Раздел: Теория вероятностей и математическая статистика

Тема: ***Математическая статистика: основные понятия, первичная обработка эмпирических данных***

Лектор Пахомова Е.Г.

2016 г.

Цель любой науки:

описание, объяснение и предсказание явлений действительности на основе установленных законов, что позволяет находить решение в типичных ситуациях.

Для обнаружения общей закономерности, которой подчиняется явление, необходимо многократно его наблюдать в одинаковых условиях.

Математическая статистика — раздел математики, в котором разрабатываются математические методы систематизации и обработки экспериментальных данных с целью изучения закономерностей массовых случайных явлений и использования их для научных и практических выводов.

Выделяют: а) описательную статистику;
б) теорию оценивания;
в) теорию проверки гипотез.

§11. Выборочный метод

Пусть для получения опытных данных необходимо провести обследование некоторых объектов.

Примеры.

1. Проверить качество выпускаемых некоторым заводом консервов.

2. Выяснить среднюю заработную плату по России.

Обычно исследуют не всю совокупность объектов, а отбирают из неё некоторое количество объектов и исследуют только их.

В этом и заключается ***выборочный метод***.

Генеральной совокупностью называют совокупность всех объектов, над которыми производят наблюдение.

Выборочной совокупностью (выборкой) называют часть отобранных из генеральной совокупности объектов.

Объёмом совокупности называют количество объектов в ней.

По выборке судят о генеральной совокупности.

Выборка должна правильно представлять генеральную совокупность, то есть быть ***репрезентативной***.

Это обеспечивается способом отбора и увеличением объёма выборки.

Способы отбора

1. Отбор, не требующий расчленения генеральной совокупности на части:

   а) простой случайный бесповторный отбор,

   б) простой случайный повторный отбор.

2. Отбор, при котором генеральная совокупность разбивается на части:

   а) типический,

   б) механический,

   в) серийный.

3. Комбинированный отбор.

Выборка, в которой меньше 30-ти элементов, называется ***малой***. В противном случае, выборка называется ***большой***.

Для выборок малого объема необходимо выбирать специально разработанные методы.

Выборочные данные делятся на: а) качественные;

б) количественные.

Качественные данные представляются (кодируются) определенным числом в соответствии с некоторым свойством.

Для работы с качественными данными используются специально разработанные методы, которые называются ***непараметрическими***.

Методы, разработанные для количественных данных (***параметрические методы***), не могут использоваться для качественных данных.

В дальнейшем будет предполагаться, что наблюдения будут представляться количественной информацией. При этом выделяют:

а) данные непрерывного типа – возможно появление любого значения из некоторого интервала;

б) данные дискретного типа – возможны лишь изолированные значения из некоторого интервала.

§12. Первичная обработка результатов наблюдений

Что такое наблюдаемые данные?
Большой массив беспорядочно расположенных чисел.

Для работы с данными удобно их группировать.

Пример.

0 1 2 2 1 2 0 0 0 0 – выборка

Объём выборки: $n = 10$

x_i	0	1	2	перечень значений
n_i	5	2	3	частота встречаемости
w_i	0.5	0.2	0.3	относительная частота

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3

Наблюдаемые значения x_i называют ***вариантами***.

Последовательность вариант, записанных в возрастающем порядке называют ***вариационным рядом***.

Частотой варианты называют число n_i, показывающее сколько раз встречается данная варианта.

Относительной частотой варианты называют отношение частоты к объёму выборки: $w_i = n_i / n$.

Статистическим распределением выборки (дискретным статистическим рядом) называется перечень вариант и соответствующих им частот или относительных частот.

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3

Замечания.

1. $\sum_i n_i = n$ сумма всех частот равна объёму выборки

2. $\sum_i w_i = 1$ сумма всех относительных частот равна 1

3. $p(X = x_i) \approx w_i$ относительная частота варианты даёт приближённое значение вероятности этой варианты

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3

Если наблюдаемые данные имеют непрерывный характер, то перечень вариант обычно очень велик и одинаковые значения вариант встречаются очень редко.

$\Rightarrow$ дискретный статистический ряд неудобен в использовании.

Тогда составляют ***интервальный статистический ряд***:

1) разбивают весь интервал, в который попадают варианты, на частичные интервалы;
2) в верхнюю строку записывают полученные интервалы;
3) в нижнюю строку записывают частоту попадания в соответствующий интервал.

27 3,5 21,1 0,8 12,3 18 11 3,4 1,2 5,2 22 17,2 18,1 11,1 0,7 7,9 19
3,2 4,9 25,4 6,1 21,6 22,3 3,4 18,4 3,4 23,2 13,1 6,5 2,4 18,4 14,1
2,1 24,8 17,4 15,1 4,8 19,8 10,4 16,1 3,7 29,4 3,1 28,7 16,4 22,2 1,7
12,4 17 15,3 3,3 14 16,8 10,1 2,4 20 14,1 19 19,8 5,4 2,5 4,1 24,4
0,4 24,7 1,3 13,7 0,1 28 24 17,1 15 3,1 19 0,4 23,1 6,7 4,6 14,8
20,7 16,2 9,4 21,3 13,4 16,1 15,7 11,3 5,1 1,9 2,8 17 2 20,8 3,4
16,7 9,3 15,2 8,7 10,7

1) разбивают весь интервал, в который попадают варианты, на частичные интервалы;

Интервал: числа от 0 до 30.

6 интервалов: [0, 5); [5, 10); [10, 15); [15, 20); [20, 25); [25, 30)

2) в верхнюю строку записывают полученные интервалы;

3) в нижнюю строку записывают частоту попадания в соответствующий интервал.

x_i	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	30	10	15	25	15	5
w_i	0.3	0.1	0.15	0.25	0.15	0.05

На сколько интервалов разбивать выборку?

$k = 1 + 3.332 \cdot \lg n$ или $k \le 5 \cdot \lg n$

n – объём выборки

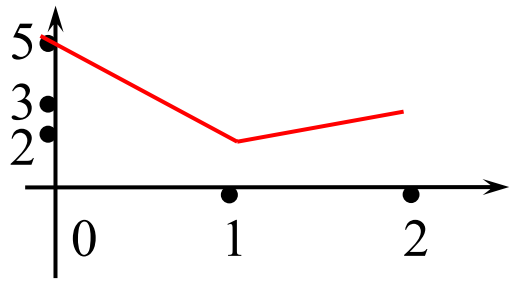
***Замечание**:* первая из этих формул носит название формула Старджеса

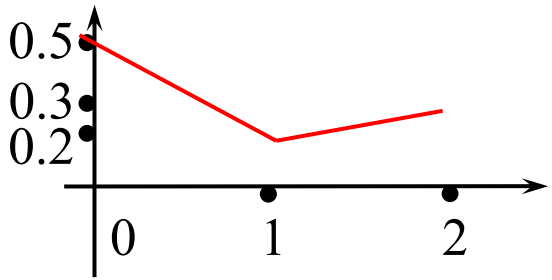
§13. Визуализация данных

Полигоном частот называют ломаную, отрезки которой последовательно соединяют точки (x_i, n_i).

Полигоном относительных частот называют ломаную, отрезки которой последовательно соединяют точки (x_i, w_i).

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3

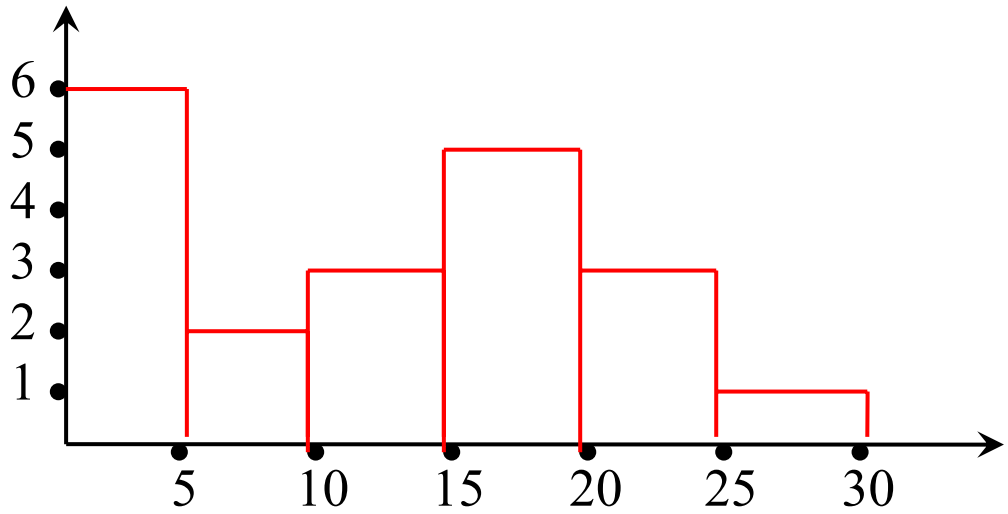




Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы, а высоты равны отношению частоты попадания в данный интервал к длине интервала.

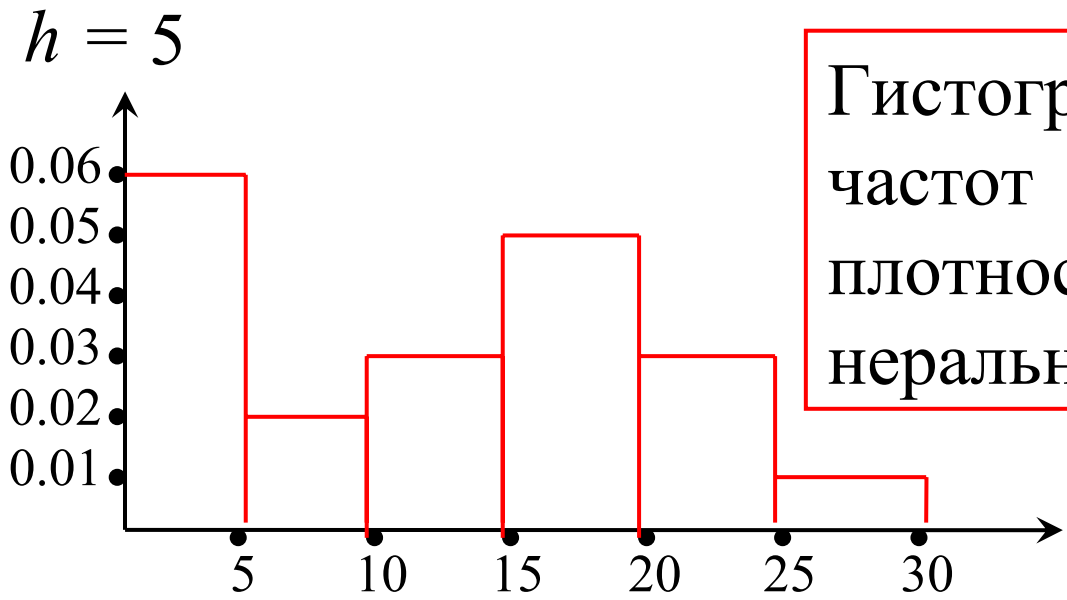
x_i	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	30	10	15	25	15	5
n_i / h	6	2	3	5	3	1

$h = 5$



Аналогично вводится понятие ***гистограммы относительных частот***.

x_i	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	30	10	15	25	15	5
w_i	0.3	0.1	0.15	0.25	0.15	0.05
w_i / h	0.06	0.02	0.03	0.05	0.03	0.01



Гистограмма относительных частот даёт представление о плотности распределения генеральной совокупности.

§14. Эмпирическая функция распределения

Функция распределения случайной величины X:

$$F(x) = p(X<x)$$

Теоретической функцией распределения называют функцию распределения генеральной совокупности.

Обозначим через n_x – частоту появления вариант, меньших x. Тогда n_x/n – относительная частота появления вариант, меньших x.

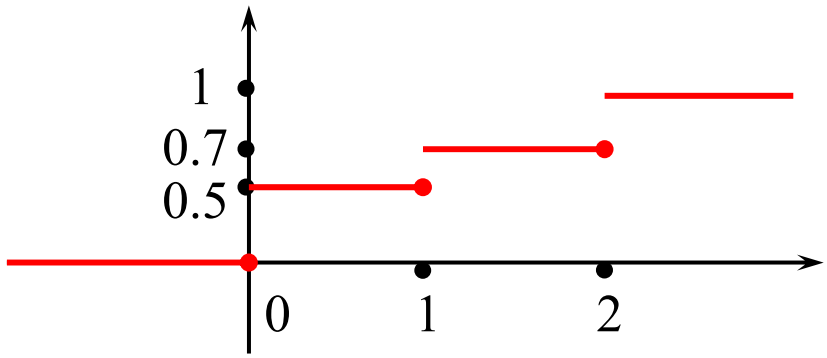
Эмпирической (выборочной) функцией распределения называют функцию

$$F^*(x) = n_x/n.$$

$F^*(x) = n_x / n.$

n_x – частота появления вариант, меньших x

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3



Объём выборки: $n = 10$

$x \le 0$ $\quad n_x = 0$

$0 < x \le 1$ $\quad n_x = 5$

$1 < x \le 2$ $\quad n_x = 7$

$x > 2$ $\quad n_x = 10$