Подробный отчёт

анализа выборочной совокупности

С О Д Е Р Ж А Н И Е

Предисловие.

А. Подробный отчёт наиболее полезен при необходимости иметь под рукой краткую теоретическую сводку при чтении, интерпретации, правке и ином использовании результатов проведённого анализа. Также, как и краткий отчёт, он содержит табличные и графические материалы, но его преимущество в наличии определений, формул, вычислений, пояснений и комментариев.

Б. **Жирным шрифтом** в тексте отчёта выделены термины, обозначения и наименования, а *курсивом* – всевозможные выводы и предположения.

В. Обозначения статистических параметров, мер, оценок и показателей, встречающихся в отчёте, которые могут сбить читателя с толку или вызвать озадаченность вследствие того, что в различных источниках они наименованы и/или выглядят по-разному, даны в таблице ниже (см. табл. А):

Таблица А

Сопоставление обозначений параметров, мер, оценок и показателей с обозначениями, применяемыми в других источниках.

Параметр/мера/оценка/показатель	Обозначение, принятое в отчёте	Иные обозначения
Выборочное среднее	$\bar{x}$	$\bar{X}, \mu, E(X), M(X)$
Среднее генеральной совокупности	$\bar{X}$	$\mu, E(X), M(X)$
Дисперсия выборки	s^2	$S^2, \sigma^2, m_2, M_2, D(x), Var(x)$
Исправленная дисперсия	S^2	$s^2, V(x)$
Дисперсия генеральной совокупности	σ^2	$D(X), Var(X)$
Выборочное СКО	s	S, SD, $\sigma, \sqrt{m_2}, \sqrt{D(x)}, \sqrt{Var(x)}$
Исправленное СКО	S	s, SD, $\sqrt{V(x)}$
СКО генеральной совокупности	σ	$\sqrt{D(X)}, \sqrt{Var(X)}$
Коэффициент вариации	V_s	V_σ, V, cv
Коэффициент асимметрии	As	$A, A_s, m'_3, M'_3, \mu'_3, \tilde{\mu}_3, \alpha_3$
Коэффициент эксцесса	Ex	$Es, E_x, E_k, m'_4, M'_4, \mu'_4, \tilde{\mu}_4, \alpha_4$
Стандартная ошибка среднего	$SE_{\bar{x}}$	$SD_{\bar{x}}$, SE, SEM
Стандартная ошибка асимметрии	SE_{As}	$\sigma_{as}, \sigma_{As}, \sigma_{A_s}\ As_{sd}$
Стандартная ошибка эксцесса	SE_{Ex}	$\sigma_{ex}, \sigma_{Ex}, \sigma_{E_x}, Ex_{sd}$

Раздел 1. Дескриптивный анализ исследуемого массива данных.

1.1. **Дескриптивный (описательный) анализ** подразумевает под собой совокупность методов обработки, систематизации, количественного исследования, а также графического представления числовых данных. Полученные результаты прилагаются в виде **статистических параметров** (числовых показателей или характеристик), которые используются в дальнейших вычислениях с целью нахождения закономерностей в анализируемых числовых совокупностях, а также в виде графиков, гистограмм, схем и таблиц.

1.1. Общая характеристика массива данных.

1.2. С целью ознакомления и проведения общей характеристики исследуемого массива данных основная информация о нём, такая как величина размера совокупности, сумма всех значений, сумма квадратов значений, а также наименьшее и наибольшее значения массива, сводится в таблицу (см. табл. 1.1)

Таблица 1.1

Общая характеристика массива данных.

Параметр/мера/оценка/показатель	Обозначение	Значение
Размер выборки	n	15
Сумма значений	Σx_i	686,30
Сумма квадратов значений	Σx_i^2	33065,24
Наименьшее значение	x_{min}	23,56
Наибольшее значение	x_{max}	67,65

1.3. **Размер совокупности** (обозначается как $\boldsymbol{n}$ для выборочных совокупностей и $\boldsymbol{N}$ для генеральных) – это количественная характеристика, показывающая число элементов, включённых в исследуемую совокупность (единиц, составляющих одно целое или измерений в одном наблюдении и т.д.).

1.4. Размер выборочной совокупности нередко позволяет говорить об ошибочности составленных выводов, ещё не глядя на них. Именно поэтому, немаловажно заранее определить оптимальный размер выборки. В зависимости от типа исследования предлагаются различные рекомендации, выражаемые как в относительных (например, $n = 5\% \cdot N$), так и в абсолютных ($n = 100$) величинах. Помимо них, даются и вовсе абстрактные (например: чем более однородным является объект исследования, тем меньшим допустимо быть размеру выборки).

1.5. Для получения визуального представления об исследуемом массиве данных все индивидуальные значения (x_i) отображены в виде кривой на **линейной диаграмме** (**графике**; см. рис. 1.1).

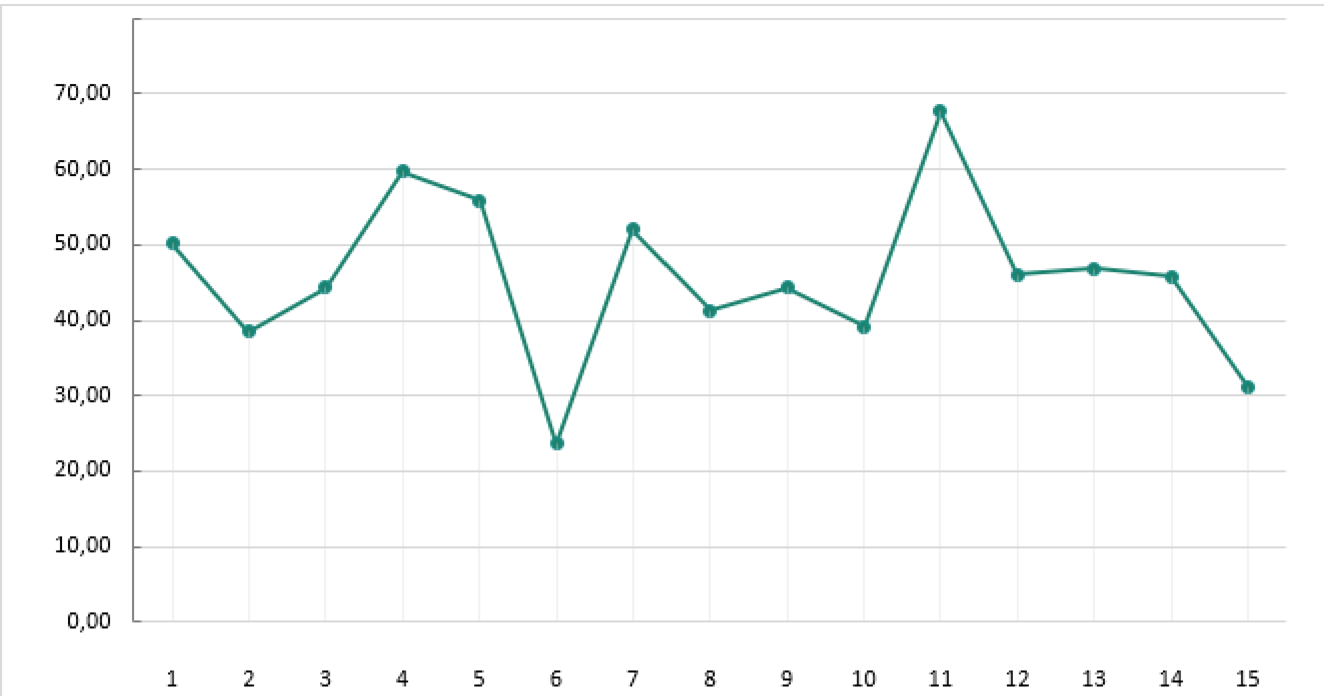

Рис. 1.1 Линейная диаграмма исследуемого массива данных.

1.6. Для отображения характера распределения частот в анализируемом массиве данных будет построена **гистограмма** (см. рис. 1.2), состоящая из прямоугольников, основанием для которых служат интервалы группировки в точно обозначенных границах, а высотой – вычисленные значения плотности распределения, т.е. частота появления значений в настоящих интервалах.

1.7. Для определения необходимого количества интервалов (***m***) в зависимости от размера выборочной совокупности, следует воспользоваться правилом Стёрджесса, записанном через десятичный логарифм. Формула и вычисление приведены ниже (1.1):

$$m = 1 + 3{,}322 \cdot \lg n = 1 + 3{,}322 \cdot \lg 15 = 4{,}91 \tag{1.1}$$

1.8. Полученный результат следует округлить до целого в меньшую сторону, т.е. в данном случае до 4. По нему, задействуя наибольшее и наименьшее значения массива, можно найти длину (шаг) интервала (***h***), расчёт ниже (1.2):

$$h = \frac{x_{max} - x_{min}}{m} = \frac{67.65 - 23.56}{4} = 11{,}02 \tag{1.2}$$

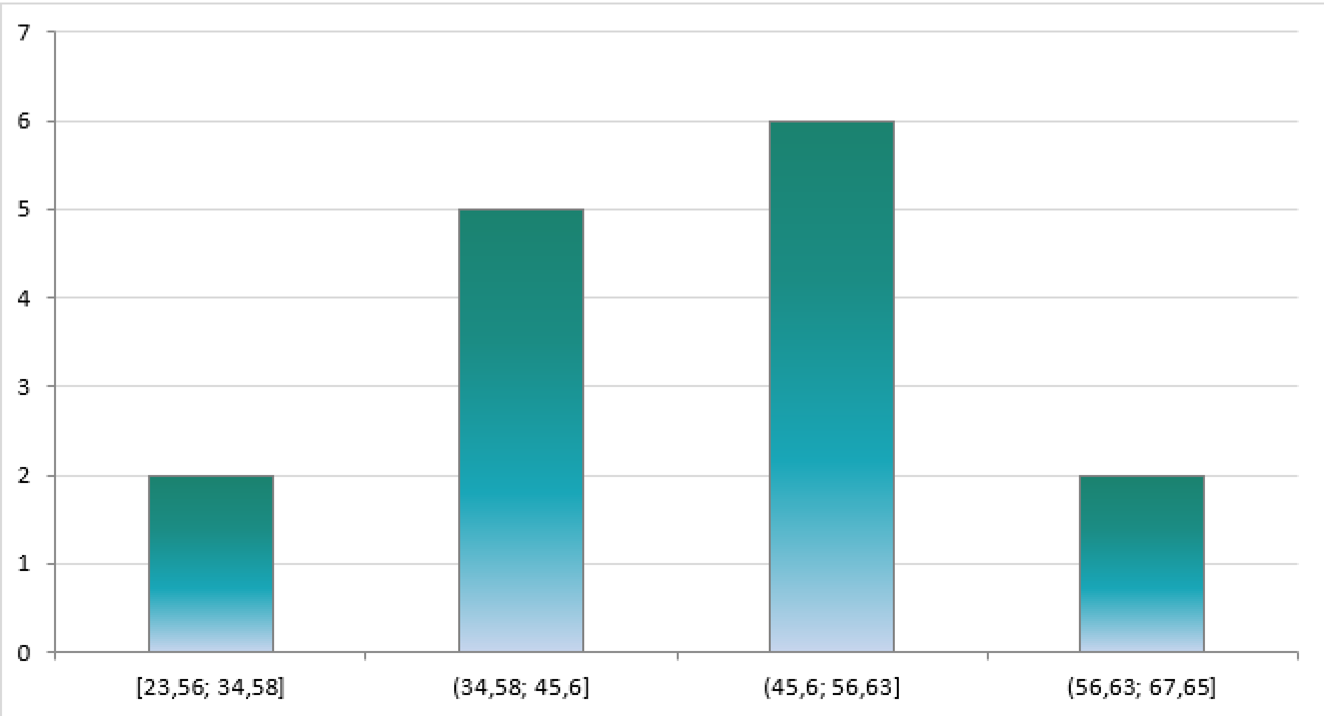

Рис. 1.2 Гистограмма распределения анализируемого массива данных.

1.2. Меры центральной тенденции массива данных.

1.9. С целью получить представление о центре распределения исследуемого массива данных, т.е. устранить индивидуальные различия в нём, а также выявить общие условия и закономерности, будут найдены следующие показатели центральной тенденции: среднее геометрическое, среднее арифметическое и среднее квадратическое значения, а также средний диапазон, мода(-ы) и медиана.

1.10. **Среднее геометрическое ($\bar{x}_{\text{геом}}$)** – это корень произведений всех индивидуальных значений массива, степень которого определяется количеством множителей, т.е. размером совокупности. Среднее геометрическое всегда меньше или равно среднему арифметическому. Применяется в качестве основной средней величины преимущественно в случаях, когда массив состоит только из положительных чисел. Формула и вычисление даны ниже (1.3):

$$\bar{x}_{\text{геом}} = \sqrt[n]{\prod x_i} = \sqrt[20]{50{,}12 \cdot 38{,}50 \cdot \ldots \cdot 31{,}00} = 44{,}44 \quad (1.3)$$

1.11. **Среднее арифметическое ($\bar{x}_{\text{арифм}}$)** – это отношение суммы всех значений выборочной числовой совокупности к её размеру (1.4). Это наиболее используемый параметр центра распределения. Его недостаток в сравнении со структурными средними заключается в высокой чувствительности к чрезвычайно большим или малым значениям в массиве, т.к. он может резко увеличиваться или уменьшаться за их счет.

$$\bar{x}_{\text{арифм}} = \frac{1}{n} \Sigma x_i = \frac{686{,}30}{15} = 45{,}75 \quad (1.4)$$

1.12. **Среднее квадратическое ($\bar{x}_{\text{квадр}}$)** – это квадратный корень из суммы квадратов всех значений массива. Применяется данный параметр, когда исходные значения признака могут быть как положительными, так и отрицательными. Среднее квадратическое всегда больше или равно среднему арифметическому. Ниже приведены формула и вычисление показателя (1.5):

$$\bar{x}_{\text{квадр}} = \sqrt{\frac{1}{n}\Sigma x_i^2} = \sqrt{\frac{33065{,}24}{15}} = 46{,}95 \qquad (1.5)$$

1.13. **Средний диапазон ($\bar{x}_m$)** – это наиболее простая для понимания и вычисления мера центральной тенденции, вычисляемая полусуммой крайних значений исследуемого массива данных (1.6). Наиболее эффективна в качестве оценочной функции центра равномерного распределения, а также в роли средней величины при грубой оценке погрешности выборки. В остальных случаях параметр нецелесообразен в применении на практике, так как является слишком чувствительным к выбросам (аномальным значениям в массиве).

$$\boldsymbol{\bar{x}_m} = \frac{x_{max}+x_{min}}{2} = \frac{67{,}65 + 23{,}56}{2} = 45{,}60 \qquad (1.6)$$

1.14. **Модой (Mo)** называется наиболее часто встречающееся значение в числовой совокупности. Следует учитывать, что в одном массиве данных мод может быть несколько или же ни одной. В исследуемой числовой совокупности моде соответствует значение 44,35.

1.15. Наконец, **медиана (Me)** – это параметр, делящий совокупность на две равные части по объему частот. Если массив данных состоит из нечетного количества чисел, то медиана будет находится ровно по середине, а если из четного, то будет вычислена путем нахождения среднего арифметического двух чисел, лежащих по середине. Наиболее удобно использовать медиану при ассиметричном распределении данных, т.к. при наличии выбросов, медиана будет устойчива к воздействию отклоняющихся данных. В исследуемом массиве данных медиана равна 45,76.

1.16. Для удобства восприятия, все результаты вычислений в подразделе сведены ниже в таблицу (см. табл. 1.2). А на рис. 1.3 представлена структура отклонений значений массива данных в положительную и отрицательную стороны от среднего арифметического.

Таблица 1.2

Меры центральной тенденции (параметры центра распределения).

Параметр/мера/оценка/показатель	Обозначение	Значение
Среднее геометрическое	$\bar{x}_{геом}$	44,44
Среднее арифметическое	$\bar{x}_{арифм}$	45,75
Среднее квадратическое	$\bar{x}_{квадр}$	46,95
Средний диапазон	$\bar{x}_{m}$	45,60
Мода	Mo	44,35
Медиана	Me	45,76

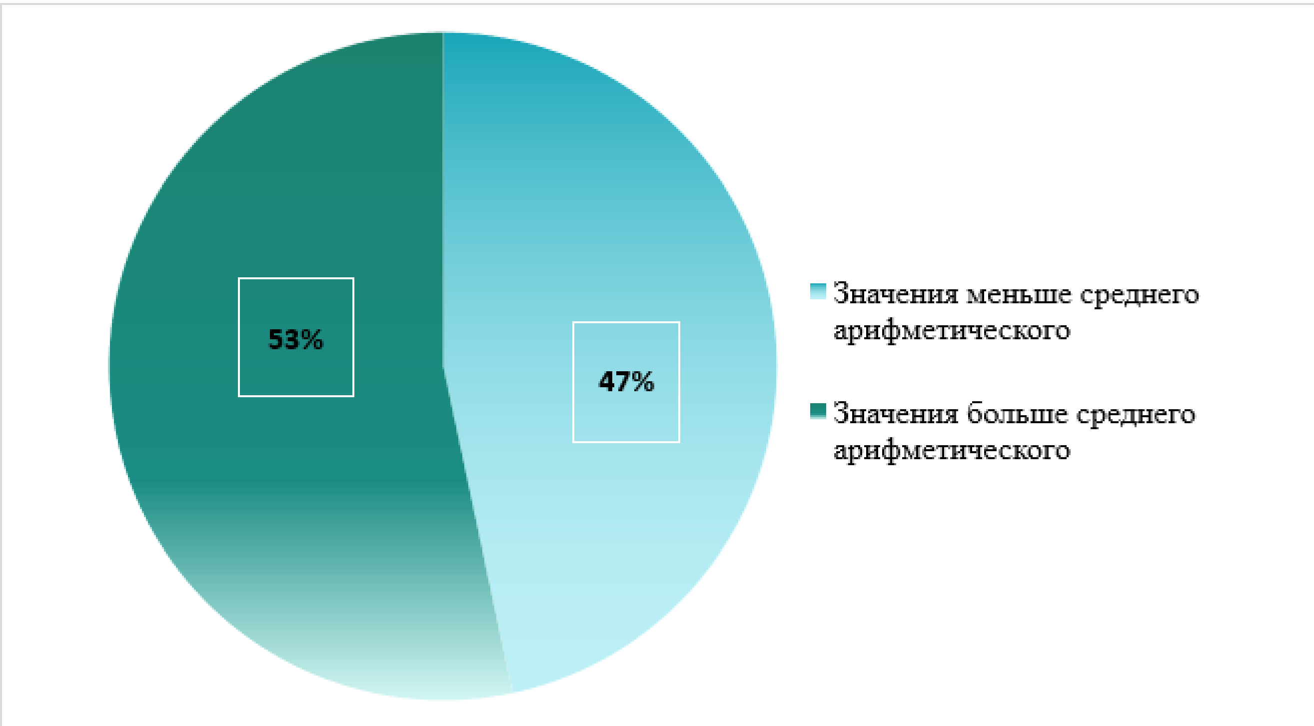

Рис. 1.3 Структура отклонений индивидуальных значений массива данных от среднего арифметического.

1.3. Показатели вариации массива данных.

1.17. Для нахождения различий между индивидуальными значениями единиц исследуемой совокупности с друг другом и со средней величиной задействуются **показатели вариации**. Их практическое предназначение – давать представление о точности и надежности вычисленного среднего арифметического в зависимости от величины различий между усредняемыми индивидуальными значениями массива.

1.18. Среди показателей вариации выделяют **абсолютные** и **относительные**. К абсолютным относятся размах вариации, среднее линейное отклонение, среднее квадратическое отклонение и дисперсия. А к относительным: коэффициент осцилляции, линейный коэффициент вариации, коэффициент вариации. Ниже каждый из них будет рассмотрен по отдельности.

1.19. **Размах вариации (R)** показывает пределы, в которых варьируются крайние значения в анализируемой числовой совокупности и, соответственно, находится разностью между максимальным и минимальным значениями (1.7). Как показатель вариации, он самый простой, но не самый оптимальный, т.к. его существенные недостатки – это зависимость от выбросов и неспособность учитывать все колебания в числовой совокупности.

$$R = x_{max} - x_{min} = 67{,}65 - 23{,}56 = 44{,}09 \qquad (1.7)$$

1.20. *Абсолютная* колеблемость *крайних значений составляет* 44,09 *в абсолютной величине (в той же единице измерения, что и все индивидуальные значения в массиве).*

1.21. Для дальнейшего анализа вариации данных следует ознакомиться с понятием **отклонения от средней величины** ($\boldsymbol{d}$), представляющего собой разность индивидуального значения массива от среднего арифметического (1.8). Сами отклонения, их модули, квадраты, кубы, биквадраты и их суммы приведены во вспомогательных таблицах для каждой позиции массива. А графическое представление отклонений в виде гистограммы приведено ниже (см. рис. 1.4).

$$d_{x_i} = x_i - \bar{x} \tag{1.8}$$

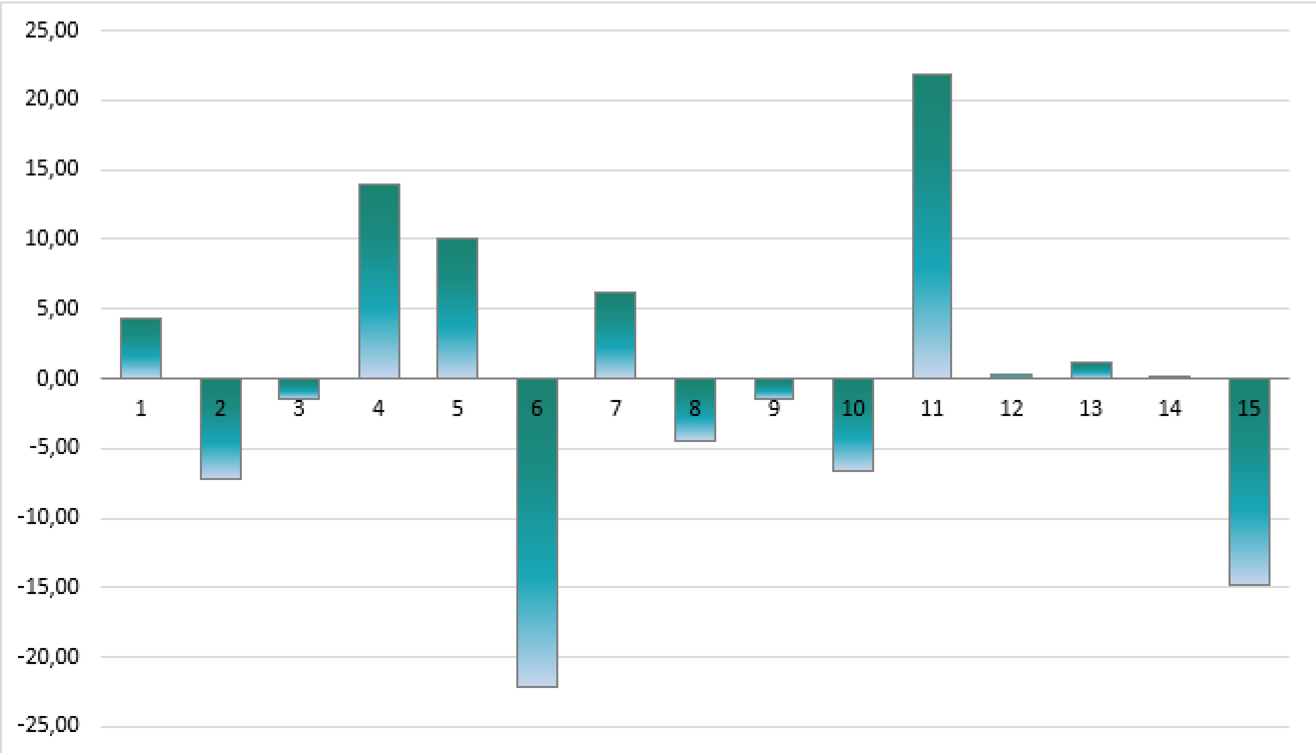

Рис. 1.4 Гистограмма отклонений индивидуальных значений массива данных от среднего арифметического.

1.22. **Среднее линейное отклонение ($\bar{d}$)** – это отношение суммы модулей отклонений индивидуальных значений от среднего арифметического к размеру совокупности. Поскольку суммирование значений здесь происходит без учёта знаков и сам показатель не соответствует ни одному вероятностному закону распределения данных, то и на практике применяется реже остальных в подразделе. Формула и вычисление приведены ниже (1.9).

$$\bar{d} = \frac{1}{n}\Sigma|x_i - \bar{x}| = \frac{116{,}07}{15} = 7{,}74 \tag{1.9}$$

1.23. *Абсолютные отклонения от среднего арифметического в данном массиве составляют в среднем* 7,74 *в той же единице измерения, что и индивидуальные значения массива данных.*

1.24. **Среднее квадратическое отклонение (s)**, иначе называемое **стандартным отклонением**, **сигмой** или сокращаемое до аббревиатуры **СКО**, находится извлечением квадратного корня из среднего квадрата отклонений всех значений числовой совокупности от среднего арифметического. Является наиболее безупречной из абсолютных характеристик вариации, поскольку применима в вероятностных расчётах, статистическом выводе и в корреляционно-регрессионном анализе. (1.10):

$$s = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2} = \sqrt{\frac{1664{,}85}{15}} = 10{,}54 \qquad (1.10)$$

1.25. **Дисперсия (s^2)** – средний квадрат отклонений индивидуальных значений от их среднего арифметического (или же СКО в квадрате). При высоком значении дисперсии можно выдвинуть предположение о существенности случайной компоненты в анализируемом массиве данных, т.е. о возможном наличии шума или выбросов. Ниже даны формула нахождения и вычисление (1.11):

$$s^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2 = \frac{1664{,}85}{15} = 110{,}99 \qquad (1.11)$$

1.26. Относительные показатели вариации вычисляются как отношения трёх из рассмотренных выше абсолютных показателей вариации к среднему арифметическому или, иными словами, представляют собой их долю в среднем значении. И первый из них – это **коэффициент осцилляции (V_R)**, который показывает относительную колеблемость крайних значений вокруг средней величины. Находится он посредством деления размаха вариации на среднее арифметическое (1.12):

$$V_R = \frac{R}{\bar{x}_{\text{арифм}}} = \frac{44{,}09}{45{,}75} = 0{,}96 \quad (1.12)$$

1.27. *Из вышеприведённого следует, что разница между крайними значениями исследуемого массива данных составляет* 96% *средней величины.*

1.28. **Линейный коэффициент вариации ($V_{\bar{d}}$)** характеризует долю усредненного значения абсолютных отклонений от среднего арифметического в нём, *здесь она достигает всего 17%, как видно из расчёта ниже* (1.13):

$$V_{\bar{d}} = \frac{\bar{d}}{\bar{x}_{\text{арифм}}} = \frac{7{,}74}{45{,}75} = 0{,}17 \quad (1.13)$$

1.29. Наконец, **коэффициент вариации (V_s)** вычисляемый по среднему квадратическому отклонению, вызывает набольший интерес в применении на практике среди рассмотренных относительных показателей вариации, потому что подлежит сравнению со следующими критическими неравенствами. Так, если $V_s < 0{,}10$, то изменчивость массива данных принято считать незначительной, следовательно, совокупность – однородной. В случае, когда коэффициент принимает значения в интервале $0{,}10 < V_s < 0{,}20$, то совокупность относится к достаточно однородной, а если $0{,}20 < V_s < 0{,}33$ – к значительно однородной. При значениях $V_s > 0{,}33$ речь идёт о неоднородности данных и необходимости исключения самых больших и самых малых значений в числовой совокупности. Ниже приведены формула и вычисление (1.14):

$$V_s = \frac{s}{\bar{x}_{\text{арифм}}} = \frac{10{,}54}{45{,}75} = 0{,}23 \quad (1.14)$$

1.30. *Коэффициент вариации, равный 0,23, позволяет предполагать о значительной однородности или умеренной колеблемости данных в исследуемом массиве*. Для наглядности данный факт отображен также и в графическом виде (см. рис. 1.5), а значения рассмотренных выше абсолютных и относительных показателей вариации сведены в одну таблицу ниже (см. табл. 1.3):

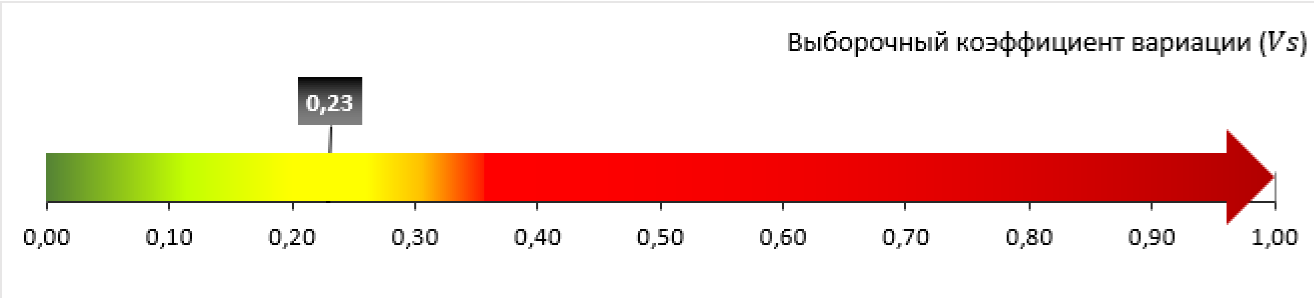

Рис. 1.5 Графическое отображение коэффициента вариации на прямой.

Таблица 1.3

Абсолютные и относительные показатели вариации.

Параметр/мера/оценка/показатель	Обозначение	Значение
Размах вариации выборки	R	44,09
Среднее линейное отклонение выборки	$\bar{d}$	7,74
Среднее квадратическое отклонение выборки	s	10,54
Выборочная дисперсия	s^2	110,99
Коэффициент осцилляции	V_R	0,96
Линейный коэффициент вариации	$V_{\bar{d}}$	0,17
Коэффициент вариации	V_s	0,23

1.4. Меры отклонения формы распределения данных.

1.31. Характеристика эмпирического (наблюдаемого) распределения предполагает вычисление параметров отклонения формы распределения данных – коэффициентов асимметрии и эксцесса. Учитывая, что названные параметры в **нормальном распределении** (или **распределении Гаусса**) равны нулю ($As = 0$; $Ex = 0$), нахождение их значений для эмпирического распределения позволяет охарактеризовать то, насколько оно отличается от нормального распределения. Функция плотности нормального распределения, формула которой дана ниже

(1.15), симметрична относительно средней величины, из-за чего распределение также называют колоколообразным, данный факт представлен графически ниже (см. рис. 1.6). Здесь также наблюдается **правило трёх сигм**: 68,26% значений нормально распределённого массива будут находится на расстоянии не более одного СКО, 95,44% – в пределах двух, и 99,72% – в пределах трёх СКО.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1.15)$$

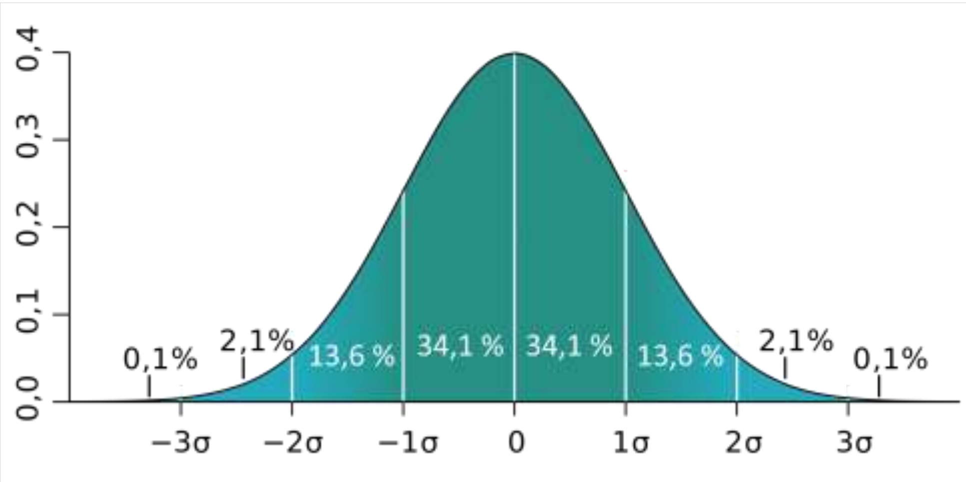

Рис. 1.6 График плотности вероятностей нормального распределения и проценты попадания случайной величины на отрезки, равные СКО.

1.32. **Коэффициент асимметрии** (**As**), представляет собой отношение центрального момента третьего порядка (иначе суммы отклонений индивидуальных значений от среднего арифметического в кубе, делённой на размер выборки) к среднему квадратическому отклонению в кубе. Для интерпретации значения данного параметра следует воспользоваться следующими неравенствами: при $As > 0$ наблюдается правостороняя асимметрия, соответственно, если $As < 0$, то асимметрия – левосторонняя. В случае, когда коэффициент асимметрии принимает нулевое значение ($As = 0$), асимметрия отсутствует, следовательно, распределение является симметричным. По величине коэффициента асимметрии также можно судить о степени скошенности кривой плотности распределения. Так, например, если $As < 0{,}25$, то асимметрия считается незначительной. Если $0{,}25 < As < 0{,}5$, то асимметрия считается умеренной, а если $As > 0{,}5$, асимметрия – значительная. Формула и нахождение (1.16), а также графическое представление эмпирического распределения с отложением среднего арифметического (см. рис. 1.7) даны ниже:

$$As = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})^3}{s^3} == \frac{\frac{1}{15} \cdot (-359{,}77)}{10{,}54^3} = -0{,}02 \qquad (1.16)$$

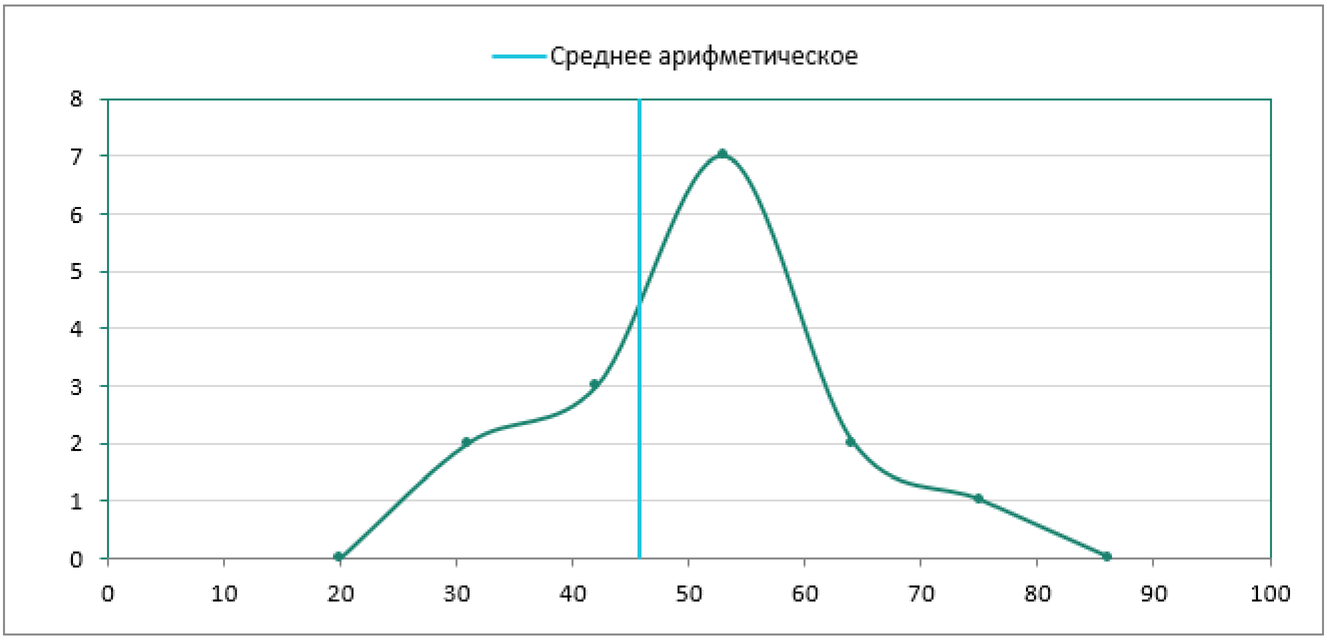

Рис. 1.7 График эмпирического распределения с отложением среднего арифметического.

1.33. *По полученному результату следует сделать вывод о наличии незначительной по величине левосторонней асимметрии в эмпирическом распределении анализируемого массива данных.*

1.34. Несколько слов следует сказать и о коэффициентах асимметрии, вычисленных по структурным средним, т.е. на основе разности среднего и моды, а также среднего и медианы. На практике они применяются реже, так как не учитывают периферию**: асимметрия по моде** $(\boldsymbol{As_{Mo}})$ описывает только наиболее встречающуюся часть распределения, а **асимметрия по медиане** $(\boldsymbol{As_{Me}})$ – лишь его центральную. Формулы, вычисления (1.17, 1.18) и графическое представление эмпирического распределения с отложением медианы и моды (см. рис. 1.8) представлены ниже:

$$As_{Mo} = \frac{\bar{x} - Mo}{s} = \frac{45{,}75 - 44{,}35}{10{,}54} = 0{,}13 \tag{1.17}$$

$$As_{Me} = 3 \cdot \frac{\bar{x} - Me}{s} = 3 \cdot \frac{45{,}75 - 45{,}76}{10{,}54} = -0{,}001 \tag{1.18}$$

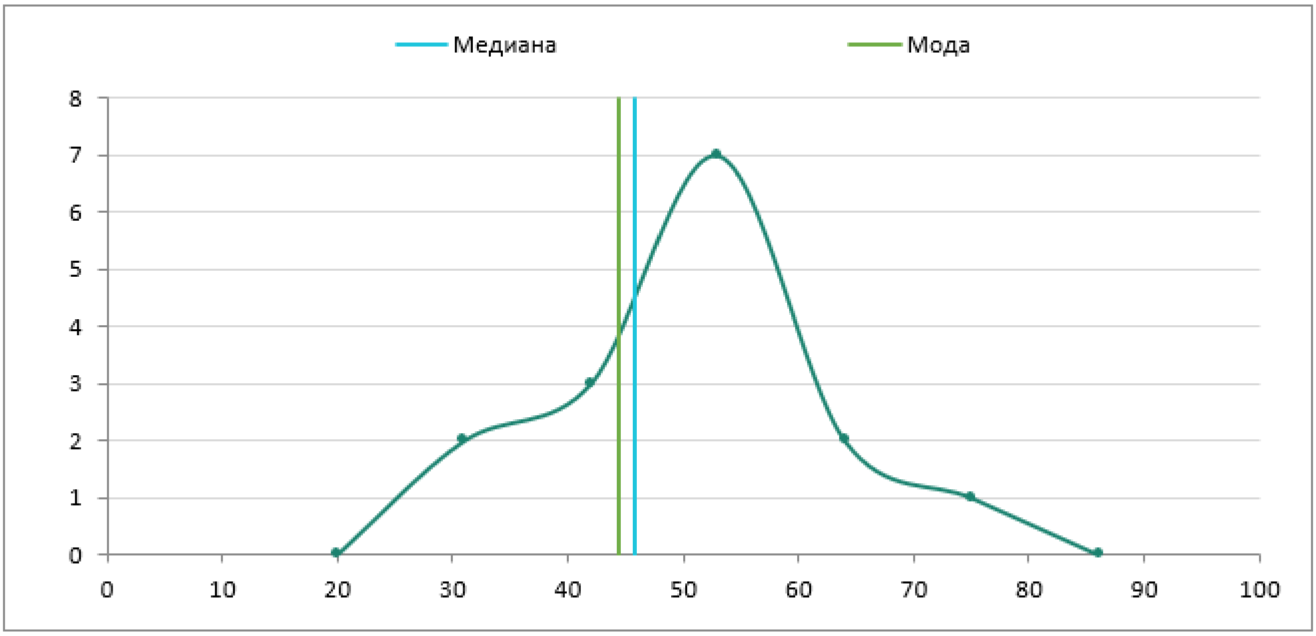

Рис. 1.8 График эмпирического распределения с отложением медианы и моды.

1.35. **Коэффициент эксцесса** $(\boldsymbol{Ex})$ следует рассматривать как меру средней толщины или относительного веса хвостов распределения. Нулевое значение параметра, как было замечено ранее, соответствует толщине хвостов нормального распределения. В случае если эксцесс больше нуля, то у эмпирического распределения наблюдаются более толстые хвосты, нежели в нормальном, а если меньше нуля, следовательно, хвосты у распределения тоньше, чем у нормального распределения. Ниже приведены формула и вычисление данного параметра (1.19):

$$Ex = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})^4}{s^4} - 3 = \frac{\frac{1}{15} \cdot 574837{,}31}{10{,}54^4} = 0{,}11 \qquad (1.19)$$

1.36. *Полученный результат свидетельствуют о том, что у эмпирического распределения наблюдаются бо́льшие по толщине хвосты в сравнении с нормальным распределением*. Все результаты вычислений подраздела сведены в таблицу ниже (см. табл. 1.4):

Таблица 1.4

Меры отклонения формы распределения данных.

Параметр/мера/оценка/показатель	Обозначение	Значение
Коэффициент асимметрии	As	–0,02
Асимметрия по моде	As_{Mo}	0,13
Асимметрия по медиане	As_{Me}	–0,001
Коэффициент эксцесса	Ex	0,67

1.5. Квартили и квартильные статистические параметры.

1.37. Описательный статистический анализ можно также провести задействуя квартили и квартильные параметры. Они используются преимущественно в случаях, когда в качестве главной меры центра распределения целесообразнее рассматривать медиану.

1.38. **Квартили** – это значения массива данных, делящие его на четыре примерно равные части. Первый квартиль, иначе называемый нижним – это значение, ниже которого располагаются 25 % значений совокупности или, иначе говоря, это медиана первой её половины. Второй квартиль соответствует медиане всей совокупности, т.е. делит её на две равные части по объему частот. Соответственно, ниже третьего (верхнего) квартиля располагаются 75% значений. Разумеется, чем больше размер совокупности и меньше повторяющихся значений в ней, тем точнее её деление квартилями на четверти.

1.39. Для графического представления данных при задействовании квартилей применяют **блочную диаграмму, иначе называемую «ящиком с усами»**, изображенную ниже (см. рис. 1.9). Данный график компактно помещает в себе минимальное и максимальное значения массива (на рисунке изображены как верхний и нижний «усы», т.е. вертикальные планки), квартили (первому квартилю соответствует низ «ящика», а третьему – верх), медиану (обозначена линией внутри ящика), среднее арифметическое (крест внутри ящика) и выбросы (точки, находящиеся за пределами усов).

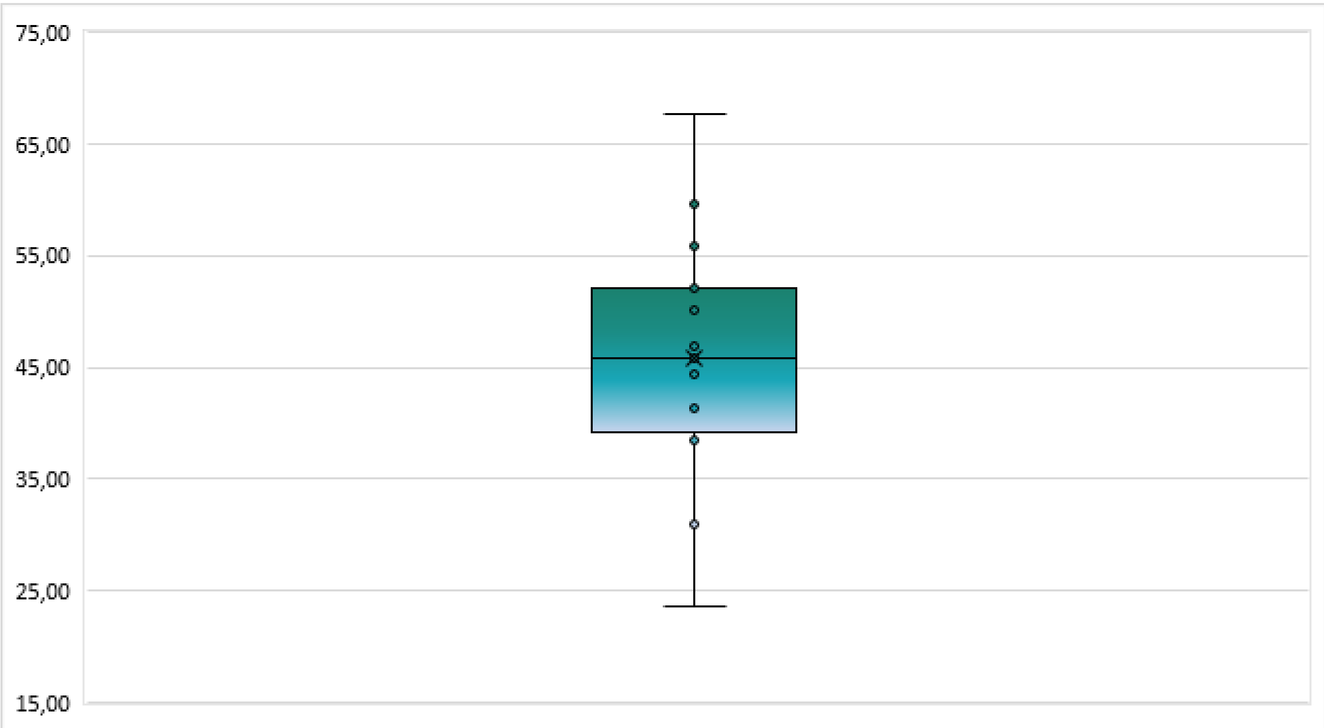

Рис. 1.9. Блочная диаграмма исследуемого массива данных («ящик с усами»).

1.40. **Интерквартильный размах (IQR)** – это разность между третьим и первым квартилями совокупности. Наиболее целесообразно использование данного параметра вместо размаха вариации при характеристики числовых совокупностей с небольшим размером. Формула и вычисление даны ниже (1.20):

$$IQR = Q_3 - Q_1 = 51{,}06 - 40{,}24 = 10{,}82 \tag{1.20}$$

1.41. **Квартильным отклонением** (d_Q) называется половина интерквартильного размаха (1.18). Практическая значимость параметра состоит в его способности характеризовать вариацию центральной части исследуемого массива данных:

$$d_Q = \frac{IQR}{2} = \frac{10{,}82}{2} = 5{,}41 \tag{1.21}$$

1.42. **Квартильный коэффициент отклонения** (K_{d_Q}), находящийся отношением интерквартильного размаха к сумме первого и третьего квартилей (1.22), также характеризует вариацию, но его преимущество в том, что он наиболее точен при сравнении разброса данных между двумя или более массивами данных за счет наличия меньшей чувствительности к выбросам:

$$K_{d_Q} = \frac{IQR}{Q_1 + Q_3} = \frac{10{,}82}{40{,}24 + 51{,}06} = 0{,}12 \tag{1.22}$$

1.43. **Квартильный коэффициент вариации** (V_Q), показывающий относительную колеблемость квартилей вокруг центральных 50% элементов анализируемой числовой совокупности, рассчитывается путём деления квартильного отклонения на медиану (1.23):

$$V_Q = \frac{d_Q}{Me} = \frac{5{,}41}{45{,}76} = 0{,}12 \tag{1.23}$$

1.44. И наконец, **квартильный коэффициент асимметрии (As_Q)**, также называемый **мерой асимметрии Боули**, необходим для установления симметричности или несимметричности исследуемого массива данных по его центру распределения. (1.24):

$$As_Q = \frac{Q_3+Q_1-2Me}{IQR} = \frac{2{,}51 + 5{,}38 - (2 \cdot 3.27)}{2{,}86} = -0{,}02 \qquad (1.24)$$

1.45. Приведённые выше параметры, рассчитанные по квартилям, можно использовать в дополнение к рассмотренными ранее дескриптивным показателям либо вместо них, следовательно, и интерпретировать полученные значения подобным образом. Их табличное представление приведено ниже (см. табл. 1.5):

Таблица 1.5

Квартили и квартильные параметры для исследуемого массива данных.

Параметр/мера/оценка/показатель	Обозначение	Значение
Первый (нижний) квартиль	Q_1	40,24
Второй квартиль (медиана)	Q_2	45,76
Третий (верхний) квартиль	Q_3	51,06
Интерквартильный размах	IQR	10,82
Квартильное отклонение	d_Q	5,41
Квартильный коэффициент отклонения	K_{d_Q}	0,12
Квартильный коэффициент вариации	V_Q	0,12
Квартильный коэффициент асимметрии	As_Q	-0,02

Раздел 2. Анализ динамики массива данных.

2.1. Числовая совокупность, значения в которой расположены в хронологической последовательности и изменяются во времени, называется **динамическим рядом** (или **рядом динамики**). Для анализа изменений отдельных **уровней** такого ряда используются **показатели динамики** (приведены во вспомогательных таблицах для каждого индивидуального значения массива). А для характеристики общей тенденции развития исследуемого объекта на протяжении всего временного интервала задействуются **средние показатели динамики** ряда данных, которые рассмотрены подробнее ниже.

2.2. **Средний абсолютный прирост ($\overline{\Delta x_i}$)** показывает, на сколько единиц увеличивался или уменьшался один уровень динамического ряда по сравнению с предыдущим в среднем за одну единицу времени (час, день, месяц, год и т.д.). Вычисляется путем деления общего прироста за весь период на длину этого периода в тех или иных единицах времени (2.1):

$$\overline{\Delta x_i} = \frac{\sum \Delta x_i}{n-1} \qquad (2.1)$$

2.3. **Средний коэффициент динамики ($\overline{К_д}$)** – это среднее геометрическое из показателей коэффициентов динамики отдельных периодов, показывающий во сколько раз в среднем в рассматриваемом периоде изменялись уровни динамического ряда (2.2).

$$\overline{К_д} = \sqrt[n-1]{К_1 \cdot К_2 \cdot \ldots \cdot К_{n-1}} \qquad (2.2)$$

2.4. **Средний темп роста ($\overline{Т_р}$)** – это не что иное, как выраженный в процентах средний коэффициент динамики (2.3). А **средний темп прироста** показывает, насколько процентов в среднем увеличивался или снижался один уровень динамического ряда по сравнению с предыдущим за единицу времени в течение всего исследуемого временного интервала (2.4).

$$\overline{\mathrm{Т_р}} = \overline{\mathrm{К_д}} \cdot 100\ \% \tag{2.3}$$

$$\overline{\mathrm{Т_{пр}}} = \overline{\mathrm{Т_р}} - 100\% \tag{2.4}$$

2.5. Для графического представления **общей тенденции** развития объекта исследования в течение всего рассматриваемого промежутка времени необходимо выполнить **выравнивание** динамического ряда данных путём построения кривой, в которой его фактические уровни заменены теоретическими, рассчитанными в настоящем исследовании тремя методами.

2.6. **Метод скользящей средней** заключается в замене фактических уровней ряда их средними арифметическими за определённое число временных единиц (обычно не более чем за три периода). **Методы среднего абсолютного прироста и среднего коэффициента динамики** подразумевают использование двух рассмотренных выше средних показателей динамики ряда (на формулах 2.1 и 2.2) для описания тенденции динамики линейной и показательной функциями: вначале первый уровень ряда, а затем каждое последующее выравненное значение суммируется с средним абсолютным приростом или в случае с коэффициентом динамики – умножается на него (приведено ниже на формулах 2.5 и 2.6):

$$x_{\text{в}1}^{\overline{\Delta x_\iota}} = x_1 + \overline{\Delta x_\iota}; \quad x_{\text{в}i}^{\overline{\Delta x_\iota}} = x_{\text{в}i-1}^{\overline{\Delta x_\iota}} + \overline{\Delta x_\iota}; \tag{2.5}$$

$$x_{\text{в}1}^{\overline{\mathrm{К_д}}} = x_1 \cdot \overline{\mathrm{К_д}}; \quad x_{\text{в}i}^{\Delta x_i} = x_{\text{в}i-1}^{\overline{\mathrm{К_д}}} \cdot \overline{\mathrm{К_д}}; \tag{2.6}$$

2.7. Задействуя приведённые выше методы выявления тенденции, также возможно осуществить **перспективную экстраполяцию** (обычно на не более 5 временных периодов). Под экстраполяцией следует понимать нахождение уровней динамического ряда за его пределами, т.е. продление его на основе выявленной тенденции. Перспективная экстраполяция означает продление в будущее (используется в простом прогнозировании), а ретроспективная – в прошлое.

2.8. Следует приступить к анализу динамики ряда с вычисления средних показателей динамики (2.7 – 2.10):

$$\overline{\Delta x_\iota} = \frac{\sum \Delta x_i}{n-1} = \frac{-19{,}12}{14} = -1{,}37 \qquad (2.7)$$

$$\overline{К_д} = \sqrt[n-1]{К_1 \cdot К_2 \cdot \ldots \cdot К_{n-1}} = \sqrt[14]{0{,}77 \cdot 1{,}15 \cdot \ldots \cdot 0{,}68} = 0{,}97 \qquad (2.8)$$

$$\overline{Т_р} = \overline{К_д} \cdot 100\ \% = 0{,}9663 \cdot 100\ \% = \ 96{,}63\ \% \qquad (2.9)$$

$$\overline{Т_{пр}} = \overline{Т_р} - 100\% = 96{,}63\% - 100\% = -3{,}37\% \qquad (2.10)$$

2.9. *Значения приведённых выше показателей позволяют заключить, что объект исследования в течение всего временного периода в среднем уменьшался на* 1,37 *в абсолютной величине или на* 3,37%. *Также можно сказать, что среднее уменьшение значений в течении всего временного периода составило* 0,97 *или* 96,63% *от исходного уровня*. Для наглядности представления результатов, они также приведены ниже в табличном виде (см. табл. 2.1):

Таблица 2.1

Средние показатели динамики ряда данных.

Параметр/мера/оценка/показатель	Обозначение	Значение
Средний абсолютный прирост	$\overline{\Delta x_\iota}$	–1,37
Средний коэффициент динамики (роста)	$\overline{К_д}$	0,97
Средний темп роста	$\overline{Т_р}$	96,63
Средний темп прироста	$\overline{Т_{пр}}$	–3,37

2.10. Выравнивания динамического ряда данных приведены в вспомогательных таблицах к отчёту, а их кривые даны ниже как единой схемой с задействованием всех трёх перечисленных методов (см. рис. 2.1), так и отдельными графиками с использованием каждого метода выравнивания индивидуально (см. рис. 2.2, 2.3 и 2.4):

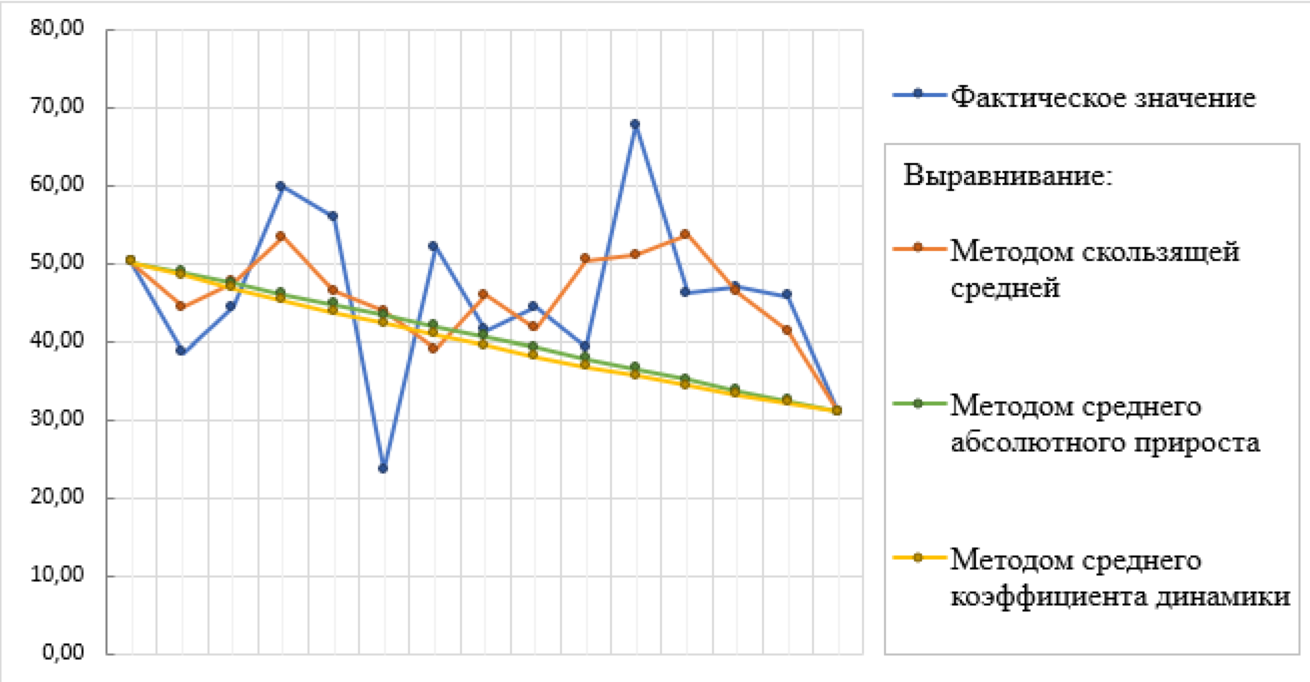

Рис. 2.1 Графики выравнивания ряда данных тремя методами одновременно.

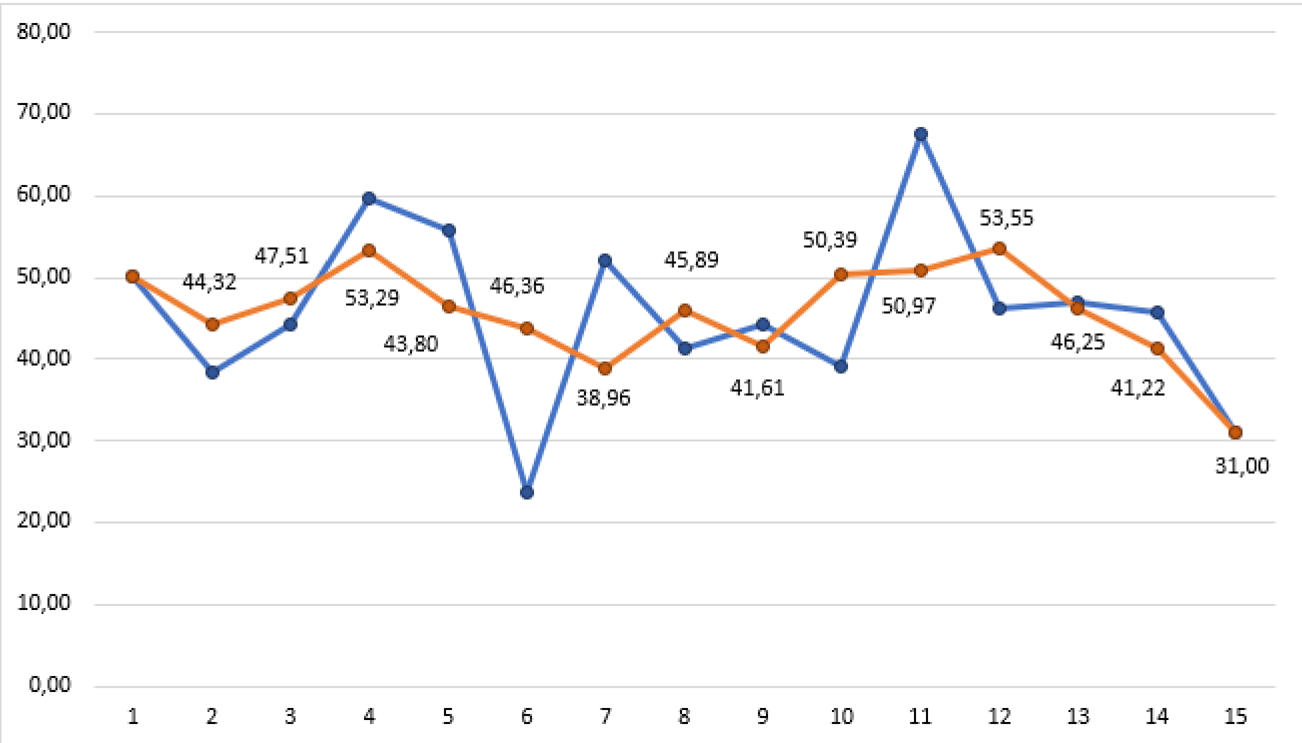

Рис. 2.2 График выравнивания методом скользящей средней.

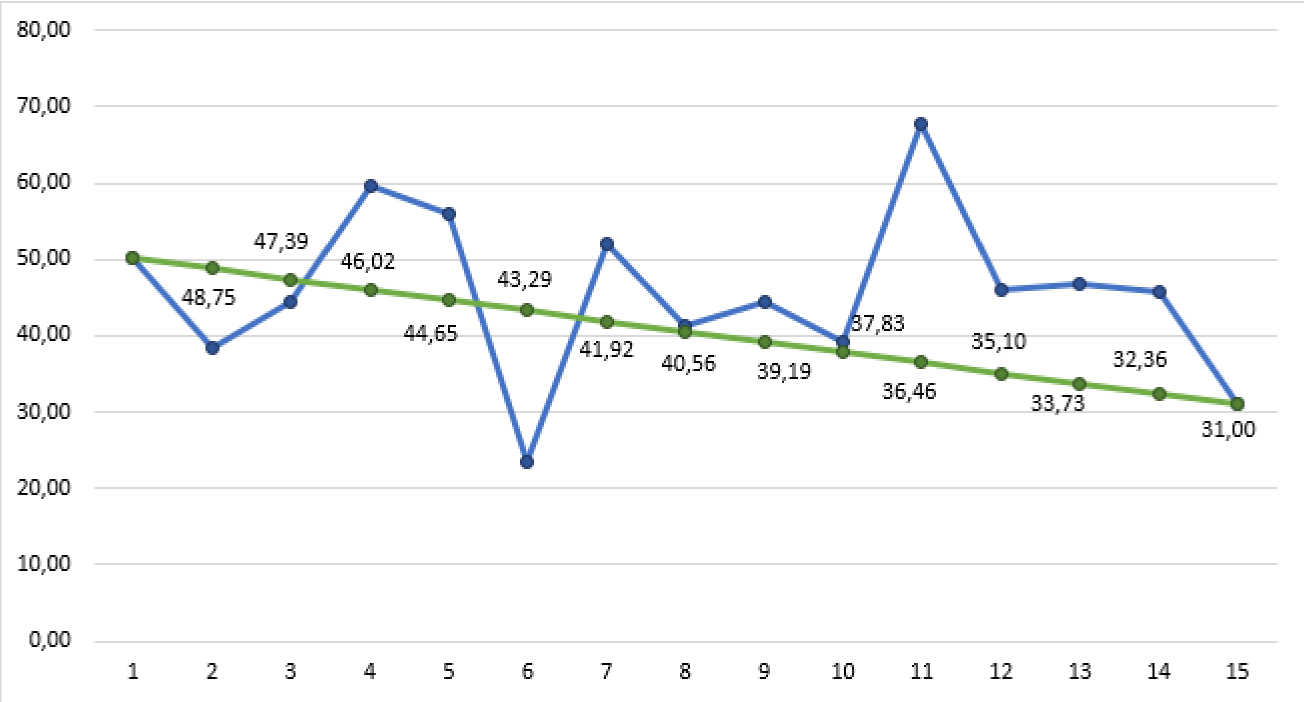

Рис. 2.3 График выравнивания методом среднего абсолютного прироста.

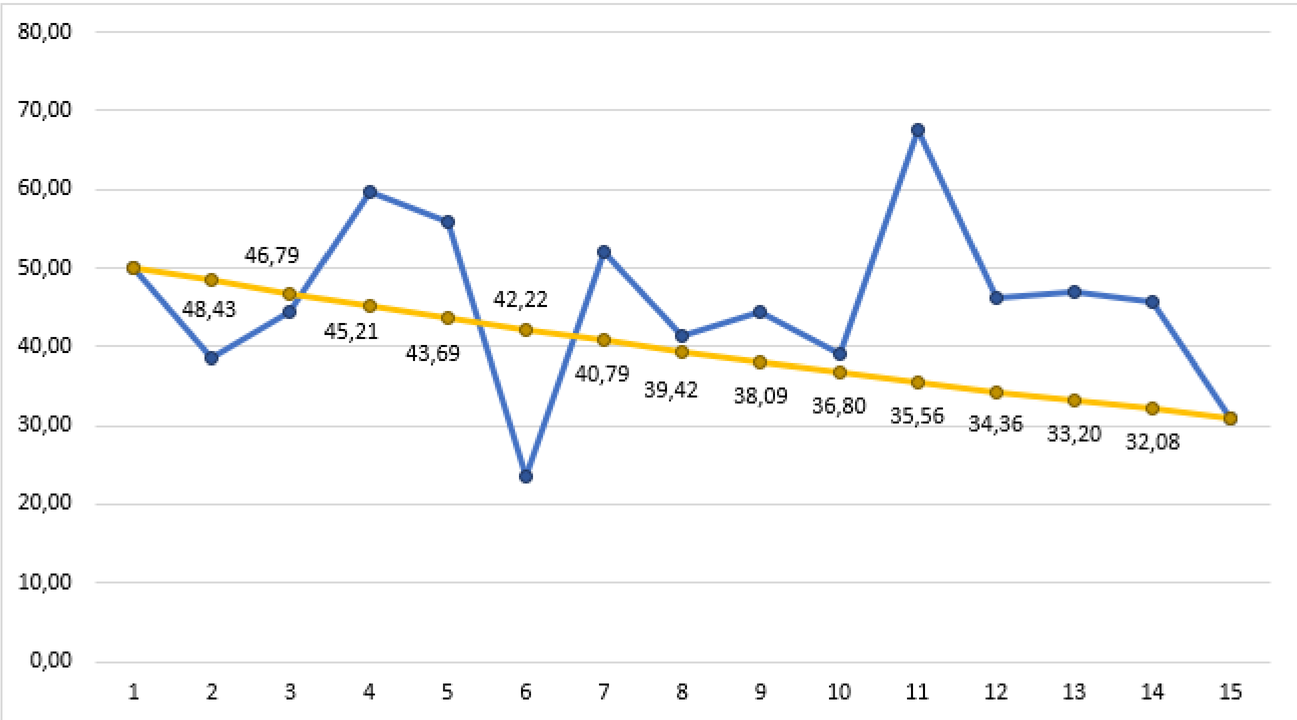

Рис. 2.4 График выравнивания методом среднего коэффициента динамики.

2.11. **Перспективная экстраполяция** на 5 временных периодов представлена в табличном (см. табл. 2.2) и графическом (см. рис. 2.5, 2.6 и 2.7) видах ниже:

Таблица 2.2

Экстраполяция ряда данных на 5 временных периодов тремя методами.

Номер периода	Методом скользящей средней (за 3 периода)	По среднему абсолютному приросту	По среднему коэффициенту динамики
1	36,30	29,63	29,95
2	39,45	28,27	28,94
3	36,63	26,90	27,97
4	36,52	25,54	27,02
5	37,50	24,17	26,11

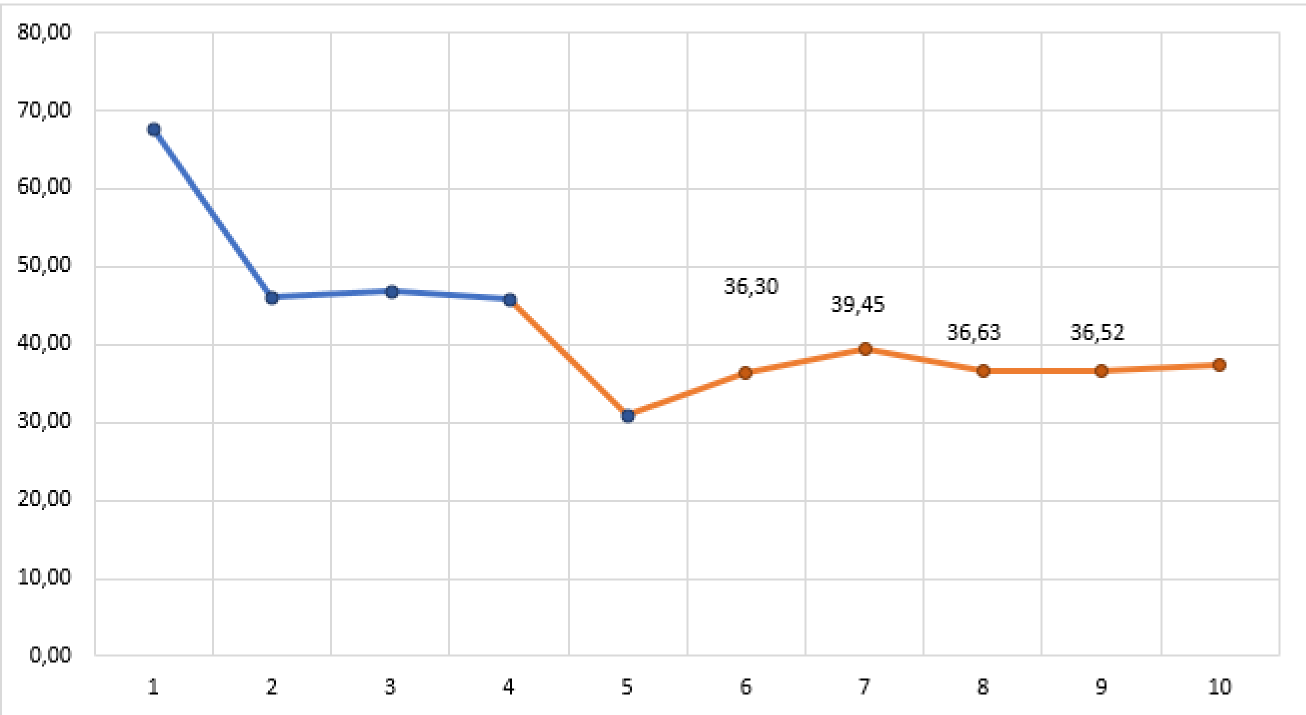

Рис. 2.5 Перспективная экстраполяция ряда данных на 5 временных периодов скользящей средней (на графике с 11 по 15).

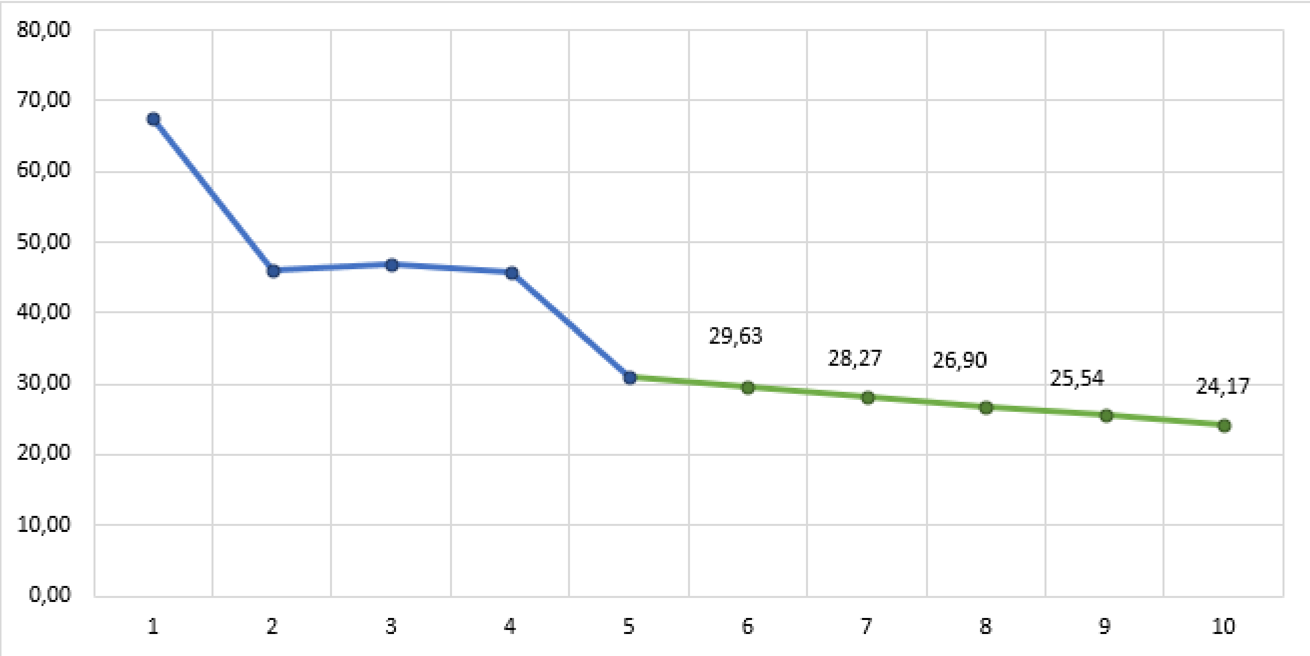

Рис. 2.6 Перспективная экстраполяция ряда данных на 5 временных периодов по среднему абсолютному приросту (на графике с 11 по 15).

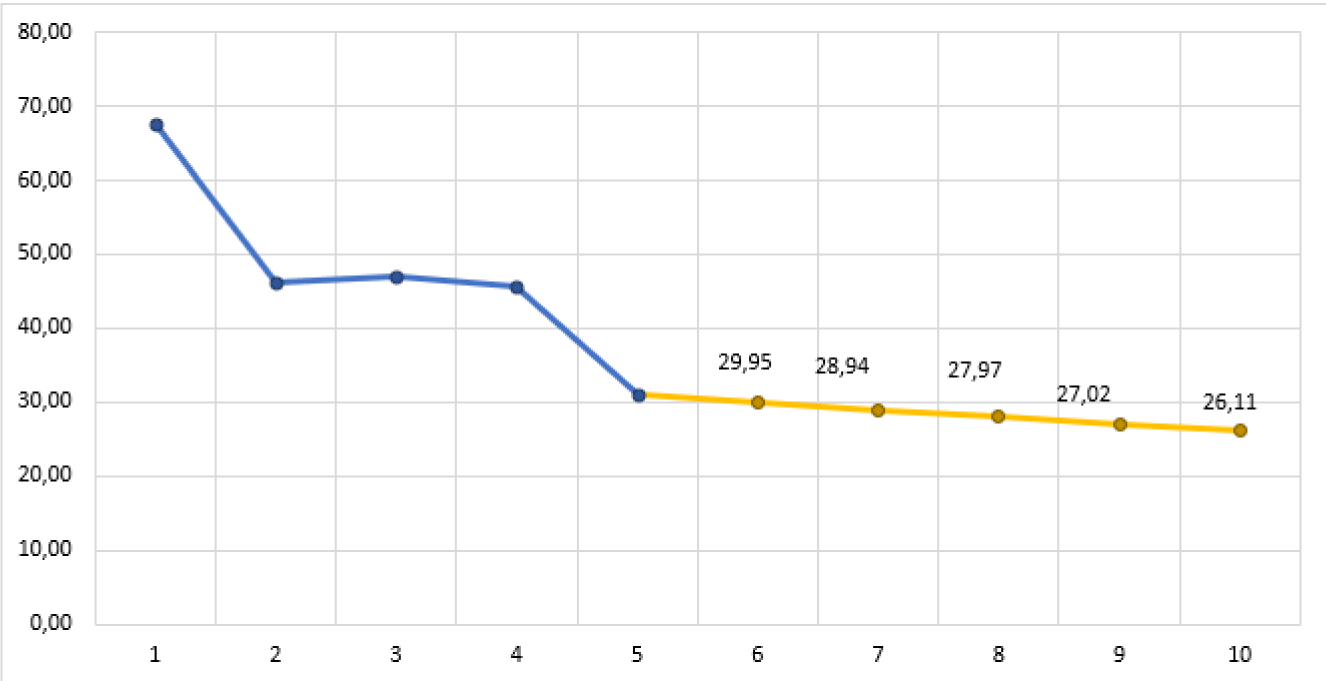

Рис. 2.7 Перспективная экстраполяция ряда данных на 5 временных периодов по среднему коэффициенту динамики (на графике с 11 по 15).

Раздел 3. Статистический вывод (оценка точности выборки).

3.1. **Статистический вывод** – это индуктивный («от частного к общему») процесс обобщения информации о выборочной совокупности для получения предположительных суждений о свойствах генеральной совокупности.

3.2. Для обозначенной цели используются **точечные** и **интервальные оценки** выборочных показателей. Под точечной оценкой понимается такая, которая имеет конкретное числовое значение, а интервальная оценка представляет собой числовое множество, внутри которого с определенной (в большинстве случаев заданной) **доверительной вероятностью** находится **истинное**, т.е. находящееся в генеральной совокупности, значение оцениваемого параметра.

3.1. Точечные и интервальные оценки средних величин.

3.3. Наипростейшим примером и точечной, и интервальной оценок выборочной средней величины будет являться **нахождение погрешности методом Корнфельда**, используемый в основном в грубых измерениях физических величин. За выборочное среднее в данном случае берётся средний диапазон эмпирических данных. Точечными оценками являются **абсолютная** и **относительная погрешности**. Абсолютная погрешность представляет собой половину разности наибольшего и наименьшего значений массива или полуразмах выборочной совокупности (3.1.). А относительная выражается отношением абсолютной к среднему диапазону (3.2.). Перед построением **двухстороннего доверительного интервала** следует найти доверительную вероятность (3.3). Верхняя граница данного интервала есть ничто иное, как сумма среднего диапазона и абсолютной погрешности, а нижняя – их разница, что, в свою очередь, равносильно интервалу, границы которого – наибольшее и наименьшее значения массива (3.4 – 3.6). Графическое представление данного интервала на прямой с отложением выборочного среднего диапазона также дано ниже (см. рис. 3.1).

$$\Delta_{К}\bar{x} = \frac{x_{max} - x_{min}}{2} = \frac{1}{2}R = \frac{44,09}{2} = 22,05 \quad (3.1)$$

$$\delta_{К}\bar{x} = \frac{\Delta_{К}\bar{x}}{\bar{x}} = \frac{22,05}{45,60} = 0,48 \quad (3.2)$$

$$P = 1 - \left(\frac{1}{2}\right)^{n-1} = 1 - 0,5^{14} = 0,9999 \quad (3.3)$$

$$\bar{X} = \bar{x} \pm \Delta_{К}\bar{x} \quad \text{или} \quad x_{min} \leq \bar{X} \leq x_{max} \quad (3.4)$$

$$\bar{X} = 45,60 \pm 22,05 \quad (3.5)$$

$$23,56 \leq \bar{X} \leq 67,65 \quad (3.6)$$

Рис. 3.1 Доверительный интервал для среднего диапазона (метод Корнфельда; $P = 0,9999$).

3.4. *Таким образом, следует предполагать, что истинное значение или средняя величина генеральной совокупности с вероятностью* 99,99% *будет лежать в замкнутом числовом промежутке* [23,56; 67,65] *при доле погрешности* 48% *от среднего диапазона выборки*. Разумеется, минусы данного метода оценки очевидны: слишком широкие границы интервала, причём известные с самого начала исследования, чувствительность к выбросам и невозможность предопределения (выбора) доверительной вероятности.

3.5. **Стандартная ошибка среднего арифметического ($SE_{\bar{x}}$)** – это параметр, величина которого приближенно характеризует отклонение выборочного среднего арифметического от неизвестного среднего генеральной совокупности. Исходя из формулы (3.7) видно, что чем больше разброс данных, тем больше и стандартная ошибка среднего, а чем больше размер выборки, тем меньше значение ошибки:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{10{,}54}{\sqrt{15}} = 2{,}72 \quad (3.7)$$

3.6. Подобный стандартной ошибке, но вычисляемый по коэффициенту вариации – это **показатель точности (C_s)**, формула и вычисление которого приведена ниже (3.8). Репрезентативность выборочного среднего по нему определяется следующим образом: чем меньше значение показателя, тем точнее выборочное среднее. Точность считается достаточной, если $\boldsymbol{C_s}$ не превышает 0,05.

$$C_s = \frac{V_s}{\sqrt{n}} = \frac{0{,}23}{\sqrt{15}} = 0{,}06 \quad (3.8)$$

3.7. В настоящем случае $C_s = 0{,}06 > 0{,}05$, следовательно, точность выборочного среднего арифметического, согласно данной оценке, следует считать недостаточной.

3.8. Для интервальной оценки выборочного среднего необходимо найти доверительную вероятность (3.9), которая зависит от заданного **уровня значимости (α)**, который в настоящем исследовании составляет 0,05.

$$P = 1 - \alpha = 1 - 0{,}05 = 0{,}95 \quad (3.9)$$

3.9. При расчёте границ двухстороннего доверительного интервала для среднего арифметического генеральной совокупности и не имея точного значения среднего квадратического отклонения генеральной совокупности, целесообразно задействовать **t-критерий** и сказать несколько слов о с **t-распределении**, иначе называемым **распределением Стьюдента**. Так, максимум его частоты совпадает с максимумом частоты нормального распределения, но высота и ширина кривой зависит от размера выборки (n): чем он меньше, тем более пологий ход имеет кривая распределения, а при большем его значении (например, $n \geq 50$), кривая t-распределения перейдет в нормальное и разница между интервалами, построенными используя оба закона распределения будет незначительна.

3.10. **Критическое значение t-критерия**, используемое в построении интервала, можно без труда найти в соответствующей справочной таблице (3.11), зная уровень значимости и **число степеней свободы** ($\boldsymbol{k}$) – т.е. разность между размером выборки и количеством констант (3.10):

$$k = n - 1 = 15 - 1 = 14 \quad (3.10)$$

$$t = t(k; \alpha) = t(14; 0{,}05) = 2{,}145 \quad (3.11)$$

3.11. Теперь, имея всё необходимое для нахождения границ доверительного интервала, следует воспользоваться формулой (3.12), и вычислить обе его границы (3.13), приведя результат в более приемлемый вид (3.14). Нахождение относительной ошибки (3.15) и графическое представление доверительного интервала на прямой с отложенным выборочным средним (см. рис. 3.2) также приведены ниже:

$$\bar{X} \in \bar{x} \pm t \cdot SE_{\bar{x}} \quad \text{или} \quad \bar{x} - t \cdot SE_{\bar{x}} < \bar{X} < \bar{x} + t \cdot SE_{\bar{x}} \quad (3.12)$$

$$\bar{x} \pm t \cdot SE_{\bar{x}} = 45{,}75 \pm 2{,}145 \cdot 2{,}72 = 45{,}75 \pm 5{,}83 \quad (3.13)$$

$$39{,}92 < \bar{X} < 51{,}59 \quad (3.14)$$

$$\delta\bar{x} = \frac{t \cdot SE_{\bar{x}}}{\bar{x}} = \frac{5{,}83}{45{,}75} = 0{,}13 \quad (3.15)$$

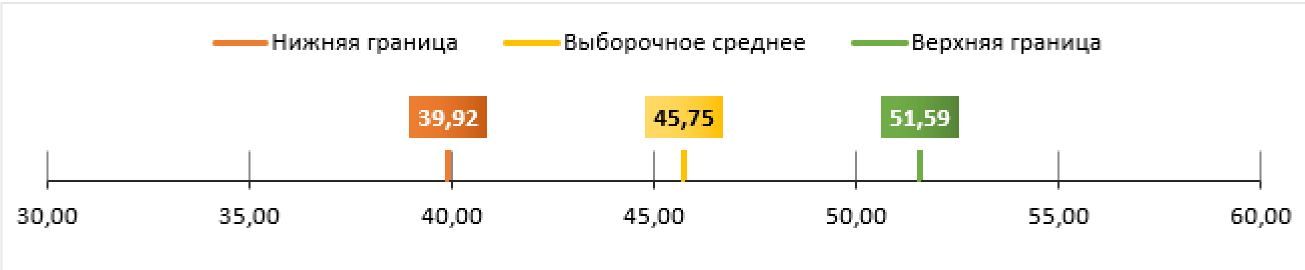

Рис. 3.2 Доверительный интервал для среднего арифметического (t-распределение; $P = 0{,}95$).

3.12. *Построенный интервал позволяет предположить, что среднее арифметическое генеральной совокупности с вероятностью* $P = 0{,}95$ *будет принадлежать открытому числовому множеству* $(39{,}92;\ 51{,}59)$ *при доле погрешности* 13% *от выборочного среднего арифметического.* Все результаты вычислений в подразделе сведены в таблицу ниже (см. табл. 3.1).

Таблица 3.1

Точечные и интервальные оценки средних величин.

<table>
<tr><td colspan="2">Параметр/мера/оценка/показатель</td><td>Обозначение</td><td>Значение</td></tr>
<tr><td rowspan="4">Средний диапазон; погрешность методом Корнфельда</td><td>Абсолютная погрешность</td><td>$\Delta_{К}\bar{x}$</td><td>22,05</td></tr>
<tr><td>Относительная погрешность</td><td>$\delta_{К}\bar{x}$</td><td>0,48</td></tr>
<tr><td>Доверительная вероятность</td><td>P</td><td>0,9999</td></tr>
<tr><td>Доверительный интервал</td><td>$\bar{X}$</td><td>[23,56; 67,65]</td></tr>
<tr><td colspan="2">Стандартная ошибка среднего арифметического</td><td>$SE_{\bar{x}}$</td><td>2,72</td></tr>
<tr><td colspan="2">Показатель точности</td><td>C_s</td><td>0,06</td></tr>
<tr><td rowspan="3">Среднее арифметическое; t-распределение $P = 0{,}95$</td><td>Абсолютная ошибка</td><td>$t \cdot SE_{\bar{x}}$</td><td>5,83</td></tr>
<tr><td>Относительная ошибка</td><td>$\delta\bar{x}$</td><td>0,13</td></tr>
<tr><td>Доверительный интервал</td><td>$\bar{X}$</td><td>(39,92; 51,59)</td></tr>
</table>

3.2. Точечные и интервальные оценки дисперсии, СКО, асимметрии и эксцесса.

3.13. Что касается оценки точности абсолютных выборочных показателей вариации, а именно дисперсии и среднего квадратического отклонения, к ним для начала следует применить методы устранения систематической ошибки, т.е. вычислить их точечные **несмещенные оценки** (3.16 и 3.17), иначе называемые **исправленной дисперсией** (S^2) и **исправленным СКО** (S):

$$S^2 = \frac{n}{k} \cdot \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{15}{14} \cdot \frac{1664,85}{15} = 118,92 \quad (3.16)$$

$$S = \sqrt{\frac{n}{k} \cdot \frac{\Sigma(x_i - \bar{x})^2}{n}} = \sqrt{\frac{15}{14} \cdot \frac{1664,85}{15}} = 10,90 \quad (3.17)$$

3.14. Получив результаты данных вычислений можно приступать к построению двухсторонних доверительных интервалов для дисперсии и СКО, но теперь уже с задействованием **распределения хи-квадрат** (χ^2) с найденным ранее числом степеней свободы. Формула и нахождение границ (3.18 – 3.20), также, как и графическое представление интервала на прямой (см. рис. 3.3) даны ниже:

$$\frac{k \cdot S^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{k \cdot S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \quad (3.18)$$

$$\frac{14 \cdot 118,92}{26,12} < \sigma^2 < \frac{14 \cdot 118,92}{5,63} \quad (3.19)$$

$$63,74 < \sigma^2 < 295,78 \quad (3.20)$$

Рис. 3.3 Доверительный интервал для дисперсии
(распределение хи-квадрат (χ^2); $P = 0{,}95$).

3.15. *Данный интервал с вероятностью* $P = 0{,}95$ *позволяет сделать предположение о том, что дисперсия генеральной совокупности принадлежит открытому числовому промежутку* $(63{,}74;\ 295{,}78)$.

3.16. Для интервальной оценки среднеквадратического отклонения необходимо лишь найти квадратный корень из границ построенного выше интервала, что и было сделано ниже (3.21 – 3.22; см. рис. 3.4):

$$\sqrt{63{,}74} < \sigma < \sqrt{295{,}78} \tag{3.21}$$

$$7{,}98 < \sigma < 17{,}20 \tag{3.22}$$

Рис. 3.4. Доверительный интервал для среднего квадратического отклонения
(распределение хи-квадрат (χ^2); $P = 0{,}95$).

3.17. *Построенный интервал позволяет с* 95%-*ной вероятностью предположить, что среднее квадратическое отклонение генеральной совокупности будет лежать в открытом числовом множестве* $(7{,}98;\ 17{,}20)$.

3.18. Для оценки значимости коэффициента асимметрии следует вычислить отношение модуля его **несмещенной оценки** ($\widehat{\boldsymbol{As}}$) к его **средней квадратической ошибке** ($\boldsymbol{SE_{As}}$). Если полученный результат больше трёх, предполагается, что распределение в генеральной совокупности несимметрично, а в случае, когда оно меньше трёх – есть причина полагать, что асимметрия несущественна, т.е. вызвана влиянием выбросов в выборке. Формулы и вычисления даны ниже (3.23 – 3.25):

$$\widehat{As} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s}\right)^3 = \frac{15}{182} \cdot (-0{,}28) = -0{,}02 \tag{3.23}$$

$$SE_{As} = \sqrt{\frac{6 \cdot (n-1)}{(n+1) \cdot (n+3)}} = \sqrt{\frac{6 \cdot 14}{16 \cdot 18}} = 0{,}52 \tag{3.24}$$

$$\frac{|\widehat{As}|}{SE_{As}} = \frac{0{,}02}{0{,}52} = 0{,}04 \tag{3.25}$$

3.19. *Рассчитанное отношение* $(0{,}04 < 3)$ *позволяет предположить, что асимметрия в распределении генеральной совокупности, так же, как и в эмпирическом распределении выборки, является незначительной.*

3.20. Аналогичным образом следует поступить с коэффициентом эксцесса: найти отношение модуля его **несмещенной оценки** ($\widehat{\boldsymbol{Ex}}$) к его **средней квадратической ошибке** ($\boldsymbol{SE_{Ex}}$). Формулы и вычисления приведены ниже (3.26 – 3.28):

$$\widehat{Ex} = \left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s}\right)^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)} = 0{,}70 \tag{3.26}$$

$$SE_{Ex} = \sqrt{\frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n-1) \cdot (n+3) \cdot (n+5)}} = \sqrt{\frac{24 \cdot 15 \cdot 13 \cdot 12}{14 \cdot 18 \cdot 20}} = 0{,}78 \tag{3.27}$$

$$\frac{|\widehat{Ex}|}{SE_{Ex}} = \frac{0{,}70}{0{,}78} = 0{,}89 \qquad (3.28)$$

3.21. *Учитывая, что полученное отношение меньше трёх* $(0{,}89 < 3)$, *отклонение относительного веса хвостов от нормального распределения следует предполагать недостаточно существенным*. В заключение, как и ранее, следует свести все результаты подраздела в таблицу (см. табл. 3.2):

Таблица 3.2

Точечные и интервальные оценки дисперсии, СКО, асимметрии и эксцесса.

Параметр/мера/оценка/показатель	Обозначение	Значение
Исправленная дисперсия	S^2	118,92
Исправленное СКО	S	10,90
Доверительный интервал для дисперсии генеральной совокупности (χ^2-распределение; $P = 0{,}95$)	σ^2	(63,74; 295,78)
Доверительный интервал для СКО генеральной совокупности (χ^2-распределение; $P = 0{,}95$)	σ	(7,98; 17,20)
Несмещённая оценка асимметрии	$\widehat{As}$	–0,02
Средняя квадратическая ошибка асимметрии	SE_{As}	0,52
Отношение модуля несмещенной оценки коэффициента асимметрии к его средней квадратической ошибке	$\frac{\lvert\widehat{As}\rvert}{SE_{As}}$	0,04
Несмещённая оценка эксцесса	$\widehat{Ex}$	0,70
Средняя квадратическая ошибка эксцесса	SE_{Ex}	0,78
Отношение модуля несмещенной оценки коэффициента эксцесса к его средней квадратической ошибке	$\frac{\lvert\widehat{Ex}\rvert}{SE_{Ex}}$	0,89