Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

Отчет по лабораторной работе №1 по дисциплине «Математическая статистика»

Тема Гистограмма и эмпирическая функция распределения

Студент Леонов В.В.

Группа ИУ7-66Б

Оценка (баллы) __________

Преподаватель Власов П.А., Андреева Т.В.

Москва — 2022 г.

Постановка задачи

Цель работы: построение гистограммы и эмпирической функции распределения.

1. Для выборки объёма n из генеральной совокупности X реализовать в виде программы на ЭВМ
   (a) вычисление максимального значения $M_{\max}$ и минимального значения $M_{\min}$;
   (b) размаха R выборки;
   (c) вычисление оценок $\hat{\mu}$ и S^2 математического ожидания MX и дисперсии DX;
   (d) группировку значений выборки в $m = [\log_2 n] + 2$ интервала;
   (e) построение на одной координатной плоскости гистограммы и графика функции плотности распределения вероятностей нормальной случайной величины с математическим ожиданием $\hat{\mu}$ и дисперсией S^2;
   (f) построение на другой координатной плоскости графика эмпирической функции распределения и функции распределения нормальной случайной величины с математическим ожиданием $\hat{\mu}$ и дисперсией S^2.
2. Провести вычисления и построить графики для выборки из индивидуального варианта.

Вариант 9:
X = (-8.47, -7.45, -7.12, -8.30, -8.15, -6.01, -5.20, -7.38, -6.76, -9.18, -6.00, -8.08, -7.96, -8.34, -6.82, -8.46, -8.07, -7.04, -7.24, -8.16, -8.20, -8.27, -7.79, -7.37, -7.02, -7.13, -6.99, -7.65, -8.18, -6.71, -8.41, -6.71, -7.04, -9.15, -7.74, -10.11, -8.20, -7.07, -7.63, -8.99, -6.62, -6.23, -7.13, -6.41, -7.06, -7.72, -8.44, -8.85, -8.02, -6.98, -6.08, -7.20, -7.48, -7.82, -9.19, -8.31, -7.95, -7.97, -6.66, -6.59, -9.10, -7.87, -9.02, -8.77, -7.62, -9.44, -8.05, -7.60, -7.33, -6.94, -8.51, -7.39, -6.44, -8.88, -8.21, -7.66, -6.91, -8.39, -7.37, -7.26, -6.04, -7.58, -7.28, -7.02, -7.10, -7.33, -8.63, -8.21, -7.12, -8.11, -9.03, -8.11, -8.79, -9.22, -7.32, -5.97, -7.26, -6.39, -7.64, -8.38, -7.67, -7.70, -7.70, -8.95, -6.25, -8.09, -7.85, -8.10, -7.73, -6.78, -7.78, -8.20, -8.88, -8.51, -7.45, -7.14, -6.63, -7.38, -7.72, -6.25)

Общие теоретические сведения

Формулы для вычисления величин

Пусть $\vec{x}$ – выборка из генеральной совокупности X.

Минимальное и максимальное значения выборки

$$\begin{aligned} M_{\max} &= x_{(n)} \\ M_{\min} &= x_{(1)}, \end{aligned} \tag{1}$$

где $x_{(i)}$ – элемент вариационного ряда.

Размах выборки

$$R = M_{\max} - M_{\min}. \tag{2}$$

Оценки математического ожидания и дисперсии

$$\begin{aligned} \hat{\mu}(\vec{x}_n) &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ S^2(\vec{x}_n) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2, \end{aligned} \tag{3}$$

где $\overline{x}_n = \hat{\mu}(\vec{x}_n)$.

Определение эмпирической плотности и гистограммы

Пусть $\vec{x} = (x_1, ..., x_n)$ – выборка из генеральной совокупности X. Если объем n этой выборки велик, то значения x_i группируют в интервальный статистический ряд. Для этого отрезок $J = [x_{(1)}, x_{(n)}]$ делят на m равновеликих частей:

$$J_i = [x_{(1)} + (i-1) \cdot \Delta, x_{(1)} + i \cdot \Delta), i = \overline{1; m-1}$$

$$J_m = [x_{(1)} + (m-1) \cdot \Delta, x_{(n)}]$$

$$\Delta = \frac{|J|}{m} = \frac{x_{(n)} - x_{(1)}}{m}$$

Интервальным статистическим рядом называют таблицу:

J_1	...	J_i	...	J_m
n_1	...	n_i	...	n_m

где n_i – количество элементов выборки $\vec{x}$, которые $\in J_i$.

Обычно выборку разбивают на $m = [\log_2 n] + 2$ интервалов, где n – размер выборки.

Гистограмма – это график эмпирической плотности.

Эмпирической плотностью, отвечающей выборке $\vec{x}$, называют функцию:

$$\hat{f}(x) = \begin{cases} \frac{n_i}{n\Delta}, x \in J_i, i = \overline{1; m} \\ 0, \text{иначе} \end{cases} \tag{4}$$

где J_i – полуинтервал статистического ряда, n_i – количество элементов выборки, входящих в полуинтервал, n – количество элементов выборки.

Определение эмпирической функции распределения

Пусть $\vec{x} = (x_1, ..., x_n)$ – выборка из генеральной совокупности X. Обозначим $n(x, \vec{x})$ – число элементов вектора $\vec{x}$, которые имеют значения меньше x.

Эмпирической функцией распределения называют функцию $F_n : \mathbb{R} \to \mathbb{R}$, определенную как:

$$F_n(x) = \frac{n(x, \vec{x})}{n} \tag{5}$$

Ход работы

Для решения поставленной задачи была использованна программа *MATLAB 2022a*.

Листинг программы

```
function main()
  function myhist(X, bins, counts, R, m)
    n = length(X);
    delta = R / m;
    middles = zeros(1, m);
    xx = zeros(1, m);

    for i = 1:m
      xx(i) = counts(i) / (n * delta);
    end

    for i = 1:m
      middles(i) = bins(i) + (delta/2);
    end

    fprintf("    delta: %f\n", delta);
    fprintf("    hist coords:\n");

    for i = 1:m
      fprintf("    [%d] : %f %f\n", i,middles(i), xx(i));
    end

    fprintf("    hist_s = %f\n", sum(xx) * delta);
    set(gca, "xlim", [min(bins) - 1, max(bins) + 1]);
    bar(middles, xx,1, "facecolor", "b", "edgecolor", "w");

    X_range = m_min:(sigma / 100):m_max;
    X_pdf = normpdf(X_range, mx, sigma);
    plot(X_range, X_pdf, "r");
    end

  function mycdf(X, bins, counts)
    n = length(X);
    xx = zeros(1, m + 3);

```

```
    bins = [(min(bins) - 0.5) bins (max(bins) + 1)];
    counts = [0 counts 0];

    m = m + 2;
    acc = 0;

    for i = 2:m
      acc = acc + counts(i);
      xx(i) = acc / n;
    end

    xx(m + 1) = 1;

    X_n = (min(X) - 0.5):(sigma / 100):(max(X) + 1.5);
    X_cdf = normcdf(X_n, mx, sigma);
    plot(X_n, X_cdf, "r");

    for i = 2:m
      fprintf("x = %f : F(x) = %f\n", bins(i), xx(i));
    end

    set(gca, "ylim", [0, 1.1]);
    stairs(gca, bins, xx, "b");
  end

  X = [-8.47, -7.45, -7.12, -8.30, -8.15, -6.01, -5.20, -7.38, -6.76, -9.18,
       -6.00, -8.08, -7.96, -8.34, -6.82, -8.46, -8.07, -7.04, -7.24, -8.16,
      -8.20, -8.27, -7.79, -7.37, -7.02, -7.13, -6.99, -7.65, -8.18, -6.71,
      -8.41, -6.71, -7.04, -9.15, -7.74, -10.11, -8.20, -7.07, -7.63, -8.99,
      -6.62, -6.23, -7.13, -6.41, -7.06, -7.72, -8.44, -8.85, -8.02, -6.98,
      -6.08, -7.20, -7.48, -7.82, -9.19, -8.31, -7.95, -7.97, -6.66, -6.59,
      -9.10, -7.87, -9.02, -8.77, -7.62, -9.44, -8.05, -7.60, -7.33, -6.94,
      -8.51, -7.39, -6.44, -8.88, -8.21, -7.66, -6.91, -8.39, -7.37, -7.26,
      -6.04, -7.58, -7.28, -7.02, -7.10, -7.33, -8.63, -8.21, -7.12, -8.11,
      -9.03, -8.11, -8.79, -9.22, -7.32, -5.97, -7.26, -6.39, -7.64, -8.38,
      -7.67, -7.70, -7.70, -8.95, -6.25, -8.09, -7.85, -8.10, -7.73, -6.78,
      -7.78, -8.20, -8.88, -8.51, -7.45, -7.14, -6.63, -7.38, -7.72, -6.25];
  m_max = max(X);
  m_min = min(X);
  fprintf("(a) \n");
  fprintf("     max_value = %f\n", m_max);
  fprintf("     min_value = %f\n", m_min);
  fprintf("----------------------------------------------------------\n");

  r = m_max - m_min;
  fprintf("(b) \n");
  fprintf("     r = %f\n", r);
  fprintf("----------------------------------------------------------\n");

  n = length(X);
```

```
mx = sum(X) / n;
dx = sum((X - mx).^2) / (n - 1);
sigma = sqrt(dx);

fprintf("(c) \n");
fprintf("    mx = %f\n", mx);
fprintf("    dx = %f\n", dx);
fprintf("----------------------------------------------------------------\n");

m = floor(log2(n)) + 2;
bins = [];
cur = m_min;

for i = 1:(m + 1)
  bins(i) = cur;
  cur = cur + r / m;
end

eps = 1e-6;
counts = [];
j = 1;

for i = 1:(m - 1)
  cur_count = 0;
  for j = 1:n
    if (bins(i) < X(j) || abs(bins(i) - X(j)) < eps) && X(j) < bins(i + 1)
      cur_count = cur_count + 1;
    end
  end
  counts(i) = cur_count;
end

cur_count = 0;

for j = 1:n
  if (bins(m) < X(j) || abs(bins(m) - X(j)) < eps) && (X(j) < bins(m + 1) || abs(bins(m + 1) - X(j)) < eps)
    cur_count = cur_count + 1;
  end
end

counts(m) = cur_count;

fprintf("(d)\n");

for i = 1:(m)
  fprintf("    [%f : %f) - %d values.\n", bins(i), bins(i + 1), counts(i));
end
```

```
    fprintf("--------------------------------------------------------\n");

    fprintf("(e) \n");

    figure;
    hold on;
    grid on;
    myhist(X, bins, counts, r, m);
    xlabel('X')
    ylabel('P')
    hold off;

    fprintf("--------------------------------------------------------\n");

    fprintf("(f) \n");

    figure;
    hold on;
    grid on;
    mycdf(X, bins, counts);
    xlabel('X')
    ylabel('F')
    hold off;
end
```

Результат работы программы

$$M_{\min} = -10.110000$$

$$M_{\max} = -5.200000$$

$$R = 4.910000$$

$$\hat{\mu}(\vec{x}_n) = -7.660917$$

$$S^2(\vec{x}_n) = 0.777892$$

$$m = 8$$

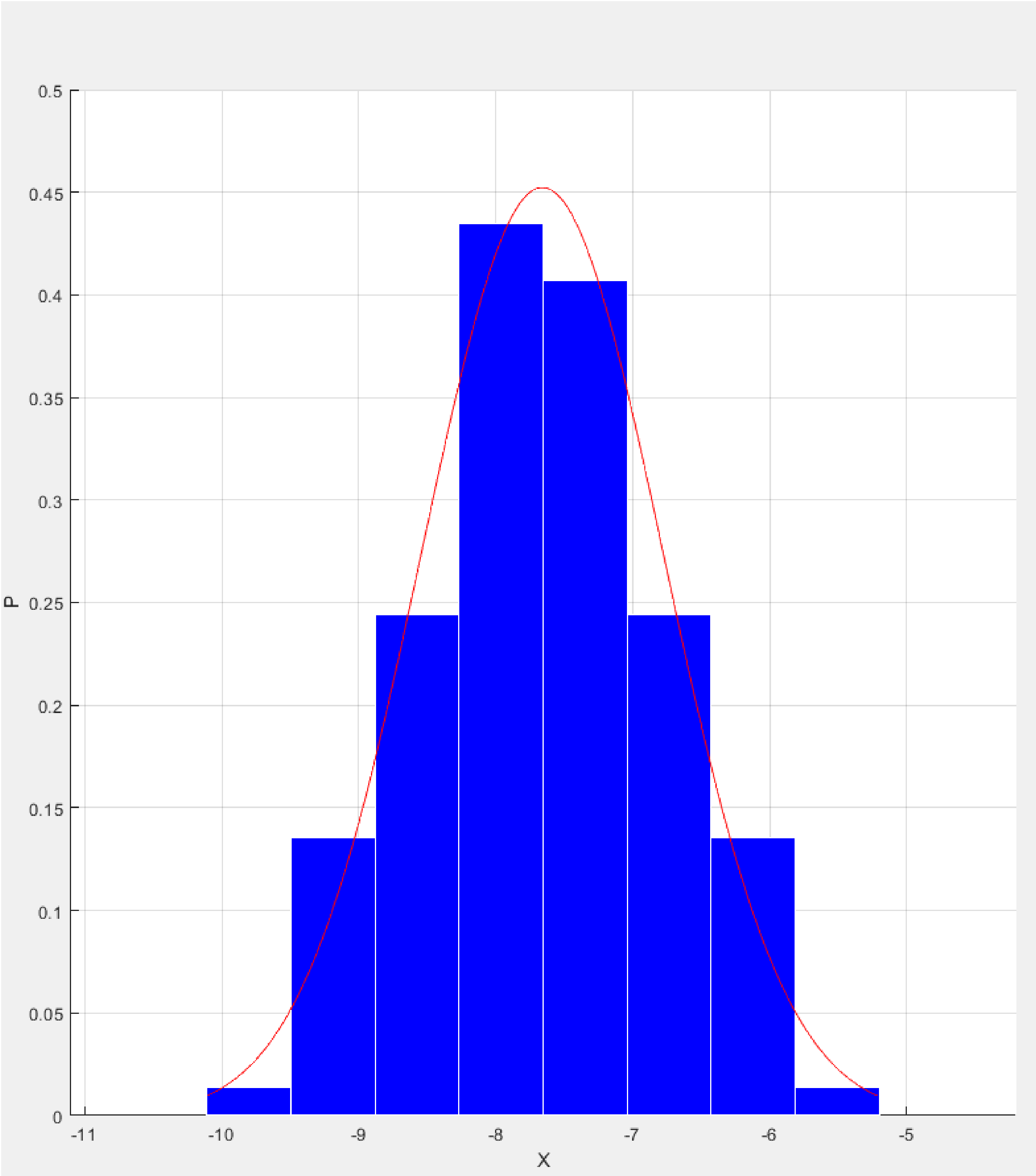

Рис. 1: Гистограмма и график функции плотности распределения вероятностей нормальной случайной величины с выборочными мат. ожиданием и дисперсией

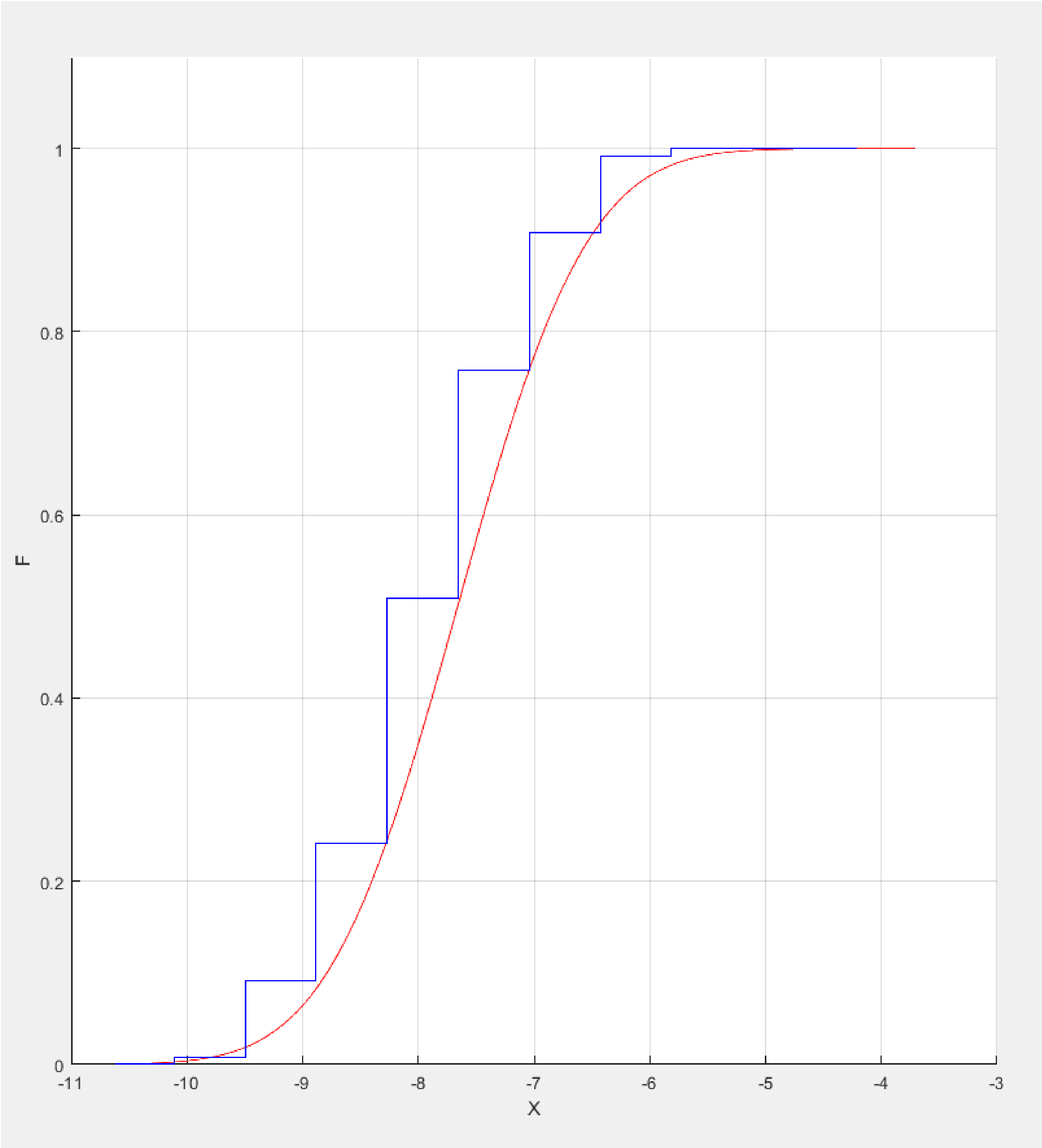

Рис. 2: График эмпирической функции распределения и функции распределения нормальной случайной величины с выборочными мат. ожиданием и дисперсией