}{\sqrt{2\pi}(1-r^2)} \int_{\bar{r}_1}^{\bar{r}_2} e^{-\frac{(n-1)(\bar{r}-r)^2}{2(1-r^2)^2}} d\bar{r} + 0\left(\frac{1}{n^{\frac{1}{4}}}\right). \tag{12}$$

Таким образом теорема, отмеченная раньше, доказана.

§ 4. Для получения формулы, удобной для изучения $D_n(\bar{r})$, в интеграле равенства (4) полезно ввести новую переменную, полагая

$$x = \tau + (1-\tau)e^{\sigma}.$$

После этого получается равенство:

$$\int_1^{\infty} \frac{dx}{(x-\tau)^{n-1}\sqrt{x^2-1}} = \frac{1}{(1-\tau)^{n-1}} \int_0^{\infty} e^{-(n-2)\sigma} \cdot \frac{1}{\sqrt{(e^{\sigma}-1)\left(e^{\sigma} + \frac{1+\tau}{1-\tau}\right)}}\, d\sigma. \tag{13}$$

Для второго множителя под интегралом в правой части нетрудно установить формулу:

$$\frac{1}{\sqrt{(e^{\sigma}-1)\left(e^{\sigma} + \frac{1+\tau}{1-\tau}\right)}} = \sigma^{-\frac{1}{2}}[A_0 + A_1\sigma + \ldots + A_{m-1}\sigma^{n-1} + 0(1)\sigma^m]. \tag{14}$$

Здесь $A_0, A_1, \ldots, A_{m-1}$ — полиномы относительно $\frac{1+\tau}{1-\tau}$ с рациональными коэффициентами. Значения последних нетрудно найти для любого числа их. Нетрудно также дать верхний предел для $0(1)$ — он зависит от τ. Подставляя (14) в (13) и интегрируя, находим:

$$\int_1^\infty \frac{dx}{(x-\tau)^{n-1}\sqrt{x^2-1}} =$$

$$= \frac{\sqrt{\pi}}{(1-\tau)^{n-1}\sqrt{n-2}}\left[A_0 + \frac{1}{2}\,\frac{A_1}{n-2} + \ldots + \frac{1.3\ldots(2m-3)}{2^{m-1}}\,\frac{A_{m-1}}{(n-2)^{m-1}} + R\right]. \quad (15)$$

Здесь

$$R = 0(1)\cdot\frac{1.3\ldots(2m-1)}{2^m}\cdot\frac{1}{(n-2)^m}.$$

Подставляя полученное в (4), найдем формулу, дающую удобное средство для изучения $D_n(\bar{r})$ при больших значениях n, если только и величина $(n-2)(1-r\bar{r})$ тоже велика.

В другом месте, где будет дано более полное изложение, будут приведены приложения этой формулы.

Поступило
6 I 1939.

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] В. Романовский, Математическая статистика (1938). [2] R. A. Fischer, Biometrica, **10**, 507—521 (1915). [3] K. Person, Biometrica, **11** (1915).