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»

НАПРАВЛЕНИЕ ПОДГОТОВКИ «09.03.04 Программная инженерия»

ОТЧЕТ

ПО ЛАБОРАТОРНОЙ РАБОТЕ №1

Название: Гистограмма и эмпирическая функция распределения

Дисциплина: Математическая статистика

Студент ИУ7-66Б ________ Т. А. Казаева

Группа — Подпись, дата — И. О. Фамилия

Преподаватель ________ Т. В. Андреева

Подпись, дата — И. О. Фамилия

Москва, 2022 г.

1. Содержание работы

Цель работы: построение гистограммы и эмпирической функции распределения.

1. Для выборки объема 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.

2. Теоретические сведения

Пусть $\vec{x} = (x_1, \dots x_n)$ – выборка из генеральной совокупности X объема n.

1. Максимальное значение выборки

$$M_{max} = X_{(1)}, \tag{2.1}$$

2. Минимальное значение выборки

$$M_{min} = X_{(n)}, \tag{2.2}$$

3. Размах выборки:

$$R = M_{max} - M_{min}, \tag{2.3}$$

4. Оценка математического ожидания:

$$\hat{\mu}(\vec{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{2.4}$$

5. Оценка дисперсии:

$$S^2(\vec{X}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{2.5}$$

Обозначим $n(t, \vec{x})$ – число компонент вектора $\vec{x}$, которые меньше чем t. *Эмпирической функцией распределения*, построенной по выборке $\vec{x}$ называется функция $F_n : \mathbb{R} \to \mathbb{R}$, определенная правилом 2.6.

$$F_n(t) = \frac{n(t, \vec{x})}{n} \tag{2.6}$$

Однако, если объем выборки достаточно велик, то элементы выборки формируют в *интервальный статистический ряд*. Для этого отрезок $J = [x_{(i)}, x_{(n)}]$ разбивают на m промежутков, ширина которых определяется согласно 2.7:

$$\Delta = \frac{|J|}{m} = \frac{X_{(n)} - X_{(i)}}{m}, \tag{2.7}$$

далее получают

$$\begin{aligned} J_i &= \left[x_{(1)} + (i-1)\,\Delta;\; x_{(i)} + i\Delta\right),\; i = \overline{1,\, m-1}, \\ J_m &= \left[x_{(1)} + (m - 1\Delta),\; x_{(n)}\right). \end{aligned} \tag{2.8}$$

Интервальный статистический ряд, отвечающий выборке $\vec{x}$ – таблица вида:

J_1	J_2	$\dots$	J_m
n_1	n_2	$\dots$	n_m

где n_i – число элементов, попавших в промежуток J_i, $i = \overline{1,\, m}$. *Эмпирической плотностью* распределения соответствующей выборке $\vec{x}$ называется функция 2.9:

$$f_n(x) = \begin{cases} \dfrac{n_i}{n \cdot \Delta} & , x \in J_i, \\ 0 & , \text{иначе}. \end{cases} \tag{2.9}$$

График эмпирической функции плотности называется *гистограммой*.

3. Текст программы

```
function main()
  pkg load statistics

  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);
    endfor

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

    fprintf("    высоты столбцов гистограммы:\n");

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

    fprintf("[проверка] площадь гистограммы s = %f\n", sum(xx) * delta);

    set(gca, "xtick", bins);
    set(gca, "ytick", xx);
    set(gca, "xlim", [min(bins) - 1, max(bins) + 1]);
    bar(middles, xx, 1, "facecolor", "g", "edgecolor", "w");

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

  function mycdf(X, bins, counts)
    n = length(X);
    xx = zeros(1, m + 3);
    xx(1) = 0;
    xx(2) = 0;
    xx(m + 1) = 1;
    xx(m + 2) = 1;
    xx(m + 3) = 1;

    acc = 0;
```

```

    for i = 3:m
      acc = acc + counts(i - 1);
      xx(i) = acc / n;
    end

    X_n = (X(1) - 1):(sigma / 100):(X(n) + 1);
    X_cdf = normcdf(X_n, mu, s_2);
    plot(X_n, X_cdf, "r");

    bins = [min(bins) - 1 bins max(bins) + 1]; % for "longer plot"

    for i = 1:(m + 3)
      fprintf("%f : %f\n", bins(i), xx(i));
    end

    set(gca, "xtick", bins);
    set(gca, "ylim", [0, 1.1]);
    set(gca, "xlim", [X(1) - 1, X(n) + 1]);
    set(gca, "ytick", xx);
    stairs(bins, xx);
  end

  X = [-0.68, 0.71, 2.27, 0.38, 0.14, 0.06, 1.21, -0.59, 0.44, 1.98, ...
       1.00, -0.88, -0.08, 1.87, -0.74, 0.83, -1.45, 0.58, 0.48, 3.26, ...
       0.02, 0.26, 2.96, 1.78, 0.58, 0.08, -1.60, 1.26, 1.28, -0.36, ...
       0.15, -0.38, -1.04, 0.95, -2.17, -0.30, 1.09, 0.39, 1.06, 0.98, ...
       -2.55, 2.62, -1.58, 3.75, -1.43, 0.92, 2.75, -0.55, 1.48, -0.96, ...
       0.50, 2.67, -0.58, 0.41, -0.46, -0.48, 1.68, -0.08, 1.76, 0.08, ...
       -1.15, 0.66, 1.54, 0.17, -0.20, 1.34, 1.08, 1.59, -0.05, 0.15, ...
       -0.35, 0.58, -0.87, 1.73, -0.27, 0.00, -0.67, 0.13, 1.75, -0.59, ...
       1.31, 1.20, 0.53, 0.14, -0.35, 1.00, -0.01, 0.21, 1.58, -0.02, ...
       1.28, 1.34, -1.66, 0.30, 0.08, 0.66, -0.26, 1.54, 1.22, 1.24, ...
       0.11, 0.79, -0.83, 1.41, 0.17, 0.55, 1.60, 1.26, 1.06, 0.39, ...
       -0.77, 1.49, 0.92, -1.58, 1.19, 0.13, 0.26, -2.14, 0.08, -1.75];

  % (a) минимальное и максимальное значение
  m_max = max(X);
  m_min = min(X);

  fprintf("(a) Максимальное значение выборки (M_max) = %f\n", m_max);
  fprintf("    Минимальное значение выборки  (M_min) = %f\n", m_min);
  fprintf("--------------------------------------------------\n");

  % (б) размах выборки
  r = m_max - m_min;

  fprintf("(б) Размах выборки (R) = %f\n", m_max);
```

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

% (в) вычисление оценок MX DX
n = length(X);
mu = sum(X) / n;
s_2 = sum((X - mu).^2) / (n - 1);
sigma = sqrt(s_2);

fprintf("(в) Оценка математического ожидания (mu) = %f\n", mu);
fprintf("    Оценка дисперсии (s_2) = %f\n", s_2);
fprintf("------------------------------------------------\n");

% (г) группировка значений выборки в m = [log_2 n] + 2 интервала

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;
        endif
    endfor

counts(i) = cur_count;
endfor

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;
  endif

```

```
    endfor

    counts(m) = cur_count;

    fprintf("(г) группировка значений выборки в m = [log_2 n] + 2 интервала:\n")
    ;

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

    fprintf("(д) построение гистограммы и графика функции плотности\n");
    fprintf("    распределения вероятностей нормальной случайной величины\n");

    figure;
    hold on;
    grid on;
    myhist(X, bins, counts, r, m);
    xlabel('X')
    ylabel('P')
    print -djpg hist.jpg
    hold off;

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

    fprintf("(е) построение графика эмпирической функции распределения\n");
    fprintf("    и функции распределения нормальной случайной величины\n");

    figure;
    hold on;
    grid on;
    mycdf(X, bins, counts);
    xlabel('X')
    ylabel('F')
    print -djpg cdf.jpg
    hold off;
end
```

4. Результаты расчетов

```
octave:1> run main
(а) Максимальное значение выборки (M_max) = 3.750000
    Минимальное значение выборки  (M_min) = -2.550000
----------------------------------------
(б) Размах выборки (R) = 6.300000
----------------------------------------
(в) Оценка математического ожидания (mu) = 0.416417
    Оценка дисперсии (s_2) = 1.339917
----------------------------------------
(г) группировка значений выборки в m = [log_2 n] + 2 интервала:
    [-2.550000 : -1.762500) - 3 вхожд.
    [-1.762500 : -0.975000) - 9 вхожд.
    [-0.975000 : -0.187500) - 22 вхожд.
    [-0.187500 : 0.600000) - 37 вхожд.
    [0.600000 : 1.387500) - 27 вхожд.
    [1.387500 : 2.175000) - 15 вхожд.
    [2.175000 : 2.962500) - 5 вхожд.
    [2.962500 : 3.750000] - 2 вхожд.
----------------------------------------
(д) построение гистограммы и графика функции плотности
    распределения вероятностей нормальной случайной величины
    высоты столбцов гистограммы:
    [1] : 0.031746
    [2] : 0.095238
    [3] : 0.232804
    [4] : 0.391534
    [5] : 0.285714
    [6] : 0.158730
    [7] : 0.052910
    [8] : 0.021164
[проверка] площадь гистограммы s = 1.000000
----------------------------------------
(е) построение графика эмпирической функции распределения
    и функции распределения нормальной случайной величины
```

Рисунок 4.1 – Результат работы программы

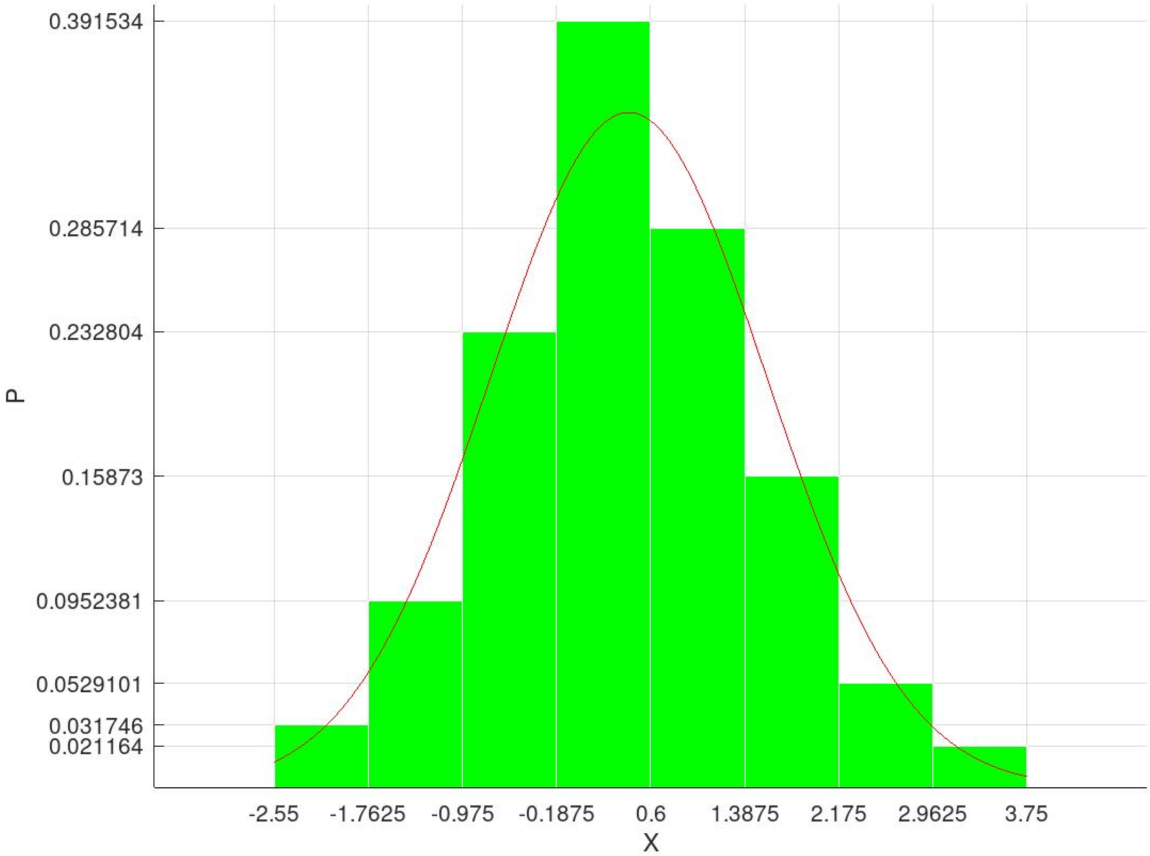

Рисунок 4.2 – Гистограмма и график функции плотности распределения вероятностей нормальной случайной величины

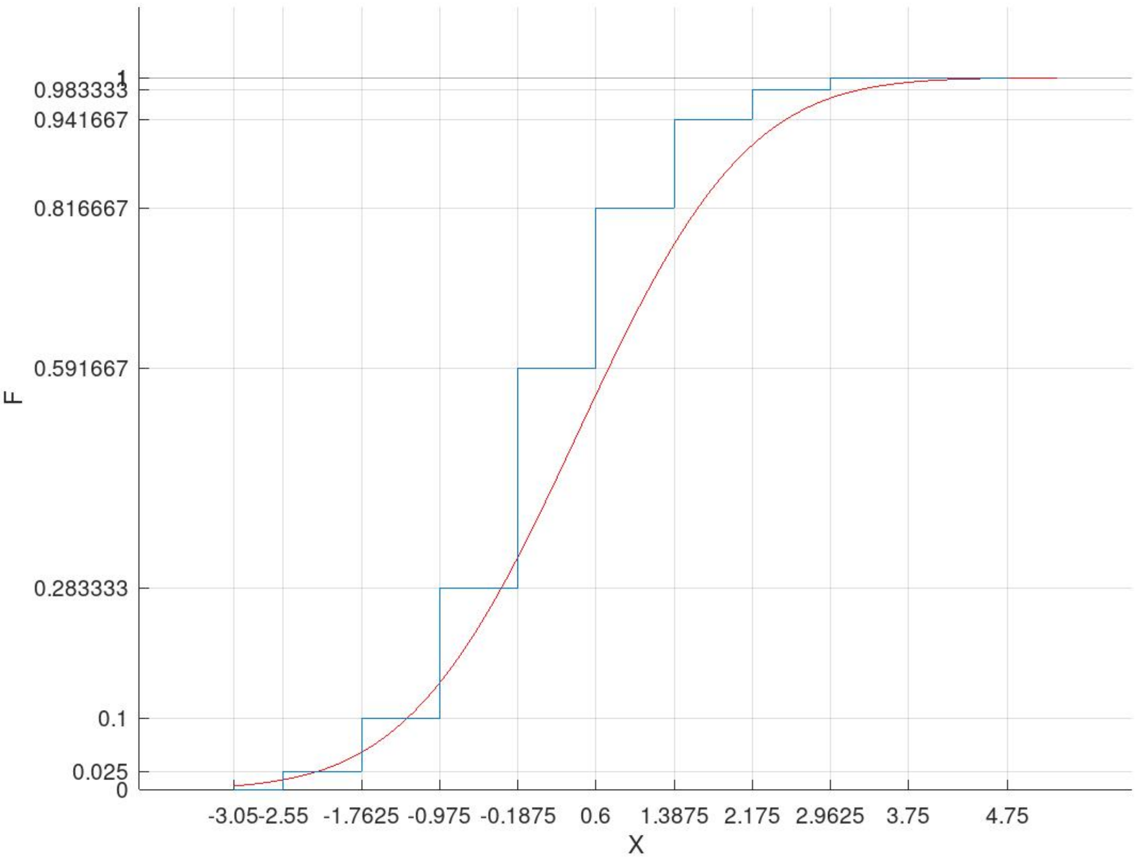

Рисунок 4.3 – График эмпирической функции распределения и функции распределения нормальной случайной величины