УНИВЕРСИТЕТ ИТМО

Теория вероятностей

Лекция 5. Статистическая и корреляционная зависимость. Коэффициент корреляции, уравнение прямой линии регрессии

Независимость случайных величин

Случайные величины называются ***независимыми***, если закон распределения одной из них не зависит от значений, принимаемой другой случайной величиной.

Теорема. *Для того чтобы случайные величины X и Y были независимы, необходимо и достаточно, чтобы функция распределения системы (X, Y) была равна произведению функций распределения ее составляющих*

$$F(x,y) = F_1(x)F_2(y).$$

Аналогичную теорему можно сформулировать и для плотности распределения.

Теорема. *Для того чтобы случайные величины X и Y были независимы, необходимо и достаточно, чтобы плотность совместного распределения системы (X, Y) была равна произведению функций плотностей распределения ее составляющих*

. Числовые характеристики системы случайных величин

Для системы нескольких случайных величин тоже можно ввести числовые характеристики. Их смысл аналогичен одномерным случайным величинам. Применять будем обозначения в соответствии с принятыми нормами.

Определение. *Начальным моментом порядка* k,s двумерной случайной величины (X,Y) называется математическое ожидание произведения X^k на Y^s:

$$a_{k,s} = M(X^k Y^s),$$

для дискретных случайных величин формула имеет вид

$$a_{k,s} = \sum_i \sum_j x_i^k y_j^s p_{ij},$$

а для непрерывных

$$a_{k,s} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^s f(x,y) dxdy.$$

Очевидно, $a_{1,0} = M(X)$, $a_{0,1} = M(Y)$ – математические ожидания случайных величин X и Y соответственно.

. Числовые характеристики системы случайных величин

Определение. *Центральным моментом порядка* k,s двумерной случайной величины (X,Y) называется математическое ожидание произведения $(X-M(x))^k$ на $(Y-M(y))^s$:

$$\mu_{k,s}=M\left((X-M(x))^k(Y-M(Y))^s\right),$$

для дискретных случайных величин

$$\mu_{k,s}=\sum_i\sum_j(x_i-M(X))^k\left(y_j-M(Y)\right)^s p_{ij},$$

а для непрерывных

$$\mu_{k,s}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-M(X))^k\left(y-M(Y)\right)^s f(x,y)dxdy.$$

Очевидно, $\mu_{2,0}=D(X)$, $\mu_{0,2}=D(Y)$ – дисперсии случайных величин X и Y соответственно.

. Числовые характеристики системы случайных величин

При $k=s=1$ в теории корреляции используется корреляционный момент $\mu=\mu_{1,1}=M\big((X-M(X)(Y-M(Y)\big)$. На практике для нахождения $M(X)$, $M(Y)$, а так же $D(X)$, $D(Y)$ переходят, суммируя в таблице p_{ij} по строкам или столбцам, к одномерным случайным величинам Y и X.

Пример 1. Вычислить числовые характеристики каждой из случайных величин.

Y	X			
	1	*3*	*4*	*10*
-2	0	0.04	0.03	0.07
1	0.1	0.15	0.15	0.25
4	0.05	0.06	0.07	0.03

Y		
-2	*1*	*4*
0.14	0.65	0.21

X			
1	*3*	*4*	*10*
0.15	0.25	0.25	0.35

Числовые характеристики системы случайных величин

Y		
-2	1	4
0.14	0.65	0.21

X			
1	3	4	10
0.15	0.25	0.25	0.35

$$M(Y)=a_{1,0}=-2\cdot 0.14+1\cdot 0.65+4\cdot 0.21=1.21;$$

$$M(X)=a_{0,1}=1\cdot 0.15+3\cdot 0.25+4\cdot 0.25+10\cdot 0.35=5.4;$$

$$D(Y)=\mu_{2,0}=M(Y^2)-M^2(Y)=(-2)^2\cdot 0.14+1^2\cdot 0.65+4^2\cdot 0.21-(1.21)^2=3.11;$$

$$D(X)=\mu_{0,2}=M(X^2)-M^2(X)=$$

$$=1^2\cdot 0.15+3^2\cdot 0.25+4^2\cdot 0.25+10^2\cdot 0.35-(5.4)^2=12.24.$$

Условный закон распределения

мы рассматривали условную вероятность $p_A(B)$ наступления события B при условии, что уже произошло событие A. Аналогично для системы случайных величин рассматривается распределение одной СВ при условии, что другая приняла конкретные значения.

Условным законом распределения одной из случайных величин, входящих в систему, называется распределение, найденное из условия, что другая случайная величина приняла определенной значение.

Условная плотность распределения вычисляется по формулам:

$$\varphi(x/y)=\frac{f(x,y)}{f_2(y)}=\frac{f(x,y)}{\int\limits_{-\infty}^{\infty}f(x,y)dx};$$

$$\psi(y/x)=\frac{f(x,y)}{f_1(x)}=\frac{f(x,y)}{\int\limits_{-\infty}^{\infty}f(x,y)dy}.$$

Условная плотность распределения обладает всеми свойствами плотности распределения одной случайной величины.

Условное математическое ожидание

Условным математическим ожиданием дискретной случайной величины Y при фиксированном $X = x$ называется произведение всех возможных значений Y на их условные вероятности

$$M(Y/X=x)=\sum_{j=1}^{m} y_j p(y_j/x),$$

для непрерывных случайных величин имеем

$$M(Y/X=x)=\int_{-\infty}^{\infty} y\psi(y/x)dy.$$

где $\psi(y/x)$ – условная плотность случайной величины Y при $X = x$.

Условное математическое ожидание

Условным математическим ожиданием дискретной случайной величины X при $Y = y$ (y – одно из возможных значений Y) называется произведение всех возможных значений X на их условные вероятности

$$M(X/Y=y)=\sum_{i=1}^{n} x_i\, p(x_i/y),$$

для непрерывных случайных величин имеем

$$M(X/Y=y)=\int_{-\infty}^{\infty} x\varphi(x/y)dx,$$

где $\varphi(x/y)$ – условная плотность случайной величины X при $Y = y$, определяемая формулой (64).

$M(Y/X=x)=g(x)$ ($M(X/Y=y)=h(y)$) – условные математические ожидания являются функциями от x (от y) и называются ***функциями регрессии*** X *на* Y (Y *на* X).

Условное математическое ожидание

Пример 2. Вычислить числовые условные характеристики каждой из случайных величин.

X	Y			
	1	*3*	*4*	*10*
-2	0	0.04	0.03	0.07
1	0.1	0.15	0.15	0.25
4	0.05	0.06	0.07	0.03

$$M(X/Y=1)=-2\cdot 0+1\cdot 0{,}1+4\cdot 0{,}05=0{,}3;$$

$$M(X/Y=3)=0.31;\ M(X/Y=4)=0{,}37;\ M(X/Y=10)=0{,}23;$$

$$M(Y/X=-2)=1\cdot 0+3\cdot 0{,}04+4\cdot 0{,}03+10\cdot 0{,}07=0{,}94;$$

$$M(Y/X=1)=3{,}65;\ M(Y/X=4)=0{,}81.$$

Статистическая и корреляционная зависимость

При работе с системами случайных величин основным вопросом является их взаимодействие, их влияние друг на друга.

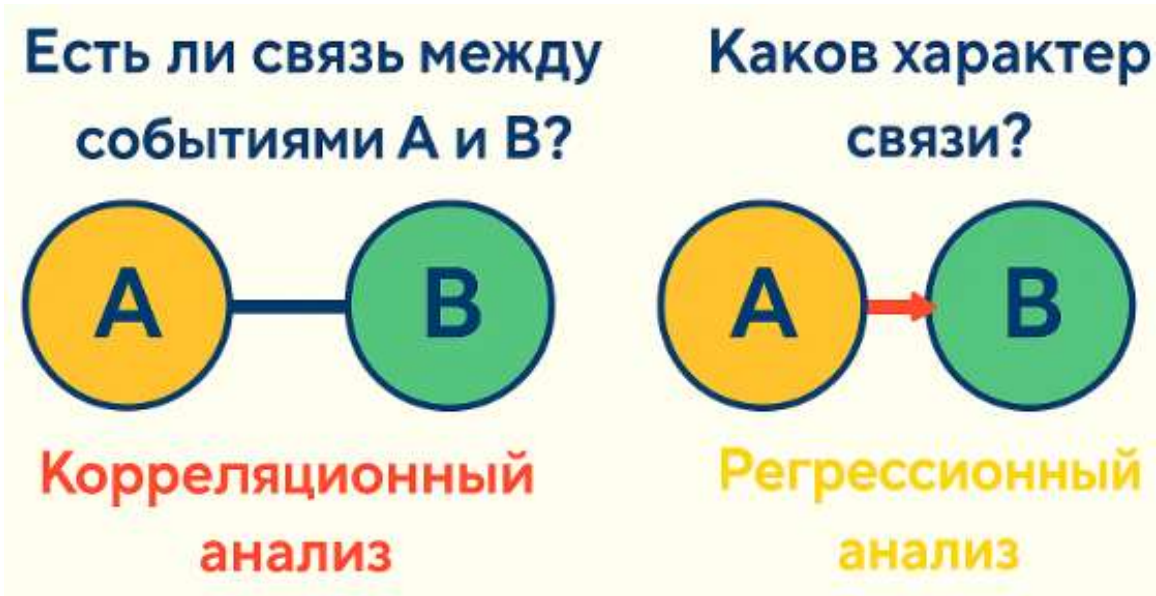
В естественных науках основной задачей является построение функциональной зависимости между переменными, т. е. когда одному конкретному значению одной переменной соответствует единственное значение другой переменной (например, зависимость расстояния от времени).

В прикладных задачах часто существуют другие зависимости, это связи, когда одному значению одной переменной может соответствовать множество значений другой переменной, что зависит от каких-то дополнительных условий, – такая зависимость называется *статистической*. Например, статистической зависимостью является зависимость успеваемости студента от его присутствия на занятиях, от его навыков, полученных в школе, от его целеустремленности.

Статистическая и корреляционная зависимость

В силу неоднозначности статистической зависимости рассматривается зависимость между случайными величинами, которая определяет зависимость среднего значения одной переменной от среднего значения другой переменной. Такая статистическая зависимость называется *корреляционной*. Она может описываться через линейные, квадратичные и другие функции, что и определяет вид связи.

На рис. указаны задачи, решаемые для систем случайных величин.



Статистическая и корреляционная зависимость

Подводя итог вышесказанному, можно сказать, что корреляционный анализ решает задачи об определении связи между двумя явлениями – скоростью в беге и ростом спортсмена или общей успеваемостью студента и его оценкой по математике и т. д., а регрессионный анализ позволяет понять характер этой зависимости. В нашем курсе рассматривается и устанавливается только линейная зависимость. При этом нужно понимать, что это далеко не единственный вид зависимости, который существует.

Корреляционная зависимость может быть представлена в виде:

$$M(Y/X=x)=\varphi(x),$$

$$M(X/Y=y)=\psi(y),$$

или в виде:

$$\overline{y_x}=\varphi(x),$$

$$\overline{x_y}=\psi(y)$$

Уравнения называются *уравнениями регрессии* соответственно Y по X и X по Y, функции $\varphi(x)$, $\psi(y)$ – *функциями регрессии*, а их графики – *линиями регрессии*.

Уравнение прямой линии регрессии

Пусть в результате n независимых опытов получены n пар чисел: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. По этим данным наблюдений надо составить уравнение прямой линии регрессии Y на X:

$$\overline{y_x} = kx + b.$$

Угловой коэффициент прямой линии регрессии Y на X называют *выборочным коэффициентом* регрессии Y на X и обозначают через ρ_{yx}.

Таким образом, будем искать выборочное уравнение прямой линии регрессии Y на X в виде уравнения

$$\boldsymbol{\overline{y_x} = \rho_{yx} x + b}$$

Подберём параметры ρ_{yx} и b так, чтобы точки (x_1, y_1), (x_2, y_2), (x_3, y_3),..., (x_n, y_n) построенные по данным наблюдения, на плоскости $x0y$ лежали как можно ближе к прямой, y_k – наблюдаемые ординаты, соответствующие x_k.

Уравнение прямой линии регрессии

Применим метод наименьших квадратов, который заключается в том, чтобы сумма квадратов отклонений была наименьшей. Так как каждое отклонение зависит от отыскиваемых параметров, то и сумма квадратов отклонений есть функция F этих отклонений $F(\rho,b)=\sum_{k=1}^{n}(\rho x_k+b-y_k)^2$, где упрощаем запись $\rho=\rho_{yx}$.

Для вычисления минимума приравниваем нулю частные производные:

$$\begin{cases}\dfrac{\partial F}{\partial \rho}=2\sum_{k=1}^{n}(\rho x_k+b-y_k)x_k=0;\\ \dfrac{\partial F}{\partial b}=2\sum_{k=1}^{n}(\rho x_k+b-y_k)=0.\end{cases}$$

Уравнение прямой линии регрессии

$$\begin{cases} \dfrac{\partial F}{\partial \rho} = 2\sum\limits_{k=1}^{n}(\rho x_k + b - y_k)x_k = 0; \\ \dfrac{\partial F}{\partial b} = 2\sum\limits_{k=1}^{n}(\rho x_k + b - y_k) = 0. \end{cases}$$

Преобразуя эти уравнения, получим систему двух линейных уравнений относительно ρ и b, далее не будем писать индексы у знаков сумм, считая k=1, 2,…, n:

$$\begin{cases} \rho\sum x_i^2 + b\sum x_i = \sum x_i y_i; \\ \rho\sum x_i + nb = \sum y_i. \end{cases}$$

Решив эту систему, найдём искомые коэффициенты в уравнении

$$\rho_{yx} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2},$$

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2},$$

Уравнение прямой линии регрессии

или, упростив запись,

$$\rho_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2},\ b = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}.$$

Аналогично можно найти выборочное уравнение прямой линии регрессии X на Y:

$$\bar{x}_y = \rho_{xy} y + c,$$

где ρ_{xy} – выборочный коэффициент регрессии X на Y. Повторяя аналогичные рассуждения, получаем коэффициенты

$$\rho_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2},\ c = \frac{\sum x \sum y^2 - \sum y \sum xy}{n\sum y^2 - (\sum y)^2}.$$

Уравнение прямой линии регрессии

Поскольку

$$\bar{x}=\frac{\sum x}{n},\ \bar{y}=\frac{\sum y}{n},\ \overline{xy}=\frac{\sum xy}{n},\ \sigma_x^2=\frac{\sum x^2}{n}-\left(\frac{\sum x}{n}\right)^2,\ \sigma_y^2=\frac{\sum y^2}{n}-\left(\frac{\sum y}{n}\right)^2,$$

тогда

$$\rho_{yx}=\frac{n\sum xy-\sum x\sum y}{n\sum x^2-(\sum x)^2}=\frac{\frac{\sum xy}{n}-\frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum x^2}{n}-\left(\frac{\sum x}{n}\right)^2}=\frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_x^2}=\frac{\sigma_y}{\sigma_x}\frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_x\sigma_y}=\frac{\sigma_y}{\sigma_x}\rho_r,$$

где коэффициент корреляции

$$\rho_r=\frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_x\sigma_y}$$

характеризует силу линейной связи между случайными величинами.

Уравнение прямой линии регрессии

Преобразуем

$$b=\frac{\sum x^2\sum y-\sum x\sum xy}{n\sum x^2-(\sum x)^2}=\frac{\frac{\sum x^2}{n}\frac{\sum y}{n}-\frac{\sum x}{n}\frac{\sum xy}{n}}{\frac{\sum x^2}{n}-\left(\frac{\sum x}{n}\right)^2}=\frac{\overline{x^2}\,\overline{y}-\overline{x}\overline{xy}}{\overline{x^2}-(\overline{x})^2}=$$

$$=\frac{\overline{x^2}\,\overline{y}-(\overline{x})^2\overline{y}+(\overline{x})^2\overline{y}-\overline{x}\overline{xy}}{\overline{x^2}-(\overline{x})^2}=\overline{y}\frac{\overline{x^2}-(\overline{x})^2}{\overline{x^2}-(\overline{x})^2}-\overline{x}\frac{\overline{xy}-\overline{x}\overline{y}}{\overline{x^2}-(\overline{x})^2}=\overline{y}-\overline{x}\frac{\sigma_y}{\sigma_x}\frac{\overline{xy}-\overline{x}\overline{y}}{\sigma_x\sigma_y}=\overline{y}-\overline{x}\frac{\sigma_y}{\sigma_x}\rho_r$$

т. е.

$$b=\overline{y}-\overline{x}\frac{\sigma_y}{\sigma_x}\rho_r\,.$$

Уравнение прямой линии регрессии

Подставив ρ_{yx} и b в уравнение $y_x = \rho_{yx} x + b$ получим

$$y_x = \frac{\sigma_y}{\sigma_x}\rho_r x + \bar{y} - \bar{x}\frac{\sigma_y}{\sigma_x}\rho_r$$

которое задает уравнение прямой линии регрессии Y на X в виде прямой с угловым коэффициентом, проходящей через заданную точку $(\bar{x};\bar{y})$:

$$y_x - \bar{y} = \rho_r \frac{\sigma_y}{\sigma_x}(x - \bar{x}).$$

Уравнение прямой линии регрессии

Аналогично уравнение прямой линии регрессии X на Y можно записать в виде:

$$x_y - \bar{x} = \rho_r \frac{\sigma_x}{\sigma_y}(y - \bar{y}) \cdot$$

Коэффициент корреляции ρ_r, входящий в уравнения прямых линий регрессии обладает важными свойствами

Свойства коэффициента корреляции

1. Коэффициент корреляции по модулю не превышает единицы, т. е. $|\rho_r| \le 1$.
2. Если случайные величины X и Y линейно независимы, то $\rho_r = 0$.
3. Если величины X и Y связаны линейной функциональной зависимостью, то $|\rho_r| = 1$.

Таким образом, по значению коэффициента корреляции оценивают, насколько тесная линейная взаимосвязь рассматриваемых случайных величин. Чем ближе $|\rho_r|$ к единице, тем теснее линейная корреляционная зависимость между X и Y, и наоборот, чем ближе $|\rho_r|$ к нулю, тем слабее линейная взаимосвязь.

величина коэффициента корреляции по абсолютной величине	название связи
$\|\rho_r\| = 0$	нет связи
$0 < \|\rho_r\| < \|0{,}3\|$	слабая
$\|0{,}3\| \le \|\rho_r\| < \|0{,}7\|$	средняя
$\|0{,}7\| \le \|\rho_r\| < \|1\|$	сильная
$\|\rho_r\| = 1$	функциональная

Уравнение прямой линии регрессии

Пример

Данные наблюдений системы случайных величин представлены в таблице.

n_{ij}		Y				
		11	13	15	17	19
X	25	4	3	1		
	35	3	5	2	2	
	45	1	4	10	4	
	55		3	4	5	2
	65			1	3	3

1. Установить силу линейной зависимости между значениями случайных величин.

2. Найти уравнения прямых линий регрессий и построить соответствующие прямые.

Уравнение прямой линии регрессии

1. Составим вспомогательную таблицу, найдем суммы частот по строкам и столбцам

n_{ij}		Y					
		11	13	15	17	19	∑
X	25	4	3	1			8
	35	3	5	2	2		12
	45	1	4	10	4		19
	55		3	4	5	2	14
	65			1	3	3	7
∑		8	15	18	14	5	60

Из таблицы вычислим числовые характеристики:

$$\overline{x} = \frac{\sum x_i n_i}{n} = \frac{25 \cdot 8 + 35 \cdot 12 + 45 \cdot 19 + 55 \cdot 14 + 65 \cdot 7}{60} = \frac{2700}{60} = 45,$$

$$\overline{x^2} = \frac{\sum x_i^2 n_i}{n} = \frac{25^2 \cdot 8 + 35^2 \cdot 12 + 45^2 \cdot 19 + 55^2 \cdot 14 + 65^2 \cdot 7}{60} = \frac{130100}{60} \approx 2168{,}33,$$

Уравнение прямой линии регрессии

n_{ij}		Y					
		11	13	15	17	19	∑
X	25	4	3	1			8
	35	3	5	2	2		12
	45	1	4	10	4		19
	55		3	4	5	2	14
	65			1	3	3	7
∑		8	15	18	14	5	60

$$\overline{y} = \frac{\sum y_j n_j}{n} = \frac{11 \cdot 8 + 13 \cdot 15 + 15 \cdot 18 + 17 \cdot 14 + 19 \cdot 5}{60} = \frac{886}{60} \approx 14{,}77,$$

$$\overline{y^2} = \frac{\sum y_j^2 n_j}{n} = \frac{11^2 \cdot 8 + 13^2 \cdot 15 + 15^2 \cdot 18 + 17^2 \cdot 14 + 19^2 \cdot 5}{60} = \frac{13404}{60} = 223{,}4,$$

$$\overline{xy} = \frac{\sum x_i y_j n_{ij}}{n} = \frac{11 \cdot 25 \cdot 4 + 11 \cdot 35 \cdot 3 + 11 \cdot 45 \cdot 1 + 13 \cdot 25 \cdot 3 + \ldots}{60} = \frac{40970}{60} \approx 682{,}83,$$

Уравнение прямой линии регрессии

$$\sigma_x^2 = \left(\overline{x^2}\right) - \left(\overline{x}\right)^2 = 2168{,}33 - (45)^2 = 143{,}33 \Rightarrow \sigma_x = \sqrt{143{,}33} \approx 11{,}97,$$

$$\sigma_y^2 = \left(\overline{y^2}\right) - \left(\overline{y}\right)^2 = 223{,}4 - (14{,}77)^2 = 5{,}25 \quad \sigma_y = \sqrt{5{,}25} \approx 2{,}29.$$

Итак, коэффициент корреляции $\rho_r = \dfrac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y} = \dfrac{682{,}83 - 45 \cdot 14{,}77}{11{,}97 \cdot 2{,}29} \approx 0{,}66.$

Из полученного значения коэффициента корреляции очевидна средняя линейная связь.

величина коэффициента корреляции по абсолютной величине	название связи
$\lvert\rho_r\rvert = 0$	нет связи
$0 < \lvert\rho_r\rvert < \lvert 0{,}3\rvert$	слабая
$\lvert 0{,}3\rvert \le \lvert\rho_r\rvert < \lvert 0{,}7\rvert$	средняя
$\lvert 0{,}7\rvert \le \lvert\rho_r\rvert < \lvert 1\rvert$	сильная
$\lvert\rho_r\rvert = 1$	функциональная

Уравнение прямой линии регрессии

Найдем уравнения регрессий:

- X на Y: $y_x - \bar{y} = \rho_r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$, тогда $y_x - 14{,}77 = 0{,}66\frac{2{,}29}{11{,}97}(x - 45)$, из

чего получим уравнение $y = 0{,}13x + 9{,}09$;

- Y на X: $x_x - \bar{x} = \rho_r \frac{\sigma_x}{\sigma_y}(y - \bar{y})$, тогда $x_x - 45 = 0{,}66\frac{11{,}97}{2{,}29}(y - 14{,}77)$, из

чего получим уравнение $x = 3{,}45y - 5{,}95$.