Основы математической статистики

Содержание

Основные понятия математической статистики

Применение большинства статистических методов основано на идее использования статистической случайной совокупности испытуемых из общего числа тех, на которых можно было бы распространить (генерализовать) выводы, полученные в результате изучения совокупности.

Математическая статистика – наука, которая на основании статистических данных занимается установлением закономерностей, присущих массовым случайным явлениям.

Статистическая совокупность – это группа, состоящая из большого числа относительно однородных объектов, взятых вместе в известных границах пространства и времени (например, дети, появившиеся в течение какого-то года или месяца). Статистическая совокупность формируется по признакам, они делятся на количественные и атрибутивные (качественные).

Признак называют количественным, если значения признака выражаются числами, например, масса, объём, производительность труда и т. п.

Признак называют качественным, если его значения характеризуют некоторое свойство или состояние элементов совокупности. Например профессия, квалификация, сорт, окраска, консистенция, форма, всхожесть семян, пол экземпляра и т. п.

Статистическая совокупность, состоящая из набора всех значений признака, интересующих исследователя, называется **генеральной совокупностью**. Для определенности генеральные совокупности обозначают прописными буквами латинского алфавита – X, Y, Z, …, а их объемы (число объектов) – N, M…

Безусловно, изучение всей генеральной совокупности позволяет получить более полную информацию об исследуемых объектах. Однако в силу различного рода причин (ограничение по пространству, времени, материальным затратам, слишком большое число объектов, разрушение объекта при его изучении и др.) вся генеральная совокупность в подавляющем числе случаев для исследователя бывает недоступна. В этом случае о закономерностях всей совокупности приходится судить, исследуя ее часть – **выборку**. Так, например, по анализу элементов, содержащихся в капле крови, медики нередко судят о составе всей крови человека, так и по выборочной совокупности учащихся изучаются явления, характерные для всей генеральной совокупности.

Когда для каждого объекта в выборке измерено значение одной переменной, популяция и выборка называются ***одномерными***. Если же для каждого объекта регистрируются значения двух или нескольких переменных, такие данные называются ***многомерными***.

В целях классификации применимости статистических методов рассмотрим следующие **типы исходных данных**:

1. Одна выборка – совокупность измерений одной переменной, произведенных в ходе эксперимента, опроса или наблюдения. Для одной выборки используются статистические методы описательной статистики.

Выборка может быть: неупорядоченная и структурированная (упорядоченная).

2. Несколько выборок – совокупность измерений нескольких переменных, произведенных в ходе эксперимента. Выборки могут быть:

– **независимые** – получены в эксперименте независимо друг от друга;

– **зависимые** – значения данных переменных каким-то образом согласованы (связаны) друг с другом в имеющихся наблюдениях.

Приведем типичные примеры зависимых переменных: рост человека связан с его массой, потому что обычно высокие индивиды тяжелее низких; IQ (коэффициент интеллекта) связан с количеством ошибок в тесте, так как люди с высоким IQ, как правило, делают меньше ошибок и т. д.

Для экспериментальной педагогики характерна постановка исследований, преследующих цель выявления эффективности педагогических средств путем сравнения достижений или свойств одной и той же группы учащихся в разные периоды времени (такие группы получили название зависимых выборок) или разных групп учащихся (независимые выборки).

3. Временной ряд или процесс представляет собой значения количественной переменной (отклика), измеренной через равные интервалы значений другой количественной переменной (параметра). Например, время измерения; в качестве исходных данных рассматриваются, как правило, значения переменной отклика.

4. Связные временные ряды – синхронные по времени измерения одной переменной в разных точках (объектах) или же измерения нескольких переменных в одной точке (объекте).

5. Многомерные данные представляются для статистического анализа в виде прямоугольной матрицы. Это могут быть измерения значений переменных у нескольких объектов или в нескольких точках, или же это могут быть измерения значений переменных у одного объекта в различные моменты времени или при различных состояниях.

Множество объектов, случайным образом отобранных из генеральной совокупности для изучения определённого признака, называется выборкой. Соответствующие выборочные совокупности обозначаются строчными буквами латинского алфавита – x, y, … Число элементов в конкретной выборке называется объёмом выборки и обозначается n, m, …

Выборочный метод. Способы отбора из генеральной совокупности

На практике применяются различные способы отбора из генеральной совокупности. Во-первых, отбор, не требующий расчленения генеральной совокупности на части: это простой случайный бесповторный отбор и простой случайный повторный отбор. Во-вторых, отбор, при котором генеральная совокупность разбивается на части: типический отбор, механический отбор, серийный отбор.

В том случае, когда не требуется предварительного исследования на предмет однородности генеральной совокупности, используется простой случайный повторный или бесповторный отбор. Простым случайным называют такой отбор, при котором объекты извлекают по одному из всей генеральной совокупности. При повторном отборе объект после извлечения из генеральной совокупности регистрируют и вновь возвращают в генеральную совокупность, откуда она опять может быть извлечена случайным образом (собственно-случайная повторная выборка). Например, при изучении методов полевого опыта. При бесповторном отборе отобранный элемент в выборку обратно не возвращают (бесповторная выборка). Например, при изучении качественных характеристик сорта, при проверке качества продуктов и т. п.

Отметим, что в практических целях чаще используется бесповторный случайный отбор. Случайный повторный отбор обычно используется, когда необходимо создать теоретическую модель исследуемой генеральной совокупности (допустим, при изучении пропорций исследуемой совокупности). Следует отметить, что если выборка составляет лишь малую часть всей совокупности, то различие между повторным и бесповторным отборами стирается.

В свою очередь, в зависимости от состава и соотношений генеральной совокупности при организации исследования применяются механический, типический, серийный (расслоенная выборка) или двухстадийный отборы.

Механическим называют отбор, при котором генеральную совокупность «механически» делят на n равных групп и из каждой группы отбирают один объект (объем выборки равен n). Механический отбор всегда бесповторный. Например, для оценки успеваемости студентов проводится 10-процентная механическая выборка. Для этого из алфавитного списка отбирается каждый десятый студент, например, 1-й, 11-й, 21-й, 31-й или 7-й, 17-й, 27-й, 37-й и т. д. Если выборка 5%, то отбору подлежит каждый двадцатый студент.

Однако в случаях, когда генеральная совокупность неоднородна и доступность для исследователя отдельных ее частей неодинакова, обеспечить репрезентативность выборки (см. пояснения в пункте «Ошибки выборочного наблюдения»), используя механический метод случайного отбора, невозможно. В подобных случаях для обеспечения представительности следует использовать расслоённую выборку (типическую или районированную), при этом внутри каждого выделенного слоя необходимо соблюдать принцип случайности.

Типическим называют отбор, при котором объекты отбираются не из всей (неоднородной) генеральной совокупности, а из каждой её «типической» части. При этом предварительно вся совокупность расчленяется на относительно однородные (типические) составные части с определением «веса» (относительной численности) каждой из них. Затем случайным образом, пропорционально «весу», из каждой части извлекается выборка. Такой метод составления выборок используют, например, в работе с гибридами.

Серийным называют отбор, при котором объекты отбирают из генеральной совокупности не по одному, а «сериями», и они подвергаются сплошному обследованию (например, таможенная служба выборочно вскрывает каждый сотый контейнер из прибывающих в порт, а в нем проверяется полностью весь груз). Серийным отбором пользуются тогда, когда обследуемый признак колеблется в различных сериях незначительно.

Особенность двухстадийного отбора состоит в том, что сначала из всей генеральной совокупности отбирают случайную выборку, а затем уже из неё производят выборку «второго порядка», которую непосредственно и анализируют. Например, при анализе качества семян сначала отбирают пробы из партии семян, а затем из неё выделяют навеску для определения всхожести, чистоты, химического состава и т. п.

Отметим, что на практике часто применяется комбинированный отбор, при котором сочетаются указанные выше способы. При этом выбираемый способ отбора объектов из генеральной совокупности должен соответствовать поставленной задаче и хорошо согласовываться с сущностью изучаемого явления.

Одной из основных задач статистического анализа является получение по имеющейся выборке достоверных сведений об интересующих исследователя характеристиках генеральной совокупности.

При осуществлении отбора возможны ошибки выборочного наблюдения: ошибки регистрации и ошибки репрезентативности. Ошибка регистрации возникает, когда истинное значение изучаемого признака не совпадает с его зарегистрированным значением. Для того чтобы избежать ошибки выборочного наблюдения – ошибки репрезентативности, необходимо выполнять условие репрезентативности (от лат. represento – представляю), то есть выборка должна полно и адекватно представлять свойства генеральной совокупности.

Ошибка репрезентативности представляет собой разность между выборочными и генеральными характеристиками изучаемой совокупности. Если эта разность равна нулю, то ошибки нет. Таким образом, информация, полученная в результате осуществления выборки, будет только тогда надежной основой для принятия решения относительно тех или иных свойств генеральной совокупности, когда структура образующих выборку элементов будет аналогична структуре элементов в генеральной совокупности.

Исходным понятием статистики является понятие **совокупность**, объединяющее обычно какое-либо множество испытуемых (учащихся) по одному или нескольким интересующим признакам. Главное требование к выделению изучаемой совокупности – это её качественная однородность, например по уровню знаний, росту, массе и другим признакам. Члены совокупности могут сравниваться между собой в отношении только того качества, которое становится предметом исследования. При этом обычно абстрагируются от других не интересующих качеств. Так, если педагога интересует успеваемость учащихся, то он не принимает во внимание, как правило, их рост, массу и другие параметры, не относящиеся непосредственно к изучаемому вопросу.

Обычно в статистике различают три типа значений переменных (признака): **количественные, номинальные и ранговые**.

Значения количественных переменных являются числовыми, могут быть упорядочены, и для них имеют смысл различные вычисления (например, среднее значение). На обработку количественных переменных ориентировано подавляющее большинство статистических методов.

Значения номинальных переменных, например пол, вид, цвет, являются не числовыми, они означают принадлежность к некоторым классам и не могут быть упорядочены или непосредственно использованы в вычислениях. Для анализа номинальных переменных специально предназначены лишь избранные разделы математической статистики, например категориальный анализ. Однако в ряде случаев для этой цели могут быть использованы и некоторые ранговые и количественные методы, если номинальные значения предварительно заменить на числа, обозначающие их условные коды.

Ранговые или порядковые переменные занимают промежуточное положение: их значения упорядочены (состояние больного, степень предпочтения), но не могут быть с уверенностью измерены и сопоставлены количественно. К анализу ранговых переменных применимы так называемые ранговые методы.

Ранг наблюдения – это тот номер, который получит данное наблюдение в упорядоченной совокупности всех данных – после их упорядочивания по определенному правилу (например, от большего значения к меньшим). Процедура перехода от совокупности наблюдений к последовательности их рангов называется ранжированием.

Пример. В группе студентов были проведены измерения массы. Полученные результаты записали в порядке их поступления:

$$48;\ 65;\ 53;\ 78;\ 69;\ 58;\ 85;\ 70;\ 68;\ 81;\ 72;\ 67.$$

При этом ранжированный ряд будет иметь вид:

$$48;\ 53;\ 58;\ 65;\ 67;\ 68;\ 69;\ 70;\ 72;\ 78;\ 81;\ 85.$$

Очевидно, что упорядоченный ряд более удобен для наблюдения и анализа.

Ранговые и номинальные значения переменных при вводе данных следует обозначать целыми числами.

Отметим, что основные понятия математической статистики предназначены для представления экспериментальных данных в удобном виде и описания информации в терминах математической статистики и теории вероятностей.

Основной величиной в статистических измерениях является единица статистической совокупности (например, любой из критериев оценки качества педагога-исследователя). Единица статистической совокупности характеризуется набором признаков или параметров. Значения каждого параметра или признака могут быть различными и в целом образовывать ряд случайных значений x_1, x_2, …, x_n.

Переменная (*variable*) – это параметр измерения, который можно контролировать или которым можно манипулировать в исследовании, так как значения переменных не постоянны, нужно научиться описывать их изменчивость.

Для этого придуманы описательные, или дескриптивные, статистики: минимум, максимум, среднее, дисперсия, стандартное отклонение, медиана, квартили, мода.

Относительное значение параметра – это отношение числа объектов, имеющих этот показатель, к величине выборки. Выражается относительным числом или в процентах (процентное значение).

Пример. Успеваемость в классе равна числу положительных итоговых отметок, деленному на число всех учащихся класса. Умножение этого значения на 100 дает успеваемость 25%.

Ошибки выборочного наблюдения

При осуществлении выборки возможны *ошибки выборочного наблюдения*: ошибки регистрации и ошибки репрезентативности. *Ошибка регистрации* возникает, когда истинное значение изучаемого признака не совпадает с его зарегистрированным значением. Для того чтобы избежать ошибки выборочного наблюдения – *ошибки репрезентативности* – необходимо выполнять условие репрезентативности (от лат. *represento* – представляю), то есть выборка должна полно и адекватно представлять свойства генеральной совокупности.

Ошибка репрезентативности представляет собой разность между выборочными и генеральными характеристиками изучаемой совокупности. Если эта разность равна нулю, то ошибки нет. Таким образом, информация, полученная в результате осуществления выборки, будет только тогда надежной основой для принятия решения относительно тех или иных свойств генеральной совокупности, когда структура образующих выборку элементов будет аналогична структуре элементов в генеральной совокупности.

Статистическое распределение выборки

Пусть при проведении эксперимента из генеральной совокупности X объёмом N извлечена выборка x объёмом n ($n \le N$), элементами (признаками) которой является ряд значений:

$$x_1, x_2, \ldots, x_k. \qquad (1)$$

Пусть признак x_1 встречается в выборке n_1 раз, x_2 встречается n_2 раза, …, x_k – n_k раз, где $\sum_{i=1}^{k} n_i = n$. Наблюдаемые значения признака называют *вариантами*. Если варианты записаны в порядке их поступления в процессе отбора, то (1) называется *простым* вариационным рядом. Если же для обнаружения закономерности при первичном анализе данные в (1) *сгруппированы* и записаны в порядке их возрастания или убывания, то вариационный ряд называется *ранжированным* (от фр. *ranger* – выстраивать).

Соответствующие вариантам x_i числа n_i называются *частотами* (или «*весами*») признака, а их отношения к объёму выборки $\omega_i = \frac{n_i}{n}$ – *относительными частотами* (*частостями*), где $i = 1, 2, \ldots, k$. Несложно проверить, что при этом выполняется условие $\sum_{i=1}^{k} \omega_i = 1$.

Для удобства представления и анализа данных *варианты* ранжированной выборки и соответствующие им частоты или частости заносятся в специальную таблицу:

Таблица 1

x	x_1	x_2	…	x_i	…	x_k
n	n_1	n_2	…	n_i	…	n_k
ω	ω_1	ω_2	…	ω_i	…	ω_k

Таблицу, содержащую все сгруппированные, ранжированные варианты выборки и соответствующие им частоты или относительные частоты, называют ***статистическим распределением выборки*** *признака.*

Пример. В группе студентов были проведены измерения массы. Полученные результаты записали в порядке их поступления:

48; 65; 53; 78; 69; 58; 85; 70; 68; 81; 72; 67.

При этом ранжированный ряд будет иметь вид:

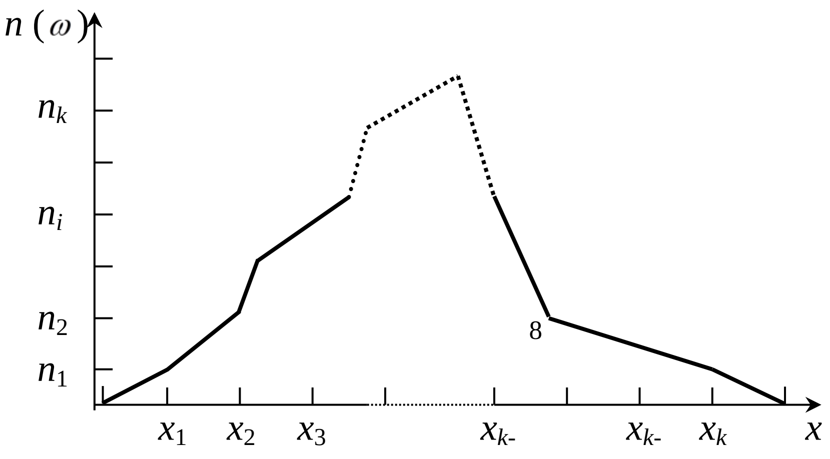
48; 53; 58; 65; 67; 68; 69; 70; 72; 78; 81; 85.

Вариация признака может быть дискретной или непрерывной. Признак называется *дискретно варьируемым*, если его отдельные значения (варианты) отличаются друг от друга на некоторую конечную величину (обычно целое число). Вариационный ряд таких признаков называется *дискретным вариационным рядом* (*статистическим дискретным рядом распределения*). Например, тарифный разряд рабочего, цена товара и т. д. Существует множество признаков, значения которых отличаются друг от друга на сколь угодно малую величину, то есть признак может принимать любые значения в некотором интервале. Такие признаки называются *непрерывно варьирующими*. К подобным признакам можно отнести среднедушевой доход, процент дневной выработки рабочего и т. п.

Графическое изображение рядов распределения облегчает их анализ и позволяет судить о форме распределения. Для графического изображения дискретного ряда распределения используют *полигон частот* (*относительных частот*).

Полигон частот **(*относительных частот*)** *– это ломаная, вершинами которой являются точки с координатами, соответствующими парам значений статистического распределения* (x_1; n_1), (x_2; n_2), …, (x_k; n_k) или (x_1; ω_1), (x_2; ω_2), …, (x_k; ω_k) (см. табл. 1). При построении полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им (относительные) частоты n_i (ω_i), где $i = \overline{1,k}$.

Рис. 1



Пример. Имеются данные о появлении числа курочек на 10 цыплят:

Количество курочек	0	1	2	3	4	5	6	7	8	9	10
Число случаев	1	10	45	120	210	252	210	120	45	10	1

Графически изобразить приведённые данные.

Решение. Полигон распределения числа курочек среди 10 цыплят (общее число – 1024) имеет следующий вид (*биномиальная кривая распределения*) (рис. 2).

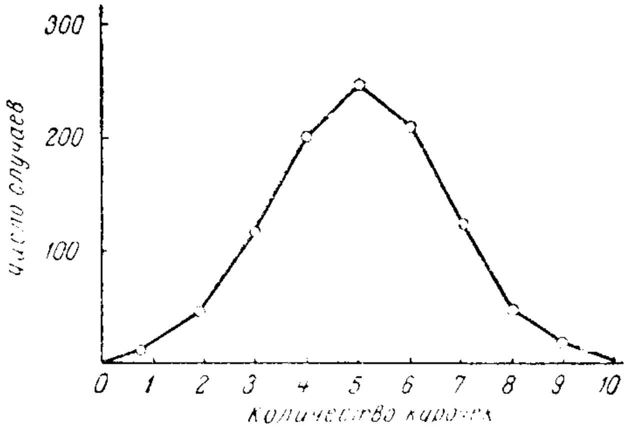


Рис. 2

Пример. Составить статистическое распределение выборки (n = 10), полученной в результате исследования действия лекарственного препарата (в единицах действия на 1 мг): 725, 740, 705, 660, 795, 865, 740, 725, 740, 705.

Решение. После ранжирования и группировки данных получим статистический дискретный ряд распределения в виде таблицы частот

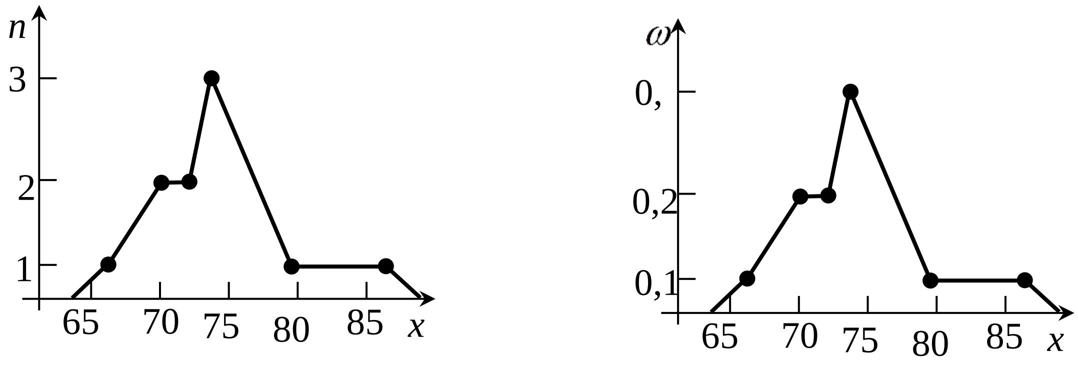
x [ед/мг]	660	705	725	740	795	865
n	1	2	2	3	1	1

или относительных частот

x [ед/мг]	660	705	725	740	795	865
ω	0,1	0,2	0,2	0,3	0,1	0,1

Отметим, что вторая таблица иллюстрирует взаимосвязь статистического дискретного ряда распределения выборки и закона распределения дискретной случайной величины (то есть понятий математической статистики и теории вероятностей).

Выполним графическое представление статистического ряда распределения, полученного при решении данного примера, которое имеет вид (рис. 3 *а*) и *б*)).



а) Рис. 3 б)

Статистический интервальный ряд распределения

Практика показывает, что для представления выборки статистический ряд распределения может быть использован в случае малочисленной выборки ($n < 30$). В случае большого числа вариант выборочной совокупности или же непрерывно-варьирующегося признака выборку представляют в виде *статистического интервального ряда распределения* (*интервальным вариационным рядом*).

При построении интервального ряда распределения используется ранжированная выборка. Вся область наблюдаемых значений разбивается на конечное число K равных интервалов (классов) длиной Δx. Затем для каждого интервала определяется относительная частота $\omega_i = \dfrac{n_i}{n}$ изучаемого признака. Здесь: n – объем выборки, а n_i – число значений признака, попадающих в отдельный классовый интервал.

Интервальный ряд распределения строится в виде таблиц соответствия интервала и одной из частотных характеристик признака:

Таблица 2

x	$(x_0; x_1)$	$(x_1; x_2)$	…	$(x_{i-1}; x_i)$	…	$(x_{k-1}; x_k)$
n	n_1	n_2	…	n_i	…	n_k

Замечания

1. Подобный простой способ разбиения может дать значительную ошибку, если его применить к расслоённой выборке. В этом случае разбиение на классовые интервалы должно соответствовать структуре изучаемой совокупности.

2. Грубая группировка данных, когда используют либо слишком широкие, либо слишком узкие классовые интервалы, может значительно исказить характерные особенности варьирования признака.

Для того чтобы получился представительный и хорошо обозримый интервальный вариационный ряд, число интервалов разбиения определяется по *формуле Стэрджесса*:

$$K = 1 + 3{,}32 \cdot \lg n \qquad (2)$$

или *Брукса* и *Карузеса*:

$$K = 5 \cdot \lg n,$$

где n – объём выборки.

Приближённые значения K в зависимости от мощности выборки представлены в таблице:

N	25-40	40-60	60-100	100-200	>200
K	5-6	6-8	7-10	8-12	10-15

3. В тех случаях, когда по вариационному ряду определяются числовые характеристики выборки (средние, дисперсия и др.), рекомендуется (Д. Юл и М. Кендэл), независимо от числа наблюдений, использовать от 15 до 20 классовых интервалов.

Размер классовых интервалов, при одинаковой их длине, находится:

$$\Delta x = \frac{x_{\max} - x_{\min}}{K}. \qquad (3)$$

4. Если объём выборки не превышает 100 единиц ($n \le 100$), то в качестве знаменателя формулы (3) следует использовать выражение (2).

Размер неравных интервалов (расслоённая выборка) определяется в каждом конкретном случае отдельно, с учётом соотношений, имеющих место в генеральной совокупности и в зависимости от задачи исследования. Такая группировка выборочных наблюдений характерна, в основном, для социально-экономических исследований. Но, например, при изучении спроса на некоторую продовольственную форму различных групп населения, в зависимости от возраста или доходов, не может быть использована равномерная группировка выборочных данных.

Основные действия при построении равноинтервального ряда распределения выборки

1. На основании первичного анализа выборочных данных (для уменьшения вероятности ошибки желательно иметь ранжированную выборку) находят минимальную $x_{\min}$ и максимальную $x_{\max}$ варианты.

2. По формуле (3) определяют длину классового интервала Δx и, если значения признака x_i выражены целыми числами и длина классового интервала равна единице ($\Delta x = 1$), то распределение выборки записывается в виде безынтервального ряда распределения. Если $\Delta x \neq 1$, то формируется интервальный ряд распределения.

Пример. Для того чтобы оценить средний возраст работников некоторой сельскохозяйственной отрасли, выбрали случайным образом 68 человек. Признак варьировался в пределах от $x_{\min} = 25{,}9$ до $x_{\max} = 53{,}6$ лет, длину классового интервала будем определять с точностью до десятых:

$$\Delta x = \frac{53{,}6 - 25{,}9}{1 + 3{,}32 \cdot \lg 68} = \frac{27{,}7}{7{,}08} = 3{,}9 \text{ лет.}$$

3. Для того чтобы интервальный ряд наиболее полно отражал природу изучаемого признака и для повышения точности вычисления числовых характеристик статистической совокупности, устанавливают границы интервалов таким образом, чтобы минимальная ($x_{\min}$) и максимальная ($x_{\max}$) варианты выборки попали, соответственно, внутрь первого и последнего классовых интервалов. К решению этого вопроса имеются разные подходы, рассмотрим их.

3.1. Чтобы варианты $x_{\min}$ и $x_{\max}$ попали внутрь, соответственно, первого и последнего интервалов, нижнюю границу x_0 первого интервала находят путём вычитания из $x_{\min}$ принятой точности измерения вариант (ε), а верхнюю границу последнего интервала находят путём прибавления к $x_{\max}$ принятой точности измерения ε. Тогда для предыдущего примера (рассмотрим случай, когда точность измерения вариант составляла $\varepsilon = 1$ год) нижняя граница первого интервала $x_0 = x_{\min} - 1 = 26 - 1 = 25$ лет, а верхняя граница последнего интервала будет равна $x_{\max} + 1 = 54 + 1 = 55$ лет. В этом случае длину классового интервала находят по формуле, следующей из выражения (3):

$$\Delta x = \frac{x_{\max} - x_{\min} + 2\varepsilon}{1 + 3{,}32 \cdot \lg 68} = \frac{55 - 25}{7{,}08} = 4 \text{ года.}$$

Далее, прибавляя к x_0 длину классового интервала Δx, получаем x_1 – верхнюю границу первого интервала: $x_1 = x_0 + \Delta x$, которая одновременно служит и нижней границей второго интервала, и так далее, пока не получим интервал с верхней границей, равной значению $(x_{\max} + 1)$.

3.2. Этот метод применяется наиболее часто. При построении интервального ряда нижнюю границу разбиения выбирают таким образом, чтобы варианта $x_{\min}$ попадала в центр первого интервала. Для выполнения этого условия нижнюю границу первого интервала находят по формуле:

$$x_0 = x_{\min} - \frac{\Delta x}{2}, \qquad (4)$$

где длина классового интервала находится по формуле (3) $\Delta x = \frac{x_{\max} - x_{\min}}{1 + 3{,}32 \cdot \lg n}$.

Возвращаясь к предыдущему примеру, имеем:

$$\Delta x = \frac{54 - 26}{1 + 3{,}32 \cdot \lg 68} = \frac{28}{7{,}08} = 3{,}95 \text{ лет.}$$

Тогда нижняя граница первого интервала $x_0 = 26 - \frac{3{,}95}{2} = 24{,}025 \approx 24$ года.

3.3. Строят интервальный ряд следующим образом:

– верхнюю границу x_1 первого интервала (она одновременно является и нижней границей второго интервала) находят путём прибавления к x_0 длины классового интервала: $x_1 = x_0 + \Delta x$;

– и так далее, пока не получится интервал, внутрь которого попадёт $x_{\max}$:

$$(x_0; x_1), (x_1; x_2), (x_2; x_3), \ldots, (x_n; b), \text{ где } x_n < x_{\max} < b.$$

4. При построении интервального ряда распределения для непрерывного признака имеет место совпадение верхней границы предыдущего интервала с нижней границей последующего. Решать проблему двойного учёта граничных вариант можно по-разному:

–допустимо к одному и тому же классу относить варианты, большие нижней и меньшие верхней границы или равные ей;

–если нижние границы входят в интервал, то верхние границы интервалов уменьшают на величину допускаемой ошибки при определении вариант выборки.

Так как определение числовых характеристик (среднее, дисперсия и др.) статистической совокупности производится на безынтервальных рядах, то интервальный ряд приходится переводить в безынтервальный заменой интервалов центральными или средними значениями величин в соответствующих интервалах $\bar{x}_i$, которые отстоят от их нижних границ $x_н$ на величину, равную половине классового интервала:

$$\bar{x}_i = x_i + \frac{\Delta x}{2}, \qquad (5)$$

то есть средние значения интервалов находятся как полусумма нижних границ соседних интервалов.

Замечание. Если выборка в силу структуры генеральной совокупности не может быть представлена в виде статистического ряда распределения с равными интервалами, то границы интервалов устанавливаются с учётом структуры выборки, а значит и генеральной совокупности. Остальные действия (разграничение интервалов, подсчёт частот, средних значений и др.) могут быть такими же, как и в случае равноинтервального разбиения.

Гистограмма и полигон

Для графического изображения интервального вариационного ряда распределения наиболее часто используют гистограмму.

Гистограммой частот *называется расположенная в прямоугольной системе координат геометрическая фигура, состоящая из прямоугольников, основаниями которых являются откладываемые по оси* Ox *классовые интервалы* (x_{i-1}; x_i) *шириной* Δx, *а соответствующими им высотами могут быть откладываемые по оси* Oy (*относительные*) *частоты*.

Гистограмма плотности относительных частот, построенная на основании данных выше приведённого примера, представлена на рис. 4.

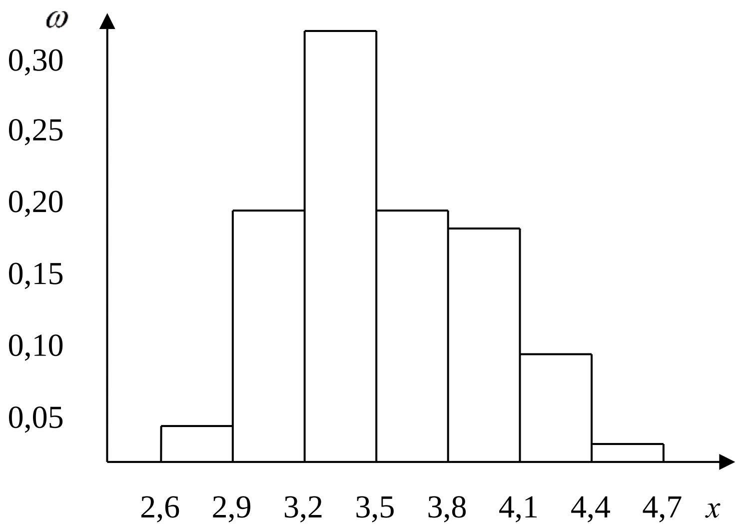


Рис. 4

Гистограмма может быть преобразована в **полигон** (рис. 4). Для этого соединяют точки, абсциссами которых являются середины интервалов, а ординатами – (относительные) частоты, соответствующие данным интервалам. Две крайние точки можно замкнуть на ось абсцисс (*см. полигон относительных частот*).

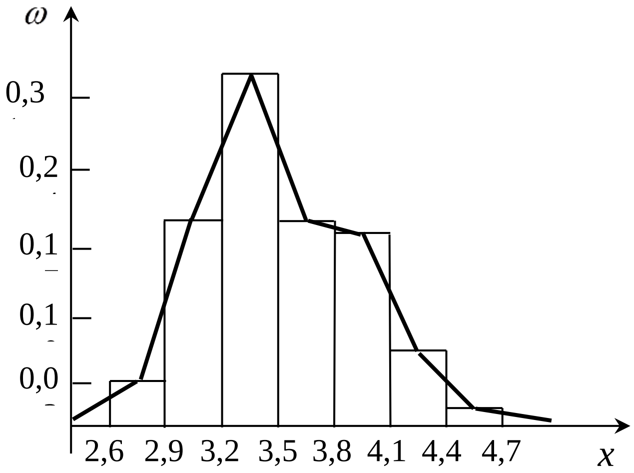


Рис. 5

Кумулята

Кумулята (от лат. *cumulo* – накапливаю) также является графическим изображением интервального ряда распределения. Её называют ещё «кривой эффекта» (*Ван-дер-Верден*, 1960) или «характеристической кривой» (*Треван*, 1927).

Кумулята строится следующим образом:

- Составляется таблица, в которой первая строка содержит значения интервалов признака, а вторая строка – накопленные частоты p_s. Накопленные частоты получаются последовательным суммированием (*кумуляцией*) частот от первого до последнего класса по формуле: $p_s = \sum_{i=1}^{s} n_i$.
- Затем по оси *OX* откладываются значения классовых интервалов (можно откладывать средние значения $\overline{x}_i$), а по оси *OY* – соответствующие накопленные частоты признака p_s.

Пример. Рассмотрим процесс кумуляции на основании следующих данных:

Интервалы	2,6-2,8	2,9-3,1	3,2-3,4	3,5-3,7	3,8-4,0	4,1-4,3	4,4-4,6
$\overline{x}_i$	2,75	3,05	3,35	3,65	3,95	4,25	4,55
Частоты n_i	3	18	30	18	17	8	2
p_s	3	21	51	69	86	94	96

Точки с координатами ($\overline{x}_i$; p_s) соединяют отрезками, получая ломаную *S*-образной формы. Данные таблицы изобразим в виде графика в прямоугольной системе координат (рис. 6):

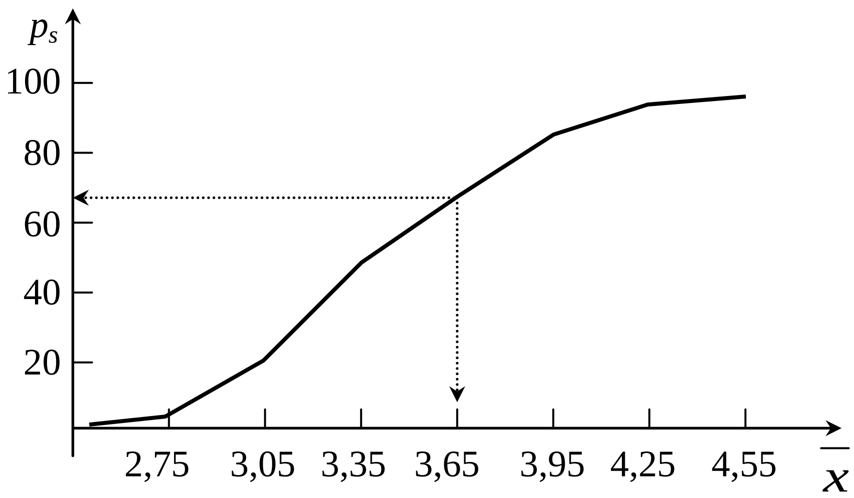


Рис. 6

Для кумуляты нормально распределённой совокупности характерно то, что её центральная точка делит ряд накопленных частот пополам, то есть совпадает с центром рассеяния. Это свойство позволяет её использовать при определении средних доз испытываемых на животных препаратов, вызывающих эффект у 50% подопытных животных. Кумулята отражает зависимость между дозами воздействующего агента (например, лекарственного или токсического вещества) и показателем ответной реакции организма, поэтому её и называют «кривой эффекта».

Литература

1. Гласс Дж. Статистические методы в педагогике и психологии / Дж. Гласс, Дж. Стенли. – М.: Прогресс, 1976. – 496 с.

2. Гланц С. Медико-биологическая статистика / С. Гланц. - М.: Практика, 1998. - 459 с.

3. Гмурман В. Е. Теория вероятностей и математическая статистика: учебное пособие для вузов / В. Е. Гмурман. – М.: Высш. шк., 2003. – 479 с.

4. Новиков Д. А. Статистические методы в педагогических исследованиях (типовые случаи) / Д. А. Новиков. – М.: МЗ-Пресс, 2004. – 67 с.

5. Новиков Д. А. Статистические методы в медико-биологическом эксперименте (типовые случаи) / Д. А. Новиков, В. В. Новочадов. – Волгоград: Изд-во ВГМУ, 2005. – 84 с.

6. Шилова З. В. Теория вероятностей и математическая статистика: учебное пособие / З. В. Шилова, О. И. Шилов. – Киров: Изд-во ВГГУ, 2015. – 158 с.