ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ОБРАЗОВАНИЯ "МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)"

Институт №8 «Компьютерные науки и прикладная математика»
Кафедра 804 «Теория вероятностей и компьютерное моделирование»

Курсовая работа на тему:
«Метод наименьших квадратов»

Вариант 48

Выполнил:
Проверил:

Оценка: _____

Дата: _____

Описание модели

Модель полезного сигнала имеет вид:

$$y(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \tag{1}$$

Рассматривается модель наблюдений:

$$y_k = \theta_0 + \theta_1 x_k + \theta_2 x_k^2 + \varepsilon_k, \quad x_k = -1 + \frac{k-1}{20}, \quad k = \overline{1,41} \tag{2}$$

где $\varepsilon_1, \ldots, \varepsilon_{41}$ – независимые центрированные и одинаково распределённые случайные величины.

Выборка $Y = (Y_1, \ldots, Y_n)^T$ для $n = 41$ находится в файле «DataN.xlsx», где «N» – номер варианта.

Задание

Для полученной выборки, полагая ошибки наблюдений нормальными $\varepsilon_k \sim \mathcal{N}(0, \sigma^2),\ k = \overline{1,n}$, выполнить следующие задания.

1. Вычислить оценки неизвестных параметров $\theta_0, \theta_1, \theta_2$ методом наименьших квадратов.

2. Построить доверительные интервалы уровней надёжности $\alpha_1 = 0.95$ и $\alpha_2 = 0.99$ для параметров $\theta_0, \theta_1, \theta_2$.

3. Вычислить оценку максимального правдоподобия дисперсии σ^2 случайной ошибки.

4. Построить доверительные интервалы уровней надёжности $\alpha_1 = 0.95$ и $\alpha_2 = 0.99$ дисперсии σ^2 случайной ошибки.

5. Построить доверительные интервалы уровней надёжности $\alpha_1 = 0.95$ и $\alpha_2 = 0.99$ для полезного сигнала (1).

6. Изобразить на одном графике

   - набор наблюдений,
   - оценку полезного сигнала, полученную в шаге 1,
   - доверительные интервалы полезного сигнала, полученные в шаге 5.

7. По остаткам регрессии построить оценку плотности распределения случайной ошибки наблюдения в виде гистограммы.

8. По остаткам регрессии с помощью χ^2-критерия Пирсона на уровне значимости 0.05 проверить гипотезу о том, что закон распределения ошибки наблюдения является нормальным.

Задание №1

Опр. *Оценкой МНК* параметра θ называют решение задачи минимизации

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{n} \big(y_k - \varphi^T(x_k)\theta\big)^2 = \arg\min_{\theta} (Y - X\theta)^T (Y - X\theta),$$

где $Y = (y_1, \ldots, y_n)^T$, а X — регрессионная матрица.

ТЕОРЕМА (Гаусса–Маркова). Если $\det(X^T X) \neq 0$, то МНК-оценка существует, единственна и задаётся формулой

$$\hat{\theta}_{\text{МНК}} = (X^T X)^{-1} X^T Y.$$

Рассмотрим модель

$$y_k = \theta_0 + \theta_1 x_k + \theta_2 x_k^2 + \varepsilon_k, \qquad x_k = -1 + \frac{k-1}{20}, \quad k = \overline{1,41}\ (n = 41).$$

Тогда

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{41} & x_{41}^2 \end{pmatrix} = \begin{pmatrix} 1 & -1.00 & 1.0000 \\ 1 & -0.95 & 0.9025 \\ 1 & -0.90 & 0.8100 \\ \vdots & \vdots & \vdots \\ 1 & 0.95 & 0.9025 \\ 1 & 1.00 & 1.0000 \end{pmatrix}, \qquad Y = \begin{pmatrix} 10{,}79206531 \\ 12{,}25630223 \\ 14{,}65453580 \\ 13{,}44450727 \\ 15{,}78195805 \\ \vdots \\ 17{,}98506447 \\ 19{,}79293941 \\ 18{,}93057624 \end{pmatrix}.$$

По формуле МНК имеем (как произведение матриц):

$$\hat{\theta}_{\text{МНК}} = (X^T X)^{-1} X^T Y = \left(\begin{pmatrix} 1 & -1.00 & 1.0000 \\ 1 & -0.95 & 0.9025 \\ 1 & -0.90 & 0.8100 \\ \vdots & \vdots & \vdots \\ 1 & 0.95 & 0.9025 \\ 1 & 1.00 & 1.0000 \end{pmatrix}^T \begin{pmatrix} 1 & -1.00 & 1.0000 \\ 1 & -0.95 & 0.9025 \\ 1 & -0.90 & 0.8100 \\ \vdots & \vdots & \vdots \\ 1 & 0.95 & 0.9025 \\ 1 & 1.00 & 1.0000 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & -1.00 & 1.0000 \\ 1 & -0.95 & 0.9025 \\ 1 & -0.90 & 0.8100 \\ \vdots & \vdots & \vdots \\ 1 & 0.95 & 0.9025 \\ 1 & 1.00 & 1.0000 \end{pmatrix}^T \begin{pmatrix} 10{,}79206531 \\ 12{,}25630223 \\ 14{,}65453580 \\ \vdots \\ 19{,}79293941 \\ 18{,}93057624 \end{pmatrix}$$

В результате вычислений получаем:

$$\hat{\theta}_{\text{МНК}} = \begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \begin{pmatrix} 23{,}84383862 \\ 3{,}62324951 \\ -8{,}82441494 \end{pmatrix}.$$

Следовательно, оценка полезного сигнала имеет вид

$$\hat{y}(x) = \hat{\theta}_0 + \hat{\theta}_1 x + \hat{\theta}_2 x^2 = 23{,}84383862 + 3{,}62324951\, x - 8{,}82441494\, x^2.$$

Вид функции $\varphi(x; \hat{\theta}_{\text{МНК}})$ (с округлением):

$$\varphi(x; \hat{\theta}_{\text{МНК}}) \approx 23.84 + 3.62x - 8.82x^2.$$

Задание №2

Опр. *Доверительным интервалом* уровня надёжности γ для параметра θ называется интервал $[\theta^-, \theta^+]$, построенный по выборке, такой что

$$\mathbb{P}\big(\theta^- \le \theta \le \theta^+\big) = \gamma, \qquad \alpha = 1 - \gamma.$$

Центральный доверительный интервал соответствует выбору квантиля $t_{1-\alpha/2}$.

Опр. Схема Гаусса–Маркова называется *нормальной регрессией*, если ошибки наблюдений имеют нормальное распределение

$$\varepsilon \sim N(0, \sigma^2 I).$$

ЛЕММА. В нормальной регрессии для каждого параметра θ_k выполняется:

$$\frac{\hat{\theta}_k - \theta_k}{\|\hat{\varepsilon}\| \sqrt{a_k}/\sqrt{n-s}} \sim t(n-s),$$

где $\hat{\varepsilon} = Y - X\hat{\theta}_{\text{МНК}}$, a_k — k-й элемент главной диагонали матрицы $(X^TX)^{-1}$, n — объём выборки, s — число оцениваемых параметров (в данной задаче $n = 41$, $s = 3$).

Тогда при $\alpha = 1 - \gamma$ получаем:

$$\mathbb{P}\left(-t_{1-\alpha/2}(n-s) \le \frac{\hat{\theta}_k - \theta_k}{\|\hat{\varepsilon}\| \sqrt{a_k}/\sqrt{n-s}} \le t_{1-\alpha/2}(n-s) \right) = 1 - \alpha,$$

откуда центральный доверительный интервал:

$$\theta_k \in \left[\hat{\theta}_k - t_{1-\alpha/2}(n-s) \frac{\|\hat{\varepsilon}\|\sqrt{a_k}}{\sqrt{n-s}},\ \hat{\theta}_k + t_{1-\alpha/2}(n-s) \frac{\|\hat{\varepsilon}\|\sqrt{a_k}}{\sqrt{n-s}} \right].$$

Решение. По результатам вычислений получены центральные доверительные интервалы:

Параметр	$\gamma = 0.95$	$\gamma = 0.99$
θ_0	$[23.35582773;\ 24.33184950]$	$[23.19017657;\ 24.49750066]$
θ_1	$[3.07359661;\ 4.17290242]$	$[2.88702159;\ 4.35947743]$
θ_2	$[-9.86408998;\ -7.78473990]$	$[-10.21699886;\ -7.43183102]$

При увеличении уровня надёжности с 0.95 до 0.99 интервалы расширяются.

Задание №3

ЛЕММА. В нормальной регрессии МП-оценка дисперсии σ^2 случайной ошибки имеет вид:

$$\hat{\sigma}^2_{\text{МП}} = \frac{1}{n} \|Y - X\hat{\theta}\|^2,$$

где $\hat{\theta}$ — МНК-оценка параметров, n — объём выборки.

Решение. Логарифм функции правдоподобия (с точностью до константы) равен

$$\ln L(Y; \theta, \sigma^2) = -\frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\theta)^T(Y - X\theta).$$

Дифференцируя по σ^2 и приравнивая к нулю, получаем

$$\frac{\partial}{\partial \sigma^2} \ln L(Y; \theta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2}\frac{(Y - X\theta)^T(Y - X\theta)}{(\sigma^2)^2} = 0,$$

$$\Rightarrow \quad \hat{\sigma}^2_{\text{МП}} = \frac{1}{n}(Y - X\theta)^T(Y - X\theta).$$

Подставляя $\theta = \hat{\theta}$ (оценка, найденная методом МНК), получаем:

$$\hat{\sigma}^2_{\text{МП}} = \frac{1}{n}(Y - X\hat{\theta})^T(Y - X\hat{\theta}) = \frac{1}{n} \|Y - X\hat{\theta}\|^2.$$

По результатам вычислений:

$$\hat{\sigma}^2_{\text{МП}} = 0.9804795454480453.$$

Задание №4

ЛЕММА в нормальной регрессии:

$$\frac{\|\hat{\varepsilon}\|^2}{\sigma^2} \sim \chi^2(n - s),$$

где $\hat{\varepsilon} = Y - X\hat{\theta}_{\text{МНК}}$ — оценка вектора остатков, n — объём выборки, s — число оцениваемых параметров.

Тогда для $\alpha = 1 - \gamma$ имеем:

$$\mathbb{P}\left(\chi^2_{\alpha/2}(n - s) \leq \frac{\|\hat{\varepsilon}\|^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}(n - s)\right) = 1 - \alpha,$$

откуда центральный доверительный интервал для σ^2:

$$\mathbb{P}\left(\frac{\|\hat{\varepsilon}\|^2}{\chi^2_{1-\alpha/2}(n-s)} \le \sigma^2 \le \frac{\|\hat{\varepsilon}\|^2}{\chi^2_{\alpha/2}(n-s)}\right) = 1-\alpha.$$

Решение. Центральные доверительные интервалы для дисперсии случайной ошибки σ^2:

а) Уровень надёжности $\gamma_1 = 0.95$ ($\alpha_1 = 0.05$):

$$\sigma^2 \in [0.70655231;\ 1.75709476].$$

б) Уровень надёжности $\gamma_2 = 0.99$ ($\alpha_2 = 0.01$):

$$\sigma^2 \in [0.62634429;\ 2.08408138].$$

Задание №5

ЛЕММА в нормальной регрессии:

$$\frac{\varphi(x;\hat{\theta}) - \varphi(x;\theta)}{\|\hat{\varepsilon}\|\sqrt{\alpha(x)}}\sqrt{n-s} \ \sim\ t(n-s),$$

где введено обозначение

$$\alpha(x) = \varphi(x)^T (X^T X)^{-1}\varphi(x), \qquad \varphi(x) = (\varphi_1(x), \ldots, \varphi_s(x))^T.$$

Тогда:

$$\mathbb{P}\left(-t_{1-\alpha/2}(n-s) \le \frac{\varphi(x;\hat{\theta}) - \varphi(x;\theta)}{\|\hat{\varepsilon}\|\sqrt{\alpha(x)}}\sqrt{n-s} \le t_{1-\alpha/2}(n-s)\right) = 1-\alpha,$$

$$\mathbb{P}\left(\varphi(x;\hat{\theta}) - t_{1-\alpha/2}(n-s)\frac{\|\hat{\varepsilon}\|\sqrt{\alpha(x)}}{\sqrt{n-s}} \le \varphi(x;\theta) \le \varphi(x;\hat{\theta}) + t_{1-\alpha/2}(n-s)\frac{\|\hat{\varepsilon}\|\sqrt{\alpha(x)}}{\sqrt{n-s}}\right) = 1-\alpha.$$

Найдём $\alpha(x)$:

$$\alpha(x) = 0.055 - 0.105x^2 + 0.249x^4.$$

Центральные доверительные интервалы для полезного сигнала:

1. Для уровня надёжности $1-\alpha_1 = 0.95$:

$$\varphi(x;\theta) \in \begin{bmatrix} 23.84 + 3.62x - 8.82x^2 - 2.082161805421579\sqrt{0.055 - 0.105x^2 + 0.249x^4}, \\ 23.84 + 3.62x - 8.82x^2 + 2.082161805421579\sqrt{0.055 - 0.105x^2 + 0.249x^4} \end{bmatrix}.$$

2. Для уровня надёжности $1-\alpha_2 = 0.99$:

$$\varphi(x;\theta) \in \begin{bmatrix} 23.84 + 3.62x - 8.82x^2 - 2.7889339796325467\sqrt{0.055 - 0.105x^2 + 0.249x^4}, \\ 23.84 + 3.62x - 8.82x^2 + 2.7889339796325467\sqrt{0.055 - 0.105x^2 + 0.249x^4} \end{bmatrix}.$$

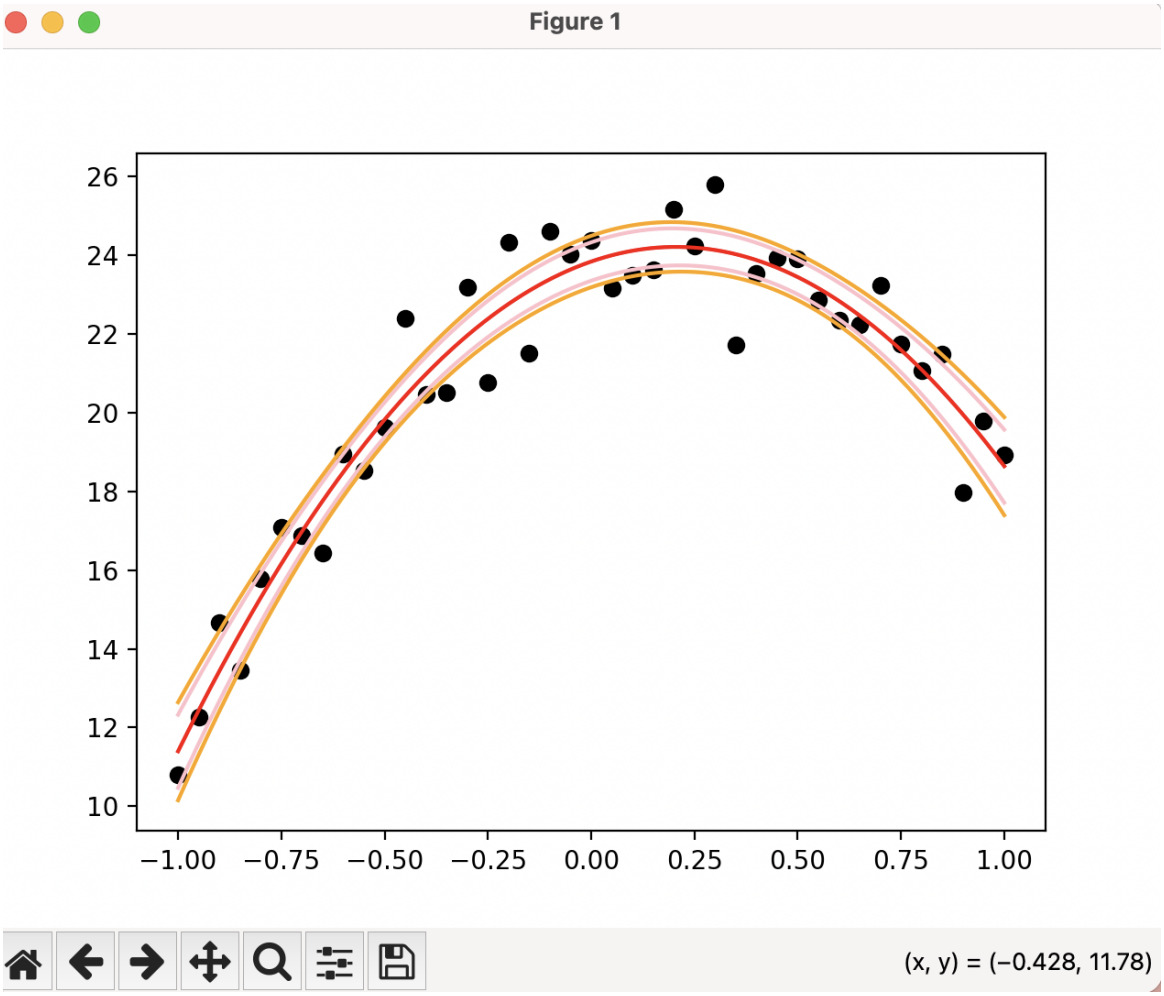

Рис. 1: Набор наблюдений, МНК-оценка полезного сигнала и доверительные интервалы. Графики построены на языке программирования Python с использованием библиотеки `matplotlib`.

Задание №6

Задание №7

По остаткам регрессии построим оценку плотности распределения случайной ошибки наблюдения в виде гистограммы. Вектор остатков (оценка вектора ошибок) вычисляется по формуле

$$\hat{\varepsilon} = Y - X\hat{\theta}_{\text{МНК}}.$$

Вектор остатков (значения округлены до 4 знаков после запятой):

$$\hat{\varepsilon} \approx \begin{pmatrix} -2.3116 & -1.9719 & -1.6196 & -1.5764 & -1.3335 & -0.9763 & -0.9439 \\ -0.8441 & -0.6602 & -0.6214 & -0.6041 & -0.5534 & -0.5272 & -0.4879 \\ -0.3360 & -0.2986 & -0.2297 & -0.2010 & -0.1814 & -0.1001 & -0.0234 \\ 0.0368 & 0.1468 & 0.2518 & 0.2879 & 0.3829 & 0.4523 & 0.4530 \\ 0.4710 & 0.4843 & 0.5379 & 0.9292 & 0.9447 & 0.9566 & 1.1898 \\ 1.2181 & 1.2194 & 1.2309 & 1.5642 & 1.6688 & 1.9751 & \end{pmatrix}.$$

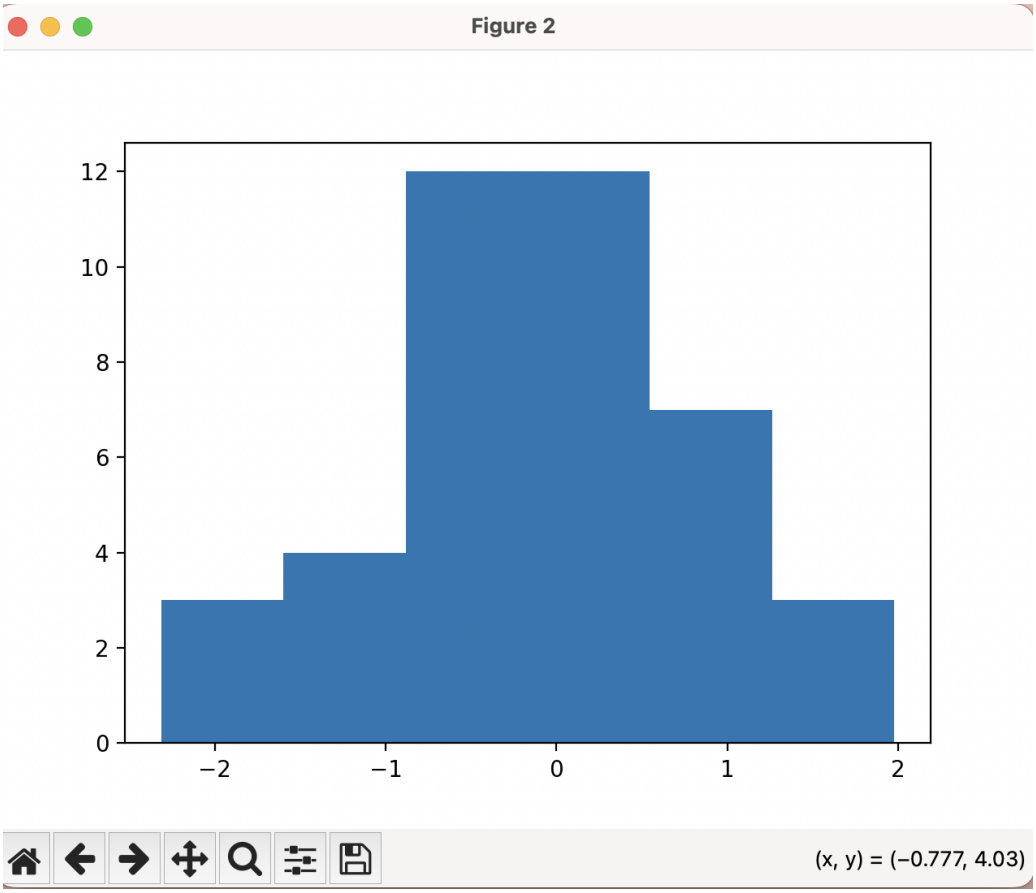

Рис. 2: Гистограмма остатков регрессии. График построен на языке программирования Python с использованием библиотеки matplotlib.

Задание №8

Проверим гипотезу о нормальности распределения ошибки наблюдения с помощью критерия Пирсона (χ^2).

Опр. Пусть по выборке остатков $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$ задано разбиение числовой оси

$$-\infty = t_0 < t_1 < \cdots < t_l < t_{l+1} = +\infty,$$

где $t_1 \leq \min\limits_k \hat{\varepsilon}_k$, $t_l > \max\limits_k \hat{\varepsilon}_k$. В нашей работе используется разбиение на 6 полуинтервалов, поэтому

$$l = 7.$$

Обозначим числа попаданий и относительные частоты:

n_i — число элементов выборки, которые попали в полуинтервал $[t_i, t_{i+1})$, $\quad i = \overline{1, l-1}$,

$$\hat{p}_i = \frac{n_i}{n}, \qquad \hat{p}_0 = \hat{p}_l = 0.$$

Рассмотрим гипотезу

$$H_0 : \ \varepsilon_k \sim \mathcal{N}(0, \sigma^2).$$

Тогда при H_0 функция распределения имеет вид

$$F_0(x) = \Phi\left(\frac{x}{\hat{\sigma}_{\text{МП}}}\right),$$

а теоретические вероятности попадания в полуинтервалы равны

$$p_i = F_0(t_{i+1}) - F_0(t_i) = \Phi\left(\frac{t_{i+1}}{\hat{\sigma}_{\text{МП}}}\right) - \Phi\left(\frac{t_i}{\hat{\sigma}_{\text{МП}}}\right).$$

Статистика критерия Пирсона:

$$T(Z_n) = n \sum_{i=0}^{l} \frac{(\hat{p}_i - p_i)^2}{p_i}.$$

ТЕОРЕМА. Если гипотеза H_0 верна, то

$$T(Z_n) \ \xrightarrow{d} \ \chi^2(l - s),$$

где s — число параметров распределения F_0, оценённых по выборке. В нашем случае(toggle) оценивается только параметр σ, поэтому $s = 1$, а значит

$$l - s = 7 - 1 = 6.$$

Критерий на уровне значимости $\alpha = 0.05$:

$$\chi^2_{\alpha/2}(l - s) \ < \ T(Z_n) \ < \ \chi^2_{1-\alpha/2}(l - s).$$

Проверка гипотезы (результат расчётов):

$$1.237344245791203 < 1.6642471553326905 < 14.44937533544792.$$

Так как двойное неравенство выполняется, **гипотеза H_0 принимается**, то есть распределение ошибки наблюдения можно считать нормальным.