Федеральное государственное бюджетное образовательное учреждение

высшего образования
«Ставропольский государственный аграрный университет»

Кафедра Математика

Составитель: доцент Литвин Д.Б.

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ РГР №8

по дисциплине

МАТЕМАТИКА

наименование дисциплины

21.03.02 Землеустройство

направление подготовки

Городской кадастр

профиль(и) подготовки

Бакалавр

Квалификация (степень) выпускника

Ставрополь, 2019

КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

1.1. Линейная корреляция

Если в результате осуществления некоторого эксперимента наблюдаются две величины X и Y, то *выборочный корреляционный момент (ковариация)* $\mu^*_{x,y}$ величин X и Y определяется формулой:

$$\mu^*_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x}\cdot\overline{y} \qquad (1)$$

где $(x_1, y_1),\ldots,(x_n, y_n)$ — n пар наблюденных значений, полученных в n независимых повторениях эксперимента, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Выборочный коэффициент корреляции r^*_B:

$$r^*_B = \frac{\mu^*_{x,y}}{\sigma^*_x \cdot \sigma^*_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \qquad (2)$$

где σ^*_x и σ^*_y - выборочные СКО

$$\sigma^*_x = \sqrt{D^*(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad \sigma^*_y = \sqrt{D^*(Y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Выборочные коэффициенты регрессий

$$Y \text{ на } X: \qquad \rho^*_{y/x} = r^*_B\frac{\sigma^*_y}{\sigma^*_x} = \frac{\mu^*_{x,y}}{(\sigma^*_x)^2}; \qquad (3)$$

$$X \text{ на } Y: \qquad \rho^*_{x/y} = r^*_B\frac{\sigma^*_x}{\sigma^*_y} = \frac{\mu^*_{x,y}}{(\sigma^*_y)^2}. \qquad (4)$$

Выборочные уравнения линейных регрессий (среднеквадратических)

$$Y \text{ на } X: \quad y - \overline{y} = \rho^*_{y/x}(x - \overline{x}) \quad \text{или} \quad \frac{y - \overline{y}}{\sigma^*_y} = r^*_B\frac{x - \overline{x}}{\sigma^*_x}, \qquad (5)$$

$$X \text{ на } Y: \quad \frac{x - \overline{x}}{\sigma^*_x} = r^*_B\frac{y - \overline{y}}{\sigma^*_y} \quad \text{или} \quad \frac{y - \overline{y}}{\sigma^*_y} = \frac{1}{r^*_B}\frac{x - \overline{x}}{\sigma^*_x}. \qquad (6)$$

Пример 3.1. Для выборки двумерной случайной величины

Таблица 3.1

i	1	2	3	4	5	6	7	8	9	10
x_i	1,2	1,5	1,8	2,1	2,3	3	3,6	4,2	5,7	6,3
y_i	4,6	5,8	7,3	10,4	12,30	14,4	14,9	14,8	15,2	16,5

вычислить выборочные средние наблюдаемых признаков, выборочные средние квадратические отклонения, выборочный коэффициент корреляции и составить выборочные уравнения линейных регрессий. Представить корреляционное поле и линейные регрессии на графике.

Решение.

Удобно исходную таблицу дополнить строкой $x_i y_i$ *и столбцом* ср.знч.

Таблица 3.2

i	1	2	3	4	5	6	7	8	9	10	ср.знч
x_i	1,2	1,5	1,8	2,1	2,3	3	3,6	4,2	5,7	6,3	3,17
y_i	4,6	5,8	7,3	10,4	12,30	14,4	14,9	14,8	15,2	16,5	11,62
$x_i y_i$	5,52	8,7	13,14	21,84	28,29	43,2	53,64	62,16	86,64	103,95	42,71

Тогда искомые выборочные параметры распределения

$$\overline{x}_B = \frac{1,2+1,5+...+6,3}{10} = 3,17;\ \ \overline{y}_B = \frac{4,6+5,8+..+16,5}{10} = 11,62.$$

$$D_B(X) = 0,1(1,2^2+1,5^2+...+6,3^2) - 3,17^2 = 2,79;\ \ \sigma_X = \sqrt{2,79} = 1,67.$$

$$D_B(Y) = 0,1(4,6^2+5,8^2+...+16,5^2) - 11,62^2 = 16,90;\ \ \sigma_Y = \sqrt{16,90} = 4,11.$$

$$\frac{1}{10}\sum_{i=1}^{10} x_i y_i = \frac{1}{10}\left(1,2\cdot 4,6+1,5\cdot 5,8+...+6,3\cdot 16,5\right) = 42,71.$$

$$\mu^*_{x,y} = 42,71 - 3,17\cdot 11,62 = 5,87; \qquad r^*_B = \frac{42,71-3,17\cdot 11,62}{1,67\cdot 4,11} = 0,85.$$

Выборочное уравнение линейной регрессии Y/X имеет вид (5):

$$y - 11,62 = 0,85\cdot\frac{4,11}{1,67}(x-3,17) \ \textit{или}\ \ y = 2,1x\ +\ 4,95.$$

Выборочное уравнение линейной регрессии X/ Y имеет вид (6):

$$y - 11,62 = \frac{1}{0,85}\cdot\frac{4,11}{1,67}(x-3,17) \ \textit{или}\ \ y = 2,88x\ +\ 2,45.$$

Поскольку линии регрессии прямые, то строим их по двум крайним точкам

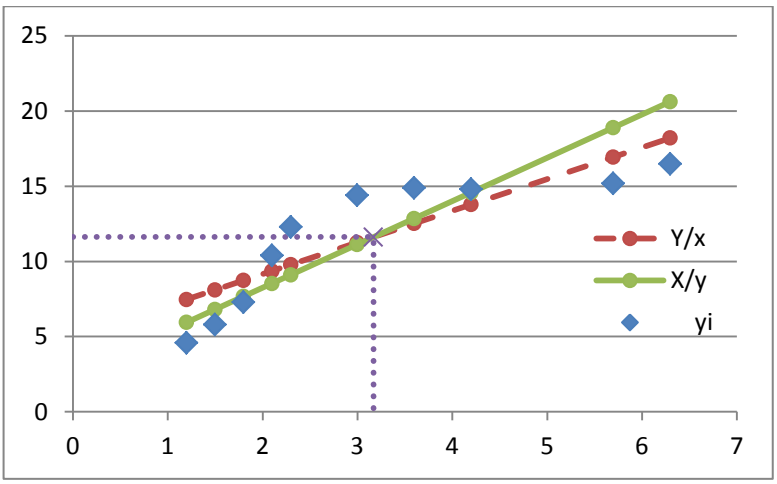


Рисунок 3.1 - Линейные регрессии

Для вычисления корреляционного момента (ковариации) и коэффициента корреляции в табличном процессоре ***Excel*** используются функции ***КОВАР(массив1;массив2)*** и ***КОРРЕЛ(массив1;массив2)*** соответственно.

1.2. Нелинейная корреляция

При совместном исследовании двух случайных величин по имеющейся выборке (x_1, y_2), $(x_2, y_2),\ldots,(x_k, y_k)$ возникает задача определения "наиболее подходящей", в том числе и нелинейной, зависимости между ними. Если вид функции $y = f(x,a,b,...)$ задан, то требуется найти такие значения коэффициентов a, b,..., при которых y_i "наименее" отличаются от $f(x_i)$. В качестве критерия оптимальности часто используют минимум суммы квадратов ошибок

$$\sum_{i=1}^{k}\left(y_i - f\left(x_i\right)\right)^2 \to \min. \qquad (7)$$

Коэффициенты a, b,... функции $y = f(x,a,b,...)$ и сама функция, обеспечивающие минимум критерия (7) называются оптимальными в смысле метода наименьших квадратов для этого класса функций. При этом, для другого класса функций, например $y = g(x,a,b,...)$, критерий (7) может принять значение еще более близкое к нулю. В этом случае функция $y = g(x,a,b,...)$, очевидно, лучше описывает статистическую взаимосвязь наблюдаемых признаков, нежели функция $y = f(x,a,b,...)$.

В качестве оценки тесноты нелинейной корреляционной связи используют коэффициент детерминации R^2 (корреляционное отношение),

который показывает долю объясненной изменением факторного признака дисперсии от общей дисперсии:

$$R^2 = \frac{\sigma_e^2}{\sigma^2} = 1 - \frac{\sigma_u^2}{\sigma^2}, \qquad (8)$$

где $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\Delta y_i^{\,2}$ - общая (total) дисперсия результативного признака y_i - относительно общей средней $\bar{y}$.

$\sigma_e^2 = \frac{1}{n}\sum_{i=1}^{n}\left(f\left(x_i\right) - \bar{y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}e_i^2$ - объясненная (межгрупповая, explained) дисперсия - дисперсия точек линии регрессии $f\left(x_i\right)$ относительно общей средней $\bar{y}$;

$\sigma_u^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(x_i\right)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}u_i^{\,2}$ - остаточная (внутригрупповая, unexplained) дисперсия - дисперсия признака y_i относительно модельной линии регрессии $f\left(x_i\right)$.

Смысл коэффициента детерминации поясняется на рисунке 3.1.

Здесь $y_i - \bar{y} = \Delta y_i = \left(y_i - y_x\right) + \left(y_x - \bar{y}\right) = u_i + e_i$.

Доказывается, что $\sum \Delta y_i = \sum u_i + \sum e_i$, т.к. $2\sum u_i e_i = 0$.

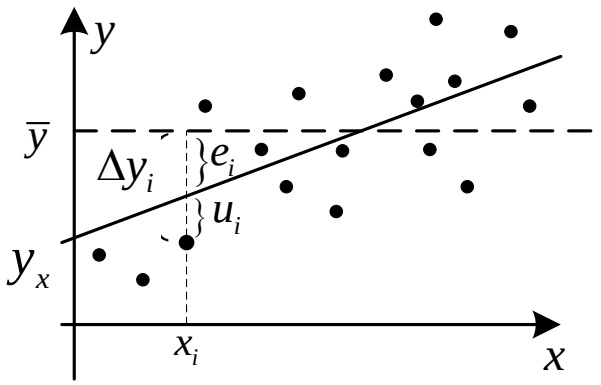


Рисунок 3.1 -
К пояснению суммы дисперсий

Поэтому имеет место *правило сложения дисперсий* - общая дисперсия σ^2 результативного признака равна сумме объясненной (межгрупповой) σ_e^2 и остаточной (необъясненной, внутригрупповой) σ_u^2 дисперсии:

$$\sigma^2 = \sigma_e^2 + \sigma_u^2 \qquad (9)$$

Из выражения (8) следует пределы изменения значений R^2:

$$0 \le R^2 \le 1.$$

Рассмотрим методику определения оптимальных по методу наименьших квадратов (МНК) параметров для некоторых классов функций (модельных регрессий).

а) Пусть модельной линией регрессии является линейная зависимость

$$y(x) = a_0 + a_1 x.$$

В качестве критерия тесноты корреляционной связи используем сумму квадратов отклонений значений признака y_i относительно линии регрессии $y(x_i)$

$$F(a_0, a_1) = \sum_{i=1}^{n} \left(y_i - (a_0 + a_1 x) \right)^2 \to \min.$$

Определим неизвестные параметры регрессии (a_0, a_1) из условия минимума $F(a_0, a_1)$: $F'_{a_0} = F'_{a_1} = 0$.

$$F'_{a_0} = 2\sum \left(y_i - a_0 - a_1 x \right) \cdot (-1) = 0; \qquad \sum \left(-y_i + a_0 + a_1 x \right) = 0;$$

$$\sum y_i = n \cdot a_0 + a_1 \sum x_i \ ; \qquad \overline{y} = a_0 + \overline{x} \cdot a_1.$$

$$F'_{a_1} = 2\sum \left(y_i - a_0 - a_1 x \right) \cdot (-x_i) = 0; \qquad \sum \left(-x_i y_i + a_0 x_i + a_1 x_i^2 \right) = 0;$$

$$\sum x_i y_i = a_1 \sum x_i^2 + a_0 \sum x_i ; \qquad \overline{xy} = \overline{x} \cdot a_0 + \overline{x^2} \cdot a_1.$$

$$\begin{cases} \overline{y} = a_0 + \overline{x} \cdot a_1 \\ \overline{xy} = \overline{x} \cdot a_0 + \overline{x^2} \cdot a_1 \end{cases} \quad \text{или} \quad \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix} = \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}. \tag{10}$$

.

б) Квадратичная зависимость $y(x) = a_0 + a_1 x + a_2 x^2$. Приравнивая частные производные суммы квадратов ошибок к нулю, по аналогии с пунктом а) получим:

$$\begin{pmatrix} \overline{y} \\ \overline{xy} \\ \overline{x^2 y} \end{pmatrix} = \begin{pmatrix} 1 & \overline{x} & \overline{x^2} \\ \overline{x} & \overline{x^2} & \overline{x^3} \\ \overline{x^2} & \overline{x^3} & \overline{x^4} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}. \tag{11}$$

в) Логарифмическая зависимость $y(x) = a_0 + a_1 \ln x$. Вводя новую переменную $z = \ln x$ приводим логарифмическую зависимость к линейной. Тогда на основании выражений (10) получим:

$$\begin{cases} \overline{y} = a_0 + \overline{z} \cdot a_1 \\ \overline{zy} = \overline{z} \cdot a_0 + \overline{z^2} \cdot a_1 \end{cases} \quad \text{или} \quad \begin{pmatrix} \overline{y} \\ \overline{zy} \end{pmatrix} = \begin{pmatrix} 1 & \overline{z} \\ \overline{z} & \overline{z^2} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}. \qquad (12)$$

Пример 3.2. Найти параметры и построить линии регрессий y от x для выборки из предыдущего примера 3.1 (табл.3.2) для случаев:

1) линейной зависимости $y(x) = a_0 + a_1 x$;

2) квадратичной зависимости $y(x) = a_0 + a_1 x + a_2 x^2$;

3) логарифмической зависимости $y(x) = a_0 + a_1 \ln x$.

Решение.

В соответствие с (10)- (12) дополним исходную таблицу необходимыми строками и последним столбцом - ср.знч. После заполнения таблица примет вид (вычисления проводились в MS Excel):

Таблица 3.3

i	1	2	3	4	5	6	7	8	9	10	ср.знач
x_i	1,2	1,5	1,8	2,1	2,3	3	3,6	4,2	5,7	6,3	3,17
y_i	4,6	5,8	7,3	10,4	12,30	14,4	14,9	14,8	15,2	16,5	11,62
xy	5,52	8,70	13,14	21,84	28,29	43,20	53,64	62,16	86,64	103,95	42,71
(x^2)y	6,62	13,05	23,65	45,86	65,07	129,60	193,10	261,07	493,85	654,89	188,68
x^2	1,44	2,25	3,24	4,41	5,29	9,00	12,96	17,64	32,49	39,69	12,84
x^3	1,73	3,38	5,83	9,26	12,17	27,00	46,66	74,09	185,19	250,05	61,53
x^4	2,07	5,06	10,50	19,45	27,98	81,00	167,96	311,17	1055,60	1575,30	325,61
z=lnx	0,18	0,41	0,59	0,74	0,83	1,10	1,28	1,44	1,74	1,84	1,01
z^2	0,03	0,16	0,35	0,55	0,69	1,21	1,64	2,06	3,03	3,39	1,31
zy	0,84	2,35	4,29	7,72	10,24	15,82	19,09	21,24	26,46	30,37	13,84
Yлин	7,48	8,11	8,74	9,37	9,79	11,26	12,52	13,79	16,94	18,20	11,62
Yкв	4,79	6,59	8,23	9,73	10,65	13,36	15,04	16,13	16,31	15,36	11,62
Yлог	5,56	7,18	8,51	9,63	10,30	12,23	13,56	14,68	16,91	17,63	11,62

Искомые векторы неизвестных параметров определим из систем линейных уравнений (10)- (12), например методом Крамера или матричным методом, с использованием последнего столбца таблицы 3.3.

В результате получим уравнения регрессий:

1) $\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1 & 3{,}17 \\ 3{,}17 & 12{,}84 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 11{,}62 \\ 42{,}71 \end{pmatrix} = \begin{pmatrix} 4{,}95 \\ 2{,}10 \end{pmatrix}$; линейной $y(x) = 4{,}95 + 2{,}10x$;

2) $\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 & 3,17 & 12,84 \\ 3,17 & 12,84 & 61,53 \\ 12,84 & 61,53 & 325,61 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 11,62 \\ 42,71 \\ 188,68 \end{pmatrix} = \begin{pmatrix} -3,84 \\ 8,17 \\ -0,81 \end{pmatrix}$;

квадратичной $y(x) = -3,84 + 8,17x - 0,81x^2$;

3) $\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1 & 1,01 \\ 1,01 & 1,31 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 11,62 \\ 13,84 \end{pmatrix} = \begin{pmatrix} 4,23 \\ 7,28 \end{pmatrix}$;

логарифмической $y(x) = 4,23 + 7,28\ln x$.

Построим графики всех полученных уравнений регрессии по точкам x_i (последние строки в таблице 3.3), как показано на рисунке 3.3

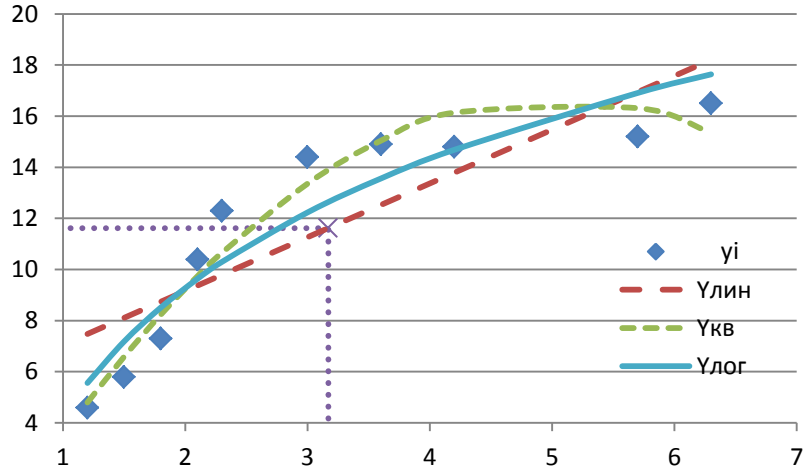


Рисунок 3.3 - Корреляционное поле и линии регрессий

Пример 3.3. Определите общую, объясненную и остаточную дисперсии для рассмотренных уравнений регрессий, а также коэффициенты детерминации.

Решение.

Вычислим общую, объясненную и остаточную дисперсии для рассмотренных уравнений регрессий с помощью MS Excel. Для этого составим следующую таблицу:

Таблица 3.4

i	1	2	3	4	5	6	7	8	9	10	ср.знач
y_i	4,6	5,8	7,3	10,4	12,30	14,4	14,9	14,8	15,2	16,5	11,62
$Yлин = y_{x1}$	7,48	8,11	8,74	9,37	9,79	11,26	12,52	13,79	16,94	18,20	11,62
$Yкв = y_{x2}$	4,79	6,59	8,23	9,73	10,65	13,36	15,04	16,13	16,31	15,36	11,62

i	1	2	3	4	5	6	7	8	9	10	ср.знач
Улог= y_{x3}	5,56	7,18	8,51	9,63	10,30	12,23	13,56	14,68	16,91	17,63	11,62
$\Delta y_i = y_i - \bar{y}$	-7,02	-5,82	-4,32	-1,22	0,68	2,78	3,28	3,18	3,58	4,88	0,00
$\Delta y_i^{\,2}$	49,280	33,872	18,662	1,488	0,462	7,728	10,758	10,112	12,816	23,814	16,90
$e_1 = y_{x1} - \bar{y}$	-4,14	-3,51	-2,88	-2,25	-1,83	-0,36	0,90	2,17	5,32	6,58	0,00
$e_2 = y_{x2} - \bar{y}$	-6,83	-5,03	-3,39	-1,89	-0,97	1,74	3,42	4,51	4,69	3,74	0,00
$e_3 = y_{x3} - \bar{y}$	-6,06	-4,44	-3,11	-1,99	-1,32	0,61	1,94	3,06	5,29	6,01	0,00
$e_1^{\,2}$	17,17	12,34	8,30	5,06	3,35	0,13	0,82	4,69	28,32	43,34	12,35
$e_2^{\,2}$	46,58	25,32	11,46	3,55	0,93	3,02	11,68	20,38	22,00	13,95	15,89
$e_3^{\,2}$	36,73	19,68	9,66	3,94	1,75	0,37	3,76	9,38	27,94	36,17	14,94
$u_1 = y_i - y_{x1}$	-2,88	-2,31	-1,44	1,03	2,51	3,14	2,38	1,01	-1,74	-1,70	0,00
$u_2 = y_i - y_{x2}$	-0,19	-0,79	-0,93	0,67	1,65	1,04	-0,14	-1,33	-1,11	1,14	0,00
$u_3 = y_i - y_{x3}$	-0,96	-1,38	-1,21	0,77	2,00	2,17	1,34	0,12	-1,71	-1,13	0,00
$u_1^{\,2}$	8,27	5,32	2,07	1,06	6,30	9,84	5,64	1,03	3,03	2,90	4,55
$u_2^{\,2}$	0,04	0,62	0,87	0,44	2,71	1,09	0,02	1,78	1,23	1,31	1,01
$u_3^{\,2}$	0,92	1,92	1,47	0,59	4,01	4,70	1,80	0,01	2,91	1,29	1,96

Легко проверить выполнение правила сложения дисперсий (9):

$$\sigma^2_{лин} = \sigma^2_{e1} + \sigma^2_{u1} = 12,35 + 4,55 = 16,90; \qquad \sigma^2_{кв} = \sigma^2_{e2} + \sigma^2_{u2} = 15,89 + 1,01 = 16,90;$$

$$\sigma^2_{лог} = \sigma^2_{e3} + \sigma^2_{u3} = 14,94 + 1,96 = 16,90.$$

Коэффициенты детерминации примут значения (8):

$$R^2_{лин} = \frac{\sigma^2_{e1}}{\sigma^2} = \frac{12,35}{16,90} = 0,73; \qquad R^2_{кв} = \frac{\sigma^2_{e2}}{\sigma^2} = \frac{15,89}{16,90} = 0,94;$$

$$R^2_{лог} = \frac{\sigma^2_{e3}}{\sigma^2} = \frac{14,94}{16,90} = 0,88, \quad \text{где } \sigma^2 = \sigma^2_{лин} = \sigma^2_{кв} = \sigma^2_{лог} = 16,90.$$

Максимальный коэффициент детерминации соответствует квадратичной регрессии, значит она и является наиболее удачной из рассмотренных.

2. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Под *статистической гипотезой* понимается всякое высказывание о генеральной совокупности (случайной величине X), проверяемое по выборочной совокупности (по результатам наблюдений).

Для проверки гипотезы определяют специальную величину (статистику) K с известным законом распределения, как функцию выборочных данных $K = K(x_1, x_2, ..., x_n)$. Величину K называют *статистическим критерием.* Полагая, что распределение признака *X* подчинено гипотезе H_0, рассчитывают область возможных значений (с заданной вероятностью) статистики *K*. Эта *область принятия основной гипотезы*. *Критической называют область* практически невозможных (с вероятностью α) значений статистики *K*, если верна гипотеза H_0.

По имеющимся наблюдениям рассчитывают выборочное значение *K*. Если оно принадлежит критической области, то основную гипотезу отвергают.

Решение о том, можно ли считать гипотезу H_0 верной для генеральной совокупности, принимается по выборочным данным, т.е. по ограниченному объему информации. Следовательно, это решение может быть ошибочным. При этом может иметь место ошибка двух родов (рисунок 4.1):

— *ошибка первого рода* совершается при отклонении верной гипотезы H_0 . Вероятность такой ошибки называют *уровнем значимости критерия* $\alpha = P(H_1 / H_0)$;

— *ошибка второго рода* совершается при принятии неверной гипотезы H_1. Вероятность ошибки второго рода обозначим как $\beta = P(H_0 / H_1)$.

Значение $1 - \beta = P(H_1 / H_1)$ - вероятность отвергнуть основную гипотезу когда она неверна, называют *мощностью критерия.*

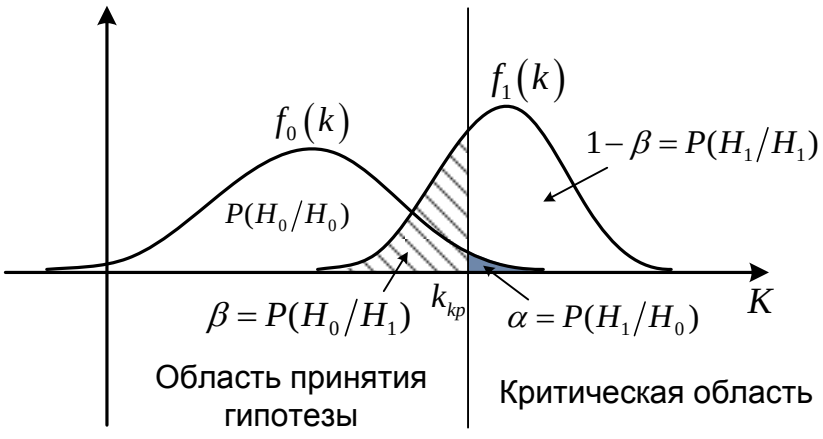
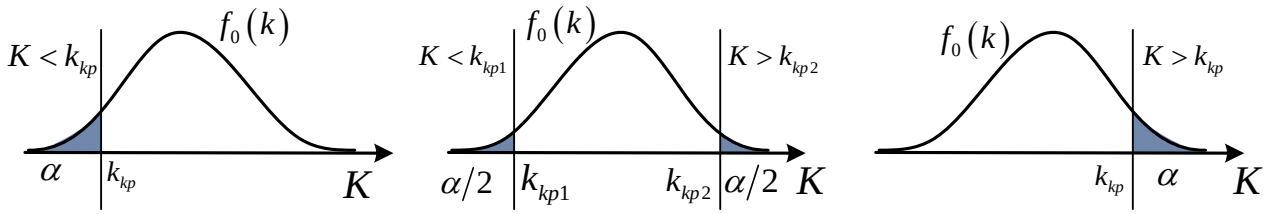


Рисунок 4.1 - К проверке гипотез

Критическую область следует выбирать из условия обеспечения максимума мощности критерия при заданном уровне значимости. Критическая область может быть *левосторонней,* если она задается неравенством ($K < k_{кр}$), *двусторонней* $\left(k_{кр1} > K\right) \cup \left(K > k_{кр2}\right)$ или *правосторонней* ($K > k_{кр}$) (рис. 4.2).



Левосторонняя Двусторонняя Правосторонняя

Рисунок 4.2 - Критические области

2.1. Проверка гипотез о числовых значениях

Проверка гипотезы о равенстве математического ожидания нормального распределения заданному значению

Проверка гипотезы о числовом значении МО ***при известной дисперсии.***

Предполагается, что $X = N(a, \sigma)$, причем значение математического ожидания *a* неизвестно, а числовое значение дисперсии σ^2 известно.

По известной выборочной средней $\overline{x}$ необходимо проверить при уровне значимости α нулевую гипотезу $H_0 : a = a_0$ о том, что математическое ожидание генеральной совокупности ***a*** равно предполагаемому значению a_0

Если дисперсия генеральной совокупности известна (или неизвестна, но выборка достаточно большая $n > 30$), то в качестве критерия *K* проверки нулевой гипотезы принимают *z-статистику* - нормально распределенную случайную величину с параметрами $N(0;1)$.

Вычислим наблюдаемое значение *z-критерия*:

$$Z_{набл} = \frac{(\overline{x} - a_0)\sqrt{n}}{\sigma} = N(0;1).$$

Критическую точку $z_{кр}$ определим в зависимости от альтернативной гипотезы (типа критической области) с использованием функции Лапласа (Приложение 2).

Правило 1. При конкурирующей гипотезе $H_1 : a \neq a_0$, *найдем критическую*

точку $z_{кр}$ двусторонней критической области из равенства

$$\Phi(z_{кр}) = \frac{1}{2} - \frac{\alpha}{2} = \frac{(1-\alpha)}{2}. \qquad (13)$$

Если $|Z_{набл}| < z_{кр}$ нет оснований отвергнуть нулевую гипотезу. Если $|Z_{набл}| > z_{кр}$ - нулевую гипотезу отвергают.

Правило 2. При конкурирующей гипотезе $H_1 : a > a_0$ критическую точку правосторонней критической области находят из равенства

$$\Phi(z_{кр}) = \frac{1}{2} - \alpha = \frac{(1-2\alpha)}{2}. \qquad (14)$$

Если $Z_{набл} < z_{кр}$ нет оснований отвергнуть нулевую гипотезу. Если $Z_{набл} > z_{кр}$ - нулевую гипотезу отвергают.

Правило 3. При конкурирующей гипотезе $H_1 : a < a_0$ сначала находят вспомогательную критическую точку $z_{кр}$, *как в пункте* 2), а затем полагают границу левосторонней критической области $z_{левост.кр} = -z_{кр}$.

Если $Z_{набл} > -z_{кр}$ - нет оснований отвергнуть нулевую гипотезу. Если $Z_{набл} < -z_{кр}$ - нулевую гипотезу отвергают.

Пример 574.

Из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma = 5,2$ извлечена выборка объема $n = 100$ и по ней найдена выборочная средняя $\bar{x} = 27,56$. Требуется при уровне значимости 0,05 проверить нулевую гипотезу $H_0 : a = a_0 = 26$ при конкурирующей гипотезе $H_1 : a \neq 26$.

Решение. Найдем наблюдаемое значение критерия:

$$z_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma} = \frac{(27,56 - 26)\sqrt{100}}{5,2} = 3.$$

По условию, конкурирующая гипотеза имеет вид $a \neq a_0$, поэтому критическая область — двусторонняя. Найдем критическую точку из равенства (13)

$$\Phi(z_{кр}) = \frac{(1-\alpha)}{2} = \frac{(1-0,05)}{2} = 0,475.$$

По таблице функции Лапласа (см. приложение 2) находим $z_{кр} = 1,96$.

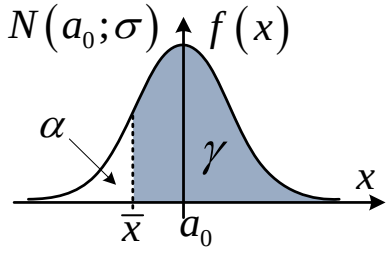
Так как $Z_{набл} > z_{кр}$ — нулевую гипотезу отвергаем. Другими словами, математическое ожидание генеральной совокупности ***a*** значимо отличается от гипотетического значения $a_0 = 26$.

В табличном процессоре Excel для проверки рассмотренной выше гипотезы используется функция ZTECT с обязательным указанием СКО. Обращение к ней имеет вид:

$$=\text{ZTECT}(\textit{массив};\ a_0;\ [\sigma]),$$

где *массив* – адреса ячеек, содержащих выборочные данные;

a_0 и σ – имеют прежний смысл, при этом, если параметр σ опущен, то используется исправленная выборочная дисперсия *s*, вычисленная по той же выборке.



Функция ZTECT численно равна вероятности того, что $P(X_0 > \bar{x})$,

где $X_0 = N(a_0;\sigma)$ - гипотетическая случайная величина.

$$ZTECT(\textit{массив};a_0;[\sigma]) = P(X_0 > \bar{x}) = \gamma.$$

Проверка гипотезы о числовом значении МО ***при неизвестной дисперсии****.*

Предполагается, что $X = N(a,\sigma)$, значения математического ожидания *a* и дисперсии σ^2 неизвестны.

Если дисперсия генеральной совокупности неизвестна (например, в случае малых выборок), то в качестве критерия проверки нулевой гипотезы принимают *t-статистику* - случайную величину

$$K = T_{n-1} = \frac{(\bar{X}_в - a_0)\sqrt{n}}{s},$$

которая имеет распределение Стьюдента с числом степеней свободы $k = n - 1$.

Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0 : a = a_0$ о равенстве неизвестной генеральной средней ***a*** гипотетическому значению a_0, вначале надо вычислить наблюдаемое значение критерия

$$T_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{s}. \qquad (15)$$

Затем, в зависимости от конкурирующей гипотезы определить критическую точку указанного распределения и критерий принятия гипотезы.

Правило 1. При конкурирующей гипотезе $H_1: a \neq a_0$, по заданному уровню значимости α, помещенному ***в верхней строке таблицы*** (прилож.5), и числу степеней свободы $k = n-1$ найти критическую точку $t_{двуст.кр}(\alpha, k)$.

Если $|T_{набл}| < t_{двуст.кр}(\alpha, k)$ — нет оснований отвергнуть нулевую гипотезу.

Если $|T_{набл}| > t_{двуст.кр}(\alpha, k)$ — нулевую гипотезу отвергают.

Правило 2. При конкурирующей гипотезе $H_1: a > a_0$ по уровню значимости α, помещенному ***в нижней строке таблицы*** (в 2 раза меньше) приложения 5, и числу степеней свободы $k = n-1$ находят критическую точку $t_{правст.кр}(\alpha, k)$ правосторонней критической области.

Если $T_{набл} < t_{правст.кр}(\alpha, k)$ — нет оснований отвергнуть нулевую гипотезу.

Если $T_{набл} > t_{правст.кр}(\alpha, k)$ — нулевую гипотезу отвергают.

Правило 3. При конкурирующей гипотезе $H_1: a < a_0$ сначала находят «вспомогательную» критическую точку (по правилу 2) $t_{правст.кр}(\alpha, k)$ и полагают границу левосторонней критической области $t_{левост.кр}(\alpha, k) = -t_{правст.кр}(\alpha, k)$.

Если $T_{набл} > -t_{правст.кр}(\alpha, k)$ — нет оснований отвергнуть нулевую гипотезу.

Если $T_{набл} < -t_{правст.кр}(\alpha, k)$ — нулевую гипотезу отвергают.

Примечание.

Для проверки указанной гипотезы можно применять критерий Стьюдента (Приложение 3), использовавшийся для интервального оценивания.

В этом случае

$$P(|T_{n-1}| < t_\gamma) = \gamma; \qquad P(|T_{n-1}| > t_\gamma) = 1-\gamma = \alpha. \tag{16}$$

Поэтому определим вероятность γ следующим образом:

для двусторонней критической области $1-\gamma = \alpha \quad \rightarrow \quad \gamma = 1-\alpha$.

для односторонней критической области $1-\gamma = 2\alpha \quad \rightarrow \quad \gamma = 1-2\alpha$;

1) Гипотеза $H_1: a \neq a_0$ - критическая область двусторонняя $t_{кр} = t_\gamma(1-\alpha, k)$.

При $|T_{набл}| > t_{кр} = t_\gamma(1-\alpha, k)$ — нулевую гипотезу отвергают.

2) Гипотеза $H_1: a > a_0$ - критическая область правосторонняя.

При $T_{набл} > t_{кр} = t_{\gamma}(1-2\alpha, k)$ — нулевую гипотезу отвергают.

3) Гипотеза $H_1 : a < a_0$ - критическая область левосторонняя.

При $T_{набл} < -t_{кр} = -t_{\gamma}(1-2\alpha, k)$ — нулевую гипотезу отвергают.

Пример 4.1. Хронометраж затрат времени на сборку узла машины n = 20 слесарей показал, что $\bar{x}_в = 77$ мин, а $s^2 = 4$ мин^2. В предположении о нормальности распределения решить вопрос: можно ли на уровне значимости $\alpha = 0.05$ считать 80 мин нормативом (математическим ожиданием) трудоемкости?

Решение. В качестве основной гипотезы принимается $H_0 : a = 80$ мин, в качестве альтернативной $H_1 : a \neq 80$ мин, т.е. имеем случай двусторонней критической области, при этом $a_0 = 80$.

Используя таблицу Приложения 5, находим по верхней строчке $\alpha = 0.05$ для $k = n - 1 = 19$ критическое значение $t_{двуст.кр} = 2,09$. Таким образом, критерий отклонения нулевой гипотезы - принадлежность наблюдаемого значения критической области $|T_{набл}| > 2,09$.

С использованием таблицы Приложения 3 получим тот же критерий отклонения нулевой гипотезы $|T_{набл}| > t_{\gamma}(1-\alpha, k) = t_{\gamma}(0,95;19) = 2,093$.

По формуле (15) вычисляем $T_{набл} = \dfrac{(77-80)\sqrt{20}}{2} = -6,71$.

Так как число –6,71 попадает в критическую область (конкретно в интервал $(-\infty, -2.09)$, то гипотеза $H_0 : a = 80$ мин отвергается.

В табличном процессоре Excel для проверки рассмотренной выше гипотезы используется уже известная функция ZTECT без указания СКО. Обращение к ней имеет вид:

=ZTECT(*массив*; a_0 ; [σ]),

где *массив* – адреса ячеек, содержащих выборочные данные;

a_0 и σ – имеют прежний смысл, при этом, если параметр σ опущен, то используется исправленная выборочная дисперсия *s*, вычисленная по той же выборке.

Проверка гипотезы о числовом значении дисперсии нормального распределения

Полагаем, что $X = N(a,\sigma)$. При этом по выборке объема n найдена исправленная дисперсия s^2.

Необходимо при заданном уровне значимости α проверить нулевую гипотезу $H_0 : \sigma^2 = \sigma_0^2$ о равенстве неизвестной генеральной дисперсии σ^2 гипотетическому (предполагаемому) значению σ_0^2.

В качестве критерия возьмем случайную величину

$$K = \chi^2_{n-1} = \frac{(n-1)s^2}{\sigma_0^2}, \qquad (17)$$

которая подчиняется χ^2-распределению с числом степеней свободы $k = n-1$.

После вычисления наблюдаемого значения критерия (17) определим его критические значения в зависимости от альтернативной гипотезы H_1 при заданном уровне значимости α и числу степеней свободы $k = n-1$ (Прил. 4):

Правило 1. При конкурирующей гипотезе $H_1 : \sigma^2 > \sigma_0^2$ критическая область правосторонняя с критической точкой $\chi^2_{кр}(\alpha;k)$.

Если $\chi^2_{набл} > \chi^2_{кр}$ — нулевую гипотезу отвергают. В противном случае, если $\chi^2_{набл} < \chi^2_{кр}$, — нет оснований отвергнуть нулевую гипотезу.

Правило 2. При конкурирующей гипотезе $H_1 : \sigma^2 \neq \sigma_0^2$ критическая область двусторонняя. Находят левую $\chi^2_{лев.кр}(1-\alpha/2;k)$ и правую $\chi^2_{прав.кр}(\alpha/2;k)$ критические точки.

Если $\chi^2_{набл} < \chi^2_{лев.кр}$ или $\chi^2_{набл} > \chi^2_{прав.кр}$ — нулевую гипотезу отвергают. Если $\chi^2_{лев.кр} < \chi^2_{набл} < \chi^2_{прав.кр}$ — нет оснований отвергнуть нулевую гипотезу.

Правило 3. При конкурирующей гипотезе $H_1 : \sigma^2 < \sigma_0^2$ критическая область левосторонняя. с критической точкой $\chi^2_{кр}(1-\alpha;k)$.

Если $\chi^2_{набл} < \chi^2_{кр}(1-\alpha;k)$ — нулевую гипотезу отвергают. Если $\chi^2_{набл} > \chi^2_{кр}(1-\alpha;k)$ — нет оснований отвергнуть нулевую гипотезу.

З а м е ч а н и е . Если число степеней свободы $k > 30$, то критическую точку $\chi^2_{kp}(\alpha,k)$ можно найти из равенства Уилсона—Гильферти [4]:

$$\chi^2_{кр}(\alpha,k)=k\left(1-\frac{2}{9k}+z_a\sqrt{\frac{2}{9k}}\right)^3, \qquad (18)$$

где z_a находят, используя функцию Лапласа, из равенства

$$\Phi(z_a)=(1-2\alpha)/2.$$

Пример 560.

Из нормальной генеральной совокупности извлечена выборка объема $n=21$ и по ней найдена исправленная выборочная дисперсия $s^2=16,2$. Требуется при уровне значимости 0,01 проверить нулевую гипотезу $H_0:\sigma^2=\sigma_0^2=15$, приняв в качестве конкурирующей гипотезы $H_1:\sigma_0^2>15$.

Решение. Найдем наблюдаемое значение критерия:

$$\chi^2_{набл}=\frac{(n-1)s^2}{\sigma_0^2}=\frac{(21-1)16,2}{15}=21,6.$$

По условию, конкурирующая гипотеза имеет вид $\sigma^2>15$, поэтому критическая область — правосторонняя (правило 1). По таблице приложения 4, по уровню значимости 0,01 и числу степеней свободы $k=n-1=21-1=20$ находим критическую точку $\chi^2_{кр}(0,01;20)=37,6$

Так как $\chi^2_{набл}<\chi^2_{кр}$ — нет оснований отвергнуть нулевую гипотезу о равенстве генеральной дисперсии гипотетическому значению $\sigma_0^2=15$. Другими словами, различие между исправленной дисперсией и гипотетической генеральной дисперсией незначимо.

Пример 565.

Партия изделий принимается, если дисперсия контролируемого размера значимо не превышает 0,2. Исправленная выборочная дисперсия, найденная по выборке объема $n=21$, оказалась равной $s_x^2=0,3$. Можно ли принять партию при уровне значимости 0,01?

Решение. Нулевая гипотеза $H_0:\sigma^2=\sigma_0^2=0,2$. Найдем наблюдаемое значение критерия:

$$\chi^2_{набл}=\frac{(n-1)s^2}{\sigma_0^2}=\frac{120\cdot 0,3}{0,2}=180.$$

Конкурирующая гипотеза имеет вид $H_1:\sigma^2>0,2$, следовательно, критическая область правосторонняя. Поскольку в таблице приложения 4 не содержится числа степеней свободы $k=120$, найдем критическую точку приближенно из равенства Уилсона — Гильферти (18):

$$\chi^2_{kp}(\alpha,k)=k\left(1-\frac{2}{9k}+z_a\sqrt{\frac{2}{9k}}\right)^3.$$

Найдем предварительно (учитывая, что по условию $\alpha=0,01$) $z_\alpha=z_{0,01}$ из равенства

$$\Phi\left(z_{0,01}\right)=\left(1-2\alpha\right)/2=\left(1-2\cdot 0,01\right)/2=0,49.$$

По таблице функции Лапласа (см. приложение 2), используя линейную интерполяцию, находим: $z_{0,01}=2,326$. Подставив $k=120, z_\alpha=2,326$ в формулу Уилсона — Гильферти, получим $\chi^2_{кр}\left(0,01;120\right)=158,85$. (Это приближение достаточно хорошее: в более полных таблицах χ^2 приведено значение 158,95). Так как $\chi^2_{набл}>\chi^2_{кр}$ — нулевую гипотезу отвергаем. Партию принять нельзя.

Проверка гипотезы о числовом значении вероятности события

Пусть по достаточно большому числу *n* независимых испытаний, в каждом из которых вероятность *p* появления события постоянна, но неизвестна, найдена относительная частота *m*/*n*. Требуется при заданном уровне значимости α проверить нулевую гипотезу $H_0: p=p_0$ о равенстве неизвестной вероятности *p* некоторому гипотетическому значению p_0.

В качестве критерия возьмем величину

$$Z=\frac{\left(m/n-p_0\right)\sqrt{n}}{\sqrt{p_0q_0}}=N(0,1), \qquad (19)$$

значение которой подчиняется стандартному нормальному распределению.

Критическое значение критерия z_{kp} определим в зависимости от альтернативной гипотезы (типа критической области) с использованием функции Лапласа (Приложение 2) по формулам (13), (14) для МО.

Замечание. Удовлетворительные результаты обеспечивает выполнение неравенства $np_0q_0>9$.

Пример 586. По 100 независимым испытаниям найдена относительна частота $m/n=0,14$. При уровне значимости 0,05 требуется проверить нулевую гипотезу $H_0: p=p_0=0,20$ при конкурирующей гипотезе $H_1: p\neq 0,20$.

Решение.

Найдем наблюдаемое значение критерия, учитывая, что $q_0=1-p_0=0,8$:

$$Z_{набл} = \frac{(0{,}14 - 0{,}2)\sqrt{100}}{\sqrt{0{,}2 \cdot 0{,}8}} = -1{,}5.$$

По условию, конкурирующая гипотеза имеет вид $H_1 : p \neq 0{,}2$, поэтому критическая область — двусторонняя. Найдем критическую точку z_{kp} по равенству (13)

$$\Phi(z_{кр}) = \frac{(1-\alpha)}{2} = \frac{1 - 0{,}05}{2} = 0{,}475.$$

По таблице функции Лапласа (см. приложение 2) находим $z_{kp} = 1{,}96$.

Так как $|Z_{набл}| < z_{кр}$ — нет оснований отвергнуть нулевую гипотезу. Другими словами, наблюдаемая относительная частота 0,14 незначимо отличается от гипотетической вероятности 0,20.

Проверка гипотезы о значимости выборочного коэффициента корреляции

Пусть двумерная генеральная совокупность (X, Y) распределена нормально. Из этой совокупности извлечена выборка объема *n* и по ней найден выборочный коэффициент корреляции $r_B \neq 0$. Требуется проверить нулевую гипотезу $H_0 : r_Г = 0$ о равенстве нулю генерального коэффициента корреляции.

Если нулевая гипотеза принимается, то это означает, что X и Y некоррелированы; в противном случае — коррелированы.

Для того чтобы при уровне значимости α проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции нормальной двумерной случайной величины при конкурирующей гипотезе $H_1 : r_Г \neq 0$, надо вычислить наблюдаемое значение критерия

$$K = T_{набл} = r_B \sqrt{n-2} \Big/ \sqrt{1 - r_B^2} \qquad (20)$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α и числу степеней свободы $k = n - 2$ найти критическую точку $t_{kp}(\alpha, k)$ двусторонней критической области. Если $|T_{набл}| < t_{kp}$ - нет оснований отвергнуть нулевую гипотезу. Если $|T_{набл}| > t_{kp}$ - нулевую гипотезу отвергают.

Пример 610.

По выборке объема *n=100*, извлеченной из двумерной нормальной генеральной совокупности (X, Y), найден выборочный коэффициент

корреляции $r_B = 0,2$. Требуется при уровне значимости 0,05 проверить нулевую гипотезу $H_0 : r_Г = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $H_1 : r_Г \neq 0$.

Решение.

Найдем наблюдаемое значение критерия (20)

$$T_{набл} = 0,2\sqrt{100-2}\Big/\sqrt{1-0,2^2} = 2,02.$$

По условию, конкурирующая гипотеза $H_1 : r_Г \neq 0$, поэтому критическая область - двусторонняя.

По таблице критических точек *t*-распределения Стьюдента (прил.5), по уровню значимости $\alpha = 0,05$ в верхней части таблицы и числу степеней свободы *k=100-2=98* методом линейной интерполяции находим критическую точку двусторонней критической области $t_{кр}(0,05;98) = 1,99$.

Поскольку $|T_{набл}| > t_{кр}$, то нулевую гипотезу отвергаем. Генеральный коэффициент корреляции значимо отличается от нуля, а генеральные совокупности (X, Y) коррелированы.

Типовой вариант задания

1. Регрессионный анализ и проверка гипотез.

1.1. i-му варианту соответствуют выборки случайных величин X и Y, расположенные в соответствующих строках.

1.2. По данным выборки определить:

- оценку вектора математического ожидания;
- оценку вектора выборочной дисперсии;
- выборочную ковариацию;
- выборочный коэффициент корреляции;
- выборочные коэффициенты линейных регрессий Y на X и X на Y;
- выборочные уравнения линейных регрессий Y на X и X на Y.

1.3. Используя метод наименьших квадратов, найти параметры линейной, квадратичной и логарифмической регрессий.

1.4. Определите общую, объясненную и остаточную дисперсии для рассмотренных уравнений регрессий, а также коэффициенты детерминации.

1.5. Построить поле корреляции и найденные линии регрессии.

1.6. Полагая в дальнейшем, что наблюдаемые признаки X и Y распределены по нормальному закону, проверить гипотезы о равенстве их математических ожиданий гипотетическим значениям $H_0 : a = a_0$ с уровнями значимости α_m при альтернативе $H_1 : a$ "*знак*" a_0 (см. таблицу).

1.7. Проверить гипотезы о равенстве генеральных дисперсий гипотетическим значениям $H_0 : \sigma^2 = D_0$ с уровнями значимости α_m при альтернативе $H_1 : \sigma^2 \neq D_0$ (см. таблицу).

1.8. Проверить гипотезу $H_0 : D(X) = D(Y)$ о равенстве генеральных дисперсий признаков X и Y по их выборкам при конкурирующей гипотезе об их неравенстве при уровне значимости α_S (см. таблицу).

1.9. Проверить гипотезу о значимости выборочного коэффициента корреляции при уровне значимости α_S.

Вариант		Номер измерения															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	x	2,3	2,1	4,6	4,9	3,7	5	2,8	1,6	4,8	3,7	1,1	2,4	4,3	4,5	3,2	2,4
	y	-1,9	-0,2	0,0	1,6	0,6	-0,9	0,0	-2,4	0,4	3,2	-4,6	-1,9	3,0	2,6	2,8	1,0

Интервал. оц.			Проверка гипотез				
γ_m	δ_m	γ_D	α_m	зн	a_0	D_0	α_S
0,95	*0,16*	*0,95*	*0,05*	*<4,8*		*0,75*	
			0,02	*>-0,3*		*7,35*	*0,05*

ЛИТЕРАТУРА

1. **Вентцель Е.**С. Теория вероятностей: учебник для студентов вузов. – 10 изд. – М.: Издательский центр «Академия», 2005. – 576с.

2. **Гмурман В.Е.** Теория вероятностей и математическая статистика: уч. пособие. – 12-е изд. – М.: Высшее образование, 2008. - 479с.

3. **Кремер Н.Ш.** Теория вероятностей и математическая статистика: учебник для вузов. – М.: ЮНИТИ-ДАНА, 2000. - 543с.

4. **Гмурман В.Е.** Руководство к решению задач по теории вероятностей и математической статистике: уч. пособие для студентов вузов. – 4-е изд. – М.: Высшая школа, 1997. - 400с.

5. **Письменный Д.Т.** Конспект лекций по теории вероятностей, математической статистике и случайным процессам. – 3-е изд. – М.: Айрис-пресс, 2008. – 288с.

6. **Гулай Т.А., Долгополова А.Ф., Литвин Д.Б., Мелешко С.В.** Теория вероятностей и математическая статистика. - 2-е изд. - Ставрополь : Агрус, 2013. - 256с.

7. **Литвин Д.Б., Мелешко С.В.** Элементы теории вероятностей: Учебное пособие. – Ставрополь: Сервисшкола, 2017. – 86 с.