ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

УПРАВЛЕНИЕ ДИСТАНЦИОННОГО ОБУЧЕНИЯ И ПОВЫШЕНИЯ КВАЛИФИКАЦИИ

Кафедра «Прикладная математика»

# МЕТОДИЧЕСКИЕ УКАЗАНИЯ
## к проведению практических занятий по дисциплине

# «Математика»

Авторы

Рябых Г.Ю.,

Фролова Н.В.

Ростов-на-Дону, 2016

## Аннотация

Методические указания к практическим занятиям по линейной корреляции: методические указания для студентов всех направлений и специальностей.

Методические указания включают краткое описание методов нахождения числовых характеристик случайных величин, а также нахождение уравнения линейной регрессии и коэффициента корреляции. Рассмотрены примеры.

## Авторы

к.ф.-м.н.,
профессор каф. «Прикладная математика»
Рябых Г.Ю.

ст.преподаватель каф. «Математика»
Фролова Н.В.

# Оглавление

# ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СИСТЕМЫ ДВУХ СЛУЧАЙНЫХ ВЕЛИЧИН. КОРРЕЛЯЦИОННЫЙ МОМЕНТ. КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Математические ожидания и дисперсии составляющих системы двух случайных величин служат для описания этой системы. Кроме них используют и другие характеристики – это корреляционный момент $m_{xy}$ и коэффициент корреляции $r_{xy}$.

**Определение 1**: Корреляционным моментом $m_{xy}$ случайных величин x и y называют математическое ожидание произведения отклонений этих величин:

$$m_{xy} = M\{[X - M(x)] \cdot [Y - M(y)]\} \qquad (1)$$

(Это второй смешанный центральный момент, момент связи).

При вычислении корреляционного момента дискретных величин используют формулу:

$$m_{xy} = \sum_{i=1}^{n} \sum_{j=1}^{m} [x_i - M(x)] \cdot [y_j - M(y)] \cdot p(x_j, y_j) \qquad (2)$$

а для её прерывных величин – формулу:

$$m_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x - M(x)][y - M(y)] f(x, y) dxdy \qquad (3)$$

Корреляционный момент служит для характеристики связи между величинами x и y.

**Теорема**: Корреляционный момент двух независимых случайных величин x и y равен нулю.

**Доказательство**: По условию случайные величины x и y независимы, значит, независимы и их отклонения x-M(x) и y-M(y). На основе свойств математического ожидания имеем:

$$m_{xy} = M\{[X - M(x)] \cdot [Y - M(y)]\} = M[x - M(x)] M[y - M(y)] \qquad (4)$$

Известно, что математическое ожидание отклонения равно нулю и, следовательно, корреляционный момент $m_{xy} = 0$.

Если корреляционный момент отличен от нуля, то это есть признак наличия связи между этими двумя случайными величинами. Эта связь не только в их зависимости, но и в рассеивании. Если одна из величин мало отличается от своего математического ожидания (почти не случайна), то $m_{xy}$ будет мал, какой бы тесной зависимостью ни были связаны величины x и y.

Очевидно, что размерность $m_{xy}$ равна произведению размерностей величин x и y, и поэтому величина корреляционного момента может иметь различные значения для одних и тех же двух величин в зависимости от того, в каких единицах они были измерены. Эта особенность корреляционного момента является недостатком этой числовой характеристики. Для его устранения вводят ещё одну характеристику – коэффициент корреляции.

**Определение** 2: Коэффициентом корреляции $r_{xy}$ случайных величин x и y называют отношение корреляционного момента к произведению средних квадратических отклонений этих величин:

$$r_{xy} = \frac{\mu xy}{\delta x \cdot \delta y} \qquad (5)$$

δx и δy имеют размерность величин x и y соответственно. Значит, $r_{xy}$ – безразмерная величина, в этом её преимущество перед корреляционным моментом. Очевидно, что коэффициент корреляции $r_{xy}$ обращается в нуль одновременно с корреляционным моментом $m_{xy}$. Величины, для которых $r_{xy}=0$ называют некоррелированными (не связанными). Коэффициент корреляции характеризует не всякую зависимость, а только так называемую линейную зависимость. Линейная вероятная зависимость двух случайных величин x и y заключается в том, что при возрастании одной величины другая имеет тенденцию возрастать (или убывать) по линейному закону. Известно, что если две случайные величины x и y связаны линейной функциональной зависимостью:

$$Y = aX + b,$$

то $r_{xy}=\pm i$; в общем случае, когда величины X и Y связаны произвольной вероятностной зависимостью, коэффициент корреляции имеет значение в пределах $-i < r_{xy} < 1$, для доказательства последнего утверждения рассмотрим две теоремы:

**Теорема 1**: Абсолютная величина корреляционного момента двух случайных величин X и Y не превышает среднего геометрического их дисперсией:

$$|\mu_{xy}| \leq \sqrt{Dx * Dy} \qquad (6)$$

**Доказательство:** Рассмотрим случайную величину $z_{1=\sigma_y X - \sigma_x Y}$ и найдём её дисперсию $D(z_1) = M[z_1 - M(z_1)]$, произведя необходимые операции, получим:

$D(z_1) = 2\sigma_x^2 * \sigma_y^2 - 2\sigma_x * \sigma_y * \mu_{xy}$.

Из того, что любая дисперсия неотрицательна следует

$\mu_{xy} \leq \sigma_x * \sigma_y$

Рассмотрим случайную величину $z_2 = \sigma_y X + \sigma_x Y$, аналогично рассуждая, получим $\mu_{xy} \geq -\sigma_x * \sigma_y (**)$ объединяя неравенства (*) и (**), получим $-\sigma_x\sigma_y \leq \mu_{xy} \leq \sigma_x\sigma_y\ (***)$ или

$|\mu_{xy}| \leq \sigma_x * \sigma_y$, итак

$\mu_{xy} \leq \sqrt{Dx * Dy}$

**Теорема 2:** Абсолютная величина коэффициента корреляции не превышает единицы

$|r_{xy}| \leq 1$

**Доказательство:** разделим обе части двойного неравенства (***) на произведение положительных чисел $\sigma_x * \sigma_y : -1 \leq r_{xy} \leq 1$ или $|r_{xy}| \leq 1$. При $r_{xy} > 0$ говорят о положительной корреляции величин X и Y, что означает, что при возрастании одной из них другая имеет тенденцию в среднем возрастать; при $r_{xy} < 0$, говорят об отрицательной корреляции, что означает, что при возрастании одной из случайных величин другая имеет тенденцию в среднем убывать.

Например:

1. $\rho$- вес осколка снаряда, ℓ - длина того же осколка. Величина ρ и ℓ имеют ярко выраженную положительную корреляцию, так более тяжёлые осколки имеют в среднем большую длину.

2. Производится два выстрела по некоторой цели; точка попадания первого выстрела регистрируется, и в прицел вводится поправка, учитывающая размеры промаха предыдущего выстре-

ла. Координаты точек попадания первого и второго выстрела связаны отрицательной корреляцией.

## ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ. НОРМАЛЬНАЯ КОРРЕЛЯЦИЯ

Пусть (X,Y) – двумерная случайная величина. Если обе функции регрессии Y на X это M(Y/x)=f(x) и X на Y это M(X/y)=φ(y) линейные, то говорят, что X и Y связаны линейной корреляционной зависимостью. Очевидно, что графики линейных функций регрессии – прямые линии.

Теорема: \‘’Если двумерная случайная величина (X,Y) распределена нормально, то X и Y связаны линейной корреляционной зависимостью”.

Двумерная плотность вероятности

$$f(x,y) = \frac{1}{2\pi * \sigma_x * \sigma_y * \sqrt{1-r^2}} * e^{-(U^2+V^2-2rUV)/(2(1-r^2))} \quad (1)$$

где $U = \frac{x-a1}{\sigma_x}, V = \frac{y-a2}{\sigma_y}$ (2)

Плотность вероятности составляющей x

$$f_1(x) = \frac{1}{\sigma_x\sqrt{2\pi}} * e^{-\frac{U^2}{a}} \quad (3)$$

Найдём функцию регрессии M(Y/x), т.е. условное математическое ожидание величины Y при условии, что x приняла одно из возможных значений x, для этого сначала найдём условный закон распределения случайной величины Y при X=x:

$$\psi(y|x) = \frac{1}{\sigma_y\sqrt{1-r^2 * \sqrt{2\pi}}} * e^{-\frac{\{y-[a2-r\frac{\sigma_y}{\sigma_x}(x-a1]\}^2}{2[\sigma_y^2(1-r^2)]}}$$

Это условное распределение нормально, имеет математическое ожидание (функцию регрессии Y на x) $M(Y|x) = a2 + r\frac{\sigma_y}{\sigma_x}(y - a2)$ и дисперсию $D = \sigma_y^2 * (1 - r^2)$.

Аналогично получается функция регрессии X на y

$$M(X|y) = a1 + r\frac{\sigma_x}{\sigma_y}(y - a2)$$

Так как эти функции линейны, то и корреляция между величинами X и Y линейная, что и требовалось доказать.

На практике, во многих задачах приходится установить и оценить зависимость изучаемой случайной величины Y от другой случайной величины X.

Зависимости эти бывают функциональными, что реализуется крайне редко, так как на обе величины действуют многие случайные факторы, статистическими, когда изменение одной из этих величин влечёт изменение распределения другой, и, наконец, корреляционными, это такие статистические зависимости, когда при изменении величин изменяется среднее значение другой.

Например, Y - урожай зерна, X–количество удобрений. При условии, что площади участков одинаковы, удобрений внесены равные количества, урожай снимают разный, т.е. величина Yне является функцией величины X, но средний урожай является функцией от количества удобрений, т.е. Yсвязана с X корреляционной зависимостью.

При обработке данных применяют условные средние – это среднее арифметическое наблюдавшихся значений случайной величины, соответствующих определенному значению другой случайной величины.

Условные средние являются оценками условных математических ожиданий, и они в свою очередь являются функциями. Если имеем M(Y/x), то условное среднее $\overline{y_x}$ является функцией от X, обозначается $\overline{y}_x = f^*(x)$ и это уравнение называют выборочным уравнением регрессии Y на X, функция$f^*(x)$называется выборочной регрессией Y на X, а её график – выборочной линией регрессии.

Уравнение $\overline{x_y} = \varphi^*(y)$ называют выборочным уравнением регрессии X на Y, $\overline{x_y}$–условное среднее, $\varphi^*(y)$–выборочная регрессии X на Y, а её график – выборочной линией регрессии X на Y.

Нахождение параметров функций $f^*(x)$ и $\varphi^*(y)$ по данным практических наблюдений, оценивание силы связей между величинами X и Y, установление их коррелированности рассмотрим ниже.

## ОПРЕДЕЛЕНИЕ КОРРЕЛЯЦИОННЫХ ХАРАКТЕРИСТИК

Если обе линии регрессии Y на X и X на Y – прямые, то корреляцию называют ***линейной***.

Выборочное уравнение прямой линии регрессии Y на X имеет вид

$$\overline{y_x} - \bar{y} = r_{\text{в}} \frac{\delta_y}{\delta_x}(x - \bar{x}) \qquad (1),$$

где $\overline{y_x}$ - условная средняя, т.е. среднее арифметическое наблюдавшихся значений Y, соответствующих X=x; $\bar{x}$ и $\bar{y}$ – выборочные средние признаков X и Y, т.е. средние арифметические значений признаков выборочных совокупностей X и Y; $\delta_x$ и $\delta_y$ – выборочные средние квадратические отклонения; $r_{\text{в}}$ - выборочный коэффициент корреляции.

Аналогично выборочное уравнение прямой линии регрессии X на Y имеет вид

$$\overline{x_y} - x = r_{\text{в}} \frac{\delta_x}{\delta_y}(y - \bar{y}) \qquad (2).$$

Если данные наблюдений над признаками X и Y заданы в виде корреляционной таблицы с равностоящими вариантами, то можно значительно упростить расчет, перейдя к условным вариантам:

$u_i = \frac{x_i - c_1}{h_1}$; $\quad v_j = \frac{y_j - c_2}{h_2}$, где

$c_1$ - варианта признака X, имеющая наибольшую частоту;

$h_1$ – шаг (разность между двумя соседними вариантами X);

$c_2$ - варианта признака Y, имеющая наибольшую частоту;

$h_2$ – шаг (разность между двумя соседними вариантами Y);

В этом случае выборочный коэффициент корреляции вычисляют по формуле:

$$r_{\text{в}} = \frac{\Sigma n_{uv} uv - n\bar{u}\bar{v}}{n\delta_u \delta_v}$$

Причем, величины $\bar{u}$, $\bar{v}$, $\delta_u$, $\delta_v$ могут быть найдены по формулам:

$$\bar{u} = \frac{\Sigma n_u u}{n}; \quad v = \frac{\Sigma n_v v}{n}; \quad \delta_u = \sqrt{\bar{u}^2 - (\bar{u})^2}; \delta_v = \sqrt{\bar{v}^2 - (\bar{v})^2}.$$

Зная эти величины, можно найти величины, входящие в уравнения регрессии (1) и (2) по формулам:

$$\bar{x} = \bar{u}h_1 + c_1; \quad \bar{y} = \bar{v}h_2 + c_2; \quad \delta_x = \delta_u h_1; \quad \delta_y = \delta_v h_2.$$

Остается указать способ вычисления $\Sigma n_{uv} uv$, где $n_{uv}$ - частота пары условных вариант (*u, v*), для чего решим задачу.

Задача 1. По данным корреляционной таблицы 1 вычислить $\Sigma n_{uv} uv$.

Таблица 1

| Y | X | | | | | | $n_y$ |
|---|---|---|---|---|---|---|---|
| | 10 | 20 | 30 | 40 | 50 | 60 | |
| 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 |

Ответ: $\Sigma n_{uv} uv = 169$.

Отметим, что выборочный коэффициент корреляции служит для оценки силы линейной корреляционной связи: чем ближе $|r_в|$ к единице, тем связь сильнее, чем ближе $|r_в|$ к нулю, тем связь слабее.

Задача 2. Вычислить выборочный коэффициент корреляции по данным таблицы 1.

Решение. Ищем коэффициент корреляции по формуле:

$$r_в = \frac{\Sigma n_{uv} uv - n\bar{u}\bar{v}}{n\,\delta_u \delta_v}.$$

Перейдя к условным вариантам, получим корреляционную таблицу 2.

Таблица 2

| *v* | u | | | | | | $n_v$ |
|---|---|---|---|---|---|---|---|
| | -3 | -2 | -1 | 0 | 1 | 2 | |
| -2 | 5 | 7 | - | - | - | - | 12 |
| -1 | - | 20 | 23 | - | - | - | 43 |
| 0 | - | - | 30 | 47 | 2 | - | 79 |
| 1 | - | - | 10 | 11 | 20 | 6 | 47 |
| 2 | - | - | - | 9 | 7 | 3 | 19 |
| $n_u$ | 5 | 27 | 63 | 67 | 29 | 9 | n=200 |

Далее найдем $\bar{u}$ и $\bar{v}$:

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

$$\bar{v} = \frac{\Sigma n_v v}{n} = \frac{12\cdot(-2)+43\cdot(-1)+47\cdot 1+19\cdot 2}{200} = -0{,}09.$$

Для нахождения $\delta_u$ и $\delta_v$ по формулам $\delta_u = \sqrt{\bar{u}^2 - (\bar{u})^2}$ и $\delta_v = \sqrt{\bar{v}^2 - (\bar{v})^2}$ вычислим вспомогательные величины:

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

$$\bar{v}^2 = \frac{\Sigma n_v v^2}{n} = \frac{12\cdot 4+43\cdot 1+47\cdot 1+19\cdot 4}{200} = 1{,}07.$$

Далее найдем $\delta_u$ и $\delta_v$:

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

$$\delta_v = \sqrt{\bar{v}^2 - (\bar{v})^2} = \sqrt{1{,}07 - (0{,}09)^2} = 1{,}209.$$

Так как ранее уже вычислена сумма $\sum n_{uv}\, uv = 169$, найдём искомый выборочный коэффициент корреляции: $r_b = \frac{\sum n_{uv}\ uv - n\bar{u}\bar{v}}{n\sigma_u\sigma_v} = \frac{169-200(-0{,}425)\cdot 0{,}09}{200\cdot 1{,}106\cdot 1{,}209} = 0{,}603$.

Ответ: $r_b = 0{,}603$.

Теперь, когда известно, как вычисляют $r_b$, решим задачу на отыскание уравнения прямой линии регрессии.

<u>Задача 3.</u> Найти выборочное уравнение прямой линии регрессии $y$ на $x$ по данным корреляции таблицы I.

<u>Решение.</u> Напишем искомое уравнение в общем виде:

$$\overline{y_x} - \bar{y} = r_b\, \frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \text{(I)}$$

Коэффициент корреляции уже вычислен в предыдущей задаче,

$r_b = 0{,}603$.

Остаётся найти $\bar{x}$, $\bar{y}$, $\sigma_x$ и $\sigma_y$:

$\bar{x} = \bar{u}h_1 + c_1 = -0{,}425*10 + 40 = 35{,}75;$

$\bar{y} = \bar{v}h_2 + c_2 = 0{,}09*10 + 35 = 35{,}9;$

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

$\sigma_y = \sigma_v h_2 = 1{,}209*10 = 12{,}09.$

Подставим введённые величины в (I), получим искомое уравнение $\bar{y}_x - 35{,}9 = 0{,}603\,\frac{12{,}09}{11{,}06}(x - 35{,}7)$, или окончательно $\bar{y}_x = 0{,}659x + 12{,}34$.

Сравним условные средние, вычисленные: а) по этому уравнению; б) по данным корреляционной таблицы I. Например, при $x = 30$:

а) $\bar{y}_{30} = 0{,}659*30 + 12{,}34 = 32{,}11;$

б) $\bar{y}_{30} = \frac{23\cdot 25+30\cdot 25+10\cdot 45}{63} = 32{,}94.$

Как видим, согласование расчётного и наблюдаемого условных средних – удовлетворительное

Задача 4. Найти выборочное уравнение прямой линии регрессии $y$ на $x$ по данным корреляционной таблицы 2.

Таблица 2

| $y \backslash x$ | 5 | 15 | 25 | 35 | 45 | 55 | 65 | $n_y$ |
|---|---|---|---|---|---|---|---|---|
| 4 | 2 | -- | 2 | -- | -- | -- | -- | 4 |
| 8 | -- | 1 | 4 | -- | -- | -- | -- | 5 |
| 12 | -- | 4 | 3 | 10 | -- | -- | -- | 17 |
| 16 | -- | 2 | -- | 2 | 3 | 6 | -- | 13 |
| 20 | -- | -- | -- | -- | 5 | 4 | -- | 9 |
| 24 | -- | -- | -- | -- | -- | 1 | 1 | 2 |
| $n_x$ | 2 | 7 | 9 | 12 | 8 | 11 | 1 | n = 50 |

Ответ: $\overline{y_x} = 0{,}25\,x + 4{,}97$.

Задача 5. Найти выборочное уравнение прямой линии регрессии $y$ на $x$ по данным корреляционной таблицы 3:

Таблица 3

| $y \backslash x$ | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | $n_y$ |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 2 | 1 | -- | -- | -- | -- | -- | - | 9 |
| 120 | 3 | 4 | 3 | -- | -- | -- | -- | -- | 10 |
| 140 | -- | -- | 5 | 10 | 8 | - | -- | -- | 23 |
| 160 | -- | -- | -- | 1 | -- | 6 | 1 | 1 | 9 |
| 180 | -- | -- | -- | -- | -- | -- | 4 | 1 | 5 |
| $n_x$ | 5 | 5 | 8 | 11 | 8 | 6 | 5 | 2 | n = 50 |

Ответ: $\bar{x}_y = 0{,}12\,y + 3{,}7$.

Задача 6. Найти выборочное уравнение прямой линии регрессии $y$ на $x$ по данным корреляционной таблицы 4:

Таблица 4

| $y$ \ $x$ | 18 | 23 | 28 | 33 | 38 | 43 | 48 | $n_y$ |
|---|---|---|---|---|---|---|---|---|
| 125 | -- | 1 | -- | -- | -- | -- | -- | 1 |
| 150 | 1 | 2 | 5 | -- | -- | -- | -- | 8 |
| 175 | -- | 3 | 2 | 12 | -- | -- | -- | 17 |
| 200 | -- | -- | 1 | 8 | 7 | -- | -- | 16 |
| 225 | -- | -- | -- | -- | 3 | 3 | -- | 6 |
| 250 | -- | -- | -- | -- | -- | 1 | 1 | 2 |
| $n_x$ | 1 | 6 | 8 | 20 | 10 | 4 | 1 | n = 50 |

Ответ: $\bar{y}_x = 4x + 57{,}8$.