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»

Отчёт по лабораторной работе №1 по курсу «Математическая статистика»

Тема Гистограмма и эмпирическая функция распределения

Студент Прянишников А.Н.

Группа ИУ7-65Б

Оценка (баллы) ____________

Преподаватели Андреева Т. В.

Москва — 2022 г.

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.

Теоретические сведения

Множество возможных значений случайной величины X называют **генеральной совокупностью** случайной величины X.

Любое возможное значение $\vec{x} = (x_1, \dots x_n)$ случайной выборки $\vec{X}_n$ будем называть **выборкой** из генеральной совокупности X. Число n характеризует объем выборки, а числа x_i представляют собой элементы выборки $\vec{X}_n$.

Пусть $\vec{x} = (x_1, \ldots x_n)$ — выборка объема n из генеральной совокупности X. Ее можно упорядочить, расположив значения в неубывающем порядке:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)} \tag{1}$$

Такую последовательность чисел из формулы 1 называют **вариационным рядом**.

Тогда $x_{(1)}$ является **минимальным** значением выборки, а $x_{(n)}$ - **максимальным** значением выборки.

Размахом выборки называют разность между максимальным и минимальным значениями.

Рассмотрим функцию $n(x, \vec{X_n})$, которая для каждого значения $x \in R$ и каждой реализации $\vec{x_n}$ случайной выборки $\vec{X_n}$ принимает значение $n(x, \vec{x_n})$, равное числу элементов в выборке $\vec{x_n}$, меньших x.

Тогда **эмпирической функцией распределения** называется функция:

$$F(x, \vec{X_n}) = \frac{n(x, \vec{x_n})}{n} \tag{2}$$

Оценку математического ожидания можно подсчитать по формуле:

$$\hat{\mu}(\vec{X}) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{3}$$

Смещённую выборочную дисперсию можно подсчитать по формуле:

$$S_n^2(\vec{X}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{4}$$

Исправленную выборочную дисперсию можно подсчитать по формуле:

$$S^2(\vec{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{5}$$

При больших объемах выборки n обычно производят группирование исходных данных следующим образом. Промежуток $J = [x_{(1)}, x_{(n)}]$, содержащий все выборочные значения, разбивают на m полуинтервалов $J_1, \ldots, J_m$, как правило, одинаковой длины δ и таких, что каждый из них, кроме последнего, содержит левую границу, а последний содержит обе границы, и подсчитывают число n_i элементов выборки, попавших в i-ый промежуток J_i, а результаты представляют в виде таблицы, которую называют **интервальным статистическим рядом**.

Эмпирической плотностью распределения соответствующей выборке $\vec{x}$ называется функция:

$$f_n(x) = \begin{cases} \dfrac{n_i}{n \cdot \Delta} & , x \in J_i, \\ 0 & , \text{иначе.} \end{cases} \tag{6}$$

График эмпирической функции плотности называется **гистограммой**.

Пусть на выборке $\vec{x}$ определён интервальный статистический ряд. Эмпирической функцией распределения интервального ряда называется функция:

$$F(J_i) = \frac{1}{n} \sum_{j=1}^{i} J_i \tag{7}$$

Код программы

На листинге 1 представлен код программы:

```
pkg load statistics

EPS = 1e-6;

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];

Xmin = min(X);
Xmax = max(X);

fprintf("------------------------------------------------\n")
fprintf("Минимальное значение выборки: %d \n", Xmin)
fprintf("Максимальное значение выборки: %d \n", Xmax)
fprintf("------------------------------------------------\n")

R = Xmax - Xmin;
printf("Размах выборки: %d \n", R);

N = length(X);
mu = mean(X);
sigma2 = var(X);
sigma = sqrt(sigma2);
correctedSigma2 = N / (N - 1) * sigma2;

fprintf("------------------------------------------------\n")
fprintf("Оценка математического ожидания: %f \n", mu)
```

```
fprintf("Смещенная оценка дисперсии: %f \n", sigma2)
fprintf("Исправленная оценка дисперсии: %f \n", correctedSigma2)

m = floor(log2(N)) + 2;

intervalBounds = [];
tmp = Xmin;
intervalDelta = R / m;
for i = 1:(m + 1)
  intervalBounds(i) = tmp;
  tmp += intervalDelta;
end

intervalValuesNum = [];

for i = 1:(m - 1)
  tmpCount = 0;

  for j = 1:N
    if ((intervalBounds(i) < X(j)) || (abs(intervalBounds(i) - X(j)) < EPS)) ...
      && (X(j) < intervalBounds(i + 1))
      tmpCount += 1;
    endif
  endfor

  intervalValuesNum(i) = tmpCount;
endfor

tmpCount = 0;

for j = 1:N
  if (intervalBounds(m) < X(j) || abs(intervalBounds(m) - X(j)) < EPS) && ...
        (X(j) < intervalBounds(m + 1) || abs(intervalBounds(m + 1) - X(j)) < EPS)
    tmpCount += 1;
  endif
endfor

intervalValuesNum(m) = tmpCount;

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

for i = 1:(m - 1)
  fprintf("     [%f : %f) - %d значений\n",intervalBounds(i), ...
                                  intervalBounds(i + 1), intervalValuesNum(i));
end
```

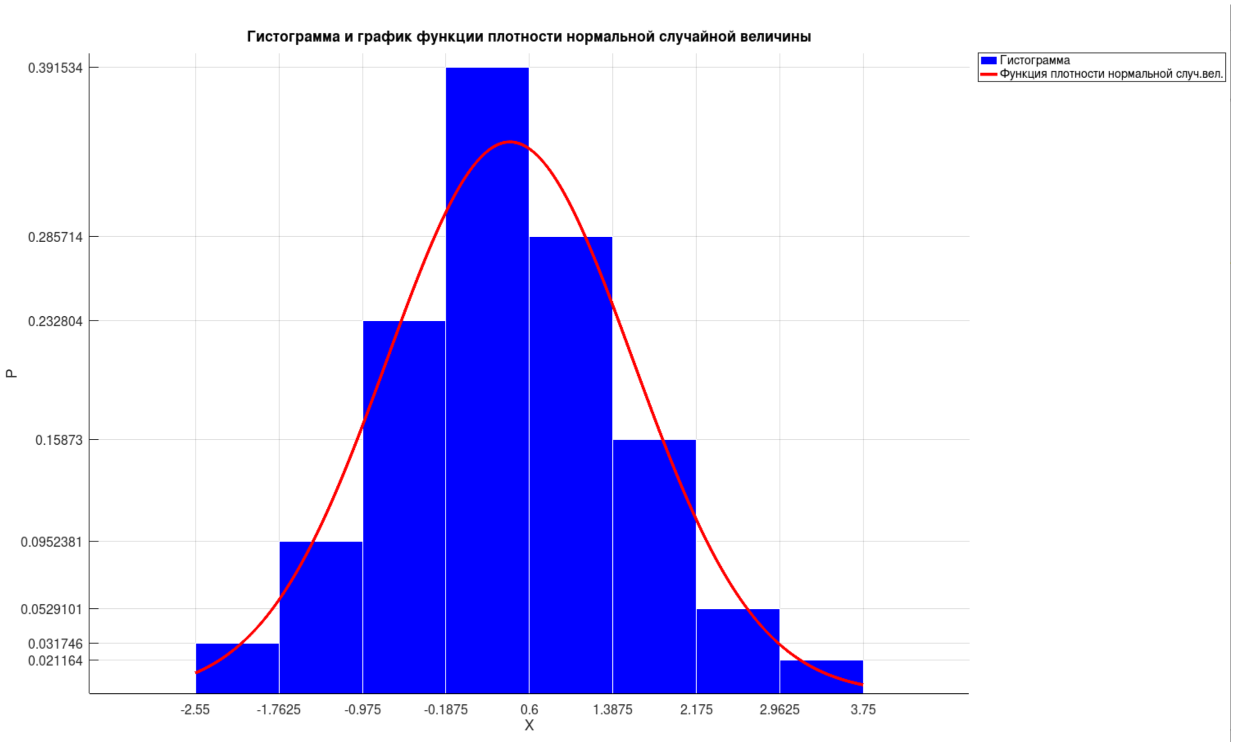
```
fprintf("      [%f : %f] - %d значений\n", intervalBounds(m), ...
                                intervalBounds(m + 1), intervalValuesNum(m));

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


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

figure('position',[100,100,1600,1200]);
title ("Гистограмма и график функции плотности нормальной случайной величины");
hold on;
grid on;

middleIntervalValues = zeros(1, m);
intervalHeight = zeros(1, m);

for i = 1:m
  intervalHeight(i) = intervalValuesNum(i) / (N * intervalDelta);
endfor

for i = 1:m
  middleIntervalValues(i) = intervalBounds(i + 1) - (intervalDelta / 2);
endfor

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

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

set(gca, "xtick", intervalBounds);
set(gca, "ytick", intervalHeight);
set(gca, "xlim", [min(intervalBounds) - 1, max(intervalBounds) + 1]);
set(gca, "fontsize", 16)
bar(middleIntervalValues, intervalHeight, 1, "facecolor", "blue", ...
    "edgecolor", "w", "displayname", "Гистограмма");


rangeX = Xmin:(sigma / 100):Xmax;
normalPdf = normpdf(rangeX, mu, sigma);
plot(rangeX, normalPdf, "color", "r", "linewidth", 4, "displayname", ...
     "Функция плотности нормальной случ.вел.");

myLegend = legend ("location", "northeastoutside");
legend(myLegend, "location", "northeastoutside");
```

```
xlabel('X')
ylabel('P')
print -djpg hist.jpg
hold off;

fprintf("----------------------------------------------------------\n");
fprintf("Значения эмпирической функции распределения в точках:\n");

figure('position',[100,100,1600,1200]);
title ("График эмпирической функции распределения и функции распределения нормальной случайной величины");
hold on;
grid on;

m += 2;
intervalCumHeigth = zeros(1, m + 1);

intervalBounds = [(Xmin - intervalDelta) intervalBounds (Xmax + intervalDelta)];
intervalValuesNum = [0 0 intervalValuesNum 0];

curHeigth = 0;

for i = 2:m
  curHeigth += intervalValuesNum(i);
  intervalCumHeigth(i) = curHeigth / N;
end

intervalCumHeigth(m + 1) = 1;

rangeNormX = (Xmin - intervalDelta):(sigma / 100):(Xmax + intervalDelta);
normCdf = normcdf(rangeNormX, mu, sigma2);
plot(rangeNormX, normCdf, "color", "r", "linewidth", 2, "displayname", ...
     "Функция распределения нормальной случ.вел.");

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

set(gca, "xtick", intervalBounds);
set(gca, "ylim", [0, 1.1]);
set(gca, "ytick", intervalCumHeigth);
set(gca, "fontsize", 16)
stairs(intervalBounds, intervalCumHeigth, "color", "blue", "linewidth", 4, ...
       "displayname", "График эмпирической функции распределения");

myLegend = legend ("location", "northeast");
legend(myLegend, "location", "northeast");
```

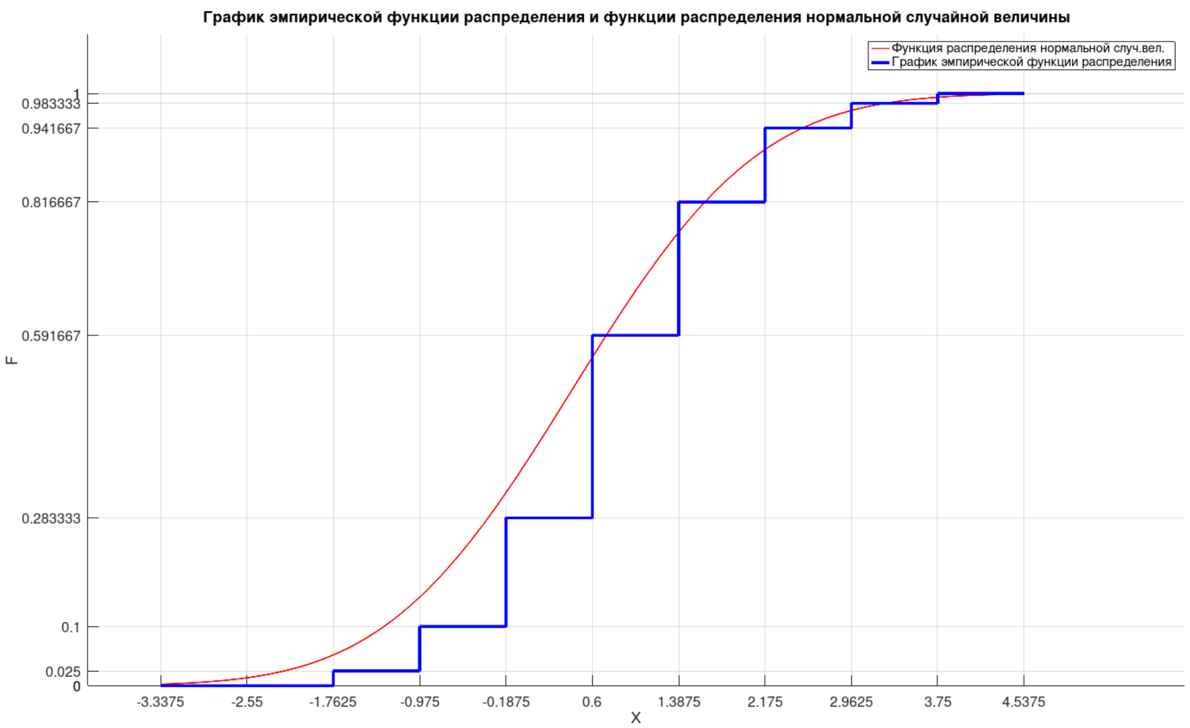
```
xlabel('X')
ylabel('F')
print -djpg cdf.jpg
hold off;
```

Результаты работы программы

На рисунке 1 представлен график гистограммы и графика функции плотности распределения вероятностей нормальной случайной величины:



На рисунке 2 представлен графика эмпирической функции распределения и функции распределения нормальной случайной величины:



Результат работы программы представлен на рисунке 3:

```
---------------------------------------
Минимальное значение выборки: -2.55
Максимальное значение выборки: 3.75
---------------------------------------
Размах выборки: 6.3
---------------------------------------
Оценка математического ожидания: 0.416417
Смещенная оценка дисперсии: 1.339917
Исправленная оценка дисперсии: 1.351177
---------------------------------------
(г) группировка значений выборки в m = [log_2 n] + 2 интервала:
    [-2.550000 : -1.762500) - 3 значений
    [-1.762500 : -0.975000) - 9 значений
    [-0.975000 : -0.187500) - 22 значений
    [-0.187500 : 0.600000) - 37 значений
    [0.600000 : 1.387500) - 27 значений
    [1.387500 : 2.175000) - 15 значений
    [2.175000 : 2.962500) - 5 значений
    [2.962500 : 3.750000] - 2 значений
---------------------------------------
(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
---------------------------------------
Значения эмпирической функции распределения в точках:
x = -2.550000 : F(x) = 0.000000
x = -1.762500 : F(x) = 0.025000
x = -0.975000 : F(x) = 0.100000
x = -0.187500 : F(x) = 0.283333
x = 0.600000 : F(x) = 0.591667
x = 1.387500 : F(x) = 0.816667
x = 2.175000 : F(x) = 0.941667
x = 2.962500 : F(x) = 0.983333
x = 3.750000 : F(x) = 1.000000
```