Решение типового варианта контрольной работы по математической статистике.

Задача 1.

Среди 16 лотерейных билетов находятся 4 выигрышных билета. Найти вероятность того, что среди четырех купленных билетов, случайным образом выбранных: а) ровно один выигрышный; б) хотя бы один выигрышный.

Решение. а) Найдем вероятность события: первый купленный билет выигрышный, а остальные – невыигрышные. Такое событие, обозначим его B_1, является совмещением четырех событий $B_1 = A_1A_2A_3A_4$, где A_1 – первый купленный билет выигрышный, A_2, A_3, A_4, соответственно, второй, третий и четвертый купленные билеты невыигрышные. Вероятность совмещения этих событий найдем по формуле (1.11):

$P\,(A_1A_2A_3A_4) = P(A_1)\cdot P(A_2/A_1)\cdot P(A_3/(A_1A_2))\cdot P(A_4/(A_1A_2A_3))$.

В силу равной возможности исходов, обеспеченной перемешиванием билетов, для вычисления как безусловных, так и условных вероятностей можно воспользоваться формулой вероятности в классической модели (1.1).

$P(A_1) = \dfrac{4}{16}$. Всего билетов 16 – общее число исходов. Выигрышных билетов 4 – число исходов, благоприятствующих событию куплен выигрышный билет.

$P(A_2/A_1) = \dfrac{12}{15}$. Найдена вероятность того, что второй билет невыигрышный, если первый был выигрышный. Общее число исходов 15, так как один билет уже куплен, осталось 15. Число исходов, благоприятствующих событию, равно 12, так как первым куплен выигрышный билет, а все невыигрышные остались.

$P(A_3/(A_1A_2)) = \dfrac{11}{14}$. Найдена вероятность того, что третий билет

невыигрышный, если первый был выигрышный, а второй – невыигрышный. После покупки двух билетов осталось 14, среди них 11 невыигрышных.

$P(A_4/(A_1A_2A_3)) = \dfrac{10}{13}$. Найдена вероятность того, что четвертый билет невыигрышный, если первый был выигрышный, а второй и третий – невыигрышные. После покупки трех билетов осталось 13, среди них 10 невыигрышных.

Тогда

$$P(B_1) = P\,(A_1A_2A_3A_4) = \frac{4}{16}\,\frac{12}{15}\,\frac{11}{14}\,\frac{10}{13}. \qquad (2.1)$$

Итак, найдена вероятность, что первый купленный билет выигрышный, остальные – невыигрышные. Но выигрышным может быть второй, третий или четвертый купленный билет при остальных невыигрышных. Возникают еще три несовместных события B_2, B_3 и B_4, вероятности каждого из которых равны (2.1) с точностью до перестановки множителей. Например, вероятность события B_2, второй купленный билет выигрышный, остальные невыигрышные, равна

$$P(B_2) = \frac{12}{16}\,\frac{4}{15}\,\frac{11}{14}\,\frac{10}{13}.$$

Искомое событие, куплен ровно один выигрышный билет, равно сумме событий $B_1 + B_2 + B_3 + B_4$. Так как события B_i, i =1, 2, 3, 4, попарно несовместные, вероятность их суммы равна сумме вероятностей событий B_i (формула (1.4)). Но вероятность каждого из событий B_i одинакова и равна (2.1). Поэтому искомая вероятность равна вероятности (2.1), умноженной на 4, т.е. $4\frac{4}{16}\frac{12}{15}\frac{11}{14}\frac{10}{13} \approx 0{,}4835$.

б) Для нахождения вероятности покупки хотя бы одного выигрышного билета (событие C) сначала найдем вероятность противоположного события – все купленные билеты невыигрышные (событие $\overline{C}$). Это событие соответствует совмещению событий – каждый из купленных билетов невыигрышный. Вероятность такого события вычисляется по формуле (1.11):

$P(\overline{C}) = \frac{12}{16}\frac{11}{15}\frac{10}{14}\frac{9}{13} \approx 0{,}272.$

Тогда $P(C) = 1 - P(\overline{C}) = 1 - 0{,}272 = 0{,}728$.

Ответ: а) 0,4835; б) 0,728.

Задача 2.

Среди населения 5 % мужчин и 0,25 % женщин являются дальтониками. а) Какова вероятность, что случайным образом выбранное лицо страдает дальтонизмом (считать, что мужчин и женщин одинаковое количество). б) Случайным образом выбранное лицо страдает дальтонизмом. Найти вероятность того, что это мужчина.

Решение. Для решения задачи а) воспользуемся формулой полной вероятности

$$P(A) = \sum_{i=1}^{N} P(H_i)\cdot P(A/H_i). \qquad (1.15)$$

Сформулируем две гипотезы: случайно выбранное лицо является мужчиной (H_1) или женщиной (H_2). Согласно условию: $P(H_1) = P(H_2) = 0{,}5$. Событие А – случайно выбранное лицо страдает дальтонизмом. Условные вероятности, как следует из условия, равны

$P(A/H_1) = 0{,}05; \quad P(A/H_2) = 0{,}0025.$

Тогда по формуле полной вероятности (1.15)

$P(A) = P(H_1)\cdot P(A/H_1) + P(H_2)\cdot P(A/H_2) = 0{,}5 \cdot 0{,}05 + 0{,}5 \cdot 0{,}0025 = 0{,}02625.$

б) Искомую вероятность найдем по формуле Байеса

$P(H_1/A) = \frac{P(H_1)\cdot P(A/H_1)}{P(A)} = \frac{0{,}5\cdot 0{,}05}{0{,}02625} \approx 0{,}952.$

Ответ: а) 0,02625; б) 0,952.

Задача 3.

Функция плотности непрерывной случайной величины задана формулой: $\varphi(x) = \begin{cases} a(4x - x^2), & x \in [0,4] \\ 0, & x \notin [0,4] \end{cases}$. Найти коэффициент a, записать функцию распределения. Вычислить M(X), $P(X > 1)$, медиану.

Решение. Коэффициент a находим из условия нормировки функции плотности

$$\int_{-\infty}^{+\infty} \varphi(x)dx = 1.$$

В нашем случае $\quad 1 = \int_0^4 a(4x - x^2)dx = a\left(4\frac{x^2}{2} - \frac{x^3}{3}\right)\Bigg|_0^4 = a\left(32 - \frac{64}{3}\right) = \frac{32}{3}a,$

Откуда получаем $a = \frac{3}{32}$.

Функцию распределения $F(x)$ находим как первообразную функции плотности

$$F(x)=\int_{-\infty}^{x}\varphi(t)dt\,.$$

В нашем случае $F(x)=0$ при $x<0$. При $x\in[0;4]$

$$F(x)=\int_0^x\varphi(t)dt=\int_0^x\frac{3}{32}(4t-t^2)dt=\frac{3}{32}\left(4\frac{t^2}{2}-\frac{t^3}{3}\right)\Bigg|_0^x=\frac{3}{32}\left(2x^2-\frac{x^3}{3}\right)=\frac{3x^2}{16}-\frac{x^3}{32};$$

и $F(x)=1$ при $x>4$.

Математическое ожидание $M(X)$ равно

$$M(X)=\int_{-\infty}^{+\infty}x\cdot\varphi(x)dx=$$

$$=\int_0^4 x\frac{3}{32}(4x-x^2)dx=\frac{3}{32}\int_0^4(4x^2-x^3)dx=\frac{3}{32}\left(4\frac{x^3}{3}-\frac{x^4}{4}\right)\Bigg|_0^x=\frac{3}{32}\left(\frac{4^4}{3}-\frac{4^4}{4}\right)=2.$$

Вероятность события $X>1$ находим, воспользовавшись формулой

$$P(X\in(x_1,\,x_2))=\int_{x_1}^{x_2}\varphi(x)dx$$

$$P(X>1)=\int_1^4\varphi(t)dt=\int_0^x\frac{3}{32}(4x-x^2)dx=\frac{3}{32}\left(2x^2-\frac{x^3}{3}\right)\Bigg|_1^4=\frac{27}{32}\approx 0,84375.$$

Медиана распределения m_x находится из уравнения $F(m_x)=0,5.$ В нашем случае

$$\frac{3m_x{}^2}{16}-\frac{m_x{}^3}{32}=\frac{1}{2};\qquad 6m_x{}^2-m_x{}^3-16=0;\qquad m_x{}^3-6m_x{}^2\_16=0;$$

откуда $m_x=2.$

Ответ: $a=\dfrac{3}{32};$ $\qquad F(x)=\begin{cases}0, & x<0\\ \dfrac{3x^2}{16}-\dfrac{x^3}{32}, & x\in[0;4];\\ 1, & x>4\end{cases}$ $\qquad M(X)=2;$

$P\approx 0,84375;$ $\qquad m_x=2.$

Задача 4.

Функция распределения непрерывной случайной величины задана в виде $F(x)=\begin{cases}0 & x\le 0\\ ax^2+b & 0<x\le 2\\ 1 & x>2\end{cases}$. Найти параметры a и b. Вычислить M(X), D(X), P(X > 1).

Решение. Найдем параметры a и b из условия непрерывности функции распределения:

$$F(0)=0=a\cdot 0^2+b;\quad\Rightarrow\quad b=0;$$

$$F(2)=1=a\cdot 2^2+b=4a;\quad\Rightarrow\quad a=\frac{1}{4};$$

откуда следует, что $F(x)=\dfrac{x^2}{4}$ при $x\in[0;2]$.

Запишем уравнение функции плотности из условия (1.46) $\varphi(x)=F'(x)$:

$$\varphi(x)=\begin{cases}0, & x<0\\ \dfrac{x}{2}, & x\in[0;2]\\ 0, & x>2\end{cases}$$

Вычислим числовые характеристики случайной величины.

$$M(X)=\int_{-\infty}^{+\infty} x\cdot\varphi(x)dx=\int_0^2 x\cdot\frac{x}{2}dx=\frac{x^3}{6}\Big|_0^2=\frac{8}{6}=\frac{4}{3};$$

$$M(X^2)=\int_{-\infty}^{+\infty} x^2\cdot\varphi(x)dx=\int_0^2 x^2\cdot\frac{x}{2}dx=\frac{x^4}{8}\Big|_0^2=\frac{16}{8}=2;$$

$$D(X)=M(X^2)-(M(X))^2=2-\left(\frac{4}{3}\right)^2=\frac{2}{9};$$

$$P(X>1)=1-F(1)=1-\frac{1}{4}=\frac{3}{4}.$$

Ответ: a = 0,25; b = 0; $\varphi(x)=\begin{cases}0, & x<0\\ \dfrac{x}{2}, & x\in[0;2];\\ 0, & x>2\end{cases}$ $M(X)=\dfrac{4}{3}\approx 1{,}333;$

$D(X)=\dfrac{2}{9}\approx 0{,}222;\quad P=0{,}75.$

Задача 5.

В таблице 1 в первом столбце записаны результаты n = 18 независимых равноточных измерений величины X. Вычислить оценки математического ожидания, дисперсии и среднего квадратического отклонения величины X.

Предполагая, что результаты измерений независимы и имеют нормальное распределение с одинаковыми параметрами, построить доверительные интервалы для математического ожидания и среднего квадратического отклонения с доверительной вероятностью P = 0,95.

Таблица 1.

Исходные данные и результаты расчетов (к задаче 1)

Значение X	Результаты расчетов		Контроль правильности расчетов	
	U	U^2	V	V^2
4,761	–19	361	–29	841
4,792	12	144	2	4
4,758	–22	484	–32	1024
4,764	–16	256	–26	676
4,810	30	900	20	400
4,799	19	361	9	81
4,797	17	289	7	49
4,790	10	100	0	0
4,747	–33	1089	–43	1849
4,769	–11	121	–21	441
4,806	26	676	16	256
4,779	–1	1	–11	121

4,785	5	25	–5	25
4,790	10	100	0	0
4,777	–3	9	–13	169
4,749	–31	961	–41	1686
4,781	1	1	–9	81
4,799	19	361	9	81
Сумма	13	6239	–167	7779

Решение. Пусть задана повторная случайная выборка $X_1, X_2, ..., X_n$. За *оценку математического ожидания* a принимается среднее арифметическое элементов выборки:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i . \qquad (3.2)$$

Оценкой дисперсии σ^2 при *неизвестном математическом ожидании* является величина S^2, которую называют *эмпирической дисперсией*:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 . \qquad (3.3)$$

О*ценкой среднего квадратического отклонения* σ при этом является, соответственно, величина

$$S = \sqrt{S^2} . \qquad (3.4)$$

Для практических расчетов формулу (3.3) целесообразно преобразовать к следующему виду:

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right). \qquad (3.5)$$

Вычисление среднего значения $\bar{X}$ и оценки дисперсии S^2 упрощается, если отсчет значений X_i вести от подходящим образом выбранного начала отсчета *C* и в подходящем масштабе, т.е. сделать линейную замену (кодирование):

$$U_i = \frac{X_i - C}{h}; \qquad X_i = C + hU_i \ (i = 1, 2, ..., n) . \qquad (3.6)$$

При такой замене формулы (3.2), (3.3), (3.4) принимают следующий вид:

$$\bar{X} = C + h\bar{U}; \qquad \bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_i . \qquad (3.7)$$

$$S^2 = \frac{h^2}{n-1}\sum_{i=1}^{n}(U_i - \bar{U})^2 = \frac{h^2}{n-1}\left(\sum_{i=1}^{n} U_i^2 - n\bar{U}^2\right). \qquad (3.8)$$

Выберем C = 4,780 и, полагая $h = 10^{-3}$, подсчитаем значения

$$U_i = (X_i - C)/h = (X_i - 4{,}780)/10^{-3} \text{ и } U_i^2 .$$

Суммы чисел второго и третьего столбца дают возможность рассчитать $\bar{X}$ и S^2:

$$\bar{U} = 13/18 = 0{,}72;$$

$$\bar{X} = 4{,}780 + 0{,}72 \cdot 10^{-3} = 4{,}7807;$$

$$S^2 = 10^{-6}(6239 - 13^2/18)/17 = 3{,}66 \cdot 10^{-4},$$

откуда $S = \sqrt{S^2} = \sqrt{3{,}66 \cdot 10^{-4}} = 1{,}91 \cdot 10^{-2}$.

В последних двух столбцах приведены расчеты при другом начале отсчета C_1 = 4,790. Новые кодированные значения обозначены как $V_i = (X_i - 4{,}790)/10^{-3}$. Эти расчеты приводят к тем же значениям $\bar{X}$ и *S*:

$$\bar{V} = 167/18 = -9{,}2, \quad \bar{X} = 4{,}790 - 9{,}28 \cdot 10^{-3} = 4{,}7807.$$

$$S^2 = 10^{-6}(7779 - (167^2)/18)/17 = 3{,}66 \cdot 10^{-4}, \quad S = 1{,}91 \cdot 10^{-2}.$$

Найдем доверительные интервалы для математического ожидания и среднего квадратического отклонения с доверительной вероятностью $P = 0{,}95$.

С вероятностью P математическое ожидание a принадлежит интервалу

$$a \in (\overline{X} - \varepsilon;\ \overline{X} + \varepsilon); \tag{3.9}$$

$$\varepsilon = t_{1-\alpha/2}(k)\, S\big/\sqrt{n}, \tag{3.10}$$

где $\overline{X}$ – оценка математического ожидания (3.2); $S = \sqrt{S^2}$ – оценка среднего квадратического отклонения σ(3.4); $t_{1-\alpha/2}(k)$ – квантиль распределения Стьюдента с k степенями свободы; n – объем выборки; k – число степеней свободы при вычислении оценки S.

Часто доверительный интервал для математического ожидания записывают символически:

$$a = \overline{X} \pm \varepsilon. \tag{3.11}$$

По формуле (3.10) $\varepsilon = t_{1-\alpha/2}(k)\, S\big/\sqrt{n}$. В таблице квантилей распределения Стьюдента находим $t_{1-\alpha/2}(k) = t_{0{,}975}(17) = 2{,}11$. Тогда

$$\varepsilon = 2{,}11 \cdot 0{,}0191\big/\sqrt{18} \approx 0{,}0095.$$

По формуле (3.11) $a = 4{,}7807 \pm 0{,}0095$, т.е. с вероятностью $P = 0{,}95$ выполняется неравенство $4{,}7712 < a < 4{,}7902$.

Доверительный интервал для среднего квадратического отклонения σ при доверительной вероятности $P = 1 - \alpha$ имеет следующий вид:

$$S\sqrt{\frac{k}{\chi^2_{1-\alpha/2}(k)}} < \sigma < S\sqrt{\frac{k}{\chi^2_{\alpha/2}(k)}}, \tag{3.12}$$

где S – оценка среднего квадратического отклонения σ(1.6) при неизвестном математическом ожидании; $\chi^2_P(k)$ – квантиль распределения Пирсона с k степенями свободы; k – число степеней свободы оценки S.

В таблице квантилей хи-квадрат распределения находим:

$\chi^2_{\alpha/2}(k) = \chi^2_{0{,}025}(17) = 7{,}56$; $\quad \chi^2_{1-\alpha/2}(k) = \chi^2_{0{,}975}(17) = 30{,}2$.

По формуле (3.12) получаем

$$0{,}0191\sqrt{\frac{17}{30{,}2}} < \sigma < 0{,}0191\sqrt{\frac{17}{7{,}56}},$$

откуда $\quad \sigma \in (0{,}0143;\ 0{,}0287)$.

Задача 6.

В двух сериях независимых экспериментов с числом измерений, соответственно, $n_1 = 15$ и $n_2 = 10$ получены оценки математического ожидания $\overline{X}_1 = 20{,}5$; $\overline{X}_2 = 18{,}7$ и оценки дисперсии $S_1^2 = 2{,}2$; $S_2^2 = 5{,}5$. Известно, что результаты измерений в каждой серии имеют нормальный закон распределения. Используя двусторонние критерии, проверить с уровнем значимости α = 0,05 гипотезу о равенстве дисперсий и гипотезу о равенстве математических ожиданий.

Решение. Пусть заданы две независимые выборки из двух нормальных генеральных совокупностей. Первая выборка имеет объем n_1, элементы выборки $X_i^{(1)} \sim N(a_1; \sigma_1)$; вторая – объем n_2, элементы выборки $X_i^{(2)} \sim N(a_2; \sigma_2)$. Необходимо проверить *гипотезу о равенстве дисперсий* этих двух генеральных совокупностей, т.е. H_0: $\sigma_1^2 = \sigma_2^2$. Математические ожидания a_1 и a_2 неизвестны.

В этом случае по каждой выборке находят несмещенные оценки дисперсий S_1^2 и S_2^2 с числами степеней свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ соответственно. Гипотезу проверяют по критерию Фишера, функция критерия

$$F = S_1^2 \big/ S_2^2 \tag{3.13}$$

имеет F-распределение Фишера с k_1 и k_2 степенями свободы, т.е. $F = F(k_1, k_2)$.

Если альтернативная гипотеза H_1: $\sigma_1^2 \neq \sigma_2^2$, то критерий Фишера рассчитывается как отношение большей по величине оценки дисперсии к меньшей:

$$F = S_{\text{бол}}^2 \big/ S_{\text{мен}}^2 > 1. \tag{3.14}$$

Гипотеза принимается при выполнении неравенства

$$F < F_{1-\alpha/2}(k_{S_{\text{бол}}}, k_{S_{\text{мен}}}), \tag{3.15}$$

в противоположном случае гипотеза отвергается. Здесь $k_{S_{\text{бол}}}$ – число степеней свободы большей оценки дисперсии; $k_{S_{\text{мен}}}$ – число степеней свободы меньшей оценки дисперсии.

Если гипотеза о равенстве дисперсий принимается, то за оценку общей σ может быть взята $S = \sqrt{S_{\text{св}}^2}$, полученная по формуле для сводной оценки дисперсии:

$$S_{\text{св}}^2 = \frac{k_1 S_1^2 + k_2 S_2^2}{k_1 + k_2}; \quad k_j = n_j - 1; \quad j = 1,\ 2, \tag{3.16}$$

где S_1^2, S_2^2 – несмещенные оценки дисперсии первой и второй выборок соответственно.

В нашем случае $F = \dfrac{S_{\text{бол}}^2}{S_{\text{мен}}^2} = \dfrac{5,5}{2,2} = 2,5.$ $F_{\text{кр}} = F_{0,975}(9;\ 14) = 3,21.$

Гипотеза о равенстве дисперсий принимается.

$$S_{св}^2 = \frac{k_1 S_1^2 + k_2 S_2^2}{k_1 + k_2} = \frac{14 \cdot 2,2 + 9 \cdot 5,5}{14 + 9} = 3,491; \quad S_{св} = 1,868.$$

Пусть заданы две независимые выборки из двух нормальных генеральных совокупностей. Первая выборка имеет объем n_1, элементы выборки $X_i^{(1)} \sim N(a_1; \sigma_1)$; вторая – объем n_2, элементы выборки $X_i^{(2)} \sim N(a_2; \sigma_2)$. Математические ожидания a_1 и a_2 неизвестны.

Проверяем *гипотезу о равенстве математических ожиданий* этих двух генеральных совокупностей, т.е. H_0: $a_1 = a_2$. По каждой выборке находим оценки математических ожиданий $\overline{X}_1$ и $\overline{X}_2$. При этом дисперсии σ_1^2 и σ_2^2 неизвестны, но гипотеза о равенстве дисперсий принимается; S_1^2 и S_2^2 – несмещенные оценки дисперсий первой и второй выборок. Находим сводную оценку дисперсии (3.16).   .

Гипотеза проверяется по критерию Стьюдента, функция критерия

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\text{св}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.17}$$

имеет t-распределение Стьюдента с $k_{\text{св}}$ степенями свободы, т.е. $t = t(k_{\text{св}})$;. $k_{\text{св}} = k_1 + k_2$ – число степеней свободы при вычислении оценки $S_{\text{св}} = \sqrt{S_{\text{св}}^2}$. При альтернативной гипотезе H_1: $a_1 \neq a_2$. гипотеза принимается при выполнении неравенства

$$|t| < t_{1-\alpha/2}(k_{\text{св}}), \tag{3.18}$$

в противоположном случае гипотеза отвергается.

В нашем случае $t = \dfrac{20,5-18,7}{1,868\sqrt{1/15+1/10}} = 2,36.$ $t_{\text{кр}} = t_{0,975} = 2,069$.

Гипотеза о равенстве математических ожиданий отвергается.

Задача 7.

Задача 1.28. В первом столбце табл. 1.13 записаны измеренные значения величины X – изменения содержания азота в стали при выпуске из конвертера по сравнению с начальным содержанием) [10^{-4} % ·]; во втором – величины Y (значения начальной концентрации углерода в этой же стали [%]). Найти оценку коэффициента корреляции по этой двумерной выборке. Вычислить выборочные параметры линейной регрессии Y на X и X на Y.

Таблица 1.17

Исходные данные и результаты расчетов к задаче 1.28.

№	X	Y	U	V	U^2	V^2	UV
1	–2,0	0,11	–4	1	16	1	–4
2	0,5	0,09	1	–1	1	1	–1
3	–1,5	0,13	–3	3	9	9	–9
4	–5,5	0,11	–11	1	121	1	–11
5	3,5	0,06	7	–4	49	16	–28
6	–1,0	0,12	–2	2	4	4	–4
7	2,0	0,08	4	–2	16	4	–8
8	0,0	0,11	0	1	0	1	0
9	1,5	0,07	3	–3	9	9	–9
$\sum$	–	–	–5	–2	225	46	–74

Решение

Вводим линейную замену $X_i = C_1 + h_1 U_i;\ Y_i = C_2 + h_2 V_i$, выбирая $C_1 = 0$, $h_1 = 0,5$; $C_2 = 0,10$, $h_2 = 10^{-2}$. Вычисляем оценки математических ожиданий $\overline{X} = C + h\overline{U}$, $\overline{U} = \frac{1}{n}\sum_{i=1}^{n} U_i$

$$\overline{U} = -\frac{5}{9} \approx 0,556;\quad \overline{X} = -0,5 \cdot 0,556 = -0,278;$$

$$\overline{V} = -\frac{2}{9} \approx -0,22;\quad \overline{Y} = 0,10 - 0,22 \cdot 10^{-2} = 0,0978.$$

Несмещенные оценки дисперсий находим по формуле

$$S^2 = \frac{h^2}{n-1}\sum_{i=1}^{n}\left(U_i - \overline{U}\right)^2 = \frac{h^2}{n-1}\left(\sum_{i=1}^{n} U_i^2 - n\overline{U}^2\right)$$

$$S_x^2 = \frac{(0,5)^2}{8}\left(225 - 9\left(-\frac{5}{9}\right)^2\right) \approx 6,94; \quad S_x \approx 2,63;$$

$$S_y^2 = \frac{10^{-4}}{8}\left(46 - 9\left(-\frac{2}{9}\right)^2\right) \approx 5,69 \cdot 10^{-4}; \quad S_y = 2,39 \cdot 10^{-2}.$$

Расчет оценки ковариации проводим по формуле $\tilde{K}_{xy} = \frac{h_1 h_2}{n-1}\left(\sum_{i=1}^{n} U_i V_i - n\overline{U}\,\overline{V}\right)$

$$\tilde{K}_{xy} = \frac{0,5 \cdot 10^{-2}}{8}\left(-74 - 9\left(-\frac{5}{9}\right)\left(-\frac{2}{9}\right)\right) \approx -4,69 \cdot 10^{-2}.$$

Оценку коэффициента корреляции находим по формуле $\rho_{xy} \approx r = \frac{\tilde{K}_{xy}}{S_x S_y}$

$$r = \frac{-4,69 \cdot 10^{-2}}{2,69 \cdot 2,39 \cdot 10^{-2}} \approx -0,746.$$

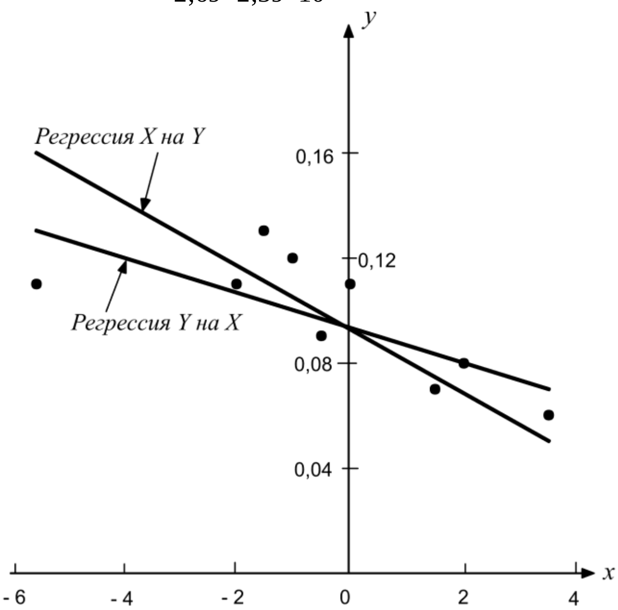

Рис. 1.5. Зависимость изменения концентрации азота в стали (y) при выпуске из конвертера от начальной концентрации углерода (x)

Выборочное уравнение линейной регрессии Y на X:

$$\frac{y - \overline{Y}}{S_y} = r\frac{x - \overline{X}}{S_x}; \qquad \frac{y - 0,0978}{2,39 \cdot 10^{-2}} = -0,746 \cdot \frac{x + 0,278}{2,63}$$

или $$y - 0,0978 = -0,00678(x + 0,278).$$

Выборочное уравнение линейной регрессии X на Y:

$$\frac{y - \overline{Y}}{S_y} = \frac{1}{r}\frac{x - \overline{X}}{S_x} \qquad \frac{y - 0,0978}{2,39 \cdot 10^{-2}} = -\frac{1}{0,746} \cdot \frac{x + 0,278}{2,63}$$

или $$y - 0,0978 = -0,0122(x + 0,278).$$

Прямые регрессии представлены на рис. 1.5, там же приведены экспериментальные точки.

Для проверки гипотезы H_0: $\rho_{xy} = 0$ при альтернативной гипотезе H_1: $\rho_{xy} \neq 0$ можно использовать следующий критерий. Гипотеза H_0 принимается с уровнем значимости α, то есть линейная зависимость между величинами не существует, если $|r| < T$, где T – значение критерия:

$$T = \frac{t_{1-\alpha/2}(n-2)}{\sqrt{n-2+t^2_{1-\alpha/2}(n-2)}}, \qquad (1.99)$$

в противном случае принимается гипотеза H_1, то есть предполагается, что линейная зависимость между величинами существует. $t_{1-\alpha/2}(n-2)$ – квантиль распределения Стьюдента с числом степеней свободы $k = n-2$.

Если принята гипотеза о существовании линейной зависимости между случайными величинами, то, зная доверительный интервал для коэффициента корреляции, можно сделать вывод о силе взаимосвязи между X и Y. Если доверительный интервал примыкает к единице или минус единице, то говорят, что связь сильная. Если доверительный интервал примыкает к нулю, то говорят, что связь слабая. Если доверительный интервал расположен примерно посередине интервала (–1; 0) или (0; 1), то говорят, что связь средней величины.

Проверим гипотезу об отсутствии линейной зависимости между величинами X и Y с помощью критерия T (1.99). По таблице квантилей распределения Стьюдента находим $t_{0,975}(7) = 2{,}365$.

$$T = \frac{t_{1-\alpha/2}(n-2)}{\sqrt{n-2+t^2_{1-\alpha/2}(n-2)}} = \frac{2{,}365}{\sqrt{7+2{,}365^2}} = 0{,}666.$$

Так как $|r| = 0{,}746 > 0{,}666$, принимаем гипотезу о существовании линейной зависимости между величинами X и Y.

Полученные результаты позволяют сделать вывод, что с увеличением одной из величин среднее значение другой величины уменьшается. Так как коэффициент корреляции значим, можно пользоваться уравнениями выборочных прямых регрессии для предсказания среднего значения одной переменной по значению другой.

Задача 8.

В таблице представлены экспериментальные данные зависимости y от x. Экспериментальные значения Y являются независимыми и равноточными. По отдельной независимой серии измерений получена несмещенная оценка дисперсии $S^2 = 0{,}32$ с числом степеней свободы k равным 20. Построить линейную модель регрессии. Проверить адекватность линейной модели регрессии с уровнем значимости $\alpha = 0{,}05$.

Таблица 1.11

Условие задачи 6 и результаты расчета

и	x	Y	$X = \frac{x-0{,}6}{0{,}2}$	$X \cdot Y$	X^2	$Y_{лин}$	$\Delta Y_{лин}$	$\Delta Y_{лин}^2$
	0,2	4,5	–2	–9,0	4	5,3	–0,8	0,64
	0,4	7,0	–1	–7,0	1	6,25	0,75	0,56
	0,6	8,0	0	0,0	0	7,2	0,8	0,64
	0,8	7,5	1	7,5	1	8,15	–0,65	0,42
	1,0	9,0	2	18,0	4	9,1	–0,1	0,04
оΣ	3,0	36,0	0	9,5	10		0	2,30

Решение:

Сначала найдем решение задачи регрессии в кодированных значениях переменной x (1.71). Введем новую переменную по формуле $X = \frac{x-\bar{x}}{h}$, где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\cdot 3 = 0{,}6.$ Если значения величины x – 0,6 поделить на число h = 0,2, то получатся целые значения, не имеющие общего множителя. Поэтому $X = \dfrac{x-0,6}{0,2}$.

Находим оценки коэффициентов линейной регрессии

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{36}{5} = 7,2; \qquad \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} = \frac{9,5}{10} = 0,95.$$

Получили линейную модель регрессии

$$\widehat{Y}_{лин} = 7,2 + 0,95X.$$

По полученной формуле вычисляем значения линейной функции регрессии $\widehat{Y}_{лин}$ при всех значениях аргумента X, а затем рассчитываем $\Delta Y_i = \widehat{Y}_i - \widehat{Y}_{i\,лин}$ отклонения экспериментальных значений Y_i от значений $\widehat{Y}_{i\,лин}$, полученных по функции регрессии. Контроль, согласно формуле $\sum_{i=1}^{n} \Delta Y_i = 0$, выполнен. Все расчеты приведены в таблице 1.11. Уравнение линейной регрессии Y от реального переменного x найдем, сделав преобразование:

$$Y_{лин} = \beta_1 + \beta_2 x = 7,2 + 0,95\frac{x-0,6}{0,2} = 4,35 + 4,75x.$$

Для проверки адекватности регрессионной модели вычисляют остаточную дисперсию (так называемую дисперсию адекватности) по формуле

$$S^2_{ад} = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{k_{ад}}; \quad k_{ад} = n - m, \tag{1.66}$$

где ΔY_i – отклонения средних Y_i от проверяемой модели регрессии; $k_{ад}$ – число степеней свободы дисперсии адекватности; n – число точек, в которых проводился эксперимент; m – число оцениваемых параметров β_j в проверяемой модели.

Если истинная функция регрессии имеет тот же вид, что и рассматриваемая модель (например, так же, как и модель, представляет собой квадратичную функцию), то дисперсия адекватности служит несмещенной оценкой истинной дисперсии эксперимента и ее можно сравнивать с другими подобными оценками. В частности, может быть проведена независимая серия измерений для получения оценки дисперсии эксперимента $S^2_{экс}$. В этом случае $S^2_{экс}$ оценивает дисперсию эксперимента $D_{экс}$, $S^2_{ад}$ характеризует степень отклонения экспериментальных точек от регрессионной модели, т.е. оценивает некую дисперсию адекватности $D_{ад}$. Проверка адекватности модели заключается в проверке гипотезы H_0: $D_{ад} = D_{экс}$ при альтернативной гипотезе H_1: $D_{ад} > D_{экс}$ (если модель неадекватна, отклонения экспериментальных точек от модели будут больше погрешностей эксперимента). Таким образом, задача сводится к проверке гипотезы о равенстве дисперсий, которая решается с помощью критерия Фишера. Вычисляем отношение

$$F = S^2_{ад} / S^2_{экс}. \tag{1.67}$$

Если при заданном уровне значимости α отношение F окажется меньше квантили $F_{1-\alpha}(k_1, k_2)$, где $k_1 = k_{ад}$, $k_2 = k_{экс}$, то рассматриваемая модель не противоречит результатам эксперимента и принимается; в противоположном случае модель отвергается с уровнем значимости α, как противоречащая результатам эксперимента.

Для нахождения дисперсии адекватности (1.66) необходимо вычислить сумму квадратов отклонений результатов эксперимента от функции регрессии $\sum \Delta Y^2$. Для линейной модели эта сумма равна 2,3 (см. последний столбец таблицы 1.11). Число точек, в которых проводился эксперимент, $n = 5$; число оцениваемых параметров $m = 2$, тогда $k_{ад} = 5 - 2 = 3$. Дисперсия адекватности (1.74) равна $S^2_{ад\ лин} = 2{,}3/3 = 0{,}767$. Для проверки гипотезы об адекватности линейной модели вычисляем критерий Фишера (1.67):

$$F = S^2_{ад} / S^2_{экс} = 0{,}767/0{,}32 = 2{,}40.$$

Квантиль распределения Фишера $F_{0,95}(3;\,20) = 3{,}10$. Так как $2{,}4 < 3{,}1$, гипотеза об адекватности линейной модели принимается с уровнем значимости $\alpha = 0{,}05$.