Лабораторная работа № 1 Статистическое моделирование, оценка основных характеристик распределения

Реальные статистические выборки получают в результате измерения значений, которые принимают интересующие нас (наблюдаемые) случайные величины. Однако в данной лабораторной работе предлагается произвести компьютерное моделирование статистической выборки из заданного распределения (варианты заданий к работе приведены ниже), а затем выполнить для получившейся выборки первоначальную статистическую обработку.

Задание

Смоделировать в системе $\mathbb{R}$ выборку $X_1, X_2, \ldots, X_N$ объемом N из распределения, указанного в списке вариантов заданий (ниже), согласно номеру своего варианта. Найти статистические оценки значений основных параметров распределения (табл. 1.1), построить асимптотические доверительные интервалы для математического ожидания и дисперсии распределения, построить графики эмпирической функции распределения, гистограммы относительных частот, сравнить с графиками теоретических функции распределения и плотности.

Таблица 1.1

Числовая характеристика случайной величины ξ	**Статистическая оценка данной характеристики**
$a = \mathbf{E}\xi$ – математическое ожидание	$\overline{X} = N^{-1}\sum_{i=1}^{N} X_i$ – выборочное среднее
$\sigma^2 = \mathbf{E}(\xi - \mathbf{E}\xi)^2$ – дисперсия	$S^2 = N^{-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$ – выборочная дисперсия $\widehat{S}_N^2 = (N-1)^{-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$ – несмещенная оценка дисперсии
$\alpha_k = \mathbf{E}\xi^k$ – момент k-го порядка, k – натуральное	$A_{N,k} = N^{-1}\sum_{i=1}^{N} X_i^k$ – выборочный момент k-го порядка
$\mu_k = \mathbf{E}(\xi - \mathbf{E}\xi)^k$ – центральный момент k-го порядка	$M_{N,k} = N^{-1}\sum_{i=1}^{N}(X_i - \overline{X})^k$ – выборочный централ-ьный k-й момент
$F_\xi(x) = \mathbf{P}\{\xi < x\}$ – функция распределения	$\widehat{F}_N(x) = \sharp\{i : X_i < x\}/N$ –эмпирическая функция распределения, где $\sharp$ – число элементов множества
$\beta = \mu_3/\sigma^3$ – параметр асимметрии	$B_N = M_{N,3}/(S^2)^{3/2}$ – выборочный параметр асимметрии
$\varepsilon = \mu_4/\sigma^4 - 3$ – параметр эксцесса	$E_N = M_{N,4}/S^4 - 3$ – выборочный параметр эксцесса

Асимптотические доверительные интервалы

Используя свойство асимптотической нормальности распределения выборочного среднего и выборочной дисперсии, которое имеет место при условии существования соответственно 2-го и 4-го моментов у наблюдаемой случайной величины ξ, можно построить асимптотические доверительные интервалы, содержащие неизвестное значение теоретических математического ожидания и дисперсии с заданной доверительной вероятностью $P = 1 - \alpha$. Обозначим $z_p = \Phi^{-1}(p)$ – квантиль порядка p стандартной нормальной функции распределения Φ. Тогда с вероятностью, близкой к $P = 1 - \alpha$, при достаточно больших N имеют место неравенства:

$$\overline{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{N}} \leqslant \ \mathbf{E}\xi \ \leqslant \overline{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{N}}\,; \tag{1.1}$$

$$S^2 - z_{1-\frac{\alpha}{2}} \frac{\sqrt{M_{N,4} - M_{N,2}^2}}{\sqrt{N}} \leqslant \ \mathbf{D}\xi \ \leqslant S^2 + z_{1-\frac{\alpha}{2}} \frac{\sqrt{M_{N,4} - M_{N,2}^2}}{\sqrt{N}}\,, \tag{1.2}$$

где $S = \sqrt{S^2} = \sqrt{N^{-1} \sum\limits_{i=1}^{N} \left(X_i - \overline{X}\right)^2}$ – квадратный корень из выборочной дисперсии.

Левые и правые части в неравенствах (1.1) – (2.2) дают нижние и верхние границы доверительных интервалов для неизвестных параметров $\boldsymbol{a} = \mathbf{E}\xi$ и $\sigma^2 = \mathbf{D}\xi$ соответственно.

Порядок выполнения работы

1. Задать объем выборки N.

2. Задать параметры соответствующего распределения (у каждого варианта свои, например: λ, x_0, h, c, l, σ, d и др.; варианты заданий приведены на с. **??**–**??**).

3. Найти аналитически для своего варианта функцию распределения F моделируемого закона и плотность распределения. Вычислить математическое ожидание и дисперсию этого распределения.

4. Смоделировать с помощью генераторов случайных чисел выборку из заданного (согласно своему варианту) распределения.

5. Построить в одном графическом окне графики функции распределения F и эмпирической функции распределения F_n, сравнить их визуально.

6. Построить в одном графическом окне график плотности распределения и гистограмму относительных частот, задав количество интервалов по формуле $m = \boldsymbol{round}(1.72 \cdot n^{1/3})$.

7. Вычислить значения оценок математического ожидания и дисперсии по полученной в п. 4 выборке. Сравнить статистические оценки математического ожидания и дисперсии с их теоретическими значениями.

8. Построить асимптотические доверительные интервалы для «неизвестных» математического ожидания и дисперсии с доверительными вероятностями 0.9, 0.95 и по правилу 3σ (т.е. с доверительной вероятностью 0.9974). Убедиться в том, что теоретические математическое ожидание и дисперсия содержатся в этих интервалах.

9. Найти статистические оценки B_N и E_N параметров асимметрии $\beta = \mu_3/\sigma^3$ и эксцесса $\varepsilon = \mu_4/\sigma^4 - 3$ (см. табл. 1.1). Являются ли эти оценки состоятельными? несмещенными?

10. Найти для своего распределения теоретические медиану и квантили уровней 0.25 и 0.75. Найти по выборке их статистические оценки, т.е. выборочную медиану и выборочные квантили тех же уровней. Сравнить оцени с соответствующими теоретическими значениями. Посмотреть, что будет происходить с увеличением объема выборки N.

ВАРИАНТЫ ЗАДАНИЙ

Вариант 1.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания – $p = 0.2$. Смешиваемые распределения:
F_1 – нормальное с параметрами $a = 1$, $\sigma^2 = 4$;
F_2 – нормальное с параметрами $a = 5$, $\sigma^2 = 1$

Вариант 2.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания – $p = 0.4$. Смешиваемые распределения:
F_1 – экспоненциальное с параметром $\lambda = 3$
F_2 – распределение Симпсона с параметром $c = 1.2, \quad l = 4.8$.

Вариант 3.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания – $p = 0.5$. Смешиваемые распределения:
F_1 – равномерное в интервале $(1, 3)$
F_2 – Распределение Лапласа с параметром $a = 1.3, \quad u = 2.7$.

Вариант 4.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания – $p = 0.3$. Смешиваемые распределения:
F_1 – распределение Вейбулла с параметрами $x_0 = 5.2, \quad d = 6.4$,
F_2 – экспоненциальное распределение с параметром $\lambda = 0.2$.

Вариант 5.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания – $p = 0.3$. Смешиваемые распределения:
F_1 – нормальное с параметрами $a = 3$, $\sigma^2 = 9$;
F_2 – распределение Парето с параметрами $a = 4, \quad x_0 = 3.5$.

Вариант 6.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания – $p = 0.2$. Смешиваемые распределения:
F_1 – распределение Рэлея с параметром $\sigma = 6$;
F_2 – Арксинус-распределение с параметром $d = 5.7$.

Вариант 7.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания – $p = 0.2$. Смешиваемые распределения:
F_1 – распределение Симпсона с параметрами $c = 2.3, \quad l = 6.8$;
F_2 – распределение Парето с параметрами $a = 4, \quad x_0 = 3.5$.

Вариант 8.

F - смесь (выпуклая комбинация) трех распределений, параметры смешивания – $p_1 = 0.2$, $p_2 = 0.3$. Смешиваемые распределения:
F_1 – нормальное с параметрами $a = 0$, $\sigma^2 = 16$;
F_2 – равномерное в интервале $(0, 2)$
F_3 – распределение Лапласа с параметрами $a = 1.3, \quad u = 2.7$.

Вариант 9.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания – $p = 0.4$.
Смешиваемые распределения:
F_1 – Арксинус-распределение с параметром $d = 3$;
F_2 – нормальное с параметрами $a = 0$, $\sigma^2 = 9$;

Вариант 10.

F - смесь (выпуклая комбинация) трех распределений, параметры смешивания – $p_1 = 0.1$, $p_2 = 0.4$.
Смешиваемые распределения:
F_1 – нормальное с параметрами $a = -3$, $\sigma^2 = 10$;
F_2 – распределение Лапласа с параметрами $a = 1.3$, $u = 2.7$
F_3 – распределение Рэлея с параметром $\sigma = 16$.

Вариант 11.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания – $p = 0.35$.
Смешиваемые распределения:
F_1 – распределение Парето с параметром $a = 4$, $x_0 = 3.5$;
F_2 – нормальное с параметрами $a = -1$, $\sigma^2 = 16$;

Вариант 12.

F - смесь (выпуклая комбинация) трех распределений, параметры смешивания – $p_1 = 0.2$, $p_2 = 0.4$.
Смешиваемые распределения:
F_1 – распределение Симпсона с параметрами $c = 1.2$, $l = 4.8$;
F_2 – экспоненциальное распределение с параметром $\lambda = 0.6$
F_3 – Арксинус-распределение с параметром $d = 6.4$.

Вариант 13.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания - $p = 0.5$.
Смешиваемые распределения:
F_1 – распределение Лапласа с параметрами $a = 1.8$, $u = 4.7$
F_2 – нормальное с параметрами $a = 0$, $\sigma^2 = 4$;

Вариант 14.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания - $p = 0.2$.
Смешиваемые распределения:
F_1 – Распределение Парето с параметрами $a = 4$, $x_0 = 0.5$,
F_2 – равномерное в интервале $(-1, 1)$;

Вариант 15.

F - смесь (выпуклая комбинация) трех распределений, параметры смешивания – $p_1 = 0.2$, $p_2 = 0.4$.
Смешиваемые распределения:
F_1 – распределение Лапласа с параметрами $a = 1$, $u = 1$
F_2 – нормальное с параметрами $a = -1$, $\sigma^2 = 8$;
F_3 – равномерное в интервале $(2, 5)$;

Вариант 16.

F - смесь (выпуклая комбинация) двух распределений, параметр смешивания - $p = 0.2$.
Смешиваемые распределения:
F_1 – Распределение Вейбулла с параметрами $x_0 = 2, \quad d = 1$,
F_2 – распределение Лапласа с параметрами $a = -1, \quad u = 3$;

Вариант 17.

F - смесь (выпуклая комбинация) трех распределений, параметры смешивания – $p_1 = 0.1$, $p_2 = 0.2$.
Смешиваемые распределения:
F_1 – нормальное с параметрами $a = 0$, $\sigma^2 = 16$;
F_2 – Арксинус-распределение с параметром $d = 3$;
F_3 – равномерное в интервале $(0, 5)$;

Формулы функций распределения

Показательное распределение:

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0 \quad (\lambda > 0).$$

Распределение Рэлея:

$$F(x) = 1 - \exp\Big(-\frac{x^2}{2\sigma^2}\Big), \ x \geq 0 \ (\sigma > 0).$$

Распределение Симпсона:

$$F(x) = \begin{cases} 0, & x \leq c - l; \\ (x - c + l)^2/(2l^2), & c - l < x \leq c; \\ 1 - (x - c - l)^2/(2l^2), & c < x \leq c + l; \\ 1, & x > c + l. \end{cases}$$

Распределение Лапласа ($a \geq 0, \quad u > 0$):

$$F(x) = \begin{cases} \exp\big((x - a)/u\big)/2, & x \leq a; \\ 1 - \exp\big(-(x - a)/u\big)/2, & x > a. \end{cases}$$

Равномерное распределение на интервале (a, b):

$$F(x) = \frac{x - a}{b - a}, \quad x \in (a, b).$$

Арксинус-распределение:

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{x}{d}, \quad |x| < d \quad (d > 0).$$

Распределение Коши:

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan\Bigl(\frac{x - x_0}{h}\Bigr).$$

Распределение Парето:

$$F(x) = 1 - \left(\frac{x_0}{x}\right)^a, \quad x \geq x_0 \ \ (\, a > 0, \ x_0 > 0).$$

Распределение Вейбулла:

$$F(x) = 1 - \exp\left(-\frac{x^d}{x_0}\right), \quad x > 0 \ \ (\, x_0 > 0, \ \ m > 0).$$

Пример выполнения задания (в MATLAB) (2012 year)

Пусть требуется смоделировать и обработать выборку $X = (X_1 \ldots, X_N)$ объема $N = 1000$ независимых случайных величин, имеющих показательное распределение с параметром $\lambda = 1.5$. Моделирующая формула для показательного распределения имеет вид:

$$X = -\ln\bigl(U\bigr)/\lambda.$$

Приведем программу, реализующую выполнение пп. 1–10 задания. Возь- мем в качестве доверительной вероятности значение $P = 1 - \alpha = 0.95$. По таблицам значений функции $\Phi(x)$ распределения стандартного нормального закона находим, что $z_{1-\frac{\alpha}{2}} = 1.96$.

Пример программы (MATLAB)

clc; clf % Чистка командного окна и графического окна
$N = 1000$; % Ввод количества элементов выборки
$L = 1.5$; % Ввод параметра показательного распределения
$P = 0.95$; % $P = 1 - \alpha = 0.95$; – доверительная вероятность
$z = 1.96$ % $z = z_{1-\frac{\alpha}{2}}$ – квантиль функции Гаусса $\Phi(x)$ (по таблицам Φ)
U=rand(1,N); % Моделирование случайных величин, равномерно распреде-
% ленных на интервале $(0, 1)$
X=-log(U)/L; % Моделирование случайных величин, имеющих
% показательное распределение, по моделирующей формуле
% и получение соответствующей выборки
$E\_t = 1/L$ % Теоретическое значение математического ожидания
$D\_t = 1/(L\hat{}2)$ % Теоретическое значение дисперсии
$E\_s = mean(X)$ % Вычисление выборочного среднего
$S^2 = var(X, 1)$ % Вычисление выборочной дисперсии (смещенной оценки)
$D\_s = var(X)$; % Несмещенная оценка дисперсии
$S = sqrt(S^2)$ % Корень из выборочной дисперсии (оценка среднего квадра-
% тического отклонения)
% Асимптотический доверительный интервал для математического ожидания
$Left\_E = E\_s - z * S/sqrt(N)$ % Нижняя граница 95%-го интервала
$Right\_E = E\_s + z * S/sqrt(N)$ % Верхняя граница
% Асимптотический доверительный интервал для дисперсии
$M_{N,4} = sum((X - E\_s).\hat{}4)/N$ % Четвертый центральный момент
$Left\_D = S\hat{}2 - z * sqrt(M_{N,4} - S\hat{}4)/sqrt(N)$ % Нижняя граница
$Right\_D = S\hat{}2 + z * sqrt(M_{N,4} - S\hat{}4)/sqrt(N)$ % Верхняя граница
$B\_N = mean((X - E\_s).\hat{}3)/S\hat{}3$; % Оценка коэффициента асимметрии
$E\_N = mean((X - E\_s).\hat{}4)/S\hat{}4 - 3$; % Оценка коэффициента экцесса
$X\_s = sort(X)$; % Сортировка выборки
% Границы интервала, в котором расположены наблюдения X_i (выборка)
$a = 0$; % Нижняя граница[1]
$b = X\_s(N) + 1$; % Верхняя граница.
$x = a : 0.01 : b$; % Точки разбиения отрезка (a, b) с шагом 0.01
$F = 1 - exp(-L * x)$; % Значения экспоненциальной функции распределения
% в точках x
$plot(x, F)$; grid % График функции распределения $F(x)$
hold on % Включение режима наложения графиков
$F\_N = (0 : (N - 1))/N$; % Эмпирическая функция распределения

[1]Экспоненциальные случайные величины неотрицательны, для других распределений это, вообще говоря, не так, и можно положить, например, $a = X\_s(1) - 1$.

$plot(X\_s, F\_N, 'r')$;   % График эмпирической функции распределения
pause; clf % пауза до нажатия любой клавиши и чистка графического окна
$f = L * exp(-L * x)$; % Плотность экспоненциального распределения в точках x
$plot(x, f)$; hold on;   % График плотности экспоненциального распределения
$hist1(X)$;   % Обращение к функции, строящей гистограмму относительных
hold off;   % частот. Текст функции $hist1$ приведен ниже, на с. 8.

Отметим, что в системе *MatLAB* есть стандартная функция ***hist***, которая позволяет строить гистограмму частот, но не относительных частот. Кроме того, недостатком функции ***hist*** является то, что при обращении к ней нужно в качестве входного аргумента задать количество столбцов гистограммы, однако рекомендуемое число столбцов зависит от объема выборки и его нужно вычислять перед обращением к функции ***hist***. Поэтому мы создадим свою универсальную функцию построения гистограммы относительных частот ***hist1***, лишенную указанных недостатков. Эта функция должна быть записана в отдельном файле с именем ***hist1.m***, и этот файл должен находиться в той же папке, что и вызывающая функцию ***hist1*** основная программа.

Функция построения гистограммы и полигона частот (MATLAB)

function m = hist1 (X)
$n = length(X)$; % Определение объема обрабатываемой выборки
$m = round(1.72 * n\hat{}(1/3))$; % Рекомендуемое количество столбцов[2].
$Xs = sort(X)$; % Сортировка массива X (выборки)
$a = Xs(1) - 0.01$; % Задание точками a и b границ диапазона значений с.в. ξ
$b = Xs(n) + 0.01$; % случайной величины, на котором будут строится графики
$hh = (b - a)/m$; % Задание длины интервала группировки
$A = a : hh : b$; % Вектор границ интервалов группировки
% Цикл для вычисление частот попадания значений в интервалы группировки[3]
$ch(1) = 0$; % $ch(i)$ – число наблюдений слева от нижней границе i-го интервала
$for\ i = 2 : (m + 1)$
  $j = ch(i - 1) + 1,$
  $while\ j <= n\ \&\ Xs(j) < A(i)$
    $j = j + 1;$
  end;
$ch(i) = j - 1;$
$nu(i - 1) = ch(i) - ch(i - 1)$; % хисло наблюдений, попавших в $(i - 1)$-й интервал
end;
$Ac = (A(1 : m) + A(2 : (m + 1)))/2$; % Координаты середин интервалов
  % группировки
$h\_n = nu/(n * hh)$; % Вектор значений гистограммы в интервалах
$stairs(Ac - hh/2,\ h\_n,\ 'r')$; % Обращение к функции $stairs$, строящей график
hold on;   % кусочно-постоянной функции (построение гистограммы)
$plot(Ac,\ h\_n,\ 'k')$; % Построение графика полигона частот (середины верхних сторон прямоугольников гистограммы MatLAB соединяет ломаной линией по умолчанию).
hold off;

---

[2]Другие эмпирическим путем найденные и хорошо себя зарекомендовавшие значения числа столбцов гистограммы: $m1 = 1 + 3.3 \cdot \log_{10}(n)$ и $m2 = 1 + \log_2(n)$, округленные до целых чисел.

[3]Приведем также второй способ вычисления частот, экономный по числу операторов (но не по времени и памяти):
$for\ i = 1 : m,$ % перебор интервалов группировки
  $nu(i) = length(find(A(i) <= Xs\ \&\ Xs < A(i + 1)))$ % Количество наблюдений, попавших в i-й интервал
end