



**ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Высшая Школа Экономики
Нижегородский филиал**

Кафедра математики

**В. В. Аниковский, Д.Е. Воронцов, В.П. Морозов,
О.М. Сольчева, В.В. Тютин**

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

**Нижний Новгород
2009**

В.В. Аниковский, Д.Е. Воронцов, В.П. Морозов, О.М. Солычева, В.В. Тютин.

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Нижний Новгород: НФ ГУ ВШЭ, 2009, 59 с.

Работа обсуждена на заседании кафедры математики 29.06.2009, протокол № 7, и одобрена Учебно-методическим советом НФ ГУ-ВШЭ.

Данная работа представляет собой учебно-методическую разработку по курсу математической статистики. Работа включает основы теории, примеры заданий (с разобранными решениями) и практические задания по 7-ми темам: «Стандартные распределения», «Вариационный ряд и порядковые статистики», «Точечные оценки», «Интервальные оценки», «Выборочные характеристики», «Проверка статистических гипотез», «Корреляция». В конце работы приведены в виде таблиц стандартные распределения: Нормализованное, Стьюдента, Фишера, Хи-квадрат. Предложенный материал рассчитан на студентов НФ ГУ-ВШЭ.

I. РАСПРЕДЕЛЕНИЯ.

1.1. Распределение Бернулли. $x : b(1, p)$

$p(x) = p^x(1-p)^{1-x}$; $x=1$, если в единичном испытании по схеме Бернулли происходит событие A и $x=0$, если происходит $\bar{A}$.

$$MX = p; DX = p(1-p)$$

1.2. Биноминальное распределение. $k : b(n, p)$

$$p_n(k) = \frac{n!}{k!(n-k)!} \cdot p^k \cdot q^{n-k} \quad (\text{где } q=1-p) - \text{вероятность того, что в } n$$

испытаниях по схеме Бернулли события A произойдет ровно k раз.

$$MK = np; DK = npq; b = \frac{q-p}{\sqrt{npq}} - \text{асимметрия}; g = \frac{1-6pq}{npq} - \text{эксцесс}$$

1.3. Отрицательное биномиальное распределение $n : \bar{b}(r, p)$

$$p_r(n) = \frac{(n-1)!}{(r-1)!(n-r)!} \cdot p^r \cdot (1-p)^{n-r}, \quad (\text{где } n=r, r+1, \dots) - \text{вероятность того, что}$$

r -ому успеху (событие A в схеме Бернулли наступает r раз) соответствует n испытаний.

$$MN = \frac{r}{p}; DN = \frac{r(1-p)}{p^2}$$

1.4. Геометрическое распределение $n : \bar{b}(1, p)$, если в 1.3. положить $r=1$,

получим $p(n) = p \cdot (1-p)^{n-1}$ $n=1, 2, \dots$ - вероятность того, что событие A наступает при n -ом испытании.

1.5. Полиномиальное распределение $(k_1, k_2, \dots, k_s) : b(n, p_1, p_2, \dots, p_s)$.

$$p_n(k_1, k_2, \dots, k_s) = \frac{n!}{k_1! k_2! \dots k_s!} \cdot p_1^{k_1} \cdot p_2^{k_2} \cdot \dots \cdot p_s^{k_s};$$

$(k_1 + k_2 + \dots + k_s = n; p_1 + p_2 + \dots + p_s = 1)$; - вероятность того, что в расширенной схеме Бернулли в n испытаниях событие A_1 поступает k_1 раз, $\dots$ $A_s - k_s$ раз.

$$MK_i = np_i; D(k_i) = np_i(1-p_i); \text{cov}(k_i, k_j) = -np_i p_j, \quad (i \neq j)$$

1.6. Гипергеометрическое распределение. $m : HG(N, M, n)$.

$$P_{N, M, n}(m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}, Mm = n \frac{M}{N}; Dm = n \frac{M}{N-1} \left(1 - \frac{M}{N}\right) \left(1 - \frac{n}{N}\right)$$

1.7. Распределение Пуассона. $k : \Pi(l)$

$$p(k) = e^{-l} \frac{l^k}{k!} \quad (\text{где } k=0, 1, \dots). MK = l; DK = l; b = \frac{1}{\sqrt{l}}; g = \frac{1}{l}$$

1.8. Нормальное распределение. $x : N(a, s)$.

$$f(x) = \frac{1}{\sqrt{2ps^2}} \cdot e^{-\frac{(x-a)^2}{2s^2}}, \quad (-\infty < x < \infty);$$

$$F(x) = F_N\left(\frac{x-a}{s}\right) - \text{функция распределения, где } F_N(z) = \frac{1}{\sqrt{2p}} \cdot \int_{-\infty}^z e^{-t^2/2} dt,$$

$$MX = a; DX = s^2; b = 0; g = 0.$$

1.9. Логарифмически-нормальное распределение. $x : LN(a, s)$

$$f(x) = \frac{1}{\sqrt{2ps^2}} \cdot \frac{1}{x} \cdot e^{-\frac{(\ln x - a)^2}{2s^2}}, \quad (x > 0);$$

$$F(x) = F_N\left(\frac{\ln(x) - a}{s}\right) - \text{функция распределения,}$$

$$MX = e^{a+s^2/2}, \quad DX = e^{2a+s^2} \cdot (e^{s^2} - 1),$$

$$b = (e^{s^2} - 1)^{\frac{1}{2}} \cdot (e^{s^2} + 2) > 0,$$

$$g = (e^{s^2} - 1) \cdot (e^{3s^2} + 3e^{2s^2} + 6e^{s^2} + 6) > 0$$

1.10. Равномерное распределение $x : U(a, b)$

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b) \\ 0, & x \notin (a, b) \end{cases},$$

$$MX = \frac{a+b}{2}; \quad DX = \frac{(b-a)^2}{12}; \quad b = 0; \quad g = -1, 2;$$

1.11. Распределение Вейбулла. $x : V(a, a)$

$$f(x) = \frac{a}{x} \cdot x^{a-1} \cdot e^{-x^a/a} \quad (x > 0),$$

$$MX = a^{1/a} \cdot \Gamma(1 + \frac{1}{a}); \quad DX = a^{2/a} \cdot \left[\Gamma(1 + \frac{2}{a}) - \Gamma^2(1 + \frac{1}{a}) \right]$$

1.12. Показательное распределение. $x : P(b, a)$

$$f(x) = \frac{1}{a} \cdot e^{-\frac{(x-b)}{a}} \quad (x > b),$$

$$MX = a + b; \quad DX = a^2; \quad b = 2; \quad g = 6$$

1.13. Распределение Лапласа. $x : L(a)$

$$f(x) = \frac{1}{2a} \cdot e^{-|x|/a} \quad (-\infty < x < \infty),$$

$$MX = 0; DX = 2a^2; b = 0; g = 3$$

1.14. Распределение Парето.

$$a) f(x) = \frac{1}{a} x^{-\frac{1}{a}-1} \quad (x > 1),$$

$$б) f(x) = \frac{1}{a} x^{\frac{1}{a}-1} \quad (0 < x < 1)$$

1.15. Гамма-распределение. $x : \Gamma_{m,r}$

$$f(x) = \frac{x^{m-1} \cdot e^{-x/b}}{\Gamma(m) \cdot b^m} \quad (x > 0),$$

$$MX = mb; DX = mb^2; b = \frac{2}{\sqrt{m}}; g = \frac{6}{m}$$

$$\Gamma(m) = \int_0^{\infty} x^{m-1} \cdot e^{-x} dx - \text{Гамма функция Эйлера}$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \quad \Gamma(1) = 1; \quad \Gamma(m+1) = m\Gamma(m)$$

1.16. Бета-распределение. $x : b(m_1, m_2)$

$$f(x) = \frac{\Gamma(m_1 + m_2)}{\Gamma(m_1) \cdot \Gamma(m_2)} \cdot x^{m_1-1} \cdot (1-x)^{m_2-1}, \quad (0 < x < 1),$$

$$MX = \frac{m_1}{m_1 + m_2}; DX = \frac{m_1 \cdot m_2}{(m_1 + m_2)^2 \cdot (m_1 + m_2 + 1)},$$

$$\int_0^1 x^{m_1-1} \cdot (1-x)^{m_2-1} dx = \frac{\Gamma(m_1) \cdot \Gamma(m_2)}{\Gamma(m_1 + m_2)} = B(m_1, m_2) - \text{Бета-функция Эйлера.}$$

1.17. “Хи – квадрат” - распределение с n степенями свободы. $U : \chi_n^2$.

Если x_i ($i=1, 2, \dots, n$) независимы и $x_i : N(0;1)$, то $U = \sum_{i=1}^n x_i^2$ имеет “Хи –

квадрат” – распределение с n степенями свободы.

$$f_{\chi_n^2}(u) = \frac{U^{(n/2)-1}}{2^{n/2} \cdot \Gamma(\frac{n}{2})} \cdot e^{-u/2}, \quad (n > 0), \quad MU = n; \quad DU = 2n; \quad b = \frac{2^{3/2}}{\sqrt{n}}; \quad g = \frac{12}{n}$$

1.18. Распределение Стьюдента с n степенями свободы. $t : t(n) = f_{st}^{(n)}(t)$

Если $z : N(0,1)$, $U : c_n^2$, то $t = \frac{z}{\sqrt{U/n}} \sim t(n)$, z и U независимы.

$$f_{t(n)}(t) = \frac{\Gamma(\frac{n+2}{2})}{\Gamma(n/2) \cdot \sqrt{pn}} \cdot \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad (-\infty < t < \infty),$$

$$M t = 0; \quad D t = \frac{n}{n-2} \quad (n > 2); \quad b = 0; \quad g = \frac{6}{n-4} \quad (n > 4)$$

1.19. Распределение Фишера. $F : F^{(m,n)} = f^{(n,m)}(F)$

Если U и V независимы и $U : c_m^2$, $V : c_n^2$, то $F = \frac{U/m}{V/n} : F^{(m,n)}$,

$$f_F^{(n,m)}(F) = \left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2}) \cdot \Gamma(\frac{n}{2})} \cdot \left(1 + \frac{m}{n} \cdot F\right)^{-\frac{n+m}{2}}, \quad (F > 0),$$

$$MF = \frac{n}{m-2} \quad (m > 2); \quad DF = \frac{2n^2(n+m-2)}{m(n-2)^2(n-4)} \quad (n > 4).$$

$$f^{(n,m)}\left(\frac{1}{F}\right) = f^{(n,m)}(F) \text{ - свойство симметрии.}$$

Если $F_{kp}^{(m,n)}(a)$ есть решение уравнения

$$\int_{F_{kp}^{(m,n)}(a)}^{\infty} f^{(m,n)}(F) \cdot dF = a, \quad \text{то } F_{kp}^{(m,n)}(1-a) = \frac{1}{F_{kp}^{(m,n)}(a)}$$

1.20. Двумерное нормальное распределение.

$$f_{X_1 X_2}(x_1, x_2) = \frac{1}{2\pi s_1 s_2 \sqrt{1-r^2}} e^{\left\{ -\frac{1}{2(1-r^2)} \left[\left(\frac{x_1-a_1}{s_1} \right)^2 + \left(\frac{x_2-a_2}{s_2} \right)^2 - 2r \frac{x_1-a_1}{s_1} \frac{x_2-a_2}{s_2} \right] \right\}},$$

$$MX_1 = a_1; \quad MX_2 = a_2; \quad DX_1 = s_1^2; \quad DX_2 = s_2^2; \quad r = \frac{\text{cov}(x_1 x_2)}{s_1 s_2},$$

$$x_1 : N(a_1, s_1); \quad x_2 : N(a_2, s_2)$$

II. ВАРИАЦИОННЫЙ РЯД И ПОРЯДКОВЫЕ СТАТИСТИКИ.

Генеральная совокупность: Множество всех возможных значений признака X исследуемого объекта.

Выборка объема n : последовательность n независимых наблюдений $x_1, x_2, \dots, x_n$ на основании которых строятся выводы о свойствах генеральной совокупности. Здесь мы будем предполагать, что случайная величина X непрерывна, и имеет функцию распределения $F(x)$ и плотность вероятность $f(x)$.

Очевидно, что $f_{X_i}(x) = f(x)$, т.е. закон распределения возможных значений x_i при i -ом наблюдении одинаков для всех i и совпадает с $f(x)$. Кроме того, $X_1, X_2, \dots, X_n$ независимы.

Вариационный ряд. Пусть $x_1, x_2, \dots, x_n$ - выборка из объема n . Построим новый ряд из элементов исходной выборки расположив x_i в порядке неубывания, обозначим член нового ряда, стоящий на i -ом месте через $x_{(i)}$. Тогда ряд будет представлен следующей последовательностью:

$$x_{(1)}, x_{(2)}, \dots, x_{(n)} \quad (2.1.)$$

Каждый член этого ряда называется порядковой статистикой, а сама последовательность (2.1.) вариационным рядом. Члены вариационного ряда уже не являются взаимно независимыми и их распределения не являются одинаковыми.

Закон распределения i -го члена вариационного ряда:

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} \cdot F^{i-1}(x) g(1-F(x))^{n-i} g f(x) \quad (2.2.)$$

Здесь $f(x)$ и $F(x)$ - плотность и функция распределения генеральной совокупности. В частности:

$$f_{X_{(1)}}(x) = n \cdot (1-F(x))^{n-1} \cdot f(x) \quad (2.3.)$$

$$f_{X_{(n)}}(x) = n \cdot F^{n-1}(x) \cdot f(x) \quad (2.4.)$$

Совместные распределения членов вариационного ряда.

Ниже приводится выражение для совместной плотности вероятности

$f_{X_{(i)}, X_{(j)}}(x, y)$ для любой пары порядковых статистик $(x_{(i)}, x_{(j)})$, где $i < j$:

$$f_{X_{(i)}, X_{(j)}}(x, y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F^{i-1}(x) [F(y)-F(x)]^{j-i-1} [1-F(y)]^{n-j} f(x) f(y) \quad (x \leq y) \quad (2.5.)$$

$$\text{В частности: } f_{X_{(i)}, X_{(n)}}(x, y) = n(n-1) \cdot [F(y)-F(x)]^{n-2} \cdot f(x) \cdot f(y) \quad (2.6.)$$

Задачи.

2.1. $x : U(0, b)$ (см. 1.10.). Найти:

а) $f_{X_{(1)}}(x)$, б) $f_{X_{(n)}}(x)$, в) $MX_{(1)}$, $DX_{(1)}$, г) $MX_{(n)}$, $DX_{(n)}$, д) $\text{cov}(x_{(1)}, x_{(n)})$.

2.2. Для распределения Парето (см. 1.14.а) Найти:

а) $f_{X_{(1)}}(x)$, б) $f_{X_{(n)}}(x)$, в) $MX_{(1)}$, $DX_{(1)}$, г) $MX_{(n)}$, $DX_{(n)}$.

2.3. Для распределения Парето (см. 1.14.б) Найти те же величины, что и в задаче 2.2.

2.4. $x : U(a, b)$ (см. 1.10.). Найти те же величины, что и в задаче 2.1.

2.5. $f(x) = e^{-(x-b)}$, ($x > b$), ($x : P(b, 1)$ (см. 1.12.))

Найти: а) $f_{X_{(1)}}(x)$, б) $MX_{(1)}$, в) $DX_{(1)}$

2.6. Для распределений (2.3.), (2.4.), (2.6.) проверить условие нормировки.

III. ТОЧЕЧНЫЕ ОЦЕНКИ.

Пусть генеральная случайная величина x имеет плотность распределения $f(x, q_1, \dots, q_s)$ в непрерывном случае или $p(x, q_1, \dots, q_s)$ в дискретном случае. Предположим, что аналитический вид закона распределения известен, но не известны параметры $q_1, \dots, q_s$. Задача состоит в том, чтобы на основании выборки $x_1, \dots, x_n$ объема n найти приближенные (точечные) значения неизвестных параметров $q_i (i = 1, 2, \dots, S)$. Полученные тем или иным образом оценки будем обозначать символом $\hat{q}_i^{\mathbf{u}}$. Очевидно, что точечные оценки зависят от результатов наблюдений, т.е. $\hat{q}_i^{\mathbf{u}} = \hat{q}_i^{\mathbf{u}}(x_1, \dots, x_n)$. Поскольку результаты наблюдений случайны, то и $\hat{q}_i^{\mathbf{u}}$ - случайные величины. В этом случае можно писать $\hat{q}_i^{\mathbf{u}} = q_i(X_1, \dots, X_n)$, где X_i - случайная величина, соответствующая i -ому наблюдению. При этом совместное распределение последовательности случайных величин $X_1, X_2, \dots, X_n$ определяется выражением:

$$f_{X_1, X_2, \dots, X_n}(x'_1, x'_2, \dots, x'_n) = \prod_{i=1}^n f(x'_i; q_1, \dots, q_s) \quad (3.2.)$$

в непрерывном случае и выражением:

$$P_{X_1, X_2, \dots, X_n}(x'_1, x'_2, \dots, x'_n) = \prod_{i=1}^n P(x'_i; q_1, \dots, q_s) \quad (3.3.)$$

в дискретном случае, где x'_i - возможные значения x_i в i -ом наблюдении.

Желательно, чтобы точечные оценки удовлетворяли условиям несмещённости, состоятельности, эффективности. Заметим, что удовлетворить всем трем условиям удастся не всегда.

Несмещенность.

Оценка называемая несмещенной, или

$$M\hat{q}_i^{\mathbf{H}} = q_i \quad (3.4.)$$

Где математическое ожидание вычисляется с помощью (3.2.) или (3.3.)

Состоятельность.

Оценка состоятельна, если

$$\lim_{n \rightarrow \infty} P\left(\left|\hat{q}_i^{\mathbf{H}} - q_i\right| \geq \varepsilon\right) = 0, \text{ т.е. } \hat{q}_i^{\mathbf{H}} \xrightarrow{P} q_i \quad (3.5.)$$

Достаточным условием состоятельности является условие

$$\lim_{n \rightarrow \infty} D\hat{q}_i^{\mathbf{H}} = 0 \quad (3.6.)$$

Эффективность.

Оценка эффективна, если её дисперсия наименьшая в классе несмещенных оценок. Часто, но не всегда качество оценки в смысле её эффективности определяется величиной (эффективностью)

$$e(\hat{q}_i^{\mathbf{H}}) = \frac{1}{ngD(\hat{q}_i^{\mathbf{H}})g(J^{-1})_{ij}} \quad (3.7.)$$

$$\text{Где } J_{ij} = -M \frac{\partial^2 \ln f}{\partial q_i \cdot \partial q_j}, \text{ (в том числе } J_{ii} = -M \frac{\partial^2 \ln f}{\partial q_i^2}). \quad (3.8.)$$

$$\text{Если информационная матрица диагональна, то } e(\hat{q}_i^{\mathbf{H}}) = \frac{1}{ngD(\hat{q}_i^{\mathbf{H}})gJ_{ii}}.$$

информационная матрица Фишера, соответствующая одному наблюдению.

Выражения (3.7.) и (3.8.) справедливы при условиях: а) J_{ij} - положительно

определенная матрица; б) Операцию взятия производной по параметру можно вносить под знак математического ожидания; в) q_i - не является границей области определения $f(x)$ или $p(x)$.

Оценка считается эффективной, если $e(\hat{q}_i^{\mathbf{H}}) = 1$

Метод наибольшего правдоподобия.

Для нахождения на основе выборки $x_1, \dots, x_n$ точечной оценки можно использовать метод наибольшего правдоподобия. Для этого строится функция правдоподобия:

$$L = \prod_{i=1}^n f(x_i; q_1, \dots, q_s) \quad \text{или} \quad (3.9.)$$
$$L = \prod_{i=1}^n P(x_i; q_1, \dots, q_s)$$

Функция правдоподобия – это распределения (3.2.) и (3.3.) для конкретной выборки, т.е. когда $x'_i = x_i$

Неизвестные параметры q_i должны быть такими, чтобы L была наибольшей.

Как правило, речь идет о нахождении таких q_i , чтобы L было максимальным:

$$\frac{\partial L}{\partial q_i} = 0 \quad \text{или} \quad \frac{\partial \ln L}{\partial q_i} = 0 \quad (i = 1, 2, \dots, S) \quad (3.10.)$$

Метод моментов.

В этом случае точечные оценки являются решением системы уравнений:

$$n_k = \mu_k \quad (k = 1, 2, \dots, S) \quad (3.11.)$$

Где $n_k = Mx^k$ - начальные моменты, вычисленные с помощью исходного распределения $f(x, q_1, \dots, q_s)$ или, а μ_k - выборочные начальные моменты порядка i :

$$\mu_i = \frac{1}{n} \sum_{i=1}^n x_i^k \quad (3.12.)$$

Обычно выбирают первые S моментов, отличных от нуля.

Задачи (к разделу точечные оценки).

3.1. $x : N(a, s)$ (см. 1.8.) На основе выборки $x_1, \dots, x_n$ методом наибольшего правдоподобия получить точечные оценки параметров a и s^2 . Проверить несмещенность, состоятельность и эффективность полученных оценок $\hat{a}$ и $\hat{s}^2$. Получить несмещенную и состоятельную оценку для s^2 :

$$\hat{s}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2; \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Получить точечные оценки $\hat{a}$ и $\hat{s}^2$ методом моментов.

3.2. $x : B(1, p)$ (см. 1.1.) На основании выборки объема $n : (x_1, \dots, x_n)$ методом наибольшего правдоподобия найти $\hat{p}$. Убедиться в несмещенности, состоятельности и эффективности. Получить точечную оценку методом моментов.

3.3. $x : P(0, a)$ (см. 1.12.) Задание для a такое же как и в 3.2. для p .

3.4. $k : \Pi(l)$ (см. 1.7.) На основании выборки $k_1, \dots, k_n$. Найти методом наибольшего правдоподобия и методом моментов $\hat{l}$. Убедиться в несмещенности, состоятельности и эффективности.

В задачах 3.5. - 3.23 При заданной плотности распределения вероятности $f = f(x)$ найти: 1) A ; 2) На основании выборки объема n методом наибольшего правдоподобия получить точечную оценку $\hat{a}$ неизвестного параметра a . 3) Проверить несмещенность, состоятельность и эффективность полученной оценки. 4) Получить точечные оценки методом моментов.

$$3.5. f = A \cdot x^{\frac{1}{a}-1}, \quad (0 < x < 1)$$

$$3.6. f = A \cdot x^{-\frac{1}{a}-1}, \quad (x > 1)$$

$$3.7. f = A \cdot e^{-x^2/a}, \quad (x > 0)$$

$$3.8. f = A \cdot x^{1/2} \cdot e^{-x/a}, \quad (x > 0)$$

$$3.9. f = A \cdot x \cdot e^{-x/a}, \quad (x > 0)$$

$$3.10. f = A \cdot |x| \cdot e^{x/a}, \quad (x < 0)$$

$$3.11. f = A \cdot e^{-|x|/a}, \quad (-\infty < x < \infty)$$

$$3.12. f = A \cdot x^2 \cdot e^{-x^3/a}, \quad (x > 0)$$

$$3.13. f = A \cdot \frac{\ln x}{x^{1+\frac{1}{a}}}, \quad (x > 1)$$

$$3.14. f = A \cdot e^{-x^2/a}, \quad (-\infty < x < \infty)$$

$$3.15. f = A \cdot \frac{1}{x^3} \cdot e^{-1/(ax)}, \quad (x > 0)$$

$$3.16. f = A \cdot \frac{1}{x^2} \cdot e^{-1/(ax)}, \quad (x > 0)$$

$$3.17. f = A \cdot \frac{1}{x^{5/2}} \cdot e^{-1/(ax)}, \quad (x > 0)$$

$$3.18. f = A \cdot e^{-\text{sh } x/a} \cdot \text{ch } x, \quad (x > 0), \quad \text{где } \text{sh } x = \frac{e^x - e^{-x}}{2}; \quad \text{ch } x = \frac{e^x + e^{-x}}{2}$$

$$3.19. f = A \cdot e^{2x} \cdot e^{e^{2x}/a}, \quad (x > 0)$$

$$3.20. f = A \cdot \frac{1}{x} \cdot (\ln x)^{-\frac{1}{a}-1}, \quad (x > e)$$

$$3.21. f = A \cdot \frac{1}{x} \cdot (\ln x)^{\frac{1}{a}-1}, \quad (1 < x < e)$$

$$3.22. x : LN(a, s), \quad (\text{см. 1.9.})$$

$$3.23. P(k) = A(1 - \frac{1}{a})^{k-1}, \quad k = 1, 2, \dots, \text{выборка: } k_1, \dots, k_n$$

3.24. $x : U(0, b)$ (см. 1.10.) 1) На основании выборки объема $n : (x_1, \dots, x_n)$ методом наибольшего правдоподобия найти точечную оценку $\hat{\theta}$. 2) С помощью распределения (2.4.) найти несмещенную оценку $\hat{\theta}^*$ (оценка $\hat{\theta}$ - смещенная). 3) Убедитесь в состоятельности оценки $\hat{\theta}^*$. 4) Найти оценку для b методом моментов.

3.25. $x : U(a, b)$ (см. 1.10.) 1) На основании выборки $x_1, \dots, x_n$ методом наибольшего правдоподобия найти точечную оценку для a и b . 2) Убедитесь в том, что $\hat{a}$ и $\hat{b}$ смещенные оценки. 3) Получить несмещенные оценки $\hat{a}^*$ и $\hat{b}^*$. 4) Найти точечные оценки $\hat{a}$ и $\hat{b}$ методом моментов.

3.26. $f(x) = e^{-(x-a)}, (x > a)$ 1) На основании выборки объема $n : (x_1, \dots, x_n)$ методом наибольшего правдоподобия найти $\hat{a}$. 2) Убедитесь в том, что оценка $\hat{a}$ смещенная. 3) Найти несмещенную оценку $\hat{a}^*$ и убедиться в том, что она состоятельна.

3.27. $F(x) = 0$ при $x \leq 0$; $F(x) = x^2/a^2$ при $0 < x \leq a$; $F(x) = 1$ при $x > a$. 1) На основании выборки объема $n : (x_1, \dots, x_n)$ методом наибольшего правдоподобия найти $\hat{a}$. Используя распределение (2.4.) проверить несмещенность полученной оценки. Если оценка смещенная, найти $\hat{a}^*$ - несмещенную оценку.

3.28. $F(x) = 0$ при $x \leq a$; $F(x) = 1 - \frac{a}{x}$ при $x > a$; 1) На основании выборки объема $n : (x_1, x_2, \dots, x_n)$ методом наибольшего правдоподобия найти $\hat{a}$. Используя распределение (2.3.) проверить несмещенность полученной оценки. Если оценка смещенная, найти $\hat{a}^*$ - несмещенную оценку.

3.29. Пусть в распределении $f(x, q)$ параметр $q = MX$ (q - математическое

ожидание случайной величины x). Показать, что оценка $\hat{q} = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$

является несмещенной и состоятельной оценкой (при существовании конечной дисперсии $s^2 = DX$).

IV. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ.

Нормальное распределение $X : N(a, s)$

$$\bar{x} - t_{кр} \cdot \frac{s}{\sqrt{n}} < a < \bar{x} + t_{кр} \cdot \frac{s}{\sqrt{n}}, \quad (4.1)$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ – (выборочное среднее) точечная несмещенная оценка

параметра a ; $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ – (исправленная выборочная дисперсия)

точечная несмещенная оценка дисперсии s^2 .

Параметр $t_{кр}$ находится по таблице распределения Стьюдента при заданном числе степеней свободы $(n-1)$ из соотношения

$$\int_{t_{кр}}^{\infty} f_{st}^{(n-1)}(t) \cdot dt = \frac{1-g}{2}. \quad (4.2)$$

Здесь и ниже g – доверительная вероятность (надежность), которая задается самим исследователем или государственным стандартом. Интервал типа (4.1) – доверительный интервал, который с доверительной вероятностью g содержит точное значение параметра a .

$$\sqrt{\frac{s^2(n-1)}{U_2}} < s < \sqrt{\frac{s^2(n-1)}{U_1}} \quad - \quad (4.3)$$

доверительный интервал для оценки параметра s нормального распределения. Критические значения U_1 и U_2 в (4.3) находятся по таблицам распределения χ^2_{n-1} в соответствии с формулами

$$\int_{U_2}^{\infty} f_{\chi^2_{n-1}}(u) \cdot du = \frac{1-g}{2}; \quad \int_{U_1}^{\infty} f_{\chi^2_{n-1}}(u) \cdot du = \frac{1+g}{2} \quad (4.4)$$

Другие распределения:

Если $\hat{q}$ – несмещенная оценка неизвестного параметра q и $D\hat{q} = \frac{c^2(q)}{n}$

(что, как правило, выполняется при достаточно больших n), то справедливо асимптотическая интервальная (при $n > 50$) оценка для q :

$$\hat{q} - z_{кр} \cdot \frac{c(\hat{q})}{\sqrt{n}} < q < \hat{q} + z_{кр} \cdot \frac{c(\hat{q})}{\sqrt{n}}, \quad (4.5)$$

где $z_{кр}$ находится из уравнения $F_N(z_{кр}) = \frac{1+g}{2}$ (4.6)

В частности:

Распределение Бернулли $X : b(1, n)$ (см. 1.1):

$$\bar{x} - z_{кр} \cdot \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n}} < p < \bar{x} + z_{кр} \cdot \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n}}, \quad (4.7)$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ – точечная несмещенная оценка параметра p .

Распределение Пуассона $k : \Pi(l)$ (см. 1.7):

$$\bar{k} - z_{кр} \cdot \frac{\sqrt{\bar{k}}}{\sqrt{n}} < l < \bar{k} + z_{кр} \cdot \frac{\sqrt{\bar{k}}}{\sqrt{n}}, \quad (4.8)$$

где $\bar{k} = \frac{1}{n} \sum_{i=1}^n k_i$ – точечная несмещенная оценка параметра l .

Показательное распределение $X : P(0, a)$ (см. 1.12):

$$\bar{x} - z_{кр} \cdot \frac{\bar{x}}{\sqrt{n}} < a < \bar{x} + z_{кр} \cdot \frac{\bar{x}}{\sqrt{n}}, \quad (4.9)$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ – точечная несмещенная оценка параметра a .

V. ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ.

Если $(x_1, x_2, \dots, x_n)$ – выборка объема n , то можно найти приближенные значения характеристик генеральной случайной величины X : математическое ожидание, дисперсию, коэффициенты асимметрии и эксцесса и т.д.

$$\mathfrak{d}_k = \frac{1}{n} \sum_{i=1}^n x_i^k - \quad (5.1)$$

выборочный начальный момент порядка k . В частности,

$$\mathfrak{d}_1 = \frac{1}{n} \sum_{i=1}^n x_i - \quad (5.2)$$

выборочное среднее.

$$\mathfrak{M}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k - \quad (5.3)$$

выборочный центральный момент порядка k . В частности,

$$\mathfrak{M}_2 = s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - \quad (5.4)$$

выборочная дисперсия.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 - \quad (5.5)$$

исправленная выборочная дисперсия, т. е. несмещенная точечная оценка дисперсии генеральной совокупности.

$$\mathfrak{B} = \frac{m_4}{\mathfrak{B}^3} - \quad (5.6)$$

выборочный коэффициент асимметрии.

$$\mathfrak{g} = \frac{m_4}{\mathfrak{B}^4} - 3 - \quad (5.7)$$

Выборочный коэффициент эксцесса.

ВЫБОРОЧНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ДИСКРЕТНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ.

Выборку объема n из генеральной совокупности представляют в виде таблицы:

x_k	x_1	x_2	$\dots$	x_r
n_k	n_1	n_2	$\dots$	n_r
$\mathfrak{P}(x_k)$	$\frac{n_1}{n}$	$\frac{n_2}{n}$		$\frac{n_r}{n}$

, где n_k - число реализаций (частота)

значений x_k , $\sum_{k=1}^r n_k = n$ (n – объем выборки).

Эмпирическая вероятность реализации значения x_k определяется

выражением (третья строка в таблице): $\mathfrak{P}(x_k) = \frac{n_k}{n}$. (5.8)

Эмпирические (выборочные) средние и дисперсия вычисляются по формулам

$$\bar{x} = \frac{1}{n} \sum_{k=1}^r n_k \cdot x_k, \quad s^2 = \frac{1}{n-1} \sum_{k=1}^r n_k \cdot (x_k - \bar{x})^2 \quad (5.9)$$

Выборочная функция распределения определяется формулой:

$$\mathfrak{P}(x) = \sum_{k: x_k < x} \mathfrak{P}(x_k) \quad (5.10)$$

Пример: Пусть $x_k = k$ - число машин, подъезжающих к автозаправке в течение часа. Наблюдения в течение 20 часов дали следующие результаты:

x_k	0	1	2	3	4	5	6
n_k	1	3	5	5	3	2	1
$\mathfrak{P}(k)$	1/20 0,05	3/20 0,15	5/20 0,25	5/20 0,25	3/20 0,15	2/20 0,1	1/20 0,05
$p^{(M)}(k)$	0,06	0,17	0,24	0,22	0,16	0,09	0,04

Здесь данные наблюдений находятся в первой и второй строках. В третьей строке – результат вычислений эмпирической вероятности $\mathfrak{P}(k)$.

Расчет с помощью формулы (3.9) дает $\bar{k} = 2,8$; $s = 1,54$.

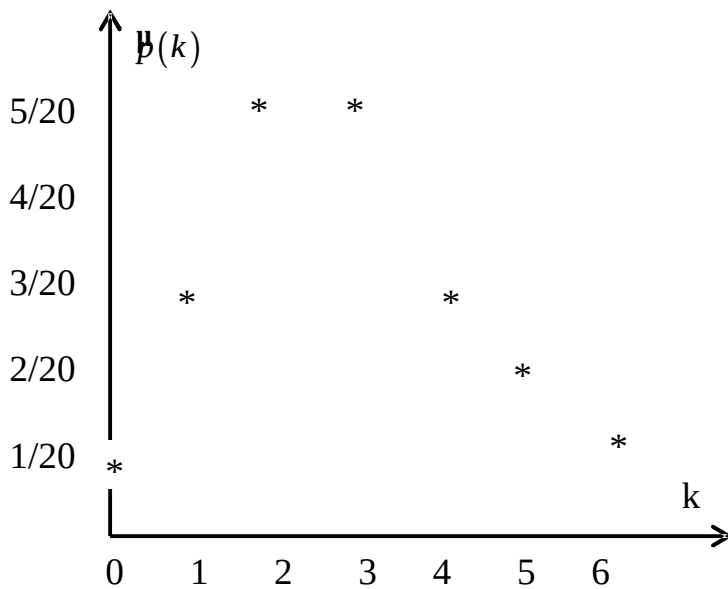


График для $\mathcal{H}(k)$ и тот факт, что величины $s = 1,54$ и

$\sqrt{k} = 1,67$ близки, позволяют сделать вывод о том, что k имеет распределение

Пуассона с параметром $\bar{k}$, т.е. модельное распределение имеет вид

$$p^{(M)}(k) = e^{-2,8} \cdot \frac{(2,8)^k}{k!}.$$

Результаты вычислений, выполненные с помощью этого распределения

приведены в 4-ой строке таблицы, из которой видно, что $\mathcal{H}(k)$ и $p^{(M)}(k)$ действительно близки. Степень близости эмпирического и модельного распределений обсуждается в разделе VI.

ВЫБОРОЧНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ НЕПРЕРЫВНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ.

Пусть $(x_1, x_2, \dots, x_n)$ - выборка объема n . Процедура обработки данных наблюдений состоит в следующем:

1. Находятся $x_{\min}$ и $x_{\max}$ - наименьшее и наибольшее значения.
2. Весь интервал полученных значений $x_{\min} \leq x \leq x_{\max}$ разбивается на r интервалов. Число интервалов при $n \gg 1$ принято (но не обязательно) находить по формуле $r \approx 1 + \log_2 n$.
3. Определяется длина каждого частичного интервала $\Delta = \frac{x_{\max} - x_{\min}}{r}$.
4. Определяются границы частичных интервалов $c_k = c_0 + k \cdot \Delta$, $(k = 1, 2, \dots, r)$. Величину Δ (или r) следует выбрать так, чтоб $c_0 \leq x_{\min} \leq x_{\max} \leq c_r$.
5. Подсчитывается число n_k тех значений x_i , которые попали в k -й полуинтервал $[c_{k-1}; c_k)$.
6. Вычисляются координаты середины каждого частичного интервала $x_k^{(0)} = \frac{c_{k-1} + c_k}{2}$, $(k = 0, 1, \dots, r)$.

Теперь данные наблюдений оформляются в виде таблицы (в 1 – 3 строки)

$c_{k-1} - c_k$	$c_0 - c_1$	$c_1 - c_2$	$c_2 - c_3$	...	$c_{r-2} - c_{r-1}$	$c_{r-1} - c_r$	
n_k	n_1	n_2	n_3	...	n_{r-1}	n_r	
$x_k^{(0)}$	$x_1^{(0)}$	$x_2^{(0)}$	$x_3^{(0)}$	...	$x_{r-1}^{(0)}$	$x_r^{(0)}$	
$\Psi(x)$	$\frac{n_1}{n \cdot \Delta}$	$\frac{n_2}{n \cdot \Delta}$	$\frac{n_3}{n \cdot \Delta}$	...	$\frac{n_{r-1}}{n \cdot \Delta}$	$\frac{n_r}{n \cdot \Delta}$	
$\mathbf{P}^H(x)$	0	$\frac{n_1}{n}$	$\frac{n_1 + n_2}{n}$	...	$\frac{n_1 + n_2 + \dots + n_{r-2}}{n}$	$\frac{n_1 + n_2 + \dots + n_{r-1}}{n}$	1

С помощью этой таблицы находим эмпирическую плотность распределения вероятностей (4-я строка)

$$\Psi(x) = \frac{n_k}{n \cdot \Delta} \text{ при } x \in (c_{k-1}; c_k) \quad (5.11)$$

График этой плотности имеет ступенчатый вид и называется **гистограммой**. Кроме того находится накопительная (кумулятивная) функция распределения

(5-я строка). $\mathbf{P}^H(x) = 0$ при $x \leq c_1$, $\mathbf{P}^H(x) = \frac{n_1 + n_2 + \dots + n_{k-1}}{n}$ при $c_{k-1} < x \leq c_k$

($k=2,3,\dots,r$) и $\mathbf{P}^H(x) = 1$ при $x > c_r$.

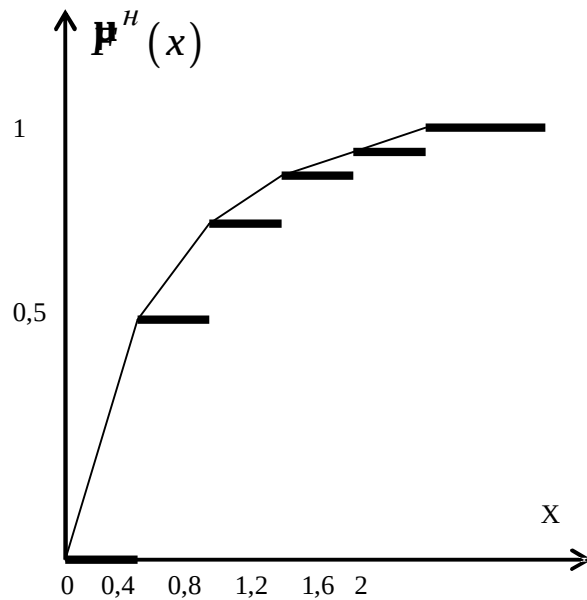
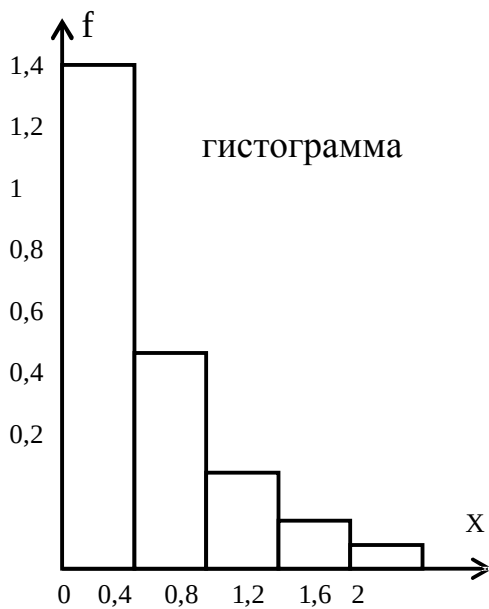
Выборочные среднее и дисперсия вычисляются по формулам:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^r n_k \cdot x_k^{(0)}, \quad s^2 = \frac{1}{n-1} \sum_{k=1}^r n_k \cdot \left(x_k^{(0)} - \bar{x} \right)^2 \quad (5.12)$$

Пример. Пусть x – время техосмотра 1-й автомашины. Наблюдения проводились над 100 автомашинами. Данные наблюдений обработаны и приведены в таблице (1-я и 2-я строки)

$c_{k-1} - c_k$	0 – 0,4	0,4 – 0,8	0,8 – 1,2	1,2 – 1,6	1,6 – 2	
n_k	56	25	12	5	2	
$x_k^{(0)}$	0,2	0,6	1	1,4	1,8	
$\Psi(x)$	1,4	0,625	0,3	0,125	0,05	
$\mathbf{P}^H(x)$	0	0,56	0,81	0,93	0,98	1
$\mathbf{p}_k$	0,56	0,25	0,12	0,05	0,02	
$\mathbf{p}_k^{(M)}$	0,559	0,247	0,108	0,048	0,021	

Здесь $\Delta = 0,4$; $r = 5$; $n = 100$



Тонкая линия на графике $F^H(x)$ является выборочной (наблюдаемой) функцией распределения $F(x)$, соответствующей указанной гистограмме:

$$F(x) = \int_{-\infty}^x f(x') \cdot dx'.$$

В шестой строке таблицы – эмпирическая вероятность попадания времени техосмотра в k -й интервал $p_k = \frac{n_k}{n}$.

Вычисления с помощью формул (5.12) дают $\bar{x} = 0,488$, $s = 0,398$

Вид гистограммы и тот факт, что $\bar{x}$ и s близки, дают основание считать, что X имеет показательное распределение $X : P(0, \bar{x})$. Т.е.

$$f^{(M)}(x) = \frac{1}{0,488} \cdot e^{-x/0,488}.$$

С помощью этого распределения вычислены $p_k^{(M)}$ – модельные вероятности

$$\text{попадания } x \text{ в } k\text{-й интервал } p_k^{(M)} = \int_{c_{k-1}}^{c_k} f^{(M)}(x) \cdot dx = e^{-c_{k-1}/0,488} - e^{-c_k/0,488}.$$

Задачи к разделам IV и V.

4.1. Для распределения Парето $f(x) = \frac{1}{a} \cdot x^{-\frac{1}{a}-1}$, ($x > 1$) найти общее выражение для построения доверительного интервала.

4.2. Для распределения Парето $f(x) = \frac{1}{a} \cdot x^{\frac{1}{a}-1}$, ($0 < x < 1$) найти общее выражение для построения доверительного интервала.

В задачах 4.3. – 4.8. найти: 1) A , 2) точечную оценку параметра a , 3) интервальную оценку параметра a (используя формулу (4.5.))

4.3. $f = A \cdot x \cdot e^{-x^2/a}$, ($x > 0$)

4.4. $f = A \cdot x^{1/2} \cdot e^{-x/a}$, ($x > 0$)

4.5. $f = A \cdot x \cdot e^{-x/a}$, ($x > 0$)

4.6. $f = A \cdot |x| \cdot e^{x/a}$, ($x < 0$)

4.7. $f = A \cdot e^{-|x|/a}$, ($-\infty < x < \infty$)

4.8. $f = A \cdot x^2 \cdot e^{-x^3/a}$, ($x > 0$)

4.9. $p(k) = \frac{1}{a} \left(1 - \frac{1}{a}\right)^{k-1}$, ($k = 1, 2, \dots$). На основании выборки объема n

($k_1, k_2, \dots, k_n$) получить точечную оценку параметра a . Найти доверительный интервал для оценки a , используя формулу (4.5)

4.10. Из 100 опрошенных покупателей 70 оказались довольны уровнем обслуживания в магазине. А) Найти 95% доверительный интервал для оценки доли покупателей, довольных уровнем обслуживания в магазине. Б) Найти минимальный объем выборки, при котором с вероятностью 0,95 точечная оценка доли покупателей, довольных обслуживанием, отличается от истинного значения этой доли не более чем на 0,06.

4.11. На примере Вашей группы найти 95% доверительный интервал для оценки доли студентов, получивших по матанализу оценку более 5 баллов.

4.12. Из очень большого ящика с микрочипами вынимают 100 чипов. Из них 2 оказались бракованными. А) Найти 98% доверительный интервал для оценки доли бракованных чипов. Б) Каков минимальный объем выборки, при котором с той же доверительной вероятностью отклонение точечной оценки доли бракованных чипов от истинного значения не превышает 0,02.

4.13. Средняя дневная выручка автостоянки по результатам 20 наблюдений равна 5000 руб. Исправленная выборочная дисперсия равна 1000000 кв. руб. А) Найти 90% доверительный интервал для оценки средней дневной выручки (считая распределение выручки $X : N(a, s)$). Б) Найти 98% доверительный интервал для оценки среднеквадратического отклонения.

4.14. Дневная выручка X небольшого киоска имеет нормальное распределение ($X : N(a, s)$). Данные наблюдений (в тыс. руб.) за 25 дней таковы: 1; 2; 2.7; 2.9; 3; 3.4; 3.9; 4.1; 4.5; 3.8; 3.5; 3.1; 2.8; 2.8; 2.1; 1.5; 1.9; 2.5; 2.3; 2.9; 3.9; 4.2; 5; 1.5; 1.5.

А) Выбрать длину частичного интервала $\Delta = 1$ и построить гистограмму.

Б) Вычислить $\bar{x} = \frac{1}{25} \sum_{k=1}^{25} x_k$ и $s^2 = \frac{1}{24} \sum_{k=1}^{24} (x_k - \bar{x})^2$

В) вычислить $\bar{x}$ и s^2 по формулам (5.12) и сравнить с результатами п. Б)

Г) Найти 95% доверительный интервал для оценки a

Д) Найти 98% доверительный интервал для оценки s

Е) Построить $F(x)$

4.15. $X : N(a, s)$ Результаты выборки обработаны, представлены в таблице:

$c_{k-1} - c_k$	0 - 1	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6
n_k	1	7	17	17	7	1

А) Построить гистограмму

Б) Найти 98% доверительный интервал для оценки a

В) Найти 99% доверительный интервал для оценки s

Г) Найти $p_k^{(M)}$ - модельные вероятности попадания случайной величины в k -й частичный интервал. Сравнить с эмпирическими вероятностями p_k

Д) построить $\mathbf{P}(x)$

4.16. $X : N(a, s)$. Данные наблюдений представлены в таблице:

$c_{k-1} - c_k$	0,5 - 1,5	1,5 - 2,5	2,5 - 3,5	3,5 - 4,5	4,5 - 5,5
n_k	3	11	20	12	3

Ответить на вопросы, сформулированные в задаче 4.15.

4.17. Линейный размер детали при ее изготовлении – случайная величина.

Пусть X – разность реальной длины детали и длины, заложенной

конструктором. $X : N(a, s)$. Данные наблюдений представлены в таблице:

k	$c_{k-1} - c_k$	n_k
1	(-0.25) - (-0.2)	2
2	(-0.2) - (-0.15)	4
3	(-0.15) - (-0.1)	9
4	(-0.1) - (-0.05)	15
5	(-0.05) - 0	19
6	0 - 0.05	19
7	0.05 - 0.1	15
8	0.1 - 0.15	8
9	0.15 - 0.2	3
10	0.2 - 0.25	2

А) Построить гистограмму

Б) Найти 99% доверительный интервал для a

В) Найти 98% доверительный интервал для s

Г) Найти $p_k^{(M)}$ - модельные вероятности попадания выборочной величины в k -й частичный интервал.

Д) Построить p_k - эмпирические вероятности попадания случайной величины в k -й интервал.

Е) Построить $\mathbf{P}(x)$

4.18. В Вашей группе измерить (воспользоваться результатом опроса) рост каждого студента. Построить гистограмму, выбрав разумное значение Δ (длины частичного интервала). Возможно, но не обязательно, визуальное вид гистограммы соответствует нормальному распределению ($X : N(a, s)$).

Найти 95% доверительные интервалы для оценки a и s .

4.19. Число заявок, поступающих на телефонную станцию за некоторое время T имеет распределение Пуассона ($K : \Pi(l)$). Данные наблюдений приведены в таблице:

k	7	8	11	14	17	20	22	23
n_k	1	2	7	10	9	4	2	1

Здесь k – число заявок, n_k – число реализаций k заявок.

А) Построить p_k – эмпирический закон распределения (см. пример п.V)

Б) Найти $p_k^{(M)}$ – модельные вероятности, и сравнить их с p_k

В) Найти 95% доверительный интервал для l

Г) Построить $\mathbf{P}(x)$

4.20. $K : \Pi(l)$. Данные наблюдений приведены в таблице:

k	8	11	14	17	20	23
n_k	6	20	30	26	13	4

Выполнить те же задания, что и в 4.19.

4.21. Число k бракованных деталей, производимых в цехе в течении дня, имеет распределение Пуассона ($K : \Pi(l)$). Данные наблюдений в таблице:

k	0	1	2	3
n_k	30	15	4	1

Выполнить те же задания, что и в 4.19.

4.22. $K : \Pi(l)$. Данные наблюдений приведены в таблице:

k	0	1	2	3	4	5	6	7	8	9
n_k	2	7	14	20	20	16	10	6	3	1

Выполнить те же задания, что и в 4.19.

4.23. Время (в миллисекундах) между заявками, поступающими на телефонную станцию имеет показательное распределение ($X : P(0, a)$), см гл.

1.12). Результаты наблюдений обработаны и представлены в таблице:

$c_{k-1} - c_k$	0 – 0.5	0.5 – 1	1 – 1.5	1.5 – 2	2 – 2.5	2.5 – 3	3 – 3.5	3.5 – 4
n_k	39	24	15	9	5	3	2	1

А) Построить гистограмму

Б) построить $\mathbf{P}(x)$

В) Найти 95% доверительный интервал для оценки a

Г) Найти эмпирические вероятностями p_k попадания в k -й интервал.

Д) Найти $p_k^{(M)}$ – модельные вероятности

4.24. Время (в месяцах) бесперебойной работы станков имеет показательное распределение ($X : P(0, a)$). Наблюдения показали:

$c_{k-1} - c_k$	0 – 2	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12	12 – 14
n_k	33	22	15	11	6	5	2

Выполнить задания, сформулированные в 4.23.

4.25. Время (в часах) безотказной работы компьютера имеет показательное распределение ($X : P(0, a)$). Наблюдения приведены в таблице:

$c_{k-1} - c_k$	0 – 2	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12
n_k	16	10	7	5	3	2

Выполнить задания, сформулированные в 4.23.

4.26. $X : P(0, a)$. Данные в таблице:

$c_{k-1} - c_k$	0 – 3.5	3.5 – 7	7 – 10.5	10.5 – 14	14 – 17.5	17.5 – 21
n_k	20	12	7	4	2	1

Выполнить задания, сформулированные в 4.23.

4.27. $X : P(0, a)$. Данные в таблице:

$c_{k-1} - c_k$	0 – 0.2	0.2 – 0.4	0.4 – 0.6	0.6 – 0.8	0.8 – 1
n_k	28	12	6	3	1

Выполнить задания, сформулированные в 4.23.

4.28. Случайная величина X имеет распределение Парето

$$f(x) = \frac{1}{a} \cdot x^{\frac{1}{a}-1}, \quad (0 < x < 1). \text{ Результаты наблюдений в таблице:}$$

$c_{k-1} - c_k$	0 – 0.2	0.2 – 0.4	0.4 – 0.6	0.6 – 0.8	0.8 – 1
n_k	4	12	19	27	35

А) Найти оценку $\hat{a}$ методом наибольшего правдоподобия

Б) Построить гистограмму

В) построить $\hat{F}(x)$

Г) Найти эмпирические вероятностями p_k попадания в k -й интервал и

сравнить их с $p_k^{(M)}$ - модельными вероятностями

Д) Найти 95% доверительный интервал для оценки a

Е) Найти оценку $\hat{a}$ методом моментов

4.29. X имеет распределение Парето $f(x) = \frac{1}{a} \cdot x^{-\frac{1}{a}-1}, \quad (x > 1)$. Результаты наблюдений в таблице:

$c_{k-1} - c_k$	1 – 4	4 – 7	7 – 10	10 – 13	13 – 15
n_k	75	11	4	2	1

Выполнить задания, сформулированные в 4.28.

4.30. X имеет распределение Вейбула: $X : V(2, a)$ (см. раздел 1.1), т.е.

$$F(x) = 1 - e^{-x^2/a}, \quad (x > 0) \text{ или } f(x) = \frac{2x}{a} e^{-x^2/a}, \quad (x > 0). \text{ Данные наблюдений}$$

в таблице:

$c_{k-1} - c_k$	0 – 0.4	0.4 – 0.8	0.8 – 1.2	1.2 – 1.6	1.6 – 2
n_k	7	16	14	8	3

Выполнить задания, сформулированные в 4.28.

4.31. $X : V(3, a)$ Данные наблюдений в таблице:

$c_{k-1} - c_k$	0 – 0.4	0.4 – 0.8	0.8 – 1.2	1.2 – 1.6
n_k	6	34	42	7

Выполнить задания, сформулированные в 4.28.

4.32. Будем считать, что время ожидания на остановках общественного транспорта имеет равномерное распределение в интервале $(0, b)$, т.е.

$X : U(0, b)$ (см. 1.10). Данные наблюдений: 0; 0,2; 0,85; 1,65; 0,3; 0,89; 1,69; 0,39; 0,41; 1,1; 1; 1,95; 0,95; 0,44; 0,7; 1,25; 0,7; 0,75; 1,5; 1,5; 1,3; 1,65; 1,69; 1,99.

А) Найти методом наибольшего правдоподобия несмещенную оценку $\hat{b}$ параметра b (см. задачу 3.24)

Б) Найти 95% доверительный интервал для b , используя формулу (4.5) и распределение (2.4)

В) Показать, что метод моментов приводит к точечной оценке $b = 2 \cdot \bar{x}$. Сравнить эту оценку с той, что получена в п.А). Используя эту оценку, построить 95% доверительный интервал для b .

Г) Построить гистограмму, выбрав $\Delta = 0,2$.

Д) Найти эмпирические p_k и модельные вероятности $p_k^{(M)}$ попадания в k -й интервал.

Е) построить $F(x)$

4.33. $X : LN(a, s)$ (см.1.9). Данные наблюдений представлены в таблице:

$c_{k-1} - c_k$	0,25 – 0,75	0,75 – 1,25	1,25 – 1,75	1,75 – 2,25	2,25 – 2,75	2,75 – 3,25	3,25 – 3,75
n_k	3	11	13	9	6	3	1

А) Построить гистограмму $H(x)$ и $F(x)$

Б) Найти $\hat{a}$ и s^2 (см. задачу 3.22)

В) Найти $\bar{x}$ и s^2

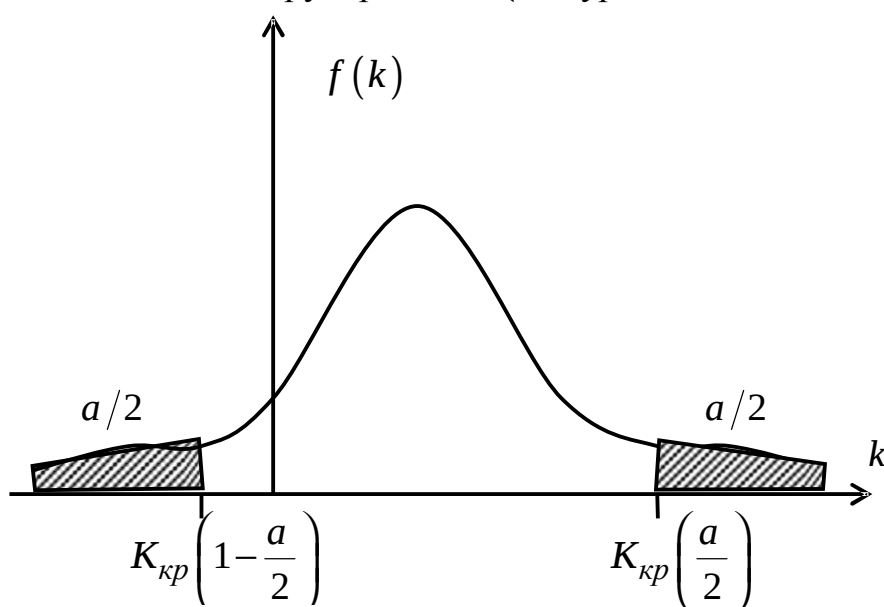
Г) Найти 95% доверительный интервал для a , используя приближенную формулу (4.5)

Д) Найти эмпирические p_k и модельные вероятности $p_k^{(M)}$.

VI. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ.

Пусть $F_X(x, q)$ - функция распределения вероятности генеральной случайной величины X с неизвестным параметром q . Одна из задач математической статистики состоит в том, чтобы принять решение о правильности НУЛЕВОЙ гипотезы $H_0: q = q_0$, или АЛЬТЕРНАТИВНОЙ гипотезы $H_1: q \neq q_0$, где q_0 определяется условием конкретной задачи. Для решения задачи такого типа на основе выборки $x_1, x_2, \dots, x_n$ строится случайная величина

$K(x_1, x_2, \dots, x_n; q_0)$. Эта величина называется критерием и предполагается, что закон распределения K – известен. С помощью этого критерия определяется критическая область (на рисунке это – заштрихованная область), вероятность попадания в которую равна α (α - уровень значимости).



$$\text{При чем } \int_{K_{kp}\left(\frac{\alpha}{2}\right)}^{\infty} f(k) dk = \frac{\alpha}{2}; \quad \int_{K_{kp}\left(1-\frac{\alpha}{2}\right)}^{\infty} f(k) dk = 1 - \frac{\alpha}{2} \quad (6.1)$$

Уровень значимости задается исследователем. Решение по предлагаемым гипотезам принимается следующим образом: Если наблюдаемое значение критерия K лежит в интервале

$$K_{kp}\left(1 - \frac{\alpha}{2}\right) < K < K_{kp}\left(\frac{\alpha}{2}\right), \quad (6.2)$$

то на уровне значимости α принимается нулевая гипотеза $H_0: q = q_0$. Если же наблюдаемое значение критерия попадает в критическую маловероятную область (обычно $\alpha \approx 0,05$),

$$\text{т.е. } K \geq K_{kp}\left(\frac{\alpha}{2}\right) \text{ или } K \leq K_{kp}\left(1 - \frac{\alpha}{2}\right), \quad (6.3)$$

то нулевая гипотеза отвергается и принимается альтернативная гипотеза $H_1: q \neq q_0$.

Изложенная процедура дает двусторонний критерий принятия решения.

Иногда приходится использовать односторонний критерий: $H_0: q = q_0$, если $K < K_{кр}(a)$; или $H_1: q \neq q_0$, если $K \geq K_{кр}(a)$ - правосторонний критерий.

Либо $H_0: q = q_0$, если $K > K_{кр}(1-a)$; или $H_1: q \neq q_0$, если $K \leq K_{кр}(1-a)$ - левосторонний критерий.

Отметим, что все вычисления проводятся при условии справедливости гипотезы H_0 .

ГИПОТЕЗЫ О РАВЕНСТВЕ НЕИЗВЕСТНОГО ПАРАМЕТРА ЗАДАННОМУ ЗНАЧЕНИЮ

Нормальное распределение: $X : N(a, s)$

1. Рассматриваются гипотезы: $H_0: a = a_0$ и $H_1: a \neq a_0$.

В этом случае критерий выбирается в виде:

$$t = \frac{\bar{x} - a_0}{s} \cdot \sqrt{n}, \quad \text{где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (6.4)$$

При выполнении нулевой гипотезы случайная величина t имеет

распределение Стьюдента с числом степеней свободы $n-1$, т.е. $t : f_{st}^{(n-1)}(t)$.

Если $\frac{|\bar{x} - a_0|}{s} \cdot \sqrt{n} < t_{кр}^{(n-1)}\left(\frac{a}{2}\right)$, то принимается гипотеза $H_0: a = a_0$ (6.5)

Если $\frac{|\bar{x} - a_0|}{s} \cdot \sqrt{n} \geq t_{кр}^{(n-1)}\left(\frac{a}{2}\right)$, то принимается гипотеза $H_1: a \neq a_0$.

Здесь $t_{кр}^{(n-1)}\left(\frac{a}{2}\right)$ находится по таблице из условия $\int_{t_{кр}^{(n-1)}(a/2)}^{\infty} f_{st}^{(n-1)}(t) dt = \frac{a}{2}$.

2. Рассматриваются гипотезы: $H_0: s = s_0$ и $H_1: s \neq s_0$.

В этом случае критерий выбирается в виде:

$$U = \frac{s^2(n-1)}{s_0^2} \quad (6.6)$$

При выполнении нулевой гипотезы случайная величина U имеет

распределение Хи - квадрат с числом степеней свободы $n-1$, т.е. $U : \chi_{n-1}^2$.

Если $U_{кр}\left(1 - \frac{a}{2}\right) < \frac{s^2(n-1)}{s_0^2} < U_{кр}\left(\frac{a}{2}\right)$, то принимается гипотеза $H_0: s = s_0$

Если $\frac{s^2(n-1)}{s_0^2} \geq U_{кр} \left(\frac{a}{2} \right)$, или $\frac{s^2(n-1)}{s_0^2} \leq U_{кр} \left(1 - \frac{a}{2} \right)$, то принимается гипотеза $H_1: s \neq s_0$ (6.7)

Здесь $U_{кр} \left(\frac{a}{2} \right)$ находится по таблице из условия $\int_{U_{кр}(a/2)}^{\infty} f_{c_{n-1}^2}(u) du = \frac{a}{2}$.

Для других распределений можно получить приближенный критерий при $n \gg 1$. Если $X: f(x, q)$ и на основании выборки $x_1, x_2, \dots, x_n$ получена точечная оценка $\hat{q}$, то критерий можно выбрать в виде

$$z = \frac{\hat{q} - q_0}{\sqrt{D\hat{q}}} \quad (6.8)$$

который при выполнении нулевой гипотезы $H_0: q = q_0$ имеет нормальное распределение, т.е. $z: N(0,1)$. При $n \gg 1$ $D\hat{q} = \frac{c^2(q)}{n}$, поэтому

$$z = \frac{\hat{q} - q_0}{c(q_0)} \sqrt{n} \quad (6.9)$$

Если $\frac{|\hat{q} - q_0|}{c(q_0)} \sqrt{n} < z_{кр} \left(1 - \frac{a}{2} \right)$, то принимается гипотеза $H_0: q = q_0$

Если $\frac{|\hat{q} - q_0|}{c(q_0)} \sqrt{n} \geq z_{кр} \left(1 - \frac{a}{2} \right)$, то принимается гипотеза $H_1: q \neq q_0$.

Здесь $z_{кр} \left(1 - \frac{a}{2} \right)$ находится из условия $F_N \left(z_{кр} \left(1 - \frac{a}{2} \right) \right) = 1 - \frac{a}{2}$.

В частности:

Распределение Бернулли $X: b(1, p)$

$$\hat{p} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i; \quad D\hat{p} = \frac{p(1-p)}{n}; \quad z = \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} \quad (6.10)$$

Если $\frac{|\bar{x} - p_0|}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{n} < z_{кр} \left(1 - \frac{a}{2} \right)$, то принимается гипотеза $H_0: p = p_0$

Если $\frac{|\bar{x} - p_0|}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{n} \geq z_{кр} \left(1 - \frac{a}{2} \right)$, то принимается гипотеза $H_1: p \neq p_0$.

Распределение Пуассона $K : \Pi(l)$

$$\mathfrak{L} = \bar{k} = \frac{1}{n} \sum_{i=1}^n k_i ; D\mathfrak{L} = \frac{l}{n} ; z = \frac{\bar{k} - l_0}{\sqrt{l_0}} \sqrt{n} \quad (6.11)$$

Если $\frac{|\bar{k} - l_0|}{\sqrt{l_0}} \sqrt{n} < z_{кр} \left(1 - \frac{\alpha}{2}\right)$, то принимается гипотеза $H_0 : l = l_0$

Если $\frac{|\bar{k} - l_0|}{\sqrt{l_0}} \sqrt{n} \geq z_{кр} \left(1 - \frac{\alpha}{2}\right)$, то принимается гипотеза $H_1 : l \neq l_0$.

Показательное распределение $X : P(0, a)$

$$\mathfrak{L} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i ; D\mathfrak{L} = \frac{a^2}{n} ; z = \frac{\bar{x} - a_0}{a_0} \sqrt{n} \quad (6.12)$$

Если $\frac{|\bar{x} - a_0|}{a_0} \sqrt{n} < z_{кр} \left(1 - \frac{\alpha}{2}\right)$, то принимается гипотеза $H_0 : a = a_0$

Если $\frac{|\bar{x} - a_0|}{a_0} \sqrt{n} \geq z_{кр} \left(1 - \frac{\alpha}{2}\right)$, то принимается гипотеза $H_1 : a \neq a_0$.

ГИПОТЕЗЫ ОДНОРОДНОСТИ

Пусть $X : N(a, s)$; $X' : N'(a', s')$, X и X' независимы. На основе выборок

$(x_1, x_2, \dots, x_n)$ и $(x'_1, x'_2, \dots, x'_{n'})$ вычислим $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$; $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$;

$\bar{x}' = \frac{1}{n'} \sum_{i=1}^{n'} x'_i$; $s'^2 = \frac{1}{n'-1} \sum_{i=1}^{n'} (x'_i - \bar{x}')^2$. Далее без ограничения общности будем

считать, что $s^2 > s'^2$.

1. На основе полученных наблюдений проверим гипотезу $H_0 : s = s'$, $H_1 : s \neq s'$. В этом случае критерий обозначается буквой F и

$$F = \frac{s^2}{s'^2} \quad (6.13)$$

F имеет распределение Фишера с числами степеней свободы $(n-1; n'-1)$,

т.е. $F : f_F^{(n-1; n'-1)}$

Если $\frac{s^2}{s'^2} < F_{kp}^{(n-1; n'-1)}\left(\frac{a}{2}\right)$, то принимается гипотеза $H_0: s = s'$

Если $\frac{s^2}{s'^2} \geq F_{kp}^{(n-1; n'-1)}\left(\frac{a}{2}\right)$, то принимается гипотеза $H_1: s \neq s'$

Здесь $F_{kp}^{(n-1; n'-1)}\left(\frac{a}{2}\right)$ находится по таблице из условия

$$\int_{F_{kp}(a/2)}^{\infty} f_F^{(n-1, n'-1)}(F) df = \frac{a}{2}.$$

2. Исследуется гипотеза $H_0: a = a'$, $H_1: a \neq a'$ при неизвестных, но равных дисперсиях ($s = s'$). В этом случае критерий выбирается в виде

$$t = \frac{\bar{x} - \bar{x}'}{s_0 \cdot \sqrt{\frac{1}{n} + \frac{1}{n'}}}, \text{ где } s_0 = \sqrt{\frac{s^2(n-1) + s'^2(n'-1)}{n-1+n'-1}} \quad (6.14)$$

При выполнении нулевой гипотезы $t: f_{st}^{(n-1+n'-1)}$

Если $\frac{|\bar{x} - \bar{x}'|}{s_0 \cdot \sqrt{\frac{1}{n} + \frac{1}{n'}}} < t_{kp}^{(n-1+n'-1)}\left(\frac{a}{2}\right)$, то принимается гипотеза $H_0: a = a'$

Если $\frac{|\bar{x} - \bar{x}'|}{s_0 \cdot \sqrt{\frac{1}{n} + \frac{1}{n'}}} \geq t_{kp}^{(n-1+n'-1)}\left(\frac{a}{2}\right)$, то принимается гипотеза $H_1: a \neq a'$

3. При неравных s и s' точного решения не существует. Однако, при $n \gg 1$ и $n' \gg 1$, хорошие результаты дает критерий

$$t = \frac{\bar{x} - \bar{x}'}{\sqrt{\frac{s^2}{n} + \frac{s'^2}{n'}}} \quad (6.15)$$

Решение по этому критерию принимается так же как и в предыдущем пункте.

Другие распределения.

Если $X: f(x, q)$; $X': f'(x', q')$, X и X' независимы; $\hat{q}$ - точечная оценка q ,

$\hat{q}'$ - точечная оценка q' ; $D\hat{q} = \frac{c^2(q)}{n}$, $D\hat{q}' = \frac{c^2(q')}{n'}$, то при $n \gg 1$ и $n' \gg 1$

можно использовать приближенный критерий:

$$z = \frac{\bar{q} - \bar{q}'}{\sqrt{\frac{c^2(\bar{q})}{n} + \frac{c^2(\bar{q}')}{n'}}} \quad (6.16)$$

Если $\frac{|\bar{q} - \bar{q}'|}{\sqrt{\frac{c^2(\bar{q})}{n} + \frac{c^2(\bar{q}')}{n'}}} < z_{kp} \left(1 - \frac{a}{2}\right)$, то принимается гипотеза $H_0: q = q'$

Если $\frac{|\bar{q} - \bar{q}'|}{\sqrt{\frac{c^2(\bar{q})}{n} + \frac{c^2(\bar{q}')}{n'}}} \geq z_{kp} \left(1 - \frac{a}{2}\right)$, то принимается гипотеза $H_1: q \neq q'$

ЗАВИСИМЫЕ ВЫБОРКИ

Наблюдения проводятся над одним и тем же объектом до и после какого-либо воздействия. Пусть до воздействия $X: N(a, s)$ и наблюдения дали выборку $(x_1, x_2, \dots, x_n)$, а после воздействия $X': N'(a', s')$ и получена выборка $(x'_1, x'_2, \dots, x'_n)$ того же объема n . Проверим гипотезы $H_0: a = a'$ и $H_1: a \neq a'$. В этом случае критерий выбирается в виде:

$$t = \frac{\bar{d}}{R} \cdot \sqrt{n}, \quad (6.17)$$

где $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$; $d_i = x_i - x'_i$; $R = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2}$.

При выполнении нулевой гипотезы $t: f_{st}^{(n-1)}$ и решение принимается так:

Если $\frac{|\bar{d}|}{R} \cdot \sqrt{n} < t_{kp}^{(n-1)} \left(\frac{a}{2}\right)$, то принимается гипотеза $H_0: a = a'$

Если $\frac{|\bar{d}|}{R} \cdot \sqrt{n} \geq t_{kp}^{(n-1)} \left(\frac{a}{2}\right)$, то принимается гипотеза $H_1: a \neq a'$

Другие распределения

Если $X: f(x, q)$; $X': f'(x', q')$ и $\bar{q} = \frac{1}{n} \sum_{i=1}^n j(x_i)$; $\bar{q}' = \frac{1}{n} \sum_{i=1}^n j(x'_i)$, то при $n \gg 1$ можно использовать приближенный критерий:

$$z = \frac{\bar{d}}{R} \cdot \sqrt{n} \quad (6.18)$$

где $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$; $d_i = j(x_i) - j(x'_i)$; $R = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2}$. Здесь $z_{кр} \left(1 - \frac{a}{2}\right)$

находится с помощью нормального распределения. Применяется критерий так же как и в предыдущем случае.

ТЕСТ ПИРСОНА.

Задача состоит в том, чтобы на основе выборки $(x_1, x_2, \dots, x_n)$ проверить гипотезу $H_0: F(x) = F^{(M)}(x)$; $H_1: F(x) \neq F^{(M)}(x)$, где модельная функция распределения $F^{(M)}(x)$ задается исследователем на основе предварительного анализа данных (гистограмма, эксцесс, асимметрия и т.д.). Исследуется критерий Пирсона:

$$K = \sum_{k=1}^s \frac{\left(n_k - n \cdot p_k^{(M)}\right)^2}{n \cdot p_k^{(M)}}, \quad (6.19)$$

где n_k - число наблюдений, попавших в интервал $[c_{k-1}; c_k)$; $p_k^{(M)}$ - вероятность попадания в интервал $[c_{k-1}; c_k)$, вычисленная с помощью модельного распределения, s - число интервалов. При $n \gg 1$ критерий имеет распределение $K: \chi^2_{s-r-1}$, где r - число параметров, которые определяются по выборке объема n (например, для нормального распределения $r=2$, поскольку неизвестны a и s^2)

Если $K < U_{кр}(a)$, то принимается гипотеза $H_0: F(x) = F^{(M)}(x)$

Если $K \geq U_{кр}(a)$, то принимается гипотеза $H_1: F(x) \neq F^{(M)}(x)$

Здесь $U_{кр}(a)$ находится по таблице по условию $\int_{U_{кр}(a)}^{\infty} f_{\chi^2_{n-1}}(u) du = a$.

Заметим, что должно выполняться $n_k \geq 5$, если это не так, то соседние частичные интервалы следует объединить в один интервал. Критерий (6.19) применим и к дискретным распределениям.

Задачи.

6.1. До прихода нового менеджера по обслуживанию, 40% покупателей магазина были удовлетворены уровнем обслуживания. После смены менеджера по результатам опроса 240 – ка покупателей было установлено, что 100 из них удовлетворены уровнем обслуживания. Можно ли, на уровне значимости 5%, считать, что при новом менеджере качество обслуживания изменилось?

6.2. Известно, что дневная выручка автостоянки имеет нормальное распределение с математическим ожиданием 5000 руб. Хозяин автостоянки совместно с участковым милиционером проводят разъяснительную работу среди жителей соседнего микрорайона о пользе хранения автомобилей на автостоянке. Наблюдения за 10 дней после разъяснительной работы дали следующий результат дневной выручки: 5200, 5050, 4950, 5020, 5120, 5080, 4970, 4900, 5030, 5100.

А) На уровне значимости 10% можно считать результаты разъяснительной работы значимыми?

Б) На уровне значимости 5% проверить гипотезу о том, что стандартное отклонение дневной выручки после беседы равно 90.

6.3. Число клиентов парикмахерской от 17 до 19 часов имеет распределение Пуассона с параметром $\lambda = 9$. Владельцем была установлена новая яркая вывеска. Наблюдения в течении 30-ти дней с 17 до 19 показали, что в среднем в это время обращалось 10 клиентов. На уровне значимости 10% проверить, можно ли считать значимыми результаты установки новой вывески.

6.4. Кассир магазина обслуживает в среднем 4 покупателя за 5 минут. Менеджер по кассовому обслуживанию предложил методы стимулирования работы кассиров. По результатам наблюдений за обслуживанием 40-ка покупателей, было установлено, что среднее время обслуживания составило 1 мин. и 20 сек. На уровне значимости 5% проверить, можно ли считать значимыми результаты внедрения мер стимулирования.

6.5. Время реакции (сек) до и после принятия энергетического напитка для 10-ти испытуемых было зафиксировано:

До:	3.0	2.4	1.2	2.3	2.4	4.2	2.7	5.2	1.5	4.1
После:	3.1	2.2	1.3	2.5	2.6	4.1	3.0	5.1	1.7	4.3

На уровне значимости $\alpha = 5\%$ проверить, можно ли считать значимыми действие энергетического напитка. Время реакции имеет нормальное распределение.

6.6. Опрос 300 избирателей до проведения дебатов показал, что 100 из них поддерживают кандидата А. Опрос 200 избирателей после проведения дебатов показал, что кандидата А поддерживают 85 избирателей. На уровне значимости 10% проверить, можно ли считать, что дебаты изменили рейтинг кандидата.

6.7. Обработка данных по двум идентичным супермаркетам, расположенным в двух городах Н и М, дала следующий результат (в тыс. руб.). Н: средний дневной объем продаж 250, выборочная дисперсия 85, число наблюдений 20 дней. М: средний дневной объем продаж 240, выборочная дисперсия 90, число наблюдений 10 дней. Считаем объем продаж нормальным.

- А) На уровне значимости 5% проверить гипотезу о равенстве дисперсий.
- Б) На том же уровне значимости проверить гипотезу о равенстве средних доходов.

6.8. Владелец крупного магазина считает, что представленный ассортимент товаров привлекает покупателей всех уровней дохода. Выборочный опрос 50-ти покупателей дал следующие результаты (в 10000 руб): 0.2; 0.18; 1.46; 0.5; 0.66; 1.02; 0.02; 0.7; 1.72; 0.68; 1.34; 0.7; 0.96; 1.52; 0.74; 1.08; 0.4; 0.96; 0.1; 1.28; 1.35; 0.94; 0.84; 1.92; 0.48; 1.6; 1.04; 0.8; 0.74; 0.16; 0.84; 0.52; 1.78; 1.06; 1.98; 0.02; 1.8; 0.5; 0.58; 0.24; 1.6; 1.58; 1.98; 1.4; 1.6; 0.3; 1.46; 1.22; 0.94; 1.28. Проверяется гипотеза о том, что наблюдается равномерное распределение на интервале [0;2].

- А) Разбить область изменения на 10 интервалов; построить гистограмму. Визуально сравнить гистограмму с модельной плотностью распределения.
- Б) На уровне значимости 1% проверить гипотезу о равномерности распределения.

6.9. Проверяется гипотеза о том, что время обслуживания (мин.) имеет показательное распределение. Данные наблюдений – в таблице:

$c_{k-1} - c_k$	0 - 1	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6
n_k	36	28	14	10	7	5

На уровне значимости $\alpha = 0.05$ проверить гипотезу о согласии наблюдений с модельным (показательным) распределением.

6.10. В задаче 4.14. при уровне значимости 5%:

- А) Проверить гипотезу $\alpha = 3$.
- Б) Используя критерий Пирсона убедиться в том, что данные наблюдений соответствуют нормальному распределению.
- В) При $\alpha = 10\%$ проверить гипотезу $s = 1.1$.

6.11. В задаче 4.15. при уровне значимости 10%:

- А) Проверить гипотезу $H_0 : \alpha = 3.2$.
- Б) Проверить гипотезу $H_0 : s = 1.1$.
- В) Проверить гипотезу о нормальности распределения.

6.12. В задаче 4.15.

- А) Проверить гипотезу $\alpha = 2.9$ на уровне значимости 5%.
- Б) проверить гипотезу $s = 1$ на уровне значимости 10%.
- В) Проверить гипотезу о нормальности распределения на уровне 2%.

6.13. В задаче 4.17.

- А) Проверить гипотезу $\alpha = 0.01$ на уровне значимости 5%.
- Б) проверить гипотезу $s = 0.1$ на уровне значимости 1%.
- В) Проверить гипотезу о нормальности распределения на уровне 5%.

6.14. В задаче 4.19.

- А) Проверить гипотезу $l = 14.9$ на уровне значимости 5%.
- Б) Проверить гипотезу о справедливости распределения Пуассона ($a = 2\%$).

6.15. В задаче 4.20.

- А) Проверить гипотезу $l = 15$ на уровне значимости 10%.
- Б) Проверить гипотезу о согласии данных наблюдений распределению Пуассона на том же уровне.

6.16. В задаче 4.21.

- А) Проверить гипотезу $H_0 : l = 0.49$ на уровне значимости 5%.
- Б) На уровне значимости 5% проверить гипотезу H_0 : данные наблюдений соответствуют распределению Пуассона.

6.17. В задаче 4.22 при $a = 5\%$:

- А) Проверить гипотезу $l = 4.1$.
- Б) Проверить гипотезу H_0 : данные наблюдений соответствуют распределению Пуассона.

6.18. В задаче 4.23:

- А) Проверить гипотезу $a = 1.1$ на уровне значимости 5%.
- Б) На уровне значимости 10% убедиться в справедливости показательного распределения.

6.19. В задаче 4.24:

- А) Проверить гипотезу $H_0 : a = 4$ на уровне значимости 10%.
- Б) На уровне значимости 5% убедиться в том, что данные наблюдений соответствуют показательному распределению.

6.20. В задаче 4.25:

- А) Проверить гипотезу $a = 4$ на уровне значимости $a = 5\%$.
- Б) На уровне значимости $a = 1\%$ проверить гипотезу о том, что данные наблюдений соответствуют показательному распределению.

6.21. В задаче 4.26:

- А) Проверить гипотезу $H_0 : a = 0.6$ на уровне значимости $a = 5\%$.
- Б) На уровне значимости $a = 5\%$ убедиться в справедливости показательного распределения.

6.22. В задаче 4.27:

- А) Проверить гипотезу $a = 0.3$ на уровне значимости $a = 6\%$.
- Б) На уровне значимости $a = 5\%$ проверить гипотезу о справедливости показательного распределения.

6.23. В задаче 4.28:

- А) Проверить гипотезу $a = 0.5$ на уровне значимости $a = 5\%$.
- Б) На уровне значимости $a = 5\%$ проверить гипотезу о том, что данные наблюдений соответствуют распределению Парето.

6.24. В задаче 4.29:

- А) Проверить гипотезу $a = 1$ при $a = 5\%$.
- Б) На уровне $a = 5\%$ проверить гипотезу о согласии данные наблюдений распределению Парето.

6.25. В задаче 4.30:

- А) Проверить гипотезу $a = 1$ при $a = 6\%$.
 Б) На уровне $a = 5\%$ убедиться в справедливости распределения Парето.

6.26. В задаче 4.30:

- А) Проверить гипотезу $H_0: a = 1$ на уровне значимости 5% .
 Б) На уровне значимости $a = 5\%$ убедиться в справедливости распределения Вейбула.

6.27. В задаче 4.31:

- А) Проверить гипотезу $a = 0.6$ при $a = 5\%$.
 Б) На уровне $a = 5\%$ убедиться в справедливости распределения Вейбула.

6.28. В задаче 4.32: Проверить гипотезу H_0 : данные наблюдений согласованы с нормальным распределением. Уровень значимости 5% .

6.29. В задаче 4.33: Проверить гипотезу о том, что X имеет распределение $LN(a, s)$. Уровень значимости 5% .

6.30. $X : N(a, s)$; $X' : N(a', s')$, X и X' - независимы. Данные наблюдений приведены в таблице:

$X :$	$c_{k-1} - c_k$	0 - 1	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6
	n_k	1	7	17	18	6	1

$X' :$	$c'_{k-1} - c'_k$	0.5 - 1.5	1.5 - 2.5	2.5 - 3.5	3.5 - 4.5	4.5 - 5.5
	n'_k	3	11	20	11	3

- А) Проверить гипотезу однородности $H_0: s = s'$ на уровне $a = 5\%$.
 Б) Проверить гипотезу однородности $H_0: a = a'$ на уровне $a = 5\%$.

6.31. $K : \Pi(l)$; $K' : \Pi(l')$, K и K' - независимы. Данные наблюдений приведены в таблице:

$K :$	k	0	1	2	3	4	5	6	7	8	9
	n_k	2	6	14	20	19	16	10	7	3	1

$K' :$	k'	1	2	3	4	5	6	7	8
	n'_k	7	13	20	20	16	10	7	3

Проверить гипотезу однородности $H_0: l = l'$; $H_1: l \neq l'$ на уровне $a = 5\%$.

6.32. $X : P(0, a)$; $X' : P(0, a')$, X и X' - независимы. Данные наблюдений приведены в таблице:

$X :$	$c_{k-1} - c_k$	0 - 2	2 - 4	4 - 6	6 - 8	8 - 10	10 - 12	12 - 14
	n_k	32	21	15	11	6	4	2

$X' :$	$c'_{k-1} - c'_k$	0 - 4	4 - 8	8 - 12
	n'_k	51	25	10

Проверить гипотезу однородности $H_0: a = a'$ на уровне значимости $a = 5\%$.

6.33. $K : \Pi(l)$; $K' : \Pi(l')$, K и K' - независимы. Данные

наблюдений приведены в таблице:

$K :$	k	7	8	11	14	17	20	22	23
	n_k	1	2	7	10	9	4	2	1

$K' :$	k'	8	11	14	17	20	23
	n'_k	6	20	30	26	13	4

Проверить гипотезу однородности $H_0 : l = l'$ на уровне значимости $\alpha = 5\%$.

6.34. $X : P(0, a)$; $X' : P(0, a')$, X и X' - независимы. Данные

наблюдений приведены в таблице:

$X :$	$c_{k-1} - c_k$	0 - 3.5	3.5 - 7	7 - 10.5	10.5 - 14	14 - 17.5	17.5 - 21
	n_k	20	12	7	4	2	1

$X' :$	$c'_{k-1} - c'_k$	0 - 3.5	3.5 - 7	7 - 10.5	10.5 - 14	14 - 17.5
	n'_k	21	12	6	4	1

Проверить гипотезу однородности $H_0 : a = a'$ на уровне значимости $\alpha = 5\%$.

6.35. В зале перед началом дебатов присутствовало 100 чел. Из них 40% поддерживало кандидата. После дебатов опрос тех же 100 чел. Показал, что 10 из них ухудшили свое мнение, 15 – улучшили, остальные свое мнение о кандидате не изменили. На уровне значимости 5% определить: изменился рейтинг кандидата или нет?

6.36. $X : V(2, a)$; $X' : V'(2, a')$, X и X' - независимы. Данные

наблюдений приведены в таблице:

$X :$	$c_{k-1} - c_k$	0 - 0.4	0.4 - 0.8	0.8 - 1.2	1.2 - 1.6	1.6 - 2
	n_k	7	16	14	8	3

$X' :$	$c'_{k-1} - c'_k$	0 - 0.4	0.4 - 0.8	0.8 - 1.2	1.2 - 1.6	1.6 - 2
	n'_k	8	15	13	7	2

Проверить гипотезу однородности $a = a'$ на уровне значимости $\alpha = 5\%$.

6.37. $X : f(x) = \frac{1}{a} \cdot e^{-\frac{1}{a}x-1}$, ($x > 1$); $X' : f'(x') = \frac{1}{a'} \cdot e^{-\frac{1}{a'}x'-1}$, ($x' > 1$),

X и X' - независимы. Данные наблюдений приведены в таблице:

$X :$	$c_{k-1} - c_k$	1 - 4	4 - 7	7 - 10	10 - 13	13 - 16
	n_k	75	11	4	2	1

$X' :$	$c'_{k-1} - c'_k$	1 - 4	4 - 7	7 - 10	10 - 13	13 - 16
	n'_k	77	12	5	3	2

Проверить гипотезу однородности $a = a'$ на уровне значимости $\alpha = 5\%$.

VII. КОРРЕЛЯЦИЯ.

Корреляционная связь – это зависимость среднего значения результативного признака (Y) от изменения факторного признака (X)¹. Основной задачей корреляционного анализа является изучение степени тесноты (силы) связи двух или более признаков (либо ее отсутствия), а также причин, вызвавших эту связь. Здесь важно понимать суть взаимосвязи, поскольку в одной ситуации она возникает вследствие причинной зависимости (X – причина, Y – следствие), а в другой оба фактора могут быть следствием одной (скрытой) причины и сами напрямую не зависеть друг от друга (как, например, длина женской юбки и объемы продаж мороженого – подумайте, почему?).

По направлению выделяют положительную (прямую) и отрицательную (обратную) связь, по аналитическому выражению – линейную и нелинейную. Мы остановимся на наиболее простом случае линейной связи при наличии двух признаков X и Y . В этом случае степень тесноты линейной связи между признаками определяется с помощью *коэффициента линейной корреляции Пирсона r* , представляющего собой объективную меру линейной взаимосвязи двух признаков. Вычисление коэффициента корреляции Пирсона, как и большинство методов измерения тесноты связи, сопряжено с сопоставлением отклонений фактических значений от средних (т.е. с вариативностью данных). Линейный коэффициент корреляции Пирсона – это безразмерная величина, носящая относительный характер, изменяющаяся в пределах от -1 до $+1$ (причем, чем r по модулю ближе к единице, тем сильнее связь). Коэффициент Пирсона может быть вычислен по одной из формул:

¹ Здесь важно понимать, что речь не идет о зависимости Y от X , а лишь о наличии (отсутствии) взаимосвязи между ними.

$$r = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) \left(\sum_{i=1}^n y_i^2 - n \bar{y}^2 \right)}}, \quad (7.1)$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}, \quad (7.2)$$

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x s_y}, \quad (7.3)$$

$$r = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{(n-1) S_x S_y}. \quad (7.4)$$

Здесь n – объем выборки, $\{x_i, y_i\}_{i=1}^n$ – двумерный набор наблюдаемых значений, $\bar{x}$, $\bar{y}$ – выборочные средние², $\overline{xy}$ – совместное среднее, S_x , S_y – (исправленные) выборочные стандартные отклонения³, s_x , s_y – стандартные отклонения генеральной совокупности⁴. К практическому вычислению коэффициента корреляции Пирсона мы вернемся несколько позже.

Парная линейная регрессия.

Регрессия, в отличие от корреляции, – это односторонняя статистическая зависимость; она определяет форму связи между признаками. Регрессионный анализ позволяет предсказывать значение одного признака

$$^2 \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$^3 S_x = \sqrt{\frac{\left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right)}{(n-1)}}$$

$$^4 S_x = \sqrt{\frac{n}{n-1}} s_x, \text{ при объеме выборки } n > 30 \text{ различия между } s_x \text{ и } S_x \text{ считают пренебрежимо}$$

малыми

(переменная Y) на основании другого (переменная X) с использованием некоторой линии, характеризующей взаимосвязь между этими признаками. Естественным считать X независимой (предсказывающей) переменной, а Y – зависимой (прогнозируемой). В регрессионном анализе независимая переменная именуется *регрессором* (или «*предиктором*»), а зависимая – *откликом*. В самом общем случае переменная X представляет собой n – мерный вектор, а линия может быть порождена произвольной функциональной зависимостью. Наиболее простой является ситуация, когда переменная X одномерна, а линия – прямая. Такую регрессию называют парной линейной регрессией.

Построение уравнения регрессии.

Предположение о парной линейной зависимости между X и Y можно описать функцией

$$Y = a + bX + e,$$

где a, b – истинные значения параметров регрессии в генеральной совокупности, e – случайная составляющая. Существует несколько причин возникновения случайной составляющей: невключение объясняющих переменных в уравнение регрессии, неправильное описание структуры модели, неверный выбор объясняющих переменных, неверная функциональная спецификация модели, ошибки данных и проч.

В случае правильной спецификации модели (т.е. если действительно имеет место линейная зависимость), по выборочным данным определяются оценки истинных параметров уравнения регрессии и случайной составляющей a, b, e – параметры выборочной регрессии. Уравнение выборочной регрессии в случае парной линейной зависимости имеет вид:

$$\hat{y}_x = a + bx, \quad (7.5)$$

где b – оценка истинного (неизвестного) наклона регрессии b , a – оценка истинного (неизвестного) сдвига a , случайная компонента (с рядом соглашений) в модель не включается.

В основе построения регрессионной прямой (т.е. нахождения параметров регрессии) лежит метод наименьших квадратов (соответствующую прямую называют линией наилучшего соответствия), минимизирующий сумму квадратов отклонений фактических значений Y от прогнозных $\hat{Y}$ (ординат соответствующих точек на прямой), что гарантирует наименьшую из возможных ошибок прогнозирования. Для рассматриваемой модели (7.5) условие минимума суммы квадратов отклонений (S_T) имеет вид

$$S_T = \sum_{i=1}^n (y_i - \hat{y}_x)^2 \rightarrow \min;$$

$$S_T = \sum_{i=1}^n (y_i - a - bx_i)^2 \rightarrow \min.$$

Необходимое условие минимума выглядит так:

$$\begin{cases} \frac{\partial S_T}{\partial a} = 2an - 2\sum_{i=1}^n y_i + 2b\sum_{i=1}^n x_i = 0 \\ \frac{\partial S_T}{\partial b} = 2a\sum_{i=1}^n x_i - 2\sum_{i=1}^n x_i y_i + 2b\sum_{i=1}^n x_i^2 = 0 \end{cases}$$

Решая систему относительно неизвестных параметров регрессии, находим, что

$$b = \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = r \frac{S_x}{S_y}, \quad (7.6)$$

$$a = \bar{y} - b\bar{x}.$$

Здесь параметр b – коэффициент наклона регрессионной прямой, характеризующий влияние, которое оказывает изменение предиктора X на отклик Y . Он показывает, на сколько единиц изменится в среднем Y при

изменении X на одну единицу. Если $b > 0$, то наблюдаем положительную линейную взаимосвязь, если $b < 0$, то отрицательную (заметьте, что знак наклона регрессии b такой же, как и знак коэффициента корреляции Пирсона r). Коэффициент b обладает размерностью отношения Y к X . Параметр a – константа в уравнении регрессии, он характеризует сдвиг регрессионной прямой вдоль оси ординат Y . Его интерпретация зависит от того, какой смысл имеют исследуемые признаки. Он обладает размерностью признака Y .

Оценка уравнения регрессии.

Для того, чтобы оценить, насколько «хороша» полученная регрессия, т.е. насколько точный прогноз она дает, отвечают два показателя – *коэффициент детерминации* и *стандартная ошибка оценки*.

Коэффициент детерминации R^2 ⁵ – это относительная мера того, как много в поведении Y удалось объяснить за счет X , т.е. это доля изменчивости Y , объясняемая изменчивостью X . В общем случае он равен отношению суммы квадратов отклонений вследствие регрессии к общей сумме квадратов (здесь y – наблюдаемые значения, $\hat{y}_x$ – прогнозные (полученные из уравнения регрессии) значения, $\bar{y}$ – среднее значение Y):

$$R^2 = \frac{\sum_{i=1}^n (\bar{y} - \hat{y}_{x_i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (7.7)$$

однако для парной линейной регрессии коэффициент детерминации вычисляется проще, а именно численно равен квадрату коэффициента корреляции

$$R^2 = r^2. \quad (7.8)$$

⁵ R^2 – это неделимый символ (обозначение), это не квадрат некоего R .

Стоит отметить, что малые значения R^2 не обязательно указывают на то, что изменчивость Y не связана с поведением X ; это может говорить о наличии *других существенных факторов, не включенных в модель*.

Стандартная ошибка оценки S_e – это абсолютная мера величины ошибок прогнозирования, т.е. это мера вариации данных относительно линии регрессии (уравнение регрессии не является идеальным в смысле предсказания, поскольку не все точки лежат на регрессионной прямой). Стандартная ошибка оценки определяется как

$$S_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_x)^2}{n-2}} = S_y \sqrt{(1-r^2) \frac{n-1}{n-2}}. \quad (7.9)$$

Стандартную ошибку оценки можно интерпретировать: если ошибки предсказания нормально распределены, то можно ожидать, что 2/3 всех точек будут лежать на расстоянии не более S_e выше или ниже линии регрессии, кроме того, 95% всех точек будут лежать в полосе не более $\pm 2S_e$ от линии регрессии.

Доверительные интервалы для оценки неизвестного генерального среднего значения $\bar{Y}_{ген}$ и индивидуального значения $\bar{Y}_i$.

Поскольку для построения регрессионной модели используются выборочные данные, то интерпретация взаимосвязи между переменными в генеральной совокупности также базируется на выборочных результатах. Прогнозное значение при заданном x (найденное посредством регрессионного уравнения) – есть ни что иное, как точечная оценка истинного среднего значения признака Y , однако возможна и интервальная оценка. Границы доверительного интервала для оценки неизвестного генерального значения $\bar{Y}_{ген}$ имеют вид

$$\hat{y}_x \pm t_{n-2}(1-g) \cdot S_e \sqrt{h_x}, \quad (7.10)$$

где

$$h_x = \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \quad (7.11)$$

n – объем выборки, $\hat{y}_x$ – предсказанное значение Y , S_e – стандартная ошибка оценки, x – заданное значение X , $t_{n-2}(1-g)$ – двустороннее t -значение критерия Стьюдента на уровне доверия g (или, что то же самое, на уровне значимости $\alpha = 1-g$) с $n-2$ степенями свободы.

Доверительный интервал для индивидуальных значений Y_x , лежащих на линии регрессии, имеет такие границы:

$$\hat{y}_x \pm t_{n-2}(1-g) \cdot S_e \sqrt{1 + h_i}, \quad (7.12)$$

все выражения, входящие в формулу такие же, как описано выше. Очевидно, доверительный интервал для отдельных значений шире, чем для значений в среднем.

Оценки коэффициентов уравнения регрессии и доверительные интервалы для оценки истинных параметров уравнения регрессии.

Поскольку истинные параметры a и b регрессии неизвестны, мы можем их лишь оценить, используя данные выборки. Найдем точечные и интервальные оценки для неизвестных параметров регрессии в генеральной совокупности.

Стандартная ошибка наклона регрессии $\hat{b}$ – это приблизительная величина отклонения выборочного наклона регрессии b от истинного b :

$$\hat{b} = \frac{S_e}{S_x \sqrt{n-1}}. \quad (7.13)$$

Стандартная ошибка сдвига $\hat{a}$ – это приблизительная величина отклонения выборочного сдвига a от истинного (неизвестного) сдвига a :

$$\hat{a} = S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2(n-1)}}. \quad (7.14)$$

Имея эти оценки нетрудно построить доверительные интервалы:

а) Двусторонние доверительные интервалы с уровнем доверия g и $n-2$ степенями свободы для наклона b и сдвига a :

$$\begin{aligned} b - t_{n-2}(1-g) \cdot \hat{\sigma}_b \leq b \leq b + t_{n-2}(1-g) \cdot \hat{\sigma}_b; \\ a - t_{n-2}(1-g) \cdot \hat{\sigma}_a \leq a \leq a + t_{n-2}(1-g) \cdot \hat{\sigma}_a \end{aligned} \quad (7.15)$$

здесь a, b – выборочные параметры регрессии, $\hat{\sigma}_a, \hat{\sigma}_b$ – их стандартные ошибки, $t_{n-2}(1-g)$ – двустороннее t-значение критерия Стьюдента на уровне доверия g с $n-2$ степенями свободы⁶.

б) Односторонние доверительные интервалы (в ряде задач нет необходимости оценки снизу и сверху одновременно, наоборот, наиболее интересна *только нижняя* или *только верхняя граница*. Так, рассчитывая себестоимость единицы продукции, интерес представляет сумма, выше которой себестоимость не поднимется с вероятностью g , т.е. верхняя граница интервала): истинные параметры регрессии b (либо a) будут или не меньше, чем $b - t_{n-2}(1-g) \cdot \hat{\sigma}_b$ (либо $a - t_{n-2}(1-g) \cdot \hat{\sigma}_a$), или не больше, чем $b + t_{n-2}(1-g) \cdot \hat{\sigma}_b$ (либо $a + t_{n-2}(1-g) \cdot \hat{\sigma}_a$) Здесь $t_{n-2}(1-g)$ – одностороннее t-значение критерия Стьюдента на уровне доверия g с $n-2$ степенями свободы.

Отметим следующее: при одних и тех же выборочных характеристиках верхняя (нижняя) граница одностороннего интервала будет ниже (выше), чем у двустороннего, т.е. оценка будет более точной, поэтому если из контекста задачи ясно, что оценивать нужно *только сверху* (*только снизу*), то «правильнее» использовать односторонние интервалы.

⁶ Не имея под рукой таблицы распределения Стьюдента, можно на уровне доверия 0,95 использовать значение $t_{\infty}(0,05) = 1,96$

Проверка гипотезы о значимости коэффициентов регрессии (о наличии линейной взаимосвязи между X и Y в генеральной совокупности).

Сформулируем гипотезы:

$H_0: b = 0$ (связи нет, это случайность),

$H_1: b \neq 0$ (связь есть).

В случае, если нулевая гипотеза будет отклонена в пользу альтернативной, мы подтвердим существование линейной взаимосвязи в генеральной совокупности. Рассмотрим два способа проверки данной гипотезы:

1) С помощью доверительного интервала для b : если

$(0 \Rightarrow) b \in (b - t_{n-2}(1-g) \cdot \frac{s}{\sqrt{n}}, b + t_{n-2}(1-g) \cdot \frac{s}{\sqrt{n}})$, то принимаем гипотезу H_0 ,
иначе H_1 ;

2) С помощью t-критерия Стьюдента: вычислим $t_{набл} = \frac{b}{\frac{s}{\sqrt{n}}}$, если

$|t_{набл}| \leq t_{табл} = t_{n-2}(1-g)$, то принимаем гипотезу H_0 , иначе H_1 .

Продемонстрируем теперь изложенные методы анализа на примере.

Пример построения уравнения регрессии и оценка его качества.

Рассмотрим методы корреляционного и регрессионного анализов на примере классической задачи прогнозирования будущих затрат, основываясь на текущих планах выпуска продукции. Имеется следующий план выпуска продукции, включающий 18 наблюдений ($n=18$):

Таблица 1.

Номер наблюдения	Объем выпуска (шт.)	Затраты на выпуск (у.е.)
1	22	3470
2	30	3473
3	26	3856
4	31	3910
5	36	4489
6	30	3876
7	22	3221
8	45	4579
9	38	4325
10	30	3789
11	38	3999
12	41	4158
13	27	3566
14	28	3885
15	31	3574
16	37	4195
17	32	3814
18	41	4430

1. Построение модели регрессии.

Для получения прогноза размера затрат применим простую модель парной регрессии, в которой используется одна факторная переменная (предиктор) – X (объем выпуска). Визуально данные, приведенные в табл. 1, представлены диаграммой разброса (рассеивания) на рис. 1 ниже. Диаграмма наглядно показывает наличие весьма тесной положительной линейной зависимости затрат от объема выпуска продукции: с увеличением числа выпущенных изделий затраты на их выпуск растут (что естественно).

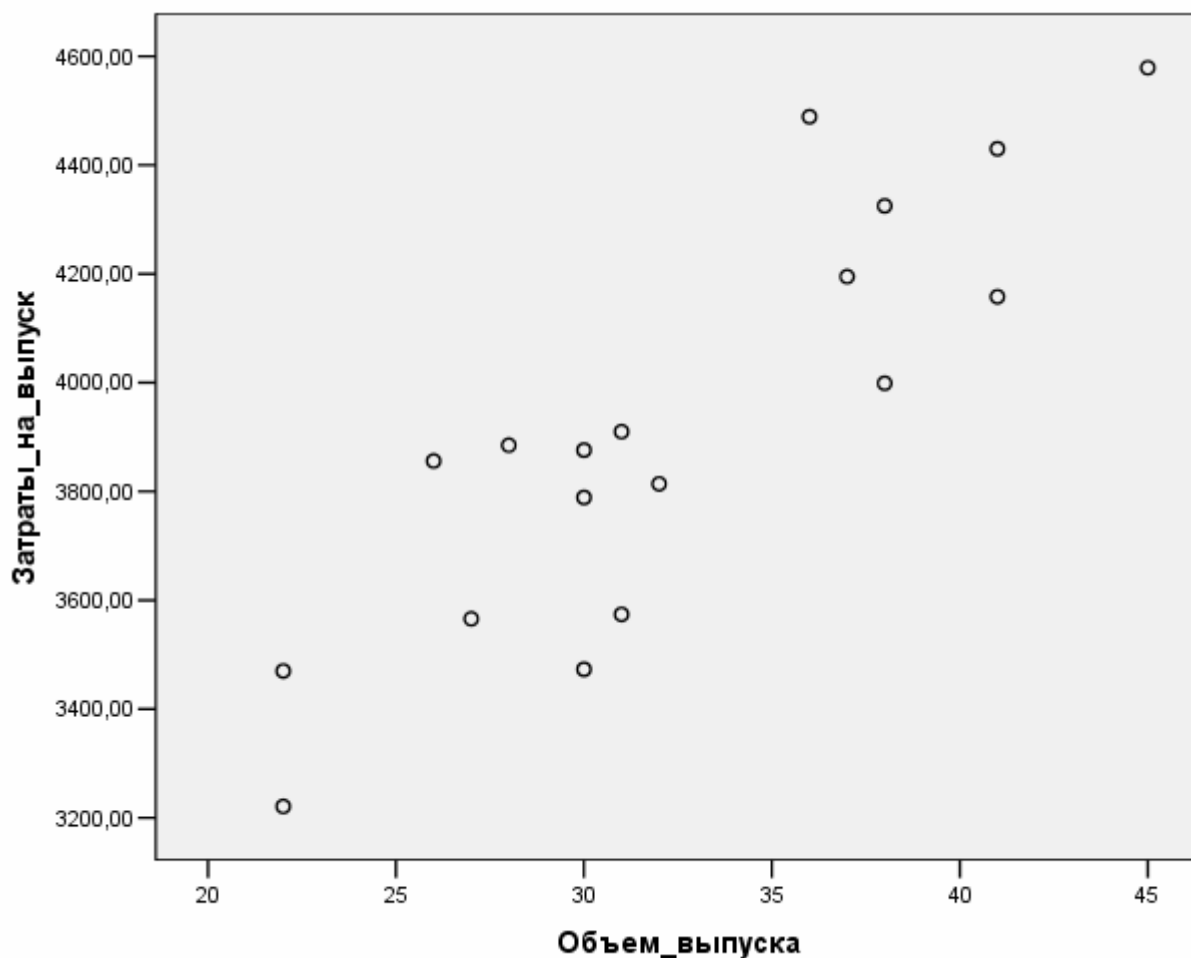


Рис. 1

Вычислим коэффициент корреляции Пирсона по формуле (7.1) в подтверждение вывода о наличии положительной весьма сильной связи, полученного при анализе диаграммы. Для упрощения расчетов здесь и далее воспользуемся вспомогательной табл. 2, приведенной ниже:

$$r = \frac{2332006 - 18 \cdot 32,5 \cdot 3922,72}{\sqrt{(19743 - 18 \cdot 32,5^2)(279471733 - 18 \cdot 3922,72^2)}} = 0,872.$$

Его значение положительно и близко к единице, что подтверждает вывод о наличии положительной тесной линейной связи.

Рассчитаем теперь параметры уравнения регрессии (т.е. найдем коэффициенты a и b уравнения), используя формулы (7.6). Снова обратимся к табл. 2. Для удобства в таблицу включены и средние значения; они как и обычно, вычисляются как среднее арифметическое.

Таблица 2.

№ набл.	Объем выпуска X	Затраты на выпуск Y	X^2	XY	Y^2
1	22	3470	484	76340	12040900
2	30	3473	900	104190	12061729
3	26	3856	676	100256	14868736
4	31	3910	961	121210	15288100
5	36	4489	1296	161604	20151121
6	30	3876	900	116280	15023376
7	22	3221	484	70862	10374841
8	45	4579	2025	206055	20967241
9	38	4325	1444	164350	18705625
10	30	3789	900	113670	14356521
11	38	3999	1444	151962	15992001
12	41	4158	1681	170478	17288964
13	27	3566	729	96282	12716356
14	28	3885	784	108780	15093225
15	31	3574	961	110794	12773476
16	37	4195	1369	155215	17598025
17	32	3814	1024	122048	14546596
18	41	4430	1681	181630	19624900
Сумма по столбцу	585	70609	19743	2332006	279471733
средние	32,5	3922,72	1096,8	129556	15526207,39

$$b = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{2332006 - 18 \cdot 32,5 \cdot 3922,72}{19743 - 18 \cdot 32,5^2} = 50,943,$$

$$a = \bar{y} - b \bar{x} = 3922,72 - 50,943 \cdot 32,5 = 2267,091$$

Таким образом, уравнение регрессии имеет вид⁷

$$\hat{y}_x = 2267,091 + 50,943x. \quad (**)$$

Коэффициент b , как говорилось ранее, характеризует наклон линии регрессии, и в данном случае означает, что *при увеличении объема выпуска X*

на одну единицу (на одно изделие) ожидаемое значение затрат на производство Y возрастет на 50,943 у.е. Таким образом, в контексте данной задачи коэффициент наклона регрессионной прямой можно интерпретировать как *переменные издержки*.

Далее, сдвиг регрессионной прямой (свободный член уравнения регрессии) – это значение отклика Y при $X = 0$. В решаемой задаче сдвиг регрессии $a = 2267,091 > 0$, т.е. при полном отсутствии производства ($X = 0$) мы имеем некие ненулевые затраты, причем не зависящие от объема производства. Поэтому можем сделать вывод о том, что здесь a есть *постоянные издержки* (аренда помещения, закупка и наладка оборудования, охрана и проч.)

Регрессионная модель может быть использована для прогноза затрат на производство продукции, однако, важно понимать, что обсуждаются лишь значения предиктора X , находящиеся в пределах от наименьшего до наибольшего его значений, используемых при создании модели. Так, в решаемой задаче наименьший объем выпуска продукции составляет 22 единицы, наибольший – 45 единиц. Поэтому мы можем сделать прогноз затрат на выпуск лишь для объемов производства, лежащих в этом диапазоне, скажем, для запланированного выпуска 24 и 39 изделий. Подставим по очереди эти значения в уравнение регрессии **(**)** и получим следующее:

$$y_x(24) = 2267,091 + 50,943 \cdot 24 = 3489,723,$$

$$y_x(39) = 2267,091 + 50,943 \cdot 39 = 4253,868.$$

Таким образом, прогнозируемые издержки для выпуска 24 единиц продукции в среднем составят 3489,723 у.е., 39 единиц – 4253,868 у.е.. Прогнозы для объемов выпуска, лежащих вне указанного диапазона, полученные с использованием *данного* регрессионного уравнения, могут давать неверные результаты, поэтому использовать это же уравнение нельзя. Так, например, по достижении определенного объема выпуска могут

⁷ Здесь мы округляем полученные значения до тысячных долей

возникнуть дополнительные затраты, скажем на привлечение новых сотрудников, увеличение производственных мощностей, увеличение арендованных площадей и проч..

2. Оценка «качества» модели.

Оценим теперь «качество» построенной регрессионной модели, вычислив стандартную ошибку оценки и коэффициент детерминации. По формуле (7.8)

$$S_e = S_y \sqrt{(1-r^2) \frac{n-1}{n-2}} = 382,89 \sqrt{(1-0,872^2) \frac{18-1}{18-2}} = 193,193,$$

что означает, что обычно фактические затраты отличались от прогнозируемых примерно на 193,193 у.е. (прогнозы не идеальны). Применяя (7.9), найдем

$$R^2 = 0,872^2 = 0,761,$$

это говорит о том, что 76,1% изменчивости производственных затрат (Y) объясняется изменением объема выпуска (X), а оставшиеся 23,9% изменчивости Y объясняется *другими причинами, не связанными с объемом выпуска* (т.е. эта доля изменчивости фактически остается необъясненной с точки зрения модели).

3. Доверительные интервалы.

Оценим теперь неизвестные параметры генеральной совокупности с помощью доверительных интервалов. Зададимся доверительной вероятностью $g = 0,95$, тогда по таблице распределения Стьюдента найдем $t_{16}(0,05) = 2,12$ – двустороннее t-значение, $t_{16}(0,05) = 1,746$ – одностороннее t-значение.

а) Найдем 95%-й доверительный интервал, в котором находятся затраты на выпуск продукции, для среднего значения объема производства, равного 35 единиц, согласно (7.10). Имеем $x_i = 35$,

$\$_{35} = 2267,091 + 50,943 \cdot 35 = 4050,096$, остальные данные возьмем из расчетной табл. 2. Тогда по формуле (7.11) имеем

$$h_{35} = \frac{1}{18} + \frac{(35 - 32,5)^2}{19743 - 18 \cdot 32,5^2} = 0,064, \text{ тогда}$$

$$4050,096 - 2,12 \cdot 193,193 \sqrt{0,064} \leq \bar{y}_{ген} \leq 4050,096 + 2,12 \cdot 193,193 \sqrt{0,064},$$

$$3946,392 \leq \bar{y}_{ген} \leq 4153,8,$$

т.е. мы на 95% уверены в том, что средние издержки на выпуск 35 единиц продукции для производственного плана в целом (т.е. всегда, когда недельный объем будет запланирован в размере 35 единиц) будут находиться в пределах от 3946,392 у.е. до 4153,8 у.е.

б) Расчитаем 95%-й доверительный интервал, в котором находятся затраты на выпуск продукции со средним объемом производства, равного 35 единиц, на отдельно взятой неделе, согласно (7.12):

$$4050,096 - 2,12 \cdot 193,193 \sqrt{1 + 0,064} \leq \bar{y}_{35} \leq 4050,096 + 2,12 \cdot 193,193 \sqrt{1 + 0,064},$$

$$3627,602 \leq \bar{y}_{35} \leq 4472,59,$$

т.е. с 95%-ой уверенностью можно утверждать, что средние издержки на выпуск 35 единиц продукции на каждой конкретной неделе будут находиться в пределах от 3627,602 у.е. до 4472,59 у.е.

с) Двусторонние доверительные интервалы для истинных наклона и сдвига регрессии. Найдем стандартные ошибки для наклона и сдвига по формулам (7.13 - 7.14):

$$\$ = \frac{193,193}{6,56 \sqrt{17}} = 7,114, \quad \mathfrak{A} = 193,193 \sqrt{\frac{1}{18} + \frac{32,5^2}{6,56^2 \cdot 17}} = 236,59,$$

тогда

$$50,943 - 2,12 \cdot 7,144 \leq b \leq 50,943 + 2,12 \cdot 7,144,$$

$$2267,091 - 2,12 \cdot 236,59 \leq a \leq 2267,091 + 2,12 \cdot 236,59,$$

откуда

$$35,8 \leq b \leq 66,087,$$

$$1765,636 \leq a \leq 2768,646,$$

т.е. можно быть уверенным на 95% в том, что в долгосрочном плане (для генеральной совокупности) размер переменных затрат (b) будет находиться в пределах от 35,8 у.е. до 66,087 у.е., постоянные издержки (a) не покинут интервал от 1765,636 у.е. до 2768,646 у.е.

d) Односторонний доверительный интервал для наклона. Говоря об издержках, естественно задаться целью не допустить перерасход денежных ресурсов, т.е. переменные издержки эффективно оценивать (ограничивать) сверху. Найдем верхнюю границу интервала для истинного наклона:

$50,943 + 1,746 \cdot 7,144 = 63,416$ у.е. – такова сумма, которую не превысят переменные издержки в долгосрочном плане с вероятностью 0,95. Заметем, что она ниже, чем верхняя граница в двустороннем интервале.

4. Проверка гипотезы о значимости коэффициента наклона регрессии.

Зададимся уровнем значимости $\alpha = 0,05$ (что соответствует доверительной вероятности $g = 0,95$) и сформулируем гипотезы:

$H_0 : b = 0$ (связи нет, это случайность),

$H_1 : b \neq 0$ (связь есть).

1) С помощью доверительного интервала для b : в предыдущем пункте мы нашли 95%-й доверительный интервал для истинного наклона регрессии:

$$35,8 \leq b \leq 66,087.$$

Очевидно, значение $b = 0$ не попадает в него, значит следует принять гипотезу H_1 , т.е. действительно имеет место линейная взаимосвязь.

2) С помощью t-критерия Стьюдента: вычислим $t_{набл} = \frac{b}{\hat{\sigma}_b} = \frac{50,943}{7,114} = 7,131$,

найдем табличное значение: $t_{табл} = t_{16}(0,05) = 2,12$. Видим, что $t_{набл} > t_{табл}$, значит принимаем гипотезу H_1 .

Как видим, оба критерия дают один результат, а именно подтверждают наличие взаимосвязи X и Y .

Задачи для самостоятельного решения.

1. Имеются данные⁸ о росте (признак X) и веса (признак Y) студентов группы. Необходимо построить диаграмму разброса и определить по нему характер зависимости, рассчитать коэффициент линейной корреляции Пирсона, сделать вывод о степени тесноты связи между двумя наборами данных, построить уравнение линейной регрессии, оценить качество регрессионной модели и сделать выводы.

2. Изучим зависимость дневной выручки от продажи баночного пива от числа посетителей магазина. Ниже приведены данные исследования 20 случайно выбранных магазинов.

Номер магазина	Число посетителей	Выручка, у.е.
1	907	1120
2	926	1105
3	506	684
4	741	921
5	789	942
6	889	1008
7	874	945
8	510	673
9	529	724
10	420	612
11	679	763
12	872	943
13	924	946
14	607	764
15	452	692
16	729	895
17	794	933
18	844	1023
19	1010	1177
20	621	741

Требуется построить диаграмму разброса и определить по нему характер зависимости, рассчитать коэффициент линейной корреляции Пирсона, сделать вывод о степени тесноты связи между двумя наборами данных, построить уравнение линейной регрессии, оценить качество регрессионной модели и сделать выводы.

3. Туристическая компания предлагает места в гостиницах приморского курорта. Менеджера компании интересует, насколько зависит привлекательность гостиницы от ее удаленности от пляжа. С этой целью по 14 гостиницам города была выяснена среднегодовая наполняемость номеров и расстояние до пляжа (см. таблицу):

⁸ Студентам предлагается собрать данные по своей группе путем письменного опроса.

Расстояние, км	0,1	0,1	0,2	0,3	0,4	0,4	0,5	0,6	0,7	0,7	0,8	0,8	0,9
Наполняемость, %	92	95	96	90	89	86	90	83	85	80	78	76	75

Требуется построить диаграмму разброса и определить по нему характер зависимости, рассчитать коэффициент линейной корреляции Пирсона, сделать вывод о степени тесноты связи между двумя наборами данных, построить уравнение линейной регрессии, оценить качество регрессии и дать интерпретацию полученных результатов.

4. Приведенная ниже таблица содержит данные о массе (г) и объеме (см³) для восьми кусочков алюминия, протестированных в химической лаборатории:

Масса	9,19	12,98	18,19	13,58	10,38	8,11	9,94	12,86
Объем	3,4	5	6,6	5	4	7	3,6	4,8

Создайте диаграмму разброса для зависимости массы от объема. Выполните анализ регрессии зависимой массы от объема без постоянного члена, поскольку кусочек с нулевым объемом должен иметь нулевую массу. Наклон линии регрессии дает оценку плотности алюминия. Найдите 95%-ный доверительный интервал для истинной плотности. Включает ли этот интервал истинное значение 2,699? Сделайте соответствующие выводы.

5. В рассказе А. П. Чехова «Анна на шее» случайно выбран отрывок, состоящий из 10 предложений, в каждом из которых подсчитано количество слов (независимая переменная X) и количество букв (зависимая переменная Y). Данные сведены в таблицу:

Количество слов	3	8	19	41	22	12	35	9	72	53
количество букв	12	41	122	203	106	52	194	42	439	247

Создайте диаграмму разброса для зависимости Y от X , по ее виду сделайте первичный вывод о наличии и тесноты связи. Выполните анализ регрессии. (найдите коэффициенты и их ошибки оценок, прокомментируйте смысл найденных коэффициентов). Найдите 95%-ный доверительный интервал для истинного числа попыток. Сделайте соответствующие выводы.

6. В книге «Основы химии» Д. И. Менделеева приведены данные о растворимости азотнокислого натрия (условные части; Y – зависимая переменная) в зависимости от температуры воды (градусы С, X – независимая переменная):

Растворимость	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1
Температура	0	4	10	15	21	29	36	51	68

Требуется прокомментировать связь между X и Y , построить уравнение регрессии, найти стандартные ошибки, построить доверительные интервалы для всех неизвестных значений генеральной совокупности и сделать выводы.

7. Компанию по прокату автомобилей интересует зависимость между пробегом (X , тыс. км) и стоимостью ежемесячного технического

обслуживания (Y , тыс. у.е.). Для выявления характера этой связи было протестировано 15 автомобилей. Данные сведены в таблицу:

X	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y	13	16	15	20	19	21	26	24	30	32	30	35	34	40	39

По диаграмме разброса определите характер зависимости, найдите коэффициент корреляции Пирсона, постройте уравнение регрессии, спрогнозируйте по нему будущую стоимость обслуживания автомобилей с 7 тыс. и 17 тыс. км пробега; оцените истинные коэффициенты регрессии и средние затраты на ежемесячное обслуживание (генеральное среднее) (с помощью 95%-го доверительного интервала). Прокомментируйте результаты.

8. Компания «N» провела рекламную кампанию в магазинах с демонстрацией антисептических свойств своего нового моющего средства. Через 10 недель компания решила проанализировать эффективность этого вида рекламы, сопоставив еженедельные объемы продаж (Y , тыс. у.е.) с расходами на рекламу (X , тыс. у.е.):

Объем продаж	72	76	78	70	68	80	82	65	62	90
Расходы на рекламу	5	8	6	5	3	9	12	4	3	10

По диаграмме разброса определите характер зависимости, найдите коэффициент корреляции Пирсона, постройте уравнение регрессии, спрогнозируйте по нему объемы продаж на следующей неделе при вложении в рекламу 7 и 11 тыс. у.е.; оцените истинные коэффициенты регрессии и средний объем продаж моющего средства (генеральное среднее) (с помощью 95%-го доверительного интервала). Прокомментируйте результаты.

9. Опрос 10 случайно выбранных студентов, проживающих в общежитии университета, позволяет выявить зависимость между средним баллом по результатам последней сессии и числом часов в неделю, затраченных студентом на самостоятельную подготовку:

Средний балл	4,6	4,3	3,8	3,8	4,2	4,3	3,8	4,0	3,1	3,9
Число часов	25	22	9	15	15	30	20	30	10	17

Выберете независимую и зависимую переменные, найдите коэффициент корреляции Пирсона и сделайте вывод о степени тесноты линейной связи; постройте уравнение регрессии, оцените истинные коэффициенты регрессии и генеральное среднее (для зависимой переменной!) (с помощью 95%-го доверительного интервала). Прокомментируйте результаты.

Литература:

1. Сигел Э. Практическая бизнес-статистика.: Пер. с англ. – М.: Издательский дом “Вильямс”, 2004. – 1056 с.
2. Ниворижкина П. П., Морозова З. А., Герасимова П. А., Житников П. В.. Основы статистики с элементами теории вероятностей для экономистов. Руководство для решения задач. Ростов-на-Дону: «Феникс», 1999. – 156 с.

Приложение. Таблица №1: значения функций

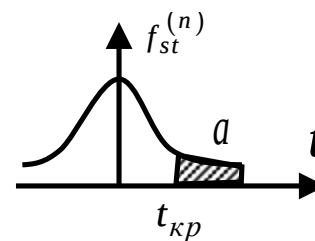
$$F_N(x) = \frac{1}{\sqrt{2p}} \int_{-\infty}^x e^{-t^2/2} dt$$

$$\Phi(x) = \frac{1}{\sqrt{2p}} \int_0^x e^{-t^2/2} dt$$

x	$\Phi(x)$	$F_N(x)$	x	$\Phi(x)$	$F_N(x)$	x	$\Phi(x)$	$F_N(x)$
0	0	0,5	1,3	0,4032	0,9032	2,6	0,4953	0,9953
0,05	0,0199	0,5199	1,35	0,4115	0,9115	2,65	0,496	0,996
0,1	0,0398	0,5398	1,4	0,4192	0,9192	2,7	0,4965	0,9965
0,15	0,0596	0,5596	1,45	0,4265	0,9265	2,75	0,497	0,997
0,2	0,0793	0,5793	1,5	0,4332	0,9332	2,8	0,4974	0,9974
0,25	0,0987	0,5987	1,55	0,4394	0,9394	2,85	0,4978	0,9978
0,3	0,1179	0,6179	1,6	0,4452	0,9452	2,9	0,4981	0,9981
0,35	0,1368	0,6368	1,65	0,4505	0,9505	2,95	0,4984	0,9984
0,4	0,1554	0,6554	1,7	0,4554	0,9554	3	0,4987	0,9987
0,45	0,1736	0,6736	1,75	0,4599	0,9599	3,05	0,4989	0,9989
0,5	0,1915	0,6915	1,8	0,4641	0,9641	3,1	0,499	0,999
0,55	0,2088	0,7088	1,85	0,4678	0,9678	3,15	0,4992	0,9992
0,6	0,2257	0,7257	1,9	0,4713	0,9713	3,2	0,4993	0,9993
0,65	0,2422	0,7422	1,95	0,4744	0,9744	3,25	0,4994	0,9994
0,7	0,258	0,758	2	0,4772	0,9772	3,3	0,4995	0,9995
0,75	0,2734	0,7734	2,05	0,4798	0,9798	3,35	0,4996	0,9996
0,8	0,2881	0,7881	2,1	0,4821	0,9821	3,4	0,4997	0,9997
0,85	0,3023	0,8023	2,15	0,4842	0,9842	3,45	0,4997	0,9997
0,9	0,3159	0,8159	2,2	0,4861	0,9861	3,5	0,4998	0,9998
0,95	0,3289	0,8289	2,25	0,4878	0,9878	3,55	0,4998	0,9998
1	0,3413	0,8413	2,3	0,4893	0,9893	3,6	0,4998	0,9998
1,05	0,3531	0,8531	2,35	0,4906	0,9906	3,65	0,4999	0,9999
1,1	0,3643	0,8643	2,4	0,4918	0,9918	3,7	0,4999	0,9999
1,15	0,3749	0,8749	2,45	0,4929	0,9929	3,75	0,4999	0,9999
1,2	0,3849	0,8849	2,5	0,4938	0,9938	3,8	0,4999	0,9999
1,25	0,3944	0,8944	2,55	0,4946	0,9946	3,85	0,4999	0,9999

Приложение. Таблица №2
Распределение Стьюдента,

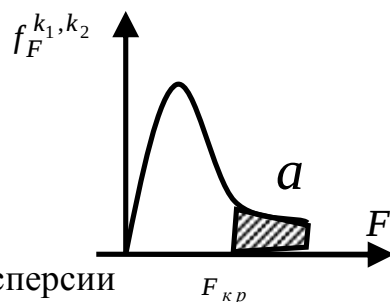
$$\int_{t_{кр}}^{\infty} f_{st}^{(n)}(t) dt = a,$$



<i>k</i>	уровень значимости <i>a</i>															
	0,5	0,4	0,3	0,2	0,15	0,1	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0,005
1	1	1,38	1,96	3,08	4,17	6,31	7,03	7,92	9,06	10,58	12,71	15,89	21,2	31,82	63,66	127,32
2	0,82	1,06	1,39	1,89	2,28	2,92	3,1	3,32	3,58	3,9	4,3	4,85	5,64	6,96	9,92	14,09
3	0,76	0,98	1,25	1,64	1,92	2,35	2,47	2,61	2,76	2,95	3,18	3,48	3,9	4,54	5,84	7,45
4	0,74	0,94	1,19	1,53	1,78	2,13	2,23	2,33	2,46	2,6	2,78	3	3,3	3,75	4,6	5,6
5	0,73	0,92	1,16	1,48	1,7	2,02	2,1	2,19	2,3	2,42	2,57	2,76	3	3,36	4,03	4,77
6	0,72	0,91	1,13	1,44	1,65	1,94	2,02	2,1	2,2	2,31	2,45	2,61	2,83	3,14	3,71	4,32
7	0,71	0,9	1,12	1,41	1,62	1,89	1,97	2,05	2,14	2,24	2,36	2,52	2,71	3	3,5	4,03
8	0,71	0,89	1,11	1,4	1,59	1,86	1,93	2	2,09	2,19	2,31	2,45	2,63	2,9	3,36	3,83
9	0,7	0,88	1,1	1,38	1,57	1,83	1,9	1,97	2,06	2,15	2,26	2,4	2,57	2,82	3,25	3,69
10	0,7	0,88	1,09	1,37	1,56	1,81	1,88	1,95	2,03	2,12	2,23	2,36	2,53	2,76	3,17	3,58
11	0,7	0,88	1,09	1,36	1,55	1,8	1,86	1,93	2,01	2,1	2,2	2,33	2,49	2,72	3,11	3,5
12	0,7	0,87	1,08	1,36	1,54	1,78	1,84	1,91	1,99	2,08	2,18	2,3	2,46	2,68	3,05	3,43
13	0,69	0,87	1,08	1,35	1,53	1,77	1,83	1,9	1,97	2,06	2,16	2,28	2,44	2,65	3,01	3,37
14	0,69	0,87	1,08	1,35	1,52	1,76	1,82	1,89	1,96	2,05	2,14	2,26	2,41	2,62	2,98	3,33
15	0,69	0,87	1,07	1,34	1,52	1,75	1,81	1,88	1,95	2,03	2,13	2,25	2,4	2,6	2,95	3,29
16	0,69	0,86	1,07	1,34	1,51	1,75	1,8	1,87	1,94	2,02	2,12	2,24	2,38	2,58	2,92	3,25
17	0,69	0,86	1,07	1,33	1,51	1,74	1,8	1,86	1,93	2,02	2,11	2,22	2,37	2,57	2,9	3,22
18	0,69	0,86	1,07	1,33	1,5	1,73	1,79	1,86	1,93	2,01	2,1	2,21	2,36	2,55	2,88	3,2
19	0,69	0,86	1,07	1,33	1,5	1,73	1,79	1,85	1,92	2	2,09	2,2	2,35	2,54	2,86	3,17
20	0,69	0,86	1,06	1,33	1,5	1,72	1,78	1,84	1,91	1,99	2,09	2,2	2,34	2,53	2,85	3,15
21	0,69	0,86	1,06	1,32	1,49	1,72	1,78	1,84	1,91	1,99	2,08	2,19	2,33	2,52	2,83	3,14
22	0,69	0,86	1,06	1,32	1,49	1,72	1,77	1,84	1,9	1,98	2,07	2,18	2,32	2,51	2,82	3,12
23	0,69	0,86	1,06	1,32	1,49	1,71	1,77	1,83	1,9	1,98	2,07	2,18	2,31	2,5	2,81	3,1
24	0,68	0,86	1,06	1,32	1,49	1,71	1,77	1,83	1,9	1,97	2,06	2,17	2,31	2,49	2,8	3,09
25	0,68	0,86	1,06	1,32	1,49	1,71	1,76	1,82	1,89	1,97	2,06	2,17	2,3	2,49	2,79	3,08
30	0,68	0,85	1,05	1,31	1,48	1,7	1,75	1,81	1,88	1,95	2,04	2,15	2,28	2,46	2,75	3,03
60	0,68	0,85	1,05	1,3	1,46	1,67	1,72	1,78	1,84	1,92	2	2,1	2,22	2,39	2,66	2,91
100	0,68	0,85	1,04	1,29	1,45	1,66	1,71	1,77	1,83	1,9	1,98	2,08	2,2	2,36	2,63	2,87
∞	0,67	0,84	1,04	1,28	1,44	1,64	1,7	1,75	1,81	1,88	1,96	2,05	2,17	2,33	2,58	2,81

Приложение. Таблица №3
Распределение Фишера

$$\int_{F_{кр}}^{\infty} f_F^{(k_1, k_2)}(F) dF = a,$$



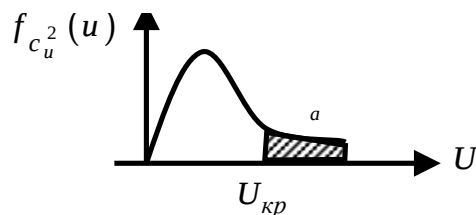
k_1 – число степеней свободы для большей дисперсии

k_2 – число степеней свободы для меньшей дисперсии

уровень значимости $\alpha = 0,01$												
k_1	1	2	3	4	5	6	7	8	9	10	11	12
k_2												
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6083	6106
2	98,5	99	99,2	99,2	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	27,1
4	21,2	18	16,7	16	15,5	15,2	15	14,8	14,7	14,5	14,5	14,4
5	16,3	13,3	12,1	11,4	11	10,7	10,5	10,3	10,2	10,1	9,96	9,89
6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,1	7,98	7,87	7,79	7,72
7	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47
8	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67
9	10,6	8,02	6,99	6,42	6,06	5,8	5,61	5,47	5,35	5,26	5,18	5,11
10	10	7,56	6,55	5,99	5,64	5,39	5,2	5,06	4,94	4,85	4,77	4,71
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,4
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,5	4,39	4,3	4,22	4,16
13	9,07	6,7	5,74	5,21	4,86	4,62	4,44	4,3	4,19	4,1	4,02	3,96
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,8
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4	3,89	3,8	3,73	3,67
16	8,53	6,23	5,29	4,77	4,44	4,2	4,03	3,89	3,78	3,69	3,62	3,55
17	8,4	6,11	5,18	4,67	4,34	4,1	3,93	3,79	3,68	3,59	3,52	3,46
уровень значимости $\alpha = 0,05$												
k_1	1	2	3	4	5	6	7	8	9	10	11	12
k_2												
1	161	199	216	225	230	234	237	239	241	242	243	244
2	18,5	19	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6	5,96	5,94	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,7	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,1	4,06	4,03	4
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,6	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,5	3,44	3,39	3,35	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,1	3,07
10	4,96	4,1	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91
11	4,84	3,98	3,59	3,36	3,2	3,09	3,01	2,95	2,9	2,85	2,82	2,79
12	4,75	3,89	3,49	3,26	3,11	3	2,91	2,85	2,8	2,75	2,72	2,69
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,6
14	4,6	3,74	3,34	3,11	2,96	2,85	2,76	2,7	2,65	2,6	2,57	2,53
15	4,54	3,68	3,29	3,06	2,9	2,79	2,71	2,64	2,59	2,54	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42
17	4,45	3,59	3,2	2,96	2,81	2,7	2,61	2,55	2,49	2,45	2,41	2,38

Приложение. Таблица №4
Распределение Хи-квадрат

$$\int_{U_{кр}}^{\infty} f_{\chi^2_u}(u) du = a,$$



число степеней свободы	уровень значимости α															
	0,5	0,4	0,3	0,2	0,15	0,1	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0,005
1	0,455	0,708	1,074	1,642	2,072	2,706	2,874	3,065	3,283	3,537	3,841	4,218	4,709	5,412	6,635	7,87944
2	1,386	1,833	2,408	3,219	3,794	4,605	4,816	5,051	5,319	5,627	5,991	6,438	7,013	7,824	9,21	10,5966
3	2,366	2,946	3,665	4,642	5,317	6,251	6,491	6,759	7,06	7,407	7,815	8,311	8,947	9,837	11,34	12,8382
4	3,357	4,045	4,878	5,989	6,745	7,779	8,043	8,337	8,666	9,044	9,488	10,03	10,71	11,67	13,28	14,8603
5	4,351	5,132	6,064	7,289	8,115	9,236	9,521	9,837	10,19	10,6	11,07	11,64	12,37	13,39	15,09	16,7496
6	5,348	6,211	7,231	8,558	9,446	10,64	10,95	11,28	11,66	12,09	12,59	13,2	13,97	15,03	16,81	18,5476
7	6,346	7,283	8,383	9,803	10,75	12,02	12,34	12,69	13,09	13,54	14,07	14,7	15,51	16,62	18,48	20,2777
8	7,344	8,351	9,524	11,03	12,03	13,36	13,7	14,07	14,48	14,96	15,51	16,17	17,01	18,17	20,09	21,955
9	8,343	9,414	10,66	12,24	13,29	14,68	15,03	15,42	15,85	16,35	16,92	17,61	18,48	19,68	21,67	23,5894
10	9,342	10,47	11,78	13,44	14,53	15,99	16,35	16,75	17,2	17,71	18,31	19,02	19,92	21,16	23,21	25,1882
11	10,34	11,53	12,9	14,63	15,77	17,28	17,65	18,07	18,53	19,06	19,68	20,41	21,34	22,62	24,72	26,7568
12	11,34	12,58	14,01	15,81	16,99	18,55	18,94	19,37	19,85	20,39	21,03	21,79	22,74	24,05	26,22	28,2995
13	12,34	13,64	15,12	16,98	18,2	19,81	20,21	20,66	21,15	21,71	22,36	23,14	24,12	25,47	27,69	29,8195
14	13,34	14,69	16,22	18,15	19,41	21,06	21,48	21,93	22,44	23,02	23,68	24,49	25,49	26,87	29,14	31,3193
15	14,34	15,73	17,32	19,31	20,6	22,31	22,73	23,2	23,72	24,31	25	25,82	26,85	28,26	30,58	32,8013
16	15,34	16,78	18,42	20,47	21,79	23,54	23,98	24,46	24,99	25,59	26,3	27,14	28,19	29,63	32	34,2672
17	16,34	17,82	19,51	21,61	22,98	24,77	25,21	25,71	26,25	26,87	27,59	28,44	29,52	31	33,41	35,7185
18	17,34	18,87	20,6	22,76	24,16	25,99	26,45	26,95	27,5	28,14	28,87	29,75	30,84	32,35	34,81	37,1565
19	18,34	19,91	21,69	23,9	25,33	27,2	27,67	28,18	28,75	29,4	30,14	31,04	32,16	33,69	36,19	38,5823
20	19,34	20,95	22,77	25,04	26,5	28,41	28,89	29,41	29,99	30,65	31,41	32,32	33,46	35,02	37,57	39,9968
21	20,34	21,99	23,86	26,17	27,66	29,62	30,1	30,63	31,22	31,89	32,67	33,6	34,76	36,34	38,93	41,4011
22	21,34	23,03	24,94	27,3	28,82	30,81	31,31	31,85	32,45	33,13	33,92	34,87	36,05	37,66	40,29	42,7957
23	22,34	24,07	26,02	28,43	29,98	32,01	32,51	33,06	33,68	34,37	35,17	36,13	37,33	38,97	41,64	44,1813
24	23,34	25,11	27,1	29,55	31,13	33,2	33,71	34,27	34,89	35,6	36,42	37,39	38,61	40,27	42,98	45,5585
25	24,34	26,14	28,17	30,68	32,28	34,38	34,9	35,47	36,11	36,82	37,65	38,64	39,88	41,57	44,31	46,9279
30	29,34	31,32	33,53	36,25	37,99	40,26	40,82	41,43	42,11	42,88	43,77	44,83	46,16	47,96	50,89	53,672
60	59,33	62,13	65,23	68,97	71,34	74,4	75,15	75,97	76,88	77,9	79,08	80,48	82,23	84,58	88,38	91,9517
90	89,33	92,76	96,52	101,1	103,9	107,6	108,5	109,4	110,5	111,7	113,1	114,8	116,9	119,6	124,1	128,299
120	119,3	123,3	127,6	132,8	136,1	140,2	141,3	142,4	143,6	145	146,6	148,4	150,8	153,9	159	163,648

**Аниковский Владимир Васильевич
Воронцов, Денис Евгеньевич
Морозов Владимир Павлович
Солычева Ольга Михайловна
Тютин Виктор Владимирович**

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА