Тема № 31 «Числовые характеристики рядов данных».

Теория вероятностей – раздел математики, который изучает количественные оценки случайных событий для прогнозирования процессов и явлений в будущем. Основой таких прогнозов являются числовые данные, накопленные в результате наблюдений в реальной жизни. Сбором, изучением и обработкой этих данных занимается наука, основанная на законах теории вероятностей - ***статистика***.

Средние характеристики числового ряда: мода и медиана.

Пример1. Пусть ученик получил в течение года следующие отметки по алгебре: 5,2,4,5,5, 4,4,5,5,5. Какую четвертную отметку поставит ему учитель?

Многих школьников волнует подобная проблема, и чаще всего ученики решают ее следующим естественным образом: складывают все отметки и делят сумму оценок на их количество. В нашем случае:

$$\frac{5+2+4+4+5+5+4+4+5+5}{10}=4{,}4$$

Число 4,4, которое получается в результате, называется ***средним арифметическим*.** Поскольку такую оценку в журнал ставить не принято, учитель, скорее всего, округлит ее до 4.

Средним арифметическим (или выборочным средним) ряда чисел называется частное от деления суммы этих чисел на их количество.

Среднее арифметическое, конечно, является важной характеристикой ряда чисел, в нашем случае — отметок за четверть, но иногда полезно рассматривать и другие средние. Например, претендуя на «5», ученик наверняка будет использовать такой аргумент: «Чаще всего в четверти я получал пятерки!». Статистик в этом случае сказал бы иначе: «Модой этого ряда является число 5».

Модой называют число ряда, которое встречается в этом ряду наиболее часто. Можно сказать, что оно в этом ряду самое «модное».

В отличие от среднего арифметического, которое можно вычислить для любого числового ряда, моды может вообще не быть. Например, пусть тот же ученик получил по русскому языку следующие отметки: 4, 2, 3, 5. Каждая отметка встречается в этом ряду только один раз, и среди них нет числа, встречающегося чаще других. Значит, у этого ряда нет моды. А вот среднее арифметическое, конечно, есть: (4 + 2 + 3 + 5) : 4 = 3,5.

Такой показатель, как мода, можно использовать не только в числовых рядах. Вы уже знакомы с социологическими опросами. Если, например, опросить большую группу учеников, какой школьный предмет им нравится больше всего, то модой этого ряда ответов окажется тот предмет, который будут называть чаще остальных.

Это одна из причин, по которой мода широко используется при изучении спроса. Например, при решении вопросов, в пачки какого веса фасовать масло, какие открывать авиарейсы и т. п., предварительно изучается спрос и выявляется мода — наиболее часто встречающийся заказ. И даже выборы президента, с точки зрения статистики, не более, чем определение моды.

Еще одной важной статистической характеристикой ряда данных является его ***медиана*.**

Пример 2. В конце года 11 учеников 8 класса сдавали норматив по бегу на 100 метров. Были зафиксированы следующие результаты:

Ученик	Результат(с)
Данила	15,3
Петя	16,9
Лена	21,8
Катя	18,4
Стае	16,1
Аня	25,1
Оля	19,9
Боря	15,5
Паша	14,7
Наташа	20,2
Миша	15,4

После того как все ребята пробежали дистанцию, к преподавателю подошел Петя и спросил, какой у него результат.

«Самый средний результат: 16,9 секунды», — ответил учитель.

«Почему? — удивился Петя. — Ведь среднее арифметическое всех результатов — примерно 18,3 секунды, а я пробежал на секунду с лишним лучше. И вообще, результат Кати (18,4) гораздо ближе к среднему, чем мой».

«Твой результат средний, потому что пять человек пробежали лучше, чем ты, и пять — хуже. То есть ты как раз посередине», — сказал учитель.

На языке статистики результат Пети называется ***медианой*** исходного ряда данных. Для того чтобы найти медиану ряда чисел, нужно сначала их упорядочить — составить ранжированный ряд. В нашем примере он выглядит так:
14,7; 15,3; 15,4; 15,5; 16,1; **16,9;** 18,4; 19,9; 20,2; 21,8; 25,1. Средним (шестым по счету) числом является 16,9: пять чисел меньше него, пять чисел больше. Значит, 16,9 — медиана.

Медианой ряда, состоящего из ***нечетного*** количества чисел, называется число данного ряда, которое окажется посередине, если этот ряд упорядочить. ***Медианой*** ряда, состоящего из ***четного*** количества чисел, называется среднее арифметическое двух стоящих посередине чисел этого ряда, если этот ряд упорядочить.

Достоинством медианы является ее большая по сравнению со средним арифметическим «устойчивость к ошибкам». Представим себе, что в наши наблюдения вкралась досадная оплошность: например, при записи одного из результатов соревнований мы пропустили десятичную запятую и вместо 20,2 написали 202. Тогда среднее арифметическое результатов возрастет с 18,1 секунды до 34,6 секунды, а медиана будет по-прежнему 16,9 секунды!

В разных ситуациях имеет смысл использовать разные средние. Поясним это на примерах. Перед нами ранжированный ряд, представляющий данные о времени дорожно-транспортных происшествий на улицах Москвы в течение одних суток (в виде час:мин): 0:15, 0:55,1:20, 3:20, 4:10, 6:10, 6:30, 7:15, 7:45, 8:40, 9:05, 9:20, 9:40, 10:15, 10:15, 11:30, 12:10, 12:15, 13:10, 13:50, 14:10, 14:20, 14:25, 15:20, 15:20, 15:45, 16:20, 16:25, 17:05, 17:30, 17:30, 17:45, 17:55, 18:05, 18:15, 18:45, 18:50, 19:45, 19:55, 20:30, 20:40, 21:30, 21:45, 22:10, 22:35.
Как и для любого ряда в данном случае мы можем найти среднее арифметическое — оно равно 13:33. Однако вряд ли имеет какой-то смысл утверждение типа «аварии на улицах Москвы происходят в среднем в 13 часов 33 минуты». В то же время,

если сгруппировать данные этого ряда в интервалы, можно найти такой временной интервал, когда происходит наибольшее количество ДТП (такую характеристику называют ***интервальной модой*)**. Получив такую характеристику, соответствующим службам имеет смысл серьезно проанализировать, почему именно в этот временной интервал происходит наибольшее количество происшествий, и попытаться устранить их причины.

Рассмотрим другой пример. Вот данные, полученные в результате измерения интервалов времени между звонками на АТС (в с):

23,12,14,20,8,24,12,15,23,20,7,2,28,8,9,14,13,19,23,16.

Здесь вполне оправдано вычисление среднего арифметического. Информация о том, что «звонки поступают в среднем через каждые 15,5 секунд, дает наглядное представление о загруженности телефонных линий. Для этого ряда можно найти также и моду, и медиану, однако практического смысла в данном случае они не имеют.

Рассмотрим теперь более трудный, но важный для практических целей вопрос. Мы знаем, что статистические данные могут быть представлены разными способами — например, может быть *дана не сама выборка, а таблица частот.* Как в этом случае найти среднее арифметическое, моду и медиану?

Конечно, можно пойти по такому пути: восстановить по таблице саму выборку (точнее, ранжированный ряд) и «свести задачу к предыдущей». К счастью, в этом случае есть более рациональный способ вычислений.

Отметка	Абсолютная частота	Относительная частота	Накопленная частота
2	1	0,1	0,1
4	3	0,3	0,4
5	6	0,6	1
ИТОГО	10	1	

Вернемся к примеру, с которого начиналась эта глава: ученик получил в течение года следующие отметки по алгебре: 5,2,4,5,5,4,4,5,5,5.

Представим эти данные в виде таблицы частот.

Мы уже знаем, что для вычисления среднего арифметического надо сложить все числа ряда и поделить полученную сумму на их количество — получится 4,4.

Но если мы знаем, сколько раз повторяется в выборке каждое значение (т. е. знаем его абсолютную частоту), ***вместо многократного сложения одного и того же числа можно умножить его на абсолютную частоту.*** Отсюда получается формула для среднего арифметического, использующая абсолютные частоты значений ряда:

$$\frac{2\cdot 1 + 4\cdot 3 + 5\cdot 6}{10} = 4{,}4$$

Поделим теперь каждое слагаемое в этой формуле на знаменатель — получим формулу для среднего арифметического с помощью относительных частот: $2\cdot 0{,}1 + 4\cdot 0{,}3 + 5\cdot 0{,}6 = 4{,}4$.

Особенно ощутим выигрыш от использования приведенных формул, когда чисел в выборке много и они многократно повторяются.

Что касается моды и медианы, то их вычисление по таблице частот происходит еще проще. Понятно, что ***для вычисления моды нужно найти максимальное зна-***

чение в столбце абсолютных или относительных частот и выбрать соответствующее ему значение числового ряда. В нашем случае максимальная частота равна 6, значит, модой выборки будет 5. Если максимальных частот в таблице несколько, то выборка не имеет моды.

Для вычисления медианы нужно найти первое значение накопленной частоты, превосходящее 0,5, и выбрать соответствующее ему значение числового ряда. В нашем случае накопленная частота впервые превосходит 0,5 только в последней строке таблицы, значит, медианой выборки будет 5.

Вычисление числовых характеристик выборки *по интервальной таблице частот* нуждается в дополнительном комментарии. Ведь в такой таблице первый столбец занимают не числовые значения ряда, а целые интервалы. Каким образом умножать их на абсолютные или относительные частоты? В этом случае вместо интервалов используют их середины, т. е. полу-суммы концов интервала.

Пример 3. Вычислим, сколько в среднем весит портфель первоклассника.

Вес портфеля (в кг)	Абсолютная частота	Относительная частота
от 1 до 2	6	0,3
от 2 до 3	10	0,5
от 3 до 4	3	0,15
от 4 до 5	1	0,05

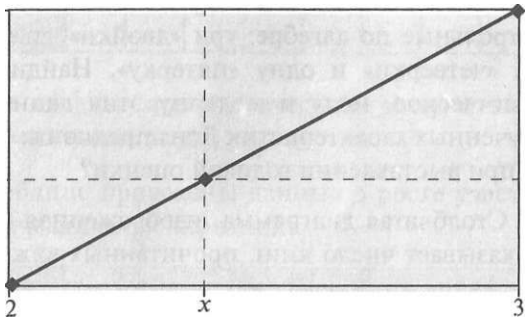
С использованием абсолютных частот:

$$\frac{1{,}5 \cdot 6 + 2{,}5 \cdot 10 + 3{,}5 \cdot 3 + 4{,}5 \cdot 1}{20} = 2{,}45$$

С использованием относительных частот: 1,5·0,3 + 2,5·0,5 + 3,5·0,15 + 4,5·0,05 = 2,45.

Конечно, при вычислении числовых характеристик выборки по интервальной таблице частот получаются только их приближенные значения, ведь мы заменяем целую группу чисел, попадающих в интервал, его серединой. Но с таким приближением вполне можно смириться: во-первых, величина интервалов небольшая; во-вторых, исходные значения выборки, как правило, лежат как слева, так и справа от середины; наконец, в-третьих, все статистические характеристики все равно носят изменчивый характер — в другой выборке они получатся иными. Так, в нашем примере с портфелями точное (до грамма) значение среднего арифметического будет 2,283 кг, в чем вы можете убедиться, если посчитаете его не по интервальной таблице частот, а по самой выборке, приведенной в примере 3. Но вряд ли такая точность имеет смысл в реальных статистических исследованиях.

Для вычисления моды и медианы по интервальной таблице частот в качестве моды берется целый интервал или его середина (в зависимости от постановки задачи), а для вычисления медианы используют *пропорциональное деление* отрезка, на котором происходит «перевал» накопленной частоты через 0,5.



Разберем это на нашем примере с портфелями. Переход накопленной частоты через 0,5 происходит на интервале от 2 до 3. При этом в левом конце интервала накопленная частота равна 0,3, а в правом — 0,8 (см. рис.). Обозначив неизвестную нам медиану через *x,* составим следующую пропорцию:

$$\frac{x-2}{3-2}=\frac{0{,}8-0{,}3}{0{,}5-0{,}3}, \qquad x \approx 2{,}4$$

Размах. Дисперсия. Среднеквадратичное отклонение.

Средние характеристики числового ряда позволяют оценить его поведение «в среднем». Но это далеко не всегда полностью характеризует выборку. Например, на планете Меркурий средняя температура +15°. Исходя из этого статистического показателя, можно подумать, что на Меркурии умеренный климат, удобный для жизни людей. Однако на самом деле это не так. Температура на Меркурии колеблется от —150° до +350°.

Значит, чтобы получить представление о поведении числового ряда, помимо средних характеристик надо знать ***характеристики разброса,*** показывающие, насколько значения ряда различаются между собой, как сильно они «разбросаны» вокруг средних. Простейшей такой характеристикой является размах.

Размах — это разность наибольшего и наименьшего значений ряда данных.

Для температуры на Меркурии, например, размах равен 350° — (-150°) = 500°. Конечно, такого перепада температур человек выдержать не может.

Размах очень просто вычисляется, но не всегда несет достоверную информацию, так как на его величину может сильно повлиять какое-то одно (возможно, ошибочное) значение статистического ряда.

Вот почему в реальных статистических исследованиях чаще используют другую характеристику разброса, которая сложнее вычисляется, но зато меньше подвержена таким колебаниям. Прежде чем определять эту величину, рассмотрим на примере, какой самый естественный способ вычисления «среднего отклонения от среднего».

Пример 4. Дан числовой ряд, который представляет собой стоимость одного литра бензина на 10-ти автозаправочных станциях (в рублях): 10,2; 9,8; 10; 9,9; 10; 10,5; 9,8; 10; 10,2; 9,8.

Найдем среднее арифметическое этих цен:

$$\frac{10{,}2+9{,}8+10+9{,}9+10+10{,}5+9{,}8+10+10{,}2+9{,}8}{10}=10{,}02.$$

Самым естественным, на первый взгляд, кажется посчитать отклонение от среднего для каждого члена ряда и затем найти их среднее арифметическое:

$$\frac{(10{,}2-10{,}02)+(9{,}8-10{,}02)+(10-10{,}02)+\dots+(9{,}8-10{,}02)}{10}=0.$$

Мы получили нуль совсем не случайно: при вычислении «среднего разброса» по такой формуле часть отклонений входит в сумму со знаком «плюс», часть — со знаком «минус», а в сумме всегда получается нуль.

Какой же выход? Можно суммировать, например, модули отклонений — тогда уж нуля точно не будет. Иногда так и поступают, но с модулем не всегда удобно ра-

ботать. Поэтому математики решили, что лучше складывать не модули отклонений, а их квадраты — они ведь тоже неотрицательные. Так появилось понятие ***дисперсии*** числового ряда.

Дисперсией числового ряда называется среднее арифметическое квадратов отклонений от среднего арифметического.

Найдем дисперсию числового ряда из нашего примера с ценами на бензин. Среднее арифметическое мы уже вычислили — оно равно 10,02. Найдем теперь дисперсию,

т. е. среднее арифметическое квадратов отклонений от среднего:

$$\frac{(10{,}2-10{,}02)^2+(9{,}8-10{,}02)^2+(10-10{,}02)^2+\cdots+(9{,}8-10{,}02)^2}{10}=0{,}0456.$$

Есть другой способ вычисления дисперсии: нужно сначала вычислить среднее арифметическое самих чисел, затем — среднее арифметическое их квадратов, и наконец, ***из среднего арифметического квадратов вычесть квадрат среднего арифметического.***

Проверим справедливость этой формулы на нашем примере:

$$\frac{10{,}2^2+9{,}8^2+10^2+\cdots+9{,}8^2}{10}=100{,}446;\quad 100{,}446-10{,}02^2=0{,}0456.$$

Действительно, мы получили тот же самый результат.

У дисперсии есть один существенный недостаток: если исходные значения ряда измеряются в каких-то единицах (например, в рублях), то у дисперсии эти единицы возводятся в квадрат («квадратные» рубли). В нашем примере среднее значение цены получилось 10 рублей 2 копейки, а вот дисперсия цен — около 4-х … «квадратных копеек».

Избавиться от таких странных единиц измерения можно, если использовать другую характеристику разброса — стандартное отклонение.

Стандартным (или средним квадратичным) отклонением числового ряда называется квадратный корень из дисперсии. Обозначают его греческой буквой **σ** («сигма»).

В рассмотренном примере стандартное отклонение будет $\sigma=\sqrt{0{,}0456}\approx 0{,}213$, т. е. приблизительно 21 коп.

Как и при изучении средних характеристик, попробуем найти характеристики разброса *по таблице частот.* Воспользуемся для этого уже знакомым нам примером 9.1 из предыдущей главы.

Пример 5. Найдем размах, дисперсию и стандартное отклонение отметок ученика из примера 1, заданных следующей частотной таблицей:

Отметка	Абсолютная частота	Относительная частота	Накопленная частота
2	1	0,1	0,1
4	3	0,3	0,4

5	6	0,6	1
ИТОГО	10	1	

Проще всего вычислить размах — он равен разности последнего и первого значений числового ряда (ведь в таблице частот эти значения упорядочены), т. е. 5-2=3.

Дисперсию, как и среднее арифметическое, можно вычислять с использованием либо абсолютных, либо относительных частот. А если вспомнить, что у нас уже есть две формулы для определения дисперсии, получаем целых четыре разных способа вычисления (среднее арифметическое мы уже вычислили в примере 1 — оно равно 4,4):

1-й способ:

$$\frac{(2-4{,}4)^2\,(1+(4-4{,}4)^2\,(3+(5-4{,}4)^2\,(6}{10}=0{,}84.$$

2-й способ:

3-й способ:

$$\frac{2^2\,(1+4^2\,(3+5^2\,(6}{10}=20{,}2;\qquad 20{,}2-4{,}4^2=0{,}84.$$

4-й способ:

Естественно, во всех четырех случаях получаем одинаковый результат: дисперсия равна 0,84. Стандартное отклонение будет $\sqrt{0{,}84}\approx 0{,}92$.

Отметим еще, что если для представления выборки используется *интервальная таблица частот,* то как и при вычислении средних характеристик, ***в качестве значений выборки берут середины интервалов.***

Математическое ожидание случайной величины.

Как мы знаем, распределение вероятностей случайной величины — это таблица, в которой указаны значения случайной величины и их вероятности. Для практики не всегда нужно изучать всю таблицу распределения. Достаточно знать некоторые ее числовые характеристики.

Рассмотрим случайную величину X. Ее математическое ожидание обычно обозначают E(X).

Пусть распределение вероятностей случайной величины X задано таблицей:

Значение величины X	x_1	x_2	x_3		x_n
Вероятность	P_1	P_2	P_3		P_n

Математическим ожиданием случайной величины X называют число

$$E(X)= x_1\cdot P_1 + x_2\cdot P_2 + \ldots + x_n\cdot P_n.$$

Математическое ожидание E(X) называют также ***ожидаемым значением*** случайной величины ***X, средним значением*** случайной величины X.

Если значения случайной величины измеряются в каких-либо единицах (например, рост—в сантиметрах, температура—в градусах), то ее математическое

ожидание измеряется в этих же единицах (средний рост —в сантиметрах, средняя температура — в градусах).

Пример 6. Чему равно математическое ожидание выпадения числа очков игральной кости? Возьмем в качестве случайной величины X число очков, выпавших на одной игральной кости. Вероятности выпадения каждой грани одинаковы и равны 1/6. Поэтому

$$E(X) = 1\cdot \quad + 2\cdot \quad + 3\cdot \quad + 4\cdot \quad + 5\cdot \quad + 6\cdot \quad = \frac{1+2+3+4+5+6}{6} = 3{,}5.$$

Этот пример показывает, что ***если все значения случайной величины равновероятны, то математическое ожидание — это просто среднее арифметическое значений.***

Рассмотрим применение математического ожидания при расчете цены лотерейного билета.

Пример 7. Для проведения лотереи изготовили 100 билетов. Из них 1 билет с выигрышем в 500 р., 10 билетов с выигрышами по 100 р. и остальные 89 билетов без выигрышей. Наудачу выбирают один билет. Найдем математическое ожидание выигрыша.

Эта случайная величина может принимать три значения: 500 р., 100 р. и 0 р. (нет выигрыша). Их вероятности равны 0,01, 0,10 и 0,89.

Математическое ожидание выигрыша равно $500\cdot 0{,}01 + 100\cdot 0{,}10 + 0\cdot 0{,}89 = 15$ (р.). Получается, что средний выигрыш на один билет равен 15 р.

Для того чтобы лотерея приносила доход своим устроителям, цена билета должна быть больше, чем средний выигрыш. Предположим, что билет стоит 20 р. Продав все билеты, устроители лотереи получат 2000 рублей. На выплату выигрышей будет потрачено 1500 рублей. Таким образом, доход от лотереи составит 500 рублей.

Разумеется, может случиться так, что на один купленный нами билет мы получим большой выигрыш. Но если бы некто решил купить все билеты, то он достоверно потерял бы 500 рублей —по 5 на каждый из 100 билетов.

Так устроены все лотереи: математическое ожидание выигрыша на один билет меньше цены этого билета. Это условие является непременным, и оно обеспечивает рентабельность лотереи и доход ее устроителям. Человек, который решил сыграть в лотерею, должен понимать это и сознательно рисковать своими деньгами.

Важное свойство математического ожидания:

Математическое ожидание суммы случайных величин равно сумме математических ожиданий этих величин.

Рассмотрим применение этого свойства на примере:

Пример 8. Чему равно математическое ожидание суммы очков, при бросании двух игральных костей?

Пусть U – число очков, выпавших на первой кости, а V- число очков, выпавших на второй кости.

Из примера 6 нам известно, что E(U) = E(V) = 3,5. Согласно свойству мы получим:

E(U+V) = E(U) + E(V) = 3,5 + 3,5 = 7.

Пример 9. Чему равна в среднем сумма очков при десяти бросаниях игральной кости?

Пусть X_1, X_2, …, X_{10} – число очков на кости во время каждого бросания, S – их сумма, тогда, используя данные предыдущих примеров, получим $E(S) = E(X_1) + … + E(X_{10}) = 3,5 \cdot 10 = 35$.