Математическая статистика.

Установление закономерностей, которым подчинены массовые случайные явления, основано на изучении статистических данных – результатах наблюдений.

Первая задача математической статистики – указать способы сбора и группировки статистических сведений.

Вторая задача математической статистики – разработать методы анализа статистических данных, в зависимости от целей исследования.

Т.е. ***задача математической статистики*** состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

Типы статистических данных.

Статистические данные представляют собой наблюдаемые или измеряемые значение одного или нескольких признаков обследуемой совокупности объектов. Различают качественные и количественные признаки. Количественные признаки могут быть непрерывными (вес, рост, цена) или дискретными (кол-во детей, комнат, продажа т/в в день).

Качественные признаки: пол, семейное положение, цвет глаз, кирпичный дом. Качественные признаки: номинальные (классификационные) и ординальные (порядковые). Говорят, что соответственные признаки измеряют в номинальной или порядковой шкале.

Признак измеряемый в номинальной шкале принимает одно значение из конечного числа установленных градаций: пол (м/ж), цвет, марка автомобиля и прочие. Часто пользуются, например, в социологических опросах. Таблицы данных записываются в виде таблиц сопряженности.

Значения качественных признаков, измеряющихся в ординальной шкале, могу быть упорядочены. Примеры: тестовые баллы, школьные оценки, качество условий жизни (очень хорошее, хорошее, удовлетворительное, плохое, российское). Для представления таких признаков используется ранг, т.е. число, которое сопоставляет номеру признака в ряду.

Для обработки данных, представленных в ординальной и номинальных шкалах, разработаны специальные методы, например, ранговая корреляция, можно проводить проверку гипотез о виде распределений, дисперсионный анализ и пр.

Представление данных в виде гистограмм, полигонов и эмпирических функций распределения дает информацию о распределении генеральной совокупности. Но часто требуется охарактеризовать генеральную совокупность по количественным показателям, которые определяют положение центра распределения, рассеяние (разброс) данных, ассиметрию. Это дает возможность сравнить одну совокупность данных с другой.

Генеральная совокупность и выборка.

Пример 1. При проверке качества производства электроламп последние должны находиться под напряжением довольно большое время, что естественно, невозможно в условиях массового производства. Поэтому для проверки на стандартность подвергают контролю только небольшую часть изготовленных ламп. Практика подтверждает, что выводы о всей совокупности объектов, сделанные на основании анализа данных наблюдения только над заведомо меньшей частью этой совокупности, бывают достаточно надежными.

Зачастую реально существующую совокупность объектов можно мысленно дополнить любым количеством таких же однородных объектов. Например, совокупность электромоторов определенной марки, изготовленных на данном заводе в течение квартала, можно дополнить гипотетической совокупностью же электромоторов, которые могут быть изготовлены II, в III и т. д. кварталах. В соответствии с этим наблюдения над объектами такой совокупности, в результате которых «снимаются» конкретные значения случайной величины (значения изучаемого признака объекта), можно мысленно продолжать в неизменных условиях как угодно долго.

Такие совокупности объектов или совокупности, соответствующие каждому из этих объектов значений определенной случайной величины, будем называть *генеральными.*

Определение. Совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины, или совокупность результатов всех мыслимых наблюдений, проводимых в неизменных условиях над одной из случайных величин, связанных с данным видом объектов, называется ***генеральной совокупностью***.

Как видно из определения, генеральная совокупность объектов данного вида и соответствующая совокупность значений случайной величины не различаются.

Генеральную совокупность будем называть *конечной* или *бесконечной* в зависимости от того, конечна или бесконечна совокупность составляющих ее элементов. Если множество значений случайной величины X бесконечно, то генеральная совокупность бесконечна. Если случайная величина дискретна ее множество значений конечно, то генеральная совокупность может быть как конечной (например, по статистическим данным оценивается доля мальчиков среди детей, родившихся за год; здесь генеральная совокупность — это все родившиеся за год дети), так и бесконечной (если рассматривать до бесконечности непрерывное воспроизводство населения).

Определение. Часть отобранных объектов из генеральной совокупности (результаты наблюдений над ограниченным числом объектов из этой совокупности) называются ***выборочной совокупностью*** или ***выборкой.***

Другими словами можно сказать, что выборка это совокупность наблюдений над случайной величиной.

Определение. Рассмотрим случайный эксперимент, связанный со случайной величиной X, имеющей функцию распределения $F_X(x)$. ***Выборкой*** объема n из генеральной совокупности с функцией распределения $F_X(x)$ называется последовательность $x_1, x_2,\ldots,x_n$ наблюдаемых значений случайной величины X, соответствующих n независимым повторениям эксперимента.

Число N объектов генеральной совокупности и числе n объектов выборочной совокупности будем называть объемами генеральной и выборочной совокупностей соответственно. При этом будем предполагать, что $N >> n$ (N значительно больше n). Как уже отмечалось выше, о свойствах генеральной совокупности (случайной величины X) можно судить по данным наблюдений над отобранными объектами, т. е. по выборке Однако не всякая выборка может быть действительным представлением о генеральной совокупности.

Пример 2. В цехе по производству специальных втулок на токарных станках работают квалифицированные токари и только начинающие. Для проверки качества продукции на контроль взята партия втулок. Если втулки изготовлены квалифицированным токарем, то, очевидно, представление о качестве всей продукции цеха будет «завышенным», а если втулки изготовлены начинающим токарем, то это представление будет «заниженным».

Для того чтобы по выборке можно было достаточно уверенно судить о случайной величине, выборка должна быть ***представительной (репрезентативной)***. Репрезентативность выборки означает, что объекты выборки достаточно хорошо представляют генеральную совокупность. Репрезентативность выборки обеспечивается случайностью отбора. Последнее означает, что любой объект выбора отобран случайно, при этом все объекты имеют одинаковую вероятность попасть в выборку. Может производиться случайная выборка с возвратом или случайная выборка без возврата. Если объем генеральной совокупности велик, то различие между выборками с возвратом и без возврата, практически не сказывается на окончательных результатах. Если объем генеральной совокупности невелик, то различие между двумя типами выборок может быть существенным. Поскольку условие независимости испытаний является существенным в дальнейшем будем предполагать, что имеет место случайная выборка с возвратом.

После того как сделана выборка, т.е. получена выборочная совокупность объектов, все объекты этой совокупности обследуют по отношению к определенной случайной величине (или

случайному событию) и в результате этого получают наблюдаемые данные. Следующая задача математической статистики заключается в обработке результатов наблюдений.

Способы записи выборки.

Полученные наблюдаемые данные представляют собой расположенные в беспорядке числа. Для изучения закономерностей опытные данные подвергают обработке.

При рассмотрении вопроса о данных эксперимента следует различать вопросы, связанные с представлением динамических рядов, и группировкой тех данных, для которых пространственные и временные факторы не играют существенной роли. Пока мы не будем заниматься анализом динамических рядов. Кроме того, следует иметь в виду, что данные могут быть разнотипны, то есть представлены в количественной или порядковых шкалах, и могут быть как одномерные, так и многомерные. Большинство обсуждаемых далее методов относятся к работе с одномерными количественными данными.

Определение 1. Операция, заключающаяся в том, что результаты наблюдений над случайной величиной, то есть наблюдаемые значения случайной величины (с.в.), располагают в порядке неубывания, называется ***ранжированием*** опытных данных**.**

Определение 2. ***Вариационным рядом*** выборки $x_1, x_2, \ldots, x_n$ называется способ ее записи, когда элементы упорядочиваются по величине $x^{(1)} \le x^{(2)} \le \ldots \le x^{(n)}$.

Определение 3 Значение с.в. соответствующее отдельной группе сгруппированного ряда наблюдаемых данных, называется ***вариантом,*** а изменение этого значения – ***варьированием.***

Определение 4. Численность отдельной группы сгруппированного ряда наблюдаемых данных называется ***частотой*** или ***весом*** соответствующего варианта и обозначается m_i (или n_i), где i - индекс варианта.

Очевидно, что $\sum_i m_i = \sum_i n_i = n$. ($n$ - объем выборки).

Определение 5. Отношение частоты данного варианта к общей сумме частом всех вариантов называется ***частостью*** или долей этого варианта и обозначается v_i $\hat{p}_i$, где i - индекс варианта. (Можно также встретить обозначение w_i - относительная частота.)

$v_i = \dfrac{m_i}{\sum_{i=1}^{v} m_i}$, где v - число вариантов. Так как $\sum_i m_i = n$, то $\hat{p}_i = \dfrac{m_i}{n}$ $v_i = \dfrac{m_i}{n}$.

Можно считать частость $\hat{p}_i$ выборочным аналогом вероятности p_i появления значения x_i случайной величины X.

Определение 6. ***Статистическим рядом*** называется последовательность пар (x_i, m_i). Записывается статистический ряд в виде таблицы: первая строка – элементы x_i, а вторая – частоты m_i.

Пример 1. Записать в виде вариационного и статистического ряда выборку 5, 3, 7, 10, 5, 5, 2, 10, 7, 2, 7, 7, 4, 2, 4.

Объем выборки n = 15. Упорядочив элементы выборки по величине, получим вариационный ряд

2, 2, 2, 3, 4, 4, 5, 5, 5, 7, 7, 7, 7, 10, 10.

Статистический ряд запишем в виде таблицы.

x_i	2	3	4	5	7	10
m_i	3	1	2	3	4	2

Можно в статистический ряд добавить строку с частостями.

Если изучаемая с.в. является непрерывной или число значений дискретной с.в. велико, то ранжирование и группировка наблюдаемых значений зачастую не позволяет выявить характерные черты варьирования ее значений. В подобных случаях следует построить интервальный (вариационный) ряд распределения. Для построения такого ряда весь интервал

варьирования наблюдаемых значений случайной величины разбивают на ряд частичных интервалов и подсчитывают частоту попадания значений в каждый частичный интервал.

Определение 7. Интервальным вариационным рядом называется упорядоченная совокупность интервалов варьирования значений случайной величины с соответствующими частотами или частостями попаданий к каждый из них значений величины.

Чтобы построить интервальный ряд необходимо установить верхнюю и нижнюю границы, то есть наибольшее и наименьшее значение с. в. Число интервалов варьирования обычно выбирают от 7 до 11.

Найдем ширину интервала варьирования $h = \frac{x_{\max} - x_{\min}}{\nu}$, где $x_{\min}, x_{\max}$ - минимальное и максимальное значение измеряемой случайной величины, ν - число интервалов. Для более точного определения величины частичного интервала можно воспользоваться формулой Стерджеса. $h = (x_{\max} - x_{\min})/(1 + 3{,}322 \lg n)$ Если h - дробное число, то за длину частичного интервала следует брать либо ближайшее целое число, либо ближайшую простую дробь. За начало первого интервала рекомендуется брать величину $x_{нач} = x_{min} - 0{,}5h$, конец последнего интервала $x_{кон}$ должен удовлетворять условию $x_{кон} - h \le x_{max} < x_{кон}$.

Пример 2. При измерении диаметра валиков после шлифовки получены следующие результаты

6,75; 6,77; 6,77; 6,73; 6,76; 6,74; 6,70; 6,75; 6,71; 6,72; 6,77; 6,79; 6,71; 6,78;
6,73; 6,70; 6,73; 6,77; 6.75; 6,74; 6,71; 6,70; 6,78; 6,76; 6,81; 6,69; 6,80; 6,80;
6,77; 6,68; 6,74; 6,70; 6.70; 6.74; 6,77; 6,83; 6,76; 6,76: 6,82; 6,77; 6,71; 6,74;
6,77; 6,75; 6,74; 6,75; 6,77; 6,72; 6,74; 6,80; 6,75; 6,80; 6,72; 6,78; 6,70: 6,75;
6,78; 6,78; 6,76; 6,77; 6,74; 6,74; 6,77; 6,73; 6,74; 6,77; 6,74; 6,75; 6,74; 6,76;
6,76; 6,74; 6,74; 6,74; 6,74; 6,76; 6,74; 6,72; 6,80; 6,76; 6,78; 6,73; 6,70: 6,76;
6,76; 6,77; 6,75; 6,78; 6,72; 6,76; 6,78; 6,68; 6,75; 6,73; 6,82; 6,73; 6,80; 6,81;
6,71; 6,82; 6,77; 6,80; 6,80; 6,70; 6,70; 6,82; 6,72; 6,69; 6,73; 6,76; 6,74: 6,77;
6,72; 6,76; 6,78; 6,78; 6,73; 6,76; 6,80; 6,76; 6,72; 6,76; 6,76; 6,70; 6,73: 6,75;
6,77; 6,77; 6,70; 6,81; 6,74; 6,73; 6,77; 6,74; 6,78; 6,69; 6,74; 6,71; 6,76; 6,76;
6,77; 6,70; 6,81; 6,74; 6,74; 6,77; 6,75; 6,80; 6,74; 6,76; 6,77; 6,77; 6,81; 6,75;
6,78; 6,73; 6,76; 6,76; 6,76; 6,77; 6,76; 6,80; 6,77; 6,74; 6,77; 6,72; 6,75; 6,76;
6,77; 6,81; 6,76; 6,76; 6,76; 6,80; 6,74; 6,80; 6,74; 6,73; 6,75; 6,77; 6,74; 6,76;
6,77; 6,77; 6,75; 6,76; 6,74; 6,82; 6,76; 6,73; 6,74; 6,75; 6,76; 6,72; 6,78; 6,72;
6,76; 6,77; 6,75; 6,78.

Для длины частичного интервала имеем:

$h = (6{,}83 - 6{,}68)/(1 + 3.322\, lg\, 200) \approx 0.0174 \approx 0.02$

$h = (6{,}83 - 6{,}68)/(1 + 3{,}322 \lg 200) \approx 0{,}0174 \approx 0{,}02$.

При этом $x_{нач} = 6{,}67$.

Просматривая результаты наблюдений, определяем, сколько значений признака попало в каждый конкретный интервал. При этом в интервал включают значения случайной величины большие или равные нижней границе и меньшие верхней границы. Результаты заносятся в таблицу.

№ п/п	Диаметр валика после шлифовки (интервалы), мм	Частота m_i	Частость $\hat{p}_i$
1	6,67 - 6,69	2	0,01
2	6,69 – 6,71	15	0,075
3	6,71 – 6,73	17	0,085
4	6,73 - 6,75	44	0,22
5	6,75 – 6,77	52	0,26
6	6,77 - 6,79	44	0,22
7	6,79 – 6,81	14	0,07

8	6,81 – 6,83	11	0,055
9	6,83 – 6,85	1	0,005

Иногда интервальный вариационный ряд для простоты исследований условно заменяют дискретным. В этом случае серединное значение i-того интервала принимают за вариант x_i, а соответствующую интервальную частоту m_i - за частоту этого варианта.

Выборочные аналоги интегральной и дифференциальной функции распределения. Полигон и гистограмма.

Представив экспериментальный материал в виде таблиц вариационных рядов мы записали результаты наблюдений в упорядоченном виде, однако человек лучше ориентируется в графическом представлении данных, чем в цифрах, поэтому желательно представить обработанные данные графически. Существует несколько типов графических изображений данных, содержащихся в вариационных рядах. Важнейшими из них являются *эмпирическая функция распределения (кривая накопленных частот)*, *полигон* и *гистограмма.*

Пусть имеется выборочная совокупность значений случайной величины X объема n и каждому варианту из этой совокупности поставлена в соответствие его частость. Пусть далее x - некоторое действительное число, а m_x - число выборочных значений с.в. X меньших x. Тогда $\frac{m_x}{n}$ является частостью наблюдаемых в выборке значений величины X меньших x, то есть частость появления $X < x$. При изменении x будет меняться и $\frac{m_x}{n}$, то есть эта величина функция аргумента x. А так как функция находится на основании выборочных данных, полученных в результате опытов, то ее называют ***выборочной*** или ***эмпирической***.

Определение. Выборочной функцией распределения (или функцией распределения выборки) называется функция $F^*(x)$, задающая для каждого значения x, относительную частоту события $X < x$.

Теоретическая функция распределения $F(x)$ и эмпирическая функция распределения $F^*(x)$ обладают одинаковыми свойствами.

1. $0 \le F^*(x) \le 1$,
2. $\hat{F}(x)$ - неубывающая,
3. $F^*(-\infty) = 0$; $F^*(+\infty) = 1$.

Пример 1. Найти эмпирическую функцию распределения по данному распределению выборки:

x_i	1	4	6
m_i	10	15	25

Найдем объем выборки $n = 10 + 15 + 25 = 50$.

Наименьшая варианта равна 1, поэтому при $x \le 1$, $F^*(x) = 0$.

Значение $X < 4$, а именно $x_1 = 1$, наблюдалось 10 раз, то есть $F^*(x) = \frac{m_x}{n} = \frac{10}{50} = 0{,}2$, при $1 < x \le 4$.

Значение $X < 6$, а именно $x_1 = 1$ и $x_2 = 4$ наблюдалось $10 + 15 = 25$ раз, то есть $F^*(x) = \frac{m_x}{n} = \frac{25}{50} = 0{,}5$, при $4 < x \le 6$.

Так как $x = 6$ - наибольшая варианта, то $\hat{F}(x) = 1$ при $x > 6$.

Запишем искомую эмпирическую функцию: $\hat{F}(x)=\begin{cases}0; & x\le 1\\ 0{,}2; & 1<x\le 4\\ 0{,}5; & 4<x\le 6\\ 1; & x>6\end{cases}$.

Осталось начертить график.

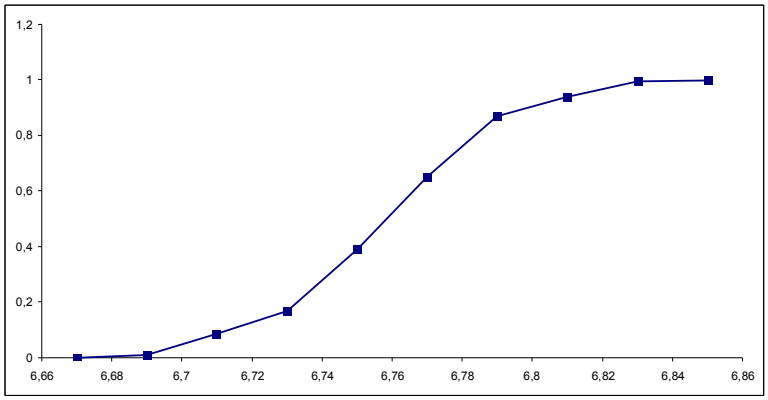
Если записан интервальный вариационный ряд, то выборочную функцию распределения построить аналогично невозможно. В этом случае функцию строят по точкам.

Пример 2. Рассмотрим интервальный вариационный ряд, построенный на основании данных примера 2 предыдущего раздела. На основании этого ряда построим эмпирическую функцию распределения.

Очевидно, что при $x\in(-\infty;6{,}67]$ $\hat{F}(x)=0$. Рассуждая аналогично предыдущему примеру запишем получаемые для $\hat{F}(x)$ величины в таблицу.

x	6,67	6,69	6,71	6,73	6,75	6,77	6,79	6,81	6,83	6,85
m_x	0	2	17	34	78	130	154	188	199	200
$\hat{F}(x)$	0	0,01	0,085	0,17	0,39	0,65	0,87	0,94	0,995	1

Так как эта таблица определяет функцию $\hat{F}(x)$ не полностью, то при графическом изображении данной функции целесообразно ее доопределить, соединив точки графика, соответствующие концам интервалов, отрезками прямой. В результате график функции $\hat{F}(x)$ будет представлять собой непрерывную линию. Для графика этой функции имеется еще одно название - ***кривая накопленных частот.***



Для графического представления данных выборки существуют еще несколько способов, в частности ***полигон*** и ***гистограмма.***

Определение. Полигоном частот (частостей) называют ломанную линию отрезки которой соединяют точки с координатами (x_i ; $m_i(v_i)$), где x_i - значение i - того варианта, а $m_i(v_i)$ - соответствующие частоты (частости).

Полигон используют для изображения дискретного ряда. Гистограмма служит только для изображения интервальных вариационных рядов. Для ее построения в прямоугольной системе координат на оси Ox откладывают отрезки, изображающие частичные интервалы варьирования, и на этих отрезках, как на основаниях строят прямоугольники с высотами, равными частотам или частостям соответствующих интервалов. В результате таких действий получают ступенчатую фигуру, состоящую из прямоугольников, которую называют ***гистограмма.*** Поскольку количество данных, по которым строятся гистограммы, может сильно отличаться, необходимо строить гистограммы не в абсолютных цифрах числа значений, попадающих в интервал, а в относительных, представляющих собой долю, которую составляет это число, от общего числа экспериментальных данных. Полигон дает представление об эмпирической дифференциальной функции распределения (плотности вероятности).

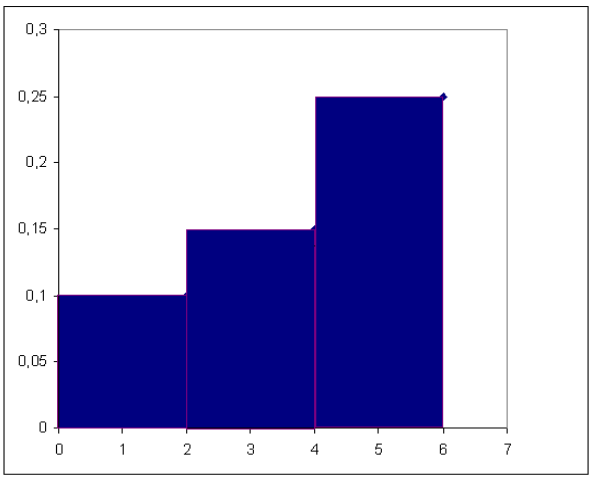
Пример 3. Построить гистограмму относительных частот по данному распределению выборки.

Номер интервала, i	Частичный интервал $x_i - x_{i+1}$	Сумма частот вариант частичного Интервала, m_i
1	0 – 2	20
2	2 – 4	30
3	4 – 6	50

Размер выборки $n = \sum m_i = 100$. Найдем относительные частоты (частости).

$v_1 = \frac{20}{100} = 0{,}2$, $v_2 = \frac{30}{100} = 0{,}3$, $v_3 = \frac{50}{100} = 0{,}5$. Высоты прямоугольников находим как отношение частостей к длине частичного интервала. Так как длина интервала $b = 2$, то $h_1 = \frac{0{,}2}{2} = 0{,}1$, $h_2 = \frac{0{,}3}{2} = 0{,}15$, $h_3 = \frac{0{,}5}{2} = 0{,}25$.

Теперь осталось изобразить гистограмму относительных частот (частостей).



Описательные статистические характеристики.

Построив вариационный ряд и изобразив его графически в виде гистограммы, полигона или эмпирической функции распределения, можно получить первоначальное представление о закономерностях, имеющих место в ряду наблюдений. Но на практике этого недостаточно. Для изучения генеральной совокупности требуются количественные показатели, которые помогут представить результаты наблюдений в компактном виде и сравнить, например, одну совокупность данных с другой. Поскольку эти характеристики вычисляются по статистическим данным (данным, полученным в результате наблюдений), их называют ***статистическими характеристиками*** или ***оценками***. Иногда применяется термин – ***статистики***.

Пусть весь собранный и обработанный статистический материал представлен в виде вариационного ряда. Анализируя эти результаты можно выделить некоторые постоянные значения, которые представляют ряд в целом: это некие средние значения, вокруг которых группируются данные, разброс данных вокруг этих средних значений (показатели рассеяния или вариации), показатели формы распределения. Основными показателями центра распределения являются среднее арифметическое, мода, медиана. Показатели вариации – размах выборки, дисперсия, среднеквадратическое отклонение. Показатели формы распределения: асимметрия, эксцесс.

Показатели центра распределения.

Среднее арифметическое.

Среднее арифметическое – наиболее часто используемый показатель центра распределения.

Определение. Пусть $x_1, x_2, \ldots, x_n$ - данные наблюдений над случайной величиной X. ***Средним арифметическим*** $\overline{X}$ наблюдаемых значений случайной величины X называется частное от деления суммы всех этих значений на их число, то есть $\overline{X} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{\sum_{i=1}^{n} x_i}{n}$. (1)

Если данные представлены в виде дискретного статистического ряда, то есть $x_1, x_2, \ldots, x_\nu$ - варианты, а $m_1, m_2, \cdots, m_\nu$ - частоты, то $\sum_{i=1}^{\nu} m_i = n$ и $\overline{X} = \dfrac{x_1 m_1 + x_2 m_2 + \cdots + x_\nu m_\nu}{m_1 + m_2 + \cdots + m_\nu} = \dfrac{\sum_{i=1}^{\nu} x_i m_i}{n}$ (2). Вычисленное по формуле (2) среднее арифметическое называют *взвешенным* так как m_i - веса.

Для интервального ряда за x_i, принимают середину i - того интервала, а за m_i соответствующую интервальную частоту.

Преобразуем формулу (2).

$$\overline{X} = \frac{\sum_{i=1}^{\nu} x_i m_i}{n} = \sum_{i=1}^{\nu} x_i \frac{m_i}{n} = \sum_{i=1}^{\nu} x_i \nu_i .$$

Пример 1. Случайная величина X - число неправильных соединений в минуту на телефонной станции. (Наблюдения велись в течении часа.) Данные наблюдений представлены в виде вариационного ряда. Найдем среднее число неправильных соединений.

i	1	2	3	4	5	6	7
x_i	0	1	2	3	4	5	7
m_i	8	17	16	10	6	2	1

$$\overline{X} = \frac{0 \cdot 8 + 1 \cdot 17 + 2 \cdot 16 + 3 \cdot 10 + 4 \cdot 6 + 5 \cdot 2 + 7 \cdot 1}{60} = 2 .$$

Для того, что бы понять, почему именно среднее арифметическое – наиболее популярная точечная оценка центра распределения сравним математическое ожидание случайной величины и среднее арифметическое:

$$MX = \sum_{i=1}^{n} x_i p_i \; ; \; \overline{X} = \sum_{i=1}^{\nu} x_i \nu_i .$$

Очевидна их внешняя схожесть. Но для математического ожидания x_i -все возможные значения случайной величины, а p_i - их вероятности; в формуле для среднего арифметического x_i - конкретные появившиеся значения случайной величины и ν_i - их относительные частоты. Математическое ожидание - это постоянная для случайной величины величина, а среднее значение может меняться от одной серии опытов к другой. Но на основании теоремы Чебышева мы знаем, что среднее арифметическое наблюдаемых значений случайной величины обладает свойством устойчивости и при увеличении числа наблюдений n сходится по вероятности к математическому ожиданию. Таким образом, несмотря на принципиальные различия, математическое ожидание и среднее арифметическое имеют много общего и естественно считать среднее арифметическое выборочным аналогом математического ожидания.

Еще две величины, характеризующие центр это медиана и мода.

Медиана (*Md*) – определяется как срединное значение в ранжированном ряду данных. Это значит, что по обе стороны он нее расположено равно по половине данных.

Вычисление медианы: если в ряду нечетное число данных, то это срединное значение, если четное, то полусумма двух срединных значений.

Мода (*Mo*) – представляет собой наиболее часто наблюдаемую величину изучаемой переменной.

Пример 2. Определить моду и медиану для выборки: 5, 6, 8, 2, 3, 1, 1, 4.

Ранжированный ряд: 1, 1, 2, 3, 4, 5, 6, 8.

Т.к. $n = 8$.

$$Md = \frac{3+4}{2} = 3{,}5 \qquad Mo = 1$$

Выборочные моменты.

Для выборки обьема n можно вычислить ***эмпирические моменты***.

Пусть $x_1, \ldots, x_n$ – выборка объема n из генеральной совокупности.

Выборочный момент (начальный) r-того порядка вычисляется по формуле

$$\alpha_r^* = \frac{1}{n}\sum_{i=1}^{n} x_i^r \; ; \; r = 1, 2, \ldots$$

Легко увидеть, что выборочный начальный момент первого порядка это среднее арифметическое - $\bar{x}$.

$$\bar{x} = \alpha_1^* = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Выборочный центральный момент r-того порядка определяется по формуле

$$\mu_r^* = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^r \; .$$

Квантили.

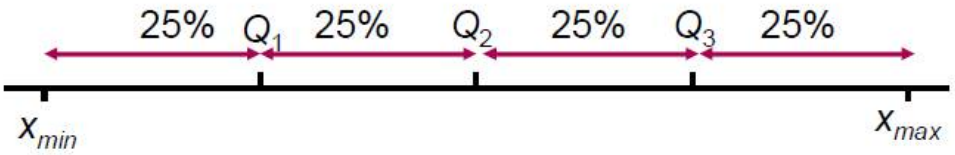
Выборочная квантиль x_p порядка p $(0 < p < 1)$ определяется как элемент вариационного ряда выборки $x^{(1)}, \ldots, x^{(n)}$ с номером $[np]+1$, где $[a]$ – целая часть числа a.

В статистической практике используется ряд квантилей, имеющий специальные названия.

Персентили: $P_1, \ldots, P_{99}$ – порядок 0,01 … 0,99.

Децили: $D_1, \ldots, D_9$ – порядок 0,1 … 0,9.

Квартили Q_1, Q_2, Q_3 – порядок 0,25; 0,5; 0,75.



Вариационный ряд делится тремя квартилями Q_1, Q_2, Q_3 на 4 равные части. Q_2 – медиана.

Показатели рассеивания.

Имея представления о центре распределения, часто требуется узнать, как данные рассеяны вокруг него. Можно, например, вычислить средний межквартильный размах: $\frac{Q_3 - Q_1}{2}$.

Но на практике чаще применяется дисперсия и среднеквадратическое отклонение.

Выборочная дисперсия.

Так как среднее арифметическое – выборочный аналог математического ожидания, следовательно, следует ввести подобную характеристику и для вариационных рядов, которая будет показывать меру рассеивания наблюдаемых данных $x_1, x_2, ..., x_n$ вокруг среднего арифметического: $x_i - \overline{X}$. Сумма всех этих значений не может быть такой характеристикой, так как она равна нулю. Чтобы этого не происходило величину отклонений возводят в квадрат.

Определение. ***Выборочной дисперсией*** случайной величины $\overline{X}$ называется среднее арифметическое квадратов отклонений наблюдаемых значений этой величины от их среднего арифметического (обозначим S_0^2 или σ^2).

$$S_0^2 = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n} \quad (3) = \mu_2^*$$

Очевидно, что одновременно выборочная дисперсия является центральным выборочным моментом второго порядка.

Если данные представлены статистическим рядом (x_i, m_i), то

$$\sigma^2 = \frac{\sum_{i=1}^{v}(x_i - \overline{X})^2 m_i}{n} \quad (4); \; n = \sum_{i=i}^{v} m_i \text{ -- объем выборки.}$$

или

$$S_0^2 = \sum_{i=1}^{v}(x_i - \overline{X})^2 \cdot \nu_i \;, \; \nu_i = \frac{m_i}{n}$$

Вычисленная по формулам (4) дисперсия называется ***взвешенной выборочной дисперсией.***

Можно также использовать упрощенную формулу для расчета выборочной дисперсии:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\overline{X}\right)^2$$

Среднее арифметическое имеет ту же размерность, что и величина, дисперсия имеет размерность квадрата величины, и это не всегда удобно для оценки разброса значений, поэтому в качестве меры рассеивания можно взять квадратный корень из дисперсии.

Определение. ***Выборочным средним квадратическим отклонением*** называется арифметический квадратный корень из выборочной дисперсии.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{v}(x_i - \overline{X})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{v}(x_i - \overline{X})^2 m_i}{n}}$$

Выборочную дисперсию можно использовать в качестве оценки генеральной дисперсии, однако она не обладает одним из важных свойств точечных оценок – несмещенностью.
Поэтому кроме выборочной дисперсии используется характеристика, называемая ***несмещенной дисперсией*** («исправленной»).

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{n}{n-1} \cdot \sigma^2$$

Также можно найти «исправленное» среднее квадратическое отклонение:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{v}(x_i - \overline{X})^2}{n-1}}$$

Пример 3. Используем данные примера 1. Так как $\overline{X} = 2$

$$\sigma^2 = ((0-2)^2 \cdot 8 + (1-2)^2 \cdot 17 + (2-2)^2 \cdot 16 + (3-2)^2 \cdot 10 + (4-2)^2 \cdot 6 + (5-2)^2 \cdot 2 +$$

$$+ (7-2)^2 \cdot 0) / 60 = 2{,}1$$

$$\sigma = \sqrt{2{,}1} \approx 1{,}45$$

$$s^2 = \frac{60}{59} \cdot 2{,}1 \approx 2{,}136 \quad s \approx 1{,}46$$