ВЫБОРКА И ЕЕ РАСПРЕДЕЛЕНИЕ

1.1. Генеральная совокупность и выборка

Математическая статистика возникла в XVIII веке в работах Я. Бернулли, П. Лапласа, К. Пирсона. В ее современном развитии определяющую роль сыграли труды Г. Крамера, Р. Фишера, Ю. Неймана и др. Большой вклад в математическую статистику внесли русские ученые П. Л. Чебышев, А. М. Ляпунов, А. Н. Колмогоров, Б. В. Гнеденко и другие.

Предметом **математической статистики является изучение случайных величин (или случайных событий, процессов) по результатам наблюдений. Полученные в результате наблюдения (опыта, эксперимента) данные сначала надо каким-либо образом обработать: *упорядочить*, представить в удобном для обозрения и анализа виде. Это первая задача. Затем, это уже вторая задача, *оценить*, хотя бы приблизительно, *интересующие* нас *характеристики* наблюдаемой случайной величины. Например, дать оценку неизвестной вероятности события, оценку неизвестной функции распределения, оценку математического ожидания, оценку дисперсии случайной величины, оценку параметров распределения, вид которого неизвестен, и т. д.**

Базовыми понятиями математической статистики являются ***генеральная совокупность*** и ***выборка***.

Определение. Генеральная совокупность – это совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины.

Генеральная совокупность может быть конечной или бесконечной в зависимости от того, конечна или бесконечна совокупность составляющих ее объектов.

Зачастую проводить *сплошное обследование*, когда изучаются все объекты (например — перепись населения), трудно или дорого, экономически нецелесообразно (например — не вскрывать же каждую консервную банку для проверки качества продукции), а иногда невозможно. В этих случаях наилучшим способом обследования является *выборочное наблюдение*: выбирают из генеральной совокупности часть ее объектов («выборку») и подвергают их изучению.

Определение. Выборкой (выборочной совокупностью) называется совокупность случайно отобранных объектов из генеральной совокупности.

Выборка должна быть **репрезентативной** (представительной), то есть ее объекты должны достаточно хорошо отражать свойства генеральной совокупности. В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если ее осуществлять *случайно*, т.е. каждый из объектов генеральной совокупности имеет одинаковую вероятность попасть в выборку.

Источники настоящих случайных чисел найти крайне трудно. **Генераторы случайных чисел бывают**:

- ***Аппаратный генератор случайных чисел** (генератор истинно случайных чисел)*
- ***Генератор псевдослучайных чисел** (ГПСЧ, англ.* **Pseudorandom number generator, PRNG***)*

Аппаратный генератор случайных чисел (**генератор истинно случайных чисел**) — устройство, которое генерирует последовательность случайных чисел на основе измеряемых параметров протекающего физического процесса. Работа таких устройств часто основана на использовании надёжных источников энтропии, таких как тепловой шум, фотоэлектрический эффект, квантовые явления и т. д. Эти процессы в теории абсолютно непредсказуемы. Физические шумы, такие как детекторы событий ионизирующей радиации, дробовой шум в резисторе или космическое излучение, могут быть такими источниками. Однако применяются такие устройства в приложениях сетевой безопасности редко, ввиду дороговизны и медленности. Сложности также вызывают грубые атаки на подобные устройства.

Криптографические приложения используют для генерации случайных чисел особенные алгоритмы. Эти алгоритмы заранее определены и, следовательно, генерируют последовательность чисел, которая теоретически не может быть статистически случайной. В то же время, если выбрать хороший алгоритм, полученная численная последовательность будет проходить большинство тестов на случайность. Такие числа называют псевдослучайными числами.

Генератор псевдослучайных чисел (**ГПСЧ**, англ. *Pseudorandom number generator, PRNG*) — алгоритм, порождающий последовательность чисел, элементы которой почти независимы друг от друга и подчиняются заданному распределению (обычно равномерному).

Современная информатика широко использует псевдослучайные числа в самых разных приложениях — от метода Монте-Карло и имитационного моделирования до криптографии. При этом от качества используемых ГПСЧ напрямую зависит качество получаемых результатов. Это обстоятельство подчёркивает известный афоризм математика Роберта Кавью : «*генерация случайных чисел слишком важна, чтобы оставлять её на волю случая*».

Никакой детерминированный алгоритм не может генерировать полностью случайные числа, он может только аппроксимировать некоторые их свойства. Как сказал Джон фон Нейман, «*всякий, кто питает слабость к арифметическим методам получения случайных чисел, грешен вне всяких сомнений*».

Любой ГПСЧ с ограниченными ресурсами рано или поздно зацикливается — начинает повторять одну и ту же последовательность чисел. Длина циклов ГПСЧ зависит от самого генератора и составляет около $2^{n/2}$, где n — размер внутреннего состояния в битах. Если порождаемая последовательность ГПСЧ сходится к слишком коротким циклам, то такой ГПСЧ становится предсказуемым и непригодным для практических приложений.

Большинство простых арифметических генераторов хотя и обладают большой скоростью, но страдают от многих серьёзных недостатков:

- ***Слишком короткий период/периоды.***

- ***Последовательные значения не являются независимыми.***
- ***Некоторые биты «менее случайны», чем другие.***
- ***Неравномерное одномерное распределение.***
- ***Обратимость.***

В частности, алгоритм RANDU, десятилетиями использовавшийся на мейнфреймах, оказался очень плохим, что вызвало сомнения в достоверности результатов многих исследований, использовавших этот алгоритм.

Наиболее распространены линейный конгруэнтный метод, метод Фибоначчи с запаздываниями, регистр сдвига с линейной обратной связью, регистр сдвига с обобщённой обратной связью.

Метод статистического исследования, состоящий в том, что на основе изучения выборочной совокупности делается заключение о всей генеральной совокупности, называется *выборочным*.

Число N объектов генеральной совокупности и число n объектов выборки называют объемами генеральной и выборочной совокупностей соответственно.

Выборка может быть повторной, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность, и бесповторной, при которой отобранный объект не возвращается в генеральную совокупность. При предположении, что $N >> n$ (значительно больше), различие между бесповторной и повторной выборками очень мало, его можно не учитывать.

Применяют различные способы получения выборки.

1) Простой отбор – случайное извлечение по одному объекту из генеральной совокупности.

2) Типический отбор, когда объекты отбираются не из всей генеральной совокупности, а из ее «типической» части (мнение о референдуме спросить у случайно отобранных людей, разделенных по признаку пола, возраста и т.д.).

3) Серийный отбор – объекты отбираются из генеральной совокупности не по одному, а сериями (выборка, в которой единицы отбора представляют собой статистические серии: семьи, бригады и другие совокупности статистически различимых единиц).

4) Механический отбор – генеральная совокупность «механически» делится на столько частей, сколько объектов должно войти в выборку и из каждой части выбирается один объект (мнение спросить у каждого шестидесятого).

На практике используют сочетание этих методов выборки.

Обычно, хотя и не всегда, статистические данные естественным образом разделены на части, отвечающие отдельным наблюдениям. Такие части мы будем выделять в наших обозначениях индексом, например, наблюдения $X_1, X_2, ..., X_n$. Каждое наблюдение трактуется как случайная величина (в простейшем случае одномерная), а ее реализовавшееся значение $x_1, x_2, ..., x_n$. Так, выражение $P(X_1 = x_1)$ следует понимать как вероятность того, что случайная величина X_1 примет значение x_1. Наряду с подобными выражениями будут употребляться и вида $P(X_1 = x)$. Здесь символом x обозначено одно из возможных значений случайной величины

X_1, которому не приписывается роль реализовавшегося. Следует особо отметить, что $X_1, X_2, ..., X_n$ является совокупностью попарно независимых одинаково распределенных случайных величин.

Совокупность наблюдений обычно упорядочена в виде последовательности. При этом номер наблюдения чаще всего имеет одно из двух толкований либо момент времени, либо номер объекта (скажем, нефтеперерабатывающего завода, банка) из совокупности одновременно рассматриваемых объектов. В первомслучае последовательность наблюдений называется time series (временной ряд), а во втором cross-section (общепринятого русского эквивалента нет, один из вариантов перевода пространственные данные). Иногда это различие удобно подчеркнуть обозначением индекса: $t = 1, 2, ..., T$ или $i = 1, 2, ..., N$. В отдельных задачах встречаются "двумерные" массивы данных X_{it} так называемые панельные данные (panel data).

1.2. Вариационные и статистические ряды

Полученные различными способами отбора данные образуют выборку, обычно это множество чисел, расположенных в беспорядке. По такой выборке трудно выявить какую-либо закономерность их изменения (варьирования).

Для обработки данных используют операцию ранжирования, которая заключается в том, что результаты наблюдений над случайной величиной, то есть наблюдаемые значения случайной величины располагают в порядке *не убывания*.

Пример 1. Дана выборка: 2, 4, 7, 3, 1, 1, 3, 2, 7, 3

❍ Проведем ранжирование выборки: 1, 1, 2, 2, 3, 3, 3, 4, 7, 7 ●

После проведения операции ранжирования значения случайной величины объединяют в группы, то есть группируют так, что в каждой отдельной группе значения случайной величины одинаковы. Каждое такое значение называется **вариантом.** Варианты обозначаются строчными буквами латинского алфавита с индексами, соответствующими порядковому номеру группы x_i , y_j , ...

Изменение значения варианта называется варьированием.

Определение. Последовательность вариантов, записанных в порядке не убывания, называется вариационным рядом.

Определение. Число, которое показывает, сколько раз встречаются соответствующие значения вариантов в ряде наблюдений, называется частотой или весом варианта и обозначается n_i, где i - номер варианта.

Определение. Отношение частоты данного варианта к общей сумме частот называется относительной частотой или частостью (долей) соответствующего варианта и обозначается $p_i^* = \left(\frac{n_i}{n}\right)$ или $p_i^* = \frac{n_i}{\sum_{i=1}^{m} n_i}$, где m - число вариантов.

Частость является статистической вероятностью появления варианта x_i.

Естественно считать частость p_i^* аналогом вероятности p_i появления значения x_i случайной величины X.

Определение. Дискретным статистическим рядом называется ранжированная совокупность вариантов x_i с соответствующими им частотами n_i или частостями p_i^*.

Дискретный статистический ряд удобно записывать в виде табл.1.

Таблица 1 (для примера 1)

x_i	1	2	3	4	7
n_i	2	2	3	1	2
$\frac{n_i}{n}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{10}$

$$\sum_{i=1}^{5} n_i = 10;$$

$$\sum_{i=1}^{5} p_i^* = 1.$$

Характеристики дискретного статистического ряда:

1. Размах варьирования $R = x_{max} - x_{min}$.

2. Мода $\left(M_0^*\right)$ - вариант, имеющий наибольшую частоту

(в примере 1. $M_0^* = 3$).

3. Медиана $\left(M_e^*\right)$ - значение случайной величины, приходящееся на середину вариационного ряда.

Пусть n - объем выборки (вариационного ряда).

Если $n = 2k$, то есть ряд имеет четное число членов, то $M_e^* = \frac{x_k + x_{k+1}}{2}$. Если $n = 2k+1$, то есть ряд имеет нечетное число членов, то $M_e^* = x_{k+1}$.

(в примере 1. $M_e^* = 3$).

Если изучаемая случайная величина X является непрерывной или число значений ее велико, то по вариационному ряду составляют *интервальный ряд распределения*.

Определение. Интервальным вариационным рядом называется упорядоченная совокупность интервалов варьирования значений случайной величины с соответствующими им частотами ν_i или частостями p_i^* попадания в каждый из них значений величины.

Интервальный ряд распределения представляет собой следующую таблицу:

j	A_j	B_j	h_j	ν_j	p_i^*	f_i^*
1	A_1	B_1	h_1	ν_1	p_1^*	f_1^*
⋮						
m	A_m	B_m	h_m	ν_m	p_m^*	f_m^*

Здесь j – номер интервала;

m – число непересекающихся и примыкающих друг к другу интервалов, на которые разбивается диапазон значений $[x_{\min}, x_{\max}]$:

$$m \approx \begin{cases} \text{int}\left(\sqrt{n}\right), n \le 100, \\ \text{int}\left((2 \div 4) \cdot \lg(n)\right), n > 100, \ \textit{формула Стерджеса} \end{cases}$$

где int(x) – целая часть числа x .

A_j, B_j – левая и правая границы j-го интервала ($B_j = A_{j+1}$ – интервалы примыкают друг к другу), причем $A_1 = x_{\min}$, $B_m = x_{\max}$;

$h_j = B_j - A_j$ – длина j-го интервала;

ν_j – количество чисел в выборке, попадающих в j-й интервал, $\sum_{j=1}^{m} \nu_j = n$;

$p_j^* = \frac{\nu_j}{n}$ – частота попадания в j-й интервал; $\sum_{j=1}^{m} p_j^* = 1$.

$f_j^* = \frac{p_j^*}{h_j} = \frac{\nu_j}{n h_j}$ – статистическая плотность вероятности в j-м интервале.

При построении интервального статистического ряда вероятностей используют следующие методы разбиения диапазона значений на интервалы:

1) *равноинтервальный*, т.е. все интервалы одинаковой длины:

$$h_j = h = \frac{x_{\max} - x_{\min}}{m}, \ \forall j \Rightarrow A_j = x_{\min} + (j-1)h, j = \overline{2,m}$$

2) *равновероятностный*, т.е. границы интервалов выбирают так, чтобы в каждом интервале было одинаковое число выборочных значений (необходимо, чтобы n без остатка делилось на m):

$$\nu_j = \nu = \frac{n}{m}, p_j^* = \frac{1}{m} \forall j \Rightarrow A_j = \frac{x_{(j-1)\nu} + x_{(j-1)\nu+1}}{2}, j = \overline{2,m}$$

Иногда интервальный статистический ряд, для простоты исследований, условно заменяют дискретным. В этом случае серединное значение i -го интервала принимают за вариант x_i, а соответствующую интервальную частоту ν_i - за частоту этого варианта.

1.3. Эмпирическая функция распределения (Оценка закона распределения)

Пусть получено статистическое распределение выборки и каждому варианту из этой выборки поставлена в соответствие его частость.

Определение. Эмпирической функцией (функцией распределения выборки) называется функция $F^*(x)$, определяющая для каждого значения x частость события $\{X < x\}$,

$$F^*(x) = p^*(X < x) = \frac{n_x}{n} = \begin{cases} 0, & x \le x_1, \\ \vdots & \\ \dfrac{\sum\limits_{k=1}^{i} n_k}{n}, & x_i < x \le x_{i+1} \\ \vdots & \\ 1, & x > x_n. \end{cases},$$

В случае, если варианты в выборке не повторяются, формула принимает вид:

$$F^*(x) = p^*(X < x) = \frac{n_x}{n} = \begin{cases} 0, & x \le x_1, \\ \vdots & \\ \dfrac{i}{n}, & x_i < x \le x_{i+1} \\ \vdots & \\ 1, & x > x_n. \end{cases}$$

где n - объем выборки, n_x - число наблюдений, меньших x $(x \in R)$.

По теореме Бернулли при увеличении объема выборки частость события $\{X < x\}$ приближается к вероятности этого события. ***При*** $n \to \infty$ ***эмпирическая функция распределения*** $F^*(x)$ ***сходится по вероятности к теоретической функции распределения*** $F(x)$**.** Статистическая эмпирическая функция $F^*(x)$ является оценкой интегральной функции $F(x)$ в теории вероятностей.

Функция $F^*(x)$ обладает теми же свойствами, что и функция $F(x)$:

1. $0 \le F^*(x) \le 1$;
2. $F^*(x)$ -неубывающая функция;
3. $F^*(x)$ -непрерывная слева функция;
3. $\lim\limits_{x \to -\infty} F^*(x) = 0$, $\lim\limits_{x \to +\infty} F^*(x) = 1$.

Пример 2. Построить эмпирическую функцию и ее график по данным табл.1

❍

$$F^*(x) = \begin{cases} 0 \quad при \quad x \le 1; \\ 0,2 \quad при \quad 1 < x \le 2; \\ 0,4 \quad при \quad 2 < x \le 3; \\ 0,7 \quad при \quad 3 < x \le 4; \\ 0,8 \quad при \quad 4 < x \le 7; \\ 1 \quad при \quad x > 7; \end{cases}$$

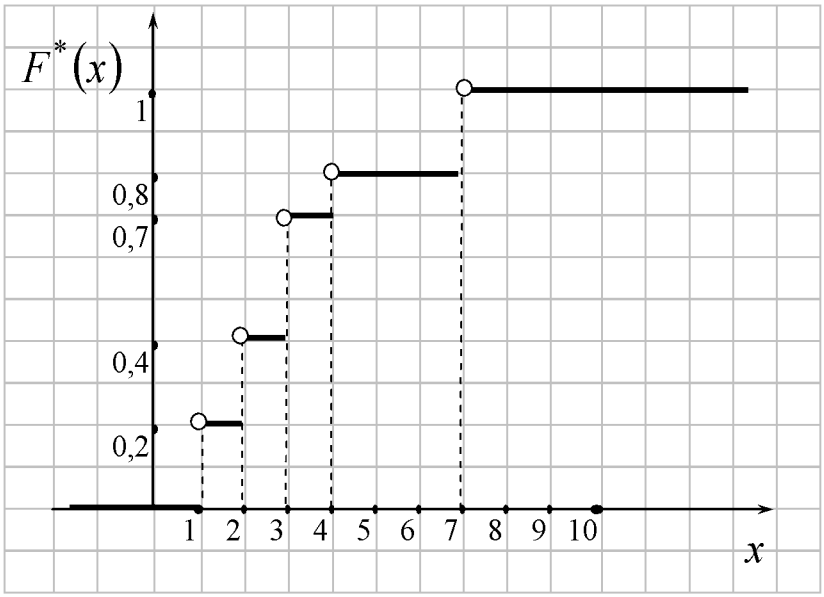

Рис. 1

1.4. Эмпирическая плотность распределения

Для интегральной функции распределения $F(x)$ справедливо равенство:

$$f(x)=\lim_{\Delta x\to 0}\frac{P(x<X<x+\Delta x)}{\Delta x}=\lim_{\Delta x\to 0}\frac{F(x+\Delta x)-F(x)}{\Delta x}=F'(x),$$

где $f(x)$- дифференциальная функция распределения (функция плотности вероятности).

Пренебрегая пределом, получаем приближенное равенство:

$$F(x+\Delta)-F(x)\approx f(x)\cdot\Delta x,$$

потому естественно выборочным аналогом функции $f(x)$ считать функцию:

$$f^*(x)=\frac{F^*(x+\Delta x)-F^*(x)}{\Delta x},$$

где $F^*(x+\Delta x)-F^*(x)$- частость попадания наблюдаемых значений случайной величины X в интервал $[x;\,x+\Delta x)$. Таким образом, значение $f^*(x)$ характеризует плотность частости на этом интервале.

Пусть наблюдаемые значения непрерывной случайной величины представлены в виде интервального вариационного ряда.

Полагая, что p_i^* - частость попадания наблюдаемых значений в интервал $[a_i;\,a_i+h_i)$, где h_i - длина i – го частичного интервала, выборочную функцию плотности $f^*(x)$ можно задать соотношением

$$f^*(x)=\begin{cases}0,\ при\ \ x<a_1,\\ \dfrac{p_i^*}{h_i},\ при\ \ a_i\le x\le a_{i+1},\ \ i=1,2,\ldots m,\\ 0,\ при\ \ x>a_{m+1},\end{cases}$$

где a_{m+1} - конец последнего m - го интервала.

Так как функция $f^*(x)$ является аналогом распределения плотности случайной величины, площадь области под графиком этой функции равна 1.

1.5. Графическое изображение статистических данных

Статистическое распределение изображается графически с помощью полигона и гистограммы.

Определение. Полигоном частот называют ломаную, отрезки которой соединяют точки с координатами (x_i, n_i); полигоном частостей – с координатами (x_i, p_i^*), где $p_i^* = \dfrac{n_i}{n}$, $i = \overline{1,m}$.

Полигон служит для изображения дискретного статистического ряда. Полигон частостей является аналогом многоугольника распределения дискретной случайной величины в теории вероятностей.

Определение. Гистограммой частот (частостей) называют ступенчатую фигуру, состоящую из прямоугольников, основания которых расположены на оси Ox и длины их равны длинам частичных интервалов h_i, а высоты равны отношению $p_j^* = \dfrac{\nu_j}{n}$ - для гистограммы частот; $f_j^* = \dfrac{p_j^*}{h_j} = \dfrac{\nu_j}{nh_j}$ - для гистограммы частостей.

Гистограмма является графическим изображением интервального ряда. Для равноинтервального метода все прямоугольники гистограммы имеют одинаковую ширину, а для равновероятностного метода – одинаковую площадь. Площадь гистограммы частот равна n, а гистограммы частостей равна 1.

Гистограмма строится по интервальному статистическому ряду и представляет собой статистический аналог графика плотности вероятности $f^*(x)$ случайной величины.

Можно построить полигон для интервального ряда, если его преобразовать в дискретный ряд. В этом случае интервалы заменяют их серединными значениями и ставят в соответствие интервальные частоты (частости). Полигон получим, соединив отрезками середины верхних оснований прямоугольников гистограммы.

Пример 3. Дана выборка значений случайной величины X объема 20:

12, 14, 19, 15, 14, 18, 13, 16, 17, 12
18, 17, 15, 13, 17, 14, 14, 13, 14, 16

Требуется: - построить дискретный статистический ряд;
- найти размах варьирования R, моду M_0, медиану M_e;
- построить полигон частостей.

❍ 1) Ранжируем выборку: 12, 12, 13, 13, 13, 14, 14, 14, 14, 14,
15, 15, 16, 16, 17, 17, 17, 18, 18, 19.

2) Находим частоты вариантов и строим дискретный статистический ряд (табл.3)

Таблица 3.

Значения вариантов x_i	12	13	14	15	16	17	18	19	
Частоты n_i	2	3	5	2	2	3	2	1	$\sum_{i=1}^{8} n_i = 20$
Частости $p_i^* = \frac{n_i}{n}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\sum_{i=1}^{8} p_i = 1$

3) По результатам таблицы 3 находим:

$$R = 19 - 12 = 7,\ M_0 = 14,\ M_e = \frac{x_{10} + x_{11}}{2} = \frac{14+15}{2} = 14{,}5$$

4) Строим полигон частостей.

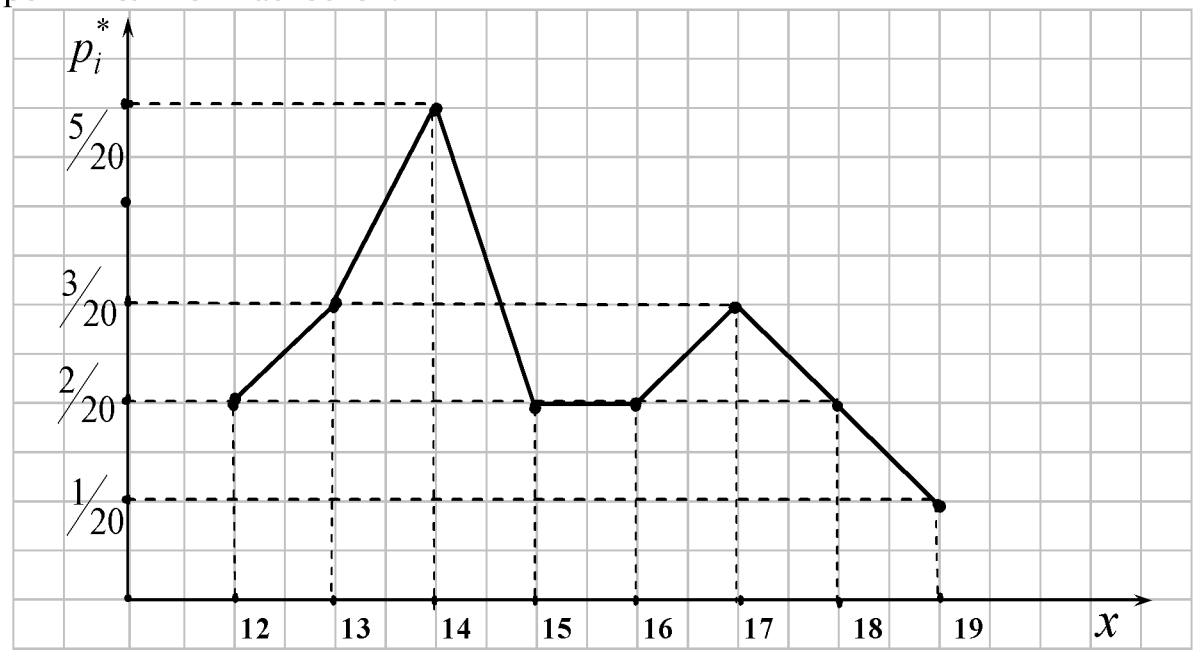

Рис. 2

Пример 4. Результаты измерений отклонений от нормы диаметров 50 подшипников дали численные значения (в мкм), приведенные в табл. 4.

Таблица 4.

-1,760	-0,291	-0,110	-0,450	0,512
-0,158	1,701	0,634	0,720	0,490
1,531	-0,433	1,409	1,740	-0,266
-0,058	0,248	-0,095	-1,488	-0,361
0,415	-1,382	0,129	-0,361	-0,087
-0,329	0,086	0,130	-0,244	-0,882
0,318	-1,087	0,899	1,028	-1,304
0,349	-0,293	0,105	-0,056	0,757
-0,059	-0,539	-0,078	0,229	0,194
0,123	0,318	0,367	-0,992	0,529

Для данной выборки: - построить интервальный вариационный ряд;
- построить гистограмму и полигон частостей.

❍ 1. Строим интервальный ряд.

По данным таблицы 4 определяем: $x_{min} = -1{,}76$; $x_{max} = 1{,}74$

Для определения длины интервала h используем формулу Стерджеса:

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322\, lg\, 50}.$$

Число интервалов $m \approx 1 + 3{,}322\, lg\, 50$.

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322\, lg\, 50} = \frac{1{,}74 - (-1{,}76)}{1 + 3{,}322\, lg\, 50} \approx \frac{3{,}5}{1 + 3{,}322\, lg\, 50} \approx \frac{3{,}5}{6{,}644} \approx 0{,}526$$

Примем h=0,6 , $m = 7$.

За начало первого интервала примем величину

$$x_{нач} = x_{min} - \frac{h}{2} = -1{,}76 - 0{,}3 = -2{,}06 .$$

Конец последнего интервала должен удовлетворять условию:

$$x_{кон} - h \le x_{max} < x_{кон} .$$

Действительно, $2{,}14 - 0{,}6 \le 1{,}74 < 2{,}14$; $1{,}54 \le 1{,}74 < 2{,}14$.

Строим интервальный ряд (табл. 5).

Таблица 5.

Интервалы	$[-2{,}06; -1{,}46)$	$[-1{,}46; -0{,}86)$	$[-0{,}86; -0{,}26)$	$[-0{,}26; 0{,}34)$
Частоты n_i	2	6	11	15
Частости p_i	$\frac{2}{50}$	$\frac{6}{50}$	$\frac{11}{50}$	$\frac{15}{50}$

Интервалы	$[0{,}34; 0{,}94)$	$[0{,}94; 1{,}54)$	$[1{,}54; 2{,}14)$
Частоты n_i	11	3	2
Частости p_i	$\frac{11}{50}$	$\frac{3}{50}$	$\frac{2}{50}$

$\sum_{i=1}^{7} n_i = 50$;

$\sum_{i=1}^{7} p_i = 1$.

Строим гистограмму частостей.

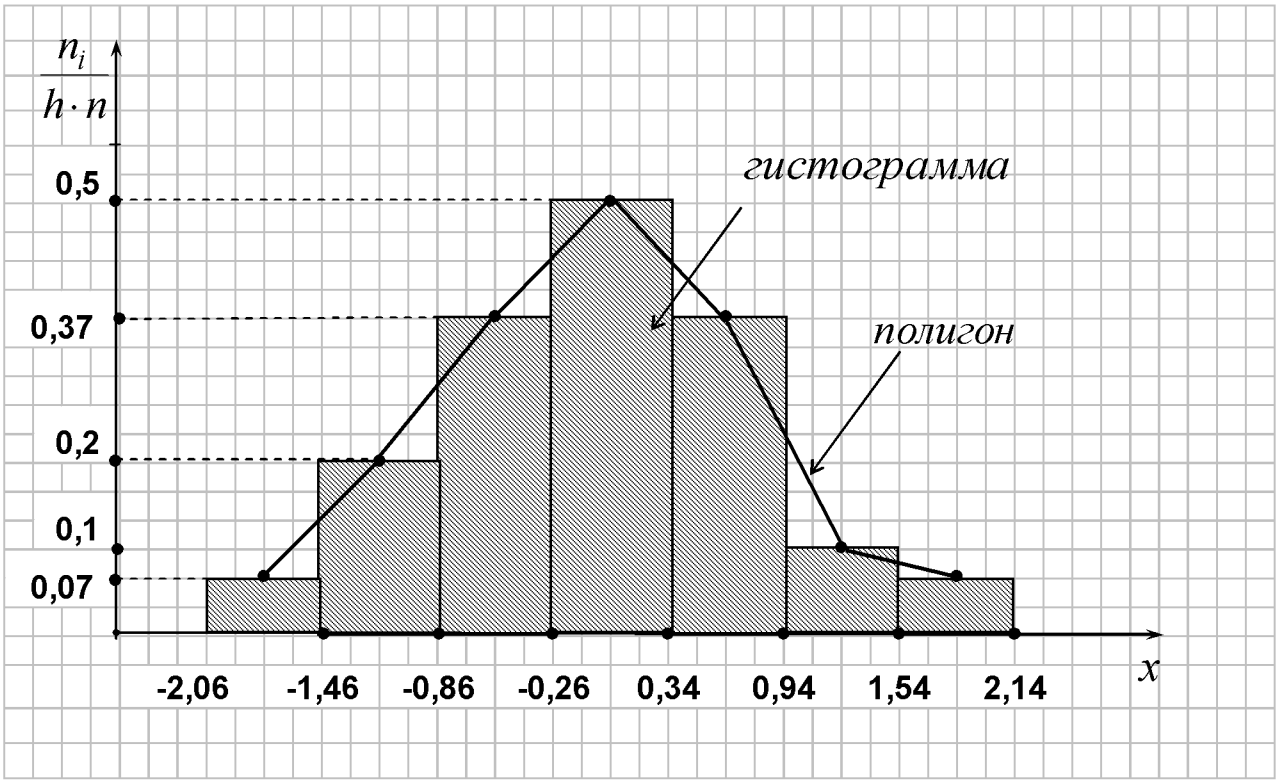

Рис.3

Вершинами полигона являются середины верхних оснований прямоугольников гистограммы.

Убедимся, что площадь гистограммы равна 1.

$$S = h \cdot \left(\frac{n_1 + n_2 + \ldots + n_m}{n \cdot h} \right)$$

$$S = 0{,}6\left(0{,}07 + 0{,}2 + 0{,}37 + 0{,}5 + 0{,}37 + 0{,}1 + 0{,}07\right) = 0{,}6 \cdot 1{,}68 = 1{,}008 \approx 1 \bullet$$