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 курсу "Математическая статистика"

Тема Гистограмма и эмпирическая функция распределения

Студент Зайцева А. А.

Группа ИУ7-62Б

Преподаватель Власов П. А.

Москва — 2022 г.

Оглавление

1 Задание

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.

2 Теоретические сведения

2.1 Формулы для вычисления величин

Пусть дана случайная выборка $\vec{X} = (X_1, ..., X_n)$ объема n из генеральной совокупности X; $(X_{(1)}, ..., X_{(n)})$ – вариационный ряд, отвечающий этой выборке.

Минимальное и максимальное значения выборки можно найти по формулам (2.1) и (2.2), соответственно:

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

$$M_{\max} = X_{(n)}. \tag{2.2}$$

Размах выборки находится по формуле (2.3):

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

Оценки математического ожидания и дисперсии можно найти по формулам (2.4) и (2.5), соответственно:

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

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

2.2 Определение эмпирической плотности и гистограммы

Пусть $\vec{x} = (x_1, ..., x_n)$ – выборка из генеральной совокупности X; вектор $(x_{(1)}, ..., x_{(n)})$ – вариационный ряд, построенный по этой выборке. Если объем n выборки велик, то значения x_i группируют в так называемый интервальный статистический ряд. Для этого отрезок $J = [x_{(1)}; x_{(n)}]$ разбивают на m равновеликих промежутков:

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

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

где Δ – ширина каждого из них:

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

Определение.

Интервальным статистическим рядом, отвечающим выборке $\vec{x}$ называют таблицу вида

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

где n_i – количество элементов выборки $\vec{x}$, попавших в промежуток J_i. Для выбора m используют формулу

$$m = [\log_2 n] + 2$$

или

$$m = [\log_2 n] + 1.$$

Пусть для данной выборки $\vec{x}$ построен интервальный статистический ряд $(J_i, n_i), i = \overline{1; m}$.

Определение.

Эмпирической функцией плотности распределения, соответсвующей выборке $\vec{x}$, называют функцию

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

Определение.

График эмпирической функции плотности называют гистограммой.

2.3 Определение эмпирической функции распределения

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

Определение.

Эмпирической функцией распределения, построенной по выборке $\vec{x}$, называют функцию $F_n : \mathbb{R} \to \mathbb{R}$, определенную правилом:

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

3 Результаты работы

3.1 Текст программы

```
X = [7.76,6.34,5.11,7.62,8.84,4.68,8.65,6.90,8.79,6.61,6.62,7.13,6.75,7.28,...
7.74,7.08,5.57,8.20,7.78,7.92,6.00,4.88,6.75,6.56,7.48,8.51,9.06,6.94,6.93,...
7.79,5.71,5.93,6.81,5.76,5.88,7.05,7.22,6.67,5.59,6.57,7.28,6.22,6.31,5.51,...
6.69,7.12,7.40,6.86,7.28,6.82,7.08,7.52,6.81,7.55,4.89,5.48,7.74,5.10,8.17,...
7.67,7.07,5.80,6.10,7.15,7.88,9.06,6.85,4.88,6.74,8.76,8.53,6.72,7.21,7.42,...
8.29,8.56,9.25,6.63,7.49,6.67,6.79,5.19,8.20,7.97,8.64,7.36,6.72,5.90,5.53,...
6.44,7.35,5.18,8.25,5.68,6.29,6.69,6.08,7.42,7.10,7.14,7.10,6.60,6.35,5.99,...
6.17,9.05,6.01,7.77,6.27,5.81,7.80,9.89,4.39,6.83,6.53,8.15,6.68,6.87,6.31,...
6.83];

n = length(X);
fprintf('n = %i\n', n);


fprintf('\na)\n');
M_max = max(X);
M_min = min(X);
fprintf('M_max = %f\n', M_max);
fprintf('M_min = %f\n', M_min);


fprintf('\бn)\n');
R = M_max - M_min;
fprintf('R = %f\n', R);


fprintf('\вn)\n');
mu = mean(X);
s_sqr = var(X);
fprintf('mu = %f\n', mu);
fprintf('s_sqr = %f\n', s_sqr);


fprintf('\гn)\n');
m = floor(log2(n)) + 2;
delta = R / m;
fprintf('m = %i, delta = %f\n\n', m, delta);


```

```
ni_array = zeros([1 m]);
ai_array = zeros([1 m]);
bi_array = zeros([1 m]);
for i = 1:m-1
    ai_array(i) = M_min + (i - 1) * delta;
    bi_array(i) = M_min + i * delta;
    ni_array(i) = count_ni(i, ai_array(i), bi_array(i), X);
end
ai_array(m) = M_min + (m - 1) * delta;
bi_array(m) = M_max;
ni_array(m) = count_ni(m, ai_array(m), bi_array(i) + 1, X);

fprintf('\n');
for i = 1:m-1
    fprintf('J%i␣=␣[%f;␣%f);␣n%i␣=␣%i\n', i, ai_array(i), bi_array(i), i, ni_array(i));
end
fprintf('J%i␣=␣[%f;␣%f];␣n%i␣=␣%i\n', m, ai_array(m), bi_array(m), m, ni_array(m));


fprintf('\дn)\n');
figure;
hold on;
h_array = ni_array * (1 / (n * delta));
ai_array(end+1) = M_max;
hist = histogram('BinEdges', ai_array,'BinCounts', h_array);
sigma = sqrt(s_sqr);
x_array = (M_min - 1):(sigma / 100):(M_max + 1);
f = normpdf(x_array, mu, sigma);
plot(x_array, f, 'r', 'LineWidth', 2);
grid on;
xlabel('x');
ylabel('fn(x)');


fprintf('\ne)\n');
figure;
hold on;
z_array = unique(X);
z_array(end + 1) = z_array(end) + 1;
nt_array = zeros([1 length(z_array)]);
for z_index = 1:length(z_array)
    nt_array(z_index) = count_nt(z_array(z_index), X) / n;
end
stairs(z_array, nt_array, 'b');
F = normcdf(x_array, mu, sigma);
```

```
plot(x_array, F, 'r');
grid on;
xlabel('x');
ylabel('Fn(x)');

function [ni] = count_ni(i, a, b, X)
    tolerance = 1e-4;
    ni = 0;
    fprintf('J%iому-␣интервалу␣принадлежат␣значения␣{', i);
    for x_index = 1:length(X)
        if (X(x_index) - a >= tolerance) && (X(x_index) - b < tolerance)
            fprintf('%f,␣', X(x_index));
            ni = ni + 1;
        end
    end
    fprintf('}\n');
end

function [nt] = count_nt(t, X)
    tolerance = 1e-4;
    nt = 0;
    for x_index = 1:length(X)
        if (X(x_index) - t < tolerance)
            nt = nt + 1;
        end
    end
end
```

3.2 Результаты расчетов

$$(a)M_{\min} = 4.39; M_{\max} = 9.89$$

$$(b)R = 5.5$$

$$(c)\hat{\mu}(\vec{x}_n) = 6.9445; S^2(\vec{x}_n) = 1.171956$$

$$(d)m = 8, \Delta = 0.6875$$

Таблица 3.1: Интервальный статистический ряд

J1 = [4.39; 5.0775)	J2 = [5.0775; 5.765)	J3 = [5.765; 6.4525)	J4 = [6.4525; 7.14)	J5 = [7.14; 7.8275)	J6 = [7.8275; 8.515)	J7 = [8.515; 9.2025)	J8 = [9.2025; 9.89]
5	12	19	37	25	10	10	2

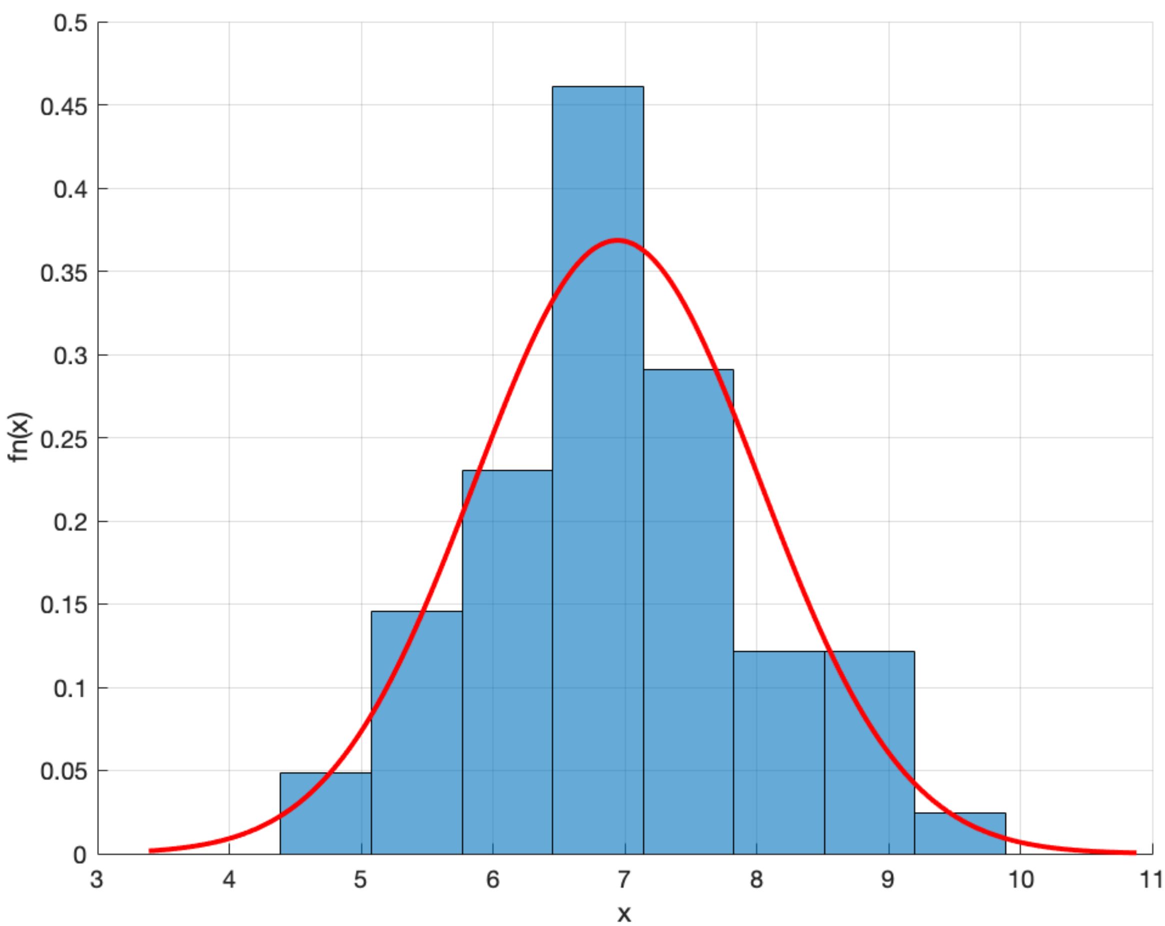

Рис. 3.1: Гистограмма (синим) и график функции плотности распределения вероятностей нормальной случайной величины с математическим ожиданием $\hat{\mu}$ и дисперсией S^2 (красным)

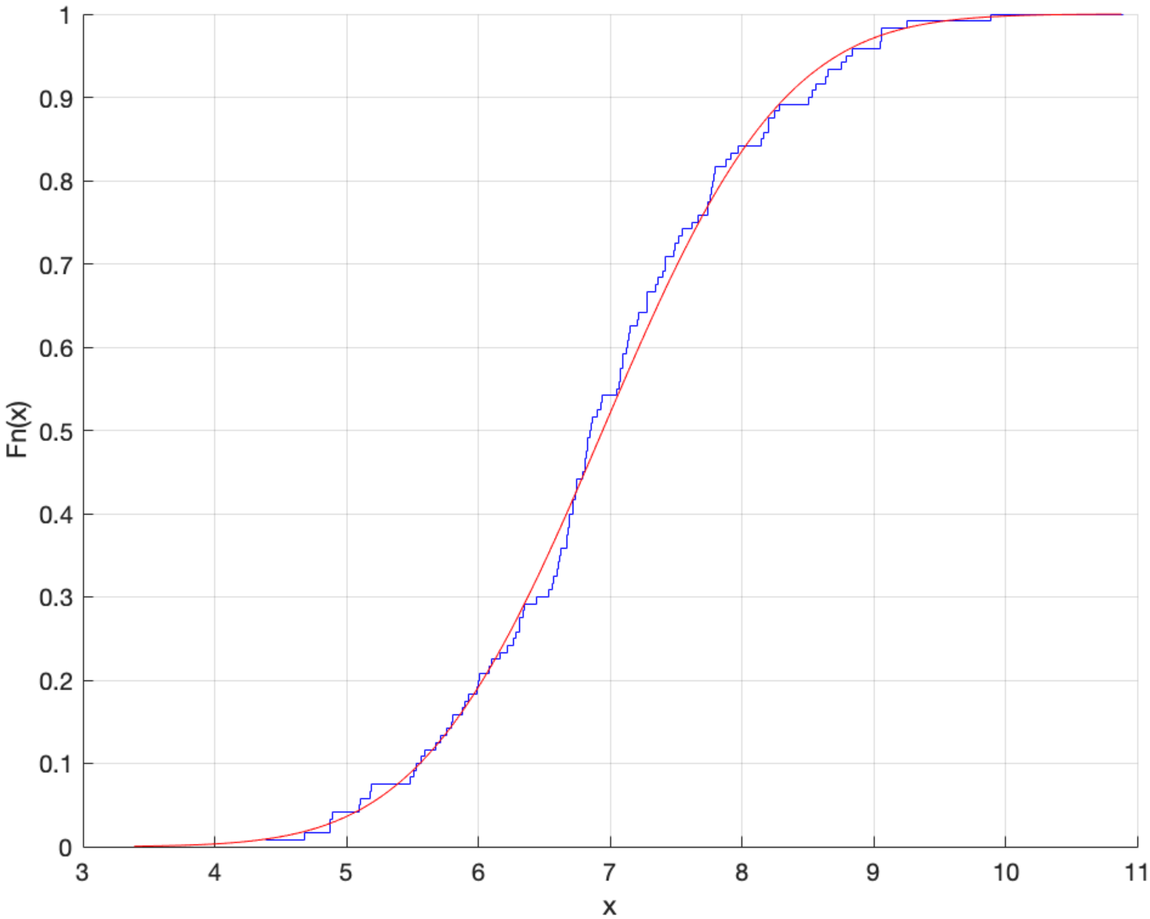

Рис. 3.2: График эмпирической функции распределения (синим) и функции распределения нормальной случайной величины с математическим ожиданием $\hat{\mu}$ и дисперсией S^2 (красным)