Лабораторная работа № 2

Часть 1. «Обработка одномерной выборки»

Краткие теоретические сведения

В результате наблюдений и регистрации массовых случайных явлений получаются эмпирические (статистические) данные или эмпирический материал. Если наблюдаемая величина является случайной, то ее можно изучать методами теории вероятностей. Для понимания характера поведения этой случайной величины нужно знать ее закон распределения.

Определение законов распределения рассматриваемых величин и оценка значений параметров распределения на основании наблюденных значений – основная задача курса «Эмпирические методы программной инженерии». Еще одной задачей курса является исследование методов обработки и анализа эмпирического материала с целью получения определенных выводов, необходимых для организации моделирования процессов с участием рассматриваемых величин.

Рассмотрим случайный эксперимент, связанный со случайной величиной X. Осуществив n независимых повторений этого эксперимента, мы получим последовательность n наблюдений данных значений величины X, которые обозначим $x_1, x_2, ..., x_n$. Эту совокупность значений называют *выборкой*, число ее элементов – *объемом выборки*, а числа x_i - *выборочными значениями*.

Статистическая совокупность G называется *генеральной*, если исследованию подлежат все элементы совокупности. Чаще всего генеральная совокупность бесконечна, выборка есть конечная ее часть, доступная исследователю. Главное требование – независимость элементов выборки, что равносильно требованию случайности извлечения элементов.

Выборочный метод - статистический метод исследования общих свойств совокупности каких-либо объектов на основе изучения свойств лишь части этих объектов, взятых на выборку.

Для анализа и последующей обработки экспериментальных данных составляются *вариационные ряды* – последовательности наблюдаемых значений, записанных в возрастающем порядке. Наблюдаемые значения случайной величины *X* называются *вариантами*.

Пусть из генеральной совокупности извлечена выборка, причем варианта x_1 наблюдалась n_1 раз, x_2 – n_2 раз, x_k – n_k раз и $\sum n_i = n$ – объем выборки. Последовательность вариант, записанных в порядке возрастания, называют *вариационным* рядом. Количество наблюдений называют *частотами*, а их отношения к объему выборки $\frac{n_i}{n} = W_i$ – *относительными частотами.*

Пример 1. Объем потребления яблок в различных регионах Украины за месяц (тыс. ящиков): 100; 115; 80; 85; 95; 112; 125; 112; 112; 110; 130; 129; 80; 90; 95; 118; 120; 95; 110; 95; 95; 110; 128; 135.

Для примера 1 вариационный ряд имеет вид: 80, 80, 85, 90, 95, 95, 95, 95, 95, 100, 110, 110, 110, 112, 112, 112, 115, 118, 120, 125, 128, 129, 130, 135.

Частотным рядом называют перечень упорядоченных вариант и соответствующих им частот или относительных частот.

Для примера 1 частотный ряд имеет вид:

x_i	80	85	90	95	100	110	112	115	118	120	125	128	129	130	135
n_i	2	1	1	5	1	3	3	1	1	1	1	1	1	1	1
W_i	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{5}{24}$	$\frac{1}{24}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$

На практике большие выборки из распределений подвергаются группировке в виде *интервального* ряда. Интервальный статистический ряд для выборки строится следующим образом:

1) отыскиваются $x_{\min}$ и $x_{\max}$ (минимальная и максимальная варианты);

2) находится $R = x_{\max} - x_{\min}$ – амплитуда (размах) изменчивости выборки;

3) промежуток $[x_{\min}, x_{\max}]$ разбивается на k интервалов, где k - количество частичных интервалов, которое рекомендуется определить по формуле Стерджеса:

$$k = [1 + 3{,}32 \lg n];$$

4) шаг интервального ряда h получают как отношение:

$$h = \frac{x_{\max} - x_{\min}}{k};$$

5) для каждого частичного интервала находится n_i, $i = 1,2,...,k$, где n_i – число вариант, попавших в i-й интервал;

6) составляется таблица:

$[x_{\min}, x_{\min} + h]$	$(x_{\min} + h, x_{\min} + 2h]$	…	$[x_{\max} - h, x_{\max}]$
n_1	n_2	…	n_k

Для примера 1: $x_{\min} = 80$, $x_{\max} = 135$, $R = 55$, $k = 5$, $h = 11$. Тогда интервальный ряд имеет вид:

	[80, 91]	(91,102]	(102,113]	(113, 124]	(124, 135]
n_i	4	6	6	3	5

Примечание 1. Дискретный вариационный ряд получается из интервального, если за новые значения признака *X* взять середины интервалов.

Дискретный вариационный ряд для примера 1 имеет вид:

x_i	85,5	96,5	107,5	118,5	129,5
n_i	4	6	6	3	5
W_i	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{5}{24}$

Для наглядности используют графические представления эмпирических значений.

Полигон частот – ломаная, отрезки которой соединяют точки $(x_1; n_1)$, $(x_2; n_2)$, …, $(x_k; n_k)$. Для построения полигона частот на оси абсцисс

откладывают варианты x_i, а на оси ординат – соответствующие им частоты n_i. Точки $(x_i;n_i)$ соединяют отрезками прямых.

Для примера 1 полигон частот представлен на рис. 1.1.

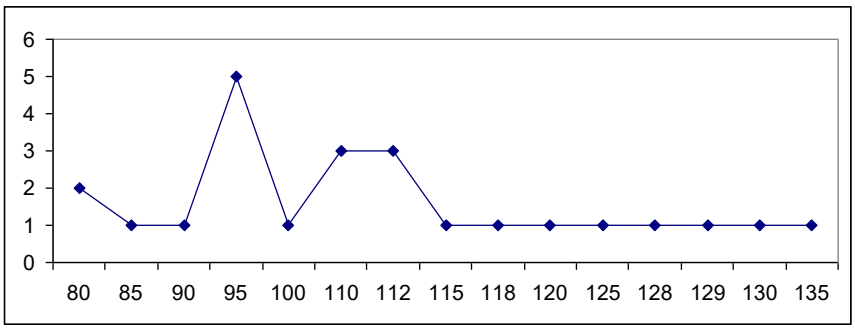

Рисунок 1.1 - Полигон частот

Полигон относительных частот – ломаная, отрезки которой соединяют точки $(x_1;W_1)$, $(x_2;W_2)$, … .

Гистограмма частот – ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{n_i}{h}$ (плотность частоты).

Для построения гистограммы частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии $\frac{n_i}{h}$.

Для примера 1 гистограмма частот представлена на рис. 1.2.

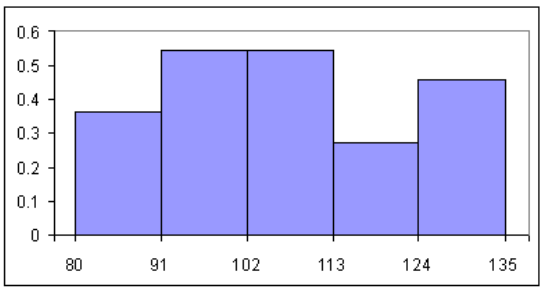

Рисунок 1.2 - Гистограмма частот

Площадь i-го частичного прямоугольника равна $\frac{h \cdot n_i}{h} = n_i$ – сумме частот вариант i-го интервала; следовательно, *площадь гистограммы частот* равна *сумме всех частот,* т. е. *объему выборки.*

Гистограмма относительных частот – ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{W_i}{h}$ (плотность относительной частоты).

Для построения гистограммы относительных частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии $\frac{W_i}{h}$. Площадь гистограммы относительных частот равна единице. Гистограмма является первым приближением плотности распределения количественного признака X генеральной совокупности.

Пусть известно статистическое распределение частот количественного признака X. Введем обозначения: n_l – число наблюдений, при которых наблюдалось значение признака, меньшее x; n – общее число наблюдений (объем выборки). Ясно, что относительная частота события $\{X < x\}$ равна $\frac{n_l}{n}$. Если x будет изменяться, то будет изменяться и относительная частота, то есть относительная частота $\frac{n_l}{n}$ есть функция от x. Так как эта функция находится эмпирическим (опытным) путем, то ее называют эмпирической.

Таким образом, эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события. Эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности.

По определению $F^*(x)=\frac{n_l}{n}$, где n_l – число вариант, меньших x, n – объем выборки.

Для примера 1 эмпирическая функция распределения имеет вид:

$$F^*(x)=\begin{cases}0, x<80\\ \frac{1}{6}, 80\le x<91\\ \frac{5}{12}, 91\le x<102\\ \frac{2}{3}, 102\le x<113\\ \frac{19}{24}, 113\le x<124\\ \frac{23}{24}, 124\le x<135\\ 1, x\ge 135\end{cases}$$

Из определения функции $F^*(x)$ вытекают следующие ее свойства:

1) значения эмпирической функции принадлежат отрезку [0,1]:

$$0\le F^*(x)\le 1;$$

2) $F^*(x)$ – неубывающая функция;

3) если x_q – наименьшая варианта, то $F^*(x)=0$, при $x\le x_q$, если x_q – наибольшая варианта, то $F^*(x)=1$ при $x>x_q$.

Контрольные вопросы

1. Что называется выборкой?
2. Дайте определение частотного ряда.
3. Как построить полигон частот?
4. Чем гистограмма частот отличается от полигона частот?
5. Дайте определение эмпирической функции распределения. Какими свойствами она обладает?

Порядок выполнения лабораторной работы

1. Для данных своего варианта построить вариационный, частотный и интервальный ряды. Количество частичных интервалов определить по формуле Стерджеса. В случае получения пустых частичных интервалов ($n_i = 0$) выполнить объединение с соседним (соседними).

2. Приписывая частоты n_i серединам интервалов, получить выборку в виде дискретного вариационного ряда.

3. Построить полигон, частотный полигон, гистограмму, эмпирическую функцию распределения $F^*(x)$. Для функции распределения проверить выполнение основных свойств.

Задания к лабораторной работе

Вариант №1

Время решения контрольной задачи (мин):

38; 60; 41; 51; 33; 42; 45; 21; 53; 60; 68; 52; 47; 46; 49; 49; 14; 57; 54; 59; 11; 47; 28; 48; 58; 32; 42; 58; 61; 90.

Вариант №2

Время восстановления диодов из одной партии (в наносекундах):

69; 73; 70; 68; 61; 73; 70; 72; 67; 70; 66; 70; 76; 68; 71; 71; 68; 70; 64; 65; 72; 70; 70; 69; 66; 70; 77; 69; 71; 74; 72; 72; 72; 68; 70; 67; 71; 67; 72; 69; 66; 75; 76; 69; 71; 67; 70; 73; 71; 74.

Вариант №3

Масса метал. заготовок для производства подшипников (в граммах):

41,6; 41,7; 41,8; 42,2; 41,2; 40,9; 41,3; 41,5; 41,7; 41,8; 41,4; 41,1; 41,4; 41,5; 42,0; 42,3; 41,6; 41,5; 41,3; 41,4; 41,3; 41,2; 41,1; 41,6; 41,9; 41,2; 42,0; 41,6; 41,7; 41,5.

Вариант №4

Данные представляют собой изменения предела прочности на разрыв, выраженные в тыс. фунтов на квадратный дюйм для низкоуглеродистого стального листа:

51,6; 51,7; 51,8; 52,2; 51,2; 50,9; 51,3; 51,5; 51,7; 51,8; 51,4; 51,1; 51,4; 51,5; 51,6; 51,7; 51,5; 52,0; 52,3; 51,6; 51,5; 51,3; 51,4; 51,3; 51,2; 51,1; 51,6; 51,9; 51,2; 52,0.

Вариант №5

Масса вещества (в граммах):

14; 17; 15; 14; 17; 21; 22; 17; 19; 23; 14; 14,2; 15; 14; 16; 20; 15; 24; 14; 15; 14; 15; 16; 15.

Вариант №6

Результаты биохимического анализа – количество эритроцитов (10^{12}/л): 3,4; 3,6; 4; 3,5; 3,7; 4,1; 3,4; 4,2; 3,7; 3,4; 4,2; 2,8; 4; 3,5; 3,6; 4; 3,7; 4,3; 3,6; 3,9; 3,6; 3,5; 3,1; 3; 3,6; 3,7; 3,2; 3,6; 3,9; 3,3.

Вариант №7

Удельное сопротивление микросхем после легирования ($Ом \cdot мм^2 / м$): 52,2; 33; 76; 32,2; 49,5; 32,5; 191,5; 112,5; 32,9; 114,8; 33,7; 69,1; 112,5; 48,5; 16,5; 50; 51; 39; 66; 40; 70.

Вариант №8

Отклонения от номинального размера имеют следующие значения (в мм): 0,02; 0,07; 0,13; 0,05; 0,11; 0,17; 0,17; 0,05; 0,03; 0,11; 0,04; 0,14; 0,10; 0,11; 0,13; 0,14; 0,16; 0,07; 0,13; 0,04; 0,03; 0,13; 0,11; 0,06; 0,10; 0,13; 0,16; 0,06; 0,14; 0,06; 0,04; 0,08; 0,14; 0,08; 0,08; 0,14; 0,13; 0,07; 0,16; 0,08; 0,02; 0,10; 0,16; 0,04; 0,09; 0,15; 0,12; 0,17; 0,17; 0,16.

Вариант №9

Данные представляют собой урожайность зерновых культур (ц/га): 28,0; 21,0; 27,6; 16,2; 29,7; 26,8; 30,3; 15,7; 25,5; 15,8; 40,6; 27,2; 16,1; 19,8; 25,3; 16,9; 31,4; 26,5; 20,4; 15,6; 20,3; 37,4; 23,2; 38,3; 28,0; 29,0; 37,7; 34,7; 30,0; 20,6.

Вариант №10

Изменения предела прочности на разрыв, выраженные в тысячах фунтов на квадратный дюйм для низкоуглеродистого стального листа:

50,5; 51,1; 50,9; 51,4; 51,7; 51,8; 51,1; 50,7; 51,2; 51,4; 50,9; 51,0; 51,4; 51,3; 51,5; 51,6; 52,2; 51,2; 51,0; 50,9; 50,7; 50,6; 51,3; 51,6; 50,7; 50,9; 51,2; 51,7; 51,8; 51,3.

Вариант №11

При токе 10мА прямое падение напряжения на диодах (в вольтах): 0,917; 0,918; 0,921; 0,909; 0,919; 0,917; 0,918; 0,909; 0,916; 0,917; 0,918; 0,919; 0,919; 0,916; 0,917; 0,923; 0,920; 0,916; 0,917; 0,922; 0,915; 0,917; 0,916.

Вариант №12

Глубина слоя диффузии для партии микросхем (в мкм):

9,8; 9,8; 8,6; 8,5; 9,2; 9,2; 9,8; 9,0; 10,0; 8,8; 10,1; 9,4; 9,0; 11,2; 10,8; 9,2; 9,4; 9,3; 10,1; 9,1; 10,0; 9,5.

Вариант №13

Масса метал. заготовок для производства обойм подшипников(в граммах)

40,5; 41,1; 40,9; 41,4; 41,7; 41,8; 41,1; 40,7, 41,2; 41,4; 40,9; 41,0; 41,4; 41,3; 41,5; 41,6; 42,2; 41,2; 41,1; 40,9; 40,7; 40,6; 41,3; 41,6; 40,7; 40,9; 41,2; 41,7; 41,8; 41,3.

Вариант №14

Затраты на питание 42 студентов в день (в гривнах):

48; 44; 40; 51; 44; 18; 46; 57; 57; 34; 38; 47; 48; 52; 39; 41; 39; 38; 43; 29; 45; 54; 38; 28; 48; 28; 47; 52; 33; 40; 45; 40; 55; 45; 32; 32; 56; 41; 52; 36; 50; 57.

Вариант №15

Недельное потребление продукции (в кг):

0,5; 0,7; 1; 0,8; 1,2; 1,3; 1,4; 0,7; 0,8; 1; 1,8; 1,2; 0,5; 1; 1,3; 0,9; 0,8; 1,4; 1,5; 1.

Вариант №16

Показатели времени для 24 бегунов на одной дистанции (в секундах):

33,5; 20,1; 23,6; 26,3; 19,9; 16,7; 23,2; 31,4; 28,2; 35,3; 29,3; 30,5; 25; 24,2; 19; 20; 28; 17,9; 24; 25; 20,2; 19,5; 20,5; 23.

Вариант №17

Результаты измерения удельного сопротивления микросхем после легирования поликремния ($Ом \cdot мм^2 / м$):

119; 17,5; 43,5; 43,5; 90,5; 40; 50; 108; 62,4; 18,5; 97,5; 96; 46; 100; 84; 35; 94; 75; 68; 70; 72.

Вариант №18

Продолжительность работы электронных ламп одного типа в часах:

13,4; 17,7; 15,2; 15,1; 13,0; 21,9; 14,0; 17,9; 15,1; 16,5; 16,6; 14,2; 16,3; 14,6; 11,7; 16,4; 15,1; 17,6; 14,1; 18,8; 11,6; 13,9; 18,0; 12,4; 17,2; 14,5; 16,3; 13,7; 15,5; 16,2; 18,4; 14,7; 13,4.

Вариант №19

Результаты биохимического анализа – количество эритроцитов (10^{12}/л):

2,7; 3,5; 3,9; 3,4; 3,8; 3,5; 3,7; 3,6; 3,8; 3,1; 3,6; 4,2; 2,9; 4,3; 3,1; 3,8; 3,7; 3,5; 4,4; 3,9; 3,8; 4; 3,9; 3,8; 3,8; 3,8; 3,2; 4,7; 3,9; 3,9.

Вариант №20

Результаты измерения предела прочности на разрыв, выраженные в тысячах фунтов на квадратный дюйм для низкоуглеродистого стального листа: 51,0; 51,3; 51,4; 51,3; 51,7; 51,9; 51,5; 51,3; 51,2; 51,0; 51,0; 51,0; 51,4; 51,4; 51,6; 51,7; 52,1; 51,8; 51,8; 51,4; 51,5; 51,2; 51,0; 51,5; 51,5; 51,4; 51,3; 51,4; 51,4; 51,5.

Вариант №21

Производительность цеха в течение 20 рабочих дней (в усл. единицах): 13,0; 13,1; 13,0; 12,5; 12,8; 12,3; 12,1; 12,2; 12,1; 12,7; 12,0; 12,6; 12,8; 12,5; 13,1; 13,2; 12,6; 12,4; 13,0; 12,9.

Вариант №22

Данные о числе тонн грузов, перевозимых еженедельно паромом некоторого морского порта в период навигации:

398; 412; 560; 474; 544; 690; 587; 600; 613; 457; 504; 477; 530; 641; 359; 566; 452; 633; 474; 499; 580; 606; 344; 455; 505; 396; 347; 441; 390; 632; 400; 582.

Вариант №23

Результаты биохимического анализа – содержание серотонина (мкг/л): 102; 94; 102; 90; 100; 79; 88; 101; 113; 99; 100; 98; 103; 99; 91; 96; 97; 115; 94; 97; 90; 102; 97; 54; 102; 79; 83; 103; 100; 79.

Вариант №24

Результаты биохимического анализа крови 30 человек – содержание витамина А (мкмоль/л):

3,9; 2,6; 2,5; 2,4; 4,3; 2,9; 2,8; 2,2; 2,8; 2,3; 2,2; 2,5; 2,3; 2,0; 2,5; 2,6; 2,1; 3,1; 2,8; 3,3; 2,5; 2,2; 1,2; 2,2; 2,1; 1,6; 2,7; 2,5; 1,9; 2,2.

Вариант №25

Данные представляют собой потребление искусственного шелка в 25 странах (в $м^2$ на душу населения): 8,9; 8,26; 7,74; 7,66; 7,11; 8,58; 7,13; 6,87;

6,47; 7,07; 5,84; 6,14; 6,77; 7,6; 7,89; 7,97; 8,08; 7,53; 8,39; 7,95; 6,45; 6,6; 6,0; 5,91; 6,5.

Часть 2: «Точечные и интервальные оценки характеристик генеральной совокупности»

Цель работы: получение практических навыков по определению основных выборочных характеристик количественного признака генеральной совокупности.

Краткие теоретические сведения

Статистические оценки параметров распределения. Обычно в распоряжении исследователя имеются лишь выборочные данные. Если из теоретических соображений удалось установить, какое именно распределение имеет признак генеральной совокупности, то возникает задача оценки параметров, которыми определяется это распределение. Для описания случайных величин используются описательные статистики: минимум, максимум, среднее, дисперсия, стандартное отклонение, медиана, мода и т.д. Статистики дают общее представление о значениях, которые принимают случайные величины. Получаемые оценки могут носить точечный и интервальный характер.

Оценка называется *точечной,* если определяется одним числом; *интервальной* – если по данным выборки строится числовой интервал, внутри которого на основании заранее выбранной вероятности находится оцениваемый параметр.

Оценка должна быть близка к оцениваемому параметру. Близость характеризуется несмещенностью оценки, ее состоятельностью и эффективностью.

Несмещенность оценки означает отсутствие систематических погрешностей в наблюдаемых данных, для этого ее математическое ожидание должно быть равно оцениваемому параметру.

Состоятельность оценки заключается в том, что с ростом числа наблюдений дисперсия стремится к нулю.

Для исследуемого параметра оценка эффективна, если имеет минимальную дисперсию среди всех возможных оценок, построенных по данной выборке.

Пусть из генеральной совокупности извлечена выборка объема n. *Выборочное среднее* (m^*) – сумма значений переменной, делённая на n (число значений переменной)

$$m^* = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Выборочное среднее может быть посчитано по частотно-вариационному ряду

$$m^* = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n},$$

где k – количество вариантов в ряду, или по интервальному ряду

$$m^* = \frac{\sum_{i=1}^{k} x'_i \cdot n_i}{n},$$

где x'_i - середина i-го интервала, k - количество интервалов.

Среднее выборочное является несмещенной, состоятельной и эффективной оценкой математического ожидания генеральной совокупности, т.е. точечная оценка математического ожидания является доброкачественной

$$\tilde{x} = m^*.$$

Выборочная дисперсия (D)*- мера изменчивости случайной величины. Вычисляется по формуле:

$$D^* = \frac{\sum_{i=1}^{n} (x_i - m^*)^2}{n}.$$

Значение 0 означает отсутствие изменчивости, т.е. переменная постоянна. Выборочная дисперсия является смещенной оценкой дисперсии генеральной совокупности, поэтому доброкачественной оценкой генеральной дисперсии является исправленная выборочная дисперсия

$$\tilde{D} = D^* \frac{n}{n-1} = \frac{\sum_{i=1}^{n} \left(x_i - m^*\right)^2}{n-1}.$$

Выборочное стандартное отклонение (*S*) - корень квадратный из дисперсии. Более удобная характеристика, так как измерена в тех же единицах, что и исходная величина. Чем выше дисперсия и стандартное отклонение, тем сильнее разбросаны значения случайной величины относительно среднего. Для оценки среднего квадратичного отклонения генеральной совокупности применяют выборочное среднее квадратичное отклонение

$$S = \sqrt{D^*}$$

или исправленное среднее квадратичное отклонение

$$\tilde{S} = \sqrt{\tilde{D}} = \sqrt{\frac{n}{n-1} D^*}.$$

Для более подробного описания свойств распределения вводятся эмпирические начальные

$$\lambda^p = \frac{\sum_{i=1}^{n} x_i{}^p}{n}$$

и центральные

$$\mu^p = \frac{\sum_{i=1}^{n} (x_i - x^*)^p}{n}$$

моменты *p*-го порядка или их комбинаций. В частности, *коэффициент асимметрии* позволяет судить о симметричности выборочных данных

$$A = \frac{n \cdot \mu^3}{\left[(n-1) \cdot (n-2) \cdot S^3\right]}$$

Если коэффициент значительно отличается от 0, распределение является асимметричным. Показатель эксцесса служит мерой крутизны (заостренности) гистограммы по отношению к кривой нормального распределения (для нормально распределенной случайной величины *E=0*).

$$E = \frac{\left[n\cdot(n+1)\cdot\mu^4 - 3\cdot\mu^2\cdot\mu^2\cdot(n-1)\right]}{\left[(n-1)\cdot(n-2)\cdot(n-3)\cdot S^4\right]}.$$

Медиана – значение, которое разбивает выборку на две равные части. Половина наблюдений лежит выше медианы, и половина – ниже. В некоторых случаях, например, при описании доходов населения медиана более удобна, чем среднее.

Медиана дает общее представление о том, где сосредоточены значения переменной, иными словами, где находится ее центр. Сумма *абсолютных* расстояний между точками выборки и медианой *минимальна*. Медиана вычисляется следующим образом. Выборка упорядочивается в порядке возрастания. Если количество элементов в выборке определяется как *2m+1* (нечетно), то медиана выборки оценивается как $Me = x_{m+1}$. Если число наблюдений четно, то медиана оценивается как $Me = (x_m + x_{m+1})/2$.

Квантиль – число t_p, ниже которого находится *p*-я часть (доля) выборки.

Процентиль – значение квантили в процентах.

Мода – наиболее часто встречающееся выборочное значение, варианта, имеющая наибольшую частоту.

Доверительным интервалом для параметра θ называется интервал $(\theta^* - \delta,\ \theta + \delta)$, который с заданной надежностью β покрывает реальное значение параметра θ, здесь θ^* – оценка параметра, δ – точность оценки. Число $\beta = 1 - \alpha$ называется доверительной вероятностью, а значение α - уровнем значимости. В качестве β, как правило, выбираются значения, близкие к единице: 0,95; 0,99; 0,999.

Точечная оценка m^* даже, если она несмещенная, состоятельная, эффективная дает приближенное значение параметра генеральной

совокупности и, особенно для выборок малого объема, отличается от истинного значения параметра, т.е. от m.

Представление о том, к каким ошибкам может привести замена параметра m на его точечную оценку m^* и с какой степенью уверенности можно ожидать, что эти ошибки не выйдут за известные пределы дает *мера достоверности* (или *интервальная оценка*).

В качестве меры достоверности принимают:

1) *доверительную вероятность* β (точный метод), с которой истинное значение параметра а будет находится в заданном относительно стат. оценки интервале;

2) *доверительный интервал* $I_\beta(m^*-\varepsilon; m^*+\varepsilon)$ (грубый метод) относительно статистической оценки, в который с заданной вероятностью β попадет истинное значение параметра m.

Понятие оценки меры достоверности. Назначим некоторую достаточно большую вероятность (β = 0,9; 0,95; 0,997) такую, что событие с этой вероятностью β можно считать практически достоверным.

Требуется найти доверительный интервал: $P(a_n^{(1)} < a < a_n{}^{(2)}) = \beta$,

где границы интервала $a_n^{(1)}; a_n^{(2)}$ – *доверительные границы*.

Интервальная оценка параметра a (доверительный интервал) - числовой интервал $I_\beta[a] = (a*_n^{(1)}; a*_n^{(2)})$ относительно статистической оценки параметра, который с заданной вероятностью β накрывает реальное значение параметра a.

Чаще всего доверительный интервал выбирают *симметричным* относительно статистического параметра (см. рис. 2.1).

$$I_\beta[a] = (a*_n^{(1)}; a*_n^{(2)}),\ a*_n^{(1)} = a_n^* - \varepsilon;\ a*_n^{(2)} = a_n^* + \varepsilon.$$

$$P(|a_n^* - a| < \varepsilon) = \beta;\ P(a_n^* - \varepsilon < a < a_n^* + \varepsilon) = \beta;\ I_\beta[a] = (a_n^* - \varepsilon; a_n^* + \varepsilon).$$

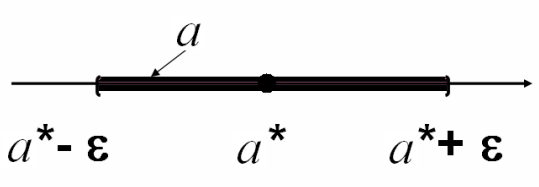

Рисунок 2.1 - Симметричный доверительный интервал

$\alpha = 1 - \beta$ - уровень значимости, вероятность того, что расхождения между параметром и его оценкой больше либо равно абсолютной величине доверительного интервала:

$$P(\left|a - a_n^*\right| \geq \varepsilon) = \alpha .$$

Чаще всего $\alpha = 0{,}05;\ 0{,}1$.

Доверительный интервал – числовой интервал значений параметра a ГС, которые не противоречат опытным данным или совместимы с опытными данными. Границы интервала и его величина получены по выборочным данным и поэтому случайны в отличие от самого параметра a.

Величина доверительного интервала ε существенно зависит:

- от объема выборки (с ростом n величина интервала уменьшается;
- от величины доверительной вероятности: чем больше доверительная вероятность β, тем больше ε.

Оценка доверительного интервала для математического ожидания. Пусть для параметра m генеральной совокупности получена доброкачественная оценка $m*$. Нужно оценить полученную при этом ошибку «грубым» и «точным» методами. Определение I_β возможно, если известен закон распределения статистической оценки, который зависит от закона распределения самой СВ, и от конкретного значения параметра ГС.

«Грубый метод» используется при следующих допущениях:

- допущение нормальности закона распределения СВ;
- замена параметров этого закона их статистическими оценками.

Пусть имеется случайная величина X – описывающая ГС, с неизвестными параметрами m, D. Найти доверительный интервал для m, если задана доверительная вероятность и получены результаты эксперимента. Т.е., дано: $x_1,...,x_n; \beta$. Найти: $I_\beta[m]: a = m_X; a^* = m^*$.

Известно, что статистическая оценка математического ожидания равна:

$$m^* = \tilde{m} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

В качестве оценки реального m по выборке принимается среднее арифметическое n независимых наблюденных значений.

x_i – некоторый экземпляр случайной величины X с параметрами m_X. Оценка m_X - это сумма n независимых одинаково распределенных СВ, тогда, по центральной предельной теореме при достаточно большом n закон распределения этой суммы близок к нормальному.

В практической статистике даже при относительно небольшом числе испытаний (от 10 до 20) считается, что закон распределения стремится к нормальному. Тогда, вероятность попадания в интервал для нормального закона равна:

$$P(a < X < b) = \Phi^*\left(\frac{b - m_X}{\sigma_X}\right) - \Phi^*\left(\frac{a - m_X}{\sigma_X}\right),$$

В симметричный интервал $\pm\varepsilon$ относительно m_X:

$$P(m_X - \varepsilon < X < m_X + \varepsilon) = \Phi\left(\frac{\varepsilon}{\sigma_X}\right) - 1 + \Phi\left(\frac{\varepsilon}{\sigma_X}\right) = 2\Phi\left(\frac{\varepsilon}{\sigma_X}\right) - 1 = 2\Phi^*\left(\frac{\varepsilon}{\sigma_X}\right)$$

.

Рассматриваемая СВ X - это оценка матожидания: $\tilde{m} = m^* \approx N(m, \frac{D}{n})$,

$$P(|\tilde{m}-m|<\varepsilon)=\beta=2\Phi(\frac{\varepsilon}{\sigma_{m^*}})-1\Rightarrow\Phi(\frac{\varepsilon}{\sigma_{m^*}})=\frac{1+\beta}{2}\Rightarrow\varepsilon=\sigma_{m^*}\cdot\arg\Phi\left(\frac{1+\beta}{2}\right)=$$

$$=\sigma_{m^*}\cdot\arg\Phi^*\left(\frac{\beta}{2}\right),\ \sigma_{m^*}=\frac{\sigma_X}{\sqrt{n}}\Rightarrow\varepsilon=\frac{\sigma_X}{\sqrt{n}}\cdot\arg\Phi(\frac{1+\beta}{2})=\frac{\sigma_X}{\sqrt{n}}\cdot\arg\Phi*(\frac{\beta}{2}).$$

Величина доверительного интервала для матожидания равна (“грубый метод”):

$$u_\beta=\arg\Phi\left(\frac{1+\beta}{2}\right)=u_{1-\alpha/2},$$

где $u_\beta=u_{1-\alpha/2}$ - квантиль нормального распределения. Тогда

$$I_\beta[m]=\left(m^*-u_\beta\cdot\sqrt{\frac{D^*}{n}};m^*+u_\beta\cdot\sqrt{\frac{D^*}{n}}\right).$$

Для примера 1 $m^*\approx107{,}33$, $D^*\approx250{,}64$, $u_{0,9}=0{,}3289$, $u_{0,95}=0{,}3340$, $u_{0,99}=0{,}3389$. Тогда $I_{0,9}[m]=(81{,}82;132{,}84)$, $I_{0,95}[m]=(81{,}42;133{,}23)$, $I_{0,99}[m]=(81{,}04;133{,}62)$.

Полученные с помощью «грубого» метода границы интервалов для математического ожидания, нанесем на полигон частот (см. рис. 2.2).

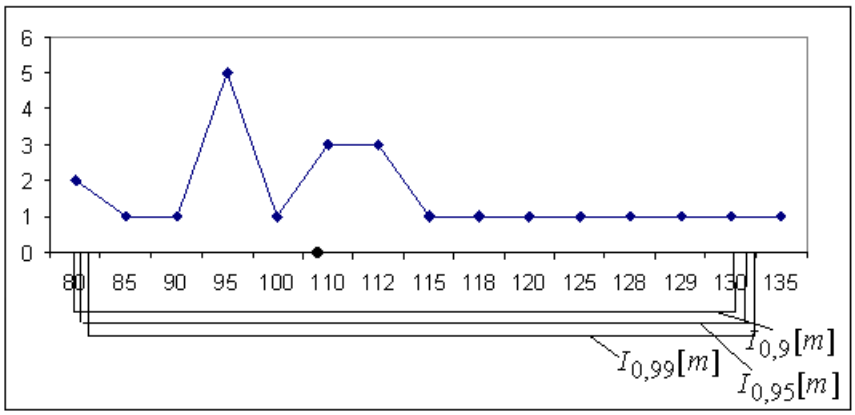

Рисунок 2.2 - Границы доверительных интервалов для мат. ожидания

“Точный ” метод оценки достоверности матожидания. Если σ не известно, то используют σ^* и вместо нормального распределения t-распределение Стьюдента:

$$\varepsilon = \frac{\sigma^*}{\sqrt{n}} t_{1-\alpha/2}(n-1),$$

где $t_{1-\alpha/2}$ - квантиль t-распределения (табличное значение).

Доверительный интервал для D_X. Дана СВ X с нормальным законом распределения и неизвестными параметрами m и D. Произведено n независимых испытаний. Требуется по заданной доверительной вероятности найти доверительный интервал для D.

В качестве оценки D принимаем:

$$\tilde{D} = \frac{\sum_{i=1}^{n}(x_i - \tilde{m})^2}{n-1}.$$

По аналогии с математическим ожиданием, ***оценка D грубым методом:***

$$I_\beta[D] = (\tilde{D} - \varepsilon; \tilde{D} + \varepsilon)\ ,\ \varepsilon = u_\beta \cdot \sigma[\tilde{D}],\ D[\tilde{D}] = \frac{\mu_4}{n} - \frac{(n-3)}{n(n-1)} \cdot D_X^2$$

Чтобы воспользоваться этими формулами вместо реальных D и μ_4 пользуются их оценками:

$$\mu_4^*[X] = \frac{\sum_{i=1}^{n}(x_i - m^*)^4}{n}$$

Нормальный закон:

$$D[\tilde{D}] = \frac{2}{n-1} \cdot D^2.$$

Равномерный:

$$D[\tilde{D}] = \frac{0{,}8n + 1{,}2}{n(n-1)} \cdot D^2.$$

Оценка D «точным» методом: если m известно, то

$$I_\beta^D = \left\{ \frac{nD^*}{\chi^2_{1-\alpha/2}(n)}; \frac{nD^*}{\chi^2_{\alpha/2}(n)} \right\};$$

если m неизвестно, то берут m^*:

$$I_{\beta}^{D} = \left\{ \frac{nD^*}{\chi^2_{1-\alpha/2}(n-1)}; \frac{nD^*}{\chi^2_{\alpha/2}(n-1)} \right\}.$$

Контрольные вопросы

1. В чем разница между точечной и интервальной оценками?
2. Дайте определение медианы, моды, квантиля и процентиля.
3. Что такое мера достоверности и доверительный интервал?
4. Что такое мат. ожидание и дисперсия?
5. Чем «точный» метод оценки отличается от «грубого» метода?

Порядок выполнения лабораторной работы

1. По выборочным значениям из задач первой части для своего варианта найти доброкачественные точечные оценки числовых характеристик параметров генеральной совокупности $M(x), D(x), \sigma(x).$

2. Найти характеристики вариационного ряда для исследуемого признака: моду M_o, медиану Me, размах варьирования R, среднее абсолютное отклонение θ, коэффициент вариации V.

3. «Грубым» и «точным» методами получить интервальные оценки для $M(x)$ и $D(x)$ генеральной совокупности, задаваясь значениями доверительной вероятности $\beta_1 = 0{,}9,\ \beta_2 = 0{,}95,\ \beta_3 = 0{,}99$.

4. Полученные границы интервалов для $M(x)$, полученные «грубым» и точным методами нанести на полигоны частот или частостей.