Математическая статистика. Лекция №1

Математическая статистика

Математическая статистика - это раздел прикладной математики, в котором рассматриваются методы отыскания законов и характеристик случайных величин по результатам наблюдений и экспериментов.

Основные задачи математической статистики*:*

1. Создание методов сбора и группировки обрабатываемого статистического материала, полученного в результате наблюдений за случайными процессами.

2. Разработка методов анализа полученных статистических данных.

3. Получение выводов по данным наблюдений.

Анализ статистических данных включает **оценку вероятностей события**, функции распределения вероятностей или плотности вероятностей, оценку параметров известного распределения, оценку связей между случайными величинами.

Математическая статистика опирается на **теорию вероятностей** и в свою очередь служит основой для разработки методов обработки и анализа статистических результатов в конкретных областях человеческой деятельности.

Генеральная совокупность

Основными понятиями математической статистики являются генеральная совокупность и выборка.

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.

Генеральная совокупность (продолжение)

Не следует **смешивать понятие** генеральной совокупности с реально существующими совокупностями. Например, на склад поступила продукция некоторого цеха за месяц, что является реально существующей совокупностью, которую нельзя назвать генеральной, поскольку выпуск продукции можно мысленно продолжить сколь угодно долго.

Выборка

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.

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.

Способы получения выборки:

1) **Простой отбор** – случайное извлечение объектов из генеральной совокупности с возвратом или без возврата.

2) **Типический отбор**, когда объекты отбираются не из всей генеральной совокупности, а из ее «типической» части.

3) **Серийный отбор** – объекты отбираются из генеральной совокупности не по одному, а сериями.

4) **Механический отбор** - генеральная совокупность «механически» делится на столько частей, сколько объектов должно войти в выборку и из каждой части выбирается один объект.

Число **N** объектов генеральной совокупности и число **n** объектов выборки - **объемы** генеральной и выборочной совокупностей соответственно. При этом предполагают, что $N>>n$ (значительно больше).

Ранжирование выборки

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**: наблюдаемые значения случайной величины располагают в порядке возрастания.

Ранжирование выборки

Пример 1. Дана выборка :

Проведем ранжирование выборки :

После проведения операции ранжирования значения случайной величины группируют так, что в каждой отдельной группе значения случайной величины одинаковы. Каждое такое значение - **вариант**.

Варианты обозначаются строчными буквами латинского алфавита с индексами, соответствующими порядковому номеру группы .

Изменение значения варианта называется **варьированием.**

Вариационный ряд

Вариационный ряд- последовательность вариантов, записанная в возрастающем порядке.

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

Отношение частоты данного варианта к общей сумме частот называется **относительной частотой** или **частность** (долей) соответствующего варианта и обозначается или ,

$$p_i^* = \left(\frac{n_i}{n} \right) \qquad p_i^* = \frac{n_i}{\sum_{i=1}^{m} n_i}$$

где **m** – чисто вариантов. Частность является статистической вероятностью появления варианта. Естественно считать частность аналогом вероятности появления значения случайной величины X.

Дискретный статистический ряд

Дискретным статистическим рядом называется ранжированная совокупность вариантов (x_i) с соответствующими им частотами (n_i) или частностями (p_i^*).

Дискретный статистический ряд удобно записывать в виде таблицы.

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** (M_o^*) – вариант , имеющий наибольшую частоту
3. **Медиана** (M_e^*) – значение случайной величины, приходящееся на середину ряду.

Пусть n - объём выборки.

Если n=2k, то есть ряд имеет чётное число членов, то

$$M_e^* = \frac{x_k + x_k + 1}{2} \quad .$$

Если n=2k+1, то есть ряд имеет нечётное число членов, то

$$M_e^* = x_{k+1} \quad .$$

Если изучаемая случайная величина X является непрерывной или число значений её велико, то составляют **интервальный статистический ряд**.

Сначала определяют число интервалов m , в зависимости от объёмов выборки с помощью таблицы:

Объем выборки	25-40	40-60	60-100	100-200	более 200
Число интервалов	5-6	6-8	7-10	8-12	10-15

Затем определяют длину частичного интервала h:

$$h = \frac{x_{max} - x_{min}}{m}$$, где **h** – шаг, **m** – число интервалов.

Более точно шаг можно рассчитать с помощью формулы Стерджеса:

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

,

число интервалов $m \approx 1 + 3{,}322 * \lg n$.

Если шаг окажется дробным, то за длину интервала берут ближайшее целое число или ближайшую простую дробь (обычно берут интервалы одинаковые по длине, но могут быть интервалы и разной длины.)

За начало первого интервала рекомендуется брать величину $x_{\text{нач}} = x_{min} - \frac{h}{2}$, а конец последнего должен удовлетворять условию: $x_{\text{кон}} - h \leq x_{max} \leq x_{\text{кон}}$.

Промежуточные интервалы получают, прибавляя к концу предыдущего интервала шаг.

Просматривая результаты наблюдений, определяют количество значений случайной величины, попавшей в каждый конкретный интервал. При этом в интервал включают значения большие или равные нижней границе интервала и меньшие – верхней границы.

В первую строку таблицы статистического распределения вписывают частичные промежутки:

$$[x_0, x_1), [x_1, x_2), \dots, [x_{m-1}, x_m).$$

Во вторую строку статистического ряда вписывают количество наблюдений n_i , (где i = 1, m), попавших в каждый интервал, то есть, частоты соответствующих интервалов.

Эмпирическая функция распределения.

Пусть получено статистическое распределение выборки, и каждому варианту из этой выборки поставлена в соответствии его частность.

Эмпирической функцией (функцией распределения выборки) называется функция F*(x), определяющая для каждого значения x частость события $F^*(x) = \frac{n_x}{n}$,

где n – число выборки , n_x – число наблюдений, меньших x $\{X < x\}$($x \in$ R). При увеличении объёма выборки частость события приближается к вероятности этого события.

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. $F^*(-\infty)=0,\ F^*(+\infty)=1.$

Эмпирическая плотность распределения

Для интегральной функции распределения F(x) справедливо приближённое равенство: $F(x+\Delta)-F(x)\approx f(x)*\Delta x$,

где f(*x*) –дифференциальная функция распределения (функция плотности вероятности).

Поэтому естественно выборочным аналогом функции *f(x)* считать функцию:

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

F**(x+Δ)-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 $[a_i; a_i + h)$, где h – длина частичного интервала, выборочную функцию плотности *f(x)* можно задать соотношением :

$$f^*(x)\begin{cases} 0 \text{ при } x < a_1 \\ \frac{p_i^*}{h} \text{ при } a_1 \le x \le a_{i+1}, i = 1, 2, \dots, m \\ 0 \text{ при } x > a_{m+1} \end{cases},$$

Где a_{m+1} – конец последнего m – интервала.

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

Графическое изображение статистических данных.

Статистическое распределение изображается графически с помощью полигона и гистограммы.

Полигоном частот называют ломаную, отрезки которой соединяют точки с координатами $(x_i;n_i)$; полигоном **частностей-** с координатами $(x_i; p_i^*)$, где $p_i^* = \frac{n_i}{n}$, i = 1, m.

Полигон служит для изображения дискретного статистического ряда.

Полигон частостей является аналогом многоугольника распределения дискретной случайной величины в теории вероятностей.

Гистограммой частот (частостей) называют ступенчатую фигуру, состоящую из прямоугольников, основания которых расположены на оси *Ox* и длины их равны длинам частичных интервалов (h), а высоты равны отношению:

$\frac{n_i}{h}$ - для гистограмных частот; $\frac{n_i}{h * n}$ - для гистограммы частостей.

Гистограмма является графическим изображением интервального ряда. Площадь гистограммы частот равна *n,* а гистограммы частостей равна 1.

Можно построить полигон для интервального ряда, если преобразовать его в дискретный ряд. В этом случае интервалы заменяют их серединными значениями и ставят в соответствие интервальные частоты (частости).

Пример 1.

Дана выборка значений случайной величины X объёма 20:

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

18, 17, 15, 13, 17, 14, 14, 13, 14, 16

Требуется: -построить дискретный вариационный ряд

-найти размах варьирования R, моду, медиану

-построить полигон частей.

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 вариационный ряд.

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

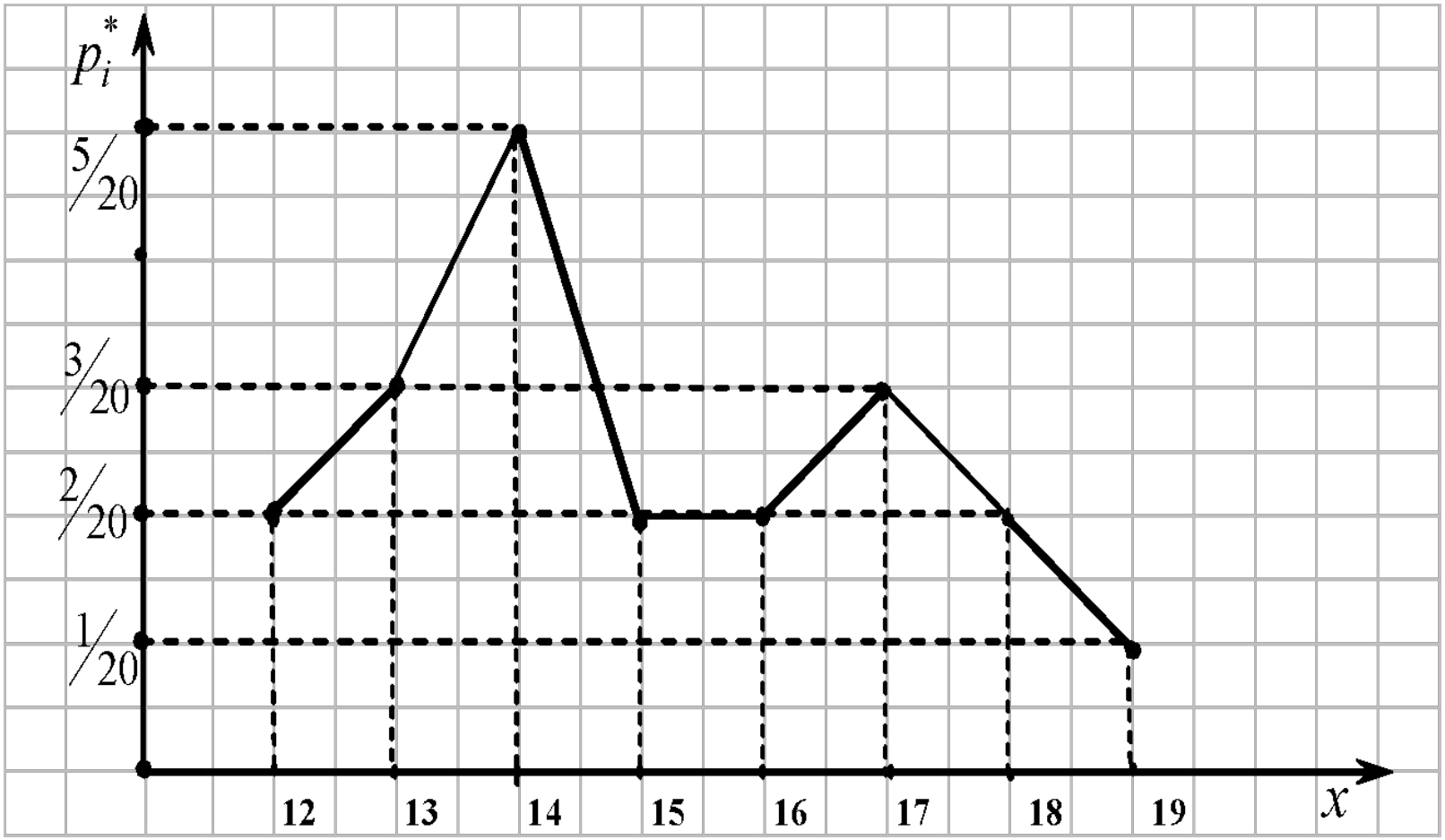
$$\sum_{i=1}^{8} n_i = 20$$

$$\sum_{i=1}^{8} p_i = 1$$

3) По результатам таблицы находим:

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

4) Строим полигон частостей.



Пример 2. Результаты измерений отклонений от нормы веса сердец кур-несушек дали численные значения (в мкм), приведённые в таблице.

-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

-построить гистограмму и полигон частостей.

1) Строим интервальный ряд.

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

Для определения длины интервала h используем формулу Стерджеса:

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

Число интервалов m ≈ 1+3,322* lg 50.

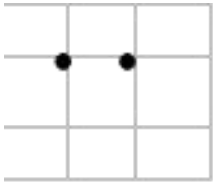
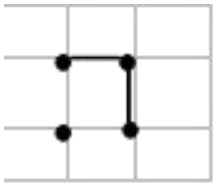
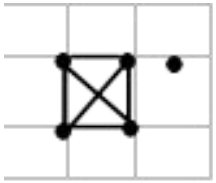
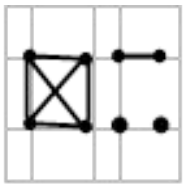
$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322 * \lg n} = \frac{1{,}74 - (-1{,}76)}{1 + 3{,}322 * \lg 50} \approx \frac{3{,}5}{1 + 3{,}322 * \lg 50} \approx \frac{3{,}5}{6{,}644} \approx 0{,}526$$

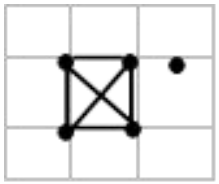
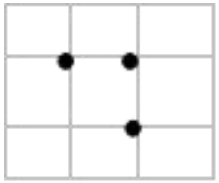
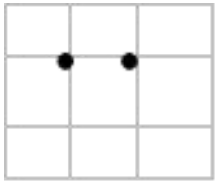
Примем h=0,6 , m = 7.

За нчало первого интервала примем величину:

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

Строим интервальный ряд:

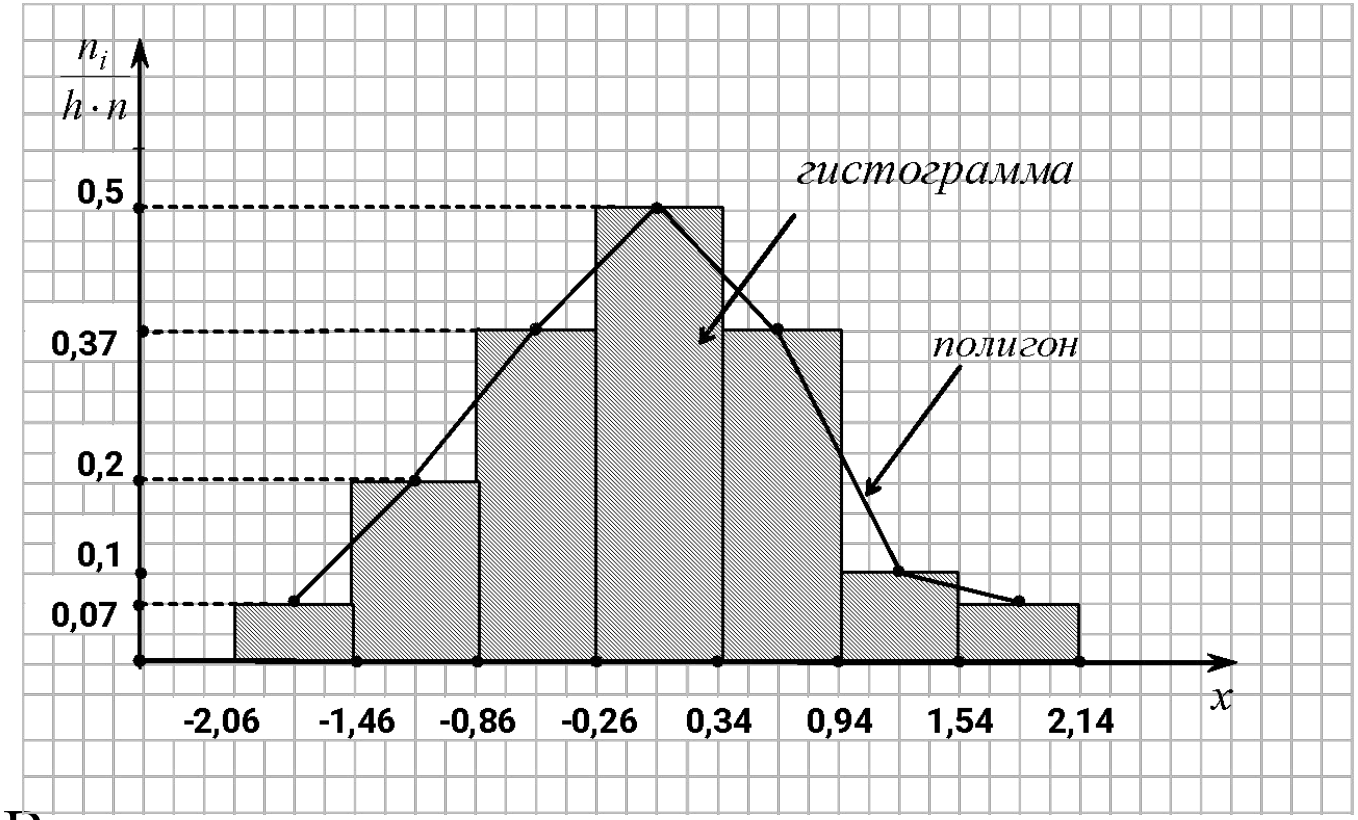
Интервалы	$[-2{,}06; -1{,}46)$	$[-1{,}46; -0{,}86)$	$[-1{,}86; -0{,}26)$	$[-0{,}26; 0{,}34)$
Подсчет частот				
Частоты n_i	2	6	11	15
Частости p_i	$\frac{2}{50}$	$\frac{6}{50}$	$\frac{11}{50}$	$\frac{15}{50}$

Интервалы	$[0{,}34; 0{,}94)$	$[0{,}94; 1{,}54)$	$[1{,}54; 2{,}14)$
Подсчет частот			
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.



Вершинами полигона являются середины верхних оснований прямоугольников гистограммы.

Убедимся, что площадь гистограммы равна 1.

$$S = h * (\frac{n_1 + n_2 + \cdots + n_m}{n * m})$$

S = 0,6*(0,07+0,2+0,37+0,5+0,37+0,1+0,07)= 0,6*1,68=1,008 ≈ 1.