1. Введение[1]

1.1 Предмет математической статистики

Математическая статистика — это раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей.

Статистические данные представляют собой данные, полученные в результате обследования большого числа объектов или явлений (то есть, математическая статистика имеет дело с массовыми явлениями).

Методы анализа массовых явлений — предмет многих научных дисциплин; но только в том случае, когда для анализа привлекаются формальные (абстрактные) математические модели, эти методы становятся статистическими.

Математическая статистика подразделяется на две обширные области:

описательная статистика	аналитическая статистика (теория статистических выводов)
методы описания статистических данных, представления их в форме таблиц, распределений и пр.	обработка данных, полученных в ходе эксперимента, и формулировка выводов, имеющих прикладное значение для конкретной области человеческой деятельности. Теория статистических выводов тесно связана с другой математической наукой — теорией вероятностей и базируется на ее математическом аппарате

Трудно найти современную область научных исследований, где бы ни использовались методы математической статистики. В последнее время они нашли широкое применение в медицине, биологии, социологии, т. е. в областях, сравнительно недавно считавшихся далекими от математики.

1.2. Понятие о выборочном методе

Сплошное наблюдение, т.е. изучение всех членов совокупности, сначала кажется единственно возможным способом получения о ней достаточно точной информации. На самом деле это не всегда так. Рассмотрим некоторые примеры.

Пусть на заводе за день изготовляется большая партия лампочек. Не контролировать срок их службы, конечно, нельзя. Однако если это будет сделано в стенах завода в отношении каждой лампочки, то, получив полную картину о долговечности лампочек, мы их все выведем из строя и ни одна не дойдет до потребителя. С такой «проверкой», безусловно, нельзя согласиться. В аналогичных условиях находятся предприятия, производящие консервы, ткани, искусственные волокна, строительные материалы и т. д.

Сплошное наблюдение нецелесообразно не только в случаях, когда оно приводит к уничтожению всех подлежащих рассмотрению объектов. Например, при составлении баланса денежных доходов и расходов населения нашей страны, при планировании денежного обращения, розничного товарооборота, транспортных тарифов, при проведении мероприятий по повышению материального и культурного уровня жизни народа необходимы данные о бюджетах семей трудящихся. Сбор этих данных осуществляется статистическими органами. Один работник-статистик в состоянии вести ежедневные записи доходов, расходов, потребления и т. д. не более чем в 20—25 семьях одновременно. Для обследования только бюджетов трудящихся нашей страны понадобилось бы несколько миллионов работников. Кроме того, для обработки собранных данных необходимо большое число специалистов. Выделить такое количество рабочей силы практически невозможно и нецелесообразно.

Средством для получения необходимой информации в подобных случаях остается **несплошное** наблюдение. Широкое применение находит *выборочный метод исследования.*

Суть этого метода: если по результатам изучения сравнительно небольшой ее части можно получить с достаточной для практики достоверностью необходимую информацию о всей совокупности, то нет необходимости в сплошном наблюдении.

Часть объектов исследования, определенным образом избранная из более обширной совокупности, называется выборкой, а исходная совокупность, из которой взята выборка, — генеральной (основной) совокупностью.

Вся подлежащая изучению совокупность объектов называется *генеральной совокупностью.*

Та часть объектов, которая попала на исследование, называется *выборочной совокупностью* (или просто *выборкой).*

[1] Основой для данного текста является учебник “Мат. статистика для институтов физкультуры”

Важнейшая характеристика выборки — объем выборки, т. е. число элементов в ней.

Число элементов в генеральной совокупности называется *объемом генеральной совокупности (*обозначается *N).* Относительно *N*, как правило, делается предположение, что он бесконечно велик, т. е. выборка получается из бесконечной генеральной совокупности.

Число элементов в выборке называется *объемом выборки (*обозначается *n).*

1.3. Способы образования выборочной совокупности

Чтобы иметь право судить о генеральной совокупности по выборке, последняя должна быть образована случайно. Этого можно достичь различными способами.

Существуют различные виды выборок:

1. собственно-случайная;
2. механическая;
3. типическая;
4. серийная…

1. Члены генеральной совокупности можно предварительно занумеровать, а каждый номер записать на отдельной карточке. Отбирая наудачу после тщательного перемешивания из пачки таких карточек по одной карточке, получим выборочную совокупность любого нужного объема, которая называется *собственно-случайной.*

Номера на отобранных карточках укажут, какие члены генеральной совокупности попали в выборку. При этом возможны два принципиально различных способа огбора карточек в зависимости от того, возвращается или не возвращается обратно вынутая карточка после записи ее номера. Выборочная совокупность, образованная по первой схеме» называется *собственно-случайной, а повторным отбором членов,* по второй — *собственно-случайной о бесповторным отбором членов.* Для краткости далее их будем часто называть соответственно *повторной* и *бесповторной* выборками.

Собственно-случайная бесповторная выборка образуется и в том случае, когда из тщательно перемешанной пачки сразу взято нужное число карточек.

Собственно-случайную выборку заданного объема *n* можно образовать и с помощью так называемых таблиц случайных чисел или генератора случайных чисел на компьютере.

При образовании собственно-случайной выборки каждый член генеральной совокупности с одинаковой вероятностью может попасть в выборку.

2. Выборка, в которую члены из генеральной совокупности отбираются через определенный интервал, называется *механической.*

Например, если объем выборки должен составлять 5% объема генеральной совокупности (5%-ная выборка), то отбирается ее каждый 20-й член, при 10%-ной выборке — каждый 10-й член генеральной совокупности и т.д. Механическую выборку можно образовать, если имеется определенный порядок следования членов генеральной совокупности, например, если они следуют друг за другом в определенной последовательности во времени. Именно так появляются готовые детали со станка, приборы с конвейера и т, п. При этом необходимо убедиться, что в следующих один за другим членах генеральной совокупности значения признака не изменяются с той же (или кратной ей) периодичностью, что и периодичность отбора элементов в выборку.

Пусть из продукции станка в выборку попадает каждая пятая деталь, а после каждой десятой детали рабочий производит смену (или заточку) режущего инструмента и подналадку станка. Эти операции рабочего направлены на улучшение качества деталей, износ режущего инструмента происходит более или менее равномерно. Следовательно, в выборочную совокупность попадут детали, на качество которых работа станка влияет в одну и ту же сторону, а показатели выборочной совокупности могут неправильно отразить соответствующие показатели генеральной совокупности.

3. Если из предварительно разбитой на непересекающиеся группы генеральной совокупности образовать собственно-случайные выборки из каждой группы (с повторным или бесповторным отбором членов), то отобранные элементы составят выборочную совокупность, которая называется *типической.*

Оказывается, что выборочная совокупность с большей достоверностью воспроизводит однородную генеральную совокупность. Качество изделий различных цехов, участков, станков и смен может оказаться существенно различным. Поэтому при изучении качества изделии, выпускаемых предприятием, целесообразно образовывать выборку не из общей массы изготовляемой предприятием продукции, а из продукции отдельно каждого цеха. смены (ночной, дневной) и даже участка, станка, т. е. образовать типическую выборку.

4. Если генеральную совокупность предварительно разбить на непересекающиеся серии (группы), а

затем, рассматривая серии как элементы, образовать собственно-случайную выборку (с повторным или бесповторным отбором серий), то все члены отобранных серий составят выборочную совокупность, которая называется *серийной.*

Пример.

Предположим, что на заводе 150 станков (10 цехов по 15 станков) производят одинаковые изделия. Если в выборку отбирать изделия из тщательно перемешанной продукции всех 150 станков, то образуется собственно-случайная выборка. Но можно отбирать изделия отдельно из продукции первого, второго и т. д. станков. Тогда будет образована типическая выборка. Если же членами генеральной совокупности считать цехи и в каждом из цехов образовать повторную или бесповторную выборку, то вся отобранная продукция составит серийную выборку.

1.4. Статистическая совокупность и статистические признаки

Все объекты (элементы), составляющие генеральную совокупность, должны иметь хотя бы один общий признак, позволяющий классифицировать объекты, сравнивать их друг с другом (например, пол, возраст, спортивная квалификация и т.п.). Наличие общего признака является основой для образования статистической совокупности. Таким образом, статистическая совокупность представляет собой результаты описания или измерения общих признаков объектов исследования.

Предметом изучения в статистике естественно являются изменяющиеся (варьирующие) признаки, которые иногда называют статистическими признаками. Они делятся на

<table>
<tr><th>качественные</th><th colspan="2">количественные</th></tr>
<tr><td rowspan="3">признаки, которыми объект обладает либо не обладает. Они не поддаются непосредственному измерению (например, цвет волос, специализация, квалификация, национальность, территориальная принадлежность, образование и т.п.).</td><td colspan="2">результаты подсчета или измерения. В соответствии с этим они делятся на</td></tr>
<tr><td>дискретные</td><td>непрерывные</td></tr>
<tr><td>могут принимать лишь отдельные значения из некоторого ряда чисел, например количество прожитых лет, число попаданий и промахов при серии выстрелов.</td><td>могут принимать любые значения в определенном интервале. Например, время работы механизма, скорость движения и т. п.</td></tr>
</table>

Отдельные числовые значения варьирующего признака называются вариантами. Варианты принято обозначать строчными латинскими буквами из конца алфавита: *x, y, z.*

1.5. Общая схема выборочного эксперимента

По эмпирическим данным, представляющим собой выборку из некоторой генеральной совокупности, оцениваются параметры, позволяющие описать всю генеральную совокупность; определяется интервал, в котором с заданным уровнем доверия находится истинное значение оцениваемого параметра; а затем проверяются те или иные утверждения и делаются выводы о свойствах всей генеральной совокупности.

2. ЭМПИРИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ

2.1. Введение

Рассмотрим методы построения эмпирических распределений, т.е. распределений элементов выборки по значениям изучаемого признака.

Выборочные данные, полученные в ходе эксперимента, называются соответственно экспериментальными (эмпирическими) данными.

2.2. Вариационные ряды

Вариационный ряд – ранжированный в порядке возрастания или убывания ряд вариант с соответствующими им весами (частотой, частостью …). То есть вариационный ряд – двойной числовой ряд, показывающий, каким образом численные значения изучаемого признака связаны с их повторяемостью в выборке. Вариационные ряды имеют большое значение при статистической обработке экспериментальных данных, поскольку дают наглядное представление о характерных особенностях варьирования признака.

Вариационные ряды бывают двух типов: интервальные и безынтервальными.

В интервальном вариационном ряду частоты (или частости), характеризующие повторяемость вариант

в выборке, распределяются по интервалам группировки. Интервальный вариационный ряд строится, если изучаемый признак варьирует непрерывно, но используется и для дискретно варьирующих признаков в тех случаях, когда признак варьирует в широких пределах.

В безынтервальном вариационном ряду частоты (или частости) распределяются непосредственно по значениям варьирующего признака. Для построения безынтервального вариационного ряда необходимо варианты выборки расположить в порядке возрастания или убывания (проранжировать) и затем подсчитать, сколько раз каждая из них встречается в выборке. Безынтервальный вариационный ряд применяется в тех случаях, когда исследуемый признак варьирует дискретно и слабо.

Пример:

Превышение разрешенной скорости движения (км/ч)	Кол-во нарушений
20-30	10
30-40	20
40-45	15
45-60	10
Больше 60	5

Признак – непрерывный

Зрение (диоптрии)	Кол-во человек
-10:-6	1
-6:-3	5
-3:-1	8
-1:+1	11
+1:+5	3
+5:+10	2

Признак дискретный, сильно варьирующийся.

Экзамена-ционная оценка	Кол-во студентов
5	5
4	8
3	12
2	5

Признак дискретный, слабо варьирующийся.

2.3. Табличное представление экспериментальных данных.

Как правило, необработанные (первичные) экспериментальные данные представлены в виде неупорядоченного набора чисел, записанных исследователем в порядке их поступления. Этот набор данных трудно обозрим, и сделать по ним какие-то выводы невозможно. Поэтому первичные данные нуждаются в обработке, которая всегда начинается с их группировки.

Группировка представляет собой процесс систематизации, или упорядочения, первичных данных с целью извлечения содержащейся в них информации. Группировка выполняется различными методами в зависимости от целей исследования, вида изучаемого признака и количества экспериментальных данных (объема выборки), но наиболее часто группировка сводится к представлению данных в виде статистических таблиц.

Группировки заключается в распределении вариант выборки по группам, или интервалам группировки, каждый из которых содержит некоторый диапазон значений изучаемого признака.

Шаг 1

Первая задача, которую необходимо решить при группировке, состоит в том, чтобы разбить весь диапазон варьирования признака в выборке (между минимальной и максимальной вариантами выборки) на интервалы группировки. Эта задача требует определения числа интервалов группировки и ширины каждого из них. Обычно предпочтительны интервалы одинаковой ширины, а при выборе числа интервалов исходят из следующих соображений.

Группировка производится для того, чтобы построить эмпирическое распределение и сформировать с его помощью предположения о форме распределения изучаемого признака в генеральной совокупности, из которой взята выборка.

При увеличении числа интервалов группировки и, следовательно, при сужении каждого из них уменьшается число экспериментальных данных, попадающих в каждый интервал. Поскольку выборочные значения случайны, они случайным образом распределяются по интервалам группировки, поэтому картина эмпирического распределения будет содержать много случайных деталей, что мешает установить общие закономерности варьирования признака.

И наоборот, при чрезмерно широких интервалах группировки нельзя получить детальной картины распределения, поэтому возникает опасность упустить важные закономерные подробности формы распределения.

Поэтому вопрос о выборе числа и ширины интервалов группировки приходится решать в каждом конкретном случае исходя из целей исследования, объема выборки и степени варьирования признака в выборке. Однако приближенно число интервалов k можно оценить исходя только из объема выборки n.

Делается это одним из следующих способов:

1) по формуле Стерджеса $k = 1 + 3.321\lg n$;

2) с помощью табл. 2.1,

Таблица 2.1

Выбор числа интервалов группировки

Объем выборки, n	Число интервалов, k
25—40	5—6
40—60	6—8
60—100	7—10
100—200	8—12
Больше 200	10—15

Шаг 2

Находим ширину каждого из интервалов (при этом все они будут одинаковой ширины) по следующей формуле:

$$h = \frac{x_{макс} - x_{мин}}{k - 1}, \qquad (2.1)$$

где h – ширина интервалов; $x_{макс}$ и $x_{мин}$ — максимальная и минимальная варианты выборки ($x_{макс}$ и $x_{мин}$ находятся непосредственно по таблице исходных данных).

Шаг 3

Наметим границы интервалов группировки. Нижняя граница первого интервала выбирается так, чтобы минимальная варианта выборки $x_{мин}$ попадала примерно в середину этого интервала. Обычно нижняя граница первого интервала определяется как

$$x_{\text{Н}1} = x_{мин} - \frac{h}{2} \qquad (2.2)$$

Прибавив к этой величине ширину интервала, найдем нижнюю границу второго интервала $x_{\text{Н}2} = x_{\text{Н}1} + h$. Это будет одновременно и верхняя граница $x_{\text{В}1}$ предыдущего (первого) интервала.

Аналогично походим $x_{\text{Н}3} = x_{\text{Н}2} + h$. и т.д для всех интервалов.

После того, как намечены границы всех интервалов, остается распределить по этим интервалам выборочные варианты. ***Однако при этом возникает следующий вопрос: как поступать в тех случаях, если какая-либо из вариант попадает точно на границу соседних интервалов группировки, т.е. варианта совпадает с нижней границей одного и верхней границей соседнего с ним интервалам. Такие варианты могут быть с одинаковыми основаниями отнесены к любому из соседних интервалов. Этот выбор оставляется на усмотрение экспериментатора.***

Шаг 4

Вычислим срединные значения интервалов группировки x_i, которые отстоят от нижних границ на величину, равную половине ширины интервалов, т. е.

$$x_i = x_{\text{Н}i} + \frac{h}{2}, \qquad (2.3)$$

где $x_{\text{Н}i}$ — нижняя граница i-го интервала.

Шаг 5

Далее на основании первичных данных распределяем варианты выборки по интервалам группировки, то есть, подсчитываем повторяемость вариант в каждом интервале. Получившиеся числа имеют в статистике определенное название. Числа, показывающие, сколько раз варианты, относящиеся к каждому интервалу группировки, встречаются в выборке, называются частотами интервалов.

Обозначим частоты символом n_i. Общая сумма всех частот всегда равна объему выборки n, что можно использовать для проверки правильности подсчетов.

Шаг 6.1

Вычисляем накопленную частоту интервала — это число, полученное последовательным суммированием частот в направлении от первого интервала к последнему, до того интервала включительно, для которого определяется накопленная частота. Накопленные частоты обозначим n_{xi}.

Шаг 6.2

Вычисляем относительную частоту интервала (отношение частоты к объему выборки). Обозначим частости символом w_i.

$$w_i = \frac{n_i}{n}. \qquad (2.4)$$

Они показывают (выражают) доли (удельные веса) членов совокупности с одинаковым значением признака (для дискретных рядов) или попадающие в один интервал (для непрерывных).

Шаг 6.3

Вычисляем относительные частости. Накопленной относительной частотой (частостью) называется отношение накопленной частоты к объему выборки.

Обозначив накопленную частность как F_i, получаем:

$$F_i = \frac{n_{x_i}}{n} \qquad (2.5)$$

Сумма всех частостей всегда равна 1.

Таблица 2.2

Табличное представление данных о результатах

Номер интервала i	Границы интервалов		Срединные значения x_i	Частоты n_i	Накопл. частоты n_{xi}	Частости w_i	Накопл. относит.частоты F_i
	x_{Hi}	x_{Bi}					

2.3. Графическое представление экспериментальных данных

Дли повышения наглядности эмпирических распределений, используется их графическое представление. Наиболее распространенными способами графического представления являются гистограмма, полигон частот и полигон накопленных частот (кумулята).

2.3.1. Гистограмма

Гистограмма используется для графического представления распределений непрерывно варьирующих признаков и состоит из примыкающих друг к другу прямоугольников, как показано на рис. 2.1. Основание каждого прямоугольника равно ширине интервала группировки, а высота его такова, что **площадь** прямоугольника пропорциональна частоте (или частости) попадания в данный интервал. Если ряд безинтервальный, то ширина всех столбцов выбирается произвольной, но одинаковые. Таким образом, высоты прямоугольников должны быть пропорциональны величинам

$$p_i = \frac{n_i}{h_i}, \qquad (2.6)$$

где n_i — частота i-го интервала группировки; h_i — ширина i-го интервала группировки.

На графике гистограммы основание прямоугольников откладывается по оси абсцисс (x), а высота — по оси ординат (y) прямоугольной системы координат.

Однако в тех случаях, когда ширина всех интервалов группировки одинакова, вид гистограммы не изменится, если по оси ординат откладывать не величины p_i, а частоты интервалов n_i.

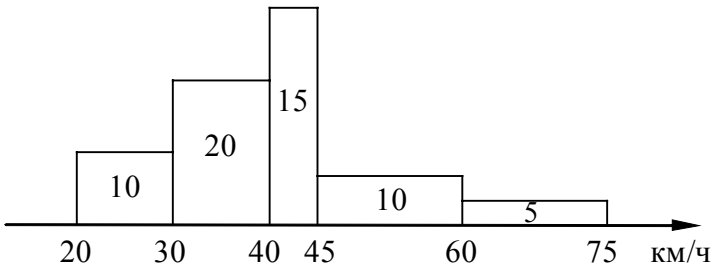

Рис. 2.1. Гистограмма распределения результатов в примере стр. 4 (когда ширина некоторых интервалов группировки неодинакова).

В этом случае чтобы не нарушить принцип построения гистограммы (площади прямоугольников пропорциональны частотам интервалов), по оси ординат уже нельзя откладывать частоты, а надо – высоты прямоугольников (которые должны быть пропорциональны отношениям $\frac{n_i}{h_i}$).

2.3.2. Полигон частот

Еще одним распространенным способом графического представления является полигон частот.

Полигон частот образуется ломаной линией, соединяющей точки, соответствующие срединным значениям интервалов группировки и частотам этих интервалов, срединные значения откладываются по оси *x*, а частоты – по оси *y*.

Из сравнения двух рассмотренных способов графического представления эмпирических распределений следует, что для получения полигона частот из построенной гистограммы (для случая моноширинного распределения) нужно середины вершин прямоугольников, образующих гистограмму, соединить отрезками прямых. Пример полигона частот представлен на рис. 2.2.

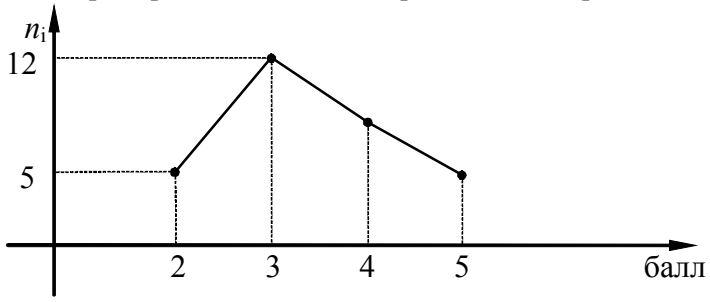

Рис. 2.2. Полигон частот в примере стр. 4 (для дискретного признака).

Полигон частот используется для представления распределений как непрерывных, так и дискретных признаков. В случае непрерывного распределения полигон частот является более предпочтительным способом графического представления, чем гистограмма, если график эмпирического распределения описывается плавной зависимостью.

2.3.3. Полигон накопленных частот

Полигон накопленных частот (кумулята) получается при соединении отрезками прямых точек, координаты которых соответствуют верхним границам интервалов группировки и накопленным частотам. Если по оси ординат откладывать накопленные частости, то полученный график называется полигоном накопленных частостей. Пример полигона накопленных частот приведен на рис. 2.3.

На практике полигон накопленных частот используется в основном для представления дискретных данных. Ему свойственна более плавная форма, чем у гистограммы или полигона частот. Данное свойство и позволяет иногда отдавать предпочтение этому способу графического представлений эмпирических распределений.

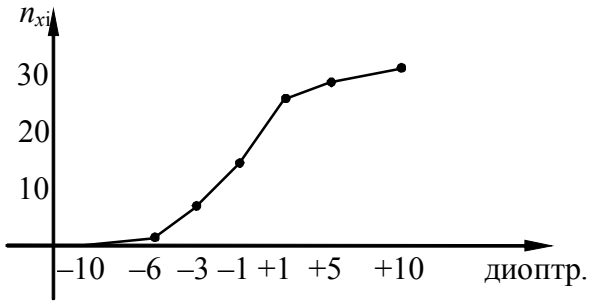

Рис. 2.3. Полигон накопленных частот в примере стр. 4.

2.3.4. Эмпирическая функция распределения

Эмпирической функцией распределения называется функция, вычисляемая для любого значения *x* по формуле

$$F(x) = \frac{n_x}{n},$$

где *n* – объем выборки, n_x – количество вариант, значения которых меньше, чем *x*.

Свойства $F(x)$:

1. Если $x \le x_{\min}$, то $F(x)$ =0;
2. Если $x > x_{\min}$, то $F(x)$ =0;
3. Если $x \in \mathbf{R}$, то $0 \le F(x) \le 1$;
4. $F(x)$ – функция неубывающая.

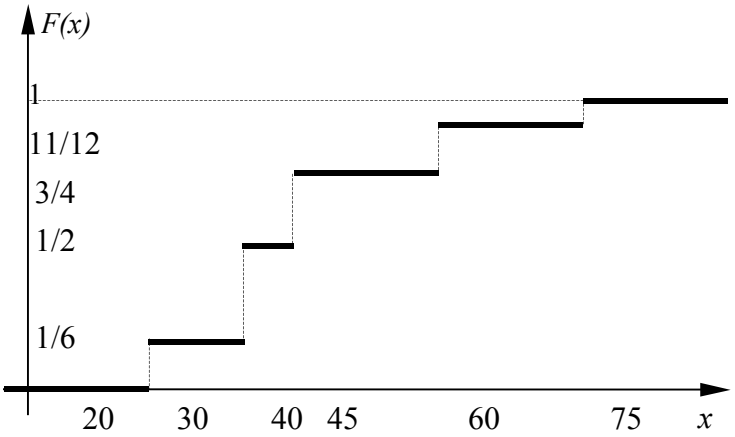

Рис. 2.4 График функции распределения

3. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ ВЫБОРКИ

3.1. Введение

Вариационные ряды и графики эмпирических распределений дают наглядное представление о том, как варьирует признак в выборочной совокупности. Но они недостаточны для полной характеристики выборки, поскольку содержат много деталей, охватить которые невозможно без применения обобщающих числовых характеристик.

Числовые характеристики выборки дают количественное представление об эмпирических данных и позволяют сравнивать их между собой. Наибольшее практическое значение имеют характеристики положения, рассеяния и асимметрии эмпирических распределений.

В этой главе рассматриваются характеристики положения и рассеяния, а также практические методы их вычисления. Характеристики асимметрии будут рассмотрены в гл. 6 применительно к проверке гипотез о виде распределения генеральной совокупности.

3.2. Характеристики положения

В этом разделе рассмотрены характеристики положения, определяющие положение центра эмпирического распределения. Чаще всего употребляются такие характеристики положения, как среднее арифметическое, медиана и мода.

3.2.1. Среднее арифметическое

Среднее арифметическое, или просто среднее, — одна из основных характеристик выборки.

Определение. Среднее арифметическое – такое значение признака, сумма отклонений от которого выборочных значений признака равна нулю (с учетом знака отклонения).

Если воспользоваться геометрической интерпретацией, то среднее арифметическое можно определить как точку на оси *x*, которая является абсциссой центра масс гистограммы.

Среднее принято обозначать той же буквой, что и варианты выборки, с той лишь разницей, что над буквой ставится символ усреднения — черта. Например, если обозначить исследуемый признак через *X*, а его числовые значения — через x_i, то среднее арифметическое имеет обозначение $\overline{x}$.

Среднее арифметическое, как и другие числовые характеристики выборки, может вычисляться как по необработанным первичным данным, так и по результатам группировки этих данных.

Для несгруппированных данных среднее арифметическое определяется по следующей формуле:

$$\overline{x}_B = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (3.1)$$

где *n* — объем выборки; x_i — варианты выборки.

Если данные сгруппированы, то

$$\bar{x}_B = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \qquad (3.2)$$

где n — объем выборки; k — число интервалов группировки; n_i — частота i-ого интервала; x_i — срединное значение i-ого интервала.

Среднее арифметическое, вычисленное по формуле (3.2), называют также взвешенным средним, подчеркивая этим, что в формуле (3.2) числа x_i, суммируются с коэффициентами (весами), равными частотам попадания в интервалы группировки.

Среднее арифметическое измеряется в тех же единицах, что и значения признаков.

Нахождение среднего арифметического непрерывного вариационного ряда осложняется если крайние интервалы не замкнуты (то есть имеют вид "менее 10" "более 60"). В этом случае считается, что ширина первого интервала равна ширине второго, а ширина последнего – ширине предпоследнего.

3.2.2. Медиана

Определение. Медианой (*Me*) называется такое значение признака *X*, когда ровно половина значений экспериментальных данных меньше ее, а вторая половина — больше.

Собственно, этим и ограничивается смысловое значение медианы. Широкое использование этой характеристики на практике объясняется простотой ее вычисления и независимостью от формы распределения эмпирических данных.

Медиана обычно несколько отличается от среднего арифметического. Так бывает всегда, когда имеет место несимметричная форма эмпирического распределения.

Для тех случаев, когда эмпирическое распределение оказывается сильно асимметричным, среднее арифметическое теряет свою практическую ценность, поскольку при этом значительно большая часть значений признака оказывается выше или ниже среднего арифметического. В этой ситуации медиана представляет собой лучшую характеристику центра распределения.

Для дискретного ряда медиана находится по определению.

1. Если данных немного (объем выборки невелик), медиана вычисляется очень просто. Для этого выборку ранжируют, т. е. располагают данные в порядке возрастания или убывания, и в ранжированной выборке, содержащей n членов, ранг R (порядковый номер) медианы определяется как

$$R_{Me} = \frac{n+1}{2}.$$

Пусть, например, имеется ранжированная выборка, содержащая нечетное число членов $n = 9$:

12 14 14 18 20 22 22 26 28.

Тогда ранг медианы

$$R_{Me} = \frac{9+1}{2} = 5,$$

и медиана, обозначаемая символом *Me*, совпадает с пятым членом ряда: $Me = 20$.

2. Если выборка содержит четное число членов, то медиана не может быть определена столь однозначно. Например, получен ряд из 10 членов:

6 8 10 12 14 16 18 20 22 24.

Ранг медианы оказывается равным

$$R_{Me} = \frac{10+1}{2} = 5{,}5.$$

Медианой в этом случае может быть любое число между 14 и 16 (5-м и 6-м членами ряда). Для определенности принято считать в качестве медианы среднее арифметическое этих значений, т. е. $Me = \frac{14+16}{2} = 15.$

Если необходимо найти медиану для сгруппированных данных, то поступают следующим образом.

Вначале находят интервал группировки, в котором содержится медиана, путем подсчета накопленных частот или накопленных относительных частот. Медианным будет тот интервал, в котором накопленная частота впервые окажется больше $n/2$ (n — объем выборки) или накопленная относительная частота — больше 0,5. Внутри медианного интервала медиана определяется по следующей формуле:

$$Me = x_{Me_h} + h_{me}\frac{0{,}5n - n_{x_{Me-1}}}{n_{Me}}, \qquad (3.3)$$

где $x_{Me\text{н}}$ — нижняя граница медианного интервала; $0{,}5n = \frac{n}{2}$ — половина объема выборки; h_{me} — ширина медианного интервала; $n_{x_{Me-1}}$ — накопленная частота интервала, предшествующего медианному, n_{Me} — частота медианного интервала.

3.2.3. Мода

Определение. Мода (*Mo*) представляет собой значение признака, встречающееся в выборке наиболее часто.

Ряд называется *унимодальным*, если в нем только одно модальное значение и *полимодальным* в противном случае. Для полимодального ряда значение моды не определяют.

1. Для дискретного ряда мода находится по определению. Например, если имеем следующие данные измерений кол-ва :

10, 15, 20, 22, 22, 23, 25, 25, 26, 26, 26, 27, 28, 28, 30, 30, 31

то чаще всего встречается (самым *модным* значением является) 26, значит данный ряд унимодальный и $Mo = 26$.

2. Интервал группировки с наибольшей частотой называется модальным. Например, в вариационных рядах, приведенных в разделе 2.2 (стр 4.) модальными интервалами является 30-40 (для превышение разрешенной скорости движения), -1:+1 (острота зрения).

Для определения моды в интервальном ряду используется следующая формула:

$$Mo = x_{Mo\text{н}} + h\frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})}, \qquad (3.4)$$

где $x_{mo\text{н}}$ — нижняя граница модального интервала; h — ширина интервала группировки; n_{Mo} — частота модального интервала; n_{Mo-1} — частота интервала, предшествующего модальному; n_{Mo+1} — частота интервала, следующего за модальным.

3.3. Характеристики рассеяния

Средние значения не дают полной информации о варьирующем признаке. Нетрудно представить себе два эмпирических распределения, у которых средние одинаковы, но при этом у одного из них значения признака рассеяны в узком диапазоне вокруг среднего, а у другого – в широком. Поэтому наряду со средними значениями вычисляют и характеристики рассеяния выборки. Рассмотрим наиболее употребительные из них.

3.3.1. Размах вариации

Определение. Размах вариации – разность между максимальной и минимальной вариантами выборки:

$$R = x_{\max} - x_{\min}$$

Как видим, размах вычисляется очень просто, и в этом его главное и единственное достоинство. Информативность этого показателя невелика. Можно привести очень много распределений, сильно отличающихся по форме, но имеющих одинаковый размах. Размах вариации используется иногда в практических исследованиях при малых (не более 10) объемах выборки, Например, по размаху вариации легко оценить, насколько различаются лучший и худший результаты в группе спортсменов. При больших объемах выборки к его использованию надо откоситься с осторожностью.

3.3.2. Дисперсия и стандартное отклонение

Дисперсия и стандартное отклонение являются важнейшими характеристиками рассеяния.

Определение. Дисперсией называется средний квадрат отклонения значений признака от среднего арифметического. Дисперсия, вычисляемая но выборочным данным, называется выборочной дисперсией и обозначается σ_B^2.

Выборочную дисперсию вычисляют по приведенным ниже формулам:

Для несгруппированных данных

$$\sigma_B^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_B)^2}{n}. \qquad (3.5)$$

В этой формуле $\sum_{i=1}^{n}(x_i - x_B)^2$ — сумма квадратов отклонений значений признака x_i от среднего арифметического x. Для получения среднего квадрата отклонений эта сумма поделена на объем выборки n.

Для сгруппированных в интервальный вариационный ряд данных:

$$\sigma_B^2 = \frac{\sum_{i=1}^{k} n_i(x_i - \bar{x}_B)^2}{n}. \qquad (3.6)$$

Здесь x_i — срединные значения интервалов группировки; $\sum_{i=1}^{k} n_i(x_i - \bar{x}_B)^2$ — взвешенная сумма квадратов отклонений.

Размерность дисперсии не совпадает с единицами измерения варьирующего признака. Дисперсия измеряется в единицах измерения признака в квадрате.

Определение. Стандартным отклонением (или средним квадратическим отклонением) называется корень квадратный из дисперсии:

$$\sigma_B = \sqrt{\sigma_B^2}. \qquad (3.7)$$

Размерность стандартного отклонения в отличие от размерности дисперсии совпадает с единицами измерения варьирующего признака, поэтому в практической статистике для того, чтобы охарактеризовать рассеяние признака используют обычно стандартное отклонение, а не дисперсию.

3.3.3. Коэффициент вариации

Стандартное отклонение выражается в тех же единицах измерения, что и характеризуемый им признак. Если требуется сравнить между собой степень варьирования признаков, выраженных в разных единицах измерения, возникают определенные неудобства. Пусть, например, результаты в беге на 100 м, показанные труппой IX классов, имеют стандартное отклонение 0,9сек (при среднем времени 14,8 сек), а исследование роста тех же учащихся показывает, что его стандартное отклонение составляет 6 см (при среднем росте 168 см). Какой из признаков варьирует сильнее? Очевидно, что только на основании сравнения стандартных отклонений на этот вопрос ответить нельзя. Требуется сопоставить стандартные отклонения со средними арифметическими этих признаков. Поэтому вводится относительный показатель

$$V = \frac{\sigma_B}{\bar{x}_B}, \qquad (3.8)$$

называемый коэффициентом вариации.

Обычно он выражается в процентном отношении:

$$V = \frac{\sigma_B}{\bar{x}_B} \cdot 100\%.$$

Коэффициент вариации является относительной мерой рассеяния признака.

Коэффициент вариации используется и как показатель однородности выборочных наблюдений. Считается, что если коэффициент вариации не превышает 10 %, то выборку можно считать однородной, т. е. полученной из одной генеральной совокупности.

Однако к использованию коэффициента вариации нужно подходить с осторожностью. Продемонстрируем возможные ошибки на следующем примере.

Если на основании многолетних наблюдений среднее арифметическое среднесуточных температур 8 марта составляет в какой-либо местности 0°С, то по формуле (3.8) получим бесконечный коэффициент вариации независимо от разброса температур. Поэтому в данном случае коэффициент вариации не применим! в качестве показателя рассеяния температур, а специфику явления более объективно оценивает стандартное отклонение 5.

Коэффициент вариации можно использовать как относительную меру рассеяния только в тех случаях, когда значения признака измерены в шкале с абсолютным нулем.

Практически коэффициент вариации применяется в основном для сравнения выборок из однотипных генеральных совокупностей.

3.3.4. Коэффициент осцилляции

С целью, аналогичной введению коэффициента вариации, вводится коэффициент осцилляции по

формуле

$$K_p = \frac{R}{\bar{x}_B} \cdot 100\% .$$

3.3.5. Коэффициент асимметрии и эксцесса

Выборочный центральный момент s-ого порядка вычисляется следующим образом:
для несгруппированных данных

$$\mu_s^* = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_B)^s}{n}, \qquad (3.9)$$

для сгруппированных в интервальный вариационный ряд данных:

$$\mu_s^* = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^s}{n}. \qquad (3.10)$$

В частности, при *s*=2 второй центральный момент случайной величины есть дисперсия (сравни с формулами 3.5 и 3.6).

На практике используются третий и четвертый центральные моменты, позволяющие судить о симметричности и остроте вершины кривой распределения случайной величины.

Применяется так называемый коэффициент асимметрии, который является безразмерной величиной и определяется как

$$\gamma_3^* = \frac{\mu_3^*}{\sigma_B^3}. \qquad (3.11)$$

Если $\gamma_3^* = 0$, то распределение симметрично относительно математического ожидания, если $\gamma_3^* > 0$, то преобладают положительные отклонения от математического ожидания, если $\gamma_3^* < 0$ — отрицательные.

Об остроте вершины кривой распределения судят по коэффициенту эксцесса:

$$\gamma_4^* = \frac{\mu_4^*}{\sigma_B^4} - 3. \qquad (3.12)$$

Если $\gamma_4^* > 0$, то распределение имеет острый пик (по сравнению с нормальным распределением), если $\gamma_4^* < 0$ (минимальное значение $\gamma_4^* = -2$), то распределение имеет плосковершинную форму (по сравнению с нормальным распределением, для которого $\gamma_4 = 0$ см. 4.9.1).