Элементы теории корреляции
3.1.Функциональная, статистическая и корреляционная зависимости

Во многих задачах требуется установить или оценить зависимость случайной величины Y от одной или нескольких других величин. Рассмотрим сначала зависимость Y от одной случайной (или неслучайной величины) X. Две случайные величины могут быть связаны либо функциональной зависимостью, либо зависимостью другого рода, называемой статистической, либо быть независимыми. При функциональной зависимости каждому значению X соответствует вполне определенное значение Y. На практике такая зависимость встречается редко, так как Y помимо X часто зависит от ряда других факторов, подчас остающихся скрытыми. Кроме того, при определении значений X и Y практически всегда присутствуют ошибки измерения. Поэтому общим видом зависимости является *статистическая* зависимость, когда изменение значений X ведет к изменению распределения случайной величины Y. В частности, статистическая зависимость может проявиться в том, что при изменении X меняется среднее значение Y. В этом случае статистическую зависимость называют *корреляционной*. Пусть, например, X – количество вносимых удобрений, а Y – урожай зерна. Тогда с ростом X урожайность в среднем увеличивается, но значение Y не определяется однозначно значением X, так как помимо количества удобрений на урожайность влияет ряд других факторов, часто случайных: погодные условия, количество осадков и т.д.

Пусть $M(Y|X = x)$ – условное математическое ожидание случайной величины Y (среднее значение случайной величины Y при фиксированном значении величины X, равном x). Функция

$$y(x) = M(Y|X = x)$$

называется *регрессией* Y на X, а ее график – *линией регрессии* Y на X.

В простейшем случае эта зависимость линейная:

$$y(x) = \rho_{yx}x + b;$$

где коэффициент ρ_{xy} называется *коэффициентом регрессии* Y на X. Ее графиком является прямая линия.

Заметим, что если X и Y – независимые случайные величины, то $M(Y|X = x) = M(Y)$ и уравнение регрессии примет вид $y(x) = b$, где b = M(Y), т.е. это будет линейная регрессия с коэффициентом регрессии, равным нулю, и горизонтальной линией регрессии.

Получение по выборке уравнения регрессии является важным элементом корреляционного анализа. В зависимости от конкретной задачи

это уравнение можно искать в классе линейных или в более широком классе уравнений. Оно будет, вообще говоря, зависеть от выборки, и поэтому называется выборочным уравнением регрессии. Но, если класс, в котором ищется уравнение, выбран правильно, то с ростом объема выборки выборочная линия регрессии, в силу закона больших чисел, будет приближаться к истинной линии регрессии.

3.2.Парная корреляция. Коэффициент корреляции

Пусть имеется выборка из совместного распределения величин (*Y,X*), в которой величина *Y* принимает значения $y_1, \cdots, y_m$, а величина *X* – значения $x_1, \dots x_k$, причем пара (y_i, x_j) встречается n_{ij} раз. Объем выборки $n = \sum_{ij} n_{ij}$.

Такую выборку удобно представить в виде *корреляционной таблицы*, строки которой соответствуют значениям величины *Y*, а столбцы – значениям *X*. В клетке, образованной *i*-ой строкой и *j*–ым столбцом, записано значение n_{ij}.

По выборке уравнение прямой линии регрессии *Y* на *X,* получим:

$$y(x) = \rho_{yx} x + b.$$

Оценивая по выборке значения ρ_{yx} и b, мы тем самым оцениваем условное математическое ожидание случайной величины Y для каждого значения *x*. Эта оценка имеет вид $M(Y|X = x) = \rho_{xy} + b.$ Как известно, наилучшей оценкой математического ожидания является величина, минимизирующая средний квадрат разности между нею и элементами выборки. Поэтому в качестве оценки величин ρ_{yx} и b берутся такие их значения, которые минимизируют сумму квадратов отклонений наблюдаемых значений от их прогнозируемых математических ожиданий:

$$F(\rho, b) = \sum_{i,j} n_{ij} \left(\rho x_j + b - y_i\right)^2 \to min$$

(ради краткости будем временно вместо ρ_{yx} писать).

Условие минимума F является обращение в нуль частных производных:

$$\frac{\partial F}{\partial \rho} = 2\sum_{ij} n_{ij}\left(\rho x_j + b - y_i\right)x_j = 0;$$

$$\frac{\partial F}{\partial b} = 2\sum_{ij} n_{ij}\left(\rho x_j + b - y_i\right) = 0.$$

Это дает систему двух линейных уравнений относительно ρ и b:

$$\begin{cases}\left(\sum_{ij} n_{ij}x_j^2\right)\rho + \left(\sum_{ij} n_{ij}x_j\right)b = \left(\sum_{ij} n_{ij}y_i x_j\right)\\ \left(\sum_{ij} n_{ij}x_j\right)\rho + \left(\sum_{ij} n_{ij}\right)b = \left(\sum_{ij} n_{ij}y_i\right).\end{cases}$$

Поделив обе части каждого из уравнений на объем выборки n, получаем:

$$\begin{cases}\overline{x^2}\rho + \bar{x}b = \overline{xy}\\ \bar{x}p + b = \bar{y}\end{cases}$$

Второе из этих уравнений показывает, что выборочная линия регрессии проходит через точку $(\bar{x}, \bar{y})$. Ее уравнение, следовательно, может быть записано в виде:

$$y(x) - \bar{y} = \rho_{yx}(x - \bar{x}).$$

Вычтя из первого уравнения системы второе, умноженное на $\bar{x}$, найдем выборочный коэффициент регрессии Y на X:

$$\rho_{yx} = \frac{\overline{xy} - \bar{x}\bar{y}}{\bar{x}^2 - (\bar{x})^2}.$$

Стоящая в знаменателе величина $\bar{x}^2 - (\bar{x})^2$ есть выборочная дисперсия величины X. Обозначим ее через $\tilde{\sigma}_x^2$, где $\tilde{\sigma}_x$ – выборочное среднее квадратическое отклонение. Через $\tilde{\sigma}_y$ обозначим выборочное среднее отклонение величины Y. Тогда

$$\rho_{yx} = \frac{\overline{xy} - \bar{x}\bar{y}}{\tilde{\sigma}_x^2}.$$

Введем величину $r_{\text{в}} = \rho_{yx}\frac{\tilde{\sigma}_x}{\tilde{\sigma}_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{\tilde{\sigma}_x\tilde{\sigma}_y} = \frac{\overline{(x-\bar{x})(y-\bar{y})}}{\tilde{\sigma}_x\tilde{\sigma}_y}$,

которая называется *выборочным коэффициентом корреляции* величин X и Y.

Выразив коэффициент регрессии через коэффициент корреляции, получим уравнение регрессии в виде:

$$\frac{y(x)-\bar{y}}{\tilde{\sigma}_y} = r_{\text{в}}\frac{x-\bar{x}}{\tilde{\sigma}_x}.$$

На практике уравнение регрессии Y на X можно рассматривать как соотношение, позволяющее прогнозировать значение случайной величины Y по известному значению величины X, используя в качестве прогноза значение $y(x) = M(Y|X = x)$.

Изучим свойства выборочного коэффициента корреляции подробнее. Коэффициент корреляции симметричен относительно X и Y. Уравнение регрессии X на Y может быть записано с его помощью как

$$\frac{x(y)-\bar{x}}{\tilde{\sigma}_x} = r_{\text{в}}\frac{y-\bar{y}}{\tilde{\sigma}_y}$$

Рассмотрим величину

$$\sum_{ij} n_{ij}\left(y_i-\bar{y}-r\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}(x_j-\bar{x})\right)^2 =$$

$$= \sum_{ij} n_{ij}(y_i-\bar{y})^2 - 2r\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}\sum_{ij} n_{ij}(y_i-\bar{y})(x_j-\bar{x}) + r^2\left(\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}\right)\sum_{ij} n_{ij}(x_j-\bar{x})^2 =$$

$$= n\tilde{\sigma}_y^2 - 2nr^2\tilde{\sigma}_y^2 + nr^2\tilde{\sigma}_y^2 = n\tilde{\sigma}_y^2 - nr^2\tilde{\sigma}_y^2 = n\tilde{\sigma}_y^2(1-r^2).$$

Исходное выражение, являясь суммой квадратов, неотрицательно. Поэтому $n\tilde{\sigma}_y^2(1-r^2) \geq 0.$ Отсюда следует, что $|r| \leq 1$ или $-1 \leq r \leq 1,$ причем $|r| = 1$ в том и только в том случае, когда все выборочные пары точек лежат на прямой регрессии. Этот случай соответствует строгой линейной функциональной зависимости величин X и Y, когда значение y однозначно определяется значением x. Как уже отмечалось, на практике он встречается редко.

Если, напротив, случайные величины X и Y независимы, то математическое ожидание выборочного коэффициента корреляции как случайной величины равно нулю, и его вычисленное по выборке значение также будет близким к нулю. Поэтому модуль выборочного коэффициента корреляции можно рассматривать как меру линейной функциональной зависимости величин X и Y. Близость модуля коэффициента корреляции к единице говорит о том, что между X и Y имеется сильная линейная связь, и предсказание значения Y по X с помощью уравнения регрессии даст высокую точность.

Здесь следует отметить, что близость коэффициента корреляции к нулю не доказывает отсутствие функциональной связи между X и Y, а говорит лишь об отсутствии линейной функциональной зависимости. В качестве примера рассмотрим случай, когда случайная величина X распределена симметрично относительно нуля, а величина Y связана с X соотношением $Y = X^2$. В этом случае коэффициент корреляции величин X и Y равен нулю, несмотря на наличие между ними жесткой функциональной связи.

На практике, когда по выборке получено некоторое отличное от нуля значение выборочного коэффициента корреляции, может возникнуть вопрос, значимо ли это различие или, другими словами, имеется ли между ними X и Y линейная корреляционная связь. Можно сказать, что если величины X и Y нормальны и независимы, то величина

$$T = r_{\text{в}} \frac{\sqrt{n-2}}{\sqrt{1-r_{\text{в}}^2}}$$

имеет распределение Стьюдента c $k = n-2$ степенями свободы. Для проверки значимости коэффициента корреляции при заданном уровне значимости α по таблице критических точек распределения Стьюдента находят $t_{kp}: P(|T| > t_{kp}) = \alpha$. Если значение величины T, вычисленное по выборке, $|T_{\text{набл}}| > t_{kp}$, то коэффициент корреляции значим, и величины X и Y зависимы.

Пример. Среди владельцев иномарок было выбрано 100 человек. Из стоимости автомашин в тыс. у.е.(X) и годового дохода владельцев также в тыс. у.е.(Y) составлена корреляционная таблица:

Y	X					n_y
	5	10	15	20	25	
10	10	5	-	-	-	15
20	5	10	5	-	-	20
30	5	5	10	5	-	25
40	-	5	5	10	-	20
50	-	5	5	5	5	20
n_x	20	30	25	20	5	$n = 100$

Найти коэффициент корреляции величин X и Y и уравнение прямой линии регрессии Y на X.

Решение. Используя корреляционную таблицу, найдем $\bar{x}$, $\tilde{\sigma}_x$, $y, \tilde{\sigma}_y$ и коэффициент корреляции и $r_в$:

$$\bar{x} = \sum_x x \cdot \frac{n_x}{n} =$$

$$= 5 \cdot 0{,}2 + 10 \cdot 0{,}3 + 15 \cdot 0{,}25 + 20 \cdot 0{,}2 +$$

$$25 \cdot 0{,}05 = 1 + 3 + 3{,}75 + 4 + 1{,}25 = 13;$$

$$\tilde{\sigma}_x^2 = \sum_x x^2 \frac{n_x}{n} - \bar{x}^2 =$$

$$= 25 \cdot 0{,}2 + 100 \cdot 0{,}3 + 225 \cdot 0{,}25 + 400 \cdot 0{,}2 + 625 \cdot 0{,}05 - 169 =$$

$$= 5 + 30 + 56{,}25 + 80 + 31{,}25 - 169 = 33{,}5;$$

$$\tilde{\sigma}_x \approx 5{,}79;$$

$$\bar{y} = \sum_y y \cdot \frac{n_y}{n} =$$

$$= 10 \cdot 0{,}15 + 20 \cdot 0.2 + 30 \cdot 0{,}25 + 40 \cdot 0{,}2 + 50 \cdot 0{,}2 =$$

$$1{,}5 + 4 + 7{,}5 + 8 + 10 = 31;$$

$$\tilde{\sigma}_y^2 = \sum_y y^2 \frac{n_y}{n} - \bar{y}^2 =$$

$$= 10 \cdot 0{,}15 + 400 \cdot 0{,}2 + 900 \cdot 0{,}25 + 1600 \cdot 0{,}2 + 2500 \cdot 0{,}2 =$$

$$= 15 + 80 + 225 + 320 + 500 - 961 = 279;$$

$$\tilde{\sigma}_y \approx 16{,}7;$$

$$\sum_{ij} n_{ij} y_i x_j = \sum_i y_i \sum_j n_{ij} x_j =$$

$$= 10(10 \cdot 5 + 5 \cdot 10) + 20(5 \cdot 5 + 10 \cdot 10 + 5 \cdot 15) +$$

$$+30(5 \cdot 5 + 5 \cdot 10 + 10 \cdot 15 + 5 \cdot 20) + 40(5 \cdot 10 + 5 \cdot 15 + 10 \cdot 20) +$$

$$+50(5 \cdot 10 + 5 \cdot 15 + 5 \cdot 20 + 5 \cdot 25) =$$

$$= 10(50 + 50) + 20(25 + 100 + 75) + 30(25 + 50 + 150 + 100) +$$

$$+40(50 + 75 + 200) + 50(50 + 75 + 100 + 125) =$$

$$= 10 \cdot 100 + 20 \cdot 200 + 30 \cdot 325 + 40 \cdot 325 + 50 \cdot 350 =$$
$$= 1000 + 4000 + 9750 + 13000 + 17500 = 45250;$$

$$r_в = \frac{n_{ij} y_i x_j - n\bar{y}\bar{x}}{n\tilde{\sigma}_x \tilde{\sigma}_y} =$$

$$= \frac{45250 - 100 \cdot 31 \cdot 13}{100 \cdot 5{,}79 \cdot 16 \cdot 7} = \frac{4950}{9670} \approx 0{,}51.$$

Проверяя значимость коэффициента корреляции по указанной выше схеме, получим:

$$T_{набл} = r_в \frac{\sqrt{n-2}}{\sqrt{1-r_в^2}} \approx 0{,}51 \cdot \frac{\sqrt{100-2}}{\sqrt{1-(0{,}51)^2}} \approx 5{,}87.$$

В предположении нормальности и независимости величин X и Y, критическое значение этой величины при уровне значимости $\alpha = 0{,}05,$ найденное по таблице с $k = 100 - 2 = 98,$ равно $t_{kp} \approx 2$, т.е. имеет место $T_{набл} > t_{кр}$, что свидетельствует о существовании линейной зависимости между X и Y.

Уравнение прямой линии регрессии Y на X запишется в виде:

$$\frac{y(x) - 31}{16{,}7} = 0{,}51 \cdot \frac{x - 13}{5.97}$$

или

$$y(x) = 1{,}5x + 12.$$

Линия регрессии представлена на графике. Черными кружками отмечены выборочные значения. Размер кружков соответствует их частотам.