Московский государственный технический университет
имени Н. Э. Баумана

Факультет: Информатика и системы управления

Кафедра: Программное обеспечение ЭВМ и информационные технологии

Дисциплина: Математическая статистика

Лабораторная работа №1
Гистограмма и эмпирическая функция распределения

Выполнила: Щербатюк Дарья Сергеевна
Группа:ИУ7-64
Вариант: 12

Москва, 2018г.

Содержание

1 Постановка задачи

Цель работы: построение гистограммы и эмпирической функции распределения.

Содержание работы

1. Для выборки объёма n из генеральной совокупности X реализовать в виде программы на ЭВМ
   - вычисление максимального значения $M_{\max}$ и минимального значения $M_{\min}$;
   - размаха R выборки;
   - вычисление оценок $\hat{\mu}$ и S^2 математического ожидания $\mathsf{M}X$ и дисперсии $\mathsf{D}X$;
   - группировку значений выборки в $m = [\log_2 n] + 2$ интервала;
   - построение на одной координатной плоскости гистограммы и графика функции плотности распределения вероятностей нормальной случайной величины с математическим ожиданием $\hat{\mu}$ и дисперсией S^2;
   - 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 Отчёт

2.1 Формулы для вычисления величин

Количество интервалов

$$m = [\log_2 n] + 2 \tag{1}$$

Минимальное значение выборки

$$M_{\min} = \min\{x_1, \dots, x_n\}, \quad \text{где} \tag{2}$$

- $(x_1, \dots, x_n)$ — реализация случайной выборки.

Максимальное значение выборки

$$M_{\max} = \max\{x_1, \dots, x_n\}, \quad \text{где} \tag{3}$$

- $(x_1, \dots, x_n)$ — реализация случайной выборки.

Размах выборки

$$R = M_{\max} - M_{\min}, \quad \text{где} \tag{4}$$

- $M_{\max}$ — максимальное значение выборки;
- $M_{\min}$ — минимальное значение выборки.

Оценка математического ожидания

$$\hat{\mu}(\vec{X}_n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,. \tag{5}$$

Выборочная дисперсия

$$\hat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\,. \tag{6}$$

Несмещённая оценка дисперсии

$$S^2(\vec{X}_n) = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\,. \tag{7}$$

2.2 Эмпирическая плотность и гистограмма

Определение 2.1. *Эмпирической плотностью распределения случайной выборки* $\vec{X}_n$ называют функцию

$$f_n(x) = \begin{cases} \frac{n_i}{n\,\Delta}, & x \in J_i,\ i = \overline{1, m}; \\ 0, & \text{иначе}. \end{cases}, \quad \text{где} \tag{8}$$

- J_i, $i = \overline{1;m}$, — полуинтервал из $J = [x_{(1)}, x_{(n)}]$, где
$$x_{(1)} = \min\{x_1, \ldots, x_n\}, \qquad x_{(n)} = \max\{x_1, \ldots, x_n\}; \tag{9}$$
при этом все полуинтервалы, кроме последнего, не содержат правую границу т. е.
$$J_i = [x_{(1)} + (i-1)\Delta, x_{(i)} + i\Delta), \quad i = \overline{1, m-1}; \tag{10}$$
$$J_m = [x_{(1)} + (m-1)\Delta, x_{(1)} + m\Delta); \tag{11}$$

- m — количество полуинтервалов интервала $J = [x_{(1)}, x_{(n)}]$;

- Δ — длина полуинтервала J_i, $i = \overline{1, m}$ равная
$$\Delta = \frac{x_{(n)} - x_{(1)}}{m} = \frac{|J|}{m}; \tag{12}$$

- n_i — количество элементов выборки в полуинтервале J_i, $i = \overline{1, m}$;

- n — количество элементов в выборке.

Определение 2.2. График функции $f_n(x)$ называют гистограммой.

2.3 Эмпирическая функция распределения

- $\vec{X}_n = (X_1, \ldots, X_n)$ — случайная выборка;

- $\vec{x}_n = (x_1, \ldots, x_n)$ — реализация случайной выборки $\vec{X}_n$;

- $n(x, \vec{x}_n)$ — количество элементов выборки $\vec{x}_n$, которые меньше x.

Определение 2.3. *Эмпирической функцией распределения* называют функцию
$$F_n \colon \Re \to \Re, \quad F_n(x) = \frac{n(x, \vec{x}_n)}{n} \tag{13}$$

Замечание. $F_n(x)$ обладает всеми свойствами функции распределения. При этом она кусочно-постоянна и принимает значения
$$0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{(n-1)}{n}, 1$$

Замечание. Если все элементы вектора $\vec{x}_n$ различны, то
$$F_n(x) = \begin{cases} 0, & x \le x_{(1)}; \\ \frac{i}{n}, & x_{(i)} < x \le x_{(i+1)},\ i = \overline{1, n-1}; \\ 1, & x > x_{(n)}. \end{cases} \tag{14}$$

Замечание. Эмпирическая функция распределения позволяет интерпретировать выборку $\vec{x}_n$ как реализацию дискретной случайной величины $\widetilde{X}$ ряд распределения которой

$\widetilde{X}$	$x_{(1)}$	$\ldots$	$x_{(n)}$
P	$1/n$	$\ldots$	$1/n$

Это позволяет рассматривать числовые характеристики случайной величины $\widetilde{X}$ как приближённые значения числовых характеристик случайной величины X.

3 Текст программы

```
function lab1()
    clear all;

    X = readFromFile();
    X = [X zeros(1,150) + 13];

    X = sort(X);

    minX = X(1);
    fprintf('Mmin = %s\n', num2str(minX));

    maxX = X(end);
    fprintf('Mmax = %s\n', num2str(maxX));

    R = maxX - minX;
    fprintf('R = %s\n', num2str(R));

    mu = expectation(X);
    fprintf('mu = %s\n', num2str(mu));

    sigmaSqr = populationVariance(X);
    fprintf('sigma^2 = %s\n', num2str(sigmaSqr));

    sSqr = unbiasedSampleVariance(X);
    fprintf('S^2: %s\n', num2str(sSqr));

    m = numberOfSubintervals(length(X));
    fprintf('m = %s\n  ', num2str(m));

    intervals(X, m);
    hold on;
    f(X, mu, sSqr, m);

    figure;
    empiricF(X);
    hold on;
    F(X, mu, sSqr, m);

end

function X = readFromFile()
    X = csvread('data.csv');
end

function m = numberOfSubintervals(size)
    m = floor(log2(size) + 2);
end

```

```
function mu = expectation(X)
    n   = length(X);
    sum = 0;

    for i = 1:n
        sum = sum + X(i);
    end

    mu = sum / n;
end

function sigmaSqr = populationVariance(X)
    n   = length(X);
    sum = 0;

    for i = 1:n
        sum = sum + (X(i))^2;
    end

    mu  = expectation(X);

    sigmaSqr = sum / n - mu^2;
end

function sSqr = unbiasedSampleVariance(X)
    sigmaSqr = populationVariance(X);
    n        = length(X);

    sSqr = n / (n - 1) * sigmaSqr;
end

function intervals(X, m)
    count = zeros(1, m+2);
    delta = (X(end) - X(1)) / m;

    J = X(1):delta:X(end);
    fprintf('%d\n', X(end));
    J(length(J)+1) = 20;

    j = 1;
    n = length(X);

    for i = 1:n
        if (j ~= m)
            if ((not (X(i) >= J(j) && X(i) < J(j+1))))
                j = j + 1;
                fprintf('[%.2f;%.2f)\t', J(j-1), J(j));
            end
        end
        count(j) = count(j) + 1;
```

```
    end
    fprintf('[%2.2f;%2.2f]\n', J(m), J(m + 1));

    Xbuf = count(1:m+2);
    for i = 1:m+2
        Xbuf(i) = count(i) / (n*delta);
    end

    stairs(J, Xbuf), grid;
end

function f(X, MX, DX, m)
    R = X(end) - X(1);
    delta = R/m;
    Sigma = sqrt(DX);

    Xn = (MX - R): delta/50 :(MX + R);
    Xn(length(Xn)+1) = 20;
    Y = normpdf(Xn, MX, Sigma);

    plot(Xn, Y);
end

function F(X, MX, DX, m)
    R = X(end) - X(1);
    delta = R/m;

    Xn = (MX - R): delta :(MX + R);

    Y = 1/2 * (1 + erf((Xn - MX) / sqrt(2*DX)));


    plot(Xn, Y, 'r');
end

function empiricF(X)
    [yy, xx] = ecdf(X);
    yy(length(yy)+1) = 1;
    xx(length(xx)+1) = 20;

    stairs(xx, yy);
end
```

4 Результаты расчётов

$$M_{\min} = 6.81;$$
$$M_{\max} = 12.41;$$
$$R = 5.6;$$
$$\hat{\mu}(\vec{X}_n) = 9.4872;$$
$$S^2(\vec{X}_n) = 1.2173.$$

Интервальная групировка значений выборки при $m = 8$:

$$[6.81; 7.51), [7.51; 8.21), [8.21; 8.91), [8.91; 9.61), [9.61; 10.31),$$

$$[10.31; 11.01), [11.01; 11.71), [11.71; 12.41]$$

5 Графики

5.1 Гистограмма и график функции плотности распределения вероятностей нормальной случайной величины с математическим ожиданием $\hat{\mu}$ и дисперсией S^2

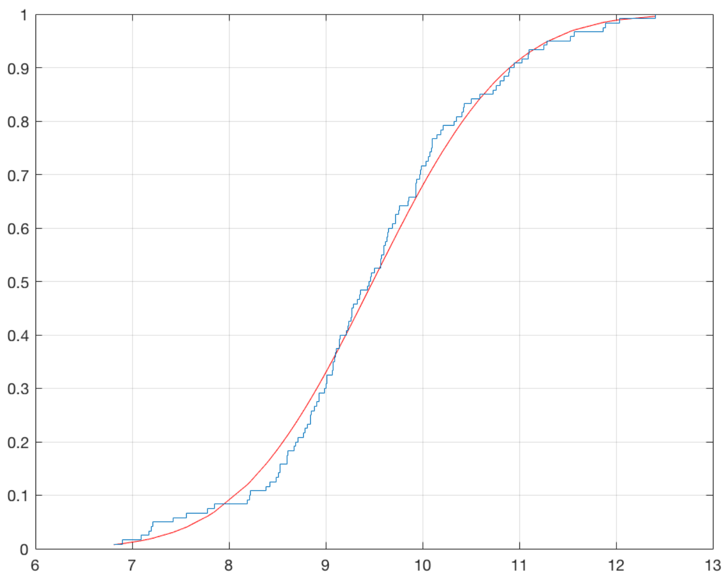

Рис. 1: Гистограмма и график функции плотности распределения нормальной случайной величины.

5.2 График эмпирической функции распределения и функции распределения нормальной случайной величины с математическим ожиданием $\hat{\mu}$ и дисперсией S^2

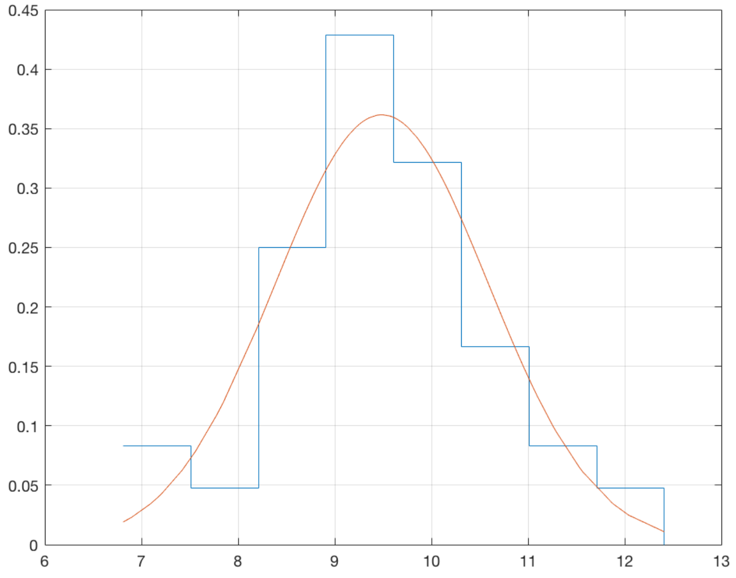

Рис. 2: График эмпирической функции распределения и функции распределения нормальной случайной величины.