Функциональная, статистическая и корреляционная зависимость

Рассмотрим систему двух случайных величин $\{X, Y\}$.

Эти случайные величины **могут быть независимыми:**

$$f(x, y) = f_1(x) \cdot f_2(y) \quad . \tag{5.1}$$

В против случае между ними может быть:

функциональная зависимость:

$$y = g(x) \, .$$

статистическая зависимость:

$$\begin{aligned} \varphi(x / y) &= f(x, y) / f_2(y) \\ \phi(y / x) &= f(x, y) / f_1(x) \end{aligned} \tag{5.2}$$

Одним из видов (частным случаем) статистической зависимости является **корреляционная зависимость.**

Корреляционная зависимость, корреляционный момент, коэффициент корреляции

Корреляционной называют статистическую зависимость двух случайных величин, при которой изменение значения одной из случайных величин приводит к изменению математического ожидания другой случайной величины.

$$\begin{aligned} M(X / y) &= q_1(y) \\ M(Y / x) &= q_2(x) \end{aligned} \tag{5.3}$$

Функции (5.3) называют функциями регрессии.

Корреляционный момент:

$$\mu_{xy} = M\{[x - M(X)] \cdot [y - M(Y)]\} \tag{5.4}$$

Коэффициент корреляции:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \sigma_y} \tag{5.5}$$

Корреляционная зависимость, коэффициент корреляции

Для коэффициента корреляции справедливо соотношение:

$$\left|r_{xy}\right| \le 1 \qquad (5.6)$$

Случайные величины X и Y называют коррелированными, если их корреляционный момент или (что тоже самое) их коэффициент корреляции отличен от нуля. В противном случае эти величины некоррелированы.

Если случайные величины X и Y коррелированы, то они зависимы.

Обратное предположение в общем случае неверно. То-есть, если случайные величины X и Y некорелированы, то они могут быть, как независимыми, так и зависимыми.

Корреляционная зависимость, коэффициент корреляции

Можно представить эти утверждения в наглядной форме:

X и Y коррелированы $\Rightarrow$ СВ X и Y зависимы

X и Y независимы $\Rightarrow$ СВ X и Y некоррелированы

Если случайные величины X и Y коррелированны и обе функции регрессии Y на X и X на Y линейны, то говорят, что X и Y связаны линейной корреляционной зависимостью. В частности, это имеет место если двумерная случайная величина $\{X, Y\}$ распределена нормально. Коэффициент корреляции, как известно, служит мерой тесноты линейной зависимости между случайными величинами X и Y. При $\left|r_{xy}\right| = 1$ эта зависимость становится функциональной.

Корреляционная таблица

Пусть имеется выборка двумерной случайной величины $\{X, Y\}$. Ее значения удобно представлять в виде так называемой корреляционной таблицы. Например,

Y	X				
	10	20	30	40	n_y
0,4	5	—	7	14	26
0,6	—	2	6	4	12
0,8	3	19	—	—	22
n_x	8	21	13	18	$n=60$

Статистическая оценка коэффициента корреляции

Значение $\overline{r}_{xy}$ - статистической оценки r_{xy} - коэффициента корреляции можно вычислить по формуле:

$$\overline{r}_{xy} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij} y_i x_j - N\overline{x}_в \overline{y}_в}{NS_x S_y} \qquad (5.7)$$

При N>50 в случае нормального распределения системы случайных величин $\{X, Y\}$ для оценки значения r_{xy} можно использовать соотношение:

$$\overline{r}_{xy} - 3\frac{1-\overline{r}_{xy}^{\,2}}{\sqrt{N}} \le r_{xy} \le \overline{r}_{xy} + 3\frac{1+\overline{r}_{xy}^{\,2}}{\sqrt{N}} \qquad (5.8)$$

Для условных вариант формула такая же

Вычисление выборочного коэффициента корреляции

При вычислении выборочного коэффициента корреляции необходимо вычислить $\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij} y_i x_j = \sum_{i=1}^{K_y} y_i \sum_{j=1}^{K_x} n_{ij} x_j = \sum_{j=1}^{K_x} x_j \sum_{i=1}^{K_y} n_{ij} y_i$.

Это удобно производить в табличной форме:

Y \ X	3	7	$X_i=\sum_{j=1}^{K_x} n_{ij}x_j$	$y_i X_i$
1	15 5 5	49 7 7	64	64
5	9 3 15	28 4 20	37	185
8	12 4 32	14 2 16	26	208
$Y_j=\sum_{i=1}^{K_y} n_{ij}y_i$	52	43		457
$x_j Y_j$	156	301	457	

Доверительный интервал для выборочного коэффициента корреляции

Распределение $\overline{r}_{xy}$ при определенных условиях можно удовлетворительно аппроксимировать нормальным законом. Однако при увеличении интенсивности корреляционной связи распределение $\overline{r}_{xy}$ становится все более ассиметричным.

С помощью преобразования Фишера перейдем к случайной величине z:

$$\overline{z} = 0.5\ln\frac{1+\overline{r}_{xy}}{1-\overline{r}_{xy}} = 1.1513\lg\frac{1+\overline{r}_{xy}}{1-\overline{r}_{xy}} \qquad (5.9)$$

Доверительный интервал для выборочного коэффициента корреляции

Распределение z при неограниченном возрастании объема выборки асимптотически нормальное со значением СКВО, равным:

$$\bar{\sigma}_z = \frac{1}{\sqrt{N-3}} \qquad (5.10)$$

В результате доверительный интервал для r_{xy} генеральной совокупности с доверительной вероятностью γ определяют по следующей схеме:

1. По формуле (5.9) вычисляют выборочное значение $\bar{z}$;
2. По формуле (5.10) вычисляют значение $\bar{\sigma}_z$;

Доверительный интервал для выборочного коэффициента корреляции

3. Доверительный интервал для генерального значения представляется в виде:

$$\left(\bar{z} - \lambda(\gamma)\bar{\sigma}_z, \bar{z} + \lambda(\gamma)\bar{\sigma}_z\right), \qquad (5.11)$$

где значение $\lambda(\gamma)$ должно удовлетворять условию:

$$\Phi\left[\lambda(\gamma)\right] = \frac{\gamma}{2} \qquad (5.12)$$

4. Для пересчета интервала (5.11) в доверительный интервал для коэффициента корреляции с тем же значением γ необходимо воспользоваться обратным преобразованием Фишера:

$$r = th(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{e^{2z} - 1}{e^{2z} + 1} \qquad (5.13)$$

Проверка гипотезы о значимости выборочного коэффициента корреляции

Пусть имеется выборка объема N значений двумерной нормально распределенной случайной величины $\{X,Y\}$ и вычислено значение выборочного коэффициента корреляции $\overline{r}_{xy} \neq 0$. Поскольку $\overline{r}_{xy}$ является случайной величиной, то это еще не значит что r_{xy} - коэффициент корреляции для генеральной совокупности тоже отличен от нуля.

Возникает необходимость проверить гипотезу $H_0 : r_{xy} = 0$.

Альтернативной будет гипотеза $H_1 : r_{xy} \neq 0$.

Если основная гипотеза H_0 отвергается, то это означает, что выборочный коэффициент корреляции $\overline{r}_{xy}$ значимо отличается от нуля (значим). В противном случае – $\overline{r}_{xy}$ незначим.

Проверка гипотезы о значимости выборочного коэффициента корреляции

В качестве критерия проверки статистической гипотезы о значимости выборочного коэффициента корреляции можно принять случайную величину:

$$T = \frac{\overline{r}_{xy}\sqrt{N-2}}{\sqrt{1-\overline{r}_{xy}^{2}}} \qquad (5.14)$$

При справедливости нулевой гипотезы H_0 случайная величина T распределена по закону Стьюдента с $k=N$-2 степенями свободы.

Критическая область для данного критерия двусторонняя.

Проверка гипотезы о значимости выборочного коэффициента корреляции

Проверка гипотезы осуществляется по стандартной схеме:

1. По формуле (5.14) вычисляется значение $T_{набл}$;
2. По заданному уровню значимости α и значению k из таблицы определяется значение $t_{крит}(\alpha, k)$;
3. Если $|T_{набл}| \le t_{крит}(\alpha, k)$ - нет оснований отвергать гипотезу H_0 .
4. Если $|T_{набл}| > t_{крит}(\alpha, k)$ - основная гипотеза H_0 с выборочными данными и должна быть отвергнута.