Лекция № 17

Корреляционный и регрессионный анализ. Продолжение

Проверка значимость коэффициента корреляции

До настоящего момента мы рассматривали линейный коэффициент корреляции, относящийся к выборке. Аналогично определяется линейный коэффициент корреляции для генеральной совокупности в целом:

$$\rho = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{x_i - m_x}{\sigma_x}\right)\left(\frac{y_i - m_y}{\sigma_y}\right)$$

где m_x, m_y, σ_x, σ_y - средние арифметические и средние квадратические отклонения признаков x и y генеральной совокупности. Схема интерпретации результатов выборочного наблюдения в корреляционном анализе такова. Пусть генеральная совокупность велика и ее исследование в полном объеме невозможно или нецелесообразно. Сформируем случайную выборку и по выборочным значениям определим линейный коэффициент корреляции. Предположим, что его значение оказалось вблизи +1 или - 1, указывая на наличие сильной связи между двумя признаками. При попытке распространить этот результат на генеральную совокупность возникает вопрос - является ли выявленная зависимость верной и для нее, или полученный результат является случайным и характерен только для отдельной выборки? Решение вопроса проводится с помощью аппарата проверки гипотез.

Проверяемые утверждения относятся к коэффициенту корреляции генеральной совокупности ρ:

$H_0 : \rho = 0;\ H_1 : \rho \neq 0$

Основная гипотеза предполагает, что генеральный коэффициент корреляции равен 0. Сформировав выборку и рассчитав ее коэффициент корреляции r, необходимо решить - является ли его значение достаточно большим, настолько, чтобы вероятность (по различным выборкам) выпадения такого значения при нулевом генеральном коэффициенте корреляции ρ была бы мала (меньше уровня значимости). Если является, то в этом случае основная гипотеза отвергается, а коэффициент корреляции и установленная зависимость между величинами полагаются значимыми.

Принятие решения будет основано на специальном критерии, рассчитываемом по коэффициенту корреляции r и длине выборки n:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

При условии, что генеральная совокупность приближенно подчиняется нормальному распределению, критерий подчиняется распределению Стьюдента с числом степеней свободы $k = n - 2$.

Пример. Исследовать значимость коэффициента корреляции в примере 1а при уровне значимости 0.05.

Решение. Расчет коэффициента корреляции для выборки из 6 человек дал значение $r = 0.897$, которое указывает на сильную прямую зависимость между возрастом и давлением.

1) Сформулируем проверяемые утверждения

$H_0 : \rho = 0$ (в генеральной совокупности нет зависимости, зависимость случайная);

$H_1 : \rho \neq 0$ (установленная зависимость справедлива для генеральной

совокупности).

2) Находим критические значения по таблице распределения Стьюдента. Для двусторонней гипотезы, уровне значимости $\alpha = 0.05$ и степени свободы $k = n - 2 = 6 - 2 = 4$ получаем $t_{\alpha/2} = 2.776$.

Критическая область задается условием $|t| > t_{\alpha/2}$, рис. 6.

3) Определяем расчетное значение критерия:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.897\sqrt{\frac{6-2}{1-0.897^2}} = 4.059.$$

4) Принятие решения. Значение критерия попадает в критическую область $t > t_{\alpha/2}$, основная гипотеза отклоняется.

5) **Вывод.** Прямая зависимость между возрастом человека и давлением является значимой, ее можно распространить на всю совокупность пациентов.

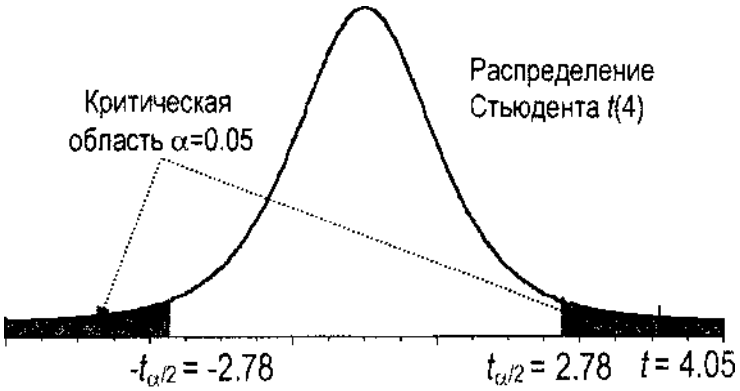

Рис. 6. *Исследование значимости коэффициента корреляции*

Интерпретация значимой зависимости между величинами

Предположим, что в исследовании установлена значимая зависимость между признаками. В различных ситуациях интерпретация этого факта может быть различной.

- Существует непосредственная причинно-следственная связь между признаками - x вызывает y. Так, например, вода вызывает рост растений, яд приводит к смерти, тепло вызывает плавление льда.
- Существует обратная причинно-следственная связь. Пусть, например, исследователь установил наличие сильной зависимости между количеством выпитых чашек кофе и нервозностью. Однако нельзя исключать, что причинным фактором является нервозность, а кофе употребляется людьми в нервозном состоянии для того, чтобы отвлечься и успокоиться.
- Зависимость между двумя признаками может быть вызвана влиянием какой-то третьей величины. Например, можно найти корреляцию между объемом продаж прохладительных напитков и числом утонувших, однако очевидно, что одновременный рост и того, и другого вызван жаркой погодой.
- Существует сложный набор зависимостей между многими факторами. Например, можно формально установить сильную зависимость между итоговой успеваемостью в школе и институте, однако вряд ли можно говорить здесь о зависимости причинно-следственного типа, так как и на ту, и на другую величину влияют многие факторы - способности человека, уровень его мотивированности, объем занятий, влияние родителей, качество преподавания и другие обстоятельства.
- Зависимость является простым совпадением. Например, можно найти прямую зависимость между количеством людей, занимающихся физической подготовкой (в широком смысле), и количеством людей, совершающих преступление. Однако здравый смысл подсказывает, что зависимость между этими факторами является совпадением.

Коэффициент корреляции рангов Спирмана

Коэффициент линейной корреляции Пирсона является наиболее распространенной мерой корреляции. Однако исследовать коэффициент Пирсона на значимость можно только при условии, что исходная генеральная совокупность приближенно подчиняется нормальному распределению. Если это условие не выполняется, то вместо нее применяют другую величину - коэффициент корреляции рангов Спирмана, значимость которого можно исследовать для любой совокупности.

Для определения коэффициента значения обеих переменных нужно упорядочить (по отдельности) в порядке убывания, присвоить им ранги (порядковые номера) N_{xi}, N_{yi} и занести в исходную таблицу. Далее, для каждой единицы совокупности вычисляется разность рангов $d_i = N_{xi} - N_{yi}$. Коэффициент корреляции рангов Спирмана вычисляется по формуле:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

Определенный таким образом коэффициент принимает значения от -1 до +1 и его интерпретация в целом совпадает с интерпретацией коэффициента Пирсона, рис. 2.

Если ранги двух признаков полностью совпадают, то сумма квадратов отклонений равна 0 и коэффициент $r_s = 1$, указывая на сильную прямую связь.

Если, с другой стороны, ранги прямо противоположны: 1 для x соответствует n для y, 2 для x и для y и т.д., то сумма квадратов примет максимальное значение $\sum_{i=1}^{n} d_i^2 = \frac{1}{3} n\left(n^2 - 1\right)$, а коэффициент корреляции -

минимальное значение -1. Это указывает на максимально сильную обратную зависимость между x и y. Следует иметь ввиду, что, поскольку коэффициент Спирмана учитывает только разность рангов, а не сами значения, то он менее точен по сравнению с коэффициентом Пирсона. Поэтому его крайние значения (±1 или 0) нельзя безоговорочно оценивать, как свидетельство функциональной связи или полного отсутствия связи между x и y. Во всех других случаях, когда r_s не принимает крайних значений, коэффициенты Пирсона и Спирмана отличаются несильно. Если же учесть простоту расчета коэффициента Спирмана, то становится понятным, почему многие исследователи отдают ему предпочтение, особенно на начальном этапе исследования.

Пример. Рассчитать коэффициент корреляции рангов Спирмана для почасовой оплаты труда x (руб/час) и текучести кадров y (%/год). Данные выборочного наблюдения представлены в табл. 5, колонки 1 и 2.

Решение. Определим ранги предприятий N_{xi}, N_{yi} по каждой из переменных, колонки 3 и 4 таблицы, найдем их разность d_i, колонка 5, вычислим квадраты полученных значений, колонка 6 и найдем их сумму. Значение коэффициента корреляции рангов будет равно:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)} = 1 - \frac{6 \cdot 164}{8 \times \left(64-1\right)} = -0.952.$$

Полученное значение $r_s = -0.952$ свидетельствует о сильной обратной зависимости между уровнем почасовой оплаты и текучестью кадров.

Таблица 5. *Коэффициент корреляции рангов почасовой оплаты и текучести кадров*

x *(руб/час)*	y *(%/год)*	Ранги		$d_i = N_{xi} - N_{yi}$	d_i^2
		N_{xi}	N_{yi}		
30	**34**	1	**7**	**-6**	**36**
40	**35**	**2**	**8**	**-6**	**36**
50	**33**	**3**	**6**	**-3**	**9**
60	**28**	**4**	**5**	**-1**	1
70	**20**	**5**	**3**	**2**	**4**
80	**24**	**6**	**4**	**2**	**4**
90	**15**	**7**	**2**	**5**	**25**
100	**11**	**8**	**1**	**7**	**49**
$n = 8$					$\sum_{i=1}^{8} d_i^2 = 164$