В.В. Листопад
АПСВ (г.Киев, Украина)

Измерение корреляционной связи для несгруппированных данных с помощью Microsoft Excel.

Окружающий нас мир полон всевозможных взаимосвязей: между уровнем зарплаты и производительностью, между производительностью и уровнем квалификации (стажем работы), между вмешательством государства и состоянием экономики, между объемом выпускаемой продукции и затратами, между спросом и предложением, между годовой прибылью и затратами на отдых и т.п. Когда мы имеем дело с двумерными данными (например, зарплата и стаж работы), то всегда преследуются три основные цели [1]:

1. Описание и понимание взаимосвязи. Знание этой информации может оказать значительную помощь в долгосрочном планировании и принятии других стратегических решений.
2. Прогнозирование и предсказывание нового наблюдения. Например, если количество заказов на определенный вид продукции в этом квартале увеличилось, то следует ожидать увеличения объема продаж. Если взаимосвязь между количеством заказов и объемом продаж обнаружена, то есть достоверная возможность сделать достоверный прогноз продаж на будущее.
3. Регулирование и управление процессом (например, регулировать процесс производства до оптимального уровня прибыли).

Существует два базовых инструмента, с помощью которых анализируют двумерные данные: *корреляционный анализ*, позволяющий оценить тесноту взаимосвязи между двумя факторами, и *регрессионный анализ*, показывающий как можно предсказать или управлять одной из двух переменных с помощью другой. Проверка статистических гипотез позволяет выяснить, является обнаруженная связь между факторами значимой, или она объяснена исключительно случайностью.

Формула для вычисления коэффициента корреляции

$$r_{yx} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \cdot \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}}, \quad (1)$$

где $\overline{X}, \overline{Y}$ - средние значения по выборке.

Заметим что для вычисления коэффициента корреляции в случае когда данные сгруппированы (то есть каждому значению Y соответсвует одно значение X) можно использовать функцию КОРРЕЛ из електронных таблиц Microsoft Excel (категория статистические).

Выборочное уравнение прямой линии регрессии Y на X имеет вид

$$\hat{Y} - \overline{Y} = r_{yx}\frac{\sigma_y}{\sigma_x}(X - \overline{X}), \quad (2)$$

, где $\hat{Y}$ - расчетное значение зависимой переменной, σ_y, σ_x - выборочные среднеквадратические отклонения [2].

$\sigma_x = \sqrt{D(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$. Выборочный коэффициент корелляции r_{yx} для несгруппированных данных вычисляется по формуле

$$r_{yx}=\frac{\sum(n_{xy}xy-n\bar{x}\cdot\bar{y})}{n\sigma_x\sigma_y}=r_{uv},\ u_i=\frac{x_i-c_x}{h_x},v_j=\frac{y_j-c_y}{h_y}\quad(3),$$ где u_i, v_j – условные варианты, C_x, C_y – «ложный нуль» для переменных X та У соотвественно (условимся принимать в качестве ложного нуля варианту имеющую наибольшую частоту), а h_x, h_y - шаг по переменным X та У соответсвенно. Напомним, что упорядочить распределение по равноотстоящим вариантам можно составив из него интервальное распределение и выбрав в качестве вариант средины интервалов.

Пример. Для данной корреляционной таблицы:

Таблица 1.

Y	X								
	5	10	15	20	25	30	35	40	n_y
100	2	1							3
120	3	4	3						10
140			5	10	8				23
160				1		6	1	1	9
180							4	1	5
n_x	5	5	8	11	8	6	5	2	n=50

Установить тесноту связи между У та Х а также построить линейное уравнение регрессии У на Х и проверить значимость полученного коэффициента корреляции для 5% уровня пользуясь електронными таблицами Ms Excel.

Решение [3]. Составим кореляционную таблицу в условных вариантах с помощью (3), выбрав в качестве ложных нулей $C_x=20$. $C_y=140$, шаги соответственно $h_x=5$, $h_y=20$. Таблица 2.

				Ui					
Vj	-3	-2	-1	0	1	2	3	4	n_v
-2	2	1							3
-1	3	4	3						10
0			5	10	8				23
1				1		6	1	1	9
2							4	1	5
n_u	5	5	8	11	8	6	5	2	n=50

Найдем $\bar{u}=\sum\frac{n_u u}{n}$ и $\bar{v}=\sum\frac{n_v v}{n}$ пользуясь функцией СУММПРОИЗВ. Получим $\bar{u}=0{,}2,\bar{v}=0{,}06$ и пользуясь формулами $\sigma_u=\sqrt{\sum\frac{n_u^2u}{n}-\bar{u}^2}$, $\sigma_v=\sqrt{\sum\frac{n_v^2v}{n}-\bar{v}^2}$ и вышеуказанной функцией - $\sigma_u\approx1{,}9,\ \sigma_v=1{,}02$. Найдем $\sum n_{uv}uv$ пользуясь самостоятельным созданием формулы (пять слагаемых в формуле равны нулю так как соответствующие варианты равны 0) или пользуясь функцией СУММПРОИЗВ предварительно выставив столбец V_j справа 8 раз. Получим $\sum n_{uv}uv$=87 и коэффициент корелляции $r_{uv}=r_{yx}\approx0{,}9$. Между У и Х существует прямая тесная связь. Из формул перехода получим $\bar{x}=\bar{u}\cdot h_x+c_x\approx21$ и , $\bar{y}=\bar{v}\cdot h_y+c_y\approx141{,}2$, а также $\sigma_x=\sigma_u\cdot h_x\approx9{,}49,$ и $\sigma_y=\sigma_v\cdot h_v\approx20{,}33$. Подставив найденные величины в формулу (2) получим уравнение прямой линии регрессии У на Х:

$\hat{y} - 141{,}2 = 0{,}9 \cdot \frac{20{,}33}{9{,}49}(x - 21)$ или окончательно $\hat{y} = 100{,}88 + 1{,}92x$. При увеличении x на 1 y увеличится на 1,92.

Значимость полученного коефициента проверим, пользуясь t критерием Стьюдента.

Вычислим наблюдаемое значение критерия пользуясь формулой $t_{íàáë} = r_{yx}\sqrt{\frac{n-2}{1-r_{yx}^2}} = 0{,}9 * \sqrt{\frac{50-2}{1-0{,}9^2}} \approx 14{,}3$ и сравним его с критическим, которое найдем по заданному уровню значимости $\alpha = 0{,}05$ (5%), и числу степеней свободы k-n-2=50 из таблиц критических точек распределения Стьюдента. Заметим, что также можно воспользоваться функцией СТЬЮДРАСПОБР. Получим $t_{êðèò} = t(\alpha; n-2) = t(0{,}05; 48) \approx 2{,}01$. Поскольку $t_{íàáë} > t_{êðèò}$ (14,3 > 2,01) то полученный коефициент корелляции значимый, то есть отвергаем гипотезу о равенстве нулю генерального коэффициента корреляции; следовательно X и У коррелированны..

Среди существенных преимуществ использования електронных таблиц Microsoft Excel при выполнении задач из раздела «Статистика» отметим:

1. Экономию аудиторного времени на практическом занятии;
2. Реализована возможность параллельного усвоения теоретического материала этой темы;
3. Значительно упрощается механизм контроля выполнения задачи;
4. Реализуются междисциплинарные связи (в частности с предметами «Информатика», «Экономика», «Количественные методы исследования социальных процессов», «Теория вероятностей и математическая статистика»).
5. Возможность использовать пакет Microsoft Excel для подготовки системы упражнений.

Литература.

1. Сигел, Эндрю. Практическая бизнес-статистика.: Пер. с англ. – М. : Издательский дом «Вильямс», 2002. – 1056 с. : ил. – Парал. тит. англ.
2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: Учеб. Пособие для студентов вузов. Изд. 5-е, стер. – М.: Высш. школа, 1999. – 400 с.
3. Лабораторний практикум з курсу «Кількісні методи дослідження соціальних процесів».Частина II\Укл. В. В. Листопад – К., 2003. 64 с.