3.5. ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ ДЛЯ ВЫБОРОЧНЫХ ЗНАЧЕНИЙ КОЭФФИЦИЕНТА ПАРНОЙ КОРРЕЛЯЦИИ

В первом параграфе этой главы были приведены формулы для вычисления выборочных значений коэффициента парной корреляции r. Поскольку рассчитывается значение этого коэффициента по некоторой выборке, то с изменением выборки следует ожидать изменения и значений коэффициента парной корреляции. Понятно, что если рассчитать его значения для генеральной совокупности, то будет получено истинное значение коэффициента ρ, но на практике с генеральной совокупностью никто не работает, а работают только с выборочными совокупностями, следовательно, каждый раз исследователь будет получать некоторую оценку r истинного значения коэффициента парной корреляции ρ. И для исследователя очень важно знать насколько близко полученное расчетное значение коэффициента парной корреляции к этому истинному значению. Например, если коэффициент парной корреляции для генеральной совокупности будет равен *+0,2*, то это означает, что если между случайными переменными и существует зависимость, то она имеет сложный нелинейный характер. Но вполне вероятно, что по выборочным значениям этой генеральной совокупности двух случайных переменных была получена оценка коэффициента парной корреляции, равная *+0,8*. Тогда исследователь будет утверждать, что две случайные переменные имеют тесную линейную корреляционную связь. На основании этого утверждения он построит линейную регрессионную модель, найдёт коэффициенты этой модели и выполнит прогноз, который окажется совсем не точным. Поэтому, получив выборочное значение коэффициента парной корреляции, необходимо оценить доверительные границы этого коэффициента.

Хотелось бы воспользоваться для этого выводами предыдущего параграфа, например, формулой (3.4.25). Для этого надо знать не только оценку r коэффициента парной корреляции, но и его дисперсию. Как это сделать? Выбирать случайным образом некоторую совокупность переменных и считать для них одно значение коэффициента парной корреляции r_1, затем случайным образом выбирать следующее множество пар случайных значений переменных, вновь считать коэффициент парной корреляции r_2, вновь случайным образом выбрать из генеральной совокупности пары выборочных значений и считать и r_3 т.п. до некоторого r_m. Затем считать среднюю оценку коэффициента парной корреляции для этого множества и вычислять дисперсию. Трудоёмкость этих расчётов в m раз превышает задачу по расчёту коэффициента парной корреляции и вряд ли может быть признана целесообразной. Но если величина выборки не меняется, и каждый раз в неё попадает случайным образом n пар значений случайных величин из генеральной совокупности, то, как было выяснено в математической статистике, дисперсия коэффициента парной корреляции σ^2_r будет зависеть только от двух показателей – самого значения коэффициента парной корреляции ρ и числа членов пар случайных величин в выборке n:

$$\sigma_r \approx \frac{1-\rho^2}{\sqrt{n}}. \qquad (3.5.1)$$

Но поскольку значение ρ коэффициента парной корреляции генеральной совокупности исследователю не известно, его заменяют на выборочное значение коэффициента парной корреляции r и оценивают выборочное значение среднеквадратичного отклонения s_r:

$$s_r \approx \frac{1-r^2}{\sqrt{n}}. \qquad (3.5.2)$$

Тогда, зная дисперсию коэффициента парной корреляции, при заданной доверительной вероятности α легко найти доверительные границы для значения этого коэффициента:

$$r - t_\alpha s_r < \rho < r + t_\alpha s_r. \qquad (3.5.3)$$

К сожалению, простата этого подхода омрачается тем обстоятельством, что все эти выкладки применимы к большому числу наблюдений $n>50$, да, как утверждают специалисты в области математической статистики, для коэффициента парной корреляции, близкого к нулю, то есть, для случая, когда линейная корреляция не диагностируется.

Дело в том, что именно для большого числа наблюдений n и для $\rho\approx 0$ распределение коэффициента парной корреляции приближается к нормальному. А если исследователь действительно имеет дело со случайной зависимостью, приближающейся к линейной, или, как чаще всего встречаются в прогнозировании социально-экономических показателей, с малой выборкой $n<50$? Тогда указанный подход не приемлем, а доверительные границы (3.5.3), как следует из выводов предыдущего параграфа, рассчитываются, исходя из априорного предположения о том, что процесс нормально распределённый.

В таком случае нужно что-то делать с коэффициентом корреляции для того, чтобы его распределение приближалось к нормальному. Оказалось, что если выполнить z-преобразование Фишера:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}, \qquad (3.5.4)$$

то полученная переменная z распределена нормально.

При этом дисперсия этой преобразованной величины, а значит, и среднеквадратическое отклонение, находится очень просто:

$$\sigma_z = \frac{1}{\sqrt{n-3}}. \qquad (3.5.5)$$

Теперь найти доверительные границы на основании выборочного значения парной корреляции для некоторого числа пар наблюдений n и заданной исследователем доверительной вероятности α довольно просто. По формуле (3.5.4) вычисляется значение величины z, затем, зная числа пар наблюдений n, вычисляется среднеквадратическое (3.5.5), и полученные значения являются основанием для вычисления доверительных границ :

$$z - t_\alpha \sigma_z < Z < z + t_\alpha \sigma_z. \qquad (3.5.3)$$

Зная эти нижнюю и верхнюю границы доверительного интервала для z, легко перейти с помощью (3.5.4) к нижнему и верхнему значениям пределов доверительных границ для коэффициента парной корреляции.

Впрочем, стоит только один раз посчитать значения *z* при разных значениях *r*, как эти подсчёты можно использовать многократно. Что и было сделано довольно давно, поскольку практически в каждом учебнике по математической статистике приводятся таблицы значений *z*-преобразований Фишера, в которых эти значения формируются так: цифра значения коэффициента парной корреляции, наблюдаемая до первого знака после запятой, начиная от 0,0 до 0,9, располагается в начале каждой отдельной строки таблицы.

Столбцы этой таблицы содержат значение коэффициента следующего знака после запятой (второй) - от 0 до 9. Всего получается десять строк и десять столбцов. На пересечении строчки и столбца находится значение *z*, соответствующее данному коэффициенту парной корреляции.

Например, если выборочное значение коэффициента парной корреляции оказалось равным 0,87, то значение *z* находится так. Первая цифра после запятой равна 8. Поэтому ищется строка с цифрой 0,8. Вторая цифра коэффициента корреляции после запятой равна 7. Поэтому ищется столбец, соответствующий этой цифре. Поскольку столбцы начинаются с 0, то по счёту это будет восьмой столбец. На пересечении строчки и столбца находится цифра 1,333. Это и есть искомая величина *z*.