Задание

1. Ознакомиться с теоретический материалом темы, составить план к лекции, законспектировать по составленному плану. Решить задания у лекции.

2. Фотоотчет и сообщение присылать на электронную почту

С уважением, Хвастова Светлана Ивановна

!!! Если возникнут вопросы обращаться по телефону 0721389311 (ватсап). Электронная почта: xvsviv@rambler.ru

Лекция «Основные понятия и задачи математической статистики»

1. Основные положения

Математическая статистика – это раздел математики, посвященный математическим методам систематизации, обработки и использования данных, полученных в результате экспериментов, опытов или наблюдений для научных и практических выводов, т.е. математическая статистика рассматривает методы сбора, анализа и обработки статистических данных.

Статистические данные – это данные, полученные в результате обследования большого числа объектов или явлений.

Введем основные понятия, полезные для математической обработки экспериментальных данных.

Экспериментальные данные в какой-либо области обычно представляют собой результаты измерения некоторых признаков, объектов, выбранных из большой совокупности объектов.

Выборка – это часть объектов исследования, определенным образом выбранная из более обширной совокупности.

Генеральная совокупность (популяция) – это исходная совокупность, из которой взята выборка.

Если рассматривать, студентов ВУЗА, как генеральную совокупность, то студенты, например, 1-го курса – выборка из более широкой генеральной совокупности.

Объектами исследования, составляющими генеральную совокупность, например, в спорте являются спортсмены, которые должны иметь хотя бы один общий признак (пол, возраст и т.д.).

Наличие общего признака является основой для образования статистической совокупности. Если статистическая совокупность получается в результате выборочного исследования, то она называется *выборочной совокупностью* или *выборкой*.

Числовые значения варьирующего (изменяющего) признака называют ***вариантами***.

Вариационный рядом выборки называется способ записи, при котором элементы выборки упорядочиваются по величине.

Вариационный ряд называется дискретным, если любые его варианты отличаются на постоянную величину и непрерывным (интервальным), если варианты могут отличаться один от другого на сколь угодно малую величину.

Медианой выборки называется элемент, находящийся в середине вариационного ряда.

Если в вариационном ряде четное количество элементов, то медиана вычисляется как среднее арифметическое двух элементов, стоящих в середине вариационного ряда.

Размахом выборки называется разность между максимальным и минимальным элементами выборки.

Если элемент x_i встречается в выборке n_i раз, то n_i называется ***частотой*** элемента x_i.

Статистическим рядом выборки называется таблица, в первой строке которой записаны все различные варианты выборки, а во второй – их частоты.

Модой выборки называется элемент, встречающийся чаще других.

Если все элементы выборки встречаются одинаковое количество раз, то моды нет. Если два различных элемента имеют максимальную частоту, то моды две, а выборка называется бимодальной.

Пример: Получены следующие результаты контрольной работы в классе N: 3, 5, 4, 3, 2, 5, 2, 4, 4, 4, 2, 5, 4, 5, 4. Охарактеризуем выборку.

Объем выборки: 15.

Размах выборки: 5 – 2 = 3.

Вариационный рад: 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5.

Мода: 4.

Медиана: 4.

Статистический ряд:

x_i	2		3	4	5
n_i	3		2	6	4

Размах, *дисперсию* и *стандартное отклонение* называют ***мерами изменчивости***. *Моду*, *среднее арифметическое* и *медиану* называют ***мерами центральной тенденции***.

По одним признакам элементы могут совпадать полностью (пол, возраст, вид спорта и т.д.), а по другим не совпадать (сила мышц, быстрота реалии, показатели систем дыхания, кровообращения и т.д.). Изменяющиеся признаки и являются предметом изучения в статистики. Они делятся на *качественные* и *количественные*.

Качественные признаки – это признаки, которыми объект обладает либо не обладает. Они не поддаются непосредственному измерению (национальность, квалификация, пол и т. п.) и представляют свойства элементов выборки.

Количественные признаки – это результаты измерений или подсчета. Они делятся на *дискретные* и *непрерывные*. Дискретные признаки могут принимать лишь отдельные значения из некоторого ряда чисел. Примером дискретных признаков могут быть число подтягиваний на перекладине, число попаданий в мишень при серии выстрелов и т. п. Непрерывные признаки могут

принимать любые значения в определённом интервале. Примером непрерывных признаков могут быть скорость движения, угол в суставе, время прохождения дистанции и т.п.

Не все признаки можно контролировать в процессе исследования.

К числу *контролируемых* факторов относятся пол, возраст, квалификация и т. п.

К числу *не контролируемых* факторов можно отнести влияние погодных условий, эмоциональное состояние испытуемых и т. п. Такие факторы называют *случайными.*

2. Способы представления экспериментальных данных

Как правило, необработанные экспериментальные данные представлены в виде неупорядоченного набора чисел. Необходима обработка.

1. Группировка. Группировка представляет собой процесс систематизации и выполняется различными методами. Основная цель группировки - представление данных в компактной форме, наглядной и удобной для анализа. Рассмотрим группировку на конкретном примере.

В таблице 1 приведены данные (бег 5 км, лыжи, девушки 1-го курса ФФК).

Таблица 1

25,7	28,2	27,2	25,5	27,2	27,3	26,4	27,1	28,3	27,4
27,0	26,7	30,0	29,6	27,8	28,4	29,3	27,5	27,3	27,0
28,1	29,5	27,1	26,7	29,4	26,1	27,2	25,6	28,1	27,5
27,1	27,3	28,2	25,3	27,5	29,3	29,1	28,1	27,1	26,0
26,4	28,4	27,4	26,4.	29,2	27,4	30,5	28,3	27,0	27,1

Объем выборки в данном случае равен 50 т. е. $n = 50$.

Группировка заключается в распределении вариант выборки по группам или интервалам группировки. Каждый интервал содержит определенное количество значений изучаемого признака. Число интервалов группировки ищется либо по Формуле Стерджеса: $к = 1+3.33*\lg(n)$, либо с помощью таблицы 2. Ширина каждого интервала определяется по формуле:

$$h = \frac{x_{max} - x_{min}}{k} \qquad (1)$$

где h - ширина интервала,

x_{max}, x_{min} - максимальная и минимальная варианты выборки.

Для нашего примера при n = 50 число интервалов группировки $к$ = 7.

$h = \frac{30,35}{7} = \frac{5,7}{7} \approx 0,8$. Округлим ширину интервала h в сторону увеличения и примем h = 0,8.

Найдём границы интервалов группировки по формулам:

$x_{н1} = x_{min} - \frac{h}{2}$; $x_{н2} = x_{н1}$+h; $x_{н2} = x_{в1}$; $x_{н3} = x_{в2}$ и т.д., где

$x_{н1}$- нижняя граница первого интервала.

$x_{н2}$ - нижняя граница второго интервала.

$x_{в1}$ - верхняя граница первого интервала и т. д.

Найдём нижние и верхние границы интервалов, причем верхнюю границу первого интервала будем записывать на 0,1 меньше нижней границы второго интервала и т.п.

$x_{н1} = 25,3 - \frac{0,8}{2} = 24,9$; $x_{н2} = 24,9 + 0,8 = 25,7$, тогда $x_{в1} = 25,8$.

После того как найдены границы всех интервалов, распределим по этим интервалам выборочные варианты, но для удобства последующей обработки вычислим *срединные значения интервалов* группировки x_{ci} по формуле $x_{ci}=x_{нi}+\frac{h}{2}$. Результаты внесем в таблицу 3.

Но прежде рассмотрим ряд определений.

Накопленная частота интервала – это число, полученное последовательным суммированием частот в направлении от 1 интервала и до того интервала включительно, для которого определяется накопленная частота.

Частостью называется отношение частоты к объему выборки и обозначается f_i , т.е. $f_i = \frac{n_i}{n}$.

Сумма всех частостей всегда равна 1

2. Графическое представление экспериментальных данных

Для наглядности результатов эксперимента используется их графическое представление: гистограмма, полигон частот и полигон накопленных частот.

Гистограмма – это примыкающие друг к другу прямоугольники. Основание каждого прямоугольника есть ширина группировки h_i. а высота - n_i. если все h_i одинаковые. На графике гистограммы основания прямоугольников откладываются по оси *Ox*, а высоты по оси *Oy*.

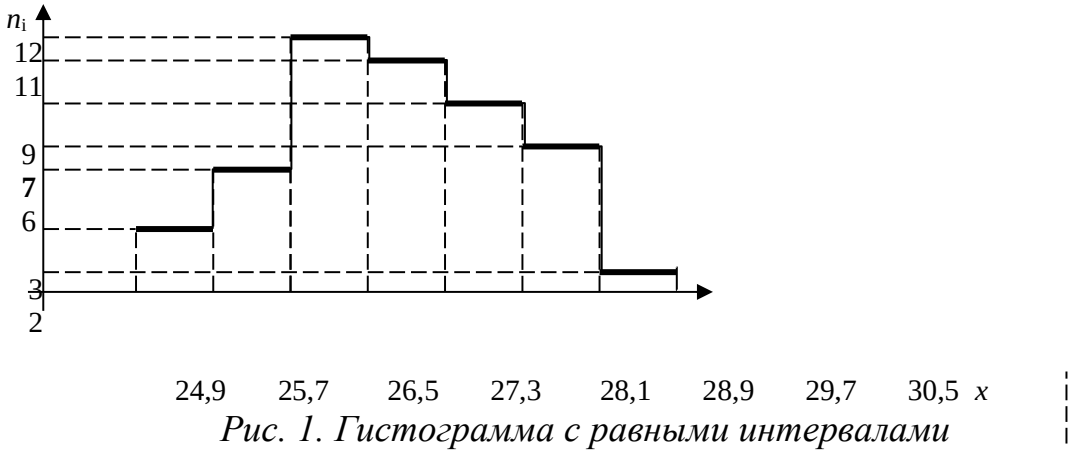


Рис. 1. Гистограмма с равными интервалами

Если h_i не одинаковые, то по оси *Oy* откладывается P_i, где $P_i = \frac{n_i}{h_i}$, т.е. высоты прямоугольников должны быть пропорциональны величинам P_i .

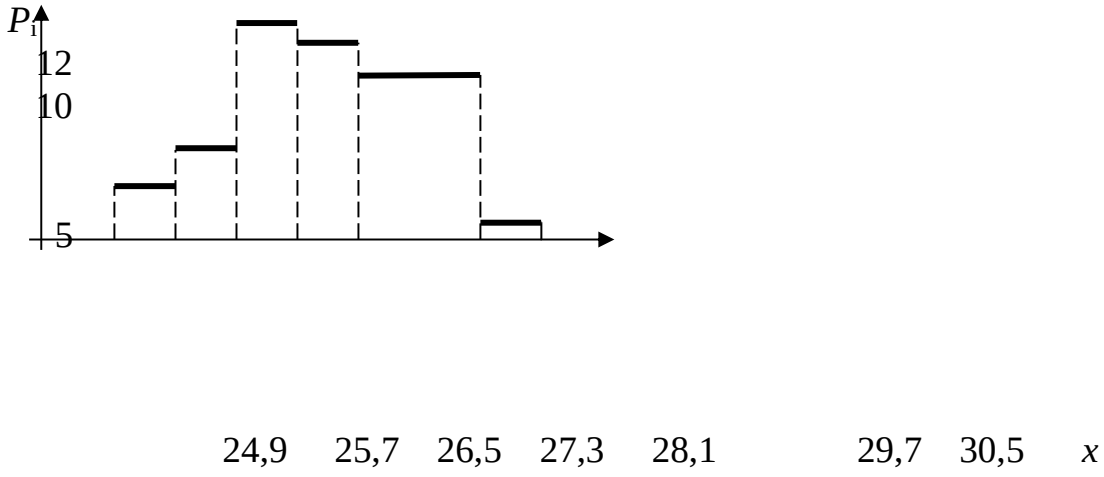


Рис. 2. Гистограмма с объединенными интервалами

На рисунке 1 гистограмма построена по данным таблицы 3 с равными интервалами. На рисунке 2 интервалы взяты не одинаковые (два объединили), поэтому по оси *Oy* откладывали P_i.

Полигон частот – это ломаная линия, соединяющая точки, соответствующие срединным значениям интервалов группировки.

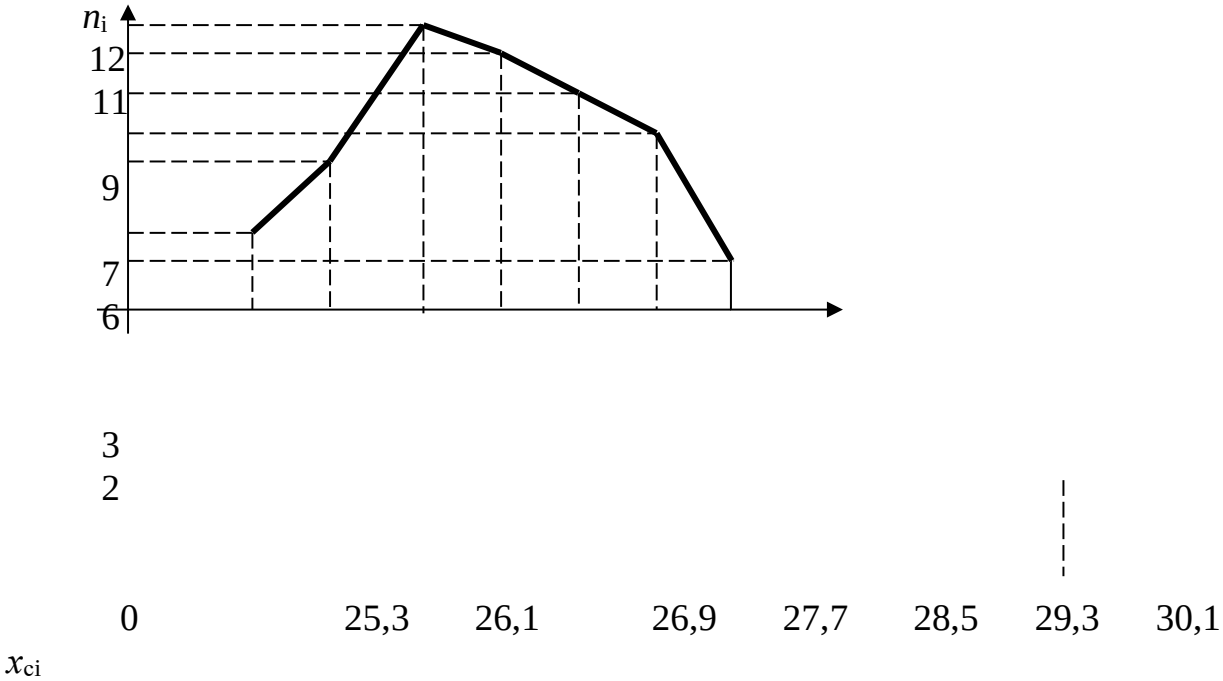
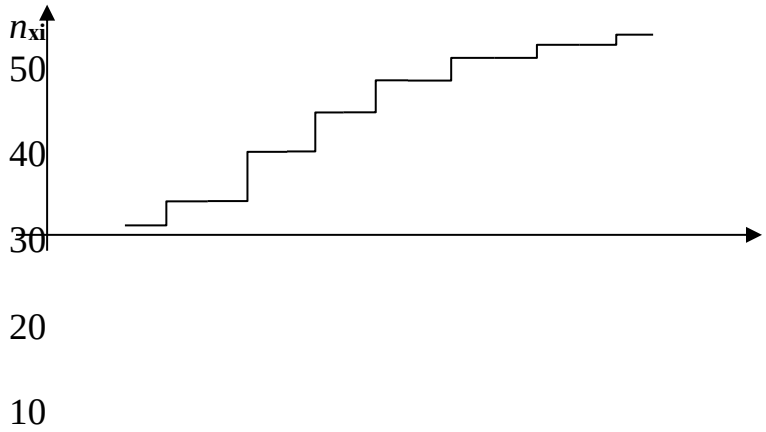


Рис. 3. Полигон частот

На оси *Ox* отложены срединные значения, по оси *Oy* – частоты. Из сравнения двух способов следует, что для получения полигона частот из построенной гистограммы нужно соединить середины сторон отрезками. На рис 1 график полигона частот построен пунктиром.

Полигон накопленных частот получается при соединении отрезком прямых точек. координаты которых соответствуют верхним границам интервалов группировки и накопленным частотам. По оси ОХ границы интервалов, по оси ОУ - накопленные частоты.



0 24,9 25,7 26,5 27,3 28,1 28,9 29,7 30,5 x

Рис.4. Полигон накопленных частот

Получили более плавную линию по сравнению с полигоном частот.

Числовые характеристики выборки дают количественные представления об эмпирических данных и позволяют сравнивать их между собой. Остановимся на некоторых из них подробнее. К характеристикам положения относятся: среднее арифметическое, медиана, мода. Они определяют положение центра эмпирического распределения.

Среднее арифметическое – одна из основных характеристик выборки и представляет собой такое значение признака, сумма отклонений от которого выборочных значений признака равна нулю. Геометрически среднее арифметическое можно определить как точку на оси *Ox*, которая является абсциссой центра масс гистограммы и обозначается $\bar{x}$.

Для *несгруппированных* данных среднее арифметическое определяется по формуле: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, (3.1), где n – объём выборки, x_i – варианты выборки.

<u>Пример</u>.

Для оценки уровня физической подготовленности студентов 1-го курса педагогического вуза были выбраны прыжки в длину с места. Результаты контрольной группы студентов в количестве 15 человек оказались следующими (в см): 213, 224, 225, 210, 226, 230, 201, 224, 230, 227, 228, 252, 238, 232, 246. Определить средний результат.

Для *сгруппированных данных* среднее арифметическое определяется по следующей формуле: $\bar{x}=\frac{1}{n}\sum_{i=1}^{k}n_i x_i$, (3.2), где n – объём выборки, n_i – частоты интервалов, x_i – срединные значения интервалов.

<u>Пример</u>:

Вычислить среднее арифметическое результатов в беге на 5 км (лыжи), сгруппированных в таблице. Для наглядности промежуточные результаты расчетов занесем в таблицу.

n	x_{ci}	n_i	$n_i x_i$
1	25,3	3	75,9
2	26,1	6	156,6
3	26,9	12	322,8
4	27,7	11	304,7
5	28,5	9	252,5
6	29,3	7	203,1
7	30,1	2	60,2
сумма			1375,8

$$\bar{x}=\frac{1}{n}\cdot 1375{,}8=27{,}5 \quad (мин)$$

Медианой называется такое значение признака X, когда одна половина значений экспериментальных данных меньше её, а вторая половина больше. Если объем выборки невелик, то для вычисления медианы выборку

ранжируют, т. е. располагают в порядке возрастание или убывания и вычисляют порядковый номер R (ранг) медианы по формуле: $R_{Me} = \frac{n+1}{2}$.

Пусть, например, имеется ранжированная выборка n=11:

13, 15, 16. 19, 19, 22, 23, 25, 27, 27, 29. Тогда $R_{Me} = \frac{11+1}{2} = 6$;

Me = 22 (6-ой член ряда).

При n = 8: 12, 15, 16, 18, 22, 24, 25, 27; $R_{Me} = \frac{8+1}{2} = 4,5$.

Медианой в этом случае может быть любое число между 18 и 22 (четвертым и пятым членами ряда). Me = $\frac{18+22}{2} = 20$.