**Тема «Элементы математической статистики»**

**Вопросы: 1. Введение**

**2. Генеральная совокупность и выборка**

**3. Вариационный и статистический ряды**

**4. Интервальный статистический ряд**

**5. Статистические характеристики выборки**

1. **Введение**

Математическая статистика — раздел математики, разрабатывающий методы регистрации, описания и анализа данных наблюдений и экспериментов с целью построения вероятностно-статистических моделей случайных явлений.

Целью математической статистики является описание, объяснение и предсказание явлений действительности на основе установленных законов.

Предметом математической статистики является изучение случайных величин (или случайных событий) по результатам наблюдений (статистическим данным).

Задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

Рассмотрим схему исследований при решении задачи математической статистики. Эти исследования делятся на два этапа.

1. На первом этапе, который называют описательной статистикой (descriptive statistics), путем наблюдений и экспериментов собираются, регистрируются статистические данные, затем они упорядочиваются, представляются в компактной, наглядной или функциональной форме. Вычисляются различные средние величины, характеризующие статистические данные.
2. На втором этапе на основе вычислений на предшествующем этапе необходимо получить достаточно обоснованные выводы о свойствах исследуемого случайного явления, используя методы оценивания и проверки гипотез.

Современная математическая статистика разрабатывает способы определения числа необходимых наблюдений, испытаний до начала исследования (планирование эксперимента), в ходе исследования (последовательный анализ) позволяет принимать решения в условиях неопределенности.

*Краткая историческая справка.* Слово «статистика» происходит от латинского слова «status» (состояние, государство) означает определенное положение вещей. Термин «статистика» впервые ввел немецкий ученый Г. Ахенвалль в 1749 г., в своей книге о «государствоведении». Несмотря на это, статистический учёт вёлся намного раньше: проводились переписи населения в Древнем Китае, осуществлялось сравнение военного потенциала государств, вёлся учёт имущества граждан в Древнем Риме и тому подобного.

Математическая статистика возникла в XVII в. и развивалась вместе с теорией вероятностей. Основоположниками науки являются Я. Бернулли, К.Гаусс, П. Лаплас. В XIX в. большой вклад внесли российские математики П. Чебышев, А. Марков, А. Ляпунов. В XX в. важные результаты были получены советскими математиками В.И. Романовским, Е.Е. Слуцким, А.Н. Колмогоровым, Н.В. Смирновым, английскими учеными Э. Пирсоном, У. Го́ссе (Стьюдентом), Р. Фишером, Г. Крамером, американскими учеными Ю. Нейманым, А. Вальдом, Р. Мизесом и другими учеными.

**2. Генеральная совокупность и выборка**

Рассмотрим абстрактный эксперимент, в результате его проведения мы наблюдаем или измеряем значение ***х*** изучаемой случайной величины***Х*** (это может быть в действительности величина инфляции, параметр детали при массовом производстве, цена на жилье в отдельных районах столичных городах, любой общий количественный признак определенного множества объектов).

***Определение 1***. *Генеральной совокупностью* называется множество возможных значений изучаемой случайной величины ***Х*** с законом распределения ***F(X)***. Возможные значения генеральной совокупности ***Х*** называются ее *элементами*. Закон распределения ***F(X)*** называется *генеральным законом* распределения, а числовые характеристики ***Х*** – *генеральными числовыми характеристиками.*

Генеральная совокупность может быть конечной (множество значений случайной величины ***Х*** конечно) или бесконечной. Например, 1) ***Х*** – число детей, родившихся в городе за определенный промежуток времени. Генеральная совокупность это множество неотрицательных чисел {0,1,2,…} с некоторым законом распределения. 2) ***Х*** – величина отклонения детали от заданного размера при массовом производстве. Генеральная совокупность это множество действительных чисел с некоторым законом распределения.

Иногда из всей генеральной совокупности случайно отбирают ограниченное число объектов и подвергают их изучению, по свойствам которой судят о свойствах генеральной совокупности.

***Определение 2*.** *Выборочной совокупностью* или *выборкой*называется совокупность случайно отобранных элементов их генеральной совокупности. *Объемом выборки* называется число ее элементов.

Выборку нельзя составить как попало. Выборка должна быть репрезентативной (представительной) и однородной. Репрезентативность обеспечивается простым случайным выбором:

1. Выбор является случайным.
2. Каждый элемент генеральной совокупности может быть выбран.
3. Каждый элемент выбирается независимо от остальных.
4. Все элементы выборки получаются в равных условиях.

Однородность означает, что условия проведения экспериментов для получения выборки не должны меняться. Но на практике простой случайный выбор не всегда осуществим (он является эталонным), применяются различные виды выбора: *механический выбор* (измерения проводят через равные промежутки времени, контролируется каждая m деталь, выбирается каждый s человек по списку); *серийный выбор* (контролируется не одна таблетка, а вся упаковка, не один человек группы, а вся группа); *типический выбор* (урожайность участка, социологический опрос, зарплата в отрасли); *выбор с помощью случайных независимых измерений* (температура среды, загрязненность воды, величина тока) и другие. Все виды выборов могут комбинироваться между собой. В математической статистике применяется только простой случайный выбор.

После того как сделана выборка, все ее элементы обследуют по отношению к генеральной совокупности и в результате получают наблюдаемые данные. Далее они упорядочиваются, представляются в компактной, наглядной или функциональной форме. Вычисляются различные средние величины, характеризующие статистические данные.

**3. Вариационный и статистические ряды**

Обычно выборка представляет собой множество чисел, расположенных в беспорядке. Для дальнейшего изучения выборку подвергают обработке.

***Определение 3*.** Наблюдаемые значения выборки называются *вариантами*. Последовательность всех вариант, записанных в возрастающем порядке, называется *вариационным рядом*.

***Пример 1.*** Взята выборка наименьших цен в тысячах рублях за 1 м2 на новое жилье в городах России на май 2016 г. (<http://www.rosrealt.ru/>):

Архангельск – 59, Барнаул – 45, Белгород – 56.5, Владивосток – 94, Волгоград – 48, Воронеж -48, Екатеренбург – 70, Казань – 64, Калининград – 60.5, Киров -47, Магнитогорск – 31, Майкоп – 36.5, Москва – 211.5, Рязань – 47, Самара – 65, Сочи – 83.5, Томск – 52, Тула – 52, Ярославль – 47.

Построим вариационный ряд.

Элемент выборки 211.5 является аномальным, что объясняется исключительным положением города. Этот элемент следует исключить из выборки. Тогда вариационный ряд примет вид:

31, 36.5, 45, 47, 48, 52, 56.5, 59, 60.5, 64, 65, 70, 83.5, 94.

Пусть из генеральной совокупности извлечена выборка: варианта ***х1***наблюдалась ***n1*** раз, варианта ***х2*** – ***n2*** раз, …, варианта ***хk*** – ***nk*** раз и *n1* + +*n2+…+ nk=n – объем выборки.* Числа вариантов *n1*, *n2*,…, *nk* называются *частотами*. Отношения частот к объему выборки , где *i=1,2,…,k*, называются относительными частотами.

***Определение 4***. *Статистическим рядом*называется вариационный ряд с указанием соответствующих частот или относительных частот.

В общем случае статистический ряд представляют в виде таблиц.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Варианты* ***xi*** | ***х1*** | ***х2*** | … | ***хk*** |
| *Частоты* ***ni*** | ***n1*** | ***n2*** | … | ***nk*** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Варианты* ***xi*** | ***х1*** | ***х2*** | … | ***хk*** |
| *Относительные частоты* |  |  | … |  |

***Пример 2.*** Преобразуем выборку из примера 1 в статистический ряд частот.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***xi*** | 31 | *36.5* | 45 | *47* | *48* | *52* | *56.5* | *59* | *60.5* | *64* | *65* | *70* | *83.5* | *94* |
| ***ni*** | 1 | 1 | 1 | 3 | *2* | *2* | *1* | *1* | *1* | *1* | *1* | *1* | *1* | *1* |

Объем выборки *n=18****.***

Статистический ряд относительных частот.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***xi*** | 31 | *36.5* | 45 | *47* | *48* | *52* | *56.5* | *59* | *60.5* | *64* | *65* | *70* | *83.5* | *94* |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Статистический ряд можно изобразить графически в виде *полигона частот* или *полигона относительных частот*, что позволяет получить наглядное представление о закономерности варьирования наблюдаемых значений случайной величины. В прямоугольной системе координат наносят точки с координатами (***xi***, ***ni***) или (***xi***, ), полученные точки соединяют отрезками, полученную ломанную называют полигоном.

Для примера 2 полигон частот можно представить в виде рисунок 1.

**Рисунок 1.**

Статистический ряд графически можно изобразить в виде *кумулятивной кривой* (кривой сумм — кумуляты). При построении кумуляты дискретного вариационного ряда на оси абсцисс откладывают варианты ***xi***, а по оси ординат соответствующие им накопленные частоты . Соединяя точки с координатами (***xi***) отрезками прямых, получаем ломаную (кривую), которую называют кумулятой. Для получения накопительных частот и дальнейшего построения точек (***xi***) составляется расчетная таблица.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Варианты* ***xi*** | ***х1*** | ***х2*** | … | ***хk*** |
| *Относительные частоты* |  |  | *…* |  |
| *Накопительные относительные частоты*  ***=*** |  |  | *…* |  |

График кумуляты дает представление о графике функции распределения ***F(X)*** генеральной совокупности.

***Пример 3.*** Для статистического ряда примера 2 составим расчетную таблицу для накопительных частот и построим кумуляту (Рисунок 2).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***xi*** | 31 | *36.5* | 45 | *47* | *48* | *52* | *56.5* | *59* | *60.5* | *64* | *65* | *70* | *83.5* | *94* |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Рисунок 2.**

**4. Интервальный статистический ряд**

Если выборка получена из непрерывной генеральной совокупности, объем наблюдаемых значений случайной величины большой, то вариационный и статистические ряды будут трудно обозримыми множествами и строят интервальный статистический ряд. Для построения такого ряда весь интервал варьирования наблюдаемых значений случайной величины разбивают на ряд частичных интервалов и подсчитывают частоту попадания значений величины в каждый частичный интервал.

Пусть , , ..., — результаты независимых наблюдений над генеральной совокупностью *Х*. Если количество наблюдений *n* достаточно большое (**), то результаты наблюдений сводят в интервальный статистический ряд, который формируется следующим образом.

Вычисляют размах варьирования *R* признака *Х*, как разность между наибольшим  и наименьшим  значениями признака, то есть

. Размах *R* варьирования признака *Х* делится на ***k*** разных частей и таким образом определяется число интервалов в таблице. Величину *k* выбирают, пользуясь следующими правилами:

1), (30≤ *n* ≤ 1000; при *n*=40 *k* =6; при *n*=100 *k* =8; при *n*=200 *k* =10, при *n*=400 *k* =12; при *n*=1000 *k* =17);

2)*k =* 1 +3,3 *lg n* (формула Старджесса).

3) 6 ≤ *k* ≤10.

Длина ***h*** каждого частичного интервала определяется по формуле . Величину *h* обычно округляют до некоторого значения *d*. Например, если результаты  признака *Х* — целые числа, то *h* округляют до целого значения, если  содержат десятичные знаки, то *h* округляют до значения *d*, содержащего такое же число десятичных знаков. Затем подсчитывается частота *ni*, с которой попадают значения  признака *Х* в *i*-ый интервал. Значение , которое попадает на границу интервала относятся к левому. За начало  первого интервала рекомендуется брать величину

,

конец .

Промежуточные интервалы получают прибавляя к концу предыдущего интервала длину частичного интервала *h*:

, , …, .

Сформированный интервальный вариационный ряд записывают в виде таблицы

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
| *Варианты-интервалы* |  |  | ... |  |
| *Частоты* ***ni*** | ***n1*** | ***n2*** | ... | ***nk*** |

Интервальный вариационный ряд изображают геометрически в виде гистограммы частот ***ni***  или гистограммы относительных частот .

*Гистограммой* называется ступенчатая фигура, для построения которой по оси *Ox* откладывают отрезки, изображающие частичные интервалы ( варьирования признака *Х*, и на этих отрезках, как на основаниях, строят прямоугольники с высотами, равными частотам или относительным частотам соответствующих интервалов.

Построение интервального статистического ряда рассмотрим на примере.

***Пример 4***. В результате независимых измерений получены данные:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2,1 | 2,3 | 4,2 | 3,7 | 5,5 | 5,3 | 3,4 | 4,7 | 4,4 | 2,6 |
| 4,3 | 4,3 | 5,6 | 4,5 | 4,8 | 5,2 | 4,8 | 4,3 | 3,4 | 4,3 |
| 4,5 | 2,2 | 3,4 | 4,5 | 4,5 | 3,4 | 3,6 | 4,1 | 3,2 | 2,8 |
| 4,3 | 3,5 | 5,3 | 4,6 | 3,9 | 3,5 | 5,7 | 5,1 | 4,2 | 5,8 |
| 2,7 | 4,2 | 4,8 | 3,6 | 3,8 | 5,9 | 3,7 | 2,4 | 4,1 | 5,1 |

1) Объем выборки *n*=50. Для построения интервального ряда возьмем *k*=6.

2) Просматривая приведенные результаты наблюдений, находим наименьшее значение выборки , наибольшее значение выборки . Размах варьирования .

3) Длина частичного интервала , *h* округлим до *d*=0,7.

4) Вычислим = 2,1-0,5∙0,63≈1,8,

получим интервалы

[1,8; 2,5), [2,5; 3,2), [3,2; 3,9), [3,9; 4,6), [4,6; 5,3), [5,3; 6,0].

5)Просматривая результаты наблюдений, определим количество значений признака в каждом полученном интервале. Это можно выполнить в виде таблицы.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№ п/п**  ***i*** | **Интервалы** | **Рабочее поле** | **Частота *ni*** | **Относительная частота** |
| 1 | [1,8; 2,5) | 1111 | 4 | 0,08 |
| 2 | [2,5; 3,2) | 111 | 3 | 0,06 |
| 3 | [3,2; 3,9) | ~~1111~~ ~~1111~~ 11 | 12 | 0,24 |
| 4 | [3,9; 4,6) | ~~1111~~ ~~1111~~ ~~1111~~ 1 | 16 | 0,32 |
| 5 | [4,6; 5,3) | ~~1111~~ 111 | 8 | 0,16 |
| 6 | [5,3; 6,0] | ~~1111~~ 11 | 7 | 0,14 |
| **Сумма** |  |  | **50** | **1** |

При вычислении относительных частот округление результатов следует проводить таким образом, чтобы общая сумма относительных частот была равна 1.

Построим гистограмму частот интервального статистического ряда.

**5. Статистические характеристики выборки**

Для дальнейшего изучения изменения значений случайной величины служат числовые характеристики выборки. Эти характеристики вычисляются по статистическим данным, т.е. данным, полученным в результате наблюдений, поэтому их называют статистическими.

Среди статистических характеристик выделяют характеристики положения выборки (медиана, мода, средняя величина), характеристики рассеяния элементов выборки относительно средних величин (дисперсия, среднее квадратическое отклонение).

Пусть выборка объема *n* представлена в виде статического ряда

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Варианты* ***xi*** | ***х1*** | ***х2*** | … | ***хk*** |
| *Частоты* ***ni*** | ***n1*** | ***n2*** | … | ***nk*** |

***Определение 5*.** *Модой* *Мо* называют варианту, которая имеет наибольшую частоту.

Например, для данного статистического ряда *Мо*=14

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *xi* | 4 | 8 | 14 | 19 |
| *ni* | 3 | 4 | 6 | 5 |

***Определение 6*.** *Медианой* *Me* называют варианту, которая делит статистический ряд на равные части.

При нечетном объеме выборки  (нечетном числе столбцов в дискретном статистическом ряде) медиана равна серединному члену статистического ряда. Например, для статистического ряда

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *xi* | 3 | 5 | 8 | 12 | 15 |
| *пi* | 6 | 2 | 4 | 5 | 8 |

*Me*=8.

При четном объеме выборки  (четном числе столбцов в дискретном статистическом ряде) медиана находится по формуле

.

Здесь — варианта, которая находится слева от середины статистического ряда, а  — слева от нее. Например, для выборки

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *xi* | 2 | 5 | 7 | 10 | 12 | 14 |
| *ni* | 3 | 4 | 8 | 2 | 3 | 6 |

.

*Выборочное среднее* .(средний показатель выборки).

*Выборочная дисперсия* или

.

При малом объеме выборки (*п≤30*) пользуются исправленной выборочной дисперсией .

*Выборочное среднее квадратическое отклонение* или (степень рассеяния значений изучаемого признака относительно средней величины).

Для интервального статистического ряда за *xi* принимают середины каждого интервала (, а за *ni*  - соответствующую интервальную частоту. Все статистические характеристики статистического ряда вычисляются по выше приведенным формулам.

Для вычисления статистических характеристик выборки можно использовать готовые компьютерные программы (например, [Microsoft Office](https://ru.wikipedia.org/wiki/Microsoft_Office) Excel).

***Пример 5***. Определим статистические характеристики для выборки из примера 4.

Представим интервальный статистический ряд в виде дискретного ряда, заменив каждый интервал на соответствующую середину.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п *i* | Интервалы | Середины интервалов  *xi* | Частота *ni* | Относительная частота |
| 1 | [1,8; 2,5) | 2,15 | 4 | 0,08 |
| 2 | [2,5; 3,2) | 2,85 | 3 | 0,06 |
| 3 | [3,2; 3,9) | 3,55 | 12 | 0,24 |
| 4 | [3,9; 4,6) | 4,25 | 16 | 0,32 |
| 5 | [4,6; 5,3) | 4,95 | 8 | 0,16 |
| 6 | [5,3; 6,0] | 5,65 | 7 | 0,14 |
| Сумма |  |  | 50 | 1 |

Вычисления можно оформить в виде таблицы.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *i* | *xi* | *ni* | *xi ni* | *xi ni xi =* |
| 1 | 2,15 | 4 | 8,6 | 18,49 |
| 2 | 2,85 | 3 | 8,55 | 24,3675 |
| 3 | 3,55 | 12 | 42,6 | 151,23 |
| 4 | 4,25 | 16 | 68 | 289 |
| 5 | 4,95 | 8 | 39,6 | 196,02 |
| 6 | 5,65 | 7 | 39,55 | 223,4575 |
| Сумма |  | **50** | **206,9** | **902,565** |

Вычислим статистические характеристики для выборки, используя полученные значения:

мода *Мо*=4,25 (т.к. наибольшая частота *n4*=16),

медиана ,

выборочное среднее ,

выборочная дисперсия

среднее квадратическое отклонение .

**Контрольные вопросы и задания**

1. Что такое генеральной совокупность, выборка?
2. Сформулируйте определение простого случайного выбора.
3. Дайте определение вариационного ряда.
4. Сформулируйте алгоритм построения статистического ряда.
5. Расскажите о графическом изображении статистического и интервального статистических рядов.
6. Дайте определение кумуляты и расскажите о ее назначении.
7. Дайте определение крайних элементов вариационного ряда, размаха варьирования.
8. По каким формулам находятся выборочные средние статистического распределения?
9. Запишите формулы для вычисления дисперсии для выборки.
10. Запишите формулы для вычисления исправленной дисперсии.
11. Что называется модой, медианой вариационного ряда?
12. Расскажите о нахождении медианы при различном объеме выборки.
13. При изучении некоторой дискретной случайной величины в результате 40 независимых наблюдений получена выборка 10, 13, 10, 9, 9, 12, 12, 6, 7, 9, 7, 8, 8, 9, 13, 14, 9, 11, 9, 8, 10, 10, 11, 11, 11, 12, 8, 7, 9, 10, 14, 13, 8, 8, 9, 10, 11, 11, 12, 12. Составьте статистический ряд, постройте полигон и кумуляту, вычислите статистические характеристики.
14. Имеются следующие данные о размерах основных фондов (в млн. руб.) 30 предприятий: 4,2; 2,4; 4,9; 6,7; 4,5; 2,7; 3,9; 2,1; 5,8; 4,0; 2,8; 7,3; 4,4; 6,6; 2,0; 6,2; 7,0; 8,1; 0,7; 6,8; 9,4; 7,6; 6,3; 8,8; 6,5; 1,4; 4,6; 2,0; 7,2; 9,1. Составьте интервальный статистический ряд, постройте полигон и кумуляту, вычислите статистические характеристики.