Закон больших чисел и предельные теоремы

Под **законом больших чисел (ЗБЧ)** в *широком* смысле понимается общий принцип, согласно которому, по формулировке академика Колмогорова А.Н., ***совокупное действие большого числа случайных факторов приводит*** (при некоторых весьма общих условиях) ***к результату, почти не зависящему от случая***.
Др. словами, ***при большом числе случайных величин их средний результат перестает быть случайным и моет быть предсказан с большой степенью определенности.***

Под **ЗБЧ** в *узком* смысле понимается *ряд математических теорем, в каждой из которых для тех или иных условий устанавливается факт приближения средних характеристик большого числа испытаний к некоторым определенным постоянным.*

Неравенство Маркова (лемма Чебышева)

Теорема. *Если случайная величина X принимает только неотрицательные значения и имеет математическое ожидание, то для любого положительного числа A верно неравенство*

$$P(x > A) \le \frac{M(X)}{A}. \qquad (1)$$

Вероятность того, что случайная величина X примет значение больше A, меньше или равно частного от деления математического ожидания этой СВ на число A.

Неравенство Маркова применимо к любым неотрицательным случайным величинам.

Пример 1. Среднее количество вызовов, поступающих на коммутатор завода в течение часа, равно 300. Оценить вероятность того, что в течение следующего часа число вызовов на коммутатор: а) превысит 400; б) будет не более 500.

Р е ш е н и е. а) По условию $M(X) = 300$. По формуле (1) $P(X > 400) \le \frac{300}{400}$, т.е. вероятность того, что число вызовов превысит 400, будет н е б о л е е 0,75.

б) По формуле (1) $P(X \le 500) \ge 1 - \frac{300}{500} = 0,4$, т.е. вероятность того, что число вызовов не более 500, будет н е м е н е е 0,4. ▶

Неравенство Чебышева

Теорема. *Для любой случайной величины, имеющей математическое ожидание и дисперсию, справедливо неравенство Чебышева:*

$$P\left(|X-a|>\varepsilon\right)\le\frac{D(X)}{\varepsilon^2}, \qquad (2)$$

Вероятность того, что случайная величина отклониться от своего математического ожидания на величину $\xi>0$ меньше или равна частного от деления дисперсии этой величины на ξ^2

Учитывая, что события $|X-a|>\varepsilon$ и $|X-a|\le\varepsilon$ противоположны, неравенство Чебышева можно записать и в другой форме:

$$P\left(|X-a|\le\varepsilon\right)\ge 1-\frac{D(X)}{\varepsilon^2}. \qquad (3)$$

Неравенство Чебышева для некоторых случайных величин

а) *для случайной величины* $X = m$, *имеющей биномиальный закон распределения с математическим ожиданием* $a = M(X) = np$ *и дисперсией* $D(X) = npq$ (см. § 4.1):

$$P\left(|m - np| \le \varepsilon\right) \ge 1 - \frac{npq}{\varepsilon^2}; \qquad (4)$$

б) *для частости* $\frac{m}{n}$ *события в n независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью* $a = M\left(\frac{m}{n}\right) = p$, *и имеющей дисперсию* $D\left(\frac{m}{n}\right) = \frac{pq}{n}$:

$$P\left(\left|\frac{m}{n} - p\right| \le \varepsilon\right) \ge 1 - \frac{pq}{n\varepsilon^2}. \qquad (5)$$

Пример 2. Оценить вероятность того, что отклонение любой случайной величины от ее математического ожидания будет не более трех средних квадратических отклонений (по абсолютной величине) — (*правило трех сигм*).

Р е ш е н и е. По формуле (3), учитывая, что $D(X)=\sigma^2$, получим:

$$P\left(|X-a|\le 3\sigma\right)\ge 1-\frac{\sigma^2}{(3\sigma)^2}=\frac{8}{9}=0{,}889,$$

Теорема Чебышева

Теорема Чебышева. *Если дисперсии n независимых случайных величин $X_1, X_2, \ldots, X_n$ ограничены одной и той же постоянной, то при неограниченном увеличении числа n средняя арифметическая случайных величин сходится по вероятности к средней арифметической их математических ожиданий $a_1, a_2, \ldots, a_n$,* т.е.

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{a_1+a_2+\ldots+a_n}{n}\right|\le\varepsilon\right)=1 \qquad (6)$$

Опуская доказательство, получаем неравенство

$$P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{a_1+a_2+\ldots+a_n}{n}\right|\le\varepsilon\right)\ge 1-\frac{D(X)}{\varepsilon^2}. \qquad (7)$$

$$P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{a_1+a_2+\ldots+a_n}{n}\right|\le\varepsilon\right)\ge 1-\frac{C}{n\varepsilon^2}. \qquad (8)$$

$$D(X_1)\le C,\ D(X_2)\le C,\ \ldots,\ D(X_n)\le C,$$

где C — постоянное число.

Суть теоремы Чебышева:

Отдельные независимые случайные величины могут принимать значения, далекие от своих математических ожиданий, среднее арифметическое достаточно большого числа случайных величин с большой вероятностью принимает значения, близкие к средней арифметической их математических ожиданий

Хотя мы и не можем предсказать конкретное значение случайной величины, ***мы можем с вероятностью, близкой к единице, определить её среднее арифметическое****, чего будет более чем достаточно на практике.*

Примеры применения теоремы Чебышева в реальной жизни

1. Проведение измерений: при достаточно большом количестве измерений, например, напряжения в сети, можно получить значение, сколько угодно близкое к истинному.
2. Проверка качества. Нет необходимости, например, проверять всю партию однообразных товаров, а достаточно выборочной проверки.
3. Страхование. Рассматривая величину страхового взноса, страховщик обладает определенной информацией о вероятности наступления страховых случаев и возможных потерях клиента от них. По теореме Чебышева найдя среднее арифметическое от этих убытков, страховщик может определить идеальную величину страхового взноса: выгодную для него и привлекательную для клиента.
4. Финансовые рынки. Проведение большого числа финансовых операций с известной средней ожидаемой доходностью лежит в основе диверсификации рисков.

Пример 3. Для определения средней продолжительности горения электроламп в партии из 200 одинаковых ящиков было взято на выборку по одной лампе из каждого ящика. Оценить вероятность того, что средняя продолжительность горения отобранных 200 электроламп отличается от средней продолжительности горения ламп во всей партии не более чем на 5 ч (по абсолютной величине), если известно, что среднее квадратическое отклонение продолжительности горения ламп в каждом ящике меньше 7 ч.

Р е ш е н и е. Пусть X_i — продолжительность горения электролампы, взятой из i-го ящика (ч). По условию дисперсия $D(X_i) < 7^2 = 49$. Очевидно, что средняя продолжительность горения отобранных ламп равна $(X_1 + X_2 + ... + X_{200})/200$, а средняя продолжительность горения ламп во всей партии $(M(X_1) + M(X_2) + ... + M(X_{200}))/200 = (a_1 + a_2 + ... + a_{200})/200$.

Тогда вероятность искомого события по формуле **(8)**

$$P\left(\left|\frac{X_1 + X_2 + ... + X_{200}}{200} - \frac{a_1 + a_2 + ... + a_{200}}{200}\right| \le 5\right) \ge 1 - \frac{49}{200 \cdot 5^2} \approx 0{,}9902,$$

т.е. н е м е н е е чем 0,9902. ▶

Теорема Бернулли и теорема Пуассона

Теорема Бернулли. *Частость события в n повторных независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью p, при неограниченном увеличении числа n сходится по вероятности к вероятности p этого события в отдельном испытании:*

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| \le \varepsilon\right) = 1 \qquad (9)$$

Теорема Пуассона. *Частость события в n повторных независимых испытаниях, в каждом из которых оно может произойти соответственно с вероятностями p_1, p_2, ..., p_n, при неограниченном увеличении числа n сходится по вероятности к средней арифметической вероятностей события в отдельных испытаниях,* т.е.

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - \frac{p_1 + p_2 + \ldots + p_n}{n}\right| \le \varepsilon\right) = 1 \qquad (10)$$

Теорема Бернулли дает теоретическое обоснование замены неизвестной вероятности события его частостью, или статической вероятностью, полученной в n повторных независимых испытаниях, проводимых при одном и том же комплексе условий (см. лекция №1).

Важная роль закона больших чисел в теоретическом обоснование методов математической статистики и ее приложений обусловила проведение ряда исследований, направленных на изучение общих условий применимости этого закона к последовательности случайных величин.

Нахождение общих условий, выполнение которых обязательно влечет за собой статистическую устойчивость средних, представляет непреходящую научную ценность исследований в области закона больших чисел.

Центральная предельная теорема

Рассмотренный выше закон больших чисел устанавливает факт приближения средней большого числа случайных величин к определенным постоянным. Но этим не ограничиваются закономерности, возникающие в результате суммарного действия случайных величин. Оказывается, что при некоторых весьма общих условиях совокупное действие большого числа случайных величин приводит к определенному, а именно — к нормальному закону распределения.

Центральная предельная теорема представляет собой группу теорем, посвященных установлению условий, при которых возникает нормальный закон распределения. Среди этих теорем важнейшее место принадлежит теореме Ляпунова.

Теорема Ляпунова. *Если* X_1, X_2, ..., X_n — *независимые случайные величины, у каждой из которых существует математическое ожидание* $M(X_i) = a$, *дисперсия* $D(X_i) = \sigma^2$, *абсолютный центральный момент третьего порядка* $M\left(\left|X_i - a_i\right|^3\right) = m_i$ и

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} m_i}{\left(\sum_{i=1}^{n} \sigma_i^{\,2}\right)^{3/2}} = 0, \qquad (11)$$

то закон распределения суммы $Y_n = X_1 + X_2 + ... + X_n$ *при* $n \to \infty$ *неограниченно приближается к нормальному с математическим ожиданием* $\sum_{i=1}^{n} a_i$ *и дисперсией* $\sum_{i=1}^{n} \sigma_i^2$.

Теорема позволяет утверждать, что всегда, когда случайная величина образуется в результате сложения большого числа независимых случайных величин, дисперсии которых малы по сравнению с дисперсией суммы, закон распределения этой случайной величины оказывается практически нормальным законом.

Так, например, потребление электроэнергии для бытовых нужд за месяц в каждой квартире многоквартирного дома можно представить в виде n различных случайных величин. Если потребление электроэнергии в каждой квартире по своему значению резко не выделяется среди остальных, то на основании теоремы Ляпунова (11) можно считать, что потребление электроэнергии всего дома, т.е. сумма n независимых случайных величин будет случайной величиной, имеющей приближенно нормальный закон распределения. Если, например, в одном из помещений дома разместится вычислительный центр, у которого уровень потребления электроэнергии несравнимо выше, чем в каждой квартире для бытовых нужд, то вывод о приближенно нормальном распределении потребления электроэнергии всего дома будет неправомерен, так как нарушено условие (11), ибо потребление электроэнергии вычислительного центра будет играть превалирующую роль в образовании всей суммы потребления.

Математическая статистика

Математическая статистика выделяется из теории вероятностей в самостоятельную область, хотя основные методы и приемы рассуждений в ней остаются теми же самыми. Причиной этого **является специфичность задач математической статистики**, являющихся в известной мере обратными к задачам теории вероятностей.

Если в теории вероятностей мы исходным материалом имели вероятностную модель случайного явления и на ее основе рассчитывали возможные реализации данного случайного явления, то в **математической статистике** мы, наоборот, *имеем некоторую реализацию случайного явления и по этой реализации устанавливаем вероятностную модель с целью получения возможности рассчитывать любые другие возможные его реализации, прогнозировать данное явление.*

Генеральная совокупность

В качестве реализации случайного явления или случайного события выступают так называемые **статистические данные**. В большинстве случаев исходные статистические данные — результат наблюдений некоторого признака, являющегося СВ или характеризующегося системой случайных величин

Совокупность всех возможных реализаций исследуемой СВ или системы случайных величин называется ***генеральной совокупностью (ГС).***

*Примерами **ГС** могут служить множество рабочих данного цеха при изучении вопроса о качестве выпускаемой продукции, множество населения страны при исследовании ее трудовых ресурсов и т.д.*

Выборка

Пусть требуется изучить некоторую **ГС**. Для этого можно провести сплошное обследование. Однако если число объектов ГС велико, то осуществить указанное обследование чрезвычайно сложно. Поэтому для изучения ГС применяется ***выборочный метод, суть которого состоит в том, что обследованию подвергают*** не все элементы ГС, а только часть их, **случайно выбранную** из данной совокупности. Выводы, полученные при исследовании этой части, распространяются на всю ГС.

Множество элементов ГС, случайным образом выбранных из нее, называется ***выборкой, а число их — объемом выборки.***

Очевидно, главное требование к выборке — она должна хорошо представлять ГС, т.е. должна быть ***репрезентативной.***
Выборка будет ***репрезентативной,*** если ее выбирать ***случайным образом*** и ***в достаточном объеме.***
Основной недостаток выборочного метода – ***ошибки исследования***, называемые ошибками репрезентативности.

Используют д в а с п о с о б а образования выборки:

- *повторный отбор* (по схеме возвращенного шара), когда каждый элемент, случайно отобранный и обследованный, возвращается в общую совокупность и может быть повторно отобран;
- *бесповторный отбор* (по схеме невозвращенного шара), когда отобранный элемент не возвращается в общую совокупность.

Математическая теория выборочного метода основывается на анализе собственно-случайной выборки. Рассмотрением этой выборки мы и ограничимся.

Обозначим:

x_i — значения признака (случайной величины X);
N и n — объемы генеральной и выборочной совокупностей;
N_i и n_i — число элементов генеральной и выборочной совокупностей со значением признака x_i;
M и m — число элементов генеральной и выборочной совокупностей, обладающих данным признаком.

Вариационный ряд

Если значения выборки записать в неубывающем порядке, то получим последовательность, называемую ***вариационным рядом, а $x_{(i)}$ называются вариантами.***

Разность между крайними членами вариационного ряда называется ***размахом варьирования.***

Вариационный ряд может быть ***дискретный*** или ***интервальный.***

Дискретный вариационный ряд — это ряд, в основу построения которого положен признак с ***прерывным изменением*** (число рабочих на предприятии, число детей в семье и т.д.).

Этот признак может принимать только конечное число фиксированных значений.

Варианта	x_1	x_2	...	x_k
Частота	n_1	n_2	...	n_k

Очевидно, сумма частот равна объему выборки n.

Если признак имеет ***непрерывное изменение***, т.е. значения могут отличаться одно от другого на сколь угодно малую величину и в определенных границах могут принимать любые значения (заработная плата рабочих, размер среднего душевого денежного дохода, стоимость основных фондов предприятия), то для этого признака нужно строить ***интервальный вариационный ряд.***

Таблица здесь также имеет две графы. В первой указывается значение признака в интервале .от — до, во второй — число выборочных значений, входящих в соответствующий интервал (частоты):

Варианта	$x_1 - x_2$	$x_2 - x_3$	...	$x_k - x_{k+1}$
Частота	n_1	n_2	...	n_k

Построение интервального ряд

Согласно *формуле Стерджеса* рекомендуемое число интервалов $m = 1 + 3{,}322 \lg n$, а *величина интервала* (*интервальная разность, ширина интервала*)

$$k = \frac{x_{max} - x_{min}}{1 + 3{,}322 \lg n},$$

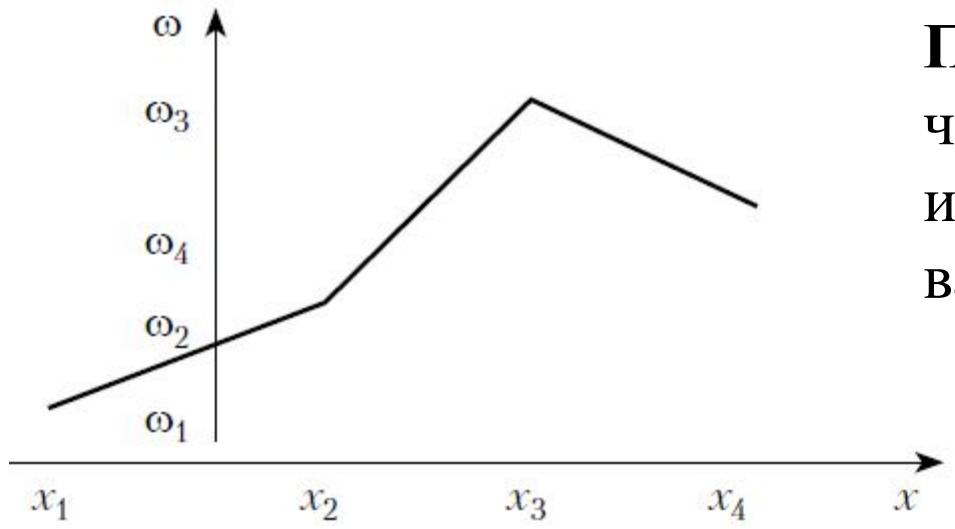
где $x_{max} - x_{min}$ — разность между наибольшим и наименьшим значениями признака.

Относительной частотой называется величина $\omega_i = \frac{n_i}{n}$

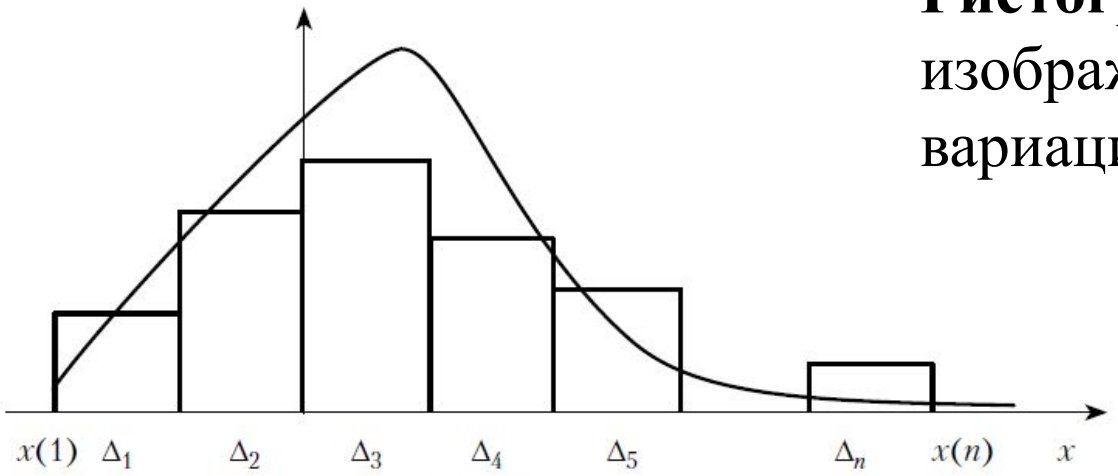
где ***n*** — *объем выборки, а* n_i — *частота выборочного значения* ***xi***

Накопленной частотой называется величина $\omega_i^{накоп} = \frac{n_i^{накоп}}{n}$

Показывающая, сколько наблюдалось вариантов со значением признака меньшим x.



Полигон относительных частот служит для изображения дискретного вариационного ряда.



Гистограмма служит для изображения интервального вариационного ряда.

Кумулятивная кривая (*кумулята*) — кривая накопленных частот (частостей). Для дискретного ряда кумулята представляет ломаную, соединяющую точки $(x_i, n_i^{\text{нак}})$ или $(x_i, w_i^{\text{нак}})$, $i = 1, 2, \ldots, m$. Для интервального вариационного ряда ломаная начинается с точки, абсцисса которой равна началу первого интервала, а ордината — накопленной частоте (частости), равной нулю. Другие точки этой ломаной соответствуют концам интервалов.

Определение. ***Эмпирической функцией распределения*** $F_n(x)$ *называется относительная частота* (*частость*) *того, что признак* (*случайная величина* X) *примет значение, меньшее заданного* x, т.е.

$$F_n(x) = w\left(X < x\right) = w_x^{\text{нак}}.$$

Другими словами, для данного x эмпирическая функция распределения представляет накопленную частость $w_x^{\text{нак}} = n_x^{\text{нак}} / n$.

Числовые характеристики выборочного распределения назовем выборочными или **эмпирическими** характеристиками.

Средние величины

О п р е д е л е н и е. ***Средней арифметической*** *вариационного ряда называется сумма произведений всех вариантов на соответствующие частоты, деленная на сумму частот:*

$$\overline{x} = \frac{\sum_{i=1}^{m} x_i n_i}{n},$$

где x_i — варианты дискретного ряда или середины интервалов интервального вариационного ряда; n_i — соответствующие им частоты; m — число неповторяющихся вариантов или число интервалов;

$$n = \sum_{i=1}^{m} n_i .$$

Очевидно, что

$$\overline{x} = \sum_{i=1}^{m} x_i w_i ,$$

где $w_i = n_i/n$ — частости вариантов или интервалов.

Средние величины

Определение. ***Медианой*** $\tilde{Me}$ ***вариационного ряда*** *называется значение признака, приходящееся на середину ранжированного ряда наблюдений.*

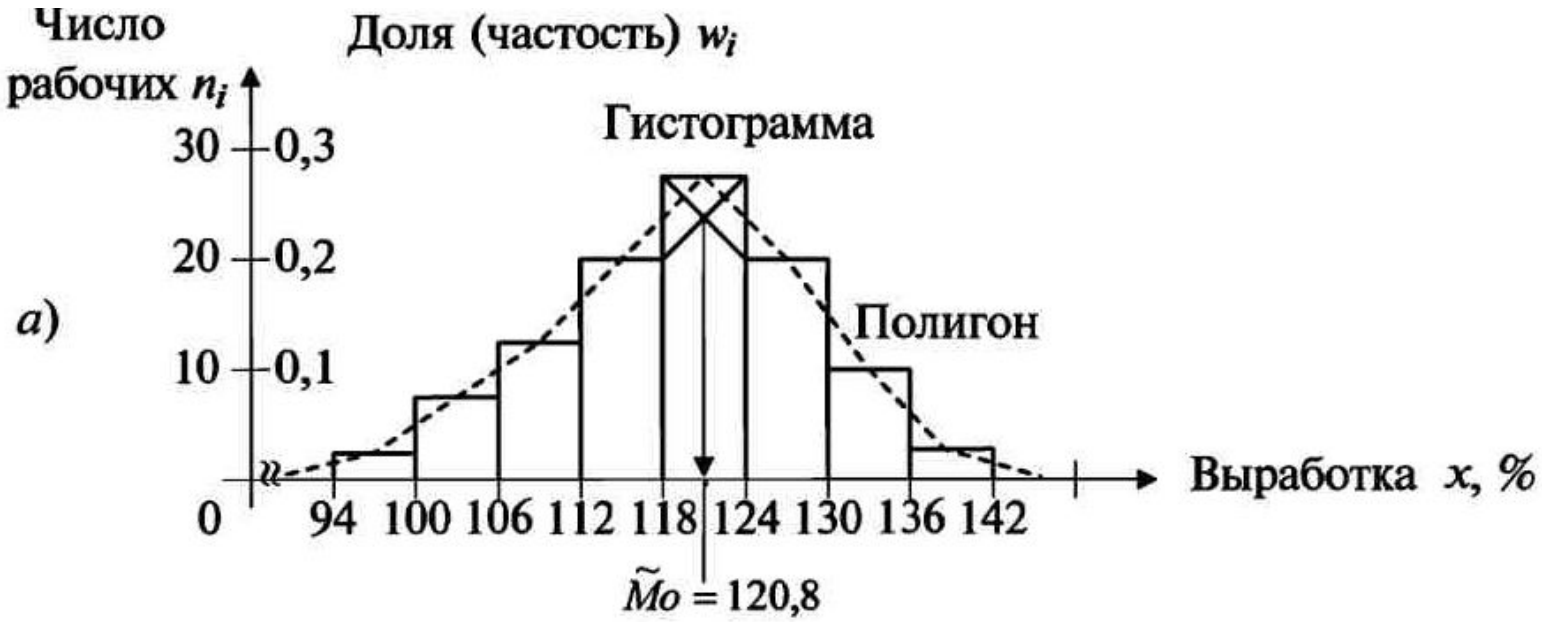
Для дискретного вариационного ряда с нечетным числом членов медиана равна серединному варианту, а для ряда с четным числом членов — полусумме двух серединных вариантов.

Для интервального вариационного ряда находится медианный интервал, на который приходится середина ряда, а значение медианы на этом интервале находят с помощью линейного интерполирования. Не приводя соответствующей формулы, отметим, что медиана может быть приближенно найдена с помощью кумуляты как значение признака, для которого $n_x^{\text{нак}} = n/2$ или $w_x^{\text{нак}} = 1/2$.

Средние величины

О п р е д е л е н и е. ***Модой*** $\tilde{M}o$ ***вариационного ряда*** *называется вариант, которому соответствует наибольшая частота.*

Особенность моды как меры центральной тенденции заключается в том, что она не изменяется при изменении крайних членов ряда, т.е. обладает определенной устойчивостью к вариации признака.



а)

Показатели вариации

Простейшим (и весьма приближенным) показателем вариации является *вариационный размах* R, равный разности между наибольшим и наименьшим вариантами ряда:

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

О п р е д е л е н и е. ***Дисперсией s^2 вариационного ряда*** *называется средняя арифметическая квадратов отклонений вариантов от их средней арифметической:*

$$s^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 n_i}{n}.$$

Формулу для дисперсии вариационного ряда можно записать в виде:

$$s^2 = \sum_{i=1}^{m} (x_i - \bar{x})^2 w_i,$$

где $w_i = n_i / n$.

Желательно в качестве меры вариации (рассеяния) иметь характеристику, выраженную в тех же единицах, что и значения признака. Такой характеристикой является ***среднее квадратическое отклонение*** *s — арифметическое значение корня квадратного из дисперсии* —

$$s = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \overline{x})^2 n_i}{n}}.$$

Рассматривается также безразмерная характеристика — ***коэффициент вариации***, равный процентному отношению среднего квадратического отклонения к средней арифметической:

$$\tilde{v} = \frac{s}{\overline{x}} \cdot 100\% \quad (\overline{x} \neq 0).$$

Если выборка представлена в виде интервального вариационного ряда, то для вычисления выборочных среднего и дисперсии вначале нужно определить середины интервалов а затем находить их по формулам в которые вместо x_i *подставить* x_i^*.

$$x_i^* = \frac{x_i + x_{i+1}}{2}$$

Свойства дисперсии

1. *Дисперсия постоянной равна нулю.*

2. *Если все варианты увеличить (уменьшить) в одно и то же число k раз, то дисперсия увеличится (уменьшится) в k^2 раз:*

$$s_{kx}^2 = k^2 s_x^2 \text{ или } \frac{\sum_{i=1}^{m}\left(x_i k - \overline{x}k\right)^2 n_i}{n} = k^2 \frac{\sum_{i=1}^{m}\left(x_i - \overline{x}\right)^2 n_i}{n}.$$

3. *Если все варианты увеличить (уменьшить) на одно и то же число, то дисперсия не изменится:*

$$s_{x+c}^2 = s_x^2 = s^2 \text{ или } \frac{\sum_{i=1}^{m}\left[\left(x_i + c\right) - \left(\overline{x} + c\right)\right]^2 n_i}{n} = \frac{\sum_{i=1}^{m}\left(x_i - \overline{x}\right)^2 n_i}{n}.$$

4. *Дисперсия равна разности между средней арифметической квадратов вариантов и квадратом средней арифметической:* $s^2 = \overline{x^2} - \overline{x}^2$

5. *Если ряд состоит из нескольких групп наблюдений, то общая дисперсия равна сумме средней арифметической групповых дисперсий и межгрупповой дисперсии:*

$$s^2 = \overline{s_i^2} + \delta^2, \qquad \delta^2 = \frac{\sum_{i=1}^{l}\left(\overline{x}_i - \overline{x}\right)^2 n_i}{n}$$

Средняя арифметическая и дисперсия вариационного ряда являются частными случаями более общего понятия — моментов вариационного ряда.

Начальный момент $\tilde{\nu}_k$ *k-го порядка вариационного ряда*[1] определяется по формуле:

$$\tilde{\nu}_k = \frac{\sum_{i=1}^{m} x_i^k n_i}{n}.$$

Очевидно, что $\tilde{\nu}_1 = \overline{x}$, т.е. средняя арифметическая является начальным моментом первого порядка вариационного ряда.

Центральный момент $\tilde{\mu}_k$ *k-го порядка вариационного ряда* определяется по формуле:

$$\tilde{\mu}_k = \frac{\sum_{i=1}^{m} \left(x_i - \overline{x}\right)^k n_i}{n}.$$

Коэффициентом асимметрии вариационного ряда называется число

$$\tilde{A} = \frac{\tilde{\mu}_3}{s^3} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^3 n_i}{ns^3}.$$

Если $\tilde{A} = 0$, то распределение имеет симметричную форму, т.е. варианты, равноудаленные от $\bar{x}$, имеют одинаковую частоту. При $\tilde{A} > 0$ $\left(\tilde{A} < 0\right)$ говорят о положительной (правосторонней) или отрицательной (левосторонней) асимметрии.

Эксцессом (или *коэффициентом эксцесса*) *вариационного ряда* называется число

$$\tilde{E} = \frac{\tilde{\mu}_4}{s^4} - 3 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^4 n_i}{ns^4} - 3.$$

Эксцесс является показателем «крутости» вариационного ряда по сравнению с нормальным распределением. Как отмечено выше эксцесс нормально распределенной случайной величины равен нулю.

Средняя арифметическая $\overline{x}$, дисперсия s^2 и другие характеристики вариационного ряда являются статистическими аналогами математического ожидания $M(X)$, дисперсии σ^2 и соответствующих характеристик случайной величины X.

$F_n(x) = w(X < x)$	Эмпирическая функция распределения	$F(x) = P(X < x)$	Функция распределения
$\overline{x} = \sum_{i=1}^{m} x_i w_i$	Средняя арифметическая	$a = M(X) = \sum_{i=1}^{n} x_i p_i$	Математическое ожидание*
$s^2 = \overline{(x - \overline{x})^2} = \sum_{i=1}^{m} (x_i - \overline{x})^2 w_i$	Дисперсия	$\sigma^2 = M[X - M(X)]^2 = \sum_{i=1}^{n} (x_i - a)^2 p_i$	Дисперсия*
$s = \sqrt{s^2}$	Среднее квадратическое отклонение	$\sigma = \sqrt{D(X)} = \sqrt{\sigma^2}$	Среднее квадратическое отклонение
$\tilde{Mo}$	Мода	$Mo(X)$	Мода
$\tilde{Me}$	Медиана	$Me(X)$	Медиана

Пример 1. Имеются данные о торгах акций некоторого акционерного общества на фондовой бирже. Количество проданных акций по курсу продаж распределилось следующим образом:

Курс продаж	900	990	1010	1015	1150
Количество проданных акций	550	650	800	700	850

Найдем оценки среднего и дисперсии курса продаж акции

Решение. В качестве оценок возьмем выборочные среднюю и дисперсию.

$$\overline{x} = \frac{1}{3550}(900 \cdot 550 + 990 \cdot 650 + 1010 \cdot 800 + 1015 \cdot 700 + 1150 \cdot 850) = 1024.$$

$$S^2 = \frac{1}{3550}[(900-1024)^2 \cdot 550 + (990-1024)^2 \cdot 650 + (1010-1024)^2 \cdot 800 + (1015-1024)^2 \cdot 700 + (1150-1024)^2 \cdot 850] = 6455{,}3.$$

Пример 2. По результатам выборочного исследования рабочих цеха были установлены их квалификационные разряды: 4, 5, 6, 4, 4, 2, 3, 5, 4, 4, 5, 2, 3, 3, 4, 5, 5, 2, 3, 6, 5, 4, 6, 4, 3. Требуется:

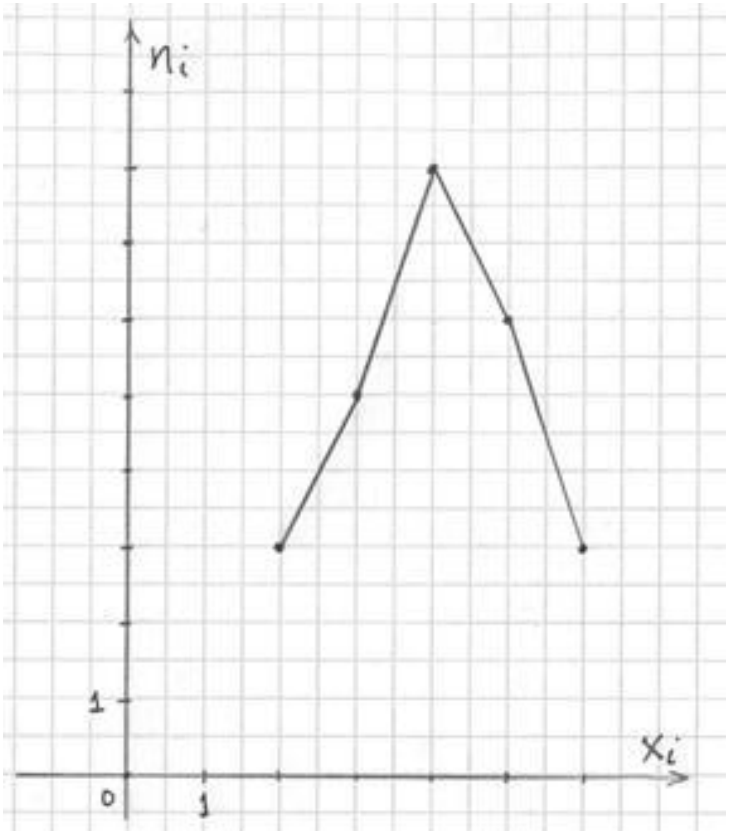
– составить вариационный ряд и построить полигон частот;

– найти относительные частоты и построить эмпирическую функцию распределения.

– найти среднее и дисперсию

Решение.

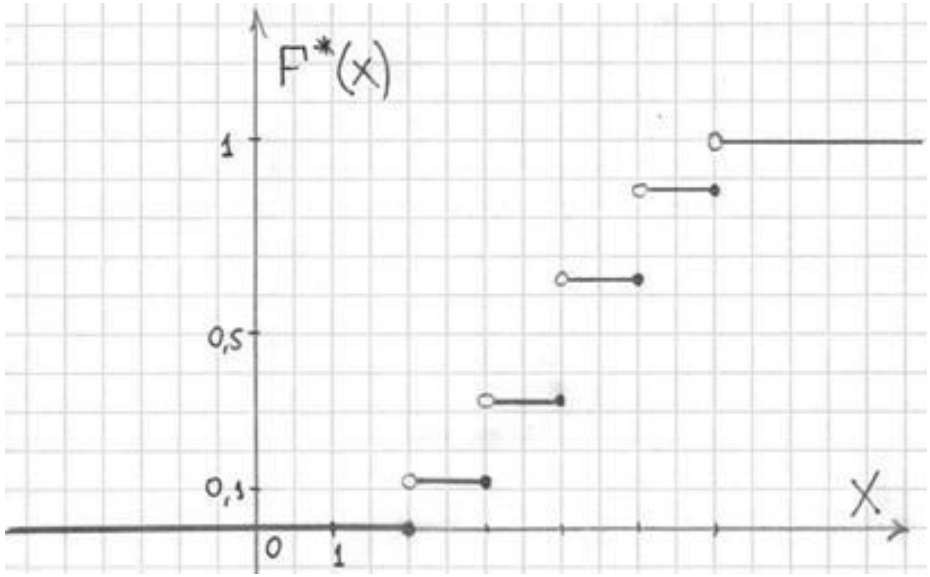
x_i	n_i
2	3
3	5
4	8
5	6
6	3
$\sum$	25



Найдём относительные частоты [illegible] и накопленные частоты

x_i	n_i	w_i	$w_н$
2	3	0,12	0,12
3	5	0,2	0,32
4	8	0,32	0,64
5	6	0,24	0,88
6	3	0,12	1
$\sum$	25	1	

[illegible]



x_i	n_i	w_i	x_i^2
2	3	0,12	4
3	5	0,2	9
4	8	0,32	16
5	6	0,24	25
6	3	0,12	36
$\sum$	25	1	

$$\overline{x^2} = \frac{\sum x_i^2 n_i}{n} = \frac{4\cdot 3 + 9\cdot 5 + 16\cdot 8 + 25\cdot 6 + 36\cdot 3}{25} = 17{,}72$$

$$S^2 = \overline{x^2} - \overline{x}^2 = 17{,}72 - 4{,}04^2 = 1{,}3984$$

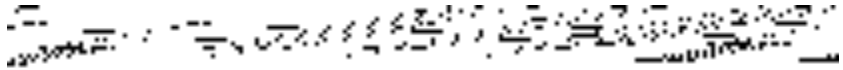
Пример 3. По результатам исследования цены некоторого товара в различных торговых точках города, получены следующие данные (в некоторых денежных единицах):

7,5	7,6	8,7
6,1	10,6	9,8
7	6	8,3
6	8,2	8,5
7,4	7,1	9,5
6,8	9,6	6,3
6,3	8,5	5,8
7,5	9,2	7,2
7	8	7,5
7,5	8	6,5

Требуется составить вариационный ряд распределения

Решение. Сначала окидываем взглядом предложенные числа и определяем примерный интервал, в который вписываются эти значения. «Навскидку» все значения заключены в пределах от 5 до 11. Далее делим этот интервал на удобные подынтервалы, в данном случае напрашиваются промежутки единичной длины.

5-6	6-7	7-8	8-9	9-10	10-11
5,8	6,1	7,5	8,7	9,8	10,6
	6	7,6	8,3	9,5	
	6	7	8,2	9,6	
	6,8	7,4	8,5	9,2	
	6,3	7,1	8,5		
	6,3	7,5	8		
	6,5	7,2	8		
		7			
		7,5			
		7,5			

Вычислим *размах вариаци*: 

Теперь его нужно разбить на *частичные интервалы*. Сколько интервалов рассмотреть? По умолчанию на этот счёт существует ***формула Стерджеса (см. выше)***:

Согласно *формуле Стерджеса* рекомендуемое число интервалов $m = 1 + 3{,}322 \lg n$, а *величина интервала* (*интервальная разность, ширина интервала*)

m=5 k=0.96

$$k = \frac{x_{max} - x_{min}}{1 + 3{,}322 \lg n},$$

К можно не округлять и использовать длину 0,96, но удобнее, 1.

И коль скоро мы прибавили 0,04, то по 5 частичным интервалам у нас получается «перебор»: 0,04·5=0,2. Поэтому от самой малой варианты 5,8 отмеряем влево 0,1 влево *(половину «перебора»)* и к значению 5,7 начинаем прибавлять по 1 , получая тем самым частичные интервалы.

При этом сразу рассчитываем их середины.

Убеждаемся в том, что самая большая варианта 10,6 вписалась в последний частичный интервал и отстоит от его правого конца на 0,1.

интервалы	x_i
5,7 - 6,7	6,2
6,7 - 7,7	7,2
7,7 - 8,7	8,2
8,7 - 9,7	9,2
9,7 - 10,7	10,2

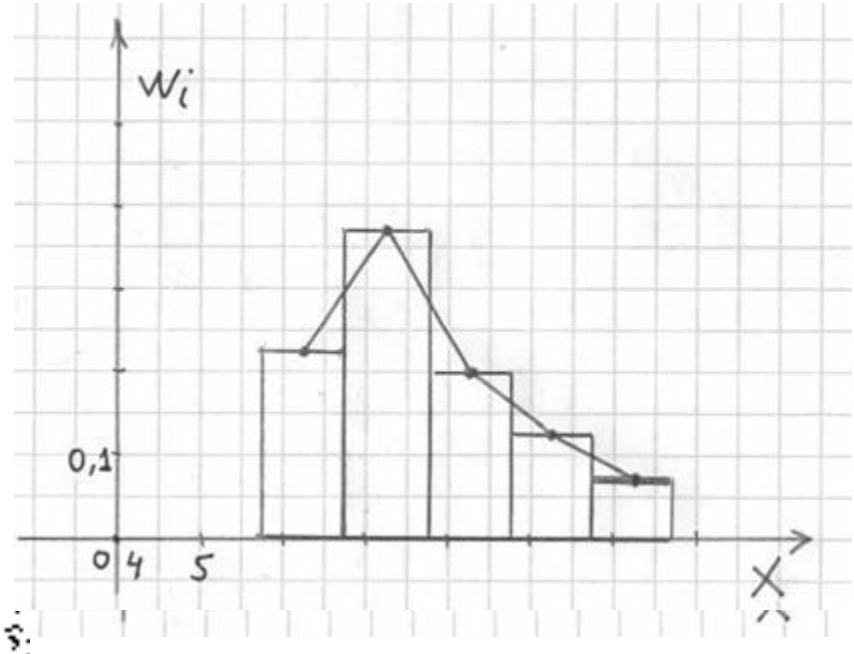
Далее подсчитываем частоты по каждому интервалу. Для этого в черновой «таблице» обводим значения, попавшие в тот или иной интервал, подсчитываем их количество и вычёркиваем:

Правило: если варианта попадает на «стык» интервалов, то её следует относить в правый интервал. У нас такая варианта встретилась одна: 8,7 – и её нужно причислить к интервалу (8,7;9,7)

В результате получаем интервальный вариационный ряд, при этом обязательно убеждаемся в том, что ничего не потеряно (n=30) и рассчитываем относительные частоты по каждому интервалу.

интервалы	x_i	n_i	w_i
5,7 - 6,7	6,2	7	0,23
6,7 - 7,7	7,2	11	0,37
7,7 - 8,7	8,2	6	0,2
8,7 - 9,7	9,2	4	0,13
9,7 - 10,7	10,2	2	0,07
суммы:		30	1

Гистограмма относительных частот – это фигура, состоящая из прямоугольников, ширина которых равна длинам *частичных интервалов*, а высота – соответствующим *относительным частотам*:



Полигон относительных частот – это ломаная, соединяющая соседние точки

середины

Важнейшей ***задачей выборочного метода*** *является оценка параметров (характеристик) генеральной совокупности по данным выборки.*

Теоретическую основу применимости выборочного метода составляет *закон больших чисел*, согласно которому при неограниченном увеличении объема выборки практически достоверно, что случайные выборочные характеристики как угодно близко приближаются (сходятся по вероятности) к определенным параметрам генеральной совокупности.