Здравствуйте!

Слово «статистика» происходит от латинского status (состояние, положение вещей)

Числовые данные о массовых явлениях получаются в результате наблюдения за **совокупностями** тех или иных явлений и измерения значений наблюдаемых **признаков** - свойств, которыми обладают явления и объекты, входящие в совокупность.

Особенностью статистических исследований является то, что статистические совокупности, как правило, состоят из огромного числа единиц. Это делает практически невозможным сплошное обследование совокупности. Возникает задача: как выявить важнейшие закономерности, присущие всей статистической совокупности, не обследуя каждый элемент этой совокупности, а только какую-то часть этих элементов?

Основной задачей статистики является выявление и исследование общих закономерностей, присущих совокупностям, состоящим из очень большого числа элементов.

Основным методом статистики является **выборочный метод,** суть которого заключается в следующем.

Из совокупности элементов выбирают некоторое количество элементов для обследования. Выбранные элементы составляют **выборку;** количество элементов в выборке называют **объемом** выборки. Совокупность, из которой сделана выборка, называют **генеральной совокупностью.**

Может быть два вариант отбора элементов: элемент может быть возвращен в генеральную совокупность, такая выборка называется повторной или он не возвращается обратно в выборку, такая выборка называется бесповторной.

При исследовании необходимо, чтобы выборка правильно представляла все свойства генеральной совокупности, то есть, чтобы свойства элементов выборки соответствовали свойствам всей генеральной совокупности. Такая выборка называется **репрезентативной** (представительной).

Далее изучают только выборку, находят ее характеристики, выявляют в ней закономерности, проверяют различные гипотезы о свойствах совокупности.

Источником статистической информации является реальный опыт, наблюдение, измерение, производимые над реальными объектами и явлениями окружающего нас мира. Этим статистика отличается от теории вероятностей, которая изучает математические модели реальных явлений.

Статистическая информация о результатах наблюдений или экспериментов может быть представлена в различных формах. Простейшей из них является запись результатов в порядке их появления (или получения) - запись в ряд: $x_1, x_2, x_3, ..., x_i, ..., x_{n-1}, x_n$, называемый **простым статистическим рядом,** или рядом данных, или выборкой. Отдельные значения x_i, составляющие этот ряд, называют **вариантами** или просто данными, или результатами наблюдений. Количество вариант в ряду n называют **объемом ряда,** или объемом выборки.

Варианты в ряду могут иметь как различные, так и одинаковые, повторяющиеся, значения. Например, игральный кубик бросили 12 раз и записали выпавшие числа в порядке их появления:

3, 4, 5, 6, 6, 6, 5, 1, 4, 6, 1,4 (n = 12).

Вариантами в ряду являются $x_1 = 3$, $x_2 = 4$, $x_3 = 5$, $x_4 = 6$, $x_5 = 6$ и т. д. Варианты x_4, x_5, x_6, x_{10} имеют одинаковые значения (6), но это разные варианты.

Запись статистической информации в форме простого ряда имеет <u>два наиболее существенных недостатка</u>:

1) громоздкость

2) труднообозримость

Второй недостаток устраняют простейшей обработкой ряда: упорядочивают ряд, располагают варианты в порядке их возрастания:

1, 1, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6.

Полученный ряд называют **вариационным рядом,** или просто упорядоченным рядом.

Первый недостаток (громоздкость) теперь тоже легко устраняется: будем записывать только значения встречающихся вариант (по одному разу), а под каждым значением будем писать число, показывающее, сколько раз это значение встречается в ряду; получим запись:

x_j	1	3	4	5	6
n_j	2	1	3	2	4

Такая «свернутая» запись статистических данных (она обычно оформляется в виде таблицы), называется **статистическим распределением ряда;** величины n_j, называются **частотами** значений варианты x_j.

Наряду с частотами n_j - широко используются **относительные частоты** $w_j = \frac{n_j}{n}$.

«Свернуть» длинный ряд в компактную таблицу распределения оказывается возможным только в тех случаях, когда наблюдаемый признак имеет небольшое число различных (дискретных) значений вариант. Если же значений вариант очень много и они редко повторяются, то для «свертывания» ряда строят так называемый **интервальный ряд:** весь диапазон наблюдаемых значений признака $x_{\max} - x_{\min}$, разбивают на небольшое число ($к$ = 6... 10) частичных интервалов, и подсчитывают количество вариант исходного ряда, попадающих в каждый частичный интервал; эти числа $n'_j \left(j = 1,2,...,k\right)$ и принимают за частоты соответствующих интервалов.

Чтобы сравнить между собой две или несколько совокупностей статистических данных, нужны показатели, характеризующие то или иное свойство совокупности данных одним числом. Такие показатели в статистике получили наименование **числовых характеристики**.

Простейшими числовыми характеристиками являются характеристики положения (среднее значение, мода, медиана) и характеристики рассеивания (размах, выборочная дисперсия, выборочное среднее квадратичное отклонение).

Среднее значение ряда наблюдений $\overline{X}$ - это центр рассеивания наблюдаемых значений, это расчетное значение, сумма отклонений всех вариант от которого равна нулю. Если варианты в ряду x_i, являются значениями непосредственно наблюдаемого (первичного) признака, то среднее значение ряда $\overline{X}$ находят по формулам среднего арифметического:

$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ (формула простой средней), $\overline{X} = \frac{\sum_{j=1}^{m} x_j \cdot n_j}{\sum_{j=1}^{m} n_j}$ (формула средней взвешенной)

В статистике при вычислении средних ставится задача заменить все индивидуальные наблюдаемые значения признака некоторой обобщающей уравненной величиной $\overline{X}$ так, чтобы при этом не изменялась **некоторая итоговая величина** для всей совокупности. Этой величиной может быть сумма всех вариант (среднее арифметическое), или их произведение (среднее геометрическое), или сумма обратных величин (среднее гармоническое), или сумма квадратов вариант (среднее квадратичное) и т. д.

Общая формула степенной средней: $\overline{X}_{ср.ст.} = \left(\frac{\sum_{j=1}^{m} x_i^k}{n} \right)^{\frac{1}{k}}$, при $k = -1$ получаем среднюю гармоническую, при $k = 1$ – среднюю арифметическую, при $k = 2$ – среднюю квадратичную, и т. д. Отдельно вводится понятие среднего геометрического $\overline{X}_{геом.} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$

Правило мажорантности средних:

$$\overline{X}_{гарм.} \le \overline{X}_{геом.} \le \overline{X}_{арифм.} \le \overline{X}_{квадр.}$$

Числовых характеристик в статистике отличаются от числовых характеристик в теории вероятностей.

В статистике числовые характеристики являются функциями результатов наблюдений, по которым они вычисляются. Бросим игральный кубик 12 раз, найдем среднее арифметическое выпавших чисел, получим значение среднего $\overline{X_1}$. Бросим кубик еще 12 раз, найдем уже другое среднее $\overline{X_2}$, в третий раз - опять другое $\overline{X_3}$ и т. д. Таким образом, средние значения, вычисляемые по выборкам конечного объема, будут колебаться, они сами будут случайными величинами. В статистике эти величины $\overline{X_1}, \overline{X_2}, \overline{X_3}$ ... рассматриваются как точечные оценки неизвестного среднего всей генеральной совокупности $\overline{X}$ (в эксперименте с бросанием игрального кубика под генеральной совокупностью понимают совокупность результатов бесконечного числа бросаний кубика). Таким образом, опытным путем мы можем найти только приближенное значение $\overline{X}$ для всей генеральной совокупности.

Но имеет место замечательная закономерность: чем больше объем выборки n (то есть чем больше мы проводим испытаний), тем меньше точечные оценки $\overline{X_n}$ отличаются одна от другой, тем больше они стабилизируются около искомого значения $\overline{X}$ среднего для всей генеральной совокупности. Нужно только иметь в виду, что такая стабилизация и «сходимость» к $\overline{X}$ имеют место при соблюдении ряда дополнительных условий, сформулированных в статистической науке. Точно так же ведут себя и другие числовые характеристики. Это является одним из проявлений «закона больших чисел».

Закон больших чисел – название собирательное. Так называют математические теоремы, которые при разных условиях утверждают, что среднее арифметическое, составленное из большого числа случайных слагаемых, мало отличается от математического ожидания этого среднего арифметического.

Пусть $X_1, X_2, \ldots, X_n$ – независимые случайные величины, имеющие одинаковое распределение и пусть a – общее для всех них математическое ожидание.

Тогда, при достаточно больших n выполняется приближенное равенство

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \approx a.$$

Это приближенное равенство тем точнее, чем больше n.

Мода Mo - это значение вариант, встречающееся в ряду чаще других. В таблице распределения ряда мода - это значение x_j, которому соответствует наибольшее значение частоты n_j.

Статистический ряд может иметь одну, две или несколько мод, может не иметь моды.

Медиана Me - это срединное в вариационном ряду значение варианты.

Медиана ряда, состоящего из **нечетного** количества чисел, называется число данного ряда, которое окажется посередине, если этот ряд упорядочить.

Медиана ряда, состоящего из **четного** количества чисел, называется среднее арифметическое двух стоящих посередине чисел этого ряда.

(Для того, чтобы найти медиану ряда, нужно сначала упорядочить элементы ряда).

Если число членов ряда n нечетное, то $Me = x_{\left[\frac{n}{2}\right]+1}$, где $\left[\frac{n}{2}\right]$ – целая часть числа $\frac{n}{2}$.

Если n четное, то $Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$.

Простейшей характеристикой рассеивания является размах: $A = x_{\max} - x_{\min}$,

размах есть разность между наибольшим и наименьшим значениями вариант в ряду.

Задача 1. В таблице приведены расходы за 4 дня:

День	Понедельник	Вторник	Среда	Четверг
Расходы (руб.)	18	25	24	25

Определить, какая статистическая характеристика находится в каждом задании:

а) 18+24+25*2=92, 92:4=23 (руб.)

б) (24+25):2 =24,5 (руб.) в) 25 (руб.)

г) 25-18=**7** (руб.)

Задача 2. На предприятии работают 100 человек: 99 служащих, каждый из которых получает по 1000 рублей в месяц и директор, получающий 100000 рублей в месяц. Служащие потребовали повысить зарплату, так как практически все работники предприятия получают по 1000 рублей. Однако директор отказал им, заявив, что средняя зарплата на предприятии и так составляет около 2000 рублей. С точки зрения статистики кто прав, директор или служащие?

Задача 3.

Как могут измениться размах и мода ряда чисел, если:

а) дополнить его числом, превосходящим все остальные;

б) вычеркнуть из него число, меньшее всех остальных;

в) дополнить его числом, равным наибольшему из чисел?

Задача 4.

В ряду чисел 8, 16, 26, __, 48, __, 46 два числа оказались стертыми. Найдите эти числа, если известно, что одно из них на 20 больше другого, а среднее арифметическое этого ряда чисел равно 32.

№	Генеральная совокупность	Цель обследования	Выборка
1.	Партия одинаковых деталей объемом 10000 штук	Определение числа бракованных изделий в партии	1) 5 рядом лежащих деталей; 2) 5 деталей, выбранных случайным образом из разных частей партии; 3) 100 деталей, выбранных случайным образом из разных частей партии
2.	Все бездомные собаки города Томска	Определение числа собак, больных чумкой	1) Одна собачья стая; 2) по нескольку случайным образом отловленных собак из каждого района города
3.	Все экзаменационные работы в форме единого теста по математике выпускников школ Томской области	Предварительное выявление соотношения между отличными, хорошими, удовлетворительными и плохими знаниями по математике	1) 5 работ, изъятых случайным образом из числа всех работ; 2) 50 работ, изъятых случайным образом из числа всех работ; 3) 50 работ выпускников одной школы
4.	Партия штампованных деталей объемом 100000 штук	Определение среднего веса детали в партии	1) 2 детали; 2) 100 деталей, отштампованных последними; 3) 50 случайным образом выбранных деталей из партии.
5.	Бидон молока	Определение процента жирности	1) Ложка молока, взятая с поверхности по прошествии 2 ч после надоя; 2) стакан молока, вылитый из бидона посл 2 ч стояния его в погребе; 3) ложка молока после тщательного перемешивания его в бидоне.

№	Цель обследования	Выборка
1.	Выявление читательских интересов	1. Дети старшей группы детского сада. 2. Курсанты роты военного училища. 3. Члены одной семьи.
2.	Выявление любимых мелодий (песен)	1. 100 учащихся музыкальной школы 2. 100 человек, случайным образом остановленных и опрошенных поздно вечером на улице города.
3.	Определение числа больных гриппом в городе в пик эпидемии	1. 100 случайным образом выбранных пациентов терапевтических кабинетов поликлиник города. 2. Жильцы одного подъезда двухэтажного дома.
4.	Определение среднего уровня доходов населения	1. Жильцы одного подъезда пятиэтажного дома 2. 300 случайным образом выбранных жильцов студенческого общежития. 3. Все жители коттеджного района города.
5.	Определение наиболее ходовых размеров джинсов	1. Все студенты хореографического училища. 2. Члены секции сумо.
6.	Определение количества домашних кошек и собак, приходящегося на душу населения в городе	1. Жильцы одного подъезда двухэтажного дома 2. Жильцы многоквартирного дома, заселенного одинокими престарелыми людьми

Задача 7.

Среди случайным образом выбранных 100 молодых людей, носящих летом кепки, провели опрос о цветовых предпочтениях для этого вида головных уборов. Результаты опроса отражены в таблице:

Цвет	Черный	Красный	Синий	Серый	Белый	Желтый	Зеленый
Частота	32	20	16	14	11	5	2

Считая выборку репрезентативной, высказать рекомендации швейной фабрике по количеству выпускаемых кепок каждого цвета, если фабрика должна подготовить к продаже 30000 кепок.

Решение

По определению репрезентативной выборки $\frac{M_i}{N} = \frac{S_i}{S}$.

В нашем случае S = 30 000, N=100, M_i - во второй строке таблицы.

$$S_i = \frac{M_i \cdot S}{N} = M_i \cdot 300.$$

Цвет	Количество кепок
черный	32*300=9600
красный	20*300=6000
синий	16*300=4800
серый	14*300=4200
белый	11*300=3300
желтый	5*300=1500
зеленый	2*300=600

Задача 8.

Пусть на экзамене по математике три класса получили

Наименование показателя	Класс А	Класс Б	Класс В
Средний балл на экзамене	4,0	4,0	4,0
Количество оценок, полученных классом на экзамене, в том числе:			
2 (не удовлетв.)	0	1	7
3 (удовлетв.)	0	4	2
4 (хорошо)	25	14	0
5 (отлично)	0	6	10

Что стоит за разбросом оценок и как оценить сам этот разброс?

Средний балл – это в данном случае средняя арифметическая величина:

$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum x}{n}$, где $x_1, x_2, \ldots, x_n$ - оценки учащихся в баллах, а n - количество учеников в классе.

Мы получили величину среднего балла по всем классам 4,0. Но что за этим стоит, ведь разброс оценок в каждом из классов оказался различным?

Для анализа разброса оценок используют дополнительные приемы обработки.

Для того, чтобы дать обобщающую характеристику распределению отклонений, исчисляют среднее линейное отклонение:

$$d = \frac{\sum \left| x - \bar{x} \right|}{n} = \frac{\left| x_1 - \bar{x} \right| + \left| x_2 - \bar{x} \right| + \ldots + \left| x_n - \bar{x} \right|}{n}$$

Но этого для анализа мало. Основными обобщающими показателями вариации в статистике являются дисперсии и среднее квадратическое отклонение

Дисперсия (средний квадрат отклонений) $S^2 = \frac{\sum \left(x_i - \bar{x} \right)^2}{n}$

Чтобы понизить размерность данных, на практике чаще всего используют среднее квадратичное отклонение представляет собой корень квадратный из дисперсии

$$S = \sqrt{\frac{\sum \left(x_i - \bar{x} \right)^2}{n}}$$

Чем меньше среднее квадратическое отклонение, тем более ровно по успеваемости учится класс, нет резкой поляризации на «двоечников» и «отличников». С другой стороны, если величина среднего квадратического отклонения близка к 0, это тоже сигнал – отсутствие каких-либо различий в уровне успеваемости обучающихся может свидетельствовать, например, о необъективности оценок. Высокое значение среднего квадратического тоже необходимо анализировать – класс разделился на успевающих и не успевающих, в классе много «запущенных» детей.

Для определения критичности ситуации используется коэффициент вариации:

$$V = \frac{S}{\bar{x}} * 100\%$$

Это наиболее распространенный индикатор колеблемости типичных средних величин. При этом, если он больше 40% - это говорит о чрезмерно большой колеблемости признака.

Получаем следующие данные

Наименование статистических операторов	Класс А	Класс Б	Класс В
Среднее арифметическое (средний балл)	4,0	4,0	4,0
Среднее линейное отклонение (d)	0,0	0,5	1,3
Дисперсия s^2	0,0	0,6	1,8
Среднее квадратическое отклонение (S)	0,0	0,75	1,35
Коэффициент вариации (V)	0,0%	18,8%	33,8%

Анализируем.

В классе А все значения операторов, характеризующих разброс величины полученных оценок от среднего балла, равны 0. Это означает, что все учащиеся класса показали на экзамене абсолютно одинаковую успеваемость. Даже теоретически такое трудно допустить. Скорее всего, это свидетельствует о необъективности выставленных оценок.

В классе Б коэффициент вариации и другие операторы больше 0, но значения их невелики и скорее всего отражают реальное положение.

В классе В коэффициент вариации приближается к критической величине. Высокие значения имеют и другие операторы. Это говорит о крайней неравномерности полученных оценок учащимися, и, как следствие, о просчетах самого учителя, преподававшего математику в этом классе. По-видимому, процесс обучения был пущен на самотек, части детей не уделялось должного внимания, и, как следствие, они не получили необходимых знаний. За высоким средним баллом скрывалось резкое разделение учащихся на «математическую элиту» и элементарно «запущенных» детей.