

КОРРЕЛЯЦИОННАЯ СВЯЗЬ И ЕЕ СТАТИСТИЧЕСКОЕ ИЗУЧЕНИЕ

Корреляционная связь — это связь, проявляющаяся не в каждом отдельном случае, а в массе случаев в средних величинах в форме тенденции.

Статистическое исследование ставит своей конечной целью получение модели зависимости для ее практического использования. Решение этой задачи осуществляется в следующей последовательности.

1. Логический анализ сущности изучаемого явления и причинно-следственных связей.

В результате устанавливаются результативный показатель (y) и факторы его изменения, характеризуемые показателями ($x_1, x_2, x_3, \dots, x_n$). Связь двух признаков (y и x) называется парной корреляцией. Влияние нескольких факторов на результативный признак называется множественной корреляцией.

По общему направлению связи могут быть прямыми и обратными. При прямых связях с увеличением признака x увеличивается и признак y , при обратных с увеличением признака x признак y уменьшается.

2. Сбор первичной информации и проверка ее на однородность и нормальность распределения.

Для оценки однородности совокупности используется коэффициент вариации по факторным признакам:

$$V_{x_i} = \frac{\sigma_{x_i}}{\bar{x}_i} \times 100.$$

Совокупность считается однородной, если коэффициент вариации не превышает 33%. Проверка нормальности распределения исследуемых факторных признаков ($x_1, x_2, x_3, \dots, x_n$) проводится с помощью правила «трех сигм». Результаты проверки на нормальность распределения следует представлять в табличной форме (табл. 23.1).

Таблица 23.1

Интервалы значений признака-фактора	Число единиц, входящих в интервал	Удельный вес единиц, входящих в интервал, в общем их числе, %	Удельный вес единиц, входящих в интервал, при нормальном распределении, %
1	2	3	4
$(\bar{x}_i - \sigma_{x_i}) - (\bar{x}_i + \sigma_{x_i})$			68,3
$(\bar{x}_i - 2\sigma_{x_i}) - (\bar{x}_i + 2\sigma_{x_i})$			95,4
$(\bar{x}_i - 3\sigma_{x_i}) - (\bar{x}_i + 3\sigma_{x_i})$			99,7

Сопоставление данных граф 3 и 4 позволяет судить о наличии или об отсутствии нормальности распределения.

На практике часто встречаются случаи отклонения от этих двух предпосылок. Однако это не означает, что следует отказаться от применения корреляционного анализа.

3. Исключение из массива первичной информации всех резко выделяющихся (аномальных) единиц по уровню признаков-факторов.

Исключаются все единицы, у которых уровень признака-фактора не попадает в интервал $\bar{x}_i \pm 3\sigma_{x_i}$ ($\bar{x}_i - 3\sigma_{x_i} \leq x_i \leq \bar{x}_i + 3\sigma_{x_i}$), и формируется новый массив для последующего анализа.

4. Установление факта наличия и направления корреляционной зависимости между результативным (y) и факторным (x) признаками.

Для установления наличия корреляционной связи используется ряд специфических методов: параллельное сопоставление рядов результативного и факторного признаков, графическое изображение фактических данных с помощью поля корреляции, построения корреляционной таблицы.

Основным методом выявления наличия корреляционной связи является метод аналитической группировки и определения групповых средних. Он заключается в том, что все единицы совокупности разбиваются на группы по величине признака-фактора и для каждой группы определяется средняя величина результативного признака. На основе данных аналитической группировки строится график эмпирической линии связи (линии регрессии), вид которой не только позволяет судить о возможном наличии связи, но и дает некоторое представление о форме корреляционной связи. Если эмпирическая линия связи по своему виду приближается к прямой линии, то можно предположить наличие прямой линейной корреляционной связи; если эмпирическая линия приближается к какой-либо кривой, то это связано с наличием криволинейной связи.

5. После установления факта наличия связи и ее формы измеряется степень тесноты связи и проводится оценка ее существенности.

Для определения степени тесноты парной линейной зависимости служит линейный коэффициент корреляции (r); при любой форме зависимости (линейной и криволинейной) — эмпирическое корреляционное отношение (η).

Для расчета **линейного коэффициента корреляции** по негруппированным данным могут быть использованы следующие формулы:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}; \quad r = \frac{\sum dx dy}{\sqrt{\sum dx^2 \sum dy^2}};$$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

где $(x - \bar{x})$ — отклонения вариантов значений признака- фактора от их средней величины; $x - \bar{x} = dx$; $(y - \bar{y})$ — отклонения вариантов значений результативного признака от их средней величины; $y - \bar{y} = dy$; n — число единиц в совокупности; σ_x , σ_y — среднее квадратическое отклонение соответственно признака-фактора и результативного признака.

Линейный коэффициент корреляции может принимать значения в пределах от -1 до $+1$. Чем ближе он по абсолютной величине к 1 , тем теснее связь. Знак при нем указывает направление связи: знак « $+$ » соответствует прямой зависимости, знак « $-$ » — обратной.

Если коэффициент корреляции равен нулю, то связи между признаками нет; если он равен единице (с любым знаком), то между признаками существует функциональная связь.

Оценка существенности линейного коэффициента корреляции при большом объеме выборки (свыше 500) проводится с использованием отношения коэффициента корреляции (r) к его средней квадратической ошибке (σ_r):

$$t_{\text{расч}} = \frac{|r|}{\sigma_r}, \quad \text{где} \quad \sigma_r = \frac{1 - r^2}{\sqrt{n - 1}}.$$

Если это отношение окажется больше значения t -критерия Стьюдента, определяемого в установленном в статистике порядке при числе степеней свободы $k = n - 2$ и с вероятностью $(1 - \alpha)$, то следует говорить о существенности коэффициента корреляции (α -уровень значимости 0,01 или 0,05).

При недостаточном большом объеме выборки величину средней квадратической ошибки коэффициента корреляции определяют по формуле:

$$\sigma_r = \frac{\sqrt{1 - r^2}}{\sqrt{n - 2}}.$$

В этом случае

$$t_{\text{расч}} = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}.$$

Полученная величина $t_{\text{расч}}$ сравнивается с табличным значением t -критерия Стьюдента.

В тех случаях, когда r получен по данным малой выборки, для проверки его существенности целесообразно использовать метод преобразованной корреляции, предложенный Р. Фишером.

Средняя квадратическая ошибка Z -распределения зависит только от объема выборки и определяется по формуле:

$$\sigma_z = \frac{1}{\sqrt{n - 3}}.$$

По таблице соотношений между u и Z по специальной таблице находят значение Z , соответствующее рассчитанному коэффициенту корреляции.

Если соотношение Z к средней квадратической ошибке (Z/σ_z) окажется больше табличного значения критерия Стьюдента при определенном уровне значимости, то можно говорить о наличии связи между признаками в генеральной совокупности.

Корреляционное отношение основано на использовании известной теории сложения дисперсий и определяется по формулам:

$$\eta = \sqrt{\frac{\delta^2}{\sigma_y^2}}; \quad \eta = \sqrt{1 - \frac{\bar{\delta}^2}{\sigma_y^2}},$$

где δ^2 — межгрупповая дисперсия результативного признака, вызванная влиянием признака-фактора; σ_y^2 — общая дисперсия результативного признака; $\bar{\delta}^2$ — средняя внутригрупповая дисперсия результативного признака:

$$\delta^2 = \frac{\sum (\bar{y}_i - \bar{y}_0)^2 n_i}{\sum n_i};$$

$$\sigma_y^2 = \frac{\sum (y - \bar{y}_0)^2 n_i}{\sum n_i};$$

$$\bar{\delta}^2 = \frac{\sum \sigma_i^2 n_i}{\sum n_i};$$

где $\bar{y}_i$ — среднее значение результативного признака в соответствующих группах, выделенных по величине признака-фактора; $\bar{y}_0$ — общая средняя для всей совокупности; n_i — число единиц в соответствующих группах; σ_i^2 — внутригрупповая дисперсия.

Вычисление корреляционного отношения требует достаточно большого объема информации, которая должна быть представлена в форме групповой таблицы или в форме корреляционной таблицы, то есть обязательным условием является группировка данных по признаку-фактору (изменяется от 0 до 1).

При недостаточном количестве данных в выделенных группах к рассчитанной величине корреляционного отношения вносится поправка:

$$\eta_{\text{скор}}^2 = 1 - (1 - \eta^2) \frac{n-1}{n-m},$$

где m — число выделенных групп.

Корреляционное отношение в квадрате (η^2) называют **коэффициентом детерминации** (причинности), он отражает долю факторной дисперсии в общей дисперсии.

В практике могут быть использованы и другие показатели для определения степени тесноты связи.

Элементарной характеристикой степени тесноты связи является **коэффициент Фехнера**:

$$K_{\phi} = \frac{n_a - n_b}{n_a + n_b}.$$

где n_a — количество совпадений знаков отклонений индивидуальных величин факторного признака x и результативного признака y от их средней арифметической величины (например, «плюс» и «плюс», «минус» и «минус», «отсутствие отклонения» и «отсутствие отклонения»); n_b — количество несовпадений знаков отклонений индивидуальных значений изучаемых признаков от значения их средней арифметической.

Коэффициент Фехнера целесообразно использовать для установления факта наличия связи при небольшом объеме исходной информации. Он изменяется в пределах $-1,0 \leq K_{\phi} \leq 1,0$.

Для определения тесноты связи как между количественными, так и между качественными признаками, при условии, что значения этих признаков могут быть проранжированы по степени убывания или возрастания, используется **коэффициент корреляции рангов Спирмэна**:

$$r = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

где d_i — разность между величинами рангов признака-фактора и результативного признака; n — число показателей (рангов) изучаемого ряда.

Коэффициент корреляции рангов Спирмэна варьирует в пределах от $-1,0$ до $+1,0$.

Ранговый коэффициент обычно исчисляется на основе небольшого объема исходной информации, поэтому необходимо выполнить проверку его существенности по таблице предельных значений коэффициента корреляции рангов Спирмэна при условии верности нулевой гипотезы об отсутствии корреляционной связи при заданном уровне значимости и определенном объеме выборочных данных.

Если полученное значение r превышает критическую величину при данном уровне значимости, то нулевая гипотеза может быть отвергнута, то есть величина r не является результатом случайных совпадений рангов.

Для исследования степени тесноты связи между качественными признаками, каждый из которых представлен в виде альтернативных признаков, может быть использован **коэффициент ассоциации Д. Юла** или **коэффициент контингенции К. Пирсона**. Расчетная таблица в этом случае состоит из четырех ячеек (таблица «четырёх полей»), статистическое сказуемое которой схематически может быть представлено в следующем виде (табл. 23.2).

Таблица 23.2

Признаки	A (да)	$\bar{A}$ (нет)	Итого
B (да)	a	b	$a+b$
$\bar{B}$ (нет)	c	d	$c+d$
Итого	$a+c$	$b+d$	n

В расчетной таблице: a, b, c, d — частоты взаимного сочетания (комбинации) двух альтернативных признаков — A — $\bar{A}$ и B — $\bar{B}$; n — общая сумма частот.

Коэффициент ассоциации исчисляется по формуле:

$$K_A = \frac{ad - bc}{ad + bc}.$$

Коэффициент контингенции находится по формуле:

$$K_A = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}},$$

где a, b, c, d — числа в четырехклеточной таблице.

Коэффициент контингенции также изменяется от -1 до $+1$, но всегда его величина для тех же данных меньше коэффициента ассоциации.

Для оценки тесноты связи между альтернативными признаками, принимающими любое число вариантов значений, применяется **коэффициент взаимной сопряженности К. Пирсона** и **коэффициент взаимной сопряженности А. А. Чупрова**.

Первичная статистическая информация для исследования этой связи располагается в форме таблицы (табл. 23.3).

Таблица 23.3

Признаки	A	B	C	Итого
D	f_{11}	f_{12}	f_{13}	A_{1j}
E	f_{21}	f_{22}	f_{23}	A_{2j}
F	f_{31}	f_{32}	f_{33}	A_{3j}
Итого	A_{1j}	A_{2j}	A_{3j}	n

В таблице f_{ij} — частоты взаимного сочетания двух атрибутивных признаков; n — число пар наблюдений.

Коэффициент взаимной сопряженности К. Пирсона определяется по формуле:

$$C = \sqrt{\frac{\varphi^2}{1 + \varphi^2}},$$

где φ^2 — показатель средней квадратической сопряженности.

Показатель φ^2 определяется как сумма отношений квадратов частот каждой клетки таблицы к произведению итоговых частот соответствующего столбца и строки за минусом единицы.

$$\varphi^2 = \left(\sum \frac{f_{ij}^2}{A_i A_j} \right) - 1,$$

где f_{ij} — частоты каждой клетки; i — номер строки; A_i — итоговые частоты по строкам; A_j — итоговые частоты по графам.

Коэффициент взаимной сопряженности А. А. Чупрова исчисляется по формуле:

$$K = \sqrt{\frac{\varphi^2}{(K_1 - 1)(K_2 - 1)}},$$

где φ^2 имеет одинаковое значение с показателем Пирсона и является показателем взаимной сопряженности; K_1 — число групп по столбцам таблицы; K_2 — число групп по строкам таблицы.

Коэффициент взаимной сопряженности Чупрова (K) является более гибким, поскольку он учитывает число образуемых по каждому признаку групп (K_1 и K_2), поэтому результат является более точным по сравнению с коэффициентом взаимной сопряженности по формуле Пирсона.

Коэффициент взаимной сопряженности изменяется от 0 до 1.

6. После установления достаточной степени тесноты связи выполняется построение модели связи (уравнения регрессии).

Тип модели выбирается на основе сочетания теоретического анализа и исследования эмпирических данных посредством построения эмпирической линии регрессии. Чаще всего используются следующие типы функций:

- а) линейная — $\hat{y}_x = a + bx$;
- б) гиперболическая — $\hat{y}_x = a + b \frac{1}{x}$;
- в) параболическая — $\hat{y}_x = a + bx + cx^2$;
- г) показательная — $\hat{y}_x = ab^x$;

Для определения численных значений параметров уравнения связи (линии регрессии) используется метод наименьших квадратов и решается система нормальных уравнений.

Для определения параметров a и b уравнения прямолинейной корреляционной связи система нормальных уравнений (для негруппированных данных) следующая:

$$\begin{cases} \sum y = an + b \sum x \\ \sum yx = a \sum x + b \sum x^2 \end{cases}$$

Параметры a и b можно определить по следующим формулам:

$$a = \bar{y} - b\bar{x}; \quad b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} \quad \text{или} \quad b = \frac{\sum d_x d_y}{\sum d_x^2}.$$

Для проверки возможности использования линейной функции определяется разность квадратов ($\eta^2 - r^2$). Если она менее 0,1, то считается возможным применение линейной функции. Для решения этой же задачи можно использовать величину ω^2 , определяемую по формуле:

$$\omega^2 = \frac{\eta^2 - r^2}{m - 2} : \frac{1 - \eta^2}{n - m},$$

где m — число групп, на которое разделен диапазон значений факторного признака.

Если ω^2 окажется меньше табличного значения F -критерия, то нулевая гипотеза о возможности использования в качестве уравнения регрессии линейной функции не опровергается. Значение F -критерия определяется по таблице в зависимости от уровня значимости $\alpha = 0,05$ (вероятность $P = 0,95$) и числа степеней свободы числителя ($k_1 = m - 2$) и знаменателя ($k_2 = n - m$).¹

Для определения параметров гиперболической функции системы нормальных уравнений следующая:

$$\begin{cases} \sum y = an + b \sum \frac{1}{x} \\ \sum y \frac{1}{x} = a \sum \frac{1}{x} + b \sum \left(\frac{1}{x} \right)^2 \end{cases}$$

¹ См.: Ефимова М. и др. Практикум по общей теории статистики. — М.: Финансы и статистика, 2002 (приложение 5).

Для определения параметров параболы второго порядка система нормальных уравнений такова:

$$\begin{cases} \sum y = an + b\sum x + c\sum x^2 \\ \sum yx = a\sum x + b\sum x^2 + c\sum x^3 \\ \sum yx^2 = a\sum x^2 + b\sum x^3 + c\sum x^4 \end{cases}$$

В качестве меры достоверности уравнения корреляционной зависимости используется процентное отношение средней квадратической ошибки уравнения (S_e) к среднему уровню результативного признака ($\bar{y}$):

$$\frac{S_e}{\bar{y}} \times 100; \quad S_e = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{n - l}},$$

где y — фактические значения результативного признака; $\hat{y}_x$ — значения результативного признака, рассчитанные по уравнению регрессии; l — число параметров в уравнении регрессии.

Если это отношение не превышает 10—15%, то следует считать, что уравнение регрессии достаточно хорошо отображает изучаемую взаимосвязь.

Полученное уравнение регрессии используется для экстраполяции. Однако ее можно применять лишь тогда, когда существенно не изменились условия формирования уровней признаков.

7. Изучение множественной корреляционной зависимости начинается с анализа матрицы парных коэффициентов корреляции. Это позволяет произвести отбор факторов, включаемых в модель множественной зависимости. Матрица имеет следующий вид (табл. 23.4).

Таблица 23.4

Признаки	y_0	x_1	x_2	...	x_k
y_0	1	r_{01}	r_{02}	...	r_{0k}
x_1	r_{01}	1	r_{21}	...	r_{k1}
x_2	r_{02}	r_{12}	1	...	r_{k2}
...	...	...	...	...	...
x_k	r_{0k}	r_{1k}	r_{2k}	1	1

Анализ первой строки матрицы позволяет выявить факторы, у которых степень тесноты связи с результативным показателем значительна, а поэтому они могут быть включены в модель. Однако при построении многофакторных моделей должно соблюдаться требование возможно меньшей коррелированности включенных в модель признаков-факторов (отсутствие мультиколлинеарности). В качестве критерия мультиколлинеарности может быть принято соблюдение следующих неравенств:

$$r_{x_j y} > r_{x_j k}; \quad r_{x_k y} > r_{x_j x_k}$$

Если приведенные неравенства (или хотя бы одно из них) не выполняются, то исключается тот фактор x_j или x_k , связь которого с результативным признаком y будет менее тесной.

8. Отобранные факторы включаются в модель множественной зависимости. При этом следует учитывать, что число факторов, включаемых в модель, должно быть в 5—6 раз меньше, чем число единиц, входящих в совокупность.

9. Для измерения степени тесноты связи между изменениями величины результативного признака (y) и изменениями значений факторных признаков определяется коэффициент множественной (совокупной) корреляции (R).

Для случая зависимости результативного признака от двух факторных признаков формула совокупного коэффициента корреляции имеет вид:

$$R_{yx_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2}}.$$

Если число факторов-признаков более двух, то совокупный коэффициент корреляции определяется следующим образом:

$$R_{yx_1x_2...x_k}^2 = 1 - \frac{\Delta}{\Delta^x},$$

где Δ — матрица парных коэффициентов корреляции; Δ^x соответствует матрице парных коэффициентов корреляции (Δ) без верхней строки и первого столбца.

Величина R^2 называется коэффициентом детерминации; она показывает, в какой мере вариация результативного признака обусловлена влиянием признаков-факторов, включенных в уравнение множественной зависимости.

Величина совокупного коэффициента корреляции изменяется в пределах от 0 до 1 и численно не может быть меньше, чем любой из образующих его парных коэффициентов корреляции. Чем ближе он к единице, тем меньше роль неучтенных в модели факторов и тем более оснований считать, что параметры регрессионной модели отражают степень эффективности включенных в нее факторов.

Для оценки существенности (значимости) совокупного коэффициента корреляции используется критерий F-Фишера.

Для этого определяется F-расчетное по следующей формуле:

$$F_{\text{расч}} = \frac{\sigma_{\hat{y}}^2}{\sigma_E^2} \times \frac{n - l}{l - 1},$$

где $\sigma_{\hat{y}}^2$ — факторная дисперсия результативного признака, обусловленная вариацией признаков-факторов;

$$\sigma_y^2 = \frac{\sum (\hat{y} - \bar{y})^2}{n},$$

где $\hat{y}$ — значения результативного признака, рассчитанные по уравнению регрессии; σ_E^2 — остаточная дисперсия:

$$\sigma_E^2 = \sigma_y^2 - \sigma_{\hat{y}}^2;$$

σ_y^2 — общая дисперсия результативного признака; n — число данных; l — число параметров уравнения.

10. Для сравнения роли различных факторов в формировании моделируемого показателя определяется коэффициент эластичности (ε_j) или β -коэффициент (β_j).

Частный коэффициент эластичности показывает, на сколько процентов в среднем изменяется результативный признак y с изменением признака-фактора x на 1%, и определяется по формуле:

$$\varepsilon_j = b_j \frac{\bar{x}_j}{\bar{y}},$$

где b_j — коэффициент регрессии при j -м факторе.

β -коэффициент показывает, на какую часть среднего квадратического отклонения изменится результативный показатель при изменении соответствующего фактора (x) на величину его среднего квадратического отклонения. Его формула имеет вид:

$$\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y}.$$

Глава 24

РЯДЫ ДИНАМИКИ

§ 24.1. ПОКАЗАТЕЛИ РЯДОВ ДИНАМИКИ

При изучении рядов динамики перед статистикой стоят следующие задачи: охарактеризовать интенсивность развития явления от периода к периоду (от даты к дате), а также среднюю интенсивность развития за исследуемый период, изучить сезонные колебания, выявить основную тенденцию в развитии явления, осуществить прогноз развития на будущее.

Для изучения **интенсивности изменения уровней ряда во времени** исчисляются следующие показатели динамики:

- ♦ абсолютные приросты;
- ♦ коэффициенты роста;
- ♦ темпы роста;
- ♦ темпы прироста;
- ♦ абсолютные значения одного процента прироста.

Перечисленные показатели динамики можно исчислять с переменной или постоянной базой. Если производится сравнение каждого уровня с предыдущим уровнем, то получают *показатели динамики с переменной базой* (цепные показатели динамики). Если каждый уровень сравнивается с начальным уровнем или каким-то другим, принятым на базу сравнения, то получают *показатели динамики с постоянной базой* (базисные показатели динамики). База сравнения должна выбираться обоснованно, в зависимости от экономических особенностей явления и задач исследования.

Методы расчета показателей динамики представлены в табл. 24.1. Они одинаковы и для моментных, и для интервальных рядов.

При расчете показателей приняты следующие условные обозначения:

y_i — уровень любого периода (кроме первого), называемый уровнем текущего периода;

y_{i-1} — уровень периода, предшествующего текущему;

y_k — уровень, принятый за постоянную базу сравнения (часто начальный уровень).