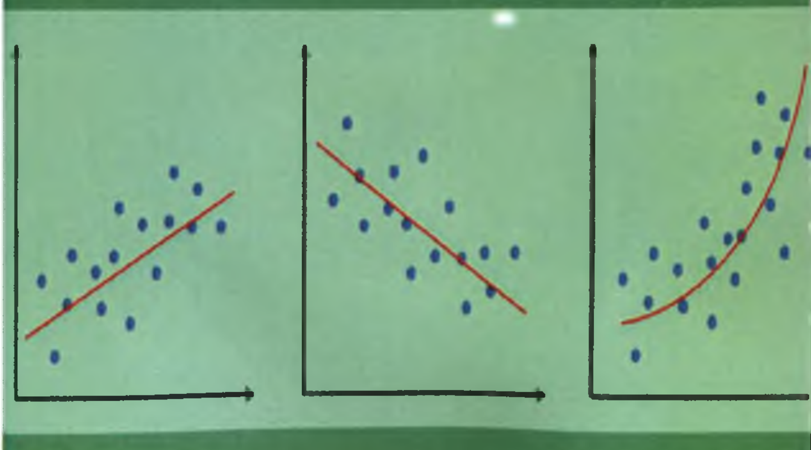
$$r = \frac{\overline{xy} \quad \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}$$

Ё.Абдуллаев,
З.Тошматов, М.Икрамов

КОРРЕЛЯЦИОННО-РЕГРЕСИОННЫЙ АНАЛИЗ

8



СЕРИЯ
УЧЕБНО-МЕТОДИЧЕСКИХ ПОСОБИЙ
ПО КУРСУ «СТАТИСТИКА»

МИНИСТЕРСТВО ВЫСШЕГО И СРЕДНЕГО
СПЕЦИАЛЬНОГО ОБРАЗОВАНИЯ
РЕСПУБЛИКИ УЗБЕКИСТАН

Ташкентский финансовый институт

СЕРИЯ учебно-методических пособий
по курсу «Статистика»

Ё. Абдуллаев,
З. Тошматов, М. Икрамов

Корреляционно-регрессионный анализ

ТАШКЕНТ
«IQTISOD-MOLIYA»
2021

УДК: 311(075.8)
ББК: 65.051
А-139

Рецензенты: *д-р экон. наук, проф.* **Т. Шодиев;**
ст. преп. **Б. Акбарова**

С 78 Корреляционно-регрессионный анализ: Учебное пособие / Ё. Абдуллаев, З. Тошматов, М. Икрамов; – Т.: «Iqtisod-Moliya», 2021. – 116 с.

Статистическое распределение характеризуется вариацией значений признака у единиц изучаемой совокупности, поэтому возникает естественный вопрос о причинах, влияющих на изменение значений признака. Изучение зависимости вариации признака от окружающих условий составляет теорию корреляции (основоположники – английские ученые Ф. Гальтон и К. Пирсон).

Содержание данного учебного пособия составляет всестороннее исследование корреляционных связей, в том числе нахождение уравнений регрессии, измерение тесноты и направления связи, а также определение возможных ошибок, как параметров уравнения регрессии, так и показателей тесноты связи. Особое внимание уделяется интерпретации полученных характеристик.

Для более глубокого и самостоятельного изучения темы в пособие приложены: сборник тестов, практикум, сборник слайдов и глоссарий; в конце текста лекций приводится список вопросов интеллектуального тренинга.

УДК: 311(075.8)
ББК: 65.051

ISBN 978-9943-13-970-1

Тексты лекций

Тема 8. Корреляционно-регрессионный анализ

Дорожная карта

8.1. Понятие о корреляционной связи

Общая постановка задачи

В научном исследовании общественных явлений первостепенное значение имеет выявление и количественное выражение связей между ними.

Связи между явлениями и отдельными их признаками весьма многообразны. Но в любой конкретной связи одни признаки выступают в качестве факторов, воздействующих на другие и обусловливающих их изменение, другие же – в качестве результатов действия этих факторов. Иными словами, одни представляют собой причину, другие – следствие. ***Первые*** принято в статистике называть признаками-факторами или факторными признаками, а ***вторые*** – результативными признаками.

Многообразие связей, в которых находятся явления, порождает необходимость их классификации, в сведении связей к определенным типам, формам по их существенным чертам и свойствам. В основе наиболее общей классификации связей в статистике рассматривают различия и сходства связей по таким их особенностям, как степень тесноты, направление, аналитическое выражение, единичность или множественность факторов и следствий (результатов). В соответствии с этим, различают следующие виды связи (рис. 8.1).

Виды связей

По типам связи	**По направлению связи**	**По формам зависимости**	**По числу факторных признаков**
• функциональные • корреляционные	• прямые • обратные	• прямолинейные • криволинейные	• однофакторные •многофакторные

Рис. 8.1. **Классификация связей**

Функциональной называют такую связь, при которой значения результативного признака (величина его вариантов) целиком и полностью определяются значениями признаков-факторов (аргументов), так что определенной системе значений факторных признаков всегда соответствует строго определенное одно или несколько (в случае многозначной функции) значений результативного признака.

Функциональная связь является жестко детерминированной, проявляется в каждом отдельном случае. Функциональная связь – полная связь, когда результативный признак (следствие) всецело зависит от факторного признака (причины) и не зависит от окружающих условий.

То есть, здесь нет места отклоняющему действию случайных причин. Если варианты факторного признака обозначить через **x**, а результативного через **y** то функциональную связь между ними можно выразить, например, в виде следующего уравнения $y = x^2$. Это может быть, например, связь между площадью квадрата и его стороной.

Существуют функциональные связи и иного рода, в которых результативный признак является функцией нескольких факторных признаков, каждый из которых в той или иной мере влияет на формирование результативного признака, причём степень влияния известна. Примером может служить связь между площадью треугольника, его основанием и высотой. Значит, установив на основе единичного исследования эту зависимость, мы можем использовать его в любых аналогичных случаях.

При функциональной связи изменение результативного признакака **y** всецело обусловлено действием факторного признака **x**:

$$y = f(x).$$

Примером функциональной связи является зависимость длины окружности от радиуса *(r)*:

$$l = 2\pi r.$$

Такая форма связи дает возможность перечислить взаимодействующие факторы и установить определенную их пропорциональность. Например, если грузоподъемность автомобиля, используемого на перевозке зерна к элеватору, равна 2,5 т, то общее количество перевезенного зерна находится в функциональной зависимости от числа рейсов.

Но в экономической практике функциональные связи практически не встречаются. Такой «чистый» вид связи между признаками обычно используется в так называемых «точных» науках (физике, математике).

Статистическая (стохастическая) *связь* – связь, при которой изменение одной переменной приводит к изменению распределения другой переменной. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой. В этом случае речь идет о корреляционной зависимости.

Корреляционная связь

При корреляционной связи изменение результативного признака **y** обусловлено влиянием факторного признака **x** не всецело, а лишь частично, так как возможно влияние прочих факторов ε

$$y = \psi(x) + e.$$

Корреляционная связь[1] – связь, при которой определенному значению факторного признака соответствует среднее значение результативного признака. Она выражается в том, что с изменением

[1] Основоположниками теории корреляции считаются английские биометрики Ф. Гальтон (1822-1911) и К. Пирсон (1857-1936). Само слово корреляция ввел в употребление в статистику английский биолог и статистик Френсис Гальтон в конце XIX в. Тогда оно писалось как «corelation» (соответствие), но не просто «связь» (relation), а «как бы связь», т.е. связь, но не в привычной в то время функциональной форме. В науке вообще, а именно, в палеонтологии термин «корреляция» применил еще раньше, в конце XVIII в., знаменитый французский палеонтолог (специалист по ископаемым останкам животных и растений прошлых эпох) Жорж Кювье. Он ввел даже «закон корреляции» частей и органов животных. «Закон корреляции» помогает восстановить по найденным в раскопках черепу, костям и т.д. облик всего животного и его место в системе: если череп с рогами, то это было травоядное животное, а его конечности имели копыта; если же лапа с когтями – то хищное животное без рогов, но с крупными клыками.

значения факторного признака меняется средняя величина результативного. Здесь между значениями признаков нет строгого соответствия в каждом отдельном случае.

Корреляционные связи проявляются, как правило, при большом числе наблюдений. Степень, полнота проявления этих связей и число случаев наблюдения, необходимое для их проявления, зависят от многих причин и, прежде всего, от степени тесноты этой связи. Чем теснее связь, чем ближе она к функциональной, тем яснее и точнее она проявляется, и тем меньше число наблюдений, необходимое для ее выявления.

В качестве примера корреляционной связи можно назвать связь между себестоимостью продукции и производительностью труда рабочих, между размерами предприятия по объему их основных производственных фондов и объемом вырабатываемой продукции, между квалификацией рабочего и его заработком. Здесь, при одном и том же значении учтенного факторного признака возможны различные значения результативного признака. Это обусловлено наличием других факторов, которые могут быть различными по составу, направлению и силе действия на отдельные (индивидуальные) единицы статистической совокупности. Поэтому для изучаемой статистической совокупности в целом здесь устанавливается такое соотношение, в котором определенному изменению факторного признака соответствует среднее изменение признака результативного. Следовательно, характерной особенностью корреляционных связей является то, что они проявляются не в единичных случаях, а в массе. Поэтому изучаются корреляционные связи по так называемым эмпирическим данным, полученным в статистическом наблюдении. В таких данных отображается совокупное действие всех причин и условий на изучаемый показатель.

При статистическом изучении корреляционной связи определяется влияние учтенных факторных признаков при отвлечении (абстрагировании) от прочих аргументов. Применяемый таким образом способ научной абстракции хотя и ведет к некоторому упрощению (аппроксимации) реального механизма связи, но делает возможным установление закономерностей взаимодействия изучаемых показателей, что позволяет, не прибегая к экспериментированию, получать количественные характеристики корреляционной связи.

При изучении корреляционной связи перед статистикой ставятся следующие основные задачи:

➢ проверка положений экономической теории о возможности связи между изучаемыми показателями и придание выявленной связи аналитической формы зависимости;

➢ установление количественных оценок тесноте связи, характеризующих силу влияния факторных признаков на результативные.

Связь прямая и обратная

По направлению, как сказано выше, обычно различают два вида связи: прямую и обратную. Прямой называют такую связь, при которой с увеличением или уменьшением значений факторного признака, соответственно, увеличиваются или уменьшаются и средние значения результативного признака. В приведенных примерах корреляционной связи две из них относятся к прямой. Это – связь между размерами предприятия по объему основных производственных фондов и объемом их продукции и связь между квалификацией рабочего и его заработками.

Чем крупнее предприятие по величине основных производственных фондов, тем больше, при прочих равных условиях, оно производит продукции, и чем выше квалификация рабочего (чем больше его тарифный разряд), тем выше, опять-таки при прочих равных условиях, его заработок.

Обратной называют такую связь, при которой значения результативного признака изменяются в противоположном направлении по отношению к изменениям значений факторного признака. При увеличении значений факторного признака значения результативного признака уменьшаются и наоборот.

Следовательно, по направлению связи различают:

✓ прямую регрессию, когда с увеличением факторного признака результативный признак увеличивается (и наоборот);

✓ обратную регрессию, когда с увеличением факторного признака результативный признак уменьшается (и наоборот).

Примером такой формы связи может служить упомянутая выше, связь между производительностью труда рабочих и себестоимостью единицы продукции. Чем выше производительность труда рабочих, тем ниже, при прочих равных условиях, себестоимость единицы продукции.

Необходимо иметь в виду, что прямая и обратная связи могут переходить одна в другую. Так, прямая связь по достижении факторным признаком определенной величины в некоторых случаях исчерпывает себя и затем становится обратной, т.е. с дальнейшим увеличением значений факторного признака, значения признака результативного уменьшаются. Такой характер имеет, например, связь между возрастом рабочего и его производительностью, между количеством удобрений, вносимых на единицу площади пашни, и величиной урожайности, получаемой с нее. Рассматривая производительность труда рабочего в отдельные периоды работы в течение рабочего дня, не трудно заметить, что в начале она не так высока, но постепенно повышается, однако к концу рабочего дня начинает сказываться усталость и производительность труда постепенно понижается.

Связи прямолинейные и криволинейные

По аналитическому выражению в общем виде обычно выделяют связи прямолинейные и криволинейные. Если данная связь может быть выражена точно или приближенно уравнением какой-либо прямой линии, ее называют прямолинейной, а если уравнением какой-либо кривой, то она называется криволинейной связью.

Аналитическими уравнениями точно могут быть выражены только функциональные связи, связи же корреляционные могут быть выражены ими лишь приближенно. Приближенное выражение при помощи аналитических уравнений корреляционных связей общественных явлений во многих случаях дает возможность получить результаты, вполне пригодные для научных и практических целей. Этим и объясняется широкое применение в статистике корреляционного метода исследования связей, основанного на аналитическом их выражении.

Таким образом, по форме зависимости различают:

- линейную регрессию, которая выражается уравнением прямой:

$$\hat{y} = a + bx;$$

- нелинейную регрессию, которая выражается уравнениями вида:

$$\hat{y} = a + bx + cx^2 \text{ – для параболы;}$$

$$\hat{y} = a + \frac{b}{x} \text{ – для гиперболы и др.}$$

Связи однофакторные и многофакторные

Изменение среднего значения результативного признака в зависимости только от одного фактора называют парной корреляцией, а в зависимости от нескольких факторов – множественной (однофакторной или многофакторной). Несмотря на то, что в массовых общественных явлениях действует множество факторов, можно исследовать зависимость результативного признака только от одного или нескольких факторов, абстрагируясь или, как говорят в статистике, элиминируя действие на него всех других факторов, называемых при этом обычно прочими факторами.

Таким образом, по числу факторных признаков различают:

- парную корреляцию, когда один факторный признак;
- множественную корреляцию, когда много факторных признаков.

Общий вид уравнений парной регрессии:

$$\hat{y} = f(x).$$

Общий вид многофакторных уравнений регрессии:

$$\hat{y} = f(x_1 x_2, \ldots, x_n).$$

8.2. Методы выявления корреляционной связи

На первом этапе статистического изучения взаимосвязей социально-экономических явлений определяют наличие связи и ее направление (если связь существует) при помощи следующих элементарных методов.

Важнейшие приемы изучения взаимосвязи общественных явлений

В установлении взаимосвязи между признаками общественных явлений определяющая роль принадлежит теоретическому анализу. Так, теоретически бесспорно, например, что уровень себестоимости продукции зависит от производительности труда, урожайности – от качества обработки почвы, количества внесенных удобрений и т.д.

Теоретически установленные зависимости одних признаков от других могут иметь различное количественное выражение в конкретных условиях места и времени.

Эти выражения взаимосвязи, имеющие большое практическое значение, могут быть статистически охарактеризованы при помощи различных методов, главнейшими из которых являются: метод приведения параллельных данных, аналитические группировки, балансовый метод. Важное место в изучении взаимосвязи общественных явлений занимает корреляционный анализ.

Метод приведения параллельных данных

Наиболее простым приемом статистической характеристики взаимосвязи является приведение в табличной форме рядов статистических данных, отображающих развитие взаимосвязанных между собой явлений. Сопоставление параллельных рядов – метод, когда ряд значений факторного признака (x), построенный в порядке возрастания, сопоставляют с рядом соответствующих значений результативного признака (y) и таким образом прослеживают их взаимосвязь (табл. 8.1).

Метод аналитических группировок

Для выявления и характеристики взаимосвязи между признаками общественных явлений широко и плодотворно используется метод аналитических группировок.

Этот метод позволяет не только констатировать связь между признаками, но и выявить причины, приведшие к тем или иным конкретным результатам.

Таблица 8.1

Сопоставление радов численности менеджеров и объемов продаж однотипных фирм в одном из регионов РУз в I кв. исследуемого года

Номер фирмы	Численность менеджеров, чел. x_i	Объем продаж, млн. сум y_i
1	15	9,50
15	17	9,54
21	8	9,58 '
7	20	10,01
5	20	10,07
26	20	9,77
11	22	9,88
30	22	9,79
2	24	10,30
4	25	10,40
8	25	10,00
16	25	10,22
6	27	10,36
10	27	9,90
25	27	10,26
29	28	10,30
9	29	10,30
19	30	10,29
13	32	10,48
18	32	10,71
17	33	10,84
12	33	10,61
23	33	10,82
28	34	10,50
14	35	10,56
20	35	10,67
27	38	10,74
3	39	10,98
24	39	10,93
22	45	11,00

Данные табл. 8.1 свидетельствуют о наличии прямой корреляционной связи между численностью менеджеров и объемом продаж фирм. По мере возрастания количества менеджеров, увеличивается и объем продаж.

Таблица 8.2

Группировка туристических фирм по затратам на рекламу

Группы туристических фирм по затратам на рекламу, усл. ден. ед.	**Число фирм в группе**	**Среднее число туристов, воспользовавшихся услугами данной группы фирм, человек**
1	2	3
8	3	790
9	5	860
10	5	966
11	4	1063
12	3	1100
Итого	20	

При аналитических группировках обычно в основу образования групп кладется факторный признак. Рассчитав затем для каждой группы обобщающие показатели результативного признака в виде средних или относительных величин, прослеживают, какое влияние оказывает изменение фактического признака на результативный. Так, например, данные приведенные в табл. 8.2 разбиваются на группы в зависимости от величины признака-фактора, и по каждой группе вычисляются средние значения результативного признака.

Примечание. Различие в величине среднего числа туристов каждой группы фирм, вычисленных в корреляционной (графа 8 табл. 8.4) и групповой таблицах (графа 3 табл. 8.2) объясняется тем, что при расчете средних в корреляционной таблице действительные значения результативного признака заменяются центральными значениями интервалов группировки.

Сравнив средние значения результативного признака по группам, можно сделать вывод, что рост затрат туристических фирм на рекламу влечет за собой увеличение числа клиентов, пользующихся услугами фирмы, т.е. в рассматриваемом примере можно предположить наличие прямой корреляционной зависимости между признаками. Корреляционная зависимость отчетливо обнаруживается только при рассмотрении средних значений результативного признака, соответствующих определенным значениям факторного признака, так как при достаточно большом числе наблюдений в каждой группе влияние прочих случайных факторов при расчете групповой средней будет взаимопогашаться, и четче выступит зависимость.

Балансовый метод

Взаимосвязь между многими статистическими показателями вполне может быть охарактеризована при помощи построения балансов в виде соотношений между наличием и распределением тех или иных ресурсов. Например, балансовая связь показателей коммерческой деятельности характеризует зависимость между источниками формирования ресурсов (средств) и их использованием. Свое проявление она получает, например, в формуле товарного баланса:

$$О_н + П = В + О_к,$$

где $О_н$ - остаток товаров на начало изучаемого периода; П - поступление товаров за период; В - выбытие товаров в изучаемом периоде; $О_н$ - остаток товаров на конец периода.

Левая часть формулы баланса характеризует предложение товаров ($О_к$ + П), а правая часть – использование товарных ресурсов (В + $О_н$). Важное практическое значение формулы товарного баланса состоит в том, что при отсутствии количественного учета продажи товаров на основе формулы баланса определяют величину розничной реализации отдельных товаров.

Компонентный метод

Компонентные связи характеризуются тем, что изменение статистического показателя определяется изменением компонентов, входящих в этот показатель, как множители:

$$a = b \cdot c.$$

Например, в статистике коммерческой деятельности компонентные связи используются в индексном методе выявления роли отдельных факторов в совокупном изменении сложного показателя. Индекс товарооборота в фактических ценах I_{qp} представляет произведение двух компонентов – индекса товарооборота в сопоставимых ценах I_q и индекса цен I_p, т.е.

$$I_{qp} = I_q \cdot I_p.$$

Важная практическая значимость показателей, состоящих в компонентной связи, в том, что она позволяет определять величину одного из неизвестных компонентов: $I_q = I_{qp} : I_p$ или $I_p = I_{qp} : I_q$.

Графический метод

Данный метод позволяет выявить наличие и направление связи между двумя признаками корреляции.

Используя данные об индивидуальных значениях признака факторных соответствующих ему значениях результативного признака можно построить в прямоугольных координатах точечный график, который называют «полем корреляции». Для нашего примера «поле корреляции» имеет следующий вид (рис. 8.2).

Положение каждой точки на графике определяется величиной двух признаков – уровнем затрат на рекламу и соответствующим ему числом туристов, пользующихся услугами данной фирмы. Точки корреляционного поля не лежат на одной линии, они вытянуты определенной полосой слева направо. Имеющийся в нашем распоряжении статистический материал был сгруппирован (табл. 8.2) и по каждому значению затрат фирм на рекламу определены значения среднего числа туристов в группе (см. гр. 3 табл. 8.2). Нанеся эти средние на график и соединяя последовательно отрезками прямых соответствующие им точки, получим так называемую эмпирическую линию связи.

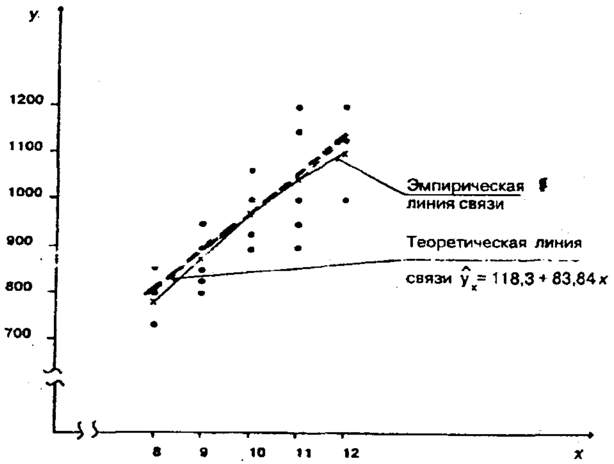


Рис. 8.2. **Зависимость числа клиентов фирмы (у) от ее затрат на рекламу** *(x)*

Если эмпирическая линия связи по своему виду приближается к прямой линии, то можно предположить наличие прямолинейной корреляционной связи между признаками. Если же имеется тенденция неравномерного изменения значений результативного признака, и эмпирическая линия связи будет приближаться к какой-либо кривой, то это может быть связано с наличием криволинейной корреляционной связи (см. гл. 3).

Метод корреляционной таблицы

Наличие большого числа различных значений результативного признака (у), сооответствующих одному и тому же значению признака и фактора (х) затрудняет восприятие таких параллельных рядов особенно при большом числе единиц, составляющих изучаемую совокупность.

В таких случаях целесообразнее воспользоваться для установления факта наличия связи статистическими таблицами – корреляционными или групповыми. Например, по 20 туристическим фирмам были установлены затраты на рекламу (факторный признак) и количество туристов воспользовавшихся услугами каждой фирмы (результативный признак).[2] В табл. 8.2 фирмы ранжированы по величине затрат на рекламу.

Можно увидеть, что в целом для всей совокупности фирм увеличение затрат на рекламу приводит к увеличению количества туристов, пользующихся услугами фирмы, хотя в отдельных случаях наличие такой зависимости может и не усматриваться. Например, сопоставим данные по фирмам с порядковыми номерами 7 и 11. Здесь мы видим даже обратное соотношение: у фирмы 11 количество туристов меньше, чем у фирмы 7 и составляет 920 человек, хотя затраты на рекламу выше, чем у фирмы 7 на 1 усл. ден. ед. В каждом отдельном случае количество туристов, воспользовавшихся услугами фирмы будет зависеть не только от размера затрат фирмы на рекламу, но и от того, как сложатся прочие факторы, определяющие величину результативного признака.

Построение корреляционной таблицы начинают с группировки значений факторного и результативного признаков. Так как в приводимом примере факторный признак представлен всего пятью вариантами повторяющихся значений, достаточно в первом столбце табл. 8.2 выписать эти результаты.

---

[2] Поскольку речь идет об изложении методологии изучения взаимосвязей, мы ограничились совокупностью малого объема. Данные в примере условные.

Таблица 8.3

Показатели, характеризующие связь между затратами на рекламу и количество туристов, воспользовавшихся услугами фирмы

Порядковые номера фирм	**Затраты на рекламу, усл. ден. ед.**	**Количество туристов, воспользо-вавшихся услугами фирмы, человек**	**Порядковые номера фирм**	**Затраты на рекламу, усл. ден. ед.**	**Количество туристов, воспользо-вавшихся услугами фирмы, чел.**
1	8	800	11	10	920
2	8	850	12	10	1060
3	8	720	13	10	950
4	9	850	14	11	900
5	9	800	15	11	1200
6	9	880	16	11	1150
7	9	950	17	11	1000
8	9	820	18	12	1200
9	10	900	19	12	1100
10	10	1000	20	12	1000

Для результативного признака необходимо определить величину интервала. Для этого воспользуемся формулой Стэрджесса:

$$h_y = \frac{y_{\max} - y_{\min}}{1 + 3{,}322 \lg n} = \frac{1200 - 720}{5} = 96 \text{ человек.}$$

В корреляционной таблице факторный признак **х**, как правило, располагают в строках, а результативный признак **у** – в столбцах (графах) таблицы. Числа, расположенные на пересечении строк и столбцов таблицы, означают частоту повторения данного сочетания значения х и у (табл. 8.4).

Данная корреляционная таблица уже при общем знакомстве дает возможность выдвинуть предположение о наличии или отсутствии связи, а также выяснить ее направление. Если частоты в корреляционной таблице расположены на диагонали из левого верхнего угла в правый нижний угол (т.е. большим значениям фактора соответствуют большие значения функции), то можно предположить наличие прямой корреляционной зависимости между признаками. Если же частоты расположены по диагонали справа налево, то предполагают наличие обратной связи между признаками.

Уместно подчеркнуть, что при рассмотрении корреляционной таблицы важно установить расположение основной части частот. Возможны варианты, когда все клетки корреляционной таблицы окажутся заполненными. Однако это обстоятельство еще не означает, что корреляционная связь между признаками отсутствует. Нужно установить, как расположена в таблице основная масса частот. Для того чтобы сделать восприятие корреляционной таблицы более доступным и в целях более четкого выявления основной тенденции связи, можно для каждой строки рассчитать средние значения результативного признака, соответствующие определенному значению признака-фактора (каждая строка таблицы дает условное распределение y при определенном значении x).

Таблица 8.4

Корреляционная таблица

Центральное значение интервала, y	**768**	**865**	**962**	**1059**	**1156**	f_x	$\bar{y}_j$
группы по y / группы по x	**720-816**	**817-913**	**914-1010**	**1011-1107**	**1108-1207**		
1	2	3	4	5	6	7	8
8	2	1				3	800
9	1	3	1			5	865
10		1	3	1		5	962
11		1	1		2	4	1035
12			1	1	1	3	1059
f_y	3	6	6	2	3	20	

Примечание: $\bar{y}_j$ – среднее значение результативного признака для j-той группы значений факторного признака; f_x – частота повторения данного варианта значения факторного признака во всей совокупности; f_y – частота повторения результативного признака во всей совокупности.

Среднее число туристов для первой группы, состоящей из трех фирм, которые тратят на рекламу 8 усл. ден. ед., будет равно 800:

$$\bar{y}_1 = \frac{768 \cdot 2 + 865 \cdot 1}{3} = 800 \text{ человек.}$$

Для следующей группы, состоящей из пяти фирм, у которых затраты на рекламу 9 усл. ден. ед. $\bar{y}_2 = \dfrac{768 \cdot 1 + 865 \cdot 3 + 962 \cdot 1}{5} = 865$ человек и т.д. (рассчитанные таким образом средние представлены в графе 8 табл. 8.4).

Итак, увеличение средних значений результативного признака с увеличением значений факторного признака еще раз свидетельствует о возможном наличии прямой корреляционной зависимости числа туристов, воспользовавшихся услугами фирмы, от затрат фирмы на рекламу.

Корреляционная таблица позволяет сжато, компактно изложить материал, поэтому все последующие расчеты (показателей тесноты связи и параметров уравнения регрессии) можно вести по корреляционной таблице.

Корреляционно-регрессионный метод

Корреляционный анализ – это количественное определение тесноты связи между факторным и результативными признаками, осуществляемой с помощью показателей корреляции.

Регрессионный анализ

Регрессионный анализ – определение теоретического выражения связи между признаками, т.е. определение формы связи (построение уравнения регрессии). Корреляционный и регрессионный виды анализа тесно взаимосвязаны между собой. Установив при помощи элементарных статистических методов наличие и направление связи между признаками, на следующем этапе анализа переходят к построению регрессионной модели зависимости результативного признака от факторного и оценке тесноты корреляционной связи, т.е. переходят к корреляционно-регрессионному анализу.

Более подробно эти два вида анализа будут рассмотрены ниже.

Метод корреляции рангов

Этот метод называется коэффициентом корреляции рангов Спирмена и рассчитывается на базе ранжированного ряда изучаемых единиц (объектов) по занимаемому месту. Формула расчета имеет следующий вид:

$$R = 1 - \frac{6(x-y)^2}{n(n^2-1)}.$$

Покажем расчет этого коэффициента на примере связи между ростом населения (х) и ростом валовой внутренней продукции на душу населения (у). табл. 8.5.

Таблица 8.5

Ранжированный ряд областей по удельному весу населения в общей численности

Ранжированный ряд		**Место областей по удельному весу в производстве ВВП республики**	**Разность мест**	**Квадрат разности мест**
областей по удельному весу населения в общей численности населения РУз	**место**			
(К)	(х)	(у)	(х–у)	$(x-y)^2$
Самаркандская	1	6	–5	25
Ферганская	2	4	–2	4
Кашкадарьинская	3	3	0	0
Андижанская	4	5	–1	1
Ташкентская	5	2	3	9
Наманганская	6	9	–3	9
г. Ташкент	7	1	6	36
Сурхандарьинская	8	10	–2	4
Бухарская	9	7	2	4
Республика КК	10	12	–2	4
Хорезмская	11	11	0	0
Джизакская	12	13	–1	1
Навоийская	13	8	5	25
Сырдарьинская	14	14	0	0
Итого	–	–		$\Sigma(x-y)^2=122$

Отсюда

$$R=1-\frac{6\Sigma(x-y)^2}{n(n^2-1)}=1-\frac{6}{14(14^2-1)}=1-\frac{732}{14\cdot 195}=$$

$$1-\frac{732}{2730}=1-0{,}268=0{,}732 \text{ или } 73{,}2\%.$$

Следовательно, связь между численностью и выпуском ВВП существует высокая связь, о чем свидетельствует шкалы Чеддока (см. табл. 8.9).

Следует отметить, что если последовательность рангов обеих переменных совпадают, то коэффициент корреляции рангов Спирмена равняется 1. Это наблюдается в Кашкадарьинской, Хорезмской и Сырдарьинской областях. Так как в этом случае:

$$x_i - y_i = 0$$

для каждого значения:

$$i = 1, 2, 3, ..., n,$$

и следовательно, числитель дроби в правой части формулы равняется нулю, поэтому

$$R = 1.$$

Можно также показать, что коэффициент корреляции рангов R = –1, если последовательность рангов, соответствующих переменным x и y, имеют противоположные тренды (один возрастает от 1 до n, а второй понижается от n до 1).

$$-1 \leq R \leq 1$$

и, следовательно, R – это, как и обычный коэффициент корреляции r – удобная мера связи признаков.

Коэффициент Спирмена можно применять и в тех случаях когда изучаемым признакам нельзя придать количественного выражения, однако есть возможность осмысленного упорядочения эмпирических наблюдений над каждым из признаков.

8.3. Задачи корреляционно-регрессионного анализа

Корреляционный анализ решают две основные задачи:

Первая задача заключается в определении формы связи, т.е. в установлении математической формы, в которой выражается данная связь. Это очень важно, так как от правильного выбора формы связи зависит конечный результат изучения взаимосвязи между признаками.

Вторая задача состоит в измерении тесноты, т.е. меры связи между признаками с целью установить степень влияния данного фактора на результат. Она решается математически, путем определения параметров корреляционного уравнения.

Затем проводятся оценка и анализ полученных результатов при помощи специальных показателей корреляционного метода (коэффициентов детерминации, линейной и множественной корреляции и т.д.), а также проверка существенности связи между изучаемыми признаками.

Выбор формы связи

Определяющая роль в выборе формы связи между явлениями принадлежит теоретическому анализу. Так, например, чем больше размер основного капитала предприятия (факторный признак), тем больше при прочих равных условиях оно выпускает продукции (результативный признак). С ростом факторного признака здесь, как правило, равномерно растет и результативный, поэтому зависимость между ними может быть выражена уравнением прямой $Y = a_0 + a_1x$, которое называется линейным уравнением регрессии.

Параметр a_1 называется коэффициентом регрессии и показывает, насколько в среднем отклоняется величина результативного признака **у** при отклонении величины факторного признака *х* на одну единицу. При $x = 0$ $a_0 = Y$. Увеличение количества внесенных удобрений приводит, при прочих равных условиях, к росту урожайности, но чрезмерное внесение их без изменения других элементов к дальнейшему повышению урожайности не приводит, а, наоборот, снижает ее. Такая зависимость может быть выражена уравнением параболы $Y = a_0 + a_1x + a_2x^2$.

Параметр $\mathbf{a_2}$ характеризует степень ускорения или замедления кривизны параболы, и при $a_2 > 0$ парабола имеет минимум, а при $a_2 < 0$ – максимум. Параметр a_1 характеризует угол наклона кривой, а параметр $\mathbf{a_0}$ – начало кривой.

Однако с помощью теоретического анализа не всегда удается установить форму связи. В таких случаях приходится только предполагать о наличии определенной формы связи. Проверить эти предположения можно при помощи графического анализа, который используется для выбора формы связи между явлениями, хотя графический метод изучения связи применяется и самостоятельно.

Аналитическое выражение связи

Применение методов корреляционного анализа дает возможность выражать связь между признаками аналитически – в виде уравнения – и придавать ей количественное выражение. Рассмотрим применение приемов корреляционного анализа на конкретном примере.

Допустим, что между стоимостью основного капитала и выпуском продукции существует прямолинейная связь, которая выражается уравнением прямой $Y = a_0 + a_1x$. Необходимо найти параметры $\mathbf{a_0}$ и $\mathbf{a_1}$, что позволит определить теоретические значения Y для разных значений x. Причем $\mathbf{a_0}$ и $\mathbf{a_1}$ должны быть такими, чтобы было достигнуто максимальное приближение к первоначальным (эмпирическим) значениям у теоретических значений Y. Эта задача решается при помощи способа наименьших квадратов, основное условие которого сводится к определению параметров a_0 и a_1, таким образом, чтобы $\Sigma (y_i - Y)^2 = \min$[3]. Математически доказано, что условие минимума обеспечивается, если параметры $\mathbf{a_0}$ и $\mathbf{a_1}$ определяются при помощи

[3] Этот метод впервые разработан К.Ф. Гаусом (1777-1865).

системы двух нормальных уравнений, отвечающих требованию метода наименьших квадратов:

$$\Sigma y = na_0 + a_1\Sigma x_i;$$
$$\Sigma xy = a_0\Sigma x + a_1\Sigma x^2.$$

Первое уравнение есть сумма всех первоначальных уравнений. Второе получается умножением обеих частей уравнения прямой на один и тот же множитель. Математически доказано, что условие $\Sigma(y_i - Y)^2 = \min$ соблюдается, если в качестве такого множителя принять значение факторного признака, т.е. если уравнение прямой умножить на x.

Кроме рассмотренных функций связи в экономическом анализе часто применяются степенная, показательная и гиперболическая функции.

Степенная функция Это функция имеет вид $У = a_0x^{a_1}$. Параметр степенного уравнения называется показателем эластичности и указывает, на сколько процентов изменится **у** *при возрастании* **х** на 1%. При **х** = 1 $a_0 = Y$.

Для определения параметров степенной функции вначале ее приводят к линейному виду путем логарифмирования: $\lg y = \lg a_0 + a_1 \lg x_1$, а затем строят систему нормальных уравнений:

$$\Sigma\lg y = n\lg a_0 + a_1\Sigma\lg x;$$
$$\Sigma\lg y\lg x = \lg a_0\Sigma\lg x + a_1\Sigma\lg x^2.$$

Решив систему двух нормальных уравнений, находят логарифмы параметров логарифмической функции $\mathbf{a_0}$ и $\mathbf{a_r}$, а затем и сами параметры $\mathbf{a_0}$ и $\mathbf{a_1}$. При помощи степенной функции определяют, например, зависимость между фондом заработной платы и выпуском продукции, затратами труда и выпуском продукции и т.д.

Если факторный признак **х** растет в арифметической прогрессии, а результативный **у** – в геометрической, то такая зависимость выражается показательной функцией $Y = a_0a_1^x$. Для определения параметров показательной функции ее также вначале приводят к линейному виду путем логарифмирования: $\lg y = \lg a_0 + x \lg a_1$, а затем строят систему нормальных уравнений:

$$\Sigma\lg y = n\lg a_0 + \lg a_1\Sigma x;$$
$$\Sigma x\lg y = \lg a_0\Sigma x + \lg a_1\Sigma x^2.$$

Вычислив соответствующие данные и решив систему двух нормальных уравнений, находят параметры показательной функции $\mathbf{a_0}$ и $\mathbf{a_1}$.

В ряде случаев обратная связь между факторным и результативным признаками может быть выражена уравнением гиперболы:

$$Y = a_0 + \frac{a_1}{x}.$$

И здесь задача заключается в нахождении параметров a_0 и a_1 при помощи системы двух нормальных уравнений:

$$\Sigma y = na + a_1 \Sigma \frac{1}{x};$$

$$\Sigma y \frac{1}{x} = a_0 \Sigma \frac{1}{x} + a_1 \Sigma \frac{1}{x^2}.$$

При помощи гиперболической функции изучают, например, связь между выпуском продукции и себестоимостью, уровнем издержек обращения (в процентах к товарообороту) и товарооборотом в торговле, сроками уборки и урожайностью и т.д.

Таким образом, применение различных функций в качестве уравнения связи сводится к определению параметров уравнения по способу наименьших квадратов при помощи системы нормальных уравнений.

В малых совокупностях значение коэффициента регрессии подвержено случайным колебаниям. Поэтому возникает необходимость в определении достоверности коэффициента регрессии.

Достоверность коэффициента регрессии определяется так же, как и в выборочном наблюдении, т.е. устанавливаются средняя и предельная ошибки для выборочной средней и доли.

Средняя ошибка коэффициента регрессии определяется по формуле

$$\mu_B = \sqrt{\frac{\sigma_0^2}{\sigma^2(n-2)}},$$

где σ_0^2 - случайная дисперсия; σ^2 - общая дисперсия, а n - число коррелируемых пар.

8.4. Измерение тесноты связи

Для определения тесноты корреляционной связи при проверке адекватности регрессионной модели (т.е. соответствия эмпирическим данным) рассчитывают следующие показатели корреляции.

1. Линейный коэффициент корреляции помимо тесноты (или силы) связи показывает и ее направление; определяется по следующей формуле:

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}}.$$

Линейный коэффициент корреляции принимает значение:

$$-1 \le r \le +1$$

Со знаком (+) – прямая связь, со знаком (–) – обратная связь, при r = 0 – линейная связь отсутствует, при r = ±1 – связь функциональная (линейная). Чем ближе линейный коэффициент корреляции к ± 1, тем корреляционная связь теснее.

2. Линейный коэффициент детерминации (r^2) – квадрат линейного коэффициента корреляции. Его числовые значения всегда заключаются в пределах от нуля до единицы:

$$0 < r^2 < +1.$$

Линейный коэффициент детерминации – более жесткий показатель тесноты связи, чем линейный коэффициент корреляции.

3. Теоретическое корреляционное отношение рассчитывается по формуле:

$$\eta_T = \sqrt{\frac{\delta_T^2}{\sigma^2}},$$

где δ_T^2 - межгрупповая дисперсия выравненных значений результативного признака, т.е. рассчитанных по уравнению регрессии; σ^2 - общая дисперсия эмпирических значений результативного признака.

Общая дисперсия определяется по уже известной формуле

$$\sigma^2_{общ} = \frac{\Sigma(y - \bar{y})^2 f}{\Sigma f}.$$

Межгрупповая дисперсия выравненных значений результативного признака определяется по формуле:

$$\delta_T^2 = \frac{\sum_{j=1}^{k}(y_j - \bar{y})^2 f_j}{\sum_{j=1}^{k} f_j},$$

где $\hat{y}_j$ - теоретическое значение результативного признака в **j**- й группе.

При расчете теоретического корреляционного отношения можно использовать правило сложения дисперсий, которое в этом случае может быть представлено следующим образом:

$$\sigma^2 = \delta_T^2 + \sigma^2_{ост},$$

где $\sigma^2_{ост}$ - остаточная дисперсия.

Тогда теоретическое корреляционное отношение можно рассчитать по следующей формуле:

$$\eta_T = \sqrt{\frac{\sigma^2 - \sigma^2_{ост}}{\sigma^2}} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2}}.$$

Теоретическое корреляционное отношение, исчисляемое по этой формуле часто называют индексом (коэффициентом) корреляции (**R**).

Теоретическое корреляционное отношение $\eta_т$ является более универсальным показателем тесноты связи, чем линейный коэффициент корреляции r, поскольку может использоваться как для прямолинейных, так и криволинейных зависимостей.

Теоретическое корреляционное отношение не следует путать с эмпирическим корреляционным отношением, которое также используется в корреляционном анализе, но строится непосредственно по фактическим данным. Эмпирическое корреляционное отношение является показателем рассеяния точек корреляционного поля относительно эмпирической линии регрессии и закономерности, изменения которой нарушаются случайными зигзагами ломаной, возникающими вследствие действия неучтенных факторов. Теоретическое корреляционное отношение, в свою очередь, характеризует рассеяние точек корреляционного поля относительно теоретической (выравненной) линии регрессии.

4.Теоретический коэффициент детерминации определяется как квадрат теоретического корреляционного отношения:

$$\eta_T^2 = \frac{\delta_T^2}{\sigma^2}.$$

По аналогии с индексом (коэффициентом) корреляции теоретический коэффициент детерминации может обозначаться как R^2.

При прямолинейной связи коэффициент корреляции по своей абсолютной величине равен индексу корреляции: $|r| = R$.

Если индекс корреляции возвести в квадрат, то получим коэффициент детерминации

$$R^2 = \frac{\sigma_T^2}{\sigma^2}.$$

Он характеризует роль факторной вариации в общей вариации и по построению аналогичен корреляционному отношению η^2. Как и корреляционное отношение, коэффициент детерминации R^2 может быть исчислен при помощи дисперсионного анализа, так как дисперсионный анализ позволяет расчленить общую дисперсию на факторную и случайную. Однако при дисперсионном анализе для разложения дисперсии пользуются методом группировок, а при корреляционном анализе – корреляционными уравнениями.

Коэффициент детерминации является наиболее конкретным показателем, так как он отвечает на вопрос о том, какая доля в общем результате зависит от фактора, положенного в основание группировки.

При прямолинейной парной связи факторную дисперсию можно определить без вычисления теоретических значений Y по следующей формуле:

$$\sigma^2 = \frac{1}{n}(a_0\Sigma y + a_1\Sigma xy) - \bar{y}^2.$$

8.5. Проверка адекватности регрессионной модели

Корреляционные связи, как ранее указывалось, проявляется в массовом статистическом материале. Поэтому при проведении корреляционно-регрессионного анализа (особенно, если исследуемая совокупность образована на основе малой выборки) могут возникнуть сомнения в том, что обнаруженная связь носит закономерный, а не случайный характер. Таким образом, необходимо оценить адекватность (соответствия эмпирическим данным) полученной регрессионной модели.

Для малых выборок и парных регрессий наиболее известен метод оценки значимости корреляционных зависимостей на основе использования критерия, предложенного английским статистиком и названного его именем – критерия Р. Фишера (критерия дисперсионного отношения):

$$F_{расч} = \frac{\sigma^2_{факт}}{\sigma^2_{ост}},$$

где $\sigma^2_{факт}$ - факторная дисперсия; $\sigma^2_{ост}$ - остаточная дисперсия.

Факторная дисперсия определяется по следующей формуле:

$$\sigma^2_{факт} = \frac{\sum_{j=1}^{k}(y_j - \bar{y})^2 f_j}{k-1},$$

где (k – 1) - число степеней свободы для $\sigma^2_{факт}$.

Остаточную дисперсию, используя правило сложения дисперсий, можно определить по формуле:

$$\sigma^2_{ост} = \frac{\sum_{j=1}^{n}(y_j - \bar{y})^2 f_j - \sum_{j=1}^{k}(y_j - \bar{y})^2 f_j}{n-k},$$

где (п – k) - число степеней свободы для $\sigma^2_{ост}$.

Число степеней свободы для общей суммы квадратов отклонений:

$$(k - 1) + (n - k) = n - 1.$$

Если расчетное значение критерия Фишера больше табличного, то гипотеза о несоответствии заложенных в уравнение регрессии связей отвергается. Табличные значения F-критерия определяются на основании уровней значимости (0,05 и 0,01), а также по числу степеней свободы.

Поясним все сказанное примером.

Задание 1

На базе данных талб. 8.1 требуется построить:

1. Линейное уравнение парной регрессии, отражающее взаимосвязь между указанными признаками.

2. График теоретической линии зависимости объемов продаж от численности менеджеров фирм.

Решение

1. Для построения линейного уравнения регрессии необходимо определить параметры этого уравнения: свободный член уравнения a_0 и коэффициент регрессии a_1 для чего построена табл. 8.6.

Таблица 8.6

Вспомогательная таблица для определения параметров линейного уравнения регрессии

Исходные данные			Расчетные данные	
номера фирм	численность менеджеров, чел., x_i	объем продаж, млн. сум, y_i	x_i^2	$x_i y_i$
1	15	9,50	**225**	142,50
15	17	9,54	289	162,18
21	18	9,58	324	172,44
5	20	10,07	400	201,40
7	20	10,01	400	200,20
26	20	9,77	400	195,40
11	22	9,88	484	217,36
30	22	9,79	484	215,38
2	24	10,30	576	247,20
4	25	10,40	625	260,00
8	25.	10,00	625	250,00
16	25	10,22	625	255,50
6	27	10,36	729	279,72
10	27	9,90	729	267,30
25	27	10,26	729	277,02
29	28	10,30	784	288,40
9	29	10,30	841	298,70
19	30	10,29	900	308,70
13	32	10,48	1024	335,36
18	32	10,71	1024	342,72
12	33	10,61	1089	350,13

17	33	10,84	1089	357,72
23	33	10,82	1089	357,06
28	34	10,50	1156	357,00
14	35	10,56	1225	369,60
20	35	10,67	1225	373,45
27	38	10,74	1444	408,12
3	39	10,98	1521	428,22
24	39	10,93	1521	426,27
22	45	11,00	2025	495,00
Итого	849	309,31	25601	8840,05

Параметры линейного уравнения регрессии определим по формулам

$$a_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\frac{8840,05}{30} - \frac{849}{30} \cdot \frac{309,31}{30}}{\frac{25601}{30} - \left(\frac{849}{30}\right)^2} = \frac{294,67 - 28,3 \cdot 10,31}{853,37 - 800,89} = \frac{2,9}{52,48} = 0,055\,;$$

$$a_0 = \overline{y} - a_1\overline{x} = 10,31 - 0,055 \cdot 28,3 = 8,754.$$

Тогда линейное уравнение парной регрессии примет вид:

$$\hat{y} = 8,754 + 0,055x.$$

2. Для построения теоретической линии зависимости объемов продаж от численности менеджеров фирм необходимо знать теоретические значения результативного признака – объемов продаж фирм.

Их определяем, используя полученное линейное уравнение парной регрессии.

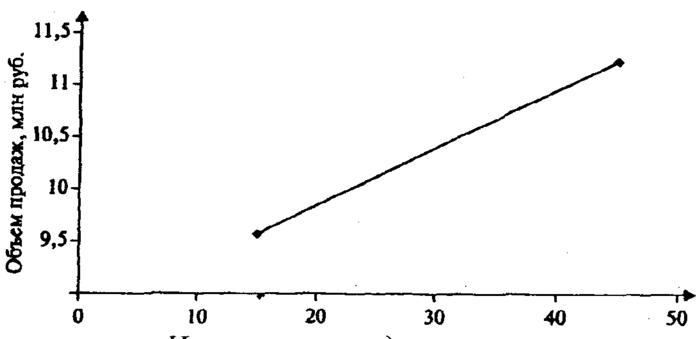


Рис. 8.3. **График теоретической линии зависимости объемов продаж от численности менеджеров фирм в одном из регионов РУз в I кв. исследуемого года**

Так как в нашем случае – линейная зависимость, то для того, чтобы получить искомые точки для построения теоретической линии зависимости, нам достаточно подставить в это уравнение два значения факторного признака (например, два крайних его значения – 15 и 45).

$$\hat{y}_{\min} = 8{,}754 + 0{,}055 \cdot 15 = 9{,}579 \text{ (млн. сум);}$$

$$\hat{y}_{\max} = 8{,}754 + 0{,}055 \cdot 45 = 11{,}229 \text{ (млн. сум).}$$

Задание 2

Используя полученные данные задания 1 сквозной задачи, требуется определить линейный коэффициент корреляции и сделать выводы о силе связи между численностью менеджеров и объемом продаж.

Решение

Прежде всего, построим вспомогательную табл. 8.7.

Используя полученные данные вспомогательной табл. 8.7, определим линейный коэффициент корреляции:

$$r = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}} = \frac{\dfrac{8840{,}05}{30} - \dfrac{849}{30} \cdot \dfrac{309{,}31}{30}}{\sqrt{\dfrac{25601}{30} - \left(\dfrac{849}{30}\right)^2}\sqrt{\dfrac{3194{,}59}{30} - \left(\dfrac{309{,}31}{30}\right)^2}} =$$

$$= \frac{294{,}67 - 28{,}3 \cdot 10{,}31}{\sqrt{853{,}37 - 800{,}89}\sqrt{106{,}49 - 10630}} = \frac{294{,}67 - 291{,}7}{\sqrt{52{,}48}\sqrt{0{,}19}} =$$

$$= \frac{2{,}90}{7{,}24 \cdot 0{,}44} = \frac{2{,}90}{3{,}19} = 0{,}91.$$

Таблица 8.7

Вспомогательная таблица для определения линейного коэффициента корреляции

Исходные данные			Расчетные данные		
номер фирмы	**численность менеджеров, чел., x_i**	**объем продаж, млн. сум, y_i**	x_i^2	y_i^2	$x_i y_i$
1	15	9,50	225	90,2500	142,50
15	17	9,54	289	91,0116	162,18
21	18	9,58	324	91,7764	172,44
5	20	10,07	400	101,4049	201,40
7	20	10,01	400	100,2001	200,20
26	20	9,77	400	95,4529	195,40
11	22	9,88	484	97,6144	217,36

30	22	9,79	484	95,8441	215,38
2	24	10,30	576	106,0900	247,20
4	25	10,40	625	108,1600	260,00
8	25	10,00	625	100,0000	250,00
16	25	10,22	625	104,4484	255,50
6	27	10,36	729	107,3296	279,72
10	27	9,90	729	98,0100	267,30
25	27	10,26	729	105,2676	277,02
29	28	10,30	784	106,0900	288,40
9	29	10,30	841	106,0900	298,70
19	30	10,29	900	105,8841	308,70
13	32	10,48	1024	109,8304	335,36
18	32	10,71	1024	114,7041	342,72
12	33	10,61	1089	112,5721	350,13
17	33	10,84	1089	117,5056	357,72
23	33	10,82	1089	117,0724	357,06
28	34	10,50	1156	110,2500	357,00
14	35	10,56	1221	111,5136	369,60
20	35	10,67	1225	113,8489	373,45
27	38	10,74	1444	115,3476	408,12
3	39	10,98	1521	120,5604	428,22
24	39	10,93	1521	119,4649	426,27
22	45	11,00	2025	121,0000	495,00
Итого	849	309,31	25601	3194,5941	8840,05

Таким образом, связь между объемом продаж и численностью менеджеров фирм прямая и близка к функциональной (линейной).

Сквозная задача

Задание 3

Используя расчетные данные заданий 1 и 2 сквозной задачи, требуется определить теоретическое корреляционное отношение и сделать выводы.

Таблица 8.8

Вспомогательная таблица для расчета межгрупповой дисперсии выравненных значений результативного признака (объема продаж)

Номер группы	**Число фирм, ед. f_{i0}**	**Центральное значение численности менеджеров в группе, чел., $x_{цj}$**	**Теоретическое значение объемов продаж фирмы, млн. сум, $\hat{y}_j$**	$(y_j - \bar{y})^2 f_j$, **где $\bar{y}$ = 10,31**
1	*3*	17,5	9,7165	1,0567

2	6	22,5	9,9915	0,6087
3	8	27,5	10,2665	0,0151
4	7	32,5	10,5415	0,3751
5	5	37,5	10,8165	1,2827
6	1	42,5	11,0915	0,6107
Итого	30	–	10,3100	3,9490

Решение

Теоретическое корреляционное отношение определяется по формуле:

$$\eta_T = \sqrt{\frac{\delta_T^2}{\sigma^2}},$$

где σ^2 - общая дисперсия результативного признака (объема продаж), полученная на базе эмпирического материала. Ее значение мы возьмем из решения $\sigma^2 = 0{,}1834$.

Для определения δ_T^2 построим вспомогательную табл. 8.8 на базе табл. 8.5, с учетом полученного линейного уравнения регрессии (см. решение задания 1).

$$\delta_T^2 = \frac{\sum_{j=1}^{k}(y_j - \bar{y})^2 f_i}{\sum_{j=1}^{k} f_i} = \frac{3{,}949}{30} = 0{,}1316;$$

$$\eta_T = \sqrt{\frac{0{,}1316}{0{,}1834}} = \sqrt{0{,}7176} = 0{,}8471.$$

На основании шкалы Чеддока можно сделать заключение о том, что связь между объемом продаж и численностью менеджеров фирм тесная (табл. 8.9).

Таблица 8.9

Шкалы Чеддока

Показания тесноты связи	0,1 -0,3	0,3–0,5	0,5–0,7	10,7–0,9	0,9–0,90
Характеристика силы связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

Сквозная задача

Задание 4

Используя полученные данные заданий 2 и 3 сквозной задачи, требуется дать оценку значимости корреляционной зависимости объемов продаж от численности менеджеров.

Решение

$$F_{расч} = \frac{\sigma^2_{факт}}{\sigma^2_{ост}} = \frac{\sum\limits_{j=1}^{k}(y_j - \bar{y})^2 f_j}{k-1} : \frac{\sum\limits_{j=1}^{k}(y_j - \bar{y})^2 - \sum\limits_{j=1}^{k}(y_j - \bar{y})^2 f_j}{n-k};$$

$$F_{расч} = \frac{3{,}9490}{6-1} : \frac{5{,}5036 - 3{,}9490}{30-6} = 0{,}7898 : 0{,}0648 = 12{,}19;$$

$F_{табл}$ = 4,53 и 9,47 (эти табличные значения F-критерия берутся из справочников по числу степеней свобод: 5 и 24 и для уровней значимости: 0,05 и 0,01).

Таким образом, $F_{расч} > F_{табл}$, т.е. гипотеза о несоответствии заложенных в уравнение, регрессии связей отвергается, или, другими словами, построенное корреляционное уравнение зависимости является значимым, и связь между численностью менеджеров и объемом продаж является существенной.

Если n > 30.

Применительно к совокупностям, у которых n < 30, для проверки типичности параметров уравнения регрессии используется t-критерий Стьюдента. При этом вычисляются фактические значения t-критерия:

для параметра a_0

$$t_{a_0} = a_0 \frac{\sqrt{n-2}}{\sigma_s};$$

для параметра a_1

$$t_{a_1} = a_1 \frac{\sqrt{n-2} \cdot \sigma_\chi}{\sigma_s},$$

где $\sigma_s = \sqrt{\frac{\Sigma(y_i - y_{xi})^2}{n}}$ – среднее квадратическое отклонение результативного признака y_i от выравненных значений y_{xi};

$\sigma_\chi = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}$ – среднее квадратическое отклонение факторного признака x_i от общей средней $\bar{x}$.

Полученные по вышеуказанным формулам фактические значения t_{a_0} и t_{a_1} сравниваются с критическим t_k, который получают по таблице Стьюдента с учётом принятого уровня значимости α и числа степеней свободы k.

Полученные в анализе корреляционной связи параметры уравнения регрессии признаются типичными, если $t_Ф$ фактическое больше $t_К$ критического:

$$t_{a_0} > t_k < t_{a_1}.$$

По проверенным на типичность параметрам уравнения регрессии производится синтезирование (построение) математической модели связи. При этом параметры примененной в анализе математической функции получают соответствующие количественные значения.

Смысловое содержание синтезированных таким образом моделей состоит в том, что они характеризуют среднюю величину результативного признака $\tilde{y}_x$ в зависимости от вариации признака-фактора x.

Важным этапом корреляционного анализа связи является оценка практической значимости синтезированных моделей. Смысл такой оценки состоит в том, чтобы обосновать применение метода функционального анализа при изучении корреляционной зависимости. Правомерность такого приема анализа будет оправданной лишь в тех случаях, если изучаемая корреляционная (соотносительная) связь не столь значительно отстоит от функциональной (жесткой) связи. При этом необходимо доказать, что применение метода функционального анализа при изучении корреляционной зависимости не дает существенных погрешностей.

Проверка практической значимости синтезированных в корреляционно-регрессионном анализе математических моделей осуществляется посредством показателей тесноты связи между признаками x и y (см. решение задач 1, 2, 3, 4).

8.6. Множественная корреляция

До сих пор мы рассматривали корреляционные связи между двумя признаками: результативным (у) и факторным (х). Например, выпуск продукции зависит не только от размера основного капитала, но и от уровня квалификации рабочих, состояния оборудования, обеспеченности и качества сырья и материалов, организации труда и т.д. В связи с этим возникает необходимость в изучении, измерении связи между результативным признаком, двумя и более факторными. Этим занимается множественная корреляция.

Построение уравнений множественной корреляции

Множественной называется такая корреляция, при которой результативный признак связан с несколькими факторными признаками. Например, зависимость между урожайностью и такими факторами, как качество почвы, сроки и способы посева, размеры внесенного удобрения, сроки уборки, зависимость дневной выработки бетонщиков от уровня механизации труда на формовке и одновременно квалификации рабочих и др.

В случае множественной корреляции, так же как и при парной, в соответствии с характером связи возможно построение как прямолинейных, так и криволинейных корреляционных уравнений. Например, выше было показано, что между дневной выработкой бетонщиков и уровнем механизации труда существует прямолинейная форма связи. Точно так же прямолинейная форма корреляционного уравнения может быть выведена и для связи между величиной дневной выработки и квалификацией рабочих. Исходя из этого, составим линейное уравнение множественной корреляции, выражающее зависимость величины дневной выработки бетонщиков (y) от уровня механизации труда на формовке (x) и квалификации рабочих (v):

$$\bar{y}_{xv} = a_0 + a_{1x} + a_{2v},$$

где a_0 - средняя дневная выработка неквалифицированного бетонщика на ручной формовке; a_x - среднее значение разницы между уровнем дневной выработки при выполнении работы механизированным и ручным способом и a_2 - среднее значение роста дневной выработки при повышении квалификации рабочих на один разряд.

Необходимо отметить, что это уравнение будет правильно выражать форму связи между рассматриваемыми признаками при условии независимости между любым из параметров и любым из факторных признаков либо такой слабой связи между параметрами и факторными признаками, которой можно пренебречь. Но для выражения формы связи не имеет значения, связаны ли отдельные факторные признаки между собой либо в такой связи находятся параметры.

Вышеуказанное уравнение содержит три параметра: $\mathbf{a_0}$, $\mathbf{a_1}$ и $\mathbf{a_2}$. Для их нахождения способом наименьших квадратов составляется и решается следующая система нормальных уравнений:

$$\left.\begin{aligned} a_0 n + a_1\Sigma x + a_2\Sigma v &= \Sigma y \\ a_0\Sigma x + a_1\Sigma x^2 + a_2\Sigma xv &= \Sigma xy \\ a_0\Sigma v + a_1\Sigma xv + a_2\Sigma v^2 &= \Sigma vy \end{aligned}\right\}.$$

В расчетной табл. 8.10 имеются все данные для составления нормальных уравнений множественной корреляции.

Таблица 8.10

Расчет сумм для определения коэффициентов системы нормальных уравнений по не сгруппированным данным

№ п/п	*x*	*v*	*y*	*xy*	*x2*	*vy*	v^2	*xv*	y^2
1	35	2	3,0	105,0	1 225	6,0	4	70	9,00

2	59	3	6,5	383,6	3 481	19,5	9	177	42,25
3	44	3	4,8	211,2	1 936	14,4	9	132	23,04
4	55	3	5,7	313,5	3 025	17,1	9	165	32,49
5	39	2	2,8	109,2	1 521	5,6	4	78	7,84
6	56	3	4,7	263,2	3 136	14,1	9	168	22,09
7	78	2	4,2	327,6	6 084	8,4	4	156	17,64
8	44	4	5,3	233,2	1 936	21 ,2	16	176	28,09
9	43	3	2,0	86,0	1 849	6,0	9	129	4,00
10	76	3	6,5	494,0	5 776	19,5	9	228	30,25
11	58	3	5,1	295,8	3 364	15,3	9	174	26,01 ‘
12	41	4	5,5	225,5	1 681	22,0	16	164	30,25
13	49	o	3,0	147,0	2 401	6,0	4	98	9,00
14	58	o	3,6	208,8	3 364	7,2	4	116	12,96
15	58	4	4,5	261,0	3 364	18,0	16	232	20,25
16	61	4	6,7	408,7	3 721	26,8	16	244	44,89
17	42	3	5,6	235,2	1 764	16,8	9	126	31,36
18	46	3	5,2	239,2	2 116	15,6	9	138	27,04
19	35	o	3,2	112,0	1 225	6,4	4	70	10,24
20	55	4	5,4	297,0	3 025	21,6	16	220	29,16
21	38	3	4,5	171,0	1 443	13,5	9	114	20,25
22	35	3	5,5	192,5	1 225	16,5	9	105	30,25
23	25	3	2,5	62,5	625	7,5	9	75	6,25
24	90	4	6,2	558,0	8 100	24,8	16	360	38,44
25	47	2	4,1	192,7	2 209	8,2	4	94	16,81
26	69	3	5,0	345,0	4 761	15,0	9	207	25,00
27	48	2	2,5	120,0	2 304	5,0	4	96	6,25
28	82	4	6,8	557,6	6 724	27,2	16	328	46,24
29	98	4	6,6	646,8	9 604	26,4	16	392	43,56
30	63	3	6,3	396,9	3 969	18,9	9	189	39,69
31	79	4	7,9	624,1	6 241	31,6	16	316	62,41
32	41	3	4,6	188,6	1 681	13,8	9	123	21,16
33	45	3	4,2	189,0	2 025	12,6	9	135	17,64
34	75	2	4,8	360,0	5 625	9,6	4	150	23,04
35	45	3	5,8	261,0	2 025	17,4	9	135	33,64
36	51	4	4,9	249,9	2 601	19,6	16	204	24,01
37	55	3	4,3	236,5	3 0.25	12,9	9	165	18,49
38	95	4	6,4	608,0	9 025	25,6	16	380	40,96
39	90	4	7,0	630,0	8 100	28,0	16	360	49,00
40	70	4	7,1	497,0	4 900	28,4	16	280	50,41
41	56	3	4,4	246,4	3136	13,2	9	168	19,36
42	57	3	5,1	290,7	3 249	15,3	9	171	26,01
43	48	2	5,0	240,0	2 304	10,0	4	96	25,00
44	72	3	6,0	439,2	5 184	18,3	9	216	37,21

45	52	3	5,9	306,8	2 704	17,7	9	156	34,81
46	33	2	3,8	125,4	1 089	7,6	4	66	14,44
47	55	3	4,6	253,0	3 025	13,8	9	165	21,16
48	30	2	3,4	102,0	900	6,8	4	60	11,56
49	67	2	5,5	368,5	4 489	11,0	4	134	30,25
50	57	3	5,9	336,3	3 249	17,7	9	171	34,81
Итого	2800	150	250	14 752	171 536	781,4	476	8 672	1 325,96

$50a_0 + 2\,800a_1 + 150a_2 = 250;$
$2800a_0 + 171\,536a_1 - 8\,672a_2 = 14\,752;$
$150a_0 + 8\,671a_1 + 476a_2 = 781,4.$

Совместным решением этих уравнений определяются параметры искомого уравнения: a_0 = 0,5; a_1 = 0,036 и a_2 = 0,835. Подставив значение найденных параметров в уравнение, получим

$$\bar{y}_{xv} = 0,5 + 0,036x + 0,835v.$$

Оно означает, что не имеющий квалификации рабочий при ручной формовке бетона вырабатывает за смену в среднем 0,5 $м^3$, возрастание уровня механизации на один процент вызывает рост дневной выработки в среднем на 0,036 $м^3$ и повышение квалификации рабочих на один разряд вызывает рост дневной выработки в среднем на 0,835 $м^3$.

Уравнение множественной корреляции дает более точные значения параметров, чем уравнения парной корреляции.

В частности, если в уравнении парной корреляции параметр a_1 показывает возрастание дневной выработки при повышении уровня механизации труда при прочих равных или средних условиях, то в уравнении множественной корреляции этот параметр показывает возрастание дневной выработки в связи с повышением уровня механизации труда при прочих равных условиях и постоянном значении уровня квалификации рабочих. Точно так же если в уравнении парной корреляции параметр a_2 показывает возрастание дневной выработки при повышении квалификации рабочих при прочих равных средних условиях, то в уравнении множественной корреляции этот параметр показывает возрастание дневной выработки в связи с повышением квалификации рабочих при прочих равных условиях и постоянном значении уровня механизации труда. Иными словами, уравнение множественной корреляции позволяет элиминировать не только влияние прочих факторов, которые не являются предметом исследования, но и влияние факторов, корреляционно связанных с *y*.

Оценка силы (тесноты) корреляционной связи

При статистическом изучении взаимосвязей между признаками общественных явлений важное значение имеет оценка тесноты связи, или, иными словами, оценка степени близости корреляционной связи к функциональной. Эта задача решается путем построения и исчисления коэффициентов корреляции.

При линейной форме множественной корреляции теснота связи между результативным *(y)* и двумя факторными *(x* и *v)* признаками измеряется с помощью показателя совокупного коэффициента корреляции:

$$R_{yxv} = \sqrt{\frac{r_{yx}^2 + r_{yv}^2 - 2r_{yx} \times r_{yv} \times r_{xv}}{1 - r_{xv}^2}},$$

где r_{yx}, r_{yv} и r_{xv} - линейные коэффициенты парной корреляции между соответствующими признаками.

Совокупный коэффициент корреляции показывает, какую часть общей колеблемости, у составляют колебания, вызванные исследуемыми факторами x и у.

Этот показатель характеризует степень связи результативного признака с двумя факторами в их совокупности и колеблется между 0 и 1. Когда он равен единице, то y связан с x и v точной линейной связью.

Когда же совокупный коэффициент корреляции равен 0, то у не может быть линейно связан с x и v, хотя возможна нелинейная корреляционная и даже функциональная связь. Следовательно, совокупный коэффициент корреляции R_{yxv} характеризует тесноту линейной связи y с x и v.

Определим, например, совокупный коэффициент множественной корреляции, характеризующий тесноту линейной связи между величиной дневной выработки бетонщиков *(y)* и уровнем механизации труда *(x)* и квалификацией рабочих *(y)* по данным табл. 8.10.

Для этого рассчитаем сначала средние значения для каждого признака и их сочетаний, а также средние квадратические отклонения:

$$\bar{x} = \frac{2800}{50} = 56; \ \bar{v} = \frac{150}{50} = 3; \ \bar{y} = \frac{250}{50} = 5;$$

$$\overline{xv} = \frac{8672}{50} = 173{,}44; \ \overline{xv} = \frac{14752}{50} = 295{,}04;$$

$$\overline{\upsilon y} = \frac{781{,}4}{50} = 15{,}63; \ \overline{x^2} = \frac{171536}{50} = 3430{,}72;$$

$$\overline{\upsilon^2} = \frac{476}{50} = 9{,}52; \ \overline{y^2} = \frac{1325{,}96}{50} = 26{,}52;$$

$$\overline{\sigma_x} = \sqrt[2]{\overline{x^2} - (\bar{x})^2} = \sqrt{3430{,}72 - 56^2} = \sqrt{3430{,}72 - 3136} = \sqrt{294{,}72} = 17{,}17;$$

$$\sigma_\upsilon = \sqrt{\overline{\upsilon^2} - (\bar{\upsilon})^2} = \sqrt{9{,}52 - 9} = \sqrt{0{,}52} = 0{,}72;$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{26{,}52 - 25} = \sqrt{1{,}52} = 1{,}23;$$

Далее найдем линейные коэффициенты корреляции между двумя признаками. Линейный коэффициент корреляции между величиной дневной выработки на формовке бетона (**у**) и уровнем механизации труда бетонщиков *(x)*:

$$r_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} = \frac{295{,}04 - 56 \times 5}{17{,}17 \times 1{,}23} = \frac{15{,}04}{21{,}12} = 0{,}712.$$

Линейный коэффициент корреляции между величиной дневной выработки и квалификацией бетонщиков (v) составляет:

$$r_{yv} = \frac{\overline{vy} - \bar{v} \cdot \bar{y}}{\sigma_v \cdot \sigma_y} = \frac{15{,}63 - 3 \times 5}{0{,}72 \times 1{,}23} = 0{,}708.$$

Наконец, линейный коэффициент корреляции между факторными признаками **х** и **v** равен:

$$r_{xv} = \frac{\overline{xv} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_v} = \frac{173{,}44 - 56 \times 3}{17{,}17 \times 0{,}72} = 0{,}44.$$

На основе произведенных расчетов определим совокупный линейный коэффициент множественной корреляции, измеряющий силу связи, между величиной дневной выработки бетонщиков, уровнем механизации труда на формовке и одновременно квалификацией рабочих:

$$R_{yxv} = \sqrt{\frac{0{,}712^2 + 0{,}708^2 - 2 \times 0{,}712 \times 0{,}708 \times 0{,}44}{1 - 0{,}44^2}} = 0{,}837.$$

Это свидетельствует о сильной степени линейной связи между рассматриваемыми признаками.

Наряду с совокупным коэффициентом множественной корреляции и в дополнение к нему при линейной множественной корреляции определяются частные коэффициенты корреляции между результативным признаком и каждым из факторных признаков при исключении влияния второго факторного признака. Эти показатели дают оценку тесноты связи между результативным признаком и одним из факторных признаков при элиминировании влияния другого факторного признака. Частный коэффициент корреляции между **у** и фактором **х** при элиминировании фактора **v** показывает, какую часть колеблемость **у**, вызванная фактором **х**, составляет в колеблемости **у**, вызванной всеми факторами, кроме ***v***. Так же как и парные коэффициенты корреляции, они могут принимать значения от -1 до $+1$.

Значение частного коэффициента корреляции, равное нулю, соответствует отсутствию линейной связи между результативным признаком и одним из факторных, хотя возможна нелинейная корреляционная и даже функциональная связь между ними. Значение частного коэффициента корреляции, равное единице, соответствует линейной функциональной связи между результативным и факторным признаками.

В рассматриваемом примере наряду с совокупным коэффициентом корреляции возможно исчисление двух частных коэффициентов корреляции: 1) частного коэффициента корреляции между величиной дневной выработки и уровнем механизации труда при исключении влияния квалификации рабочих и 2) частного коэффициента корреляции между величиной дневной выработки и квалификацией рабочих при исключении влияния уровня механизации труда.

Частные коэффициенты множественной корреляции при двух факторных признаках, влияние одного из которых исключается, определяются по следующим формулам:

$$\left.\begin{aligned} r_{yx}(\upsilon) &= \frac{r_{xy} - r_{x\upsilon} \times r_{\upsilon y}}{\sqrt{\left(1 - r_{x\upsilon}^2\right)\left(1 - r_{\upsilon y}^2\right)}}; \\ r_{y\upsilon}(x) &= \frac{r_{\upsilon y} - r_{x\upsilon} \times r_{xy}}{\sqrt{\left(1 - r_{x\upsilon}^2\right)\left(1 - r_{xy}^2\right)}}; \end{aligned}\right\}$$

где $\mathbf{r_{yv(x)}}$ - измеряет тесноту связи между **у** ***и*** **х** при исключении влияния **v**;_а $\mathbf{r_{yv\ (x)}}$ - между **у** и **v** при исключении влияния **х** (корень в этих формулах берется со знаком плюс).

Так, для рассматриваемого примера частный коэффициент линейной множественной корреляции между величиной дневной выработки и уровнем механизации труда при исключении влияния квалификации рабочих составляет:

$$r_{yx(\upsilon)} = \frac{0{,}712 - 0{,}44 \times 0{,}708}{\sqrt{\left(1 - 0{,}44^2\right)\left(1 - 0{,}708^2\right)}} = 0{,}635.$$

Частный коэффициент линейной множественной корреляции между величиной дневной выработки и квалификацией рабочих при исключении влияния уровня механизации труда на формовке бетона составляет:

$$r_{y\upsilon(x)} = \frac{0{,}708 - 0{,}44 \times 0{,}712}{\sqrt{\left(1 - 0{,}44^2\right)\left(1 - 0{,}712^2\right)}} = 0{,}625.$$

Если линейный коэффициент корреляции между величиной дневной выработки (у) и уровнем механизации труда на формовке бетона (х) равен 0,712, то частный коэффициент линейной множественной корреляции между этими показателями только 0,635; если линейный коэффициент корреляции между величиной дневной выработки и квалификацией рабочих *(v)* составляет 0,708, то соответствующий частный коэффициент линейной множественной корреляции только 0,625. Меньшая величина частного коэффициента линейной множественной корреляции по сравнению с линейным коэффициентом корреляции – результат исключения влияния второго факторного признака.

8.7. Измерение связей неколичественных переменных

Коэффициент ассоциации

Показателем силы связи двух явлений, выраженных качественными признаками, каждый из которых, в свою очередь, является альтернативным, служит коэффициент ассоциации.

Он выводится из четырехклеточной аналитической таблицы. Например, обследуют группу населения одного из регионов РУз в отчетном периоде.

I вопрос – о месте проживания (следует выбрать правильный ответ):

1. Проживаю в городе (*a*).
2. Проживаю в сельской местности (b).

II вопрос – о принадлежности к полу (следует выбрать правильный ответ):

1. Мужчина (c).
2. Женщина (d).

Представив суммарную численность ответов на каждый вопрос буквенными символами (*a*, *b*, **c**, d), покажем, как можно построить комбинационную четырехклеточную таблицу, которая может помочь нам в дальнейших расчетах (табл. 8.11)

Таблица 8.11

Взаимосвязь между ответами на два вопроса

Ответы на I вопрос	**Ответы на II вопрос**		**Итого**
	мужчина B	**женщина** $\overline{B}$	
Проживают в городе A	a	b	a + b
Проживают в сельской местности $\overline{A}$	c	d	c + d
Итого	a + c	b + d	a+b+c+d

Здесь A и B – признаки, связь между которыми анализируется; $\overline{A}$ и $\overline{B}$ обозначают противоположные признаки; a, b, c и d – частоты соответствующих комбинаций признаков.

На основании данных табл. 8.6 можно построить коэффициенты контингенции и ассоциации, предложенные английскими статистиками К. Пирсоном и Д. Юлом, соответственно. Коэффициент контингенции вычисляется по следующей формуле:

$$K_{конт} = \frac{ad - bc}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}}.$$

Коэффициент ассоциации

Коэффициент ассоциации рассчитывается по формуле

$$K_{асс} = \frac{ad - bc}{ad + bc}.$$

Свойства этих коэффициентов такие же, как линейного коэффициента корреляции.

Покажем расчет этих коэффициентов на примере аналитической табл. 8.12, характеризующей связь между уровнем образования и процентом выполнения нормы выработки.

Таблица 8.12

Аналитическая таблица

Группа рабочих по уровню образования	Выполнение нормы	Невыполнение нормы	Всего
Среднее специальное	78 (a)	22(b)	100 (a+b)
Среднее	32(c)	68(d)	100 (c+d)
Итого	110(a+c)	90(b+d)	200

$$K_{асс} = \frac{76 \cdot 68 - 32 \cdot 22}{78 \cdot 68 - 32 \cdot 22} = \frac{4600}{6008} = 0{,}766;$$

$$K_{конт} = \frac{76 \cdot 68 - 32 \cdot 22}{(78+22)(22+68)(78+32)(32+68)} = \frac{5304 - 704}{\sqrt{99000000}} = 0{,}46$$

Если $К_{асс} \geq 0{,}5$ и $К_{конт}$, наличие связи подтверждается.

Если ad > bc ($К_{асс} = 0$), то связь прямолинейная.

Если ad < bc, то это значит связь криволинейная.

Если ad = bc фв ($К_{асс} = 0$), то связь отсутствует.

Коэффициент взаимной сопряженности

Показателем силы связи двух атрибутивных признаков, каждый из которых имеет более двух вариантов, служит коэффициент взаимной сопряженности. Это одна из мер тесноты связи для качественных признаков. Предложены разные формулы расчета КВС англичанином К. Пирсоном и русским статистиком А.И.Чупровым.

КВС Пирсона находится по формуле:

$$КВС = \sqrt{\frac{\varphi^2}{1-\varphi^2}},$$

где φ^2 - показатель средней квадратической сопряженности. Эта величина получается путем вычитания единицы из суммы отношений квадратов частот каждой клетки корреляционной таблицы, к произведению итоговых частот соответствующего столбца и строки.

Таблица 8.13

Распределение фермерских хозяйств по степени удобрения посевов и урожайности

Степень удобрения посевов	Всего хозяйств	В том числе, получивших урожайность		
		низкую	**среднюю**	**высокую**
Слабая	20	8	1ы1	1
Средняя	60	11	43	6
Хорошая	20	1	6	13
Итого	100	20	60	20

КВС по формуле А.Н.Чупрова

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

где K_1 - число групп по столбцам статистической таблицы;

K_2 - число групп по строкам.

Покажем расчет КВС Чупрова на примере аналитической таблицы, характеризующей связь между размерами внесенного удобрения и величиной урожайности хлопка сырца (данные условные).

Для получения величины φ^2 составим вспомогательную табл. (8.14).

Таблица заполняется следующим образом. Сначала в верхнем углу каждой клетки записываются частоты. Затем в середине каждой клетки, кроме графы «Итого», указываются квадраты частот.

Таблица 8.14

Расчет показателей средней квадратической сопряженности

Урожайность / Степень удобрения	Низкая	Средняя	Высокая	Итого	
Слабая	8 64 3,2	11 121 2,02	1 1 0,05	20 – 5,57	0,264
Средняя	11	43	6	60	

	121	1 849	36	–	
	6,05	30,82	1,68	38,67	0,644
	1	6	13	20	
Хорошая	1	36	169	–	
	0,05	0,6	8,45	9,10	0,455
Итого	20	60	20	100	1,363

В третьей строке каждой клетки проставляются частные от деления квадратов частот на сумму частот по соответствующему столбцу: 3,2 как результат деления 64 на 20; 6,05 как результат деления 121 на 20 и т.д. Сумма этих чисел заносится в нижнюю часть клеток графы «Итого»: 3,2 +2,02+ 0,05 = 5,27 – в первой клетке и т.д. Наконец, частное от деления этих сумм на итог частот по каждой строке проставляется в последнем столбце таблицы: 5,27 : 20 = 0,264 по первой клетке; такой же расчет применяется и по второй и третьей клеткам. Сумма чисел последнего столбца без единицы и равна φ^2. Таким образом, φ^2= 1,363–1,0 = 0,363.

Подставляя найденное значение в формулу находим искомый коэффициент:

$$KBC = \sqrt{\frac{\varphi^2}{1-\varphi^2}} = \sqrt{\frac{0{,}363}{1{,}363}} = 0{,}51.$$

Коэффициент взаимной сопряженности характеризует наличие сопряженности двух данных признаков. Величина его варьирует от нуля до единицы. Чем больше сопряженность изучаемых признаков, тем его значение ближе к единице.

Из произведенного расчета видно, что величина коэффициента взаимной сопряженности зависит исключительно от величин частот рассматриваемых признаков и не связана со значениями самих признаков. Такой характер данного коэффициента делает его удобным при изучении связи между атрибутивными признаками.

Интеллектуальный тренинг

1.В чем сущность корреляционной связи между признаками?

2. Назовите основные методы выявления корреляционной связи.

3. В чем состоит основное содержание корреляционно-регрессионного анализа?

4. Как определяются параметры уравнения парной регрессии при линейной зависимости?

5. Что показывают линейный коэффициент корреляции и линейный коэффициент детерминации и как они рассчитываются?

6. Какие показатели используются для оценки силы корреляционной связи при криволинейной зависимости и как они рассчитываются?

7. В чем сущность метода оценки значимости корреляционной связи с использованием критерия Фишера?

8. С какой целью, и каким образом рассчитывают ошибку аппроксимации?

9. Какие вы знаете формы статистической связи?

10. Что такое уравнение регрессии?

11. Каковы предельные значения корреляционного отношения?

12. Что такое множественная корреляция?

13. Даст ли количественную оценку качества модели средняя ошибка аппроксимации?

14. Коэффициент корреляции r_{xy} применяется в тех случаях, когда между явлениями существует прямолинейная связь. Так ли это? Если нет, то докажите.

15. Что Вы понимаете под детерминированной зависимостью?

16. Помимо факторного признака (К) на результативный признак влияют и другие факторы, в том числе и неучтенные. Как они называются и обозначаются?

17. Каковы особенности измерения связей неколичественных переменных?

18. Как определяется коэффициент взаимной сопряженности А.И.Чупрова?

19. Как определяется коэффициент контингенции К.Пирсона? Что он характеризует?

20. Покажите порядок расчета коэффициента ассоциации?

21. Можно ли осуществлять прогноз на основе парной регрессионной модели? Если да, то как?

22. Когда признаются типичными полученные при анализе корреляционные связи параметры уравнения регрессии?

23. Для чего используется t-критерий Стьюдента?

24. Что означает $F_{расч.}$ и $F_{табл.}$?

25. В чем состоит отличие между корреляционной и функциональной связью?

26. Какие основные проблемы решает исследователь при изучении корреляционных зависимостей?

27. Какова роль групповых и корреляционных таблиц при анализе взаимосвязей?

28. Какие методы целесообразно использовать для выявления возможного наличия связи между факторным и результативным признаком при небольшом объеме фактических данных?

29. Какие показатели являются мерой тесноты связи между двумя признаками?

30. Как сцепить существенность линейного коэффициента корреляции?

31. Какие показатели используют для измерения степени тесноты связи между качественными признаками?

32. В чем состоит значение уравнения регрессии?

33. Что характеризует коэффициент регрессии?

34. Какая существует связь между линейным коэффициентом корреляции и коэффициентом регрессии?

35. Какое значение имеет расчет индекса корреляции?

36. Способы расчета средней квадратической ошибки уравнения, ее роль в оценке надежности уравнения регрессии.

37. Как осуществить прогноз значений результативного признака, опираясь на использование для этой цели уравнения регрессии?

38. Как измерить долю общей вариации результативного признака, которая объясняется влиянием вариации признака фактора x?

39. Если в случае линейной зависимости между признаками 60% вариации результативного признака объясняется влиянием факторного признака, чему будет равна величина коэффициента корреляции?

40. Как подходить к отбору факторов для включения их в уравнение множественной регрессии?

41. Каким образом можно выделить факторы, в изменении которых заложены наибольшие возможности в управлении изменением результативного признака?

42. Что означает величина коэффициента эластичности 0,58%.

43. Для чего рассчитываются частные коэффициенты корреляции?

44. В каких пределах заключена величина совокупного коэффициента корреляции и как она соотносится с величиной парных коэффициентов корреляции?

Тестовые задания по теме 8

1. Слово «корреляция» (от английского correlation) означает:

а) соотношение
б) соответствие
в) стохастическая, вероятная, возможная связь между двумя или несколькими случайными величинами
г) неправильных ответов нет

2. Само слово корреляция ввел в употребление в статистику английский биолог и статистик Френсис Гальтон:

а) в конце XVII века
б) в конце XVIII века
в) в конце XIX века
г) в середине XX века

3. Содержание теории корреляции составляет:

а) изучение зависимости вариации признака от окружающих условий
б) установление факта наличия связи, определение ее направления и формы
в) измерение степени тесноты связи между признаками и оценка адекватности модели
г) неправильных ответов нет

4. Между явлениями различают взаимосвязи:

а) функциональные (полная связь)
б) корреляционные (неполная связь)
в) стохастические (статистические)
г) а + б + в

5. Если определенному значению переменной X строго соответствует одно значение другой переменной Y и с изменением значения X значение Y меняется строго определенно, то такая связь между двумя переменными X и Y называется:

а) стохастической
б) функциональной (жестко детерминированной)
в) корреляционной
г) неправильных ответов нет

6. Связь, проявляющаяся при большом числе наблюдений в виде определенной зависимости между средним значением результативного признака и признаками - факторами, называется:
а) корреляционной
б) функциональной
в) детерминированной
г) б + в

7. Если рассматривается связь средней величины результативного показателя Y с одним признаком - фактором X, корреляция называется:
а) прямой
б) парной
в) множественной
г) обратной

8. По характеру изменений X и Y в парной корреляции различают:
а) обратную связь
б) множественную связь
в) прямую связь
г) а + в

9. В обратной зависимости значения факторного и результативного признаков:
а) изменяются в разных направлениях
б) изменяются в одном направлении
в) изменяются однозначно, т.е. с увеличением значений X увеличиваются и значения Y, с уменьшением значений факторного признака уменьшаются и значения результативного признака
г) б + в

10. Под общим термином «корреляционно-регрессионный анализ» подразумевают:
а) нахождение уравнений регрессии

б) измерение тесноты и направления связи
в) определение возможных ошибок, как параметров уравнений регрессии, так и показателей тесноты связи
г) а + б + в

11. Под корреляционным анализом подразумевают:
а) определение уравнения регрессии
б) определение математической модели, в которой среднее значение результативного признака Y рассматривается как функция одной или нескольких переменных факторных признаков
в) измерение тесноты связи между двумя (и более) признаками с помощью специальных коэффициентов
г) неправильных ответов нет

12. Для выявления наличия и характера корреляционной связи в статистике используется:
а) графический метод и применение параллельных данных (значений X и Y у n единиц)
б) метод аналитических группировок и корреляционных таблиц
в) расчет коэффициентов корреляции
г) а + б + в

13. К непараметрическим показателям статистики относятся:
а) коэффициент корреляции знаков
б) коэффициент регрессии
в) коэффициент корреляции
г) коэффициент детерминации

14. К параметрическим показателям статистики относятся:
а) коэффициент корреляции рангов
б) коэффициент множественной корреляции
в) коэффициент ассоциации
г) коэффициент взаимной сопряженности

15. Если коэффициент корреляции равен нулю, то X и Y называют:
а) некоррелированными
б) коррелированными
в) связь значимыми
г) б + в

16. При изменении значений факторного признака функциональная зависимость проявляется в изменении:

а) значений результативного признака

б) распределения единиц совокупности по результативному признаку

в) распределения единиц совокупности по факторному признаку

г) средних значений результативного признака

17. Линейный коэффициент корреляции R в состоянии удовлетворительно измерять:

а) лишь криволинейную связь

б) лишь прямолинейную связь

в) лишь долю среднего квадратического отклонения результативного признака

г) б + в

18. Для оценки значимости коэффициента корреляции R применяется:

а) t-критерий Стьюдента $\left(t_r = r\sqrt{\frac{n-2}{1-R^2}}\right)$

б) F-критерий Фишера $\left(F_r = \frac{R_2}{1-R^2} \cdot \frac{n-m}{m-1}\right)$

в) доверительные интервалы оценки коэффициента корреляции ($t_{расч} > t_{кр}$), где $t_{расч} = R : \sigma_r$

г) а + в

19. Корреляционное отношение (η) используется для:

а) определения факторной вариации

б) определения остаточной вариации

в) определения тесноты связи

г) выявления направления связи

20. С помощью F-критерия проверяется:

а) правильно ли построена группировка

б) какова теснота связи между признаками

в) значимости уравнения регрессии в целом или проверка адекватности модели

г) а + б

21. Значение F-критерия, вычисленное по данным

аналитической группировки, характеризующей связь между заработной платой рабочих и стажем работы, равно 5,3. Критическое значение для уровня значимости 0.05 равно 3,8. Это позволяет сделать вывод:

а) связь между признаками измерена точно
б) связь между признаками существенна
в) связь функциональная
г) группировка построена правильно

22. Чтобы определить, на сколько изменится среднее значение результативного признака при увеличении факторного признака на единицу, необходимо:

а) построить комбинационное распределение
б) вычислить коэффициент корреляции
в) построить аналитическую группировку
г) вычислить параметры уравнения регрессии

23. Теоретическими значениями называются:

а) групповые средние
б) значения результативного признака, вычисленного по уравнению регрессии
в) фактические значения результативного признака
г) фактические значения факторного признака

24. Путем решения системы нормальных уравнений вычисляются параметры уравнения регрессии, при которых является минимальной:

а) сумма отклонений теоретических значений результативного признака от эмпирических значений этого признака
б) сумма квадратов этих отклонений
в) сумма отклонений эмпирических значений факторного признака от теоретических значений результативного признака
г) а + в

25. Вычислено уравнение регрессии между стоимостью основных фондов и выпуском продукции (млн. сум): $\overline{Y}_x = 17 + 0{,}4x$. Это означает, что при увеличении стоимости основных фондов на 1 млн. сум выпуск продукции в среднем увеличивается:

а) на 17,4 млн. сум
б) на 0,4 млн. сум
в) на 40 %

г) на 0,54 млн. сум

26. Вычислено уравнение регрессии между процентом брака и себестоимостью 1т литья (сум): $\overline{Y}_x = 5600 + 8x$. Это означает, что:

а) увеличение брака на 1% увеличивает себестоимость на 6400 сум

б) увеличение брака на 1% увеличивает себестоимость на 8%

в) если процент брака не изменится, то себестоимость составит 5600 сум

г) если брак увеличится на 1%, то себестоимость 1т увеличится на 800 сум

27. Коэффициент детерминации характеризует:

а) форму связи

б) существенность связи

в) меру тесноты связи между признаками

г) какой удельный вес в общей дисперсии ряда **у** занимает дисперсия, вызываемая вариацией фактора x_0

28. Если коэффициент детерминации между Y и X равен нулю, то:

а) различие между групповыми средними отсутствует

б) отклонение эмпирических значений от теоретических равны нулю

в) линия регрессии проходит через все эмпирические точки

г) теоретические значения совпадает со средними значениями результативного признака

29. Коэффициент эластичности показывает, на сколько процентов изменяется в среднем результативный показатель у при изменении факторного признака х на 1% и рассчитывают по формуле:

а) $Э_x = b_j \dfrac{\overline{x}_j}{y}$

б) $Э_x = \dfrac{\Delta x_j}{\overline{x}_i} : \dfrac{\Delta y}{\overline{y}}$

в) $b = r \dfrac{\sigma_y}{\sigma_x}$

г) $a = \overline{y} - b\overline{x}$

30. По уравнению регрессии: $\overline{Y}_x = a + bx$ между процентом механизации работ (х) и производительности труда (Y) вычислен коэффициент детерминации R^2=0,37. Это означает, что:

а) 37% работ механизировано

б) при увеличении уровня механизации на 1% производительности труда увеличится на 0,37%

в) 37% вариации производительности линейно связано с вариацией процента механизации работ

г) 37% вариации производительности труда связано с различиями в уровне механизации работ

31. Линейный коэффициент корреляции ($a_1 \frac{\sigma_x}{\sigma_y}$ или $\frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y}$) характеризует:

а) существенность связи

б) направление связи

в) тесноту связи

г) неправильных ответов нет

32. Коэффициент корреляции знаков Г. Фехнера - это коэффициент, определяющий связь между:

а) двумя количественными признаками

б) двумя качественными признаками

в) количественным и качественными признаками

г) альтернативными признаками

33. Коэффициент корреляции знаков Г. Фехнера определяется по формуле:

а) $K_ф = 1 - \frac{6\sum d^2}{n^3 - n}$;

б) $K_ф = \frac{2S}{n \cdot (n-1)}$

в) $K_ф = \frac{\sum c - \sum н}{\sum c + \sum н}$

г) $K_ф = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}$

34. Для измерения тесноты связи между качественными признаками применяются:

а) коэффициент ассоциации

б) критерий «Хи-квадрат»

в) коэффициент корреляции знаков Г. Фехнера
г) а + б

35. Коэффициент контингенции:
а) всегда больше коэффициента ассоциации
б) всегда меньше коэффициента ассоциации
в) всегда равен коэффициенту ассоциации
г) неправильных ответов нет

36.Связь считается достаточно значительной и подтвержденной, если:
а) $|K_{асс}| > 0{,}5$
б) $|K_{конт}| > 0{,}3$
в) $|K_{асс}| > 0{,}3$
г) а + б

Практикум по теме 8

8.1. Методические указания

Изучение взаимосвязи – одна из важнейших задач курса общей теории статистики. Исследование объективно существующих связей между явлениями имеет большое значение в социально-экономической статистике. Изучение любого явления будет неполным, если не исследованы его связи с другими явлениями и процессами. В основе исследования связей лежит, общефилософское положение о всеобщей связи в природе и обществе. В процессе статистического исследования зависимостей вскрываются причинно-следственные отношения между явлениями, что позволяет выявлять факторы (признаки), оказывающие основное влияние на вариацию изучаемых явлений.

Статистикой выработаны специальные приемы установления и измерения такого вида связей, получившие название метода корреляционного анализа. Они применимы к измерению связей между двумя признаками – парная корреляция или к измерению связей между тремя и большим числом признаков – множественная корреляция.

Методы корреляции позволяют решить следующие основные задачи:

1) Определить среднее изменение результативного признака под влиянием одного, или комплекса факторов (в абсолютном или относительном измерении); 2) охарактеризовать меру зависимости результативного признака и одного из факторов при среднем значении других; 3) определить тесноту связи результативного признала со всем комплексом включенных в анализ факторов или с отдельным фактором при исключении влияния других; 4) статистически оценить выборочные показатели корреляционной связи. Каждая из этих задач решается путем расчета определенных показателей.

Применение метода корреляционного анализа включает ряд этапов:

1. Постановка задачи и установление причин связи. Для этого требуется глубокое понимание сущности изучаемых взаимосвязей, так как сам метод не позволяет установить причины возникновения связей между явлениями, его назначение заключается в их количественном измерении. На данном этапе осуществляется общее ознакомление с изучаемым объектом, уточняются задачи исследования, устанавливается теоретическая возможность причинно-следственной связи.

2. Ограничение объекта исследования и отбор необходимых признаков. При ограничении объекта следует иметь в виду, что корреляционный анализ должен проводиться лишь в пределах качественно однородных (в социальном, экономическом или производственно-техническом отношении) достаточно многочисленных совокупностей. Отбираемые для корреляционной модели факторные и результативные признаки должны, быть существенными, первые должны оказывать непосредственное влияние на вторые. Нежелательно включение в одну модель, частных и общих факторов, а также нескольких факторных признаков, находящихся в тесной связи друг с другом.

Серьезную помощь для отбора показателей могут оказать результативные и факторные статистические группировки с соответствующим анализом влияния факторов на результативный признак.

3. Установление формы связи и подбор математического уравнения модели связи. Этот вопрос решается на основании теоретического анализа или предшествующим практическим опытом соответствующих исследований. Если форма связи неизвестна, то проводится группировка статистических данных и изучение изменения средних по группам, сопоставление параллельных рядов, построение графиков и таблиц распределения численностей. Уравнения, выражающие статистическую связь, называются уравнениями регрессии или корреляции.

Связь между результативным и факторным признаками может носить линейный или криволинейный (параболический, синусоидальный т. п.) характер. При линейной парной связи между признаками используется уравнение прямой: $\bar{y}_x = a_0 + a_1 x$, где $\bar{y}_x$ - зависимая переменная; x_i - независимая переменная; a_o - начало отсчета, a_i - коэффициент регрессии, показывающий среднее изменение x_i при изменении x_1 на единицу. При этом единицы измерения коэффициента регрессии соответствуют единицам измерения величин $\bar{y}_x$ и x_1.

В случае линейной взаимосвязи результативного признака с несколькими факторами используется множественное линейное уравнение:

$$\bar{y}_x = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n,$$

где a_o- свободный член (начало отсчета);

$x_1, x_2, \ldots, x_n$ - факторные признаки;

$a_1, a_2, \ldots, a_n$ - коэффициент регрессии показывает степень среднего изменения зависимой переменной при изменении факторного признака на единицу при условии, что остальные факторы, включенные в уравнение остаются неизменными.

4. Расчет числовых характеристик корреляционной связи. Этот этап заключается в нахождении параметров корреляционного уравнения a_o, a_1, a_2,..., a_n. Для этого составляется система нормальных уравнений, при этом сумма квадратов отклонений фактических данных от исчисленных по уравнению должна быть минимальной (МНК), т.е. его суть заключается в следующем требовании: искомые теоретические значения результативного признака y должны быть такими, при которых бы обеспечивалась минимальная сумма квадратов их отклонений от эмпирических значений, т.е:

$$S = \sum(y - \bar{y}_x)^2 = \min \qquad (8.1)$$

(минимизируются квадраты отклонений, поскольку $\sum(y - \bar{y}_x) = 0$).

Если данное требование соблюдается, легко определить; при каких значениях a_0, a_2 и т.д. для каждой аналитической кривой эта сумма квадратов отклонений будет минимальной.

Параметры a_0 и a_1 отыскиваются по методу наименьших квадратов (МНК) следующим образом.

Согласно требованию МНК при линейной зависимости в формуле (8.1) вместо y записываем его конкретное выражение:

$a_0 + a_1x$. Тогда

$$S = \sum (y - a_0 - a_1x)^2 = min.$$

Дальнейшее решение сводится к задаче на экстремум, т.е. к определению того, при каком значении a_0 и a_1 функция двух переменных S может достигнуть минимума.

Как известно, для этого надо найти частные производные S по a_0 и a_1, приравнять их к нулю и после элементарных преобразований решить систему двух уравнений с двумя неизвестными.

$$\begin{cases} \sum y = na_0 + a_1 \sum x \\ \sum yx = a_0 \sum x + a_1 \sum x^2 \end{cases} \qquad (7.28)$$

Эта система называется системой нормальных уравнений для линейного уравнения регрессии.

Линейный коэффициент парной корреляции определяется по формуле:

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y},$$

где σ_x, σ_y – соответственно, среднее квадратическое отклонение в ряду x и в ряду y.

При множественной линейной зависимости, если известны коэффициенты парных связей используют следующую формулу:

$$R_{y_{xv_2}} = \sqrt{\frac{r^2_{yx_1} + r^2_{y_2} - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r^2_{x_1x_2}}}$$

Линейный парный коэффициент корреляции меняется в пределах от -1 до +1, а множественный коэффициент и индекс корреляции рассматривается только как положительная величина и изменяется в пределах от 0 до 1. Квадрат коэффициента корреляции и индекса корреляции называется коэффициентом детерминации и показывает, на сколько процентов результативный признак зависит от одного или нескольких факторных признаков, включенных в анализ.

Корреляция и регрессия. Традиционные методы корреляционно-регрессионного анализа позволяют не только оцедить тесноту связи, но и выразить эту связь аналитически. Применению корреляционно-регрессионного анализа должен предшествовать качественный, теоретический анализ исследуемого социально-экономического явления или процесса.

Связь между двумя факторами аналитически выражается уравнениями:

прямой $\bar{y}_x = a_0 + a_1 x$;

гиперболы $\bar{y}_x = a_0 + \frac{a_1}{x}$;

параболы $\bar{y}_x = a_0 + a_1 x + a_2 x^2$ (или другой ее степени);

степенной функции $\bar{y}_x = a_0 x^{a_1}$.

Параметр a_0 показывает усредненное влияние на результативный признак, неучтенных (не выделенных для исследования) факторов. Параметр a_1 – коэффициент регрессии показывает, на сколько изменяется в среднем значение результативного признака при увеличении факторного на единицу. На основе этого параметра вычисляются коэффициенты эластичности, которые показывают изменение результативного признака в процентах в зависимости от изменения факторного признака на 1% $Э = a_1 \frac{\bar{x}}{\bar{y}}$ (см. задачу 8.1).

Для определения параметров уравнений используется метод наименьших квадратов, на основании которого строится соответствующая система уравнений.

Теснота связи при линейной зависимости измеряется с помощью линейного коэффициента корреляции: (см. задачу 8.1).

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y},$$

а при криволинейной зависимости с помощью корреляционного отношения (см. задачу 8.29)

$$r_1 = \sqrt{\frac{\delta^2}{\sigma^2}}.$$

Удобной формой изложения данных о взаимосвязанных признаках является корреляционная таблица, представляющая собой комбинационную статистическую таблицу (см. задачу 8.4).

Показатели тесноты связи. Для оценки тесноты связи применяется ряд показателей, одни из которых называются эмпирическими или непараметрическими, другие (выводимые строго математически) – теоретическими.

Коэффициент знаков (коэффициент Фехнера) вычисляется на основании, определения знаков отклонений вариантов двух взаимосвязанных признаков от их средних величин.

Если число совпадений знаков обозначить через а, число несовпадений – через b, а t сам коэффициент – через i, то можно написать формулу этого коэффициента так:

$$i = \frac{\Sigma a - \Sigma b}{\Sigma a + \Sigma b}.$$

Коэффициент корреляции рангов (коэффициент Спирмена) рассчитывается не по значениям двух взаимосвязанных признаков, а по их рангам следующим образом:

$$\rho_{x/y} = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)},$$

где d_i^2 - квадраты разности рангов; п - число наблюдений (число пар рангов) (см. задачу 8.5).

При исследовании социальных явлений и процессов большое значение имеет изучение качественных показателей и признаков, не имеющих количественной оценки.

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c +d

Для определения тесноты связи двух качественных признаков, каждый из которых состоит только из двух групп, применяются коэффициенты ассоциации и контингенции. Для их вычисления строится таблица, которая показывает связь между двумя явлениями, каждое из которых должно быть альтернативным, т.е. состоящим из двух качественно отличных друг от друга значений признака (например, хороший, плохой).

Коэффициенты вычисляются по формулам:

$$K_A = \frac{ad - bc}{ad + bc} \text{ – ассоциации;}$$

$$K_K = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}} \text{ – контингенции.}$$

Коэффициент контингенции всегда меньше коэффициента ассоциации. Связь считается подтвержденной если $K_A \geq 0{,}5$, или $K_K \geq 0{,}3$ (см. задача 8.6).

8.2. Решение типовых задач

Задача 8.1

По данным о стоимости основных производственных фондов и объеме валовой продукции 10 малых предприятий нужно определить уравнение связи и тесноту связи. Связь предполагается линейной.

Решение

Принимая для этой связи уравнение прямой линии, определим его параметры на основе метода наименьших квадратов, решив следующую систему нормальных уравнений:

$$\begin{cases} na_0 + a_1\Sigma x = \Sigma y; \\ a_0\Sigma x + a_1\Sigma x^2 = \Sigma yx. \end{cases}$$

Расчеты указанных в системе уравнений сумм произведем в табличной форме:

Стоимость основных производственных фондов, млн. сум, x	Объем валовой подукции, млн. сум, y	xy	x^2	y^2	$\bar{y}_x$
1	20	20	1	400	19,4
2	25	50	4	625	25,0
3	31	93	9	961	30,6
4	31	124	16	961	36,2
5	40	200	25	1 600	41,8
6	56	336	36	3 136	47,4
7	52	364	49	2 704	53,0
8	60	480	64	3 60D	58,6
9	60	540	81	3 600	64,2
10	70	700	100	4 900	69.8
2=55	Σ=445	Σ= 2967	Σ=385	Σ=2248	2 = 446

$$10a_0 + 55a_1 = 445;$$
$$55a_0 + 385a_1 = 2907;$$
$$a_0 = 13{,}8;\ a_1 = 5{,}6;$$
$$\bar{y}_x = 13{,}8 + 5{,}6_x.$$

Следовательно, с увеличением стоимости основных производственных фондов на 1 млн. сум объем валовой продукции увеличивается в среднем на 5,6 млн. сум или с увеличением стоимости основных производственных фондов на 1% объем валовой продукции увеличивается на 0,69%.

Определим коэффициент эластичности:

$$Э = 5{,}6 \cdot \frac{5{,}5}{44{,}5} = \frac{30{,}8}{44{,}5} = 0{,}6921.$$

Рассчитаем величину линейного коэффициента корреляции:

$$r = \frac{\frac{2907}{10} - \frac{55}{10} \cdot \frac{445}{10}}{\sqrt{\left[\frac{22487}{10} - \left(\frac{445}{10}\right)^2\right] \cdot \left[\frac{385}{10} - \left(\frac{55}{10}\right)^2\right]}} = 0{,}98.$$

Расчет коэффициентов регрессии несколько осложняется, если ряды по исследуемым факторам сгруппированы, а связь криволинейная.

Задача 8.2

Имеется следующая информация по однотипным предприятиям торговли (возрасте, продолжительности эксплуатации) типового оборудования и затратах на его ремонт.

Номер предприятия	Возраст оборудования, лет	Затраты на ремонт, млн. сум
1	2	3
1	4	1,5
2	5	2,0
3	5	1,4
4	6	2,3
5	8	2,7
6	10	4,0
7	8	2,3
8	7	2,5
9	11	6,6
ю	6	1,7

В целях нормирования расходов средств на ремонт оборудования произвести синтезирование адекватной экономико-математической модели. Для чего определить:

1. Параметры парной корреляции уравнений регрессии.

2. Выровненные значения $\bar{y}_x$.

3. Среднее квадратическое отклонение результативного признака y.

4. Фактическое значение t-критерия для параметра a_0.

5. Среднее квадратическое отклонение факторного признака x от общей средней $\bar{x}$.

6. Фактическое значение t-критерия для параметра a_1.

7. Проверить параметры на типичность.

8. Оценить практическую значимость синтезированной модели.

9. Значимость коэффициента корреляции. Сделать соответствующие выводы.

Решение

При статистическом изучении связи показателей коммерческой деятельности нередко постулируется прямолинейная форма зависимости между признаками x и y перед применением формулы

$$Yx=a_0+a_1x.$$

Для определения параметра уравнения на основе требований метода наименьших квадратов составляется система нормальных уравнений:

$$\begin{cases} na_0 + a_1\Sigma x = \Sigma y; \\ a_0\Sigma x + a_1\Sigma x^2 = \Sigma yx. \end{cases}$$

Для решения системы применяется способ определителей, позволяющий сводить к минимуму неточности округлений в расчетах параметров уравнений регрессии:

$$a_0 = \frac{\sum y \sum x^2 - \sum xy \sum x}{n\sum x^2 - \sum x \sum x}$$

$$a_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \sum x \sum x}$$

Применительно к анализируемым данным для решения алгоритмов и составляется расчетная таблица.

№ п/л	y	x	x^2	xy
1	2	3	4	5
1	1,5	4	16	6,0
2	2,0	5	25	10,0
3	1,4	5	25	7,0
4	2,3	6	36	13,8
5	2,7	8	64	21,6
6	4,0	10	100	40,0
7	2,3	8	64	18,4

8	2,5	7	49	17,5
9	6,6	11	121	72,6
10	1,7	6	36	10,2
	27,0	70	536	217,1

1

По итоговым данным табл. 8.23 определим параметры уравнения регрессии (8.1):

$$a_0 = \frac{27 \cdot 536 - 217{,}1 \cdot 70}{10 \cdot 536 - 70 \cdot 70} = -1{,}576;$$

$$a_1 = \frac{10 \cdot 217{,}1 - 70 \cdot 27}{10 \cdot 536 - 70 \cdot 70} = 0{,}611$$

Вычисленные значения параметров a_0 = -1,576 и a_1 = 0,611 необходимы для синтезирования математической модели зависимости расходов на ремонт от возраста оборудования. Подставляя значения вычисленных в анализе параметров в уравнение регрессии получаем:

$$y_x = -1{,}576 + 0{,}611\ x.$$

Но прежде чем использовать модель в последующем анализе, необходима проверка ее параметров на типичность, что осуществляется по формулам:

Для параметров a_0 $t_{a_0} = a_0 \dfrac{\sqrt{n-2}}{\sigma_y}$.

Для параметров a_1 $t_{a_1} = a_1 \dfrac{\sqrt{n-2} \cdot \sigma_x}{\sigma_y^*}$.

В этих формулах содержится :

$$\sigma_y = \sqrt{\frac{\Sigma\left(x_1 - Y_{x_1}\right)^2}{n}}\ ; \ \sigma_x = \sqrt{\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n}}.$$

2

Для определения σ_x на основе модели определяются выровненные значения $\bar{y}_x$

$$y_{x1} = -1{,}576 + 0{,}611 \cdot 4 = 0{,}868;$$

$$y_{x2,3} = -1{,}576 + 0{,}611 \cdot 5 = 1{,}479;$$

$$y_{x4,10} = -1{,}576 + 0{,}611 \cdot 6 = 2{,}09;$$

$$y_{x5,7} = -1{,}576 + 0{,}611 \cdot 8 = 3{,}312;$$
$$y_{x6} = -1{,}576 + 0{,}611 \cdot 10 = 4{,}536;$$
$$y_{x8} = -1{,}576 + 0{,}611 \cdot 7 = 2{,}700;$$
$$y_{x9} = -1{,}576 + 0{,}611 \cdot 11 = 5{,}145.$$

3

Определяется среднее квадратическое отклонение результативного признака $\bar{y}_y$ от выравненного значения $\bar{y}_x$:

$$\sigma_y = \sqrt{\frac{\sum (y_i - y_{x_i})^2}{n}} = \sqrt[3]{\frac{4{,}712}{10}} = 0,69$$

4

Определяются фактические значения t-критерия для параметра a_0

$$t_{a_0} = a_0 \cdot \frac{\sqrt{n-2}}{\sigma_y} = \frac{1,576\sqrt{10-2}}{0,69} = 6,46$$

5

Определяется среднее квадратическое отклонение факторного признака X_i от общей средней $\bar{X}$

$$\sigma_x = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{146}{10} = 2{,}14.$$

6

Определяется фактическое значение t-критерия для параметра a_1:

$$t_{a_1} = a_1 \cdot \frac{\sqrt{n-2} \cdot \sigma_x}{\sigma_y} = \frac{0,611\sqrt{10-2} \cdot 2,14}{0,69} = 5,36$$

7

С учетом принятых в экономико-статистических исследованиях значимости α = 0,05 и числа степеней свободы k = 10-2 табличное критическое значение t_k = 2,3.

Сравнение фактических и табличных значений t-критерия:

$$t_{a_0} > t_k < t_{a_1}$$
$$6,46 > 2,3 < 5,36$$

Это позволяет признать, вычисленные по уравнению:

$$Y_x = a_0 + a_1 x,$$

параметры типичными.

8

Далее произведем оценку практической значимости синтезированной модели y_x =-1,576 + 0,611x. Для прямолинейной связи это выполняется посредством показателя коэффициента корреляции *r*. По формуле определяется значение *r*:

$$r = \frac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\left[\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}\right]} = \frac{217{,}1 - \dfrac{70 \cdot 27}{10}}{\sqrt{536 - \dfrac{70^2}{10}}\left[94{,}78 - \dfrac{27^2}{10}\right]} = 0{,}89.$$

Полученная величина r = 0,89 означает, что в соответствии со шкалой Чеддока установленная по уравнению регрессии: $\bar{y}_x = a_0 + a_1 x$ связь между затратами на ремонт и возрастом оборудования высокая.

9

Проверка коэффициента корреляции на существенность

Интерпретируя значение коэффициента корреляции, следует иметь в виду, что он рассчитан для ограниченного числа наблюдений и подвержен случайным колебаниям, как и сами значения х и у, на основе которых он рассчитан, т.е. как любой выборочный показатель, он содержит случайную ошибку и не всегда однозначно отражает действительно реальную связь между изучаемыми показателями.

Оценка значимости коэффициента корреляции осуществляется по t-критерию.

Фактическое значение этого критерия t, определяется по формуле:

$$t_r = r\sqrt{\frac{n-2}{1-r^2}} = 0{,}89 \cdot \sqrt{\frac{10-2}{1-0{,}89^2}} = 3{,}69.$$

При критическом значении t_k= 2,3 получается, что

$$tr > tk,$$

$$3{,}69 > 2{,}3.$$

Резюме

Из значения R^2 = 0,792 следует, что 79,2% общей вариации объясняется изменением факторного признака. Поэтому синтезированная по уравнению $\bar{y}_x = a_0 + a_1 x$ математическая модель:

$$\bar{y}_x = 1{,}576 + 0{,}611x$$

может быть использована для практических целей.

Другой способ оценки значимости коэффициента корреляции

Для того чтобы оценить существенность (значимость) самого r и, соответственно, реальность измеряемой связи между x и y, необходимо рассчитать среднюю квадратическую ошибку коэффициента корреляции σ_r.

Оценка существенности (значимости) линейного коэффициента корреляции основана на сопоставлении значения r с его средней квадратической ошибкой:

$$\frac{|r|}{\sigma_r}$$

Проверим на значимость линейный коэффициент корреляции. Так как n –0,95, средняя ошибка коэффициента корреляции

$$\sigma_r = \frac{\sqrt{1-r^2}}{n-2} = \frac{\sqrt{1-0{,}89^2}}{\sqrt{10-2}} = 0{,}12.$$

Отсюда

$$t_{расч} = \frac{|r|}{\sigma^2} = \frac{0{,}89}{0{,}12} = 7{,}6.$$

По таблице находим $t_{табл}$ (при $\alpha = 0{,}05$ и числе степеней свободы $v = n–2–8)$

$$t_{табл} = 2{,}3.$$

Так как полученное $t_{расч} = 7{,}6$ больше $t_{табл} = 2{,}3$, то нулевая гипотеза об отсутствии связи между x и y в генеральной совокупности отвергается, т.е. коэффициент корреляции значим и существенно отличается от нуля, подтверждая тем самым реальную связь между x и y.

Задача 8.3

Взаимосвязь между среднегодовой стоимостью основных производственных фондов, относительным уровнем затрат на реализацию продукции и стоимостью реализованной продукции по 10 малым предприятиям характеризуется следующими данными:

Номер предп-риятия	Стоимость основных производственн ых фондов, млн. сум	Уровень затрат на реализацию (в % к стоимости реализован-ной продукции)	Объем реализо-ванной продукции, млн. сум
1	3	4	20
2	3	3	25

3	5	3	20
4	6	5	30
5	7	10	32
6	6	12	25
7	8	12	29
8	9	11	37
9	9	15	36
10	10	15	40

Считая зависимость между этими показателями линейной, определим параметры уравнений регрессии a_0, a_1, a_2 и вычислим множественный и частные коэффициенты корреляции.

Решение

Составим вспомогательную таблицу для расчета коэффициентов регрессии и корреляции:

Номер предприятия	x	z	y	x^2	z^2	xz	xy	yz	y^2
1	3	4	20	9	16	12	63	80	400
2	3	3	25	9	9	9	75	75	625
3	5	3	20	25	9	15	100	60	400
4	6	5	30	36	25	30	180	150	900
5	7	10	32	49	100	70	224	320	1 024
6	6	12	25	36	144	72	150	300	625
7	8	12	29	64	144	96	232	348	841
8	9	11	37	81	121	99	333	407	1 369
9	9	15	36	81	225	135	324	540	1 296
10	10	15	40	100	205	150	400	600	1 600
Итого	66	90	294	490	1 018	683	2 078	2 880	9 080

$$10a_0 + 66a_1 + 90a_2 = 294;$$
$$66a_0 + 490a_1 + 688a_2 = 2078;$$
$$90a_0 + 688a_1 + 1018a_2 = 1880.$$

Отсюда $a_0 = 12{,}508$; $a_1 = 2{,}672$; $a_2 = 0{,}0826$.

Следовательно: $\bar{y}_{xz} = 12{,}508 + 2{,}672x - 0{,}0826z$;

$$r_{xy} = 0{,}89;\ r_{xz} = 0{,}88;\ r_{yz} = 0{,}78.$$

Множественный коэффициент корреляции равен:

$$R = \sqrt{\frac{r_{xy}^2 + r_{yz}^2 - 2r_{yx} \cdot r_{yz} \cdot r_{xz}}{1 - r_{xz}^2}} =$$

$$= \sqrt{\frac{0,7921 + 0,6084 - 2 \cdot 0,89 \cdot 0,78 \cdot 0,88}{1 - 0,7714}} = \sqrt{\frac{0,1787}{0,2256}} = \sqrt{0,7921} = 0,89.$$

Рассчитаем частные коэффициенты:

$$r_{yx(z)} = \frac{r_{xy} - r_{yz} \cdot r_{xz}}{\sqrt{(1 - r_{yz}^2)(1 - r_{xz}^2)}} = \frac{0,89 - 0,78 \cdot 0,88}{\sqrt{(1 - 0,6084) \cdot (1 - 0,7744)}} = 0,69;$$

$$r_{yz(z)} = \frac{r_{yz} - r_{xy} \cdot r_{xz}}{\sqrt{(1 - r_{xy}^2)(1 - r_{xz}^2)}} = \frac{0,78 - 0,89 \cdot 0,88}{\sqrt{(1 - 0,7921) \cdot (1 - 0,7744)}} =$$

$$= \frac{-0,0032}{\sqrt{0,2079 \cdot 0,2256}} = \frac{-0,0032}{\sqrt{0,0469}} = \frac{-0,0032}{0,2166} = -0,01477.$$

Как видим, связь между среднегодовой стоимостью основных производственных фондов, относительным уровнем затрат на реализацию продукции и стоимостью реализованной продукции связь очень тесна, так как коэффициент множественной корреляции составляет 0,890, а детерминации – 0,792, т.е. 79,2% колебаний реализации продукции в данных условиях зависит от исследованных факторов и только 20,8 – от других факторов, которые не были учтены в анализе.

Задача 8.4

Имеются следующие данные, характеризующие связь между количеством станков, обслуживаемых одной работницей, и часовой выработкой ткани.

Часовая выработка ткани, м y	**5–7**	**7–9**	**9–11**	**11–13**	**13–15**	**15–17**	**17–19**	**Итого**
10–15	7	4		2	1			14
15–20	3	8	5	4				20
20–25		2	11	8		2		23
25–30			5	13	7	1		26
30–35				1	16	3		20
35–40				2	6	19	3	30
40-45					3	7	18	28
Итого	10	14	21	30	33	32	21	161

Таблица показывает, что частоты концентрируются у диагонали, идущей из левого верхнего угла в правый нижний. Это указывает на то, что связь между количеством обслуживаемых работницей станков и ее часовой выработкой ткани прямая (с увеличением числа обслуживаемых станков увеличивается выработка) или близкая к прямой (концентрация частот идет почти по прямой линии).

По данным таблицы можно рассчитать среднюю выработку по каждой из семи групп работниц, выделенных по числу обслуживаемых станков. Обозначив эти средние значения через $\bar{y}$ и произведя расчеты, получаем: $\bar{y}_1 = 14{,}0$; $\bar{y}_2 = 16{,}79$; $\bar{y}_3 = 22{,}51$; $\bar{y}_4 = 24{,}67$; $\bar{y}_5 = 32{,}65$; $\bar{y}_6 = 36{,}88$; $\bar{y}_7 = 41{,}79$.

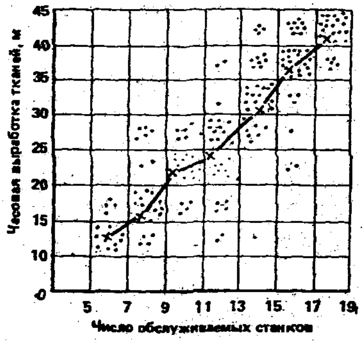


Рис. 8.1. **Поле корреляции и эмпирическая линия регрессии**

Ломанная линия на графике (линия значений $\bar{y}_x$) называется эмпирической линией регрессии.

Задача 8.5

По данным одной группы однотипных «предприятий о реализованной продукции (x, млн. сум) и накладных расходах по реализации этой продукции (у, тыс. сум) рассчитать коэффициент Спирмена:

Номер пред-приятия	x	y	Ранжирование				Сравнение рангов		Разность рангов	d_i^2
			x	y	ранг					
					R_x	R_y	R_x	R_y	$d_i=R_x-R_y$	
1	12,0	462	11,0	462	t	1	2	1	1	1
2	18,8	939	12,0	506	2	2	5	6	–1	1
3	11,0	506	15,4	765	3	3	1	2	–1	1

4	29,0	1 108	17,5	804	4	4	9	9	0	0
5	17,5	872	18,8	872	5	5	4	5	–5	1
6	23,4	765[г]	20,7	939	6	6	7	3	4	16
7	35,6	1 368	23,4	998	7	7	10	10	0	0
8	15	1 002	26,1	1 002	8	8	3	8	–5	25
9	26,1	998	29,0	1 108	9	9	8	7	1	1
10	20,7	804	35,6	1 368	10	10	6	4	2	4
										Σ=50

Решение

Составим вспомогательную вышеприведенную таблицу для расчета коэффициента корреляции рангов (коэффициента Спирмена). Он рассчитывается не по значению двух взаимосвязанных признаков, а по их рангам следующим образом:

$$P_{x/y} = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = \frac{6,50}{10 \cdot 99} = 1 - \frac{300}{990} = 0,687.$$

Следовательно, связь между исследуемыми показателями достаточно тесная.

Для определения тесноты связи между тремя и более признаками применяется ранговый коэффициент согласия – коэффициент конкордации, который вычисляется по формуле

$$W = \frac{12S}{m^2(n^3-n)},$$

где m - количество факторов; n - число, наблюдений; S - сумма квадратов отклонений рангов (см. задачу 8.6).

Задача 8.6

Исследовалась связь между выполнением норм выработки рабочими и уровнем их образования. Результаты обследования характеризуются следующими данными:

Группы рабочих	**Выполняющие нормы**	**Не выполняющие нормы**	**Всего**
Окончившие колледж Не окончившие колледж	78(а) 32(с)	22(в) 68 (d)	100 100
Итого	110	90	200

Определить коэффициент ассоциации и контингенции.

Решение

При исследовании степени тесноты связей между качественными признаками, каждый из которых представлен в виде альтернативного признака, используют коэффициент ассоциации или коэффициент контингенции.

Построенная в такой форме таблица носит название «Таблица четырех полей», частоты которой обозначим, соответственно, a, b, c, d. Коэффициент ассоциации ($К_А$) определяется по формуле:

$$K_A = \frac{ad - bc}{ad + bc} = \frac{78 \cdot 68 - 32 \cdot 22}{78 \cdot 68 + 32 \cdot 22} = \frac{4600}{6008} = 0,776.$$

Таким образом, можно сделать убедительный вывод о связи между уровнем образования рабочих и уровнем выполнения нормы выработки, поскольку теснота связи высока (77,6%).

В тех случаях, когда хотя бы один из четырех показателей и в таблице «четырех полей» отсутствует, величина коэффициента ассоциации будет равна единице, что дает преувеличенную оценку степени тесноты связи между признаками, и предпочтение следует отдать коэффициенту контингенции ($К_k$):

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

В приведенном примере его величина будет равна 0,46.

$$K_k = \frac{76 \cdot 68 - 32 \cdot 22}{\sqrt{(78+22)(22+68)(78+32)(32+6)8}} = \frac{5304 - 704}{\sqrt{99000000}} = 0,46.$$

В этом случае вряд ли можно сделать убедительный вывод о тесной связи между изучаемыми признаками, поскольку степень тесноты связи невелика. Связь считается подтвержденной, если $К_А \geq 0,5$, или $К_k \geq 0,3$.

Задача 8.7

Имеются следующие данные по вилоятам республики об урожайности картошки и бахчевых (см. таблицы). Определите коэффициент знаков, т.е. коэффициенты Фехнера.

Решение

Как было указано выше коэффициент Фехнера вычисляется на основе определения количества совпадений знаков двух взаимосвязанных признаков от их средних величин

$$i = \frac{\Sigma a - \Sigma b}{\Sigma a + \Sigma b}.$$

Средняя урожайность картошки и бахчевых в вилоятах РУз

Регионы	Урожайность (ц/га)		Знаки, показывающие отклонения от средней республиканской	
	картошка (x)	бахчевые (y)	картошка	бахчевые
Республика Узбекистан	92	193	x	x
Республика КК	37	89	–	–
Андижан	93	207	+	+
Бухара	101	160	+	–
Джизак	82	160	–	–
Кашкадарья	72	174	–	–
Наманган	114	237	+	+
Самарканд	98	147	+	–
Сурхандарья	158	169	+	–
Сырдарья	44	192	–	–
Ташкент	79	218	–	+
Фергана	128	179	+	–
Хорезм	62	140	–	–

Здесь a - знаки, которые совпадают в обоих случаях;
b - знаки, которые не совпадают в обоих случаях;
n - единицы наблюдения (число областей).

В 12 областях средняя урожайность картошки и бахчевых в семи случаях по обоим продуктам совпадали. Так, например, в Республике КК средняя урожайность картошки 37 ц/га и ниже среднереспубликанской (–), так же по бахчевым ниже, чем среднереспубликанской 89 ц/га (–). Знаки совпадали ($\Sigma a = 7$). В пяти случаях знаки не совпадали (+, –). Это в областях: Бухара, Самарканд, Сурхандарья, Ташкент, и Фергана ($\Sigma b = 5$).

Принимая вышеприведенные формулы, рассчитываем коэффициент Фехнера:

$$i = \frac{\Sigma a - \Sigma b}{\Sigma a + \Sigma b} = \frac{7-5}{7+5} = \frac{2}{12} = 0{,}17.$$

Этот коэффициент можно рассчитать, применяя коэффициент соответствия ($К_с$)

$$K_c = \frac{a-b}{n} = \frac{7-5}{12} = \frac{2}{12} = 0{,}17.$$

Следовательно, связь между этими признаками очень слабая и прямолинейная.

8.3. Задачи для аудиторных занятий и самостоятельной работы

Задача 8.8

По 10 малым предприятиям одной и той же отрасли известны следующие данные за месяц:

Номер малого пред-приятия	Валовая продукция, млн. сум	Фонд заработной платы, млн. сум	Номер пред-приятия	Валовая продукция, млн. сум	Фонд заработной платы, млн. сум
1	500,0	110,0	6	1000,0	180,0
2	600,0	130,0	7	1100,0	200,0
3	700,0	145,0	8	1200,0	230,0
4	800,0	160,0	9	1300,0	240,0
5	900,0	150,0	10	1400,0	250,0

По этим данным: а) найдите уравнение линейной регрессии фонд заработной платы от валовой продукции этих предприятий; б) дайте экономическую интерпретацию параметра aj; в) определите коэффициент корреляции фонда заработной платы по валовой продукции; г) изобразите графически эмпирическую и теоретическую зависимость фонда заработной платы от валовой продукции.

Ответ: а) Y_x =2^52+0,158 х; б) R_{xy} =0,99.

Задача 8.9

По 25 малым промышленным предприятиям, производящим один вид продукции, известны следующие данные:

№ п/п	Энерговооруженность труда, кВт. ч (х)	Производительность труда, шт. (у)	№ п/п	Энерговооруженность труда, кВт. ч	Производительность труда, шт. (у)
1	7,1	14	13	11,6	19
2	7,2	15	14	12,3	20
3	7,9	18	15	12,4	18

4	8,3	16	16	12,6	21
5	8,5	14	17	12,8	21
6	9,0	15	18	13,2	20
7	9,3	16	19	13,4	19
8	9,6	17	20	13,7	20
9	9,7	18	21	13,8	21
10	10,2	20	22	14,0	18
11	10,5	17	23	14,1	22
12	11,0	19	24	14,3	21
			25	14,4	22

По этим данным: а) найдите линейное уравнение регрессии, выражающее зависимость производительности труда от параметра энерговооруженности труда и охарактеризуйте экономический смысл параметра a_1; б) определите коэффициент корреляции производительности труда по энерговооруженности труда; в) постройте график эмпирической и теоретической зависимости производительности труда от энерговооруженности труда.

Ответ: а) $\mathbf{Y_x = 0{,}90 + 1{,}46x}$; б) $\mathbf{R_{xy} = 0{,}85}$.

Задача 8.10

Определите тесноту связи с помощью коэффициента знаков Фехнера между урожайностью картофеля и количеством внесенных минеральных удобрений по следующим данным:

Номер фермерского хозяйства	**Урожайность, ц/га**	**Внесено минеральных удобрений на 1 га, кг**	**Номер фермерского хозяйства**	**Урожай ность, ц/га**	**Внесено минеральных удобрений на 1 га, кг**
1	128	140	в	183	197
2	179	262	7	201	246
3	221	289	8	195	276
4	136	191	9	Н1	187
5	164	202	10	19?	253

Ответ: 0,30

Задача 8.11

Имеются следующие данные о стоимости основных промышленно-производственных фондов и среднесуточной переработки сырья:

Стоимость промышленно-Производственных основных фондов, млн. сум	Среднесуточная переработка сырья, тыс. ц				Итого
	4-6	6–8	8–10	10–12	
2,5–3,5	2				2
3,5–4,5	6	3			9
4,5–5,5	2	5	7		14
5,5–6,5		2	2	3	7
6,5–7,5			1	7	8
Итого	10 -	10	10	10	40

Определите вид корреляционной зависимости, найдите параметры уравнения регрессии, определите тесноту связи. Дайте анализ полученным результатам.

Ответ: $\bar{y}_x = -0{,}441{,}607x$

Задача 8.12

По следующим данным постройте линейное уравнение регрессии, вычислите линейный коэффициент корреляции: $\overline{xy} = 106; \bar{x} = 11; \bar{y} = 9;$ $\bar{x}^2 = 137; \bar{y}^2 = 85; a_0 = 4{,}8.$

Ответ: $\bar{y}_x = 48 + 0{,}44x$

Задача 8.13

По следующим данным постройте линейное уравнение регрессии: $a_0 = 2{,}8; r = 0{,}9; \sigma_y^2 = 25; \sigma_x^2 = 36.$

Задача 8.14

Вычислите корреляционное отношение, если известно, что общая дисперсия равна 35,68, групповые дисперсии $\sigma_1^2 = 12; \sigma_2^2 = 6; \sigma_3^2 = 18,$ а численность групп – соответственно, 36, 43, 21 единиц.

Задача 8.15

Определите величину корреляционного отношения, если известны: общая дисперсия $\sigma^2 = 8{,}4,$ общая средняя $\bar{x} = 13{,}0,$ групповые средние $\bar{x}_1 = 10, \bar{x}_2 = 15, \bar{x}_3 = 12{,}0,$ численность групп, соответственно, равна 32, 53, 45.

Задача 8.16

Преподавателям по курсам «Экономическая теория» и «Высшая математика» было предложено ранжировать 10 студентов по уровню знания ими соответствующих предметов:

Студенты	1	2	3	4	5	6	7	8	9	10
Ранг по высшей математике	5	6	3	2	10	7	4	1	8	9
Ранг по экономической теории	1	4	3	7	5	8	6	2	9	10

Для оценки степени тесноты связи между уровнем знания экономической теории и высшей математики вычислите коэффициенты корреляции рангов Спирмена и Кендела. Сформулируйте вывод, следующий из исчисленных вами коэффициентов.

Ответ: а) $К_{СП}$ = 0,53; б) $К_{К}$ = 0,42.

Задача 8.17

При подведении итогов экзаменационной сессии в группе были получены следующие данные о зависимости между количеством пропущенных обязательных занятий студентом без уважительных причин и средним баллом его успеваемости по пяти предметам:

№ п/п	Количество пропущенных обязательных занятий, (ч)	Средний балл по всем предметам	№ п/п	Количество пропущенных обязательных занятий, (ч)	Средний балл по всем предметам
1	38	3,8	16	24,4	4,3
2	0	4,8	17	16	4,7
3	6	5,0	18	24	4,2
4	26	3,7	19	34	3,8
5	18	3,4	20	56	3,0
6	56	3,0	21	4	5,0
7	28	4,1	22	2	3,6
8	35	3,9	23	38	4,0
9	14	4,6	24	54	3,2
10	32	3,9	25	16	4,5
11	12	5,0	26	14	4,2
12	38	3,9	27	12	4,7
13	10	4,6	28	36	3,9
14	54	3,5	29	52	3,4
15	48	3,2	30	60	3,3

Разбейте численность студентов на три группы по количеству пропущенных обязательных занятий, укажите средний балл для каждого студента. Методами дисперсионного анализа исследуйте, влияют ли непосещения занятий на успеваемость. Определите вид зависимости, постройте уравнение регрессии и рассчитайте параметры. Определите тесноту связи.

Задача 8.18

Исследовалась связь между тремя факторами. При этом парные коэффициенты корреляции получили следующие значения: $r_{yx} = 0{,}71$; $r_{yz} = 0{,}60$; $r_{xz} = 0{,}48$. Вычислите совокупный и частный коэффициенты корреляции. Объясните результаты решения задачи.

Задача 8.19

Имеются следующие данные по 8 малым предприятиям о стоимости основных производственных фондов, x (млн. сум) и суточной переработке хлопка сырца, y (т):

x	y
2,0	8,9
2,3	10,0
2,4	9,9
2,9	10,3
2,9	10,0
3,7	13,0
3,7	12,8
4,1	13,1

1. Найти уравнение регрессии y по x и определить значимость его параметров (с помощью t-критерия).
2. Проверить остаточные величины на автокорреляцию.
3. Измерить тесноту зависимости между x и y с помощью:
а) корреляционного отношения;
б) линейного коэффициента корреляции;
в) коэффициентов корреляции рангов Спирмэна и Кендэла.
4. С помощью F-критерия Фишера проверить уравнение регрессии на значимость (существенность).

Ответ: 1) $\bar{y}_x = 4{,}64 + 2{,}12x$, параметры значимы, так как $t_ф > t_{табл}$;
2) автокорреляция отсутствует;
3) а) $\eta = 0{,}96$, б) $r = 0{,}96$, в) $\rho = 0{,}94$, $\tau = 0{,}87$;
4) уравнение значимо, так как $F_{факт} > F_{табл}$.

Задача 8.20

У 8 учащихся колледжа зафиксировано следующее количество баллов, полученных за самостоятельные работы по математике (x) и по гуманитарным предметам (y).

Студент	X	$У$
А	90	75
Б	60	69
В	46	45
Г	68	49
Д	82	58
Е	71	54
Ж	66	59
З	78	70

Для характеристики корреляции между успеваемостью по математике и гуманитарным предметам рассчитать:

а) коэффициент Фехнера;

б) коэффициент корреляции рангов Спирмэна;

в) коэффициент корреляции рангов Кендэла.

Ответ: а) $K_ф = 0{,}25$; б) $\rho = 0{,}548$; в) $\tau = 0{,}428$.

Задача 8.21

Совокупность разбита по определенному признаку z на 3 группы, численность которых: $n_1 = 10$, $n_2 = 20$, $n_3 = 20$.

Групповые средние исследуемого показателя x равны: $\bar{x}_1 = 5$, $\bar{x}_2 = 8$, $\bar{x}_3 = 15$.

Определить величину эмпирического корреляционного отношения, если общая дисперсия показателя x равна 18,5.

Ответ: $\eta_{эмп} = 0{,}946$.

Задача 8.21

На основе опроса 400 работников коммерческих структур и 400 работников бюджетных организаций получено следующее их распределение по ответам на вопрос, довольны ли они своей заработной платой:

Работающие	Довольные зарплатой	Недовольные зарплатой	Итого
В коммерческих структурах В бюджетных организациях	360 140	40 260	400 400
Итого работников	500	300	800

1. С помощью критерия Пирсона χ^2 определить, случайно или неслучайно данное распределение.

2. Рассчитать коэффициенты ассоциации и контингенции.

Ответ: 1) $\chi^2 = 258{,}2$; 2) $k_{асс} = 0{,}887$, $k_{континг} = 0{,}568$.

Задача 8.23

Записать уравнение регрессии y по x, имея следующие данные:

$\bar{x} = 20$, $\overline{x^2} = 436$, $\bar{y} = 60$, $\overline{y^2} = 3700$, $r_{xy} = 0{,}75$.

Ответ: $\bar{y}_x = 35 + 1{,}25x$

Задача 8.24

Имеются следующее данные о распределении 200 молочных ферм области по производительности труда и себестоимости молока:

Производительность / Себестоимость	Высокая	Средняя	Низкая	Итого
Высокая	10	10	30	50
Средня	30	30	10	70
Низкая	50	20	10	80
Итого	90	60	50	200

1. С помощью критерия χ^2 проверить, случайно ли данное распределение, т.е. существует ли зависимость между производительностью труда и себестоимостью молока.

2. Измерить тесноту зависимости между указанными показателями с помощью коэффициентов взаимной сопряженности Пирсона и Чупрова.

Ответ: 1) $\chi^2 = 51{,}4$; 2) С= 0,45, $K_ч = 0{,}36$.

8.4. Рекомендации преподавателям

1. *Практические занятия.* Необходимо решение задач на определение эмпирического корреляционного отношения (с выяснением смысла дисперсий и отношения межгрупповой и общей дисперсий) и на вычисление теоретического корреляционного отношения. Следует обратить внимание на выяснение смысла параметров уравнения регрессии.

Очень полезно решение задач, аналогичных задачам 8.16 и 8.28. Эти задачи способствуют уяснению связи показателей, на основе которых измеряется теснота связи.

2. *Задание для самостоятельной внеаудиторной работы студентов.* В качестве такой работы можно дать задание 8.26, 8.18, 8.20, 8.21, 8.22, 8.24.

3. *Аудиторная контрольная работа.* Может быть составлена из одной-двух задач типа задачи 8.29 и двух-трех – типа 8.30,8.31, 8.32.

Постановка такого рода задач дает возможность контролировать усвоение студентами темы и экономно расходовать учебные часы (работу можно дать на 1 час или даже на 20-30 минут).

Так же по теме необходимо провести часовой текущий контроль, используя для этого «Сборник тестов по курсу «Статистика» (см. Приложение 1к учебнику «Статистика» Тема 8). Провести тестирование – преподаватель имеет возможность оценить уровень знаний студентов за очень короткое время.

Слайды по теме 8 для проведения «Мастер класса»

Дорожная карта

Слайд 8.1. Понятие корреляционной зависимости.

Слайд 8.2. Виды связей и методы их изучения.

Слайд 8.3. Виды зависимостей на корреляционном поле. Корреляционная таблица.

Слайд 8.4. Условия правильного применения методов корреляционного анализа.

Слайд 8.5. Задачи исследования корреляционных зависимостей.

Слайд 8.6. Виды уравнения прямолинейной и криволинейной связи.

Слайд 8.7 – 8.10. Парная корреляция.

Слайд 8.11 – 8.12. Множественная корреляция.

Слайд 8.13 – 8.14. Непараметрические показатели оценки связи.

Слайд 8.15 – 8.16. Решение типовой задачи.

СЛАЙД 8.1

ПОНЯТИЕ КОРРЕЛЯЦИОННОЙ СВЯЗИ

Содержание теории корреляции составляет изучение зависимости вариации признака от окружающих условий

ВИДЫ СВЯЗЕЙ

Балансовая связь
ОН + П = В + ОК
Компонентные связи
$a = в \cdot c$
$J_{qp} = J_q \cdot J_p$
Факторные связи

Функциональные	Корреляционные
Y = f(x) $l = 2\pi r$	$\bar{y}_x = a + bx$

Слово «корреляция» (англ. - correlation) означает соотношение, соответствие
Основоположниками теории корреляции считаются английские биометрики Ф. Гальтон (1822-1911) и К. Пирсон (1857-1936)

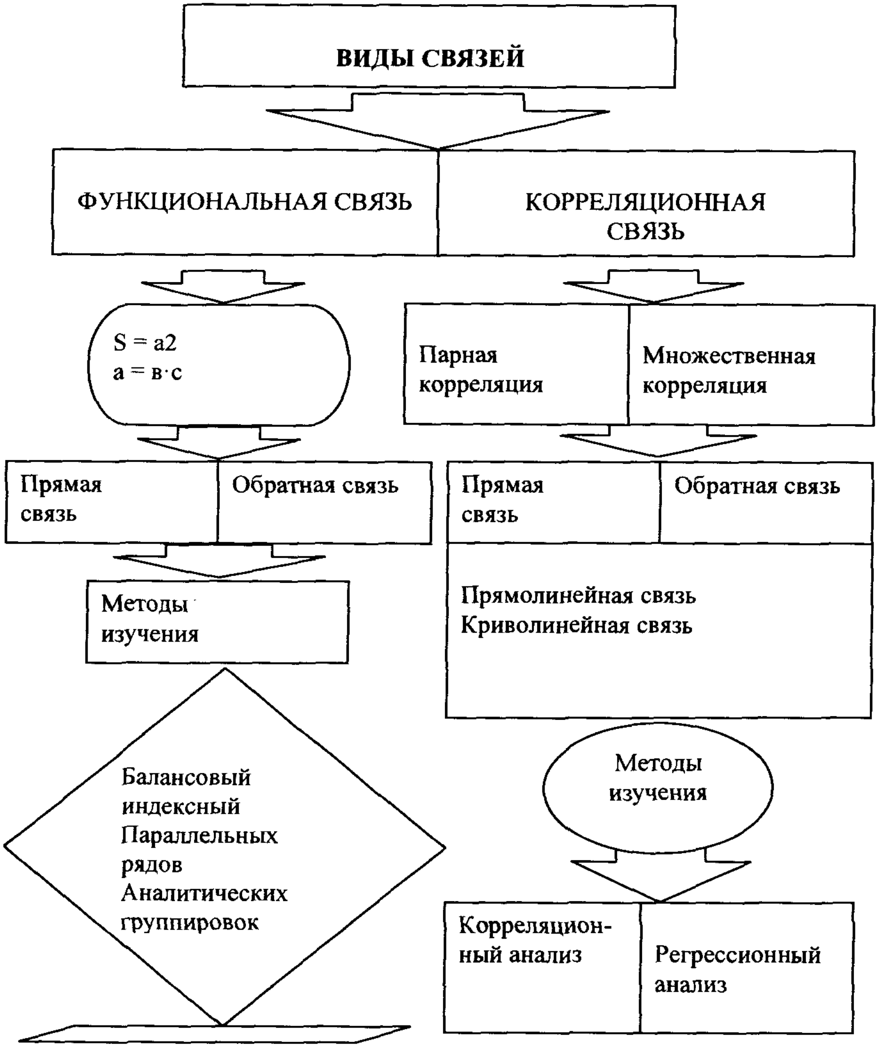
Ф.Гальтон

(1822-1911)

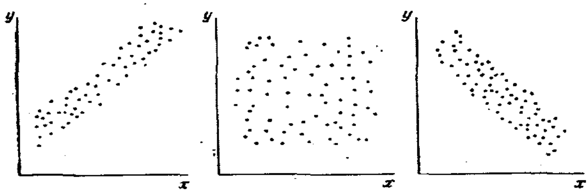
К.Пирсон

(1857-1936)

СЛАЙД 8.2


ВИДЫ СВЯЗЕЙ
ФУНКЦИОНАЛЬНАЯ СВЯЗЬ
КОРРЕЛЯЦИОННАЯ СВЯЗЬ
S = a2
a = в·с
Парная корреляция
Множественная корреляция
Прямая связь
Обратная связь
Прямая связь
Обратная связь
Методы изучения
Прямолинейная связь
Криволинейная связь
Балансовый
индексный
Параллельных рядов
Аналитических группировок
Методы изучения
Корреляцион-ный анализ
Регрессионный анализ

СЛАЙД 8.3
ВИДЫ ЗАВИСИМОСТЕЙ НА КОРРЕЛЯЦИОННОМ ПОЛЕ



Прямая корреляционная зависимость *Отсутствие зависимости* *Обратная корреляционная зависимость*

Корреляционная таблица
Группировка значений факторного и результативного признаков

Центральное значение интервала	768	865	962	1059	1156	fx	$\bar{y}_j$
Группы по y / Группы по x	720-816	817-913	914-1010	1011-1107	1108-1207		
1	2	3	4	5	6	7	8
8	2	1				3	800
9	1	3	1			5	865
10		1	3	1		5	962
11		1	1	1	2	4	1035
12			1		1	3	1059
fy	3	6	6	2	3	20	

Примечание: $\bar{y}_j$ - среднее значение результативного признака j-ой группы значений факторного признака;

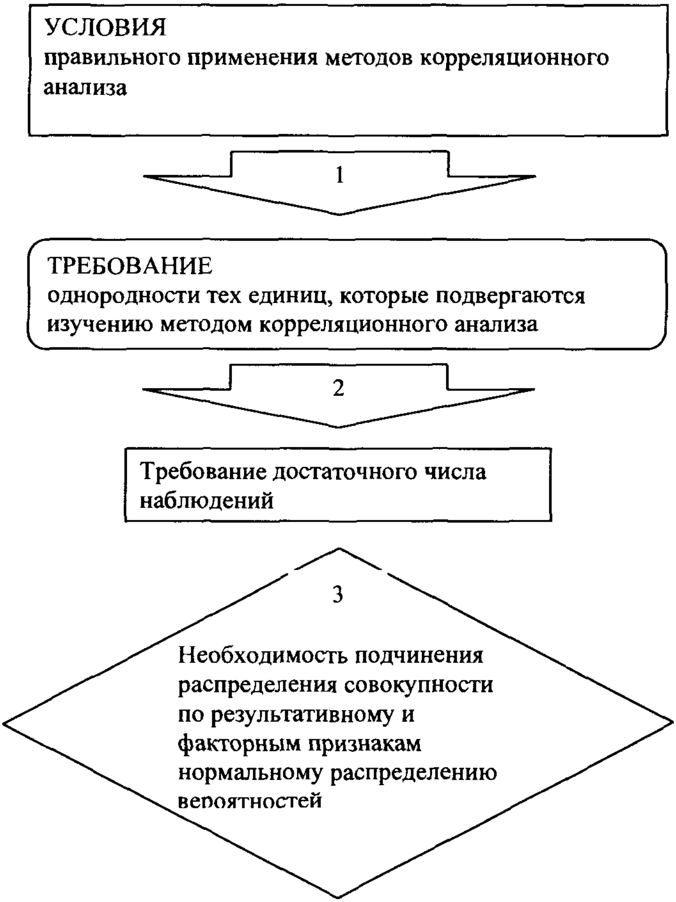
fx - частота повторения данного варианта значения факторного признака для всей совокупности;

fy - частота повторения результативного признака по всей совокупности.

В корреляционной таблице факторный признак X, как правило, располагают в строках, а результативный признак Y – в столбцах (графа X) таблицы.

Как видно из таблицы, по мере увеличения значений X групповые средние значения Y, т.е. $\bar{y}_i$ тоже увеличиваются от группы к группе, что позволяет сделать вывод о том, что между X и Y существует корреляционная связь.

СЛАЙД 8.4


УСЛОВИЯ
правильного применения методов корреляционного анализа
1
ТРЕБОВАНИЕ
однородности тех единиц, которые подвергаются изучению методом корреляционного анализа
2
Требование достаточного числа наблюдений
3
Необходимость подчинения распределения совокупности по результативному и факторным признакам нормальному распределению вероятностей

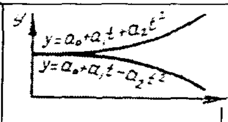
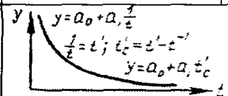
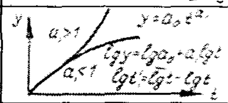
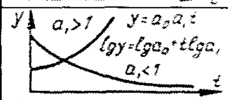
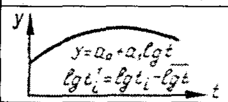
СЛАЙД 8.5

Этапы исследования корреляционных зависимостей

1	Предварительный анализ свойств моделируемой совокупности единиц
2	Установление факта наличия связи, определение ее направления и формы
3	Измерение степени тесноты связи между признаками
4	Построение регрессионной модели, т.е. нахождение аналитического выражения связи
5	Оценка адекватности модели, ее экономическая интерпретация и практическое использование

СЛАЙД 8.6

ВИДЫ УРАВНЕНИЯ ПРЯМОЛИНЕЙНОЙ И КРИВОЛИНЕЙНОЙ СВЯЗИ

№ п/п	Формы связи	Формулы связи	Параметры		
			a0	a1	a2
	А. Уравнения прямолинейной связи				
1	Прямая		$\frac{\sum y}{n}$	$\frac{\sum ty}{\sum t^2}$	-
	Б.Уравнения криволинейной связи				
2	Парабола 2-го порядка		$\frac{\sum y - a_2\sum t^2}{n}$	$\frac{\sum ty}{\sum t^2}$	$\frac{n\sum t^2y-\sum t^2\sum y}{n\sum t^4-\left(\sum t^2\right)^2}$
3	Гипербола		$\frac{-\sum y}{n}$	$\frac{\sum t'_c\, y}{\sum (t'_c)^2}$	-
4	Показательная функция		$\lg a_0 = \frac{\sum \lg y}{n}$	$\lg a_1 = \frac{\sum \lg t \lg y}{\sum (\lg t)^2}$	-
5	Логарифмическая функция		$\lg a_0 = \frac{\sum \lg y}{n}$	$\lg a_1 = \frac{\sum \lg t \lg y}{\sum t^2}$	-
6.	Полулогарифмическая функция		$\frac{\sum y}{n}$	$\frac{\sum y \lg t'}{\sum (\lg t')^2}$	-

Вычисление параметров осуществляется способом выравнивания эмпирических данных методом наименьших квадратов (МНК). В основу этого метода положено требование минимальности сумм квадратов отклонений эмпирических данных Yi от выровненных Yxi.

$$\sum\left(Y_i - Y_{xi}\right)^2 = \min.$$

СЛАЙД 8.7

ПАРНАЯ КОРРЕЛЯЦИЯ (1)

1. Уравнения прямой

$\bar{y} = a + bx$

2. Система двух линейных уравнений с двумя неизвестными

$$na + b\sum x = \sum y$$

$$a\sum x + b\sum x^2 = \sum xy$$

3. Определение параметров

$$\boxed{a = \frac{\sum y \sum x^2 - \sum xy \sum x}{n\sum x^2 - \sum x \sum x};} \quad \boxed{b = r\frac{\sigma_y}{\sigma_x};} \quad \boxed{b = \frac{n\sum xy - \sum y \sum x}{n\sum x^2 - \sum x \sum x}}$$

4. Оценка значимости коэффициента регрессии (b):

$$\boxed{t = \frac{b}{\mu_b};} \quad \boxed{\mu_b = \frac{\sigma_{ост}}{\sigma_x \sqrt{n-2}};} \quad \boxed{\sigma_{ост} = \sqrt{\frac{\sum (Y - \bar{Y})^2}{n}};} \quad \boxed{t = \frac{b \cdot \sigma_x \sqrt{n-2}}{\sigma_y \sqrt{1 - r^2}};}$$

• Для большого числа наблюдений «b» считается значимым, если: $t > 3$

• Для малого числа наблюдений, т.е. n < 30. Фактическое (расчетное) t сопоставляется с табличным (критическим) t критерием Стьюдента, определяемым для числа степеней свободы r = n-2 заданного уровня значимости.

$\alpha \cdot (0{,}05$ или $0{,}01)$

Если tфак>tтабл. ← то параметр считается значимым

СЛАЙД 8.8

ПАРНАЯ КОРРЕЛЯЦИЯ (2)

1. Коэффициент парной корреляции

$$R = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y}$$

$$\sigma_y = \sqrt{\overline{y^2} - (\overline{y})^2};$$

$$\sigma_x = \sqrt{\overline{x^2} - (\overline{x})^2};$$

2. Корреляционное отношение η

- $\sigma_y^2(\sigma^2) = \dfrac{\sum (y - \overline{y})^2 f}{\sum f}$ - дисперсия результативного признака
- $\sigma_{\overline{y}_x}^2 (\overline{\delta^2}) = \dfrac{\sum (\overline{y}_x - \overline{y})^2 f}{\sum f}$ - дисперсия признака-фактора (межгрупповая)
- $\sigma^2 (y - \overline{y}_x)(\overline{\sigma_i^2}) = \dfrac{\sum (y - \overline{y})^2 f}{\sum f}$ - остаточная дисперсия (внутригрупповая)

$$\eta = \sqrt{\frac{\sigma_{\overline{y}_x}^2}{\sigma_y^2}} \quad \text{или} \quad \eta = \frac{\sigma_{\overline{y}_x}}{\sigma_y}$$

3. Индекс корреляции

$$\eta = \sqrt{\frac{\sigma_{\overline{y}}^2 - \sigma^2 (y - \overline{y}_x)}{\sigma_y^2}} = \sqrt{1 - \frac{\delta^2 (y - \overline{y}_x)}{\delta^2{}_y}}$$

Примечание:

При прямолинейной связи r и η совпадают. Несовпадение этих показателей свидетельствует, что связь между изучаемыми признаками не прямолинейная, а криволинейная.

η имеет перед r такое важное преимущество, что оно может быть использовано для оценки тесноты связи любой формы зависимости (а не только прямолинейной)

СЛАЙД 8.9

ПАРНАЯ КОРРЕЛЯЦИЯ (3)

1. Критерий надежности r (средняя ошибка)

$$\boxed{\mu = \frac{|r|}{\sigma_r} \geq 3} \qquad \boxed{\sigma_r = \frac{1-r^2}{\sqrt{n}}} \qquad \boxed{\sigma_r = \frac{1-r^2}{\sqrt{n-1}}}$$

• Если число пар значений $\leq$ 100 и r не близок к единице

Величина r при достаточно большом числе наблюдений (пар значений) должна превышать среднюю ошибку μ в 2,5 - 3 раза

$$\boxed{\mu = \frac{|r|}{\sigma_r} \geq 3}$$

• Рассчитаем μ для зависимости часовой заработной платы от производственного стажа по 50 рабочим, т.е.

r=0,96 и n=50

$$\sigma_r = \frac{1-r^2}{\sqrt{n-1}} = \frac{1-0{,}96^2}{\sqrt{49}} = \frac{1-0{,}92}{7} = \frac{0{,}08}{7} \cong 0{,}01$$

$$\mu = \frac{|r|}{\sigma_r} = \frac{0{,}96}{0{,}01} \cong 96$$

• Здесь численное значение коэффициента корреляции (0,96) превосходит среднюю ошибку (σ_r) во много раз, откуда следует, что r не случаен, полученная по нему теснота связи весьма существенна

• Если

$$\frac{|r_1 - r_2|}{\mu_{r_1 - r_2}} \geq 3$$

то разность r1 и r2 считается существенной, а не случайной

СЛАЙД 8.10

МНОЖЕСТВЕННАЯ КОРРЕЛЯЦИЯ (1)

Уравнения прямой

$$\overline{Y}_{x,z} = a + bx + cz$$

Система трех линейных уравнений с двумя признаками-факторами

$$na + b\sum x + c\sum z = \sum y$$
$$a\sum x + b\sum x^2 + c\sum xz = \sum yx$$
$$a\sum z + b\sum xz + c\sum z^2 = \sum yz$$

Определение параметров

$$a = \overline{y} - b\overline{x} - c\overline{z}$$
$$b = \frac{r_{yx} - r_{yz} \cdot r_{xz}}{1 - r_{xz}^2} \cdot \frac{\sigma_y}{\sigma_x}$$
$$c = \frac{r_{yz} - r_{yx} \cdot r_{xz}}{1 - r_{xz}^2} \cdot \frac{\sigma_y}{\sigma_x}$$

Коэффициент множественной корреляции

$$R_{yxz} = \sqrt{\frac{r_{yx}^2 + r_{yz}^2 + 2r_{yx}^2 \cdot r_{yx} \cdot r_{xz}}{1 - r_{xz}^2}}$$

$$r_{yx} = \frac{\overline{yx} - \overline{y} \cdot \overline{x}}{\sigma_y \cdot \sigma_x}; \qquad r_{yz} = \frac{\overline{yz} - \overline{y} \cdot \overline{z}}{\sigma_y \cdot \sigma_x}; \qquad r_{xz} = \frac{\overline{xz} - \overline{x} \cdot \overline{z}}{\sigma_x \cdot \sigma_z};$$

Коэффициент парной корреляции (ryx или ryz) отражает влияние на результативный признак не только исследуемого признака фактора (X или Z), но и других не включенных в расчет факторов

СЛАЙД 8.11

МНОЖЕСТВЕННАЯ КОРРЕЛЯЦИЯ (2)

Индекс множественной детерминации (R2)

$$R^2_{y_1x_1,x_2..x_n} = 1 - \frac{\sigma^2_{фактор}}{\sigma^2_y}$$

$$R_{y_1x_1,x_2...x_n} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_y}$$

где $\delta^2_{фактор} = \frac{1}{n}\sum\left(\left(\overline{Y}_x\right)_i - \overline{Y}\right)^2$ - факторная дисперсия, характеризующая вариацию результативного признака, обусловленную вариацией включенных в анализ факторов;

σ^2_y - общая дисперсия результативного признака;

$\sigma^2_{ост} = \frac{1}{n}\sum\left(Y_i - \left(\overline{Y}\right)_i\right)^2 = \sigma^2_y - \sigma^2_{фактор}$ - остаточная дисперсия, характеризующая отклонения фактических уровней результативного признака Y_i от рассчитанных по уравнению множественной регрессии $\left(\overline{Y}_x\right)_i$.

R2 характеризует долю вариации результативного признака, обусловленную изменением всех факторов, входящих в уравнение множественной регрессии.

Индекс множественной корреляции (R)

$$R = \sqrt{\frac{\sigma^2_{y_x}}{\sigma^2_y}} \quad \text{или} \quad R = \frac{\sigma_{y_x}}{\sigma_y}$$

СЛАЙД 8.12

МНОЖЕСТВЕННАЯ КОРРЕЛЯЦИЯ (3)

Коэффициенты частной корреляции

$$z^{r_{yx}} = \frac{r_{yx} - r_{xz} \cdot r_{xy}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{xy}^2\right)}}$$

$$x^{r_{yz}} = \frac{r_{yz} - r_{xz} \cdot r_{xy}}{\sqrt{\left(1 - r_{yx}^2\right)\left(1 - r_{yz}^2\right)}}$$

$$y^{r_{xz}} = \frac{r_{xz} - r_{yx} \cdot r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{xy}^2\right)}}$$

Частный коэффициент корреляции – показатель, характеризующий тесноту связи между признаками при элиминации всех остальных признаков

Оценка значимости коэффициента множественной корреляции вытекает из оценки значимости коэффициента (индекса) множественной детерминации

Оценка значимости коэффициента (индекса) детерминации определяется с использованием критерия Фишера

$$F_{\text{расч.}} = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}$$

где F - критерий, который представляет собой соотношение оценок факторной и случайной дисперсий, рассчитанных на одну степень свободы

Число степеней свободы для факторной дисперсии $V1 = K$, для случайной $V2 = n - k - 1$

Если

$$F_{\text{расч.}} > F_{\text{табл}}$$

то связь признается существенной

СЛАЙД 8.13

НЕПАРАМЕТРИЧЕСКИЕ ПОКАЗАТЕЛИ ОЦЕНКИ СВЯЗИ (1)

Коэффициент корреляции рангов Спирмена

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

где di2 - квадраты разности рангов, связанных величин x и y
п - число наблюдаемых пар значений x и y

Коэффициент Фехнера

$$K_{\Phi} = \frac{\sum C - \sum H}{\sum C + \sum H}$$

Коэффициент ассоциации К. Пирсона

$$K_{acc} = \frac{Aa \cdot Bb - Ав \cdot Ba}{\sqrt{\sum A \sum B \sum a \sum b}}$$

Коэффициент ассоциации Д. Юла

$$K_{acc} = \frac{ad - bc}{ad + bc}$$

Коэффициент контингенции Дж. Кендэла

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

СЛАЙД 8.14

НЕПАРАМЕТРИЧЕСКИЕ ПОКАЗАТЕЛИ ОЦЕНКИ СВЯЗИ (2)

1. Коэффициент корреляции рангов Кендэла

$$r = \frac{S}{\frac{n(n-1)}{2}} \quad \text{или} \quad r = \frac{2S}{n(n-1)}$$

$S = P + Q$

где P - число «правильных» следований (+)

Q - число «неправильных» следований (–)

2. Коэффициент взаимной сопряженности К. Пирсона

$$\mathbf{P} = \sqrt{\frac{\chi^2}{n - \chi^2}} \quad \text{или} \quad \mathbf{C_1} = \sqrt{\frac{q^2}{1 - q^2}}$$

3. Коэффициент взаимной сопряженности А. Чупрова

$$\mathbf{C} = \sqrt{\frac{\chi^2}{n(k_1 - 1)(k_2 - 1)}} \quad \text{или} \quad \mathbf{K} = \frac{\varphi^2}{\sqrt{(k_1 - 1)(k_2 - 1)}}$$

4. Коэффициент конкордации М. Кендэля и Б. Смита

$$\mathbf{W} = \frac{12S}{m^2(n^3 - n)}$$

5. Коэффициент совпадения знаков (Ксовп)

$$\mathbf{K}_{совп} = \frac{a - b}{n}$$

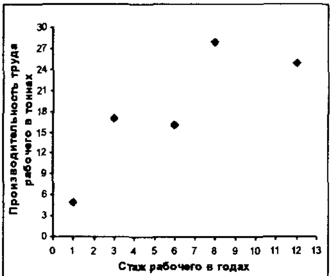
СЛАЙД 8.15

РЕШЕНИЕ ТИПОВОЙ ЗАДАЧИ (1)

1. Эмпирический ряд и необходимые расчеты по ним

Стаж (в годах) x	Произво-дительность труда (в т.), y	Стаж x произво-дительности труда, xy	Квадрат стажа, x2
1	7	7	1
3	17	51	9
6	16	96	36
8	29	232	64
12	26	312	144
30	95	698	254
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$

2. Корреляционное поле



Зависимость производительности труда от стажа работы

3. Находим на основе построения нормальных уравнений прямую линию, которая бы ближе других прямых подходила к эмпирическим данным:

$$\sum y = an + b\sum x;$$
$$\sum xy = a\sum x + b\sum x^2.$$

Решив эту систему уравнений, получим следующие значения параметров:

$$b = \frac{\sum xy - \bar{x}\cdot\bar{y}\cdot n}{\sum x^2 - (\bar{x})^2}; \qquad a = \bar{y} - \bar{в x}.$$

$$\bar{x} = \frac{30}{5} = 6 \text{ лет}; \qquad \bar{y} = \frac{95}{5} = 19 \text{ m}.$$

Подставляем в формулу полученные величины:

$$b = \frac{698 - 6\times19\times5}{254 - 6^2\times5} = \frac{128}{74} = 1{,}73 \text{ т}; \qquad a = 19 - 1{,}73\times6 = 8{,}62 \text{ т}.$$

Найдя параметры a и b, подставляем их значения в уравнение прямой линии: $\bar{y}_x = 8{,}62 + 1{,}73x$,

где $\bar{y}_x$ - теоретические значения уровня производительности труда.

Величина параметра а показывает начальный исходный уровень при x = 0. В данном примере это означает, что 8,62 т. – это уровень производительности труда новичка, только приступившего к практическому освоению своей профессии (стаж его в годах равен нулю).

Величина параметра b показывает, насколько возрастает производительность труда при увеличении стажа на 1 год.

Приведенное уравнение дает возможность установить, каков должен быть теоретический уровень производительности труда при заданном стаже. Каждый год стажа дает рост производительности труда на 1,73 т. Следовательно, при трехлетнем стаже вправе ожидать следующий уровня производительности труда: $\bar{y}_x = 8{,}62 + 1{,}73 \times 3 = 13{,}81$ т.

СЛАЙД 8.16

ЭМПИРИЧЕСКАЯ И ТЕОРЕТИЧЕСКАЯ ЛИНИЯ РЕГРЕССИИ (СВЯЗИ) У ПО Х

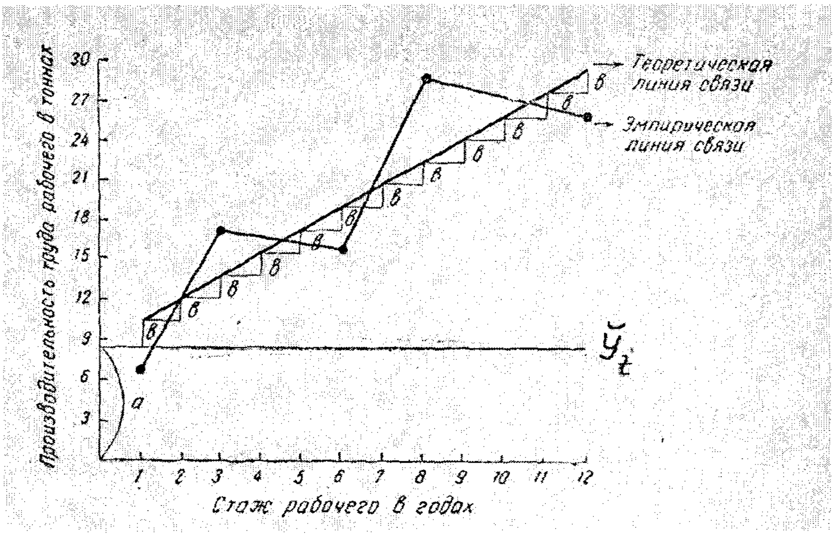


Рис. 8.1. **Теоретическая и эмпирическая линии связи стажа и производительности труда**

Установив теоретические значения $\bar{y}_x$, исчислим квадрат отклонений теоретического ряда динамики от эмпирического.

Стаж (в годах) x	Производительность труда (т)		Отклонения в уровне производительности труда эмпирического от теоретического $y-\bar{y}_x$	Квадраты отклонений $(y-\bar{y}_x)^2$
	эмпирический ряд, y	теоретический ряд, $\bar{y}_x$		
1	7	10,35	-3,35	11,2225
3	17	13,81	+3,19	10,1761
6	16	19,00	-3,00	9,0000
8	29	22,46	+6,54	42,7716
12	26	29,38	-3,38	11,4244
Итого	95	95,00	-	84,5946

Исчисленная сумма квадратов отклонений теоретического ряда от данного эмпирического ряда является минимальной при выравнивании по прямой

$$\sum\left(y-\bar{y}_x\right)^2=\min.$$

Минимизируются квадраты отклонений, поскольку $\sum(y-\bar{y}_x)=0$. Если данное требование соблюдается, легко определить, при каких значениях а0, а1 и т.д. Для каждой аналитической кривой эта сумма квадратов отклонений будет минимальной.

Найденное уравнение отражает закон связи между стажем и производительностью труда, закон, отвлекающийся от влияния прочих факторов, создающих отклонения в ту или иную сторону, и исходящий из принципа «при прочих равных условиях». Если уравнение функциональной зависимости выражает закон связи для каждого случая в отдельности, то уравнение корреляционной зависимости выражает закон связи для всех случаев, взятых вместе, для всей массы в целом, для всей изучаемой совокупности.

Глоссарий по теме 8

1. Корреляция (англ. «correlation» – означает соответствие) – это слово ввел в употребление в статистику английский биолог и статистик Фрэнсис Гальтон в конце XIX в. Тогда оно писалось как «correlation», но не просить «связь» (relation), а «как бы связь», т.е. связь, но не в привычной в то время функциональной форме. Прямое толкование термина корреляция статистическая, вероятная, возможная связь между двумя (парная) или несколькими (множественная) случайными величинами. В науке вообще, а именно, в палеонтологии, термин «корреляция» применен еще раньше, в конце XVIII века, знаменитый французский палеонтолог (специалист по ископаемым останкам животных и растений прошлых эпох) Жорж Кювье, ввел даже «закон корреляции» частей и органов животных.

2. Балансовая связь – характеризует зависимость между источником формирования ресурсов (средств) и их использованием. Свое проявление она получает, например, в формуле товарного баланса:

$$O_н + П = В + O_к,$$

где $O_н$ - остаток товаров на начало изучаемого периода; П - поступление товаров за период; В - выбытие товаров в изучаемом периоде; Ок - остаток товаров на конец года.

Левая часть формулы характеризует предложение товаров (Он+П), а правая часть – использование товарных ресурсов (В+Ок).

3. Компонентные связи – характеризуются тем, что изменение статистического показателя определяется изменением компонентов, входящих в этот показатель, как множители:

$$а = в \cdot с.$$

Компонентные связи используются в индексном методе выявления роли отдельных факторов в совокупном изменении сложного показателя. Так,

$$J_{pq} = J_p \cdot J_q.$$

4. Факторные связи – характеризуются тем, что они проявляются в согласованной вариации изучаемых показателей. При этом одни показатели выступают как факторные, а другие – как результативные.

По своему характеру этот вид связи является причинно-следственной (детерминированной) зависимостью. В свою очередь, факторные связи могут рассматриваться как функциональные и корреляционные.

5. Корреляционная связь – такая связь, когда изменению одного признака (у) на единицу соответствует изменение другого признака (х) на строго определению величину. Это тесно детерминированная связь:

$$\bar{y}_x = a + вx,$$

где у - результатный признак (напр., количество израсходованного топлива);

х - факторный признак (напр., объем продукции в тыс.сум);

а - параметры прямой (напр., расход топлива на непроизводственные нужды);

в - параметр прямой (напр., коэффициент, показывающий увеличение расхода топлива с ростом объема продукции на одну тыс.сум).

По своему характеру корреляционные связи – это связи соотносительные. Здесь помимо факторного признака влияют и другие факторы, в том числе и неучтенные «Σ». Поэтому корреляционные связи не являются полными зависимостями.

6. Корреляционная таблица – строится по типу «шахматной», т.е. в подлежащем таблицы выделяются группы по факторному признаку X, в сказуемом – по результативному признаку. «Общий вид такой таблицы показан на примере обратной зависимости между себестоимостью зерна и урожайностью зерновых по условным данным 80 хозяйств.

7. Корреляционное поле – корреляционное поле отражает не только общую зависимость между х и у, но и концентрацию индивидуальных точек вокруг линии регрессии показателя $\bar{y}_i$.

8. Эмпирическое корреляционное отклонение (η) – на основе аналитических группировок и корреляционных таблиц можно не только выявить наличие зависимости между двумя коррелируемыми показателями, но и измерить тесноту этой связи, в частности, с помощью эмпирического корреляционного отношения:

$$\eta = \sqrt{\frac{\delta^2}{\sigma_y^2}}.$$

Здесь δ^2 и σ_y^2, соответственно, межгрупповая и общая дисперсия результативного признака. Рассчитываются как:

$$\delta^2 = \frac{\sum_1^m \left(\bar{y}_j - \bar{y}\right)^2 f_j}{\sum f_j} \quad \text{и} \quad \sigma_y^2 = \frac{\sum_1^n \left(\bar{y}_i - \bar{y}\right)^2 f_i}{\sum f_i},$$

где m - число групп по факторному признаку x;

n - число единиц совокупности;

$\bar{y}_j$ - среднее значение результативного признака по группам;

$\bar{y}$ - общее среднее значение результативного признака;

$\bar{y}_i$ - индивидуальные значения результативного признака;

fj = fy - частота в i- й группе y;

fi = fx - частота в i- й группе x.

9. Эмпирический коэффициент детерминации – это квадрат корреляционного отношения

$$\eta^2 = \frac{\delta^2}{\sigma_y^2},$$

где δ^2 - межгрупповая дисперсия; σ^2- общая дисперсия. Общую дисперсию можно рассчитать и по формуле:

$$\sigma_y^2 = \overline{y^2} - \left(\bar{y}\right)^2$$

$$\overline{y^2} = \frac{\sum y_i^2 f_i}{\sum f_i} \quad \text{и} \quad \bar{y} = \frac{\sum y_i f_i}{\sum f_i}.$$

10. Эмпирическая линия регрессии – это ломаная линия, соединенная последовательно, нанесенными на ее основе фактических данных.

11. Коэффициент регрессии – это параметр «в» в уравнении линейной регрессии. Показывает, на сколько (в абсолютном выражении) результативный признак У изменяется при изменении факторного признака x на единицу:

$$в = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2}.$$

12. Свободный член уравнения (а) – вычисляется по формуле:

$$а = \bar{y} - \Sigma\, в_j\, \bar{x}_j$$

или

$$а = \bar{y} - в\, \bar{x}$$

13. Коэффициент условно-чистой регрессии – означает, что каждая из величин $в_j$ измеряется как среднее по совокупности отклонения результативного признака от его средней величины, при отклонении данного фактора x_j от своей средней величины на единицу его измерения и при условии, что все прочие факторы, входящие в уравнение регрессии, закреплены на средних значениях, не изменяются и не варьируют.

14. Стандартизованный коэффициент регрессии или коэффициент β –коэффициенты условно-чистой регрессии $в_j$ являются именными числами, выраженными в разных единицах измерения, и поэтому несравнимы друг с другом.

Для преобразования их в сравнимые относительные показатели применяется то же преобразование, что и для получения коэффициента парной корреляции. Полученную величину называют стандартизированным коэффициентом регрессии или β- коэффициентом.

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

β- коэффициент при факторе x_j, определяет меру влияния вариации фактора x_j на вариацию результативного признака У при отвлечении от сопутствующей вариации двух факторов, входящих в уравнение регрессии.

Коэффициенты условно-чистой регрессии полезно выразить в виде относительных сравнимых показателей связи, коэффициентом эластичности.

15. **Коэффициент эластичности** (Э) – показывает на сколько процентов изменяется в среднем результативный признак **У** при изменении факторного признака x_j на 1%. Обычно Э рассчитывают как отношение прироста (в %) результативного признака к приросту (в %) факторного признака:

$$э_x = a_1 \frac{\overline{x_j}}{\overline{y}} \quad \text{или} \quad \frac{\Delta x_j}{\overline{x_j}} : \frac{\Delta y}{\overline{y}}.$$

Зная линейный коэффициент корреляции, оценивающий степень тесноты связи между изменениями факторного и результативного признаков, можно определить коэффициент регрессии в уравнении:

$$\overline{y} = a_0 + a_1 \overline{x}$$

по формуле:

$$a_1 = r \cdot \frac{\sigma_y}{\sigma_x}; \qquad a_0 = \overline{y} - a_1 \overline{x},$$

где r - коэффициент корреляции.

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\left[\overline{x^2} - (\overline{x})^2\right]\left[\overline{y^2} - (\overline{y})^2\right]}}$$

или

$$r_{xy} = a_1 \cdot \frac{\sigma_y}{\sigma_x} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \cdot \sum (y_i - \overline{y})^2}}.$$

16. Коэффициент корреляции – этот показатель представляет собой стандартизированный коэффициент регрессии, т.е. коэффициент, выраженный не в абсолютных единицах изменения признаков, а в долях среднего квадратического отклонения результативного признака и является измерителем тесноты связи только при прямолинейной зависимости.

Коэффициент корреляции был предложен английским статистиком и философом Карлом Пирсоном (1857-1936). Его интерпретация такова: отклонение признака-фактора от его среднего значения и на величину своего среднего значения, на величину своего среднего квадратического отклонения $(x-\bar{x}) \le \sigma_x^2$, в среднем по совокупности приводит к отклонению признака-результата от своего среднего значения $(y-\bar{y})$ на R_{xy} его среднее квадратическое отклонение σ_y.

В отличие от коэффициента регрессии коэффициент корреляции не зависит от принятых единиц измерения признаков, а стало быть, он сравним для любых признаков.

17. Индекс детерминации (причинности) – выражает долю факторной дисперсии в общей дисперсии, т.е. характеризует какая часть общей вариации результативного признака у объясняется изучаемым фактором x:

$$R^2 = \frac{\sigma_{y_x}^2}{\sigma_y^2}.$$

Соотношение между факторной $\sigma_{y_x}^2$

т.е. $\sigma_{y_x}^2 = \frac{\sum (y_{x_i} - \bar{y})^2}{n}$

и общей δ_y^2, $т.е.\ \sigma_y^2 = \frac{\sum (y_i - \bar{y})^2}{n}$

дисперсиями характеризует меру тесноты связи между признаками x и y.

18. Индекс корреляции (R) – на основе формулы индекса детерминации определяется индекс корреляции R:

$$R = \sqrt{\frac{\sigma_{y_x}^2}{\sigma_y^2}} \quad \text{или} \quad R = \frac{\sigma_{\bar{y}}}{\sigma_{\bar{y}}}.$$

В отличие от коэффициента корреляции этот индекс пригоден для измерения тесноты связи любой формы зависимости.

19. «Правило сложения дисперсий» – позволяет исчислить независимую дисперсию, если известны любые две из трех, оно даёт возможность определять долю общей дисперсии, складывающуюся под влиянием признака, положенного в основу группировки.

$$\delta_y^2 = \delta_{yx}^2 + \delta_{ост}^2.$$

Отсюда получают формулу индекса корреляции:

$$R = \sqrt{\frac{\delta_y^2 - \delta_{ост}^2}{\delta_y^2}} = \sqrt{1 - \frac{\delta_{ост}^2}{\delta_y^2}},$$

где $\delta_{ост}^2$ - остаточная дисперсия. Его можно исчислить и по формуле:

$$R = \frac{\delta_{\bar{y}}}{\delta}.$$

20. Средняя ошибка коэффициента корреляции (δ_r) – интерпретируя значение коэффициента корреляции, следует иметь в виду, что как любой выборочный показатель, он содержит случайную ошибку и не всегда однозначно отражает действительно реальную связь между изучаемыми показателями.

Для того чтобы оценить существенность (значимость) самого r и соответственно, реальность измеряемой связи между Х и У, необходимо рассчитать среднюю ошибку коэффициента корреляции δ_r.

Особенность расчета этого критерия в зависимости от числа наблюдений (объема выборки) – п заключается в следующем:

Если число наблюдений достаточно велико (п>50) и есть основания полагать, что выборка осуществлена из нормальной совокупности, то δ_r рассчитывается по следующей приближенной формуле:

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

Обычно при большом п, если коэффициент корреляции r превышает свою среднюю ошибку δ_r больше чем в 3 раза т.е.:

$$\frac{|r|}{\delta_r} > 3$$

он считается значимым (существенным), а связь реальной.

Задавшись определенной вероятностью, можно определить доверительные, границы r. Так, например, при вероятности 0,997, для которой коэффициент доверия t=3, доверительные границы r составят:

$$r \pm 3\frac{1 - r^2}{\sqrt{n}} = r \pm 3\delta_r.$$

При небольшом числе наблюдений (n<30) средняя ошибка линейного коэффициента корреляции определяется по формуле:

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

21. Оценка значимости коэффициента корреляции (r) – проверяется на основе t-критерия Стьюдента.

22. t- критерий Стьюдента – при этом выдвигается и проверяется нулевая гипотеза о равенстве коэффициента корреляции нулю, т.е. об отсутствии связи между Х и У в генеральной совокупности.

23. Расчетные значения критерия ($t_{расч}$) – определяется по формуле:

$$t_{расч} = \frac{|r|}{\delta_r} = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

и сопоставляются с $t_{табл}$.

Если нулевая гипотеза верна, т.е. r = 0, то распределение t-критерия подчиняется закону Стьюдента (с заданными параметрами: уровнем значимости α, принимаемым обычно за 0,05 и числом степеней свободы γ = n–2).

Поэтому в каждом конкретном случае по таблице распределения t – критерия Стьюдента находится такое критическое значение t, которое допустимо при справидливости нулевой гипотезы, и с ним сравнивается фактическое (расчетное) значение t.

Если $t_{расч} > t_{табл}$, то нулевая гипотеза отвергается и линейный коэффициент считается значимым, а связь между х и у – существенной. Если $t_{расч} < t_{табл}$, то нулевая гипотеза не отвергается и коэффициент корреляции считается незначимым, т.е. считается, что связь между Х и Y отсутствует, и значение r, отличное от нуля, получено случайно.

Допустим, n = 8, r = – 0,95 средняя ошибка коэффициента корреляции равна:

$$\delta_r = \frac{\sqrt{1-r^2}}{\sqrt{n-2}} = \frac{\sqrt{1-0{,}95}}{\sqrt{8-2}} = 0{,}13.$$

Отсюда $t_{расч} = \frac{|r|}{\delta_r} = \frac{0{,}95}{0{,}13} = 7{,}3$.

По таблице Стьюдента находится t-табл (при α=0,05 и числе степеней свободы v = n – 2 = 6)

$$t_{табл} = 2{,}4469.$$

Так полученное tрасч=7,3 больше tтабл =2,4469, то нулевая гипотеза об отсутствии связи между х и у в генеральной совокупности отвергается, т.е. мы делаем вывод, что коэффициент корреляции значим и существенно отличается от нуля, подтверждая, тем самым, реальную связь между x и y.

24. F-критерий Фишера – применяется для оценки значимости индекса корреляции (R). Фактическое значение критерия FR – определятся по формуле:

$$F_R = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m-1},$$

где m - число параметров уравнения регрессии. Величина FR сравнивается с критическим значением Fк, которое определяется по таблице F-критерия с учетом принятого уровня значимости x и числа степеней свободы R1=m-1 и R1=n-m.

Если FR>Fk , то величина индекса корреляции признается существенной.

25. Таблица интеграла вероятностей Лапласа – применяется для оценки значимости связи в совокупностях достаточно большого объема:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int \frac{1}{2} t^2.$$

При уровне значимости α=0,05 табличная величина t=2.

По значению показателя тесноты связи можно посредством t-критерия произвести оценку значимости коэффициента регрессии a1

$$t_{a_1} = \frac{a_1 \sigma_x \sqrt{n-2}}{\sigma_y \sqrt{1-R^2}}.$$

Сравнивая исчисленное по этой формуле значение ta1 с табличным tk получают заключение о существенности связи коэффициента регрессии (a1).

Полученные в анализе корреляционной связи параметры уравнения регрессии признаются типичными, если t критического

$$t_{a_o} > t_k < t_{a_1}.$$

26. Шкала Чеддока – для получения выводов о практической значимости синтезированных в анализе моделей, показаниям тесноты связи даётся качественная оценка. Это осуществляется на основе шкалы Чеддока.

Показания тесноты связи	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-0,99
Характеристика силы связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

При значениях показателей тесноты связи превышающих 0,7 зависимость результативного признака Y является высокой, а при значениях более 0,9 – весьма высокой. Это, в соответствии с показаниями индекса детерминации R2, означает, что более половины объема вариации результативного признака У объясняется влиянием изучаемого фактора X.

27. Множественная корреляция – при показаниях тесноты связи ниже 0,7 величина индекса детерминации R2 всегда будет меньше 50%. Это означает, что на долю вариации факторного признака X приходится меньшая часть по сравнению с прочими признаками, влияющими на изменение общей дисперсии результативного признака. Это способ изучения зависимости результативного признака от ряда признаков – факторов.

28. Линейное уравнение множественной корреляции – данное уравнение зависимости результативного признака у от двух факторных (x, z) можно записать так:

$$\overline{Y}_{xz} = a + вx + cz.$$

Если признаков факторов три (x,z,l), то

$$\overline{Y}_{xzl} = a + вx + cz + kl \text{ и т.д.,}$$

где a, в, c, k, … - являются параметрами уравнений.

29. Коэффициент множественной корреляции – применительно к влиянию на Y двух факторов (X,Z), этот коэффициент будет следующим:

$$R_{yxz} = \sqrt{\frac{R_{yx}^2 + R_{yz}^2 + 2R_{yx} \cdot R_{yz} \cdot R_{xz}}{1 - R_{xz}^2}},$$

где

$$R_{yx} = \frac{\overline{yx} - \overline{y} \cdot \overline{x}}{\sigma_y \cdot \sigma_x}; \quad \mathrm{R}_{\mathrm{yz}} = \frac{\overline{yz} \cdot \overline{y} \cdot \overline{z}}{\sigma_x \cdot \sigma_z};$$

$$\mathrm{R}_{\mathrm{xz}} = \frac{\overline{\mathrm{xz}} - \overline{x} \cdot \overline{z}}{\sigma_x \cdot \sigma_z}.$$

Коэффициент парной корреляции (Ryx или Ryz) отражает влияние на результативный признак не только исследуемого признака-фактора (X или Z), но и на другие не включенные в расчет факторы. Поэтому при корреляционном анализе большую роль играют коэффициенты так называемой частной корреляции.

30. Коэффициент частной корреляции – это показатель тесноты связи между двумя признаками, исчисленный в условиях неизменности всех остальных факторов. Порядок вычисления коэффициентов частной корреляции состоит в том, что последовательно устраняется влияние каждого из рассмотренных факторов одного за другим. Формула коэффициента частной корреляции результативного признака Y с факторным признаком X при исключении влияния факторного признака Z будет такова:

$$ {}_{z}R_{yx} = \frac{R_{yx} - R_{xz} \cdot R_{xy}}{\sqrt{\left(1 - R_{xz}^{2}\right)\left(1 - R_{zy}^{2}\right)}}. $$

То же – зависимости Y от Z при исключении влияния X:

$$ {}_{x}R_{yz} = \frac{R_{yz} - R_{xz} \cdot R_{xy}}{\sqrt{\left(1 - R_{xz}^{2}\right)\left(1 - R_{xy}^{2}\right)}}. $$

Можно рассчитать взаимосвязь факторных признаков при устранении влияния результативного признака:

$$ {}_{y}R_{xz} = \frac{R_{xz} - R_{yx} \cdot R_{yz}}{\sqrt{\left(1 - R_{yx}^{2}\right)\left(1 - R_{yz}^{2}\right)}}. $$

31. Коллинеарность – совокупный коэффициент множественной корреляции зависит не только от корреляции результативного признака с факторными, но и от корреляции факторных признаков между собой. Наличие между двумя факторами весьма тесной линейной связи (парный коэффициент корреляции Rji превышает по абсолютной величине 0,8) называется коллинеарностью.

32. Мультиколлиарность – характеризует весьма тесные связи не 0,8 между несколькими факторами, т.е. это линейная зависимость между аргументами множественной регрессии. Если x_1, x_2, ... x_n – аргументы, то при строгой мультиколлиарности, существуют такие не все равные нулю a_1, a_2, ...a_n, что

$$ a_1x_1+a_2x_2 + ... +a_nx_n= 0. $$

Проблема отбора факторных признаков и проблема мультиколлиарности могут быть решены на основе многомерных статистических методов анализа (например, с помощью пошаговой регрессии).

33. Показатель средний ошибки аппроксимации ($\bar{\varepsilon}$) – применяется для оценки адекватности уравнения регрессии (проф. И.П. Суслов).

$$ \bar{\varepsilon} = \frac{1}{n}\sum\frac{|Y_i - Y_{xi}|}{Y_i}\cdot 100, $$

где Yi - Yxi - линейные отклонения абсолютных величин эмпирических и выровненных точек регрессии.

«Лучший, т.е. более адекватной считается та модель, в которой меньше среднее отклонение теоретических уровней от эмпирических» (И.П. Суслов).

34. Параболическая корреляция – применяется для изучения нелинейных связей. Если при линейной связи среднее изменение результативного признака на единицу фактора постоянно по всей области вариации фактора, то при параболической корреляции изменение признака X на единицу признака Y меняется равномерно с изменением величины фактора.

В этом случае зависимость между коррелируемыми величинами может быть выражена в виде параболы 2-го порядка:

$$\overline{Y}_x = a_0 + a_1 x + a_2 x^2.$$

Параметры которой, находят путем решения системы нормальных уравнений:

$$\begin{cases} na_0 + a_1\sum x + a_2\sum x^2 = \sum y \\ a_0\sum x + a_1\sum x^2 + a_2\sum x^3 = \sum xy \\ a_0\sum x^2 + a_1\sum x^3 + a_2\sum x^4 = \sum x^2 y \end{cases}.$$

35. Гиперболическая корреляция – применяется при наличии обратно пропорциональной зависимости и выравнивание производится по гиперболе, уравнение которой:

$$\overline{Y}_x = a_0 + a_1\frac{1}{x}.$$

Параметры гиперболы a0 и a1 находятся на основе следующей системы уравнений по способу наименьших квадратов:

$$\begin{cases} na_0 + a_1\sum\frac{1}{x} = \sum y \\ a_0\sum\frac{1}{x} + a_1\sum\frac{1}{x^2} = \sum\frac{y}{x} \end{cases}.$$

Следовательно, для отыскания параметров нужно по сравнению с выравниванием по прямой линии вместо $\sum x$ найти $\sum\frac{1}{x}$, а вместо $\sum x^2$ найти $\sum\frac{1}{x^2}$ и вместо $\sum xy$ найти $\sum\frac{y}{x}$.

36. Сопряженные уравнения – часто зависимость между коррелируемыми показателями X и Y такова, что каждый из них можно рассматривать в качестве и факторного, и результативного признака.

Такими показателями, например, могут быть производительность и оплата труда.

Если первый показатель обозначить X, а второй – Y, то уравнение регрессии можно записать и как Y по X, т.е. $\overline{Y}_x$ и как X по Y, т.е. $\overline{X}_y$.

В случае линейной зависимости это будет, соответственно:

$$\bar{Y}_x = a_0 + a_1 x;$$
$$\bar{X}_y = a_0 + a_1 Y.$$

Такие уравнения называются сопряженными. При линейной зависимости коэффициент корреляции в обоих случаях будет одинаков:

$$R_{y|x} = R_{x|y},$$

но параметры сопряженных уравнений, естественно, разные. Используя следующую формулу, легко определить коэффициенты регрессии a_1 и a_1^1 для сопряженных уравнений на основе линейного коэффициента корреляции R:

$$\mathbf{a_{1(y|x)} = R\frac{\sigma_y}{\sigma_x}}$$

и $$a^1_{1(x|y)} = R\frac{\sigma x}{\sigma y}$$

37. Коэффициент корреляции рангов – основан на корреляции не самих значений коррелируемых признаков, а их рангов.

38. Ранг – это порядковый номер, присваиваемый каждому индивидуальному значению X и Y (отдельно) в ранжированному ряду. Оба признака необходимо ранжировать (нумеровать) в одном и том же порядке: от меньших значений к большим и наоборот. Чаще нумерация (присвоение ранга) от 1 до n идет по возрастанию значений признака.

Если встречается нескольких одинаковых значений X (или Y), то каждому из них присваивается ранг, равный частному от деления суммы рангов (мест в ряду), приходящихся на эти значения, на число равных значений.

39. Коэффициент корреляции рангов Спирмена – был использован им в начале XX в. Основан на рассмотрении разности рангов значений факторного и результативного признаков:

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

где d_i^2 - квадраты разности рангов, связанных величин X и Y;

n - число наблюдаемых пар значений X и Y.

Данный коэффициент может принимать значения от 0 до ± 1. Когда ранги двух признаков полностью совпадают, т.е. каждое значение Nx =Ny, то $\sum d^2 = 0$. Соответственно, P=1, что характеризует максимально тесную прямую связь. Если ранги двух признаков имеют строго противоположное направление, т.е. первому рангу X соответствует n-й (последний) ранг Y, второму – (n-1) – й ранг Y и т.д., то в этом случае максимальная величина $\sum d^2$ равна $\frac{n(n^2-1)}{3}$ и, следовательно, $\frac{6\sum d^2}{n(n^2-1)}$ может иметь максимальное значение 2. Тогда по формуле Спирмена Р= –1, что характеризует полную (максимально тесную) обратную связь между изменениями значений X и Y. Если же связь между X и Y отсутствует, то, очевидно, должно соблюдаться равенство:

$$\sum d^2 = \frac{n(n^2-1)}{6},$$

и тогда Р=0.

40. Коэффициент контингенции Дж. Кендэла (Кконт) – принимается в тех случаях, когда хотя бы один из четырех показателей в «таблице четырех полей» отсутствует, величина коэффициента ассоциации будет равна единице:

$$K_{конт} = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Коэффициент контингенции всегда меньше коэффициента ассоциации. Связь считается достаточно значительной и подтвержденной если

$$|K_{acc}| > 0{,}5 \text{ или } |\text{К}_{\text{конт}}| > 0{,}3.$$

41. Коэффициент конкордации М. Кендэля и Б. Смита – используется тогда, когда число ранжируемых признаков (факторов) больше двух. Поэтому этот коэффициент называют также множественным коэффициентом корреляции:

$$W = \frac{12s}{m^2(n^3-n)},$$

где s - сумма квадратов отклонений суммы m рангов от их средней величины;

m - число ранжируемых признаков;

n - число ранжируемых единиц (число наблюдений).

Это формула применяется для случая, когда ранги по каждому признаку не повторяются. Если же есть связанные ранги, то коэффициент конкордации рассчитывается с учетом числа таких повторяющихся (связанных) рангов по каждому фактору:

$$W = \frac{12S}{m^2(n^3-n) - m\sum_{1}^{m}(t^3-t)},$$

где t - число одинаковых рангов по каждому признаку.

42. Коэффициент Фехнера (Кф) (коэффициент корреляции знаков)

– это мера тесноты связи в виде отношения разности числа пар совпадающих и несовпадающих пар знаков к сумме этих чисел:

$$K_ф = \frac{\sum c - \sum H}{\sum c + \sum H}.$$

Он основан на сравнении поведения отклонений индивидуальных значений каждого признака (X и Y) от своей средней величины. При этом во внимание принимаются не величины отклонений $(X_i - \overline{X})$ и $(Y_i - \overline{Y})$, а их знаки («+» или «-»). Определив знаки отклонения от средней величины в каждом ряду, рассматривают все пары знаков и подсчитывают число их совпадений и несовпадений.

Кроме того неточность показателя Фехнера заключается в том, что при определении размера Кф разные по абсолютной величине отклонения имеют равный вес. Поэтому данный показатель, несмотря на простоту вычисления, применяется редко.

43. Автокорреляция – это корреляция между уровнями ряда или отклонениями от тренда, взятыми со сдвигом во времени: на 1 период (год), на 2, на 3 и т.д., поэтому говорят о коэффициентах корреляции разных порядков: первого, второго и т.д. (см. IX главу «Динамические ряды»).

Тематика самостоятельной работы по теме 8 с представлением презентации

1. Линейный корреляционный и регрессионный анализ двух переменных.
2. Нелинейная парная корреляция и регрессия.
3. Применение многофакторного корреляционно-регрессионного анализа для изучения производительности труда.
4. Экономическая интерпретация результатов регрессионного анализа.
5. Статистические методы выражения зависимости между экономическими явлениями.
6. Методы изучения существенности связи критерий Стьюдента и Р.Фишера.
7. Производственные функции.
8. Методы изучения связи качественных признаков.
9. Использование результатов проверки по критерию Хи-квадрат.
10. Ранговая корреляция.
11. Параметрические и непараметрические методы определения тесноты связи.
12. О ложной корреляции.

ЛИТЕРАТУРА

1. Бородкин Ф.М. Статистическая оценка связей экономических показателей. М.: Статистика, 1968.
2. Вайну Я.Я. Корреляция рядов динамики. М.: Статистика, 1977.
3. Дюран Б., Одел П. Кластерный анализ. М.: Статистика, 1977.
4. Езикиел М., Факс К. Методы анализа корреляции и регрессии. М.: Статистика, 1966.
5. Елисеева И.И. Статистические методы измерения связей. Л.: Ленинградский университет, 1982.
6. Кендэл М. Ранговые корреляции. М.: Статистика, 1975.
7. Ковалева Л.Н. Многофакторное прогнозирование на основе рядов динамики. М.: Статистика, 1980.
8. Кайкин В.П., Найденов В.С., Галуза С.Г. Корреляция и статистическое моделирование в экономических расчетах. М.: Экономика, 1964.
9. Крастинов О.П. Разработка и интерпретация моделей корреляционных связей в экономике. Рига: Занятие, 1983.
10. Льюис К.Д. Методы прогнозирования экономических показателей. М.: Финансы и статистика. 1986.
11. Лукомский Я.И. Теория корреляции и ее приминение к анализу производства. М.: Госстатиздат, 1958.
12. Окунь Д. Факторный анализ. М.: Статистика, 1974.
13. Статистические методы исследования корреляции в экономике. М.: Статистика, 1972.
14. Сиськов В.И. Корреляционный анализ в экономических исследованиях. М.: Статистика, 1975.
15. Урланис Б.Ц. Статистические методы изучения зависимости явлений. М.: Госстатиздат, 1959.
16. Чупров А.А. Основные проблемы теории корреляции. М.: Госстатиздат, 1960.
17. Четвериков Н.С. О ложной корреляции. В кн. Применение методов корреляции в экономических исследованиях. М.: Наука, 1969.
18. Фестер Э., Ренц Б. Методы корреляционного и регрессионного анализа. М.: Финансы и статистика, 1983.

ДЛЯ ЗАМЕТОК

Ёркин Абдуллаев,
Зоир Тошматов, Махмуджан Икрамов

КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Учебное пособие

Редактор **Э. Хуснутдинова**
Художник **К. Бойхужаев**
Компьютерная верстка **З. Улугбекова**

Лиц. изд. AI № 305. Подписано в печать 22.02.2021.
Формат 60x84 1/16. Усл.печ.л. 6,7. Уч.-изд.л. 6,9.
Тираж 100 экз. Заказ №. 8.

Издательство «IQTISOD-MOLIYA».
100000, Ташкент, ул. Амира Темура, 60А.

Отпечатано в типографии
«DAVR MATBUOT SAVDO» ООО.
100198, Ташкент, Куйлюк, массив 4, 46.