Характеристики вариационного ряда

План:

- 1. Абсолютные показатели вариации.
- 2. Относительные показатели вариации.
- 3. Понятие, виды и свойства дисперсий.
- 4. Методы расчета сводных характеристик выборки.
- 5. Эмпирические моменты.
- 6. Эмпирические и выравнивающие теоретические частоты.
- 7. Критерий согласия Пирсона.

ПОКАЗАТЕЛИ ВАРИАЦИИ ПРИЗНАКОВ

Различия индивидуальных значений признака внутри изучаемой совокупности называется **вариацией признака**.

Выделяют три **формы вариационного ряда**:

- ранжированный ряд;
- дискретный ряд;
- интервальный ряд.

АБСОЛЮТНЫЕ ПОКАЗАТЕЛИ ВАРИАЦИИ

Размах вариации – разность между наибольшим и наименьшим значениями варьирующего признака, показывает, в каких пределах колеблется размер признака, образующего ряд распределения.

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

АБСОЛЮТНЫЕ ПОКАЗАТЕЛИ ВАРИАЦИИ

1. Средняя - характеризует среднее арифметическое вариант вариационного ряда.

а) **простая**

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

б) взвешенная

$$\bar{X} = \frac{\sum x_i \cdot f_i}{\sum f_i}$$

АБСОЛЮТНЫЕ ПОКАЗАТЕЛИ ВАРИАЦИИ

Дисперсия – средний квадрат отклонений индивидуальных значений признака от их средней величины и в зависимости от исходных данных вычисляется по формулам:

а) простая

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n};$$

б) взвешенная

$$\sigma^2 = \frac{\sum_{i=1}^{k}(X_i - \bar{X})^2 \cdot f_i}{\sum_{i=1}^{k} f_i}.$$

АБСОЛЮТНЫЕ ПОКАЗАТЕЛИ ВАРИАЦИИ

Среднее квадратическое отклонение (нормированное или стандартизированное отклонение) рассчитывается как корень квадратный из дисперсии. Оно может быть

а) простое

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}};$$

б) взвешенное

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(X_i - \bar{X})^2 \cdot f_i}{\sum_{i=1}^{k} f_i}}.$$

АБСОЛЮТНЫЕ ПОКАЗАТЕЛИ ВАРИАЦИИ

Среднее линейное отклонение – средняя арифметическая из абсолютных значений отклонений вариант признака от их средней. Эта величина вычисляется как средняя арифметическая из абсолютных значений отклонений вариант.

а) простое

$$\bar{d} = \frac{\sum_{i=1}^{n} \left| X_i - \bar{X} \right|}{n};$$

б) взвешенное

$$\bar{d} = \frac{\sum_{i=1}^{k} \left| X_i - \bar{X} \right| \cdot f_i}{\sum_{i=1}^{k} f_i}.$$

ОТНОСИТЕЛЬНЫЕ ПОКАЗАТЕЛИ ВАРИАЦИИ

Коэффициент осцилляции – процентное отношение размаха вариации к средней величине признака.

$$V_R = \frac{R}{\overline{X}} \cdot 100\%$$

Линейный коэффициент вариации – процентное отношение среднего линейного отклонения к средней величине признака.

$$V_{\bar{d}} = \frac{\bar{d}}{\overline{X}} \cdot 100\%$$

Коэффициент вариации – процентное отношение среднего квадратического отклонения к средней величине признака.

$$V_\sigma = \frac{\sigma}{\overline{X}} \cdot 100\%$$

ПОНЯТИЕ, ВИДЫ И СВОЙСТВА ДИСПЕРСИЙ

Общая дисперсия измеряет вариацию признака во всей совокупности под влиянием всех факторов, обусловивших эту вариацию.

$$\sigma_0^2 = \frac{\sum_{i=1}^{k}(X_i - \bar{X}_0)^2 \cdot n_i}{\sum_{i=1}^{k} n_i}$$

Межгрупповая дисперсия характеризует систематическую вариацию, т.е. различия в величине изучаемого признака, возникающие под действием признака-фактора, положенного в основу группировки.

$$\delta_x^2 = \frac{\sum_{j=1}^{k}(\bar{X}_j - \bar{X}_0)^2 \cdot n_j}{\sum_{j=1}^{k} n_j}$$

ПОНЯТИЕ, ВИДЫ И СВОЙСТВА ДИСПЕРСИЙ

Внутригрупповая дисперсия отражает случайную вариацию, т.е. часть вариации, происходящей под влиянием неучтенных факторов и не зависящую от признака-фактора.

$$\sigma_j^2 = \frac{\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2}{n_j}$$

По совокупности в целом вариация значений признака под влиянием прочих факторов характеризуется **средней из внутригрупповых дисперсий**.

$$\bar{\sigma}^2 = \frac{\sum_{j=1}^{k} \sigma_j^2 \cdot n_j}{\sum_{j=1}^{k} n_j}$$

Общая дисперсия равна сумме средней из внутригрупповых и межгрупповой дисперсий (**правило сложения дисперсий**):

$$\sigma_0{}^2 = \bar{\sigma}^2 + \delta^2$$

Методы расчета сводных характеристик выборки

Предположим, что варианты выборки
расположены в возрастающем порядке,
т.е. в виде вариационного ряда.

Равноотстоящими называют варианты, которые образуют арифметическую прогрессию с разностью *h*.

Методы расчета сводных характеристик выборки

Условными называют варианты, определяемые равенством:

$$u_i = (x_i - C)/h,$$

где *C*—ложный нуль (новое начало отсчета);

h—шаг, т.е. разность между любыми двумя соседними первоначальными вариантами (новая единица масштаба).

Методы расчета сводных характеристик выборки

Упрощенные методы расчета сводных характеристик выборки основаны на замене первоначальных вариант условными.

Условные варианты.
Замечания

В качестве ложного нуля можно принять любую варианту. Максимальная простота вычислений достигается, если выбрать в качестве ложного нуля варианту, которая расположена примерно в середине вариационного ряда (часто такая варианта имеет наибольшую частоту).

Варианте, которая принята в качестве ложного нуля, соответствует условная варианта, равная нулю.

Условные варианты. Пример.

Найти условные варианты статистического распределения:

варианты . . .	23,6	28,6	33,6	38,6	43,6
частоты . . .	5	20	50	15	10

Условные варианты. Решение

Выберем в качестве ложного нуля варианту 33,6 (эта варианта расположена в середине вариационного ряда).

Найдем шаг:

$$h = 28{,}6 - 23{,}6 = 5.$$

Условные варианты. Решение

Найдем условную варианту:

$$u_1 = (x_1 - C)/h = (23,6 - 33,6)/5 = -2.$$

Аналогично получим: $u_2 = -1$, $u_3 = 0$, $u_4 = 1$, $u_5 = 2$. Мы видим, что условные варианты – небольшие целые числа. Разумеется, оперировать с ними проще, чем с первоначальными вариантами.

Эмпирические моменты

Для вычисления сводных характеристик выборки удобно пользоваться эмпирическими моментами.

Эмпирические моменты вычисляют по данным наблюдений.

Обычные, начальные и центральные эмпирические моменты

Обычным эмпирическим моментом порядка k называют величину:

$$M'_k = \left(\sum n_i (x_i - C)^k\right)/n,$$

где x_i — наблюдаемая варианта, n_i — частота варианты, $n = \sum n_i$ — объем выборки, C — произвольное постоянное число (ложный нуль).

Обычные, начальные и центральные эмпирические моменты

Начальным эмпирическим моментом порядка k называют обычный момент порядка *k* при С = 0

$$M_k = \left(\sum n_i x_i^k\right)/n.$$

Обычные, начальные и центральные эмпирические моменты

В частности,

$$M_1 = \left(\sum n_i x_i\right)/n = \bar{x}_\text{в},$$

т.е. начальный эмпирический момент первого порядка равен выборочной средней.

Обычные, начальные и центральные эмпирические моменты

Центральным эмпирическим моментом порядка k называют обычный момент порядка *k* при $C = \bar{x}_B$

$$m_k = \left(\sum n_i (x_i - \bar{x}_{\text{в}})^k\right)/n.$$

Обычные, начальные и центральные эмпирические моменты

В частности,

$$m_2 = \left(\sum n_i (x_i - \bar{x}_в)^2\right)/n = D_в,$$

т.е. центральный эмпирический момент второго порядка равен выборочной дисперсии.

Асимметрия и эксцесс

Для оценки отклонения эмпирического распределения от нормального используются асимметрия и эксцесс эмпирического распределения:

- Асимметрия

$$a_s = \frac{m_3}{\sigma_B^3}$$

- Эксцесс

$$e_k = \frac{m_4}{\sigma_B^4} - 3$$

где m_3 и m_4 - центральные эмпирические моменты 3 и 4-го порядков.

Эмпирические и выравнивающие теоретические частоты

Пусть из генеральной совокупности X извлечена выборка объема n с наблюдаемыми частотами n_i:

x_i	x_1	$\cdots$	x_m
n_i	n_1	$\ldots$	n_m

где $n = n_1 + \ldots + n_m$.

Рассматриваемая выборка получена случайным образом, т.е. другая выборка такого же объема будет иметь другие частоты $\tilde{n}_i$, третья выборка будет иметь новые частоты. Некоторые из частот будут совпадать, другие отличаться.

Таким образом, наблюдаемые в выборке частоты n_i называются ***эмпирическими.***

Предположим, что закон распределения генеральной совокупности X неизвестен, но есть основание считать, что X обладает некоторым законом распределения.

Теоретическими частотами n_i' называются частоты, вычисленные в предположении, что генеральная совокупность X имеет то или иное распределение.

Сравнение теоретических и эмпирических частот позволяет при их близости сделать вывод об истинности предположения и предполагаемом законе распределения X.

Эмпирические и выравнивающие теоретические частоты

Теоретические частоты находят по формуле

$$n_i' = n \cdot P_i$$

где n - число испытаний (объем выборки), а вероятности P_i вычисляются следующим образом:

а) в случае дискретного признака X:

$$P_i = P(X = x_i)$$

вероятность наблюдаемого значения x_i, в предположении, что X имеет то или иное распределение.

б) в случае непрерывного признака X:

$$P_i = P(X \in \Delta x_i)$$

вероятность того, что значения признака X попадут в i-й частичный интервал. Из теории вероятностей следует, что

$$P_i \approx hp(\bar{x}_i)$$

где h - длина частичного интервала, $p(\bar{x}_i)$ - значения плотности распределения вероятностей в середине $\bar{x}_i$ интервала Δx_i.

Таким образом, в дискретном случае

$$n_i' = n \cdot P(X = x_i)$$

а в непрерывном

$$n_i' \approx nhp(\bar{x}_i)$$

Так как наиболее часто на практике встречается нормальное распределение, то рассмотрим вычисление n_i' по формуле для непрерывного случая в предположении, что X имеет нормальное распределение.

Эмпирические и выравнивающие теоретические частоты

Для непрерывного случая в случае, если случайная величина X подчиняется стандартному нормальному распределению, теоретические частоты вычисляются по формуле:

$$n_i' \approx \frac{nh}{\sigma_B} P_{0,1}(u_i)$$

где $u_i = \dfrac{\overline{x}_i - \overline{x}_B}{\sigma_B}$.

Построение гипотетической нормальной кривой

Применим последнюю формулу для n_i' и построим нормальную кривую по данному распределению выборки. Для этого:

1. вычисляем $\bar{x}_B$ и σ_B, например, методом произведений;
2. вычисляем $u_i = \dfrac{x_i - \bar{x}_B}{\sigma_B}$;
3. вычисляем $y_i = \dfrac{nh}{\sigma_B} P_{0,1}(u_i)$,

где n - объем выборки, h - разность между двумя соседними вариантами x_i; $P_{0,1}(u_i)$ вычисляется по таблице значений плотностей стандартного нормального распределения;

4. на плоскости строим точки (x_i, y_i) и соединяем их плавной кривой.

Согласие теоретического и статистического распределения

Если между теоретической кривой распределения F(X) и эмпирической функцией распределения существуют различия, то возникает вопрос: объясняются ли эти расхождения некоторыми случайными обстоятельствами, или же они связаны с тем, что эмпирическая функция распределения не описывается теоретической кривой?

Для ответа на этот вопрос используются критерии согласия.

Нулевая гипотеза H: случайная величина X подчиняется определенному закону распределения.

Критерий согласия Пирсона

Как бы точно не вычислялись теоретические частоты они, как правило, не совпадают с эмпирическими частотами ряда. Отсюда возникает необходимость сопоставления эмпирических частот с вычисленными, или ожидаемыми, частотами, с тем, чтобы установит достоверность или случайность наблюдаемого между ними расхождения. Нулевая гипотеза сводится к предположению, что несоответствие эмпирических частот частотам, вычисленным по тому или иному закону распределения, - совершенно случайное, т. е. между вычисленными и эмпирическими частотами никакой разницы нет. Для проверки нулевой гипотезы используются особые критерии. Одним из наиболее часто применяемых служит критерий χ2, предложенный к. Пирсоном в 1900 г. Этот критерий представляет сумму квадратов отклонений эмпирических частот (р) от частот теоретических или ожидаемых (р'), отнесенную к теоретическим частотам (р')

Критерий согласия Пирсона

Критерием согласия Пирсона называется критерий проверки гипотезы о нормальном распределении генеральной совокупности.

Для проверки гипотезы H_0 о том, что генеральная совокупность X имеет нормальный закон распределения, извлекают выборку объема n:

x_i	x_1	…	x_m
n_i	n_1	…	n_m

$n = n_1 + \ldots + n_m$

где n_i – эмпирические частоты.

Предположим, что X имеет нормальный закон распределения. Тогда можно вычислить теоретические частоты:

x_i	x_1	…	x_m
n'_i	n'_1	…	n'_m

Если гипотеза H_0 верна, то в идеальном случае $n_i = n'_i$, однако на практике $n_i \neq n'_i$. В этом случае требуется оценить их близость друг к другу по критерию:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n'_i)^2}{n'_i},$$

где число степеней свободы $k = m - 3$.

Сравнение теоретических и эмпирических распределений

- Нулевая гипотеза. **Согласно этой гипотезе первоначально принимается, что между эмпирическим и теоретическим распределением признака в генеральной совокупности достоверного различия нет.**
- **H_0 – нулевая гипотеза**
- **H1 –альтернативная гипотеза**

- Число степеней свободы **– это общее число величин, по которым вычисляются соответствующие статистические показатели, минус число тех условий, которые связывают эти величины, то есть уменьшают возможности вариации между ними. Число степеней свободы определяется по следующей** формуле:

  $df=k-r-1$, **где k – число интервалов, r – число параметров предполагаемого распределения. Для нашего случая** $r=2$, **следовательно,** $df=k-3$.
- **По заданному уровню значимости** (α) **и числу степеней свободы** df, **находим критическое значение** $\chi^2_{кр}$ (α,df).
- Если $\chi^2_{эмп} < \chi^2_{кр}$ **гипотеза о согласии эмпирического и теоретического распределения** не отвергается.

Критерий согласия Пирсона

Таким образом, для того чтобы при заданном уровне значимости α проверить гипотезу о нормальном распределении генеральной совокупности необходимо:

а) Вычислить наблюдаемое значение критерия

$$\chi^2_{набл} = \sum_{i=1}^{m} \frac{\left(n_i - n'_i\right)^2}{n'_i},$$

где n_i - эмпирические (опытные) частоты, n'_i - теоретические частоты, m - число вариант выборки.

б) Найти $\chi^2_{кр}(\alpha, k)$, где k = m - 3, по таблице критических точек распределения χ^2 (таблица 5).

в) Если $\chi^2_{набл} < \chi^2_{кр}$ – нет оснований отвергнуть нулевую гипотезу о нормальном распределении генеральной совокупности. Другими словами, эмпирические и теоретические частоты различаются незначимо. Если $\chi^2_{набл} > \chi^2_{кр}$ – нулевую гипотезу отвергаем. Другими словами, эмпирические и теоретические частоты различаются значимо.

«χ-квадрат» критерий