УДК 311.16

А.В. САНЮКЕВИЧ

О ПРЕПОДАВАНИИ КОРРЕЛЯЦИОННОГО АНАЛИЗА НА ГЕОГРАФИЧЕСКОМ ФАКУЛЬТЕТЕ

В настоящее время при изучении корреляционной связи внимание в основном концентрируется на линейной корреляции двух переменных, в первую очередь на технике вычисления выборочного коэффициента корреляции и определении линейных уравнений регрессии. При этом, чтобы не нарушить одно из основных требований выборочного метода исследования, необходимо оценить точность или значимость этих выборочных оценок.

При большом объёме выборки n из нормально распределенных случайных величин X и Y, среднее квадратическое отклонение выборочного коэффициента корреляции приближенно определяется по формуле

$$\sigma_r = \frac{1-r^2}{\sqrt{n}}. \qquad (1)$$

Полагая, что выборочный коэффициент корреляции при значениях, не близких к единице, и большом объеме выборки приближенно следует нормальному закону, можно построить приближенный доверительный интервал для r генерального:

$$r_n - t_p\sigma_r \le r \le r_n + t_p\sigma_r. \qquad (2)$$

Однако формула (1) основана на r генеральном, которое обычно неизвестно. Замена его выборочной оценкой может быть совсем неудовлетворительной, особенно при малом объеме выборки. Распределение выборочного коэффициента корреляции, даже при условии нормального распределения, асимметрично.

Поэтому в ряде учебников для оценки точности коэффициента корреляции предлагают использовать z-преобразование Фишера:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r},$$

которое имеет приблизительно нормальное распределение. Средняя ошибка z не зависит от r:

$$\sigma_z = \frac{1}{\sqrt{n-3}}.$$

Построив доверительный интервал для z можно найти доверительные границы для r, используя равенство $r = (e^{2x}-1)/(e^{2x}+1)$. Для упрощения перехода от r к z и обратно существуют специальные таблицы.

При малых значениях выборочного коэффициента корреляции можно оценить его значимость проверив гипотезу о некоррелированности величин *X* и *Y*. Для этого составляется произведение $|r_n|\sqrt{n-1}=\alpha$, которое сравнивается с табличным распределением критических значений этого произведения при заданных надежности *p* и объеме выборки *n*.

Все эти вычисления с давних пор проводятся студентами вручную с помощью калькуляторов. Однако в последние десятилетия бурное развитие вычислительной техники и программного обеспечения для неё привело к появлению подробных программ статистического и корреляционного анализа. Теперь при изучении этого раздела математической статистики студентам, основной специальностью которых не является математика, потребуется не столько умение вычислять, сколько умение анализировать полученные результаты и практически их использовать.

Ставя себе задачу подготовить к этому студентов, мы должны научить их самым общим методам корреляционного анализа, использованию универсальных критериев оценки значимости полученных результатов.

Очевидно, первой задачей корреляционного анализа следует считать задачу определения интенсивности связи и лишь при наличии значительной связи имеет смысл говорить о ее форме.

Выборочный коэффициент корреляции является весьма условным показателем даже линейной связи. Более естественным показателем степени тесноты связи является корреляционное отношение, так как использует непосредственно эмпирические данные, и более общим, так как не связано с формой зависимости.

Данные наблюдений в географических исследованиях чаще всего представлены корреляционной таблицей. В этом случае общая дисперсия каждой из двух случайных величин может быть выражена в виде суммы межгрупповой и внутригрупповой дисперсий:

$$\sigma_y^2=\sigma_{\overline{y_x}}^2+\sigma_{y_x}^2.$$

Так как внутри каждой группы признак *X* фиксирован, то внутригрупповая дисперсия не может быть обусловлена его изменчивостью. В зависимости от *X* изменяется только межгрупповая дисперсия. Корреляционное отношение и выделяет ту часть изменчивости *Y*, которая обусловлена изменчивостью *X*: $\eta_{y/x}=\dfrac{\sigma_{\overline{y_x}}}{\sigma_y}$.

Аналогично, $\eta_{x/y}=\dfrac{\sigma_{\overline{x_y}}}{\sigma_x}$.

Значимость корреляционного отношения проверяется критерием F (отношение межгрупповой дисперсии к остаточной внутригрупповой) с учетом числа степеней свободы:

$$F = \frac{\sigma^2_{\overline{y_x}}}{\sigma^2_{y_x}} \cdot \frac{n-k}{k-1} = \frac{\sigma^2_{\overline{y_x}}}{\sigma^2_y - \sigma^2_{\overline{y_x}}} \cdot \frac{n-k}{k-1} = \frac{\eta^2}{1-\eta^2} \cdot \frac{n-k}{k-1}.$$

Основная гипотеза состоит в утверждении, что межгрупповая дисперсия, обусловленная изменчивостью x, значительно превосходит остаточную, внутригрупповую, т.е. $F \gg 1$. Полученное значение F сравнивается с критическим значением F_p по таблице Фишера-Снедекора. Если $F > F_p\left(\frac{n-k}{k-1}\right)$, то корреляция существенна. Если $F < F_p\left(\frac{n-k}{k-1}\right)$, то принимается конкурирующая гипотеза – корреляционное отношение незначимо, нуль-гипотеза о некоррелированности случайных величин не отвергается.

Если между случайными величинами установлена значительная корреляция, то для установления характера зависимости надо выяснить форму связи, которая в общем случае нелинейна. Линейную зависимость можно рассматривать как частный случай нелинейной. Если значения r и η близки, то можно принять гипотезу о линейной корреляцией.

Форма связи может быть изначально неизвестна. Тогда на компьютере с помощью статистических программ можно подобрать множество различных уравнений, каждое из которых в своем семействе будет наилучшим образом, например по методу наименьших квадратов, описывать эмпирические данные. Для того, чтобы выделить лучшее из этих уравнений, можно использован критерий F, сопоставляющий общую дисперсию с остаточной для данного уравнения.

Студентам на конкретных примерах нужно показать практическое значение полученного уравнения регрессии для прогноза средних значений одной переменной по заданным значениям другой.

Общий метод оценки интенсивности связи с помощью корреляционного отношения с оценкой существенности полученного результата, понятие о криволинейной связи с оценкой практической ценности полученных уравнений и отборе из них наиболее соответствующего эмпирическим данным, изучение линейной корреляции как частного случая криволинейной и, наконец, понятие о множественной корреляции, имеющей большое практическое значение, – такое построение раздела «Корреляционный анализ» улучшит его изучение студентами и облегчит им переход к практическим вычислениям на компьютерах.