Статистическое изучение взаимосвязей социально-экономических явлений

В процессе исследования зависимостей вскрываются причинно-следственные отношения, что позволяет выявить факторы (причины), оказывающие существенное влияние на вариацию изучаемых явлений и процессов. *Причина* – это совокупность условий, обстоятельств, действие которых приводит к появлению *следствия*.

На основе проведения качественного анализа появляется возможность разделить признаки на два класса:

а) *факторные признаки* (*факторы*), которые обусловливают изменение других признаков;

б) *результативные признаки*, которые изменяются под действием факторных признаков.

Связи между явлениями классифицируют по различным направлениям:

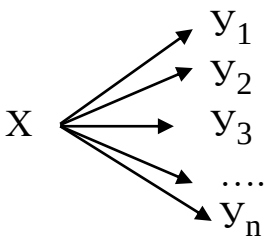
по характеру зависимости различают функциональную и стохастическую связь. Связь между признаками называют *функциональной* (*детерминированной*), если каждому значению одного из них соответствует одно (или несколько, в случае множественных связей), вполне определенное значение другого. Такая зависимость является строгой, точной, полной.

Схематично функциональную связь можно представить следующим образом: $X \Rightarrow Y$.

В общем виде функциональную связь можно записать: $y_i = f(x_i)$.

Для социально-экономических явлений характерно то, что наряду с существенными факторами, определяющими в основном величину результативного признака, на него оказывают воздействие многие другие, в том числе и случайные факторы. Такая зависимость называется *стохастической*.

При стохастической зависимости изменение факторного признака приводит к изменению закона распределения результативного признака. Схему стохастической связи можно представить следующим образом:



Частным случаем стохастической связи является *корреляционная*, при которой изменение среднего значения результативного признака обусловлено изменением факторного (факторных) признака.

Корреляционная связь является неполной, нестрогой и проявляется лишь при достаточно большом числе случаев. Схематично ее можно представить следующим образом: $X \Rightarrow \overline{Y}$.

В общем виде корреляционную связь можно записать: $\overline{y}_i = f(x_i)$.

по степени тесноты связи делятся на *слабые*, *умеренные* и *сильные* (*тесные*). **по направлению** различают связи *прямые* и *обратные*.

по аналитическому выражению выделяют связи *прямолинейные* (линейные) и *криволинейные* (нелинейные).

в зависимости от количества факторов, влияющих на результат, различают *парную* и *многофакторную* (*множественную*) связь.

К основным методам изучения функциональных связей относятся: графический, индексный, балансовый, аналитических группировок и другие.

К методам изучения корреляционных связей относятся: графический, аналитических группировок, параллельных рядов и др., а также дисперсионный, корреляционный и регрессионный анализ и др.

Графический метод позволяет изобразить взаимосвязь между признаками с помощью корреляционного поля («поля рассеяния»), которое является наглядным изображением корреляционной таблицы. В системе координат на оси абсцисс откладываются значения факторного признака, а на оси ординат – результативного (рис. 1 – 6).

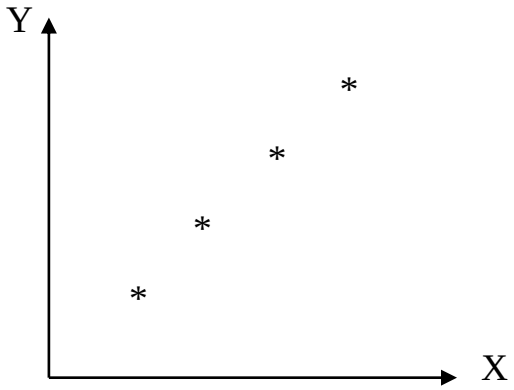


Рис. 1. **График функциональной зависимости**
Связь функциональная, прямая, линейная

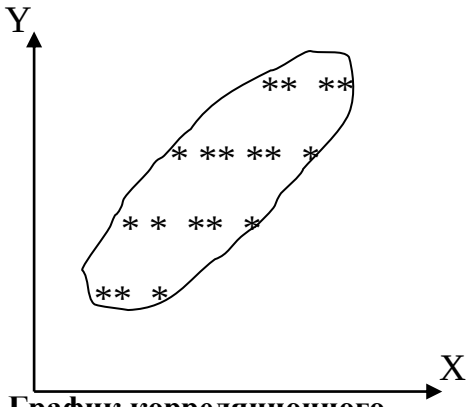


Рис. 2. **График корреляционного поля**
Связь корреляционная, прямая, линейная

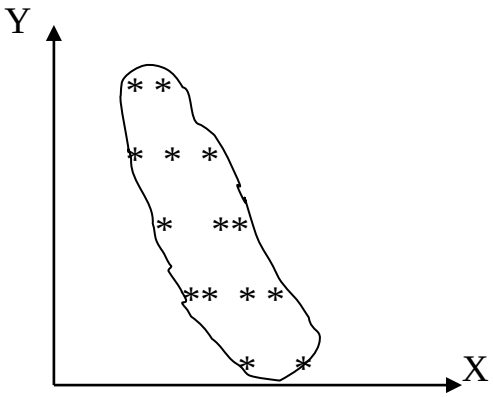


Рис. 3. **График корреляционного поля**
Связь корреляционная, обратная, линейная

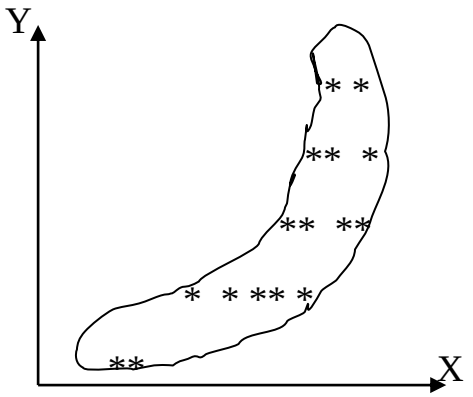


Рис. 4. **График корреляционного поля**
Связь корреляционная, прямая, параболическая

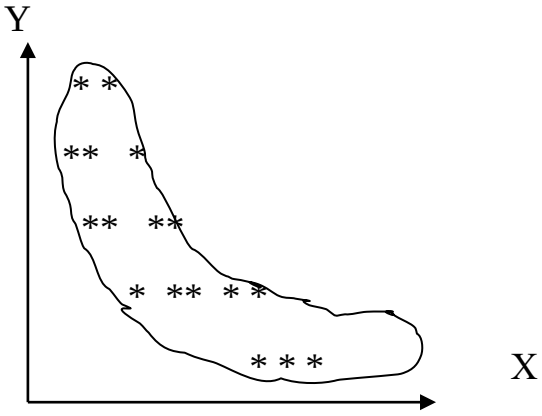


Рис. 5. **График корреляционного поля**
Связь корреляционная, обратная, гиперболическая

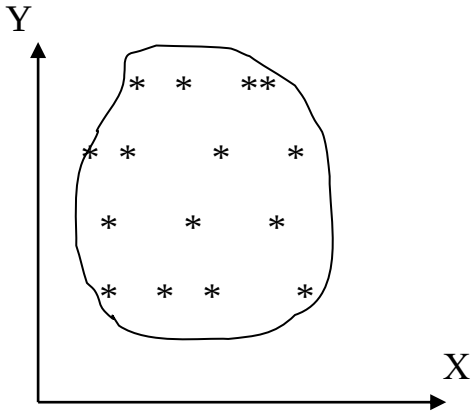


Рис. 6. **График корреляционного поля**
Зависимость между X и Y отсутствует

По расположению точек, их концентрации в определенном направлении можно судить о наличии связи.

Корреляционно-регрессионный анализ
Корреляционный анализ связей

Корреляция – это статистическая зависимость между случайными величинами, не имеющими строгого функционального характера, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.

Основные задачи корреляционного анализа сводятся к следующим:

1. Количественно охарактеризовать тесноту связи между результативным и факторными признаками.
2. Выявить направление изменения результативного признака в зависимости от роста или снижения факторного.
3. Ответить на вопрос: случайна или неслучайна выявленная связь?

При практическом применении корреляционный анализ включает в себя несколько этапов:

1) постановка задачи и выбор факторных и результативных признаков;
2) сбор статистического материала и его первичная обработка;
3) предварительное изучение взаимосвязей;
4) исследование парных зависимостей;
5) исследование многофакторных связей;
6) оценка результатов исследования;
7) анализ результатов проведения корреляционного анализа.

Если связь между признаками является прямолинейной, то для ее определения используют *линейный коэффициент корреляции*:

$$r = \frac{\overline{(X - \overline{X}) \times (Y - \overline{Y})}}{\sigma_X \times \sigma_Y}. \qquad (1)$$

Линейный коэффициент корреляции изменяется от (– 1) до (+ 1). На практике приняты следующие пределы качественной характеристики тесноты связи по абсолютной величине (табл. 1).

Таблица 1

Количественные критерии оценки тесноты связи

Величина коэффициента корреляции (по модулю)	Характер связи
0 – 0,1	Связь практически отсутствует или не подчиняется уравнению прямой
0,1 – 0,3	Связь слабая
0,3 – 0,65	Связь средней тесноты (умеренная)
0,65 – 0,8	Связь тесная (сильная)
0,8 – 0,95	Связь очень тесная, практически изменение результативного признака определено изменением факторного
0,95 – 1,0	Связь функциональная, т. е. все точки (X, Y) лежат на прямой линии, имеет место строго пропорциональная зависимость в изменении Y и X

Если линейный коэффициент корреляции принимает положительные значения, то связь между признаками прямая, а если отрицательные – обратная. Если r = 0, то линейная корреляция отсутствует, т. е. признаки X и Y являются независимыми.

Если связь между признаками является криволинейной, то используется *индекс корреляции* (или *корреляционное отношение* η):

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}}, \qquad (2)$$

где σ^2 – вариация результативного признака за счет всех факторов – общая дисперсия;

$\overline{\sigma_i^2}$ – вариация результативного признака за счет всех факторов, кроме фактора X – средняя из групповых дисперсий;

δ^2 – вариация результативного признака за счет анализируемого фактора X – межгрупповая дисперсия ($\delta^2 = \sigma^2 - \overline{\sigma_i^2}$).

В случае функциональной связи индекс корреляции равен 1, а при полном отсутствии связи он принимает значение 0. Следовательно, $0 \le \eta \le 1$.

Величина η^2 называется *коэффициентом детерминации*. Его экономическое значение заключается в том, что он измеряет, насколько колеблемость результативного признака объяснена изменением факторного.

Оценка результатов исследования парной зависимости заключается в проверке выявленной связи на случайность с помощью корреляционной поправки:

$$\sigma_r = \sqrt{\frac{1-r^2}{n-1}}, \qquad (3)$$

где σ_r – среднеквадратическая ошибка выборочного коэффициента парной корреляции.

Если совокупность является малой, то корреляционная поправка рассчитывается по формуле:

$$\sigma_r = \sqrt{\frac{1-r^2}{n-2}}. \qquad (4)$$

С большой уверенностью можно утверждать, что коэффициент корреляции при достаточно большом числе наблюдений должен превышать среднюю ошибку не менее чем в 3 раза, т. е. $\frac{|r|}{\sigma_r} \geq 3$. Если это неравенство не выполняется, то существование связи между явлениями нельзя признать доказанным.

Значимость линейного коэффициента корреляции проверяется на основе t-критерия Стьюдента.

Критерий $t_{расч}$ определяется по формуле:

$$t_{расч} = \frac{|r|\sqrt{n-1}}{\sqrt{1-r^2}}, \qquad (5)$$

а для малой совокупности он равен:

$$t_{расч} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (6)$$

Данный критерий подчиняется закону распределения Стьюдента с числом степеней свободы $k = n - 2$.

В случае если $|t_{расч}| > |t_{табл}|$, то связь считают существенной, неслучайной. Если $|t_{расч}| < |t_{табл}|$, то связь считают несущественной, случайной.

Регрессионный анализ связей

Парная регрессия характеризует изменение среднего уровня результативного признака Y в зависимости от изменения признака-фактора X. Уравнение регрессии, в общем виде выражается функцией:

$$\overline{y_x} = f(x). \quad (7)$$

Уравнение линейной регрессии имеет широкое применение, его параметры легче определить и истолковать, но в действительности линейная связь существует относительно редко, поэтому выбор прямой линии может рассматриваться как некое упрощение, как эмпирический прием.

Уравнение линейной регрессии имеет вид:

$$\overline{y_x} = a_0 + a_1 x, \quad (8)$$

где $\overline{y_x}$ – рассчитанные выравненные значения результативного признака (переменная средняя) после подстановки в уравнение значений x;

a_i – неизвестные параметры уравнения регрессии;

a_0 – свободный член уравнения регрессии;

a_1 – коэффициент регрессии при факторе X.

Коэффициенты регрессии связаны с коэффициентами корреляции следующим образом:

$$a_{1(YX)} = r_{YX} \frac{\sigma_y}{\sigma_x}; \quad (9)$$

$$a_{1(XY)} = r_{XY} \frac{\sigma_x}{\sigma_y}; \quad (10)$$

Если выбрана форма связи, то далее проблема заключается в оценке параметров уравнения регрессии. Она может осуществляться различными методами, наибольшее распространение из них получил *метод наименьших квадратов* (МНК), который основывается на предположении независимости друг от друга отдельных наблюдений.

Сущность метода наименьших квадратов заключается в том, что отыскиваются такие значения параметров уравнения регрессии, при которых

сумма квадратов отклонений фактических значений результативного признака от вычисленных по уравнению будет наименьшей из всех возможных:

$$\Sigma\left(y_{\text{факт}} - \overline{y_x}\right)^2 \to \min . \qquad (11)$$

Применение метода наименьших квадратов для определения параметров уравнения сводится к задаче на экстремум.

Для линейной зависимости получаем:

$$\sum (y - a_0 - a_1 x)^2 \to \min .$$

Для линейной зависимости получаем следующую систему нормальных уравнений:

$$\begin{cases} a_0 n + a_1 \sum x = \sum y; \\ a_0 \sum x + a_1 \sum x^2 = \sum x y. \end{cases}$$

В этой системе n – объем исследуемой совокупности (число единиц наблюдений), или, с другой стороны, n – число пар значений x и y.

Задача заключается в решении системы двух уравнений с двумя неизвестными a_0 и a_1, имеющей единственное решение.

Если предварительно проведен корреляционный анализ, то для упрощения нахождения параметров a_0 и a_1 можно воспользоваться взаимосвязью коэффициента корреляции и коэффициента регрессии. В результате получим:

$$a_{1(YX)} = r_{YX} \frac{\sigma_y}{\sigma_x}, \qquad (12)$$

$$a_0 = \overline{y} - a_1 \overline{x} . \qquad (13)$$

Теоретической *линией регрессии* называется линия, вокруг которой группируются точки корреляционного поля и которая указывает на основное направление, основную тенденцию связи:

$$\Sigma\left(y_{\text{факт}} - \overline{y_x}\right) = 0 . \qquad (14)$$

Следовательно, теоретическая линия регрессии расположена так, что сумма отклонений точек поля корреляции от фактических (эмпирических)

точек линии регрессии равняется нулю, а сумма квадратов расстояний по вертикали между точками и этой линией минимальна.

Следующим этапом является проверка *адекватности моделей*, построенных на основе уравнений регрессии, которая начинается с определения значимости каждого коэффициента регрессии. Значимость коэффициента a_1 оценивается с помощью t-критерия Стьюдента:

$$t_{\text{расч}} = \frac{|a_1|}{\sigma_{a_1}}, \tag{15}$$

где σ_{a_1} – квадратическая ошибка коэффициента уравнения регрессии, рассчитываемая по формуле:

$$\sigma_{a_1} = \frac{\sum(y - \bar{y})^2}{(n-2) \times \sum(x - \bar{x})^2}. \tag{16}$$

Если $t_{\text{расч}} > t_{\text{табл}}$, значение параметра a_1 значимо.

Значение $t_{\text{табл}}$ берется из таблицы для заданного уровня значимости α, т. е. вероятности $P = 1 - \frac{\alpha}{2}$ и числа степеней свободы $\nu = n - 2$.

Качество построенной модели можно проверить по *коэффициенту детерминации*:

$$\eta^2 = 1 - \frac{\sum(y_{\text{факт}} - \overline{y_x})^2}{\sum(y_{\text{факт}} - \bar{y})^2}. \tag{17}$$

Коэффициент детерминации показывает, какую часть вариации результативного признака объясняет построенная модель.

Величина $(1 - \eta^2)$ характеризует долю дисперсии результативного признака Y, вызванную влиянием остальных не учтенных в модели факторов.

Проверка адекватности модели осуществляется с помощью абсолютной $(y - \overline{y_x})$, относительной $(\frac{|y - \overline{y_x}|}{y} \times 100\,\%)$ и средней ошибки аппроксимации, которая не должна превышать 12 – 15 %.

Средняя ошибка аппроксимации рассчитывается по формуле:

$$\bar{\varepsilon} = \frac{1}{n}\sum \frac{\left| y - \overline{y_x} \right|}{y} \times 100\,\%. \qquad (18)$$

Завершающим этапом регрессионного анализа является практическое применение построенного уравнения. С помощью уравнения можно:

1. Вскрыть резервы производства, рассмотрев причины существенных различий в эмпирических и теоретических значениях результативного показателя.

2. Составить прогноз результативного показателя, взяв конкретные значения факторного.

Рассчитаем среднюю квадратическую ошибку уравнения регрессии, предложенную Ф. И. Эджвортом:

$$S_{\overline{y_x}} = \sqrt{\frac{\sum (y_{факт} - \overline{y_x})^2}{n - m}}, \qquad (19)$$

где m – число параметров в уравнении (в случае прямой m = 2).

Относительная ошибка уравнения регрессии – коэффициент вариации, равна:

$$K_{S_{\overline{y_x}}} = \frac{S_{\overline{y_x}}}{\bar{y}} \times 100\,\%. \qquad (20)$$

Если коэффициент вариации имеет значение менее 33 %, то построенным уравнением регрессии можно пользоваться для принятия управленческих решений.