Практическая работа 1. Анализ одномерной случайной величины.

Цель работы. Произвести анализ одномерной случайной величины на основе ее N независимых измеренных значений.

Теоретические сведения. Первичная обработка полученных в результате случайного эксперимента данных включает в себя:

- построение статистического ряда распределения;

- построение эмпирической функции распределения;

- получение точечных статистических оценок;

- предварительное предположение о характере распределения случайной величины X.

При статистической обработке экспериментальных данных распределение случайной величины X заменяется так называемым выборочным распределением, т.е. выборкой $x_1, x_2, \dots\ x_N$ с вероятностями $P\{X = x_i\} = \frac{1}{N}$.

Если выборка небольшого объема, то статистический ряд распределения представляет собой дискретный ряд распределения (ряд распределения выборочной случайной величины X^*). Если выборка большого объема, то строится интервальный (группированный) статистический ряд.

Для построения интервального ряда распределения необходимо:

1. Упорядочить выборку, т.е. составить вариационный ряд

$$x_1^* \le x_2^* \le x_3^* \dots \le x_{N-1}^* \le x_N^*, \qquad (1)$$

в котором упорядоченные значения $x_1^*, x_2^*, x_3^* \dots, x_N^*$ называют порядковыми статистиками.

2. Найти диапазон выборки $[x_1^*; x_N^*]$ и размах выборки R_B по формуле

$$R_B = x_N^* - x_1^*. \qquad (2)$$

3. Для заданного объема выборки N найти оптимальное число интервалов l, на которые разбивается диапазон выборки. Рекомендуется выбирать

$$l = \log_2 N + 1, \quad 5 \le l \le 25, \qquad (3)$$

4. Найти длину каждого интервала h по формуле

$$h = \frac{R_B}{l}. \qquad (4)$$

5. После этого записать полуоткрытые интервалы $I_1 = [a_0, a_1)$, …, $I_i = [a_{i-1}, a_i)$, …, $I_l = [a_{l-1}, a_l]$, на которые разбит весь диапазон выборки $[x_1^*; x_N^*]$ и границы которых определяются формулами

$$a_1 = x_1^*, a_i = a_0 + i \cdot h, i = 1, \dots l. \quad (5)$$

Замечание. Интервалы выбирают полуоткрытыми, чтобы исключить случай, когда какое-то выборочное значение попадает на границу интервала, и приходится решать, к какому интервалу его отнести. В последнем интервале $I_l = [a_{l-1}, a_l]$ должно быть $a_l = a_0 + l \cdot h \geq x_N^*$. Поэтому его длина может оказаться больше, чем h.

6. Для каждого интервала I_i, $i = 1, \dots l$ с помощью вариационного ряда (1) вычислить числа N_i – количество выборочных значений, попавших в этот интервал. Очевидно, что $\sum_{i=1}^{l} N_i = N$.

7. Все выборочные значения, попавшие в интервал I_i, $i = 1, \dots l$, принимаются равными его середине, а середины интервалов $\tilde{x}_i$ группированного ряда вычисляются по формуле

$$\tilde{x}_i = \frac{a_{i-1}+a_i}{2}, i = 1, \dots l. \quad (6)$$

8. Вычислить частоты p_i^* по формулам

$$p_i^* = \frac{N_i}{N}, \quad (7)$$

где N_i - число выборочных значений, попавших в интервал I_i. Очевидно, что $\sum_{i=1}^{l} p_i^* = 1$.

9. После этого записать интервальный ряд (табл.1), в котором указаны интервалы, середины интервалов, количество выборочных значений в каждом интервале и частоты вычисленные по формуле (7).

Табл.1

I_i	I_1	I_2	…	I_l	Σ
$\tilde{x}_i$	$\tilde{x}_1$	$\tilde{x}_2$	…	$\tilde{x}_l$	-
N_i	N_1	N_2	…	N_l	N
p_i^*	p_1^*	p_2^*	…	p_l^*	1

Построение эмпирической функции распределения. Для этого наряду с интервальным строится *дискретный статистический ряд* (табл.2), а также накопленные частоты z_i, которые вычисляются по формулам

$z_1 = p_1^*, z_2 = p_1^* + p_2^*, \dots, z_i = \sum_{k=1}^{i} p_k^*$.

Табл.2

$\tilde{x}_i$	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	…	$\tilde{x}_l$	Σ
p_i^*	p_1^*	p_2^*	p_3^*	…	p_l^*	1
z_i	z_1	z_2	z_3	…	1	

Построенный дискретный статистический ряд представляет собой приближенное выборочное распределение, а частоты p_i^* являются статистическими оценками вероятностей того, что выборочное значение равно $\tilde{x}_i$.

В качестве приближения функции распределения $F(x)$ генеральной совокупности в статистике рассматривают *эмпирическую функцию распределения*, которая определяется формулой

$$F_N^*(x) = \sum_{i:\tilde{x}_i<x} p_i^*. \quad (8)$$

Аналитическое выражение $F_N^*(x)$ через накопленные частоты z_i имеет вид

$$F_N^*(x) = \begin{cases} 0, \ x \le \tilde{x}_1, \\ z_1, \ \tilde{x}_1 < x \le \tilde{x}_2, \\ z_2, \ \tilde{x}_2 < x \le \tilde{x}_3, \\ \dots\dots\dots\dots\dots\dots\dots \\ z_{l-1}, \ \widetilde{x}_{l-1} < x \le \tilde{x}_l, \\ 1, \ x > \tilde{x}_l. \end{cases} \quad (9)$$

Эмпирическая функция распределения $F_N^*(x)$ ставит в соответствие каждому значению x частоту события $X < x$ и представляет собой кусочно-постоянную функцию со скачками в серединах интервалов $\tilde{x}_i$, $i = 1, \dots l$.

Если X – непрерывная случайная величина, то в качестве статистического аналога функции распределения генеральной совокупности используют *кумуляту*. Кумуляту строят как непрерывную ломаную линию, состоящую из отрезков, соединяющих точки

$(a_0, 0)$ и (a_1, z_1), (a_1, z_1) и (a_2, z_2), (a_2, z_2) и (a_3, z_3), …, (a_{l-1}, z_{l-1}) и (a_l, z_l), а также двух полупрямых: $y = 0$ при $x \le a_0$ и $y = 1$ при $x > a_l$.

Построение гистограммы и полигона частот.

Гистограмма является статистическим аналогом плотности распределения непрерывной генеральной совокупности X. Это кусочно-постоянная функция f_N^*, значения которой на каждом интервале $I_i = (a_{i-1}, a_i)$ определяются формулой

$$f_i^*(x) = \frac{p_i^*}{h}, \quad (10)$$

где частота p_i^* выбирается из таблицы 2; h - длина интервала.

Соединив точки гистограммы с абсциссами $\tilde{x}_i = \frac{a_{i-1}+a_i}{2}$, $i = 1,2, \dots l$, на этом же рисунке можно построить полигон частот.

По виду полигона выдвигается основная гипотеза о характере распределения генеральной совокупности X: нормальное (гауссовское) распределение, равномерное распределение и т.д.

Получение точечных статистических оценок:

- выборочное среднее

$$\bar{x} = \hat{\mu}_x = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad (11)$$

$$\bar{x} = \sum_{i=1}^{l} \tilde{x}_i p_i^*; \quad (12)$$

- несмещенную (исправленную) выборочную дисперсию

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2; \quad (13)$$

$$s^2 = \sum_{i=1}^{l} \tilde{x}_i^2 p_i^* - \bar{x}^2; \quad (14)$$

- несмещенное среднеквадратическое отклонение

$$s_x = +\sqrt{s^2}. \quad (15)$$

Описание работы. Имеется набор (выборка) экспериментальных данных x_1, x_2, … x_N. Требуется произвести обработку этих данных для получения эмпирических характеристик одномерной случайной величины.

Этапы выполнения работы:

1. Составить интервальный ряд распределения;

2. Построить эмпирическую функцию распределения, ее график и кумуляту;

3. Построить гистограмму и полигон, выдвинуть гипотезу о законе распределения;

4. Получить точечные статистические оценки параметров распределения.