Специальная математика и основы статистики

Показатели вариации. Дисперсионный метод анализа

Вопрос 1. Показатели вариации и способы их расчета

Одна из причин, по которой существует необходимость в проведении статистического анализа и постоянном сборе и обновлении информации о социальных и экономических явлениях состоит в том, что данные изменчивы.

Вариацию, или ***изменчивость*** можно определить как степень различия между отдельными значениями признака или показателя.

Ситуация, в которой присутствует изменчивость, всегда связана с долей риска и неопределенностью в будущем. Систематическое воздействие различных факторов и условий вызывает изменение отдельных вариантов признаков или показателя в целом. В большинстве случаев обнаружить такое воздействие и тем самым снизить риск возможно, изучая колебания или индивидуальные различия значений, а не обобщающие величины.

Для количественной характеристики колеблемости признака необходимо оценить расстояние между каждым индивидуальным значением признака и общей для них средней величиной, которое определяется как разность между их значениями. Эту разность в статистике называют **отклонением от средней величины** (рис.1).

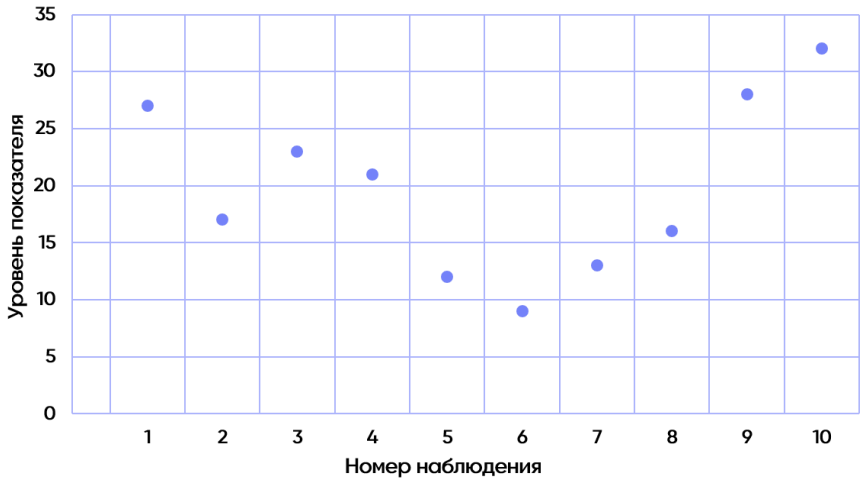

Рис.1 – График отклонения наблюдений от среднего значения

Для количественного измерения степени близости значений отдельных единиц к средней используется система абсолютных и относительных показателей.

А) абсолютные показатели вариации

1) Размах вариации - это разность между наибольшим (X_{max}) и наименьшим (X_{min}) значениями вариантов.

$$R = X_{max} - X_{min}$$

Размах измеряется в тех же абсолютных единицах, что и значения признака. Размах устанавливает ширину интервала, занимаемого значениями данных, но не отражает отклонений всех вариантов признака в ряду. Поскольку размах зависит только от крайних значений, его величина в большей мере подвержена воздействию случайности, так как в совокупности аномально большие или маленькие значения данных могут быть получены под влиянием случайных причин.

Размах вариации хорошо применять в случаях, когда минимальный или максимальный вариант признака имеет особое значение. Например, при определении пределов, в которых могут колебаться размеры отдельных параметров деталей, оценки пределов точности измерения приборов.

Чтобы дать обобщающую характеристику распределению отклонений, исчисляют среднее линейное отклонение d, которое учитывает различие всех единиц изучаемой совокупности.

2) Среднее линейное отклонение показывает стандартное отличие значения каждого варианта от общей средней величины и определяется как средняя арифметическая из отклонений индивидуальных значений от средней, без учета их знака (по модулю): $\overline{d} = \frac{\sum |x-\overline{x}|}{n} = \frac{|x_1-\overline{x}|+|x_2-\overline{x}|+\cdots+|x_n-\overline{x}|}{n}$.

Среднее линейное отклонение измеряется в тех же абсолютных единицах, что и значения признака.

Порядок расчета среднего линейного отклонения следующий:

1. по значениям признака находят среднюю арифметическую;
2. определяют отклонения каждого значения x; от средней $|x_i - \overline{x}|$;
3. рассчитывается сумма абсолютных величин отклонений: $\sum |x_i - \overline{x}|$;
4. сумма абсолютных величин отклонений делится па число значений.

Если данные наблюдения представлены в виде дискретного или интервального ряда распределения с частотами, среднее линейное отклонение исчисляется по формуле средней арифметической взвешенной: $\overline{d} = \frac{\sum |x_i-\overline{x}|f_i}{\sum f_i} = \frac{|x_1-\overline{x}|f_1+|x_2-\overline{x}|f_2+\cdots+|x_n-\overline{x}|f_n}{f_1+f_2+\cdots+f_n}$

Порядок расчета взвешенного среднего линейного отклонения:

1. вычисляется средняя арифметическая взвешенная;

2. определяются абсолютные отклонения вариантов от средней $|x_i - \overline{x}|$;
3. полученные отклонения умножаются на частоты $|x_i - \overline{x}|f_i$;
4. находится сумма взвешенных отклонений без учета знака:$\sum |x_i - \overline{x}| f_i$;
5. сумма взвешенных отклонений делится на сумму частот: $\frac{\sum |x_i - \overline{x}| f_i}{\sum f_i}$.

Для нахождения $\overline{d}$ по интервальному ряду распределения вначале находят середины каждого интервала (как среднюю арифметическую из значений границ) и используют их для расчета отклонений от общей средней аналогичным образом.

3) Дисперсия - это средняя арифметическая из квадратов отклонений каждого значения признака от общей средней. Дисперсия обычно называется средним квадратом отклонений и обозначается σ^2. В зависимости от исходных данных дисперсия может вычисляться как средняя арифметическая простая или взвешенная.

Если каждый вариант признака повторяется только один раз, используют простую (невзвешенную) формулу: $\sigma^2 = \frac{\sum (x_i - \overline{x})^2}{n}$

Если варианты признака повторяются неодинаковое количество раз, используют взвешенную формулу: $\sigma^2 = \frac{\sum (x_i - \overline{x})^2 f_i}{\sum f_i}$

Порядок расчета дисперсии по взвешенной формуле следующий.

1. определяют среднюю арифметическую взвешенную;
2. рассчитывают отклонения вариантов от средней:$(x_i - \overline{x})$;
3. возводят в квадрат отклонение каждого варианта от средней: $(x_i - \overline{x})^2$;
4. умножают квадраты отклонений на веса (частоты): $(x_i - \overline{x})^2 f_i$;
5. суммируют полученные произведения:$\sum (x_i - \overline{x})^2 f_i$;
6. Полученную сумму делят на сумму весов: $\frac{\sum (x_i - \overline{x})^2 f_i}{\sum f_i}$.

Если исходные данные представлены в виде интервального ряда распределения, то сначала надо определить средины каждого интервала, и далее рассчитывать показатели вариации аналогичным образом.

4) Среднее квадратическое отклонение (в *литературе также часто используется термин* ***«стандартное отклонение»)*** представляет собой квадратный корень из дисперсии и обозначается σ.

Формулы для его расчета следующие:

- среднее квадратическое отклонение невзвешенное: $\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$

- среднее квадратическое отклонение взвешенное: $\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2 f_i}{\sum f_i}}$

Математические преобразования формул приводят к упрощенному виду формулы, которая часто оказывается более удобной на практике, особенно при расчете σ по__**несгруппированным** данным: $\sigma = \sqrt{\overline{x^2} - (\overline{x})^2}$, где $\overline{x^2} = \frac{\sum x_i^2 f_i}{\sum f_i}$

Среднее квадратическое отклонение является обобщающей характеристикой абсолютных размеров вариации признака в совокупности. Выражается оно в тех же единицах измерения, что и признак (в метрах, тоннах, процентах, гектарах и т.д.) и показывает, насколько в среднем отличается значение каждого варианта признака от среднего значения для данной совокупности.

Среднее квадратическое отклонение всегда больше среднего линейного отклонения. Между ними имеется соотношение: $\boldsymbol{\sigma = \overline{d} \times 1,25}$.

Б) относительные показатели вариации

1) Линейный коэффициент вариации характеризует долю среднего линейного отклонения от общего размера средней величины и рассчитывается по формуле: $K_d = \frac{\overline{d}}{\overline{x}} * 100\%$

2) Коэффициент вариации рассчитывается как отношение среднего квадратического отклонения к средней величине: $v = \frac{\sigma}{\overline{x}} * 100\%$

Учитывая, что среднеквадратическое отклонение дает обобщающую характеристику колеблемости всех вариантов совокупности, коэффициент вариации является наиболее распространенным показателем, используемым для оценки типичности средних величин. **Если V больше 40%,** то это говорит о большой колеблемости признака в изучаемой совокупности.

Совокупность считается однородной, если коэффициент вариации **не превышает 33%.**

Следует отметить, что коэффициент вариации может быть более 100%, что бывает при наличии значений сильно отличающихся от средней величины, например, отрицательных значений или аномально больших или малых значений отдельных вариантов. Такой результат означает, что в исследуемой совокупности сильна вариация признаков по отношению к средней величине и для более качественного ее исследования необходимо разбить ее на более однородные части или с помощью специальных методов исключить влияние на среднюю аномальных вариантов.

Пример: Вычислить показатели вариации для следующих данных:

Таблица 1.

Затраты на производство единицы продукции по корпорации

Отделение корпорации	Затраты на единицы продукции, руб. (x_i)	Средний объем производства продукции, единиц продукции в месяц
1 отделение	10	2000
2 отделение	12	2000
3 отделение	14	2000
Итого	-	**6000**

Так как каждое отделение производит одинаковое количество продукции, для расчета обобщающих показателей - средней, дисперсии, среднего квадратического отклонения - можно воспользоваться невзвешенной формулой. Промежуточные расчеты показателей представлены в таблице 2.

Таблица 2

Затраты на производство единицы продукции по корпорации

Отделение корпорации	(x_i)	$x_i-\bar{x}$	$(x_i-\bar{x})^2$
1 отделение	10	-2	4
2 отделение	12	0	0
3 отделение	14	+2	4
Итого	-	**0**	**8**

$$\bar{x}=36:3 = 12 \text{ руб.}$$
$$\sigma^2 = 8:3 = 2{,}6667$$
$$\sigma = \sqrt{2{,}667} = 1{,}632 \text{ руб.}$$
$$R = 14 - 10 = 4 \text{ руб.}$$
$$V = 1{,}632:12 * 100 = 13{,}6\%$$

Средние затраты на единицу продукции корпорации составляют 12 руб. Риск превышения затрат на единицу продукции составляет 1,632 руб. - величину среднего квадратического отклонения. Общие производственные затраты в месяц составляют в среднем 12 * 6000 = 72 000 руб.

■ ■ ■

Вопрос 2. Математические свойства показателей вариации

Свойство 1. Уменьшение или увеличение всех значений признака на одинаковую величину не меняет величины дисперсии, среднего квадратического отклонения и размаха вариации: $\sigma^2_{(x\pm A)} = \sigma^2_x;\ \sigma_{(x\pm A)} = \sigma_x;\ R_{(x\pm A)} = R_x$

Свойство 2. Увеличение всех значений признака в к раз (к - любое число) увеличивает дисперсию в k^2 раз, среднее квадратическое отклонение и размах - в k раз. (Если к<0, то коэффициент берется по

модулю).Коэффициент вариации при этом не меняется.: $\sigma^2_{(x\times k)} = \sigma^2_x \times k^2$; $\sigma_{(x\times k)} = \sigma_x \times k$; $R_{(x\times k)} = R_x \times k$

Свойство 3. Дисперсия отклонений значений признака от произвольного числа А (X - A)^2 увеличивает дисперсию отклонений от средней $(X_i - \overline{X})$ на число, равное возведенной в квадрат разнице между средней и этим числом А, т.е. на$(\overline{X} - A)^2$: $\sigma^2{}_{(A)} = \sigma^2{}_x + (\overline{X} - A)^2$ или $\sigma^2{}_x = \sigma^2{}_{(A)} - (\overline{X} - A)^2$

Вопрос 3. Однофакторный дисперсионный анализ

Изучая вариацию интересующего нас признака в пределах совокупности, разделенной на группы, для более качественного анализа необходимо проследить количественные изменения признака внутри каждой группы, а также между группами. Такой анализ возможен с помощью вычисления различных видов дисперсии. Выделяют три вида дисперсии - общую, внутригрупповую и межгрупповую.

Общая дисперсия (σ^2) характеризует вариацию признака по всей совокупности как результат влияния всех факторов, определяющих индивидуальные различия единиц совокупности: $\sigma^2 = \frac{\sum(x_i - x)^2 f_i}{\sum f_i}$.

Внутригрупповая дисперсия (σ_i^2) отражает случайную вариацию, т.е. ту часть вариации признака, которая обусловлена действием всех прочих неучтенных факторов, кроме фактора, по которому осуществлялась группировка. Внутригрупповая дисперсия рассчитывается отдельно по каждой выделенной группе по формуле: $\sigma_i^2 = \frac{\sum(x_i - \bar{x}_i)^2}{n_i}$, где x_i- значения признака у единиц, входящих i- ю группу; $\overline{x_i}$ - среднее значение признака в i-ой группе; n_i- число единиц в i-ой группе.

Для всех групп в целом вычисляется **средняя из внутригрупповых** дисперсий $\overline{\sigma_i{}^2}$ по формуле: $\overline{\sigma_i{}^2} = \frac{\sum \sigma_i^2 n_i}{\sum n_i}$

Межгрупповая дисперсия ($\boldsymbol{\sigma_x{}^2}$) характеризует вариацию, обусловленную влиянием на значения исследуемого признака признака-фактора, положенного в основание группировки: $\sigma_x{}^2 = \frac{\sum(\bar{x}i - \bar{x})^2 n_i}{\sum n_i}$, где $\overline{x}$ - общая средняя величина признака; $\overline{x_i}$ - среднее значение признака в i-ой группе; n_i – число единиц в i-ой группе, при этом $\sum n_i = \sum f_i$

Взаимосвязь между тремя видами дисперсий получила название **правило сложения дисперсий.** Согласно ему, общая дисперсия равна сумме средней из внутригрупповой и межгрупповой дисперсий: $\sigma^2 = \overline{\sigma_i{}^2} + \sigma_x{}^2$

Таким образом, зная два вида дисперсий, всегда можно определить третий. Кроме того, при качественно проведенной группировке на основании соотношения между данными видами дисперсий можно судить о степени влияния группировочного признака на изменение значений других, зависящих от него признаков. Такие соотношения называются эмпирическими коэффициентами.

Эмпирический коэффициент детерминации (η^2) *(от лат. determinatio — «ограничение, определение»)* характеризует долю межгрупповой дисперсии в общей дисперсии и рассчитывается по формуле: $\eta^2 = \frac{{\delta_x}^2}{\sigma^2}$

Он показывает долю общей вариации изучаемого признака, которую вызывает (определяет) вариация группировочного признака.

Если извлечь квадратный корень из коэффициента детерминации, получим **эмпирическое корреляционное отношение η** *(от позднелат. correlatio - «соотношение, взаимозависимость, взаимное соответствие»)* - коэффициент, при помощи которого можно оценить тесноту связи между групнировочным (факторным) и результативным признаками. Он рассчитывается по формуле: $\eta_э = \sqrt{\frac{{\delta_x}^2}{\sigma^2}}$

Данный коэффициент может принимать значения от 0 до 1. Чем ближе к 1 будет его величина, тем сильнее взаимосвязь между рассматриваемыми признаками.

Пример: Исследуем зависимость объема выполненных работ от формы собственности проектно-изыскательских организаций (табл. 3.).

Таблица 3

Выполнение работ проектно-изыскательскими организациями разной формы собственности

№ группы	Форма собственности	Количество организаций	Объем выполненных работ (млн. руб.)	Итого
1	Государственная	4	10,30,20,40	100
2	Негосударственная	6	30,45,60,20,65,50	270
Итого	-	10	-	-

1. Рассчитаем общую среднюю: $\overline{X}$ =370:10 = 37 млн. руб.

2. Рассчитаем средний объем выполненных работ по каждой группе организаций.

Государственные:

$\overline{X_1}$ = (10 + 30 + 20 + 40) : 4 = 100 : 4 = 25 млн.руб.

Негосударственные:

$\overline{X_2}$ = (30 + 45 + 60 + 20 + 65 + 50) : 6 = 270: 6 = 45 млн. руб.

3. Рассчитаем внутригрупповые дисперсии объема выполненных работ.

По 1 группе - государственным организациям - ${\sigma_1}^2$

$$\sigma_1{}^2 = \frac{\sum(x_i - \overline{x_1})^2}{n_1}$$
$$= \frac{(10-25)^2 + (30-25)^2 + (20-25)^2 + (40-25)^2}{4}$$
$$= \frac{(-15)^2 + (5)^2 + (-5)^2 + (15)^2}{4}$$
$$= \frac{225 + 25 + 25 + 225}{4} = 500 : 4 = 125$$

По 2 группе - негосударственным организациям - $\sigma_2{}^2$

$$\sigma_2{}^2 = \frac{\sum(x_i - \overline{x_2})^2}{n_2}$$
$$= \frac{(30-45)^2 + (45-45)^2 + (60-45)^2 + (20-45)^2\,(65-45)^2\,(50-45)^2}{6}$$
$$= \frac{(-5)^2 + (0)^2 + (15)^2\,(-25)^2\,(20)^2\,(5)^2}{6}$$
$$= \frac{25 + 0 + 225 + 625 + 400 + 25}{6} = 1500 : 4 = 250$$

4. Рассчитаем среднюю из внутригрупповых дисперсий:

$$\overline{\sigma_i{}^2} = \frac{\sum \sigma_i^2 n_i}{\sum n_i} = \frac{\sigma_1{}^2 n_1 + \sigma_2{}^2 n_2}{n_1 + n_2} = \frac{125 \times 4 + 250 \times 6}{4 + 6} = \frac{500 + 1500}{10}$$
$$= 200$$

5. Рассчитаем межгрупповую дисперсию:

$$\delta_x{}^2 = \frac{\sum(\overline{x}i - \overline{x})^2 n_i}{\sum n_i} = \frac{(\overline{x_1} - \overline{x})^2 n_1 + (\overline{x_2} - \overline{x})^2 n_2}{n_1 + n_2}$$
$$= \frac{(25-37)^2 \times 4 + (45-37)^2 \times 6}{4+6}$$
$$= \frac{(-12)^2 \times 4 + (8)^2 \times 6}{10} = \frac{144 \times 4 + 64 \times 6}{10}$$
$$= \frac{576 + 384}{10} = 960 : 10 = 96$$

6. Рассчитаем общую дисперсию по правилу сложения дисперсий:

$$\sigma^2 = \overline{\sigma_i{}^2} + \delta_x{}^2 = 200 + 96 = 296$$

7. Для проверки правильности расчетов определим общую дисперсию по формуле: $\sigma^2 = \frac{\sum(x_i - \overline{x})^2}{n}$. Вспомогательные расчеты представим в табл. 4.

Таблица 4

Вспомогательная таблица для расчета общей дисперсии

№ группы организации	Объем работ, млн. руб. (x_i)	$(x_i-\overline{x})$	$(x_i-\overline{x})^2$
1	10	-27	729
1	30	-7	49
1	20	-17	289
1	40	3	9
2	30	-7	49
2	45	8	64
2	60	23	529
2	20	-17	289
2	65	28	784
2	50	13	169
Итого:	370	0	2960

$$\sigma^2 = \frac{\sum(x_i - \overline{x})^2}{n} = 2960 : 10 = 296$$

$$\sigma^2 = {\sigma_i}^2 + {\delta_x}^2 = 200 + 96 = 296$$

Вычисление дисперсии по обычной формуле и по правилу сложения дисперсий дает одинаковый результат.

8. Проверим, в какой мере группировочный признак - форма собственности организации - оказывает влияние на вариацию объема выполненных работ, и оценим степень их взаимосвязанности.

Рассчитаем эмпирический коэффицент детерминации:

$$\eta^2 = \frac{{\delta_x}^2}{\sigma^2} = \frac{96}{296} = 0{,}3243\ (32{,}43\%)$$

Это значит, что объем выполненных работ на 32,43% зависит от формы собственности проектно-изыскательной организации и на 67,57% от ее внутриорганизационных возможностей.

Рассчитаем эмпирическое корреляционное отношение:

$$n_э = \sqrt{\frac{{\delta_x}^2}{\sigma^2}} = \sqrt{\frac{96}{296}} = \sqrt{0{,}3243} = 0{,}57$$

Степень близости отношения к единице говорит о гом, что форма собственности оказывает среднее по силе, но существенное влияние на объем работ, выполняемый проектно-исследовательскими организациями.

■ ■ ■