Статистические методы обработки экспериментальных данных

Лекция №5

Санкт-Петербург
2022

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.

$$M(X/y)=q_1(y)$$

$$M(Y/x)=q_2(x) \qquad (5.3)$$

Функции (5.3) называют функциями регрессии.

Корреляционный момент:

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

Коэффициент корреляции:

$$r_{xy}=\frac{\mu_{xy}}{\sigma_x\sigma_y} \qquad (5.5)$$

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

Корреляционная зависимость, коэффициент корреляции

Для коэффициента корреляции справедливо соотношение:

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

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 $|r_{xy}|=1$ 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:

$$\overline{r}_{xy} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x}\left[n_{ij}y_ix_j\right] - N\overline{x}_в\overline{y}_в}{NS_xS_y} \qquad (5.7)$$

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

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

Вычисление выборочного коэффициента корреляции

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

Это удобно производить в табличной форме:

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

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

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 $\overline{z}$;
2. По формуле (5.10) вычисляют значение $\overline{\sigma}_z$;

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 $\bar{r}_{xy} \neq 0$. Поскольку $\bar{r}_{xy}$ 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 $\bar{r}_{xy}$ значимо отличается от нуля (значим). В противном случае – $\bar{r}_{xy}$ незначим.

Проверка гипотезы о значимости выборочного коэффициента корреляции

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

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

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.

Критические точки распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29