Лекция 15. Элементы теории корреляции.

§1. Функциональная, статистическая и корреляционная зависимости.

Две случайные величины могут быть связаны функциональной зависимостью, т.е. изменение одной из них по определенному закону влечет изменение другой, или зависимостью другого рода, называется статистической, или быть независимыми.

Определение 1. Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой.

Определение 2. Если при изменении одной из величин изменяется среднее значение другой, то в этом случае статистическая зависимость.

*Пример.*называется корреляционной.

Пусть У- урожай зерна, Х- количество удобрений. С одинаковых по площади участков земли при равных количествах внесенных удобрений снимают различный урожай, т.е. У не функция от Х. Это объясняется случайными факторами (осадки, агротехника и т.д.). Как показывает опыт, средний урожай зависит от количество удобрений. Точка У связана с Х корреляционной зависимостью.

§2. Условные средние.

Пусть каждому X отвечает несколько Y. Например, значение X

$x_1 = 2,\ Y - y_1 = 5,\ y_2 = 6,\ y_3 = 10$. Найдем их среднее $\bar{y}_2 = \frac{5+6+10}{3} = 7$ - условное среднее.

Определение 1. Условными средним $\bar{y}_x$ называется среднее арифметическое значение Y, соответствующее значению X.

Определение 2. Корреляционной зависимостью Y от X называют функциональную зависимость условной средней

$\bar{y}_x$ от x: $\bar{y}_x = f(x)$. (1)

Уравнение (1) называется уравнением регрессии Y на X, функцию $f(x)$ называется регрессией Y на X, а ее график - линией регрессии Y на X.

Аналогично определяется корреляционная зависимость X от Y.

§3. Основные задачи теории корреляции.

Таких задач две.

Первая задача теории корреляции – установить форму корреляционной связи, т.е. вид функции регрессии (линейная, квадратичная, показательная и т.д.). Наиболее часто встречаются линейные. Если обе зависимости x от y и y от x линейны, то корреляция – линейная, в противном случае нелинейная.

Вторая задача теории корреляции – оценить тесноту (силу) корреляционной связи. Теснота корреляционной зависимости Y от X оценивается по величине

рассеяния значений *У* вокруг словного среднего $\bar{y}_x$. Большое рассеяние говорит о слабой зависимости *У* от *Х* или об ее отсутствии. Малое рассеяние говорит о сильной связи, возможно и функциональной. Аналогично *Х* от *У*.

§4. Отыскание параметров выборочного уравнение прямой линии регрессии по не сгруппированным данным.

Допустим, корреляционная зависимость между *У, Х* линейная, тогда линии регрессии будут прямыми.

Для отыскания уравнений этих прямых проведено n независимых испытаний, в результате получено n пар чисел

$$(x_1, y_1),\ (x_2, y_2),...,\ (x_n, y_n). \qquad (1)$$

Будем рассматривать их как случайную выборку из генеральной совокупности. Тогда величины и уравнения, найденные по этим данным, называются **выборочными**.

В простейшем случае, когда x соответствует $1y$, можно искомое уравнение $\bar{y}_x = kx + b$ записать в виде $Y = kx + b$, где $k -$ угловой коэффициент прямой линии регрессии и обозначать $k = \rho_{yx}$, тогда уравнение регрессии будет: $Y = \rho_{yx} x + b$. (2)

Выберем параметры ρ_{yx} и b так, чтобы (1) были как можно ближе к (2) на плоскости $X_0 Y$.

Назовем отклонением разность $Y_i - y_i$, $\left(i = \overline{1,n}\right)$, где Y_i вычисляется по уравнению (2), y_i - наблюдаемая из (1). Подберем ρ_{yx} и b так, чтобы сумма квадратов отклонений была минимальной (т.е. это метод наименьших квадратов). Строим функцию $F(\rho, b) = \sum_1^n (Y_i - y_i)^2$ или $F(\rho, b) = \sum_{i=1}^n \left(\rho_{yx} x_i + b - y_i\right)^2$. Для отыскания минимума составим $\begin{aligned} \frac{\partial F}{\partial \rho} &= 2\sum_1^n (\rho x_i + b - y_i) x_i = 0 \\ \frac{\partial F}{\partial \rho} &= 2\sum_1^n (\rho x_i + b - y_i) = 0 \end{aligned}$ и отсюда получим систему 2-х линейных уравнений относительно ρ и b $\left.\begin{aligned} \left(\sum x^2\right)\rho + \left(\sum x\right) b &= \sum xy \\ \left(\sum x\right)\rho + \sum nb &= \sum y \end{aligned}\right\}$, решив ее, найдем $\rho = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2};\ b = \frac{\sum x^2 \sum y - \sum x \sum yx}{n\sum x^2 - \left(\sum x\right)^2}$ получим искомое уравнение $y = \rho x + b$.

Пример. Найти выборочное уравнение прямой линии регрессии *У* на *Х* по данным наблюдений $n = 5$.

x	1	1,5	3	4,5	5
y	1,25	1,4	1,5	1,75	2,25

Решение: составим расчетную таблицу

	x_i	y_i	x_i^2	$x_i y_i$
1	1	1,25	1	1,25
2	1,5	1,4	2,25	2
3	3	1,5	9	4,5
4	4,5	1,75	20,25	4,875
5	5	2,25	25	11,25
$\sum x_i = 15$		$\sum y_i = 8{,}15$	$\sum x_i^2 = 5{,}75$	$\sum x_i y_i = 26{,}975$

По известным формулам вычислим $\rho_{yx} = \dfrac{5 \cdot 26.975 - 15 \cdot 8{,}15}{5 \cdot 57{,}5 - 15^2} = 0{,}202$

$b = \dfrac{57{,}5 \cdot 8{,}15 - 15 \cdot 26{,}975}{62{,}5} = 1{,}024$

Напиши искомое уравнение регрессии

$$Y = 0{,}202x + 1{,}024 .$$

Для того, чтобы получить представление, насколько хорошо вычисленные по этому уравнению значения Y_i согласуются с наблюдаемыми значениями y_i, найдем отклонение $Y_i - y_i$.

x_i	Y_i	y_i	$Y_i - y_i$
1	1,226	1,25	-0,024
1,5	1,327	1,4	-0,073
3,00	1,630	1,5	0,130
4,5	1,933	1,75	0,083
5	2,034	2,25	-0,216

Из таблицы видно, что не все отклонения малы. Это объясняется числом наблюдений.

§5. Корреляционная таблица.

При большом числе наблюдений одно и то же значение *x* может встретиться n_x раз, одно и то же значение *y* n_y раз, одна и та же пара чисел (x,y) может наблюдаться n_{xy} раз. Поэтому данные наблюдений группируются, т.е. подсчитывают частоты n_x, n_y, n_{xy}. Все сгруппированные данные записывают в виде таблицы, которую называют корреляционной.

Пример.

X	10	20	30	40	n_y
У					

0,4	5	-	7	14	26
0,6	-	2	6	4	12
0,8	3	19	-	-	22
n_x	8	21	13	18	$n = 60$

§6. Отыскание параметров выборочного уравнения прямой линии регрессии по сгруппированным данным. Выборочный коэффициент корреляции.

Для определения параметра уравнение прямой линии регрессии *У* на *Х* была получена система уравнений

$$\begin{cases} \left(\sum x^2\right)\rho_{yx} + \left(\sum x\right)b = \sum xy \\ \left(\sum x\right)\rho_{yx} + nb = \sum y \end{cases} \qquad (*)$$

ρ_{yx} - коэффициент регрессии.

Предполагалось, что значения *Х* и соответствующие им значения *У* наблюдались по одному разу.

Теперь же допустим, что получено большое число данных (≈ 50 наблюдений), среди них есть повторяющиеся и они сгруппированы в виде корреляционной таблицы. Запишем систему (*) так, чтобы она отражала данные корреляционной таблицы

$$\sum x = n\bar{x} \text{ (следствие из } \bar{x} = \frac{\sum x^2}{n})$$

$$\sum y = n\bar{y}$$

$$\sum x^2 = n\bar{x}^2 \text{ (из } \bar{x}^2 = \frac{\sum x^2}{n})$$

$$\sum xy = \sum n_{xy}x_y \text{ (учтено, что (ху) наблюдается } n_{xy} \text{ раз)},$$

тогда (*) примет вид

$$\left.\begin{array}{l} \left(n\bar{x}^2\right)\rho_{xy} + \left(n\bar{x}\right)b = \sum n_x xy \\ n\left(\bar{x}\right)\rho_{xy} + b = \sum \bar{y} \end{array}\right\} \qquad (**)$$

Решив эту систему, найдем ρ_{yx} и b и получим $\bar{y}_x = \rho_{yx}x + b$.

Однако, удобнее, введя новую величину - коэффициент корреляции, написать уравнение регрессии в ином виде. Сделаем это $b = \bar{y} - \rho_{yx}\bar{x}$, подставив в уравнение регрессии, получим

$$\bar{y}_x = \rho_{yx}x + \bar{y} - \rho_{yx}\bar{x}$$

$$\bar{y}_x - \bar{y} = \rho_{yx}\left(x - \bar{x}\right) \qquad (***)$$

найдем из (**) ρ_{yx}, учитывая, что $\bar{x}^2 - \left(\bar{x}\right)^2 = \sigma_x^2$

$$\rho_{yx} = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\left[\bar{x}^2 - \left(\bar{x}\right)^2\right]} = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\sigma_x^2}$$

$$\rho_{yx} \cdot \frac{\sigma_x}{\sigma_y} = \frac{\sum n_{xy} xy - n\overline{x}\overline{y}}{n\sigma_x^2}$$

r_B – выборочный коэффициент корреляции

$$\rho_{yx} = r_B \cdot \frac{\sigma_y}{\sigma_x}$$

$\overline{y}_x - \overline{y} = r_B \frac{\sigma_y}{\sigma_x}(x - \overline{x})$ - уравнение прямой линии регрессии *У* на *Х*.

§7. Свойства выборочного коэффициента корреляции.

Цель введения этой величины – оценка тесноты линейной корреляционной зависимости. Это следует из его свойств.

$$S_y = D_y\left(1 - r_b^2\right), \qquad S_x = D_x\left(1 - r_b^2\right),$$

где S_y - дисперсия наблюдавшихся значений у вокруг соответствующих условных средних $\overline{y}_x$;

D_y - дисперсия наблюдений вокруг общей средней $\overline{y}$. Аналогично S_x и D_x.

1. абсолютная величина r_B не превосходит 1.

Доказательство: $S_y = D_y\left(1 - r_b^2\right) \geq 0$

$$1 - r_B^2 \geq 0, \quad -1 \leq r_B \leq 1 \text{ или } |r_B| \leq 1.$$

2. если r_B=0 и выборочные линии регрессии – прямые, то *Х* и *У* не связаны линейной корреляционной зависимостью.

Доказательство: если r_B=0 $\qquad \overline{y}_x - \overline{y} = r_B \frac{\sigma_y}{\sigma_x}(x - \overline{x})$

$$\overline{y}_x - \overline{y} = 0 \Rightarrow \overline{y}_x = \overline{y}$$

$\overline{y}_x$ - сохраняет постоянное значение.

3. Если $|r_B| = 1$, то $S_y = D_y\left(1 - r_b^2\right) = 0$.
4. С возрастанием $|r_B|$ линейная корреляционная зависимость становится более тесной и при $|r_B| = 1$ переходит в функциональную зависимость.

Из этих всех свойств следует смысл r_B: выборочный коэффициент корреляции характеризует тесноту связи между количественными признаками в выборке, чем ближе $|r_B| \to 1$, тем связь сильнее, чем ближе $|r_B| \to 0$, тем слабее.

Для оценки коэффициента корреляции r_B нормально распределенной генеральной совокупности ($n \geq 50$) можно воспользоваться формулой

$$r_B - 3\frac{1 - r_B}{\sqrt{4}} \leq r_2 \leq r_B + 3\frac{1 + r_B^2}{\sqrt{4}}$$

8. Метод четырех полей вычисления коэффициента корреляции.

Дана корреляционная таблица. Вычислить r_B -?

Перейдем к угловым вариантам $u_i = \frac{x_i - a}{n_1}$, $\upsilon_j = \frac{y_j - a}{n_2}$, тогда $r_B = \frac{\sum u_{n_n} u\upsilon - n\overline{u}\,\overline{\upsilon}}{n\sigma_n\sigma_\upsilon}$.

$\overline{u}, \overline{\upsilon}, \sigma_n$ и σ_υ - вычисляются по методу произведений.

Для вычисления $\sum n_{nu} u \cdot \upsilon$ применяется метод 4-х полей. Название метода связано с тем, что строка и столбец, пересекающиеся в клетке, содержащей наибольшую частоту, делят корреляционную таблицу на Н-части, которые называются полями. Рассмотрим метод на примере.

X / У	10	20	30	40	50	60	n_y
15	5	7	-	-	-	-	12
25	-	20	23	-	-	-	43
35	-	-	30	47	2	-	79
45	-	-	10	11	20	6	47
55	-	-	-	9	7	3	19
n_x	5	27	63	67	29	9	$n = 200$

В качестве C_1 взят вариант 40, имеющий наибольшую частоту

$$\left.\begin{array}{l} h_1 = 20-10 = 10 \\ C_2 = 35,\ \ h_2 = 25-15 = 10 \end{array}\right\rangle u = \frac{x-40}{10},\ \upsilon = \frac{y-35}{10}$$

u / υ	-3	-2	-1	0	1	2	n_u	I	II
-2	6 | 5	4 | 7					12	12	
-1		2 | 20	1 | 23				43	63	
0			30	47	2	-	79	III	IV
1			-1 | 10	11	11 | 20	2 | 6	47	-10	32
2				9	2 | 7	4 | 3	19	-	26
n_u	5	27	63	67	29	9	200		
I	30	68	23	II			121		
III			-10	IV					

u / υ	-3	-2	-1	1	2	I	II
-2	6 | 5	4 | 7				58	
-1		2	1			63	

		20	23				
1			-1 10	1 20	2 6	-10	32
2				2 7	4 3		26
I	30	68	23			121	
III			-10	34	24	-10	58

121+58=169

$\bar{u}, \bar{\upsilon}, \sigma_n$ и σ_υ,

$\sigma_u = \sqrt{\bar{u}^2 - (\bar{u})^2}, \qquad \sigma_\upsilon = \sqrt{\bar{\upsilon}^2 - (\bar{\upsilon})^2}$

$$\bar{u} = \frac{\sum n_u u}{n} = \frac{5(-3) + 27(-2) + 63(-1) + 29 \cdot 1 + 9 \cdot 2}{200} = 0{,}425$$

$$\bar{\upsilon} = \frac{\sum n_\upsilon \upsilon}{n} = \frac{12(-1) + 43(-2) + 47 \cdot 1 + 19 \cdot 2}{200} = 0{,}09$$

$$\bar{u}^2 = \frac{5 \cdot 9 + 27 \cdot 4 + 63 \cdot 1 + 29 \cdot 1 + 9 \cdot 4}{200} = 1{,}405$$

$$\sigma_u = \sqrt{\bar{u}^2 - (\bar{u})^2} = \sqrt{1{,}405 - 0{,}425^2} = 1{,}106$$

$\sigma_\upsilon = 1{,}209$

$\sum n_{u\upsilon} u\upsilon = 121 - 10 + 58 = 169.$

$$r_B = \frac{\sum n_{u\upsilon} u\upsilon - n\bar{u}\,\bar{\upsilon}}{n\sigma_u \sigma_\upsilon} = \frac{169 - 200(-0{,}425) \cdot 0{,}09}{200 \cdot 1{,}106 \cdot 1{,}203} = 0{,}603$$

$r_B = 0{,}603$.

§9. Пример на отыскание выборочного уравнения прямой линии регрессии.

Так как имеются r_B и $\bar{u}, \bar{\upsilon}, \sigma_n$ и σ_υ, то $\sigma_x = h_1 \sigma_u$, $\sigma_y = h_2 \sigma_\upsilon$, $\bar{x} = \bar{u} h_1 + C_1$, $\bar{y} + \bar{\upsilon} h_2 + C_2$.

Пример. Общий вид уравнения

$$\bar{y}_x - \bar{y} = r_B \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$\bar{x} = \bar{u} h_1 + C_1 = -0{,}425 \cdot 10 + 40 = 35{,}75$

$\bar{y} = \upsilon h_2 + C_2 = 0{,}09 \cdot 10 + 35 = 35{,}9$

$\sigma_x = \sigma_u h_1 = 1{,}106 \cdot 10 = 11{,}06$

$\sigma_y = \sigma_\upsilon h_2 = 1{,}209 \cdot 10 = 12{,}09$

$$\bar{y}_x - 35{,}9 = 0{,}603 \cdot \frac{12{,}09}{11{,}06}(x - 35{,}79)$$

$$\bar{y}_x = 0{,}659x + 12{,}34$$