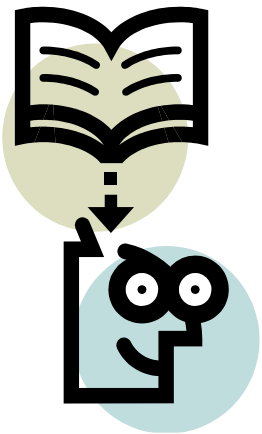
Министерство науки и высшего образования Российской Федерации

федеральное государственное бюджетное образовательное учреждение
высшего образования
«Мичуринский государственный аграрный университет»

В.Б. ПОПОВА

СТАТИСТИЧЕСКИЙ АНАЛИЗ ЭКОНОМИЧЕСКИХ ДАННЫХ

Учебное пособие



МИЧУРИНСК – НАУКОГРАД, 2023

УДК 311:631.1
ББК65.051.532

Рецензенты:
Черемисина Н.В., доктор экономических наук, профессор,
зав. кафедрой бухгалтерского учета и налогового контроля
ФГБОУ ВО «Тамбовский государственный университет имени Г.Р. Державина».
Карамнова Н.В., доктор экономических наук, профессор, зав. кафедрой управления и делового администрирования ФГБОУ ВО Мичуринский ГАУ.

Попова В.Б.
Статистический анализ экономических данных: учебное пособие/В.Б. Попова – Мичуринск: Изд-во Мичуринского ГАУ, 2022. – 120 с.

В учебном пособии представлены сущность и области применения наиболее распространенных методов статистического анализа и прогнозирования экономических данных с применением пакетов прикладных программ. Рассмотрены процедуры статистических вычислений по изучению распределения, динамики, взаимосвязи; проверке статистических гипотез; применению экстраполяции и адаптивных методов прогнозирования.

Рекомендуется к применению в учебном процессе при изучении дисциплин (модулей) «Статистика», «Статистический анализ и прогнозирование с использованием пакетов прикладных программ», «Статистический анализ и прогнозирование с использованием пакетов прикладных программ». Может использоваться обучающимися, преподавателями, региональными органами управления для обработки массовых данных при решении наиболее востребованных экономических задач.

УДК 311:631.1
ББК65.051.532

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Статистические методы представляют собой важнейший аналитический инструмент научного исследования, широко используемый при изучении разнообразных массовых процессов и явлений: социальных, экономических, демографических, исторических, биологических, технических, медицинских и др. Они позволяют анализировать состояние и взаимосвязи изучаемых явлений, выявлять закономерности и тенденции их развития.

Всесторонний качественный статистический анализ обеспечивает получение научно-обоснованных, логически выверенных результатов исследования.

В учебном пособии рассмотрены особенности статистического анализа экономических данных, уделено внимание обоснованию вероятностного характера статистического вывода, представлены традиционные методы статистического исследования (основы дисперсионного, корреляционного, регрессионного анализа, анализа временных рядов), даны понятия непараметрической статистики.

При изложении материала термин «*статистические методы*» и термин «*методы статистического анализа данных*» авторами используются как равноправные. При том, что «технологически» математическая статистика и анализ данных практически не различаются, следует все же указать, что оба эти подхода к анализу базируются на различных моделях получения данных и, соответственно, определяют различие в подходах к интерпретации полученных результатов. Однако при обработке экономической информации оправданы оба подхода.

В основе *математико-статистического* подхода – вероятностная модель, предполагающая, что имеющаяся статистическая совокупность представляет собой выборку из некоторой генеральной совокупности, на которую и должны распространяться полученные выводы. Эта модель довольно хорошо соответствует данным, действительно представляющим собой

выборочные совокупности. Например, бюджеты домохозяйств определенного населенного пункта представляют собой лишь часть более обширной совокупности бюджетов домохозяйств данного региона, а личные карточки работников некоторого предприятия – часть личных карточек работников некоторой отрасли и т.п. Изучая такие совокупности, исследователь действительно стремится расширить свои выводы на весь регион или отрасль, и в этом случае оправдан вероятностный подход.

Иная модель лежит в основе *анализа данных*: не предполагается, что изучаемая информация получена из более обширной генеральной совокупности, и полученные выводы интерпретируются без какого-либо расширительного толкования. Например, если изучается экономическая структура городского населения, и исследователь располагает соответствующими данными по всем городам, нет смысла расширять полученные результаты на генеральную совокупность, так как она совпадает с изучаемой совокупностью. Хотя в рамках вероятностного подхода сформированная совокупность может быть рассмотрена лишь как один из возможных случайных результатов реализации процесса.

Следует отметить, что проведение исследований в экономической сфере сопряжено с большими затратами финансовых средств, затрачиваемых на сбор и обработку данных, подготовку и оплату кадров, материально– технических ресурсов, трудовых ресурсов, привлекаемых к обследованию на всех его этапах, времени, затрачиваемого как на получение первичной информации, так и на последующую её обработку. Это предопределяет в большинстве случае интерпретацию результатов исследования на основе метода статистического вывода, который позволяет по данным выборок делать заключение о большей совокупности, по которой не имеется исчерпывающих наблюдений.

1. ОСОБЕННОСТИ СТАТИСТИЧЕСКОГО АНАЛИЗА ЭКОНОМИЧЕСКИХ ДАННЫХ

1.1 Экономические и статистические принципы анализа данных

Статистический анализ экономических данных основывается на применении статистических и статистико-математических методов с целью адекватного отражения исследуемых явлений и процессов и выявления действующих в них закономерностей.

В качестве этапов статистического анализа выделяются:

- формулировка цели анализа;
- критическая оценка данных;
- сравнительная оценка и обеспечение сопоставимости данных;
- формирование обобщающих показателей;
- фиксация и обоснование существенных свойств, особенностей, сходств и различий, связей и закономерностей изучаемых явлений и процессов;
- формулировка заключений, выводов и практических предложений о резервах и перспективах развития изучаемого явления [16].

Выбор статистических методов зависит от задач исследования, от характера изучаемых процессов, их специфики, особенностей и форм проявления.

Статистический анализ экономических данных проводится в неразрывной связи теоретического, качественного анализа сущности исследуемых явлений и процессов и соответствующего количественного инструментария изучения их структуры, связей и динамики. Основные трудности, связанные с применением количественных статистико-математических методов, заключаются в том, что они достаточно нейтральны к исследуемым экономическим процессам. Поэтому важным этапом статистического исследования на информационной базе, характеризующей деятельность реальных субъектов экономического процесса, является

критическая оценка исходных данных с точки зрения их достоверности и научной обоснованности.

Под критической оценкой статистического материала следует понимать его соответствие целям и задачам исследования.

Надежность выводов и заключений по статистическому анализу экономических данных обеспечивается соответствием исходной информации следующим параметрам качества:

–*целостность*– соблюдение научно-обоснованных методик сбора, обработки и распространения данных;

– *востребованность (релевалентность)* – способность информации соответствовать потребностям пользователей; степень практической применимости результата;

– *достоверность*– степень объективного отображения данными сущности изучаемых явлений и процессов; соблюдение научных правил сбора и обработки информации;

–*точность*– степень соответствия значения показателя, полученного по материалам наблюдения действительной его величине;

–*своевременность*– поступление данных в сроки, соответствующие целям проводимого наблюдения;

– *доступность*– состояние готовности данных к официальному распространению и информированность пользователей о возможности и средствах получения интересующих их данных;

– *интерпретируемость*– строгость используемых в публикациях статистических терминов;

– *согласованность*– степень полноты данных и логической взаимосвязи между результатами различных обследований;

– *сопоставимость*– единообразие данных во времени или в пространстве по тем или иным параметрам;

– *конфиденциальность* – необходимость предотвращения утечки (разглашения) информации;

–*объективность* – независимость от чьего-либо мнения или сознания, а также от методов получения;

–д*оступность* – мера возможности получить ту или иную информацию;

–*оперативность* – быстрота сбора и передачи данных, способность информации отражать происходящие изменения в изучаемом явлении или процессе.

При проведении экономико-статистического анализа следует учитывать экономические и статистические принципы его проведения.

Экономика представляет собой производственную систему, которая контролируется и управляется хозяйствующим субъектом, в которой труд и капитал участвуют в преобразовании одних товаров и услуг для создания других товаров и услуг. Функционирование и развитие данной системы осуществляется с учетом определенных принципов:

- *планомерность*- проектирование желаемых показателей деятельности и эффективных путей их достижения;

- *целенаправленность* – постановка задач исходя из объективных потребностей экономики на основе выбора некоторой совокупности средств их достижения;

- *системность* – планирование деятельности производственного объекта с учетом его места в экономической системе, его связей с другими объектами и подсистемами;

-*комплексность* – рассмотрение мероприятий в учетом всех затрагиваемых ими сфер, видов используемых ресурсов, их последствий в экономической, технологической, социальной, экологической и др. областях;

-*эффективность* – соотношение полученного результата с использованными ресурсами или затратами;

- *альтернативность* – рассмотрение множества возможных вариантов достижения поставленных целей;

- *оптимальность* – выбор направлений развития, обеспечивающих максимальную эффективность функционирования системы;

- *иерархичность* – наличие горизонтальных и вертикальных связей, обеспечивающих взаимодействие составных частей системы, обладающих собственными интересами и механизмом принятия решений;

- *динамичность* – изменение во времени потребностей в товарах и услугах, характеристик, условий и режимов функционирования производственных объектов, относительной и абсолютной ценности ресурсов и продукции и т.п.;

- *инерционность* – тенденция к сохранению длительных хозяйственных и технологических связей, традиционного ассортимента выпускаемой и потребляемой продукции;

- *непрерывность* – многократная корректировка планов развития по мере поступления новой информации;

- *адаптация* – способность эффективно приспосабливаться к непредсказуемым вариациям внешних условий и внутренних технико-экономических характеристик;

- *управляемость* – поиск оптимальных управленческих решений на основе анализа результатов ретроспективного и стратегического анализа[12].

Данные особенности производственных систем предопределяют следующие экономические принципы проведения статистического анализа:

- соответствие экономическим законам и положениям концепции расширенного воспроизводства;

- адекватное отражение сущности экономической политики современного этапа общественно- экономического развития;

- ориентация на конечные экономические результаты;

- учет специфики изучаемого объекта, вида экономической деятельности и т.д.;

-согласование интересов субъектов различных иерархических уровней как составных элементов единого экономического механизма.

Одной из основополагающих предпосылок проведения научно обоснованного статистического анализа, адекватно отражающего причинно-

следственные связи и зависимости, тенденции развития реальных явлений и процессов в статике и динамике, является однородность статистической совокупности. Это связано с необходимостью обработки больших массивов информации, формируемой по разнородным экономическим единицам. Общепринято, что какая-либо статистическая обработка данных производится только в однородных группах наблюдений. В статистической теории и практике разработаны различные подходы к оценке степени однородности. Содержание основных подходов к выделению однородных групп заключается в следующем.

1. *Вероятностно- статистический подход* предполагает выделение групп, каждая из которых представляет собой реализацию некоторой случайной величины. В классическом виде подход называется методом разделения (расщепления) смесей, и формально задача ставится так: предполагается, что исходная совокупность представляет собой смесь нескольких выборок (обычно считается, что выборки представляют собой реализации нормальных случайных величин, отличающихся как минимумом вектором средних) и требуется при некоторых предположениях (о числе классов, о матрице ковариации и др.) эти выборки разделить.

2. *Структурный подход* (кластерный анализ и визуализация данных) предполагает выделение компактных групп объектов, удалённых друг от друга, отыскивает естественное разбиение совокупности на области скоплений объектов. Этот подход используется для двух видов исходных данных: матриц близости или расстояний между объектами и объектов, представленных как точки в многомерном пространстве. Наиболее распространены данные второго вида, для них структурный подход можно назвать геометрическим, так как он ориентирован на выделение некоторых геометрически удаленных групп, внутри которых объекты близки.

3. *Вариативный (нормативный)* подход заключается в разделении совокупности по некоторому признаку на группы в соответствии с определёнными интервалами, причём характер распределения объектов на

выбор интервалов и число групп практически не влияет. В одномерном случае подход реализуется структурной группировкой, в многомерной ситуации – в форме комбинационной группировки [6].

Экономические объекты являются многомерными, и только совокупное взаимодействие признаков способно в той или иной мере отражать разбиение объектов на классы по объективному критерию. В связи с этим разбиение многомерных объектов на однородные по основным производственно-экономическим характеристикам подмножества хорошо реализуется на основе методов многомерной классификации, в частности кластерного анализа.

При изучении экономических явлений и процессов измерению подлежат различные проявления их свойств. Некоторые свойства при этом проявляются количественно, другие – качественно. То есть признаки бывают качественными (атрибутивными, описательными) и количественными (дискретными и непрерывными). К дискретным относятся количественные признаки, которые могут принимать только целочисленные значения без промежуточных значений между ними. Непрерывные количественные признаки способны принимать любые значения в определенных границах.

Полученные в результате измерения экономических явлений и процессов переменные могут принадлежать к различным статистическим шкалам. Такая принадлежность во многом предопределяет возможности их статистического анализа. Различают 4 шкалы измерений: номинальную, порядковую (ординальную), интервальную и относительную (шкалу отношений). Качественные признаки имеют номинальную или порядковую шкалу, количественные – интервальную или шкалу отношений.

Номинальные переменные могут быть измерены только в терминах принадлежности к некоторым существенно различным классам. Градации на номинальной шкале могут быть как некоторыми высказываниями, так и числами – цифровыми метками каждой категории. При этом отдельным числам не соответствует никакого эмпирического значения. То есть они играют на этой шкале роль ярлыков и к ним неприменимы обычные правила арифметики.

Возможности обработки переменных, относящихся к номинальной шкале, очень ограничены. Эти переменные используются для группировки, с помощью которых совокупность разбивается на группы по категориям этих переменных и затем производится частотный анализ. Поэтому часто номинальные переменные называют категориальными.

Порядковые переменные позволяют упорядочивать объекты, если указано, какие из них в большей или меньшей степени обладают качеством, выраженной данной переменной. То есть порядковые переменные сортированы в порядке значимости: устанавливается некоторый порядок следования объектов. При этом возможно только сравнивать порядковые переменные (больше, меньше), но невозможно придать этим терминам точный количественный смысл (т.е. измерить величину различий между ними) или сравнить между собой эти различия. Данные, выраженные в порядковой шкале, обрабатываются с помощью частотного анализа, допускаются также вычисление определенных статистических характеристик, таких как медианы; могут применяться непараметрические тесты, формулы которых оперируют рангами. Если должна быть установлена связь (корреляция) с другими переменными такого рода, для этой цели может быть использован коэффициент ранговой корреляции.

Переменные**,** у которых разность (интервал) между двумя значениями имеет эмпирическую значимость, относятся к интервальной шкале. Интервальные переменные позволяют не только упорядочивать объекты измерения, но и численно выражать и сравнивать различия между ними. Примерами интервальных шкал могут служить шкалы измерения большинства экономических характеристик. Могут обрабатываться любыми статистическими методами без ограничений.

Когда на шкале можно указать абсолютный нуль, мы имеем шкалу отношений. К переменным, измеренным в шкале отношений, применимы соотношения эквивалентности и порядка - операции вычитания и умножения (шкалы отношений 1-го рода - пропорциональные шкалы), а во многих случаях

и суммирования (шкалы отношений 2-го рода - аддитивные шкалы). Для них являются обоснованными утверждения, что одна величина больше другой в определенное количество раз. По шкале отношений можно измерять такие характеристики, как стаж работы, заработная плата, потребление сырья, окупаемость инвестиций и т.п. В шкалах отношений допустимы все арифметические и статистические операции.

Анализ экономических данных основан на комплексном применении традиционных и многомерных статистических методов. Традиционными статистическими методами являются дисперсионный анализ, корреляционный анализ, регрессионный анализ, анализ временных рядов.

К методам многомерного статистического анализа относятся: кластерный анализ, компонентный анализ (метод главных компонент), факторный анализ, дикриминантный анализ.

Таким образом, востребованность статистического инструментария при исследовании деятельности экономических субъектов предопределяется стохастическим принципом действия экономической системы и обработкой больших массивов информации, формируемой по разнородным экономическим единицам. Комплексное применение статистических методов позволяет охарактеризовать различные направления анализа: изучение структуры, динамики, выявление и характеристика взаимосвязей, прогнозирование, что, в свою очередь, обеспечивает наиболее полное раскрытие сущности, закономерностей и тенденций развития изучаемых явлений и процессов.

1.2 Обзор средств статистического анализа данных в пакетах прикладных программ

Эффективность аналитической работы значительно повышает применение компьютерных технологий. Это достигается за счет сокращения сроков проведения анализа; более полного охвата влияния факторов на изучаемые явления; замены приближенных или упрощенных расчетов точными

вычислениями; постановки и решения многомерных задач анализа, практически не выполнимых вручную и традиционными методами.

Большие возможности для использования методов статистики открылись в результате появления персональных компьютеров. Использование ПК для обработки данных стало особенно эффективным с появлением электронных таблиц (табличных процессоров) – пакетов прикладных программ для автоматизации табличных расчётов. В настоящее время с этой целью широко применяется табличный процессор Excel, работающий на платформе Windows, который представляет большие возможности по осуществлению статистических вычислений.

Основными средствами анализа статистических данных в Excel являются статистические процедуры (инструменты) надстройки *Пакет анализа* и статистические функции библиотеки встроенных функций. В Пакет анализа входят следующие статистические процедуры:

1) генерация случайных чисел;
2) выборка;
3) гистограмма;
4) описательная статистика;
5) ранг и персентиль;
6) двухвыборочный z-тест для средних;
7) двухвыборочный t –тест для средних с одинаковыми дисперсиями;
8) двухвыборочный t –тест для средних с различными дисперсиями;
9) парный двухвыборочный t –тест для средних;
10) двухвыборочный F-тест для дисперсий;
11) ковариация;
12) корреляция;
13) регрессия;
14) однофакторный дисперсионный анализ;
15) двухфакторный дисперсионный анализ без повторений;
16) двухфакторный дисперсионный анализ с повторением;

17) скользящее среднее;

18) экспоненциальное сглаживание;

19) анализ Фурье.

Статистические функции являются одним из разделов библиотеки встроенных функций рабочего листа Excel. В этот раздел входит функции, предназначенные для решения некоторых наиболее востребованных задач теории вероятностей и статистики. Большая часть статистических функций дублирует (в несколько упрощённом виде) некоторые процедуры, входящие в надстройку *Пакет анализа*. Однако другая часть функций вполне «самостоятельна» и среди процедур этого пакета аналогов не имеет.

Современный этап характеризуется также наличием специализированных и универсальных пакетов прикладных программ, предназначенных для статистического анализа данных. Статистический пакет - программный продукт, предназначенный для статистической обработки данных. Из зарубежных пакетов известны STATGRAPHICS, SPSS, SYSTAT, BMDP, SAS, CSS, STATISTICA, S-plus, и др. Из отечественных можно назвать такие пакеты, как STADIA, ЭВРИСТА, МЕЗОЗАВР, ОЛИМП, Стат-Эксперт, Статистик-Консультант, САНИ, КЛАСС-МАСТЕР и др.

Специализированные пакеты обычно содержат методы из одного - двух разделов статистики или методы, используемые в конкретной предметной области (контроль качества промышленной продукции, расчет страховых сумм и т.д.). Чаще всего встречаются пакеты для анализа временных рядов (например, ЭВРИСТА, МЕЗОЗАВР, ОЛИМП, Стат-Эксперт), регрессионного и факторного анализа. Обычно эти пакеты содержат весьма полный набор традиционных методов в своей области, а иногда включают также и оригинальные методы и алгоритмы, созданные разработчиками пакета. Как правило, специализированный пакет и его документация ориентированы на специалистов, хорошо знакомых с соответствующими методами.

Универсальные пакеты (пакеты общего назначения) предлагают широкий диапазон статистических методов, характеризуются отсутствием прямой ориентации на специфическую предметную область. Примерами универсальных статистических пакетов могут служить американские пакеты SPSS, STATGRAPHICS, SYSTAT, STATISTIKA, S-plus и отечественный пакет STADIA.

Выбор статистического пакета для анализа данных зависит от характера решаемых задач, объема и специфики обрабатываемых данных, квалификации и категории пользователей, имеющегося оборудования, ценового фактора и др.

Для того чтобы статистический пакет общего назначения был удобен и эффективен в работе, он должен удовлетворять определенным требованиям:

· содержал достаточно полный набор стандартных статистических методов;

· был достаточно прост для быстрого освоения и использования;

· отвечал высоким требованиям к вводу, преобразованиям и организации хранения данных;

· имел широкий набор средств графического представления данных и результатов обработки;

· предоставлял удобные возможности для включения в отчеты таблиц исходных данных, графиков, промежуточных и окончательных результатов обработки;

· имел подробную документацию, доступную для начинающих и информативную для специалистов-статистиков.

Система пакета прикладных программ состоит из ряда модулей, работающих независимо. Каждый модуль включает определенный класс процедур. Почти все процедуры являются интерактивными, т.е. для запуска обработки необходимо выбрать из меню переменные и ответить на ряд вопросов системы. Модули и процедуры могут различаться в зависимости от

редакций, версий ППП (базовая, основная, специальная), от модификации продукта.

Пакет STATISTICA – универсальный статистический пакет фирмы StatSoft. Пакет был создан в начале 1990-х годов сразу для среды Windows. Основные модули и процедуры: описательные статистики, анализ многомерных таблиц, подгонка распределений, корреляционный анализ, регрессионный анализ, дисперсионный анализ, кластерный анализ, дискриминантный анализ, факторный анализ, анализ соответствий, многомерное шкалирование, анализ выживаемости, структурные модели, деревья классификаций, прогнозирование временных рядов, непараметрическая статистика, анализ Монте-Карло и др. Работа со всеми модулями происходит в рамках единого программного пакета, для которого можно выбирать один из нескольких предложенных интерфейсов пользователя.

Статистический пакет для социальных наук - Statistical Package for Social Science (SPSS) – модульный, полностью интегрированный, обладающий всеми необходимыми возможностями программный комплекс, охватывающий все этапы аналитического процесса: планирование, сбор данных, доступ к данным и управление данными, анализ, создание отчетов и распространение результатов. Основные функциональные возможности программы по обработке статистических данных и презентации результатов: кодирование статистических данных; проведение тестов на наличие нормального распределения; построение таблиц сопряженности и расчет статистических характеристик для них; частотный анализ; проведение непараметрических тестов для выборок; корреляционный анализ; простая линейная регрессия (пример из маркетинга); многомерная регрессия; нелинейная регрессия, логистическая регрессия; дискриминантный анализ; факторный анализ; кластерный анализ.

Пакет STATGRAPHICS (не русифицирован) имеет следующие основные процедуры: основные статистики и разведочный анализ данных, дисперсионный анализ и регрессии, SPC (анализ возможностей, контрольные

карты, анализ измерительных систем), планирование экспериментов, шесть сигма, надежность и анализ жизненного цикла данных, многомерный анализ, непараметрические методы, анализ временных рядов и прогнозирование.

Параллельно с разработкой статистических пакетов шёл интенсивный процесс включения статистических функций в математические пакеты общего назначения (Mathcad, Math lab, Maple, Mathematica (США) и др.).

Перспективным инструментом решения трудноформализуемых задач статистического анализа и прогнозирования являются пакеты, построенные по технологии обучающихся нейронных сетей, в частности пакет STATISTICA Neural Network. Известны применения нейрокомпьютеров (CNAPS PC/128), имитаторов нейронных сетей (Qnet for WIndows) для прогнозирования финансовой деятельности и пр.

Таким образом, в настоящее время средства программной поддержки методов статистических исследований представлены табличными процессорами (Excel), статистическими пакетами, математическими пакетами с включёнными в них статистическими функциями и пакеты, построенные по технологии обучающихся нейронных сетей. Они сделали анализ экономических данных более доступным и наглядным.

2. МЕТОД СТАТИСТИЧЕСКОГО ВЫВОДА В АНАЛИЗЕ ЭКОНОМИЧЕСКИХ ДАННЫХ

2.1 Формирование выборки

Любое статистическое исследование начинается со сбора данных об исследуемом объекте. Этот этап работы называется *наблюдением*. Данные, собранные и зафиксированные в ходе наблюдения, называются результатами наблюдения. Объектом исследования в терминологии математической статистики выступает случайная величина. *Случайной величиной* называется переменная величина, которая в зависимости от случайного исхода испытаний принимает какое-то одно из своих возможных значений, причем заранее неизвестно, какое именно.

Выполнив *n* независимых наблюдений над исследуемой случайной величиной *X*, получим *n* чисел: $x_1, x_2 ..... x_j, ... x_n$, которые называют наблюдаемыми значениями или реализациями этой случайной величины.

То есть *реализацией* случайной величины *X* называется числовое значение *x*, которое приняла эта случайная величина в каком-то конкретном испытании. Множество значений, которые может принимать случайная величина *X*, называется областью возможных значений этой случайной величины.

Наблюдаемые значения исследуемой случайной величины в статистике принято рассматривать как *случайную выборку* из генеральной совокупности реализаций этой случайной величины, которые могли бы быть получены при проведении всех мыслимых наблюдений над этой случайной величиной. При этом числа $x_1, x_2 ..... x_j, ... x_n$, образующие выборку, называются *элементами выборки*, а число *n* этих элементов – *объемом выборки*.

Таким образом, выборка является основным исходным объектом любого статистического исследования.

Применение выборочного метода обусловлено тем, что часто генеральная совокупность слишком многочисленна или малодоступна. Поэтому для

исследования из нее формируют выборочные совокупности (выборки), по которым судят обо всей генеральной совокупности (то есть с заданным уровнем вероятности распространяют результаты выборочного обследования на всю изучаемую совокупность). Для полного, адекватного представления информации о генеральной совокупности выборка должна быть представительной (репрезентативной). Это можно сделать только в том случае, когда выборка достаточно точно отображает пропорции генеральной совокупности, то есть когда распределение исследуемого случайного признака в выборке достаточно близко к распределению этого признака в генеральной совокупности. Репрезентативность выборки достигается отсутствием всякой предвзятости по отношению к отбираемым элементам. Каждый элемент генеральной совокупности должен иметь равную со всеми остальными элементами возможность включения в выборку (принцип обеспечения случайности отбора).

Порядок отбора единиц из генеральной совокупности называется способом отбора. Различают два способа отбора: повторный и бесповторный.

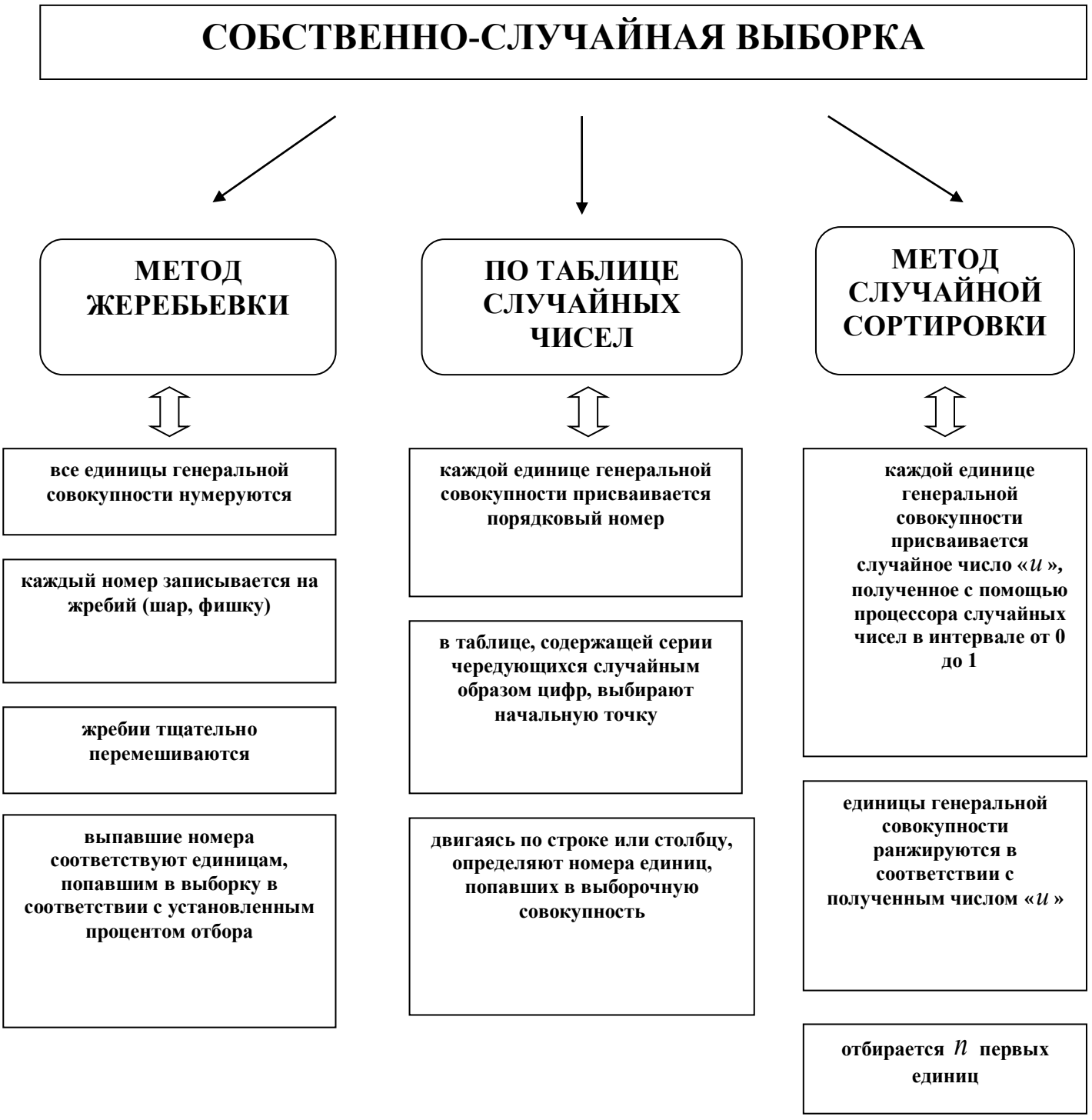
При *повторном способе* каждая отобранная в случайном порядке единица после её обследования возвращается в генеральную совокупность и при последующем отборе может снова попасть в выборку. Вероятность попадания любой единицы в выборку равна $1/N$, и она остаётся той же самой на протяжении всей процедуры отбора.

При *бесповторном способе* отбора попавшая в выборочную совокупность единица после регистрации значений наблюдаемых признаков не возвращается в совокупность, из которой осуществляется дальнейший отбор. По мере производства отбора вероятность попасть в выборку для каждой единицы генеральной совокупности увеличивается, тем самым повышается репрезентативность выборки.

В зависимости от методики формирования выборочной совокупности различают следующие *основные виды выборки*:

- ✓ собственно– случайная;

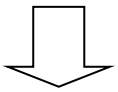
- ✓ механическая;
- ✓ типическая (стратифицированная, расслоенная, районированная);
- ✓ серийная (гнездовая);
- ✓ многоступенчатая;
- ✓ многофазная;
- ✓ комбинированная;
- ✓ взаимопроникающая.



Собственно-случайная выборка заключается в отборе единиц из генеральной совокупности случайным образом, наугад или наудачу, без каких либо элементов системности.

Механическая (периодическая) выборка состоит в том, что отбор единиц в выборочную совокупность производится из генеральной совокупности, которая каким-либо образом упорядочена, то есть имеется определённая последовательность в расположении её единиц. Размер интервала отбора равен обратной величине доли выборки.

ТИПИЧЕСКАЯ ВЫБОРКА



генеральная совокупность расчленяется на однородные типические группы (страты или слои)

из каждой типической группы собственно-случайным или механическим способом производится индивидуальный отбор единиц в выборочную совокупность

1) пропорционально объему типических групп

2) пропорционально внутригрупповой вариации (дифференциации) признака

При *серийной (гнездовой) выборке* из генеральной совокупности отбираются не отдельные единицы, а целые их серии (гнёзда), группы. Отбор отдельных серий в выборочную совокупность осуществляется посредством собственно-случайной или механической выборки. Внутри каждой из попавших в выборку серии обследуются все без исключения единицы, т.е. производится сплошное наблюдение единиц.

При *многоступенчатой выборке* выборочная совокупность формируется постепенно, по ступеням отбора. На каждой ступени используются разные единицы отбора: более крупные на начальных ступенях, на последней ступени единица отбора совпадает с единицей наблюдения.

В отличие от многоступенчатой *многофазная выборка* предполагает сохранение одной и той же единицы отбора на всех этапах (стадиях) его проведения; при этом выборка на каждом этапе отличается подробностью программы наблюдения: на каждой последующей стадии отбора программа обследования расширяется, т. е выборка включает несколько фаз. Например, 25% всей генеральной совокупности обследуется по краткой программе, каждая четвёртая из этой выборки обследуется по более полной программе и т.д.

Комбинированная выборка предполагает объединение нескольких видов отбора. Например, можно комбинировать типическую и собственно-случайную, типическую и механическую, серийную и собственно-случайную выборки.

2.2 Основные понятия теории статистического оценивания и статистической проверки гипотез

Для характеристики отдельных свойств распределения данных наблюдения используются специальных числовые параметры, отражающие в сжатом виде основные, существенные черты распределения выборочных данных. Эти числовые параметры называются *эмпирическими (выборочными) числовыми характеристиками*.

Для распространения данных выборки на всю генеральную совокупность используются ***статистические выводы***. В статистическом выводе делается предположение о свойствах, параметрах генеральной совокупности, используя данные выборки (результаты наблюдения за объектом исследования в течение конечного промежутка времени). Результатом статистического вывода является статистическое суждение: точечная оценка, интервальная оценка, принятие или отклонение гипотезы. То есть параметры генеральной совокупности по данным

выборки могут быть установлены на основе статистического оценивания и статистической проверки гипотез.

Идея ***статистического оценивания*** параметров генеральной совокупности по выборочным данным сводится к тому, что выборочная характеристика какого-либо параметра является не точным, а приближенным значением – *оценкой* – этого же параметра в генеральной совокупности.

Качество эмпирических (выборочных) оценок характеризуют такими основными свойствами как состоятельность, несмещенность и эффективность.

Оценка θ^* числовой характеристики θ называется *состоятельной*, если с увеличением объема выборки (т.е. при $n \to \infty$) эта оценка стремиться к значению оцениваемого параметра, т.е. сходится по вероятности к оцениваемой числовой характеристике θ. Согласно этому определению при достаточно большом объеме n выборки состоятельная оценка с высокой вероятностью практически равна оцениваемой числовой характеристике. Для состоятельности оценки θ^* достаточно, чтобы при $n \to \infty$ ее математическое ожидание M (θ^*) стремилось к оцениваемой числовой характеристике θ, а ее дисперсия D(θ^*) стремилась к нулю. Таким образом оценка является состоятельной, если она удовлетворяет закону больших чисел.

Оценка θ^* называется *несмещенной,* если ее выборочное математическое ожидание M (θ^*) равно оцениваемой числовой характеристике θ (оцениваемому параметру генеральной совокупности). Несмещенность оценки особенна важна при малом объеме выборки. Оценка θ^*, математическое ожидание которой M (θ^*) не совпадает с оцениваемой характеристикой θ, называется *смещенной*. Разность b (θ^*) = M (θ^*) $-\theta$ называется *смещением* или *систематической ошибкой* оценки θ^*. В том случае, когда смещение найдено, его легко устранить введением соответствующей поправки.

Оценка θ^* называется *эффективной* для параметра θ, если она имеет наименьшую возможную дисперсию при заданном объеме выборки n. То есть качество несмещенной оценки определяется величиной ее дисперсии: чем меньше дисперсия, тем лучше оценка. Из двух несмещенных оценок числовой

характеристики θ, найденных по одной и той же выборке, предпочтительной (более эффективной) считается оценка с меньшей дисперсией. Использование такой оценки позволяет добиться необходимой точности при меньшем объеме выборке.

В процессе наблюдения получаются неупорядоченные данные, просматривая которые трудно оценить отражаемые им закономерности. Упорядочение данных наблюдения осуществляется для того, чтобы данные наблюдения сделать более наглядными и упросить их дальнейший анализ. Упорядочение производится путем *ранжирования* – расположения значений изучаемого признака в порядке возрастания или убывания и группировки – разделение единиц изучаемой совокупности на группы по определённым существенным для них признакам. При группировке получают *эмпирические ряды распределения*. Ряды распределения, построенные по количественному признаку, называются *вариационными*. Различают дискретные вариационные ряды распределения - построенные по дискретному количественному признаку, и интервальные вариационные ряды – построенные по непрерывному количественному признаку. Они состоят из двух элементов: 1) возможные значения признака (варианты) или интервалы; 2) частота или частоть - число (доля) единиц совокупности в интервальной группе.

При построении вариационных рядов:

- определяется число групп;
- вычисляется величина равного или неравного интервала;
- определяются границы интервалов;
- осуществляется распределение единиц совокупности по полученным интервальным группам.

Для определения числа групп могут быть использованы следующие формулы:

а) $k = 1 + 3{,}322 \lg n = 1 + 1{,}443 \ln n$ (формула Стерджесса);

б) $k = 1{,}72\sqrt[3]{n}$;

в) $k = \left|\sqrt{n} - 0{,}013n + 0{,}5\right|$;

г) $k = \left|\sqrt{n} + 1\right|$, где

n - число единиц исследуемой совокупности.

Наиболее важными числовыми характеристиками являются характеристики положения, вариации (рассеивания), асимметрии и эксцесса.

Для характеристики *положения* используются показатели центра распределения данных наблюдения– средняя (арифметическая), мода и медиана.

Средняя (математическое ожидание) – характеризует типичный размер признака у единиц исследуемой совокупности. Для дискретного вариационного ряда распределения рассчитывается по формуле:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}, \text{ где}$$

x_i –варианты значений признака;

f_i– частота повторения данного варианта.

В интервальном вариационном ряду средняя определяется по формуле:

$$\bar{x} = \frac{\sum x'_i f_i}{\sum f_i}, \text{ где}$$

x' – середина соответствующего интервала;

f_i– частота интервала.

Мода распределения– это наиболее часто встречающееся значение признака в совокупности. Выборочная мода непрерывной случайной величины и дискретной случайной величины с большим числом возможных значений определяется по сгруппированным данным.

В рядах распределения с равными интервалами рассчитывается по формуле:

$$Mo = x_0 + i\frac{(f_{Mo} - f_{Mo-1})}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})}$$

x_0 – нижняя граница модального интервала;

i– величина модального интервала;

f_{Mo} – частота модального интервала;

f_{Mo-1} – частота интервала, предшествующего модальному;

f_{Mo+1} – частота интервала, следующего за модальным.

В рядах распределения с неравными интервалами применяется формула:

$$M_o = x_0 + i\frac{\frac{f_{M_0}}{i_{M_0}} - \frac{f_{M_{0-1}}}{i_{M_{0-1}}}}{\left(\frac{f_{M_0}}{i_{M_0}} - \frac{f_{M_{0-1}}}{i_{M_{0-1}}}\right) + \left(\frac{f_{M_0}}{i_{M_0}} - \frac{f_{M_{0+1}}}{i_{M_{0+1}}}\right)}$$

x_0 – нижняя граница модального интервала;

i– величина модального интервала;

$\frac{f_{M_0}}{i_{M_0}}$ – плотность модального интервала;

$\frac{f_{M_{0-1}}}{i_{M_{0-1}}}$ – плотность интервала, предшествующего модальному;

$\frac{f_{M_{0+1}}}{i_{M_{0+1}}}$ – плотность интервала, следующего за модальным.

Медиана – это значение признака, расположенное в середине (в центре) ранжированного ряда. Медиана делит совокупность на две равные части– со значениями признака меньше медианы и со значениями признака больше медианы.

$$Me = x_0 + i\frac{\frac{\sum f}{2} - S_{Me-1}}{f_{Me}}$$

x_0 – нижняя граница медианного интервала;

i– величина медианного интервала;

$\sum f$ – сумма всех частот ряда;

S_{Me} – накопленная частота интервала, предшествующего медианному;

f_{Me} – частота медианного интервала.

Основными характеристики *вариации* признака являются дисперсия, среднее квадратическое (стандартное) отклонение и коэффициент вариации. Они характеризуют степень рассеивания данных наблюдения относительно центра распределения. Абсолютные показатели вариации основаны на учете отклонений индивидуальных значений признака от средней арифметической.

Дисперсия рассчитывается по формуле: $\sigma^2{}_x = \frac{\sum\left(x_i - \bar{x}\right)^2}{n}$. Стандартное отклонение σ равно корню квадратному из дисперсии. Коэффициент вариации равен: $V_\sigma = \frac{\sigma_x}{\bar{x}} \times 100\%$.

Стандартное отклонение показывает, на сколько в среднем отклоняются индивидуальные значения признака от их средней величины. Размерность отклонения σ совпадает с размерностью самого признака. Коэффициент вариации измеряет интенсивность вариации. Для нормальных и близких к нормальному распределений показатель вариации служит индикатором однородности совокупности: принято считать, что при выполнимости неравенства $V \leq 33$, % совокупность является количественно однородной по данному признаку. Коэффициент вариации V используется для сравнения вариации разных признаков в одной и той же совокупности и вариации одного и того же признака в различных совокупностях, имеющих разные средние x.

При резко асимметричном распределении более удобной характеристикой «центра» распределения является медиана. Она более устойчива к резким выбросам данных, чем средняя, что позволяет использовать ее при работе с распределениями, имеющими «хвосты». В этом случае для измерения вариации признака применяются коэффициент вариации, определяемый делением стандартного отклонения на медиану.

Асимметрия характеризует меру несимметричности (скошенности) распределения. Если коэффициент асимметрии больше нуля, то асимметрия правосторонняя, если меньше нуля – левосторонняя. Выборочный коэффициент асимметрии, основанный на определении центрального момента третьего порядка μ_3 (в нормальном распределении его величина равна нулю) вычисляется по формуле: $A_S = \frac{\mu^3}{\sigma^3}$; $\mu_3 = \frac{\sum\left(x_i - \bar{x}\right)^3}{n}$.

Эксцесс представляет собой выпад вершины эмпирического распределения вверх или вниз от вершины кривой нормального распределения,

имеющей куполообразную форму. Эксцесс характеризует островершинность (плосковершинность) распределения. Если эксцесс больше нуля, то распределение островершинное, если меньше нуля – плосковершинное. Наиболее точным является выборочный коэффициент эксцесса, основанный на использовании центрального момента четвёртого порядка: $E = \frac{\mu^4}{\sigma^4} - 3$; $\mu_4 = \frac{\sum\left(x_i - \bar{x}\right)^4}{n}$ Для нормального распределения E равен нулю, так как $\frac{\mu_4}{\sigma^4} = 3$.

Коэффициенты асимметрии и эксцесса используются для оценки степени отклонения распределения исследуемой величины от нормального распределения.

Если $\frac{/ A_{\varsigma} /}{\sigma_{A\varsigma}} \le 2$ - асимметрия несущественна (ее наличие объясняется наличием случайных обстоятельств) и распределение признака в генеральной совокупности можно считать нормальным.

$\sigma_{A\varsigma} = \sqrt{\frac{6n \times (n-1)}{(n-2) \times (n+1) \times (n+3) \times (n+3)}}$ - средняя квадратическая ошибка коэффициента асимметрии.

Если $\frac{|E|}{\sigma_E} \le 2 -$ эксцесс несущественен и распределение можно отнести к разряду нормального распределения.

$\sigma_E = \sqrt{\frac{24 \times n(n-1)^2}{(n-3) \times (n-2) \times (n+3) \times (n+5)}}$ - средняя квадратическая ошибка коэффициента эксцесса.

Точечные оценки параметров генеральной совокупности по данным выборки рассчитываются по соответствующим формулам и характеризуются одним конкретным числом. Для выборок большого объема (100 и более наблюдений) при условии несмещенности, эффективности и состоятельности точечных оценок их точность признается достаточной. Тем не менее, они, как правило, применяются в качестве ориентировочных, первоначальных результатов обработки выборочных данных. А для выборок небольшого объема точность точечных оценок вообще является недостаточной.

В связи с этим применяются *два основных вида* статистических выводов:

1) *интервальное оценивание,* при котором с заданной вероятностью строится доверительный интервал, в котором находится оцениваемый параметр генеральной совокупности;

2) *проверка статистических гипотез* (вероятностный вывод о том, что определенные параметры выборочной совокупности отображают (или не отображают) параметры генеральной совокупности.

В основе *интервального оценивания* лежит корректировка выборочных оценок на величину ошибки выборки (репрезентативности). *Ошибкой выборки* называется разница между значением показателя, полученного по выборке, и генеральным параметром. Ошибки репрезентативности бывают неслучайными (систематическими) и случайными. *Систематические* ошибки возникают, из-за нарушения схемы и принципов отбора данных. При проведении наблюдения необходимо исключить или минимизировать появление систематических ошибок, которые приводят к нарушению репрезентативности: при наличии систематической ошибки структурные части генеральной совокупности неодинаково представлены в выборке. Величину систематических ошибок определить очень сложно, иногда невозможно.

Статистическому измерению подлежат только *случайные ошибки,* которые обусловлены действием случайных факторов и не содержат каких-либо элементов системности в направлении воздействия на рассчитываемые выборочные характеристики.

Предельной ошибкой выборки (Δ) принято считать максимально возможное расхождение между выборочными и генеральными параметрами, то есть максимум ошибки при заданном уровне вероятности. Надежность выборочной оценки принято задавать перед процессом оценивания параметра генеральной совокупности. На практике при исследовании экономических явлений и процессов достаточным (доверительным) уровнем вероятности (надежности) $p= 1-\alpha$, где α – уровень значимости, считается вероятность, равная 0,95. Это означает, что только в 5-ти случаях из 100 ошибка может

выйти за установленные границы. Для повышения надежности статистических выводов берут α=0,01, чему соответствует доверительная вероятность 0,99.

При изучении экономических типов данных статистическое оценивание, как правило, производится для двух основных видов обобщающих показателей (параметров): средней величины количественного признака и относительной величины альтернативного признака (доли единиц, обладающих заданным значением признака или определенным качеством).

Предельные ошибки выборки для основных видов формирования выборочной совокупности

Способ отбора / Вид отбора	Повторный		Бесповторный	
	для средней	для доли	для средней	для доли
1. Собственно-случайный	$t\sqrt{\frac{\sigma_0^2}{n}}$	$t\sqrt{\frac{\omega(1-\omega)}{n}}$	$t\sqrt{\frac{\sigma_0^2}{n}\left(1-\frac{n}{N}\right)}$	$t\sqrt{\frac{\omega(1-\omega)}{n}\left(1-\frac{n}{N}\right)}$
2. Механический	–	–	$t\sqrt{\frac{\sigma_0^2}{n}\left(1-\frac{n}{N}\right)}$	$t\sqrt{\frac{\omega(1-\omega)}{n}\left(1-\frac{n}{N}\right)}$
3. Типический (пропорционально объёму групп)	$t\sqrt{\frac{\overline{\sigma_i^2}}{n}}$	$t\sqrt{\frac{\overline{\omega_i(1-\omega_i)}}{n}}$	$t\sqrt{\frac{\overline{\sigma_i^2}}{n}\left(1-\frac{n}{N}\right)}$	$t\sqrt{\frac{\overline{\omega_i(1-\omega_i)}}{n}\left(1-\frac{n}{N}\right)}$
4. Серийный (гнездовой)	$t\sqrt{\frac{\delta_x^2}{r}}$	$t\sqrt{\frac{\delta_\omega^2}{r}}$	$t\sqrt{\frac{\delta_x^2}{r}\left(1-\frac{r}{R}\right)}$	$t\sqrt{\frac{\delta_\omega^2}{r}\left(1-\frac{r}{R}\right)}$

где t – критерий Стьюдента при заданном уровне значимости α и числе степеней свободы $\nu = n-2$;

$\sigma_0^2 = \frac{\sum(x_i-\bar{x})^2}{n}$ -выборочная общая дисперсия количественного признака;

$\omega(1-\omega)$ - выборочная общая дисперсия альтернативного признака;

$w = \frac{m}{n}$ - доля единиц, обладающих заданным значением признака или определенным качеством в выборочной совокупности;

n -объем выборочной совокупности;

N - объем генеральной совокупности.

$\overline{\sigma_i^2} = \frac{\sum \sigma_i^2 n_i}{\sum n_i}$ – выборочная средняя из внутригрупповых дисперсий для количественного признака;

$\overline{\sigma_\omega^2} = \overline{\omega_i(1-\omega_i)} = \frac{\sum \omega_i(1-\omega_i)n_i}{\sum n_i}$ –выборочная средняя из внутригрупповых дисперсий для доли;

$\delta_x^2 = \frac{\sum(\tilde{x}_i-\tilde{x})^2}{r}$ –выборочная межгрупповая дисперсия для количественного признака;

где $\tilde{x}_i$ – среднее значение признака в i-ой серии;

$\tilde{x}$ – общая средняя по всей выборочной совокупности;

r – число отобранных серий;

R – общее число серий;

$\delta_\omega^2 = \dfrac{\sum\left(\omega_i - \bar{\omega}\right)^2}{r}$ –выборочная межгрупповая дисперсия для доли.

Для распространения результатов выборки на изучаемую совокупность строятся доверительные интервалы, нижняя и верхняя границы которых определяются путем корректировки выборочных характеристик на величину предельной ошибки выборки:

$$\tilde{x} - \Delta_x \le \bar{x} \le \tilde{x} + \Delta_x;$$
$$\omega - \Delta_\omega \le p \le \omega + \Delta_\omega.$$

Второй важнейшей задачей при решении прикладных задач статистического анализа экономических данных после статистического оценивания параметров распределения является статистическая проверка гипотез.

Под *статистической гипотезой* принято понимать любое (разумное с точки зрения теории вероятностей) предположение о закономерностях, которым подчиняется исследуемый случайный объект.

Статистическая гипотеза – это любое предположение о виде неизвестного закона распределения или о параметрах известного распределения. Статистическую гипотезу принято обозначать символом H (по первой букве греческого слова *hypothesis* – предположение).

Гипотезы о значениях параметров распределения или о сравнительной величине параметров двух или нескольких распределений называются *параметрическими* (например, математические ожидания двух нормальных совокупностей равны между собой). Гипотезы о законе (виде) распределения называются *непараметрическими* (например, генеральная совокупность подчиняется закону нормального распределения).

Проверить статистическую гипотезу– значит проверить, согласуются ли выборочные (эмпирические) данные наблюдения с выдвинутой гипотезой.

Статистическая гипотеза выдвигается на основании теоретических соображений, вытекающих из сущности исследуемого случайного явления, или исходя из результатов предварительного анализа данных наблюдения над этим явлением.

Если исследуемый случайный объект характеризуется l параметрами и гипотеза H задает конкретные числовые значения всех этих параметров, то она называется *простой*. Если хотя бы один из l параметров задан не одним конкретным числом, а указанием интервала его возможных значений, то гипотеза называется *сложной*.

Статистическая гипотеза, подлежащая проверке, называется основной или *нулевой гипотезой* и обозначается H_0. Любая другая гипотеза относительно исследуемого случайного объекта называется *альтернативной* (конкурирующей) *гипотезой*. Например, если основная гипотеза H_0 содержит предположение, что среднее значение случайной величины X равно 3,5, то в качестве альтернативы может фигурировать одно из следующих предположений: 1) $H_1 : M(x) > 3{,}5$ 2) $H_1 : M(x) < 3{,}5$, 3) $H_1 : M(x) \neq 3{,}5$.

Конечная цель проверки всякой статистической гипотезы состоит в том, чтобы принять или отклонить проверяемую гипотезу. Решение этого вопроса зависит от того, согласуется проверяемая гипотеза с фактическими данными наблюдения или нет. Если гипотеза не противоречит фактическим данным, то ее принимают, если же она противоречит реальным данным результатам наблюдения, ее отклоняют. Правило, по которому принимается или отклоняется статистическая гипотеза, называется *статистическим критерием*.

Статистические критерии носят названия в соответствие с распределением: F-критерий, Т-критерий, χ^2-критерий и др.

В основе всякого статистического критерия лежит детерминированная (неслучайная) функция $U = \varphi(x_1, x_2, ...x_n)$ данных наблюдения $x_1, x_2, ...x_n$, выбранная таким образом, чтобы она могла служить мерой расхождения между

проверяемой гипотезой H_0 и реальными данными наблюдения. Обычно функцию φ выбирают таким образом, чтобы она принимала малые значения, когда H_0 верна, и большие значения, когда гипотеза H_0 ошибочна. Мера расхождения U, лежащая в основе статистического критерия, называется статистикой этого критерия, или *критериальной статистикой*.

Наблюдаемое (расчетное) значение статистического критерия u – это значение статистики U, найденное по данным конкретной выборки.

Критическое значение статистического критерия $u_{(\alpha)}$ – это значение статистики U при заданном уровне значимости α.

Если расчетное значение u статистики U меньше ее критического значения $u_{(\alpha)}$, то считают, что гипотеза H_0 не противоречит данным наблюдения и ее следует принять: если же u больше или равно $u_{(\alpha)}$, то проверяемую гипотезу отклоняют как противоречащую реальным данным (говорят, что гипотеза H_0 отклоняется на уровне значимости α).

В соответствии с эти правилом интервал $\Omega_{кр.} = \lfloor u_{(\alpha)}, \infty)$ называют областью отклонения проверяемой гипотезы или *критической областью* порядка α, а интервал $\Omega_{пр.} = (0, u_{(\alpha)})$ – областью принятия проверяемой гипотезы или *областью допустимых значений*.

Критической областью называется область, попадание значения статистического критерия в которую приводит к отклонению H_0. Вероятность попадания значения критерия в эту область равна уровню значимости α, (величина которого в социально-экономических исследованиях обычно составляет 0,05).

Область допустимых значений (принятия проверяемой гипотезы) дополняет критическую область. Если рассчитанное по эмпирическим данным значение критерия попадает в область допустимых значений, это

свидетельствует о том, что выдвинутая гипотеза H_0 не противоречит фактическим данным (H_0 не отклоняется).

Проверка любой статистической гипотезы основана на случайной выборке, объем которой всегда конечен. Поэтому проверка простых гипотез сопряжена с возможностью допустить следующие ошибки:

1) отклонить проверяемую гипотезу H_0, когда она верна (ошибка 1-го рода);

2) принять проверяемую гипотезу H_0, когда верна гипотеза H_1(ошибка 2-го рода).

Вероятность ошибки 1 рода равна уровню значимости критерия α. Ошибка 1-ого рода совершается в том случае, когда гипотеза H_0 верна, а выборочное значение критерия попало в критическую область. То есть, чем меньше уровень значимости α, тем меньше риск ошибочного отклонения проверяемой гипотезы.

Ошибка 2-го рода происходит, когда верна альтернативная гипотеза H_1, а выборочное значение критерия попало в область допустимых значений. Вероятность ошибки 2-го рода принято обозначать символом β.

Вероятность $1 - \beta$ отклонения гипотезы H_0, когда она неверна (а справедлива конкурирующая гипотеза), называется *мощностью критерия* проверки гипотезы.

То есть α– вероятность ошибочного отклонения гипотезы H_0 (вероятность ошибки 1 –го рода);

β– вероятность ошибочного принятия гипотезы H_0 (вероятность ошибки 2 –го рода);

$1 - \beta$ – вероятность правильного отклонения гипотезы H_0 (мощность критерия).

Мощность критерия тем больше, чем больше число *n* данных наблюдения и уровень значимости α и чем сильнее отличается значение параметра

исследуемого случайного объекта, заданного гипотезой H_0 от значения этого же параметра, заданного альтернативой H_1.

Различают одностороннюю и двустороннюю критические области. Когда при справедливости гипотезы H_0 маловероятны большие значения статистики U используется *правосторонняя критическая область*, определяемая неравенством $U \geq u_{(\alpha)}$. В тех случаях, когда при справедливости гипотезы H_0 маловероятны малые значения статистики U используется левосторонняя критическая область вида $U \leq u_{н(\alpha)}$, где $u_{н(\alpha)}$ - нижнее критическое значение порядка α статистики U.

При использовании некоторых критериев «опасными» для проверяемой гипотезы H_0 являются как слишком большие, так и слишком малые значения статистики U. В таких случаях приходится определять нижнее и верхнее критические значения статистики U порядка α/2. Эти критические значения делят область возможных значений статистики U на три части: нижнюю критическую область $\Omega_{н.кр.}$, область принятия проверяемой гипотезы $\Omega_{пр.}$ и верхнюю критическую область $\Omega_{в.кр.}$ Критическая область, представляющая собой объединение нижней и верхней критических областей, называется *двусторонней*.

Статистические критерии чрезвычайно разнообразны по своему назначению. Однако их объединяет общность логической схемы, по которой они строятся и которая формирует *основные этапы проверки статистических гипотез*:

1) исходя из теоретических соображений, вытекающих из природы исследуемого случайного объекта, или на основании анализа результатов первичной обработки данных наблюдения над этим объектом формулируется проверяемая основная H_0 и конкурирующая (альтернативная) H_1 гипотезы;

2) принимается уровень значимости α, контролирующий допустимую вероятность попадания в критическую область (автоматически устанавливается вероятность ошибки 1-го рода);

3) выбирается статистика U, которая будет использоваться в качестве меры расхождения между проверяемой гипотезой H_0 и реальными данными наблюдения $x_1, x_2, ...x_n$. Имеются подробные таблицы распределения наиболее употребительных статистик, позволяющие находить их критические значения, соответствующие заданному уровню значимости α;

4) с помощью формулы по конкретным данным наблюдения определяется расчётное значение статистического критерия;

5) в соответствии с альтернативной гипотезой H_1 выбирается вид критической области $\Omega_{кр.}$. По уровню значимости α с помощью таблиц распределения статистики U определяются границы этой области;

6) принимается решение о принятии или отклонении проверяемой гипотезы. Если расчётное значение критерия попало в область допустимых значений $\Omega_{пр.}$, то гипотеза H_0 принимается. Если же расчетное значение критерия попало в критическую область $\Omega_{кр.}$, то гипотеза H_0 - отклоняется на уровне значимости α как противоречащая реальным данным наблюдения.

7) если альтернатива H_1 простая. То вычисляются вероятность β ошибки 2-ого рода и мощность критерия $1 - \beta$.

Процедуры проверки гипотез при обработке случайных выборок позволяют следующие задачи их статистической обработки:

– выявление различий между выборками; выполняется с использованием *критериев различия: t*-критерия Стьюдента, критерия Фишера и др.;

– определение меры соответствия выборки какому-либо теоретическому распределению. Выполняется с использованием *критериев согласия: хи*-квадрат – Пирсона, Колмогорова, Крамера- Мизеса- Смирнова и др.

2.3 Проверка основных видов статистических гипотез

К основным типам гипотез, проверяемых в ходе статистической обработки данных, относятся:

- гипотезы *о типе закона распределения* признака (чаще всего проверяется соответствие нормальному закону распределения);
- гипотезы *о числовых значениях параметров* совокупности;
- гипотезы *о типе зависимости* признаков (например, о линейной зависимости).

Проверка гипотезы о законе распределения случайной величины

Каждая случайная величина подчиняется определенному закону распределения, т.е. правилу, по которому задается соответствие между значением случайной величины и вероятностью ее появления. Закон распределения может быть задан таблично или в виде функции распределения.

Гипотезы о законе распределения заключаются в предположении о том, что распределение в генеральной совокупности подчиняется какому-то определенному закону. Каждый закон распределения описывает процессы разной вероятностной природы и характеризуется специфическими параметрами:

– *равномерное распределение* – n случайных чисел выпадает с одной и той же вероятностью $p=1/n$; характеризуется нижней и верхней границей;

– *биномиальное распределение* моделирует взаимосвязь числа успешных испытаний m и вероятностей успеха каждого испытания p при общем количестве испытаний n; частным случаем является распределение Бернулли; используется при статистическом контроле качества продукции массового производства, в теории массового обслуживания, теории стрельбы и др. областях практической деятельности;

– *нормальное (гауссово) распределение* описывает процессы, в которых на результат воздействует большое число независимых случайных факторов, среди которых нет сильно выделяющихся; характеризуется двумя параметрами - математическим ожиданием и стандартным отклонением;

нормальное распределение с $\mu = 0$ и $\sigma = 1$ называется *стандартным нормальным распределением*; логарифмически нормальное распределение одномодально, имеет правостороннюю асимметрию и положительный эксцесс, при уменьшении параметра σ асимметрия и эксцесс уменьшаются. Логнормальное распределение используется для описания многих социально-экономических ситуаций (например, зарплата работника, доход семьи, сумма банковских вкладов, размер наследства).

– *экспоненциальное (показательное) распределение* моделирует временные задержки между событиями, описывает процессы в задачах массового обслуживания и в задачах с «временем жизни»; описывается параметром $\lambda = 1/x^*$; характеризуется близостью значений выборочных оценок средней и стандартного отклонения;

– *распределение Пуассона* характеризуется параметром λ; предсказывает число случайных событий на определенном отрезке времени или на определенном пространстве; тесно связано с показательным распределением и распределением Эрланга; позволяет аппроксимировать биномиальное распределение; используется для описания числа сбоев ЭВМ, отказов сложной системы, заявок на обслуживание, несчастных случаев, редких заболеваний и др.

– *хи-квадрат распределение Пирсона* с ν степенями свободы; одномодально; используется для описания расхождения выборочных распределений от нормального, построения интервальных статистических оценок и статистических критериев;

– *распределение Стьюдента* с ν степенями свободы используется для описания расхождения выборочных распределений от нормально распределенных результатов наблюдения, построения интервальных оценок параметров и статистических критериев проверки гипотез при малом объеме выборки;

– *F-распределение* (*Фишера- Снедекора, распределение дисперсионного отношения*) характеризуется параметрами ν_1 - число степеней свободы числителя ν_2 -число степеней свободы знаменателя, имеет правостороннюю асимметрию; при возрастании степеней свободы асимметрия уменьшается; используется при проверке гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей, в дисперсионном, регрессионном и многомерном статистическом анализе;

– *гамма-распределение* задается параметрами α и β; используется для изучения случайных величин, имеющих умеренную правостороннюю асимметрию (в экономике – для описания доходов и сбережений населения); гамма распределение с параметром масштаба β =1 называют стандартным гамма-распределением. При α =1 гама – распределение совпадает с экспоненциальным распределением. При α = m, где m – целое положительное число, гамма распределение называется распределением Эрланга m-- го порядка.

– *бета-распределение* задается параметрами α и β; широко используется в математической статистике, в сетевом планировании и управлении для описания времени выполнения различных работ.

Проверка гипотезы состоит в том, чтобы на основе сравнения фактических (эмпирических) и предполагаемых (теоретических) частот сделать вывод о соответствии фактического распределения гипотетическому распределению. Имеются данные наблюдения $x_1, x_2 \ldots\ldots x_j, \ldots\ldots x_n$ над случайной величиной X, функция распределения которой неизвестна. Выдвигается гипотеза о том, что истинной функцией распределения исследуемой случайной величины *X* является некоторая заданная функция F(x).

Если гипотеза верна, то найденная по данным наблюдения эмпирическая функция распределения $F^*(x)$ не должна сильно отличаться от гипотетической функции распределения, и с увеличением объема *n* выборки различие между ними должно уменьшаться. В связи с этим вопрос о принятии или отклонении проверяемой гипотезы решается в зависимости от того насколько хорошо

согласуются эмпирическая $F^*(x)$ и гипотетическая $F(x)$ функции распределения. Статистические критерии, базирующиеся на таком подходе, называются *критериями согласия*. В основе этих критериев лежит выбранная соответствующим образом статистика, которая может служить *мерой расхождения* между эмпирическим и гипотетическим законами распределения исследуемой случайной величины. Различные критерии согласия отличаются друг от друга видом функциональной зависимости меры расхождения статистики от элементов выборки $x_1, x_2 \ldots\ldots x_j, \ldots\ldots x_n$.

Существует несколько критериев, используемых для проверки соответствия эмпирического и теоретического распределений. Например, критерии К. Пирсона (хи- квадрат), В.И. Романовского, А.Н. Колмогорова, критерий омега –квадрат, Крамера–Мизеса–Смирнова, Андерсона- Дарлигна, Шапиро-Уилка, и др. Каждый из этих критериев имеет свои особенности применения. Так, проверка с помощью *критерия хи-квадрат Пирсона* осуществляется по предварительно сгруппированным данным, при *большом* объёме выборки и теоретических частотах в интервалах более 5. В *критерии Колмогорова* за «расстояние» принимается экстремальное значение (максимальное) значение разности между накопленными эмпирическими и теоретическими частотами. *Критерий Крамера–Мизеса–Смирнова* же учитывает отклонение между эмпирическим и теоретическим распределениями при всех возможных значениях исследуемой величины, то есть является более мощным критерием. *Критерий Шапиро-Уилка* используется для проверки гипотезы о нормальном и показательном распределениях при малом объёме выборки.

Проверка гипотез о числовых значениях параметров совокупности

Наиболее распространены в практике экономических исследований проверка гипотез о математических ожиданиях (средних величинах) и о дисперсиях.

Основные гипотезы о математических ожиданиях (средних величинах) следующие:

–гипотеза о значении математического ожидания нормальной случайной величины с известной и неизвестной дисперсией:

– гипотеза о разности математических ожиданий двух независимых нормальных случайных величин с известными или неизвестными дисперсиями;

–гипотеза о разности математических ожиданий двух коррелированных нормальных случайных величин с неизвестными дисперсиями.

В большинстве случаев при анализе генеральных совокупностей дисперсии сравниваемых величин неизвестны. Поэтому перед проверкой гипотезы о математических ожиданиях проверяется гипотеза о равенстве дисперсий.

Проверка гипотезы о значении математического ожидания нормальной случайной величины с неизвестной дисперсией

Для проверки гипотезы $H_0 : \mu = \mu_0$ о том, что среднее μ нормальной случайной величины X равно заданному числу μ_0, используется статистика: $T = \frac{\bar{x}^* - \mu_0}{\sigma_x} \cdot \sqrt{n}$, называемая отношением Стьюдента, где n – объем выборки, $\bar{x}^*$ и σ_x– выборочные оценки среднего μ и стандартного отклонения σ нормальной случайной величины X.

Если проверяемая гипотеза $H_0 : \mu = \mu_0$ верна, то статистика T имеет распределение Стьюдента (t –распределение) с n-1 степенями свободы.

Расчетное значение t статистики T вычисляется по выше приведенной формуле подстановкой в нее гипотетического значения μ_0 математического ожидания исследуемой случайной величины X и числовых значений выборочных оценок $\bar{x}^*$ и σ_x, найденных по данным конкретной выборки.

а) При альтернативе $H_1 : \mu > \mu_0$, для H_0 «опасны» большие значения статистики T. Поэтому критическая область размера α имеет вид $\Omega_{кр.}(\alpha) = [t(\alpha; n-1), \infty)$, где $t(\alpha; n-1)$- критическое значение распределения Стьюдента при заданном уровне значимости α и с (n-1) степенями свободы. При $t \geq t(\alpha; n-1)$- проверяемая гипотеза отклоняется, при $t < t(\alpha; n-1)$- принимается (t– расчетное значение статистики T).

б) при альтернативе $H_1: \mu < \mu_0$ для H_0 «опасны» большие по абсолютной величине отрицательные значения статистики *T*. При этом критическая область порядка α $\Omega_{кр.}(\alpha) = (-\infty; -t(\alpha; n-1)]$.

в) при альтернативе $H_1: \mu \neq \mu_0$ опасны большие по абсолютной величине значения статистики *T*.При этом критическая область порядка α состоит из двух интервалов $(-\infty; -t(\alpha/2; n-1)]$ и $[t(\alpha/2; n-1); \infty)$. То есть область принятия гипотезы $\Omega_{пр.} = (-t(\alpha/2; n-1); +t(\alpha/1; n-1))$, где $t(\alpha/2; n-1)$ - критическое значение распределения Стьюдента значимости α/2 (0,025) и с (n-1) степенями свободы.

Проверка гипотезы о разности математических ожиданий (средних значений) двух независимых нормальных случайных величин с различными неизвестными дисперсиями

При проверке гипотезы $H_0: \mu_x - \mu_y = \delta$ о том, что разность между математическими ожиданиями μ_x и μ_y независимых нормальных случайных величин X и Y с различными неизвестными дисперсиями равна заданному числу дельта δ, используется статистика Фишера - Беренса:

$$T = \frac{\bar{x}^* - \bar{y}^* - \delta}{\sqrt{\frac{\sigma_x^{\,2}}{n} + \frac{\sigma_y^{\,2}}{m}}},$$

где $\bar{x}^*$ и $\bar{y}^*$; σ_x^2 и σ_y^2 – выборочные средние значения и дисперсии исследуемых случайных величин X и Y; $\delta = \mu_x - \mu_y$ – гипотетическое значение разности математических ожиданий этих величин; n и m – объёмы выборок, по которым проверяется гипотеза H_0.

В том случае, когда проверяемая гипотеза $H_0: \mu_x - \mu_y = \delta$ верна, статистика T имеет распределение Стьюдента с числом степеней свободы:

$$\nu = (\sigma_x^2/n + \sigma_y^2/m) / \left[(\sigma_x^2/n)^2/(n-1) + (\sigma_y^2/m)^2/(m-1)\right].$$

Расчётное значение t статистики T вычисляется по вышеприведённой формуле подстановкой в неё гипотетического значения δ и конкретных

числовых значений выборочных оценок $\bar{x}^*, \bar{y}^*, \sigma_x^2, \sigma_y^2$, найденных по данным конкретных выборок.

а) При альтернативной гипотезе $H_1 : \mu_x - \mu_y > \delta : \Omega_{кр}(\alpha) = [t(\alpha;\nu); \infty)$, где $t-$ расчётное значение статистики T; $t(\alpha;\nu)-$ критическое значение -критерия Стьюдента при уровне значимости α и ν степенями свободы.

б) При альтернативе $H_1 : \mu_x - \mu_y < \delta : \Omega_{кр}(\alpha) = (-\infty; -t(\alpha;\nu)]$.

в) При альтернативе $H_{1;} : \mu_x - \mu_y \neq \delta : \Omega_{пр}(-t\alpha/2;\nu), t(\alpha/2;\nu))$, $\Omega_{кр}(\alpha) = (|T| \geq t(\alpha/2;\nu)$, где $t(\alpha/2;\nu)-$критическое значение t-критерия Стьюдента при уровне значимости $\alpha/2$ и ν степенями свободы.

В тех случаях, когда коррелированность (сопряженность) пар наблюдений не вызывает сомнения проверяется *гипотеза о разности математических ожиданий двух коррелированных нормальных случайных величин с неизвестными дисперсиями* $H_0 : \mu_x - \mu_y = \delta$. В формуле критерия проверки данной гипотезы используются не сами реализации x_i и y_i исследуемых случайных величин X и Y, а их разности $z_i = x_i - y_i$, которые рассматриваются как реализации случайной величины *Z*. Такая процедура позволяет нейтрализовать все случайные факторы, влияющие на эти величины, кроме того фактора, воздействие которого составляет предмет исследования.

$$T = \frac{\bar{z}^* - \delta}{\sigma_z^2} \cdot \sqrt{n}$$

Сопряженные пары наблюдения образуются, например, когда фиксируются два значения исследуемой характеристики до воздействия на объект и после (обследование одной и той же группы обучающихся (работающих) в начале и в конце срока обучения (работы)). Корреляционная зависимость между двумя группами данных выявляется на основе выборочного коэффициента корреляции.

Если проверяемая гипотеза $H_0 : \mu = \mu_0$ верна, то статистика Z имеет распределение Стьюдента (*t* –распределение) с *n*-1 степенями свободы.

Проверка гипотезы аналогична проверке гипотезы о разности математических ожиданий двух независимых нормальных случайных величин.

Виды гипотез о дисперсиях:

1-*гипотеза о значении дисперсии нормальной случайной величины*: $H_0 : \sigma^2{}_x = \sigma_0^2$ о том, что дисперсия $\sigma^2{}_x$ нормальной случайной величины X равна заданному числу σ_0^2. При ее проверки используется статистика $Z = \frac{\sigma_x^2}{\sigma_0^2} \times (n-1)$.

Если проверяемая гипотеза верна, то статистика Z имеет χ^2-распределение с n-1 степенями свободы.

а) При альтернативе $H_1 : \sigma_x^2 > \sigma_0^2$ критическая область размера α имеет вид $\Omega_{кр.}(\alpha) = \left[\chi^2(\alpha; n-1), \infty\right)$, где $\chi^2(\alpha; n-1)$- критическое значение хи-квадрат -распределения при заданном уровне значимости α и с (n-1) степенями свободы.

б) при альтернативе $H_1 : \sigma_x^2 < \sigma_0^2$ $\Omega_{пр.}(\alpha) = \left(0; \chi^2{}_н(\alpha; n-1)\right]$, где $\chi^2{}_н(\alpha; n-1)$- нижнее критическое значение порядка α хи-квадрат -распределения с (n-1) степенями свободы.

в) при альтернативе $H_1 : \mu \neq \mu_0$ область принятия гипотезы $\Omega_{пр.} = \left(\chi_н^2(\alpha/2; n-1); \chi^2(\alpha/1; n-1)\right)$, где $\chi_н^2(\alpha/2; n-1)$ и $\chi^2(\alpha/2; n-1)$- нижнее и верхнее критические значения хи-квадрат-распределения значимости α/2 (0,025) и с (n-1) степенями свободы.

2–*гипотеза о равенстве дисперсий двух независимых нормальных случайных величин* X и Y $H_0 : \sigma^2{}_x = \sigma_y^2$. При проверке этой гипотезы используется статистика $F = \frac{\sigma_x^2}{\sigma_y^2}$, называемая дисперсионным отношением. $\sigma^2{}_x$ и $\sigma^2{}_y$ - несмещенные оценки дисперсий исследуемых нормальных случайных величин X и Y, найденные по данным двух независимых выборок объемов n и m.

В том случае, когда проверяемая гипотеза верна, статистика *F* имеет распределение Фишера – Снедекора (*F*- распределение) с *n* -1 и *m-1* степенями свободы.

Расчетное значение *f* статистики *F* вычисляется по формуле путем подстановки в нее числовых значений выборочных дисперсий $\sigma^2{}_x$ и $\sigma^2{}_y$. На практике расчетное значение *f* определяется как отношение большей выборочной дисперсии к меньшей (то есть случайная величина с большей выборочной дисперсией обозначается через X, а случайная величина с меньшей выборочной дисперсией –через Y). При этом отпадает необходимость рассматривать альтернативу $H_1 : \sigma_x^2 < \sigma_y^2$ и упрощается проверка при альтернативе $H_1 : \sigma_x^2 \neq \sigma_y^2$.

а) при альтернативе $H_1 : \sigma_x^2 > \sigma_y^2 : \Omega_{кр.} = \left[f(\alpha; n-1; m-1); \infty \right)$, где $f(\alpha; n-1; m-1)$ - критическое значение распределения Фишера – Снедекора. (если расчетное значение *f* попадает в критическую область, это значит, что нулевая гипотеза противоречит фактическим данным наблюдения и ее следует отклонить и принять альтернативную гипотезу $H_1 : \sigma_x^2 > \sigma_y^2$.

б) при альтернативе $H_1 : \sigma_x^2 \neq \sigma_y^2 : \Omega_{кр.} = \left[f(\alpha/2; n-1; m-1); \infty \right)$.

3– *гипотеза о равенстве дисперсий* $H_0 : \sigma^2{}_x = \sigma_y^2$ *двух коррелированных нормальных случайных величин* X и Y. Для проверки используется статистика $T = \dfrac{\sigma_x^2 - \sigma_y^2}{\sqrt{4\sigma_x^2 \sigma_y^2 (1-r^2)/(n-2)}}$,где *r*– выборочный коэффициент корреляции нормальных случайных величин X и Y.

Если проверяемая гипотеза верна, статистика *T* имеет распределение Стьюдента с *n*-2 степенями свободы.

4- *гипотеза о равенстве дисперсий нескольких независимых нормальных случайных величин*. Наиболее распространенными критериями проверки

гипотезы $H_0 : \sigma^2_1 = \sigma^2_2 = ... = \sigma^2_k$ о равенстве дисперсий k независимых случайных величин являются критерии Бартлета и Кокрена.

Следует отметить, что многие критерии чувствительны к отклонениям распределений исследуемых случайных величин от нормального распределения. Значимость используемых статистик может указывать не на отсутствие однородности дисперсий, а на отклонение от нормальности.

Другие задачи статистической обработки выборки имеют самостоятельное значение и решаются в рамках соответствующих статистических методов. К ним относятся:

– оценка влияния на выборки одного, двух или более качественных факторов – производится на основе *дисперсионного анализа*;

– выявление степени связи между выборками - реализуется посредством *корреляционного анализа*;

– установление формы зависимости между выборкой (случайной переменной Y) и одной или несколькими независимыми переменными величинами – осуществляется в процессе *регрессионного анализа*.

3. ОСНОВЫ ДИСПЕРСИОННОГО АНАЛИЗА

Дисперсионный анализ - статистический метод, который применяется для исследования влияния одного или нескольких качественных факторов, измеренных в порядковой или номинальной шкале, на исследуемую случайную величину - зависимую количественную переменную (результативный признак).

В зависимости от числа факторов модели дисперсионного анализа подразделяются на *однофакторные* и *многофакторные* (двухфакторные и т.д.).

Конкретные значения каждого из факторов принято называть *уровнями*. Множество уровней – свое для каждого из факторов, оно фиксировано и конечно. Допускаются и факторы с количественными характеристиками, но область изменения каждого из таких факторов – это тоже индивидуальное, фиксированное и конечное числовое множество.

Основная задача дисперсионного анализа заключается в том, чтобы по результатам наблюдений над исследуемой случайной величиной *Y* оценить зависимость ее математического ожидания (средней величины) от рассматриваемых факторов. Эта задача решается путем сравнения выборочной дисперсии, вызванной воздействием рассматриваемого фактора (факторов), с выборочной дисперсией, обусловленной чисто случайными причинами (воздействием неконтролируемых в данном исследовании факторов, ошибками измерений и т.п.). Если различие между дисперсиями значимо, то считают, что рассматриваемый фактор (факторы) оказывает существенное влияние на исследуемую случайную величину (результативный признак).

То есть процедура дисперсионного анализа основана на проверке гипотезы о равенстве дисперсий двух независимых нормальных случайных величин $H_0 : \sigma^2_1 = \sigma^2_2$, где для сравнения дисперсий применяется критерий, предложенный Рональдом Фишером, который называется дисперсионным отношением или *F*-критерием: $F = \frac{\sigma_1^2}{\sigma_2^2}$.

F-критерий строится так, что в числителе стоит большая дисперсия, поэтому величина дисперсионного отношения может быть равна или больше 1. Если *F*-критерий=1, то это указывает на равенство дисперсий, и вопрос об оценке существенности их расхождений снимается. Если величина *F*-критерия > 1, то возникает необходимость оценить, случайно ли расхождение между дисперсиями. При этом очевидно, что, чем больше величина F-критерия, тем значительнее расхождения между дисперсиями.

Для определения границ случайных колебаний отношения дисперсий Фишером разработаны специальные таблицы *F* –распределения. В этих таблицах указываются критические (предельные) значения *F*-критерия для различных комбинаций числа степеней свободы числителя (т.е. большей дисперсии) df_1 и знаменателя (меньшей дисперсии) df_2, которые могут быть превзойдены с вероятностью α = 0,05 или 0,01.

Рассчитанная по фактическим данным наблюдения величина дисперсионного отношения сопоставляется с критическим значением F-критерия. Если фактическое дисперсионное отношение будет больше предельного, то лишь с вероятностью 0,05 или 0,01 можно утверждать, что различие между дисперсиями определяется случайными факторами. Однако, события, имеющие столь малую вероятность, считаются практически невозможными, а поэтому в этом случае с вероятностью 1- α (0,95 или 0,99) можно утверждать существенность различий в величине дисперсий.

Если же фактическое (расчетное) значение дисперсионного отношения меньше соответствующего критического значения, то с вероятностью 95% или 99% можно утверждать, что расхождение между дисперсиями несущественно.

Гипотеза, которая проверяется посредством данных действий, состоит в том, что сравниваемые дисперсии характеризуют вариацию признака в выборках, отобранных из одной и той же нормально распределенной генеральной совокупности, или же отобранных из нормально распределенных генеральных совокупностей с одинаковой дисперсией. То есть каждую из этих дисперсией можно рассматривать как оценку генеральной дисперсии.

В *однофакторном дисперсионном анализе* данные подразделяются по признаку-фактору (x) на m групп – по числу уровней воздействия данного фактора на результативный признак. При этом результаты наблюдений образуют m совокупностей, которые можно рассматривать как независимые случайные выборки из m нормально распределенных генеральных совокупностей с неизвестными средними μ_1, μ_2…. μ_m и одинаковыми дисперсиями. Считается, что различия в средних значениях результативного признака в выделенных группах обусловлены воздействием признака-фактора. Так как рассматриваемый фактор влияет только на среднюю величину μ результативного признака, но не влияет на его дисперсию, то при дисперсионном анализе проверяется нулевая гипотеза $H_0 = \mu_1 = ... = \mu_m$ о равенстве математических ожиданий нескольких нормальных генеральных совокупностей с одинаковыми дисперсиями ($(m \geq 3)$.

Если данная гипотеза верна, то есть расхождения между средними в разных группах несущественны, то это означает, что рассматриваемый фактор не влияет на исследуемую случайную величину *Y*. Если нулевая гипотеза не верна (хотя бы две из m средних не равны друг другу), то рассматриваемый фактор оказывает влияние на результативный признак.

Существенность различия между группами в дисперсионном анализе доказывается на основе *F*-критерия, представляющего отношение двух выборочных дисперсий: межгрупповой и внутригрупповой.

При применении дисперсионного анализа общая вариация результативного признака разбивается на две составляющие: факторную вариацию, которая характеризует влияние рассматриваемого фактора, по которому исследуемая совокупность разбита на группы, и случайную вариацию, которая характеризует влияние случайных причин, не связанных с воздействием изучаемого фактора.

Общая вариация, показывающая колеблемость результативного признака под влиянием всех возможных факторов, выражается как о*бщая (полная) сумма*

квадратов отклонений наблюдаемых значений y_i от общей средней результативного признака $\bar{y}$ – $SS_{общ.} = \sum(y_i - \bar{y})^2$.

Для характеристики вариации между группами, обусловленную влиянием рассматриваемого фактора, положенного в основу группировки, вычисляется *межгрупповая (факторная) сумма квадратов отклонений* групповых средних $\bar{y}_i$ от общей средней результативного признака $\bar{y}$ – $SS_{мг.} = \sum(\bar{y}_i - \bar{y})^2$.

Для измерения вариации внутри групп, возникающую под влиянием прочих причин, кроме влияния фактора, положенного в основу группировки, рассчитывается *внутригрупповая (остаточная, случайная) сумма квадратов отклонений* наблюдаемых значений y_i от групповых средних $\bar{y}_i$ – $SS_{вг.} = \sum(y_i - \bar{y_i})^2$.

Математически эти суммы связаны *правилом сложения*: общая сумма квадратов отклонений = межгрупповая (факторная) сумма квадратов отклонений + внутригрупповая (остаточная, случайная) сумма квадратов отклонений: $SS_{общ.} = SS_{мг.} + SS_{вг.}$.

Каждая из этих сумм имеет определенное число степеней свободы. Под *числом степеней свободы* рассматриваемой суммы понимается число независимых слагаемых, входящих в эту сумму. Число независимых слагаемых равно разности между общим числом слагаемых, по которым вычисляется рассматриваемая сумма, и числом условий, наложенных на эти слагаемые.

Разделив суммы квадратов отклонений на соответствующее число степеней свободы, получим общую, межгрупповую и внутригрупповую дисперсии. Для расчета общей дисперсии число степеней свободы равно $df_{общ.} = n - 1$, межгрупповой дисперсии $-df_{мг.} = m - 1$, внутригрупповой дисперсии – $df_{вг.} = n - m$. Как и суммы квадратов отклонений, числа степеней свободы связаны между собой равенством: $df_{\text{общ}} = df_{\text{мг.}} + df_{вг.}$.

Сравнение величин межгрупповой и внутригрупповой дисперсий, рассчитанных по конкретным данным наблюдения, дает возможность оценить существенность влияния признака-фактора на результативный признак с помощью *F*-критерия, который используется в качестве меры их расхождения: $F = \frac{\sigma^2_{мг.}}{\sigma^2_{вг.}}$. Если группировочный признак не оказывает влияние на вариацию результативного признака (а в этом случае $\mu_1 = ... = \mu_m$), то межгрупповая дисперсия будет отражать только влияние тех же самых прочих факторов, вариацию под влиянием которых измеряет внутригрупповая дисперсия. В этом случае отношение дисперсий будет близко к 1 или отличаться от нее в силу случайных колебаний, предельный размер которых можно установить по таблице *F* –распределения.

Это означает, что обе выборочные дисперсии (и межгрупповая, и внутригрупповая) являются независимыми несмещенными оценками генеральной средней σ^2.

Если группировочный признак оказывает влияние на вариацию результативного признака (в этом случае наблюдается различие между средними μ_1, μ_2…. μ_m) , то различие межгрупповой и внутригрупповой дисперсии будет существенно, то есть $F_{расч.} > 1$. Причем, чем сильнее влияние рассматриваемого фактора на исследуемую случайную величину *Y*, тем больше вероятность появления больших значений статистики *F*. Поэтому при проверке гипотезы H_0 используется правосторонняя критическая область.

Если $F_{расч.} > F_{крит.}$, то гипотеза H_0 о равенстве средних μ_1, μ_2…. μ_m отклоняется на уровне значимости α., то есть с вероятностью 0,95 или 0,99 можно утверждать, что влияние рассматриваемого фактора является существенным или, иначе говоря, статистически значимым. Если $F_{расч.} < F_{крит.}$ H_0 принимается, что означает, что рассматриваемый фактор не влияет на результативный признак.

Степень влияния фактора на исследуемый результативный признак определяется путем нахождения *коэффициента детерминации*, равного отношению межгрупповой суммы квадратов отклонений к общей сумме квадратов отклонений: $R^2 = \frac{SS_{мг.}}{SS_{общ.}}$.

При *двухфакторном дисперсионном анализе* выявляют и оценивают влияние на случайную величину *Y* двух факторов: 1) фактора *x*, который имеет *m* уровней воздействия, расположенных по строкам; 2) фактора *z*, который имеет *p* уровней воздействия, расположенных по столбцам.

Уровни фактора	Уровни фактора z			
	z_1	z_2	…	z_p
x_1	y_{111} y_{112} y_{113} …..	y_{121} y_{122} y_{123} …..		
x_2	y_{211} y_{212} y_{213} …..	y_{221} y_{222} y_{223} …..		
….				

Дисперсионный анализ может быть выполнен при предположении, что взаимодействие между факторами *x* и *z* отсутствует и что факторы *x* и *z* взаимодействуют между собой.

В первом случае общая вариация результативного признака складывается из вариации под влиянием фактора *x*, вариации под влиянием фактора *z* и вариации под влиянием неучтенных факторов.

Во втором случае – добавляется еще вариация, обусловленная взаимодействием факторов *x* и *z*.

Как и в случае однофакторного анализа, каждая вариация измеряется соответствующей суммой квадратов отклонений.

Правило сложения суммы квадратов отклонений может быть представлено в виде:

$$SS_{общ.} = SS_X + SS_Z + SS_{XZ} + SS_{ост.} = SS_{факт.} + SS_{ост.}$$

$SS_{общ.}$ – сумма квадратов отклонений всех наблюдаемых значений от общей средней $\bar{y}$;

SS_X – сумма квадратов отклонений выборочных средних по уровням фактора *x* (средним по строкам) от общей средней $\bar{y}$;

SS_Z – сумма квадратов отклонений выборочных средних по уровням фактора *z* (средним по столбцам) от общей средней $\bar{y}$;

SS_{XZ} – сумма квадратов отклонений средних по ячейкам от общей средней $\bar{y}$;

$SS_{ост.}$ - остаточная сумма квадратов отклонений.

Число степеней свободы для каждой суммы квадратов отклонений составляет:

$df_{общ.} = n-1$; $df_X = m-1$; $df_Z = p-1$; $df_{XZ} = (m-1)\cdot(p-1) = mp-m-p+1$;

$df_{факт.} = df_X + df_Z + df_{XZ} = mp-1$; $df_{ост.} = df_{общ.} - df_{факт.} = n-mp$.

Делением сумм квадратов отклонений на соответствующее им число степеней свободы определяются факторные дисперсии и остаточная дисперсия. Путем отношения факторной дисперсии по каждому фактору и их взаимодействию к остаточной дисперсии определяется расчетное значение F-критерия.

Критическое значение F- критерия определяется при заданном уровне значимости α и числе степеней свободы для фактора *x* - $df_1 = df_X = m-1$ и $df_2 = df_{ост.} = n-mp$; для фактора z – $df_1 = df_Z = p-1$ и $df_2 = df_{ост.} = n-mp$; для взаимодействия факторов – $df_1 = df_{XZ} = mp-m-p+1$ и $df_2 = df_{ост.} = n-mp$

Влияние соответствующего фактора признается значимым, если $F_{расч.} > F_{крит.}$.

В пакете STATISTICA в разделе «Основные статистические модули и таблицы» есть модуль «Дисперсионный анализ», где рассматриваются различные схемы дисперсионного анализа.

4. ОСНОВЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА

4.1 Понятие, условия применения и основные характеристики корреляционной связи количественных переменных

Современная наука исходит из взаимосвязи всех явлений природы и общества. Невозможно управлять явлениями, предсказывать их развитие без изучения характера, силы и других характеристик связей. Поэтому методы измерения связей составляют чрезвычайно важную часть методологии научного исследования, в том числе и статистического.

Различают два типа связей между явлениями и их признаками: функциональную или жестко детерминированную и статистическую, вероятностную или стохастически детерминированную.

Если с изменением значения одной переменной (факторного признака) вторая переменная (результативный признак) изменяется строго определенным образом, то есть значению одной переменной обязательно соответствует одно или несколько точно заданных значений другой переменной, связь между ними является *функциональной*. В теории вероятности о функциональной зависимости между случайными величинами X и $У$ говорят, когда значение y случайной величины $У$ является функцией значения x, принятого случайной величиной X.

В реальной природе и тем более обществе функциональная связь в чистом виде не существует, так как все явления и процессы реального мира связаны между собой, и нет такого конечного числа переменных, которое абсолютно полно определяло бы собой зависимую величину $У$. Тем не менее многие науки (как правило точные: механика, электротехника, акустика, астрономия), в том числе и экономика успешно используют представление связей как функциональных в аналитических целях, а нередко и в целях прогнозирования. Это допустимо в тех случаях, когда изучаемая переменная (результативный признак) зависит в основном (на 99% и более) от одной или

немногих других переменных, то есть связь является хотя и не абсолютно функциональной, но практически очень близкой к таковой.

Связь является *статистической (вероятностной)*, если с изменением значения одной из переменных (факторного признака) вторая переменная (результативный признак) может в определенных пределах принимать любые значения с некоторыми вероятностями, но ее среднее значение или иные статистические характеристики изменяются по определенному закону.

То есть сущность вероятностной зависимости состоит в том, что распределение одной случайной величины зависит от того, какое значение приняла другая случайная величина. При вероятностной зависимости между случайными величинами X и Y, зная значение x, которое приняла в данном наблюдении случайная величина X, нельзя сказать, какое значение y примет в этом же наблюдении случайная величина Y. Можно указать только условный закон распределения случайной величины Y.

Вероятностная зависимость между двумя случайными явлениями возникает в тех случаях, когда наряду со случайными факторами, различными для обоих явлений, имеются и общие случайные факторы, влияющие как на одно, так и на другое случайное явление.

Чем сильнее вероятностная связь, тем она ближе к функциональной связи. То есть функциональная зависимость является крайним, предельным случаем вероятностной зависимости. (Другой предельный случай – полная независимость случайных величин).

Случайные величины X и Y называются *независимыми*, если закон распределения одной из этих случайных величин не зависит то того, какое значение приняла другая случайная величина. В противном случае они называются *зависимыми*.

Таким образом, в настоящее время наука не знает более широкого определения связи. Все связи, которые могут быть измерены и выражены численно, подходят под определение «статистические, вероятностные связи», в том числе и функциональные. Последние представляют собой частный случай

статистических связей, когда значениям одной переменной соответствуют «распределения» значений другой переменной, состоящие из одного или нескольких значений и имеющие вероятность, равную единице.

Однако, качественное различие действительно вероятностных распределений и отдельных значений, имеющих вероятность единицы (достоверных), настолько велико, что хотя функциональные связи и могут рассматриваться как предельный случай статистической связи, все же с полным обоснованием можно говорить о двух типах связи.

Важной разновидностью стохастически детерминированных связей является корреляционная зависимость между случайными величинами, при которой математическое ожидание одной случайной величины зависит от того, какое значение приняла другая случайная величина.

Корреляционной называют связь, при которой разным значениям одной переменной (факторного признака) соответствуют различные средние значения другой переменной. То есть с изменением значения признака x закономерным образом изменяется среднее значение признака y, в то время как в каждом отдельном случае значение признака y (с различными вероятностями) может принимать множество различных значений.

Если же изменение значения признака x не ведет к закономерному изменению среднего значения признака y, но приводит к закономерному изменению другой статистической характеристики (показателей вариации, асимметрии, эксцесса и др.), то связь не считается корреляционной, но является статистической.

Статистическая связь между двумя переменными (признаками) предполагает, что каждая из них имеет случайную вариацию индивидуальных значений относительно средней величины. Если же такую вариацию имеет только один из признаков, а значения другого являются жестко детерминированными, то говорят лишь о регрессии (например, при анализе динамических рядов можно измерять регрессию уровней ряда на номера лет, но

нельзя говорить о корреляции между ними и применять показатели корреляции с соответствующей интерпретацией).

Корреляционная связь между признаками может возникать разными путями. Первый (важнейший) путь – это причинная зависимость результативного признака (его вариации) от вариации факторного признака. Например, зависимость урожайности сельскохозяйственной культуры от балла оценки плодородия почв. Здесь совершенно ясно логически, какой признак выступает как независимая переменная (фактор) x, а какой – как зависимая переменная (результат) - y.

Второй путь – сопряженность признаков, возникающая при наличии общей причины. Известен классический пример, приведенный крупнейшим статистиком России 20 века Чупровым: при рассмотрении зависимости между числом пожарных команд (x) и суммой убытков за год от пожаров (y) в совокупности городов России, им была выявлена высокая прямая корреляция между данными признаками: чем больше пожарных команд в городе, тем больше и убытков от пожаров. Данную корреляцию нельзя интерпретировать как связь причины и следствия; оба признака – следствия общей причины – размера города. Вполне логично, что в крупных городах больше пожарных частей, но и больше пожаров и убытков от них за год, чем в малых городах.

Третий путь возникновения корреляции – взаимосвязь признаков, каждый из которых и причина, и следствие. Например, корреляция между уровнем производительности труда рабочих и часовым уровнем заработной платы. С одной стороны – уровень заработной платы – следствие производительности труда: чем выше выработка, тем больше и оплата. Но с другой стороны, установленные тарифные ставки и расценки играют стимулирующую роль: при правильной организации системы оплаты труда они выступают в качестве фактора, от которого зависит производительность труда. В такой системе признаков допустимы обе постановки задачи: каждый признак может выступать в роли независимой переменной x и в качестве независимой переменной y.

То есть корреляционная связь не всегда должна иметь причинно – следственный характер.

Поскольку корреляционная связь является статистической, первым условием ее изучения является наличие достаточного количества наблюдений для изучения. На практике считается, что число наблюдений должно не менее чем в 5-6 раз (а лучше – не менее чем в 10 раз) превышать число факторов. В случае если число наблюдений превышает количество факторов в десятки или сотни раз, закон больших чисел обеспечивает эффективное взаимопогашение случайных отклонений от закономерного характера связи переменных.

Вторым условием закономерного проявления корреляционной связи служит качественная и количественная однородность совокупности. Качественная однородность предполагает близость условий формирования факторного и результативного признаков. Количественная однородность устанавливается по 33%-ному значению коэффициента вариации. Условие однородности совокупности (наряду с большим числом единиц совокупности) обеспечивает надежное выражение закономерности в средней величине.

Третье условие – необходимость подчинения распределения значений результативного и факторных признаков нормальному закону распределения вероятностей. Это условие связано с применением метода наименьших квадратов (МНК) при расчете параметров корреляции. В основу этого метода (разработан Гауссом (1777-1855)) положено требование минимальности суммы разности квадрата отклонений фактически измеренных значений зависимой переменной у от ее значений, вычисленных по уравнению связи с факторным признаком, одни или несколькими.: $\Sigma(y_i - y_{xi})^2 \to \min$. Нормальность распределения проверяется на основе критериев согласия.

Только при нормальном распределении метод наименьших квадратов дает оценки параметров, отвечающих принципам максимального правдоподобия. На практике эта предпосылка чаще всего выполняется приближенно, но и тогда метод наименьших квадратов дает неплохие результаты. Однако при значительном отклонении распределений признаков от

нормального закона нельзя оценивать надежность выборочного коэффициента корреляции, используя параметры нормального распределения вероятностей или распределения Стьюдента.

Различают следующие виды корреляционной зависимости: парная, частная и множественная.

Корреляционные связи между переменными имеют ряд характеристик: форма, характер, степень тесноты (сила). *По форме* (аналитическому выражению) корреляционные связи между признаками могут быть линейными (прямолинейными) и нелинейными (криволинейными). При *линейной* форме равномерное изменение значений одного признака сопровождается более или менее равномерным изменением значений другого признака. Математически она выражается уравнением прямой $y_x = a + вх$, графически - прямой линией. При *нелинейной* форме равномерному изменению значений одного признака соответствует неравномерное изменение значений другого. Выражается уравнением какой- либо кривой линии: параболы, гиперболы, показательной, степенной, логарифмической, логической функции и др.

По направлению (характеру изменения) корреляционные связи бывают прямыми и обратными. *Прямой* (положительной) является зависимость, при которой направление изменения значений факторного и результативного признаков совпадает, то есть с увеличением факторного признака, результативный также возрастает, и, наоборот, при уменьшении факторного признака результативный тоже убывает. *Обратной* (отрицательной) называется связь, при которой изменение значений факторного и результативного признаков осуществляется в разных направлениях, то есть с ростом факторного результативный признак убывает или при убывании факторного признака результативный возрастает.

Степень тесноты корреляционной связи измеряется с помощью показателей корреляции и оценивается по специальным шкалам, содержащим условные градации значений этих показателей.

4.2 Основные этапы корреляционного анализа

Статистическим исследованием зависимостей занимаются такие разделы математической статистики как корреляционный и регрессионный анализ. В этих видах анализа много общих вычислительных процедур. Различие между ними заключается в том, что корреляционный анализ оценивает силу корреляционной зависимости, тогда как регрессионный анализ исследует ее конкретную форму.

Корреляционным анализом называются методы обнаружения и оценки корреляционной зависимости между случайными величинами или признаками по статистическим данным, полученным в процессе наблюдения. Он включает в себя следующие основные практические приемы:

1) построение корреляционного поля и составление корреляционной таблицы;

2) вычисление выборочных коэффициентов корреляции и корреляционных отношений;

3) проверка статистических гипотез о значимости корреляционной зависимости.

В случае исследования зависимости между двумя случайными величинами X и Y, мы имеем n пар совместных реализаций этих величин (x_1, y_1), (x_2, y_2) и т.д., то есть двумерные выборочные данные.

Как правило, первым шагом к систематизации этих данных является построение корреляционного поля (поле или диаграмма рассеивания). *Корреляционное поле* – это точечный график, построенный в координатной плоскости *Oxy*. Положение каждой точки на графике определяется величиной двух признаков – факторного и результативного. Характер расположения точек этого поля дает наглядное представление о направлении, форме и силе зависимости между исследуемыми случайными величинами.

Если точки корреляционного поля вытянуты определенной полосой слева на право, то это указывает на положительную корреляцию (с ростом

значений факторного признака значения результативного признака тоже увеличиваются), если справа налево - отрицательная корреляция.

По виду воображаемой кривой, проходящей «наилучшим образом» через эти точки, можно сделать предположение о форме связи (прямолинейная, параболическая, гиперболическая, степенная и др.).

Сравнительно небольшой разброс точек относительно воображаемой кривой, проходящей «наилучшим образом» через эти точки, говорит о довольно сильной зависимости между X и Y, и наоборот.

В случае небольшого объема выборки (меньше 50) на основе визуального анализа поля корреляции может быть проверена нормальность распределения. Если в расположении точек наблюдается линейная тенденция, то можно предположить, что совокупность исходных данных подчиняется закону нормального распределения.

В тех случаях, когда корреляция между признаками имеет явно выраженный нелинейный характер и объём выборки велик, данные наблюдения группируют и представляют их в виде корреляционной таблицы, состоящей из $k+2$ строк и $l+2$ столбцов, где $k-$ число интервалов группировки по факторному признаку и $l-$ число интервалов группировки по результативному признаку.

Это обусловлено тем, что при нелинейной зависимости вычисляются выборочные корреляционные отношения, которые могут быть определены только по сгруппированным данным.

Построение корреляционной таблицы начинают с группировки значений факторного и результативного признаков.

Корреляционная таблица

X	Y					n_{i*}
	$\overline{y_1}$	…	$\overline{y_j}$	…	$\overline{y_l}$	
$\overline{x_1}$	n_{11}	…	n_{1j}	…	n_{1l}	n_{1*}
…	…	…	…	…	…	…
$\overline{x_i}$	n_{i1}	…	n_{ij}	…	n_{il}	n_{i*}
…	…	…	…	…	…	…
$\overline{x_k}$	n_{k1}	…	n_{kj}	…	n_{kl}	n_{k*}
n_{*j}	n_{*1}	…	n_{*j}	…	n_{*l}	n

$\overline{x_i}$ – середина i-го интервала группировки по факторному признаку;

$\overline{y_j}$ – середина j-го интервала группировки по результативному признаку;

n_{ij} – групповая частота «клетки», находящейся на пересечении строки $\overline{x_i}$ и столбца $\overline{y_j}$ корреляционной таблицы;

n_{i*} – групповая частота i-го интервала группировки по факторному признаку (число наблюдений в i-й строке);

n_{*j} –групповая частота j-го интервала группировки по результативному признаку (число наблюдений в j-м столбце);

n –объём изучаемой совокупности (общее число наблюдений).

Заполнение корреляционной таблицы даёт довольно наглядное представление о характере зависимости между изучаемыми признаками

Если частоты в корреляционной таблице расположены на диагонали из левого верхнего угла в правый нижний угол (как в примере), то можно предположить наличие прямой корреляционной зависимости между признаками. Если же частоты расположены по диагонали справа налево, то предполагают наличие обратной связи между признаками.

Для определения силы связи необходимо произвести расчет числовых характеристик: выборочных коэффициента корреляции и корреляционного отношения.

Показатели тесноты связи дают возможность охарактеризовать зависимость вариации результативного признака от вариации одного или нескольких факторных признака. Знание показателей тесноты корреляционной связи позволяет решать следующие вопросы:

1) обосновать необходимость изучения данной связи между признаками и целесообразность ее практического применения;

2) на основе сопоставления показателей тесноты связи для различных выборок, можно судить о степени различий в ее проявлении для конкретных условий;

3) на основе сопоставления показателей тесноты связи результативного признака с различными факторами, можно выявить те факторы, которые в данных конкретных условиях являются решающими и главным образом воздействуют на формирование величины результативного признака.

Числовой характеристикой линейной корреляционной зависимости двумерной выборки является выборочный коэффициент корреляции. Линейный коэффициент корреляции был предложен в начале 1890-х гг. английским статистиком Пирсоном (поэтому известен под его именем).

Выборочный коэффициент корреляции может быть вычислен по данным корреляционной таблицы и по первичным (несгруппированым) данным. В теории разработаны и на практике применяются различные модификации формулы расчёта данного коэффициента:

$$r = \frac{(x-\bar{x})\cdot(y-\bar{y})}{\sigma_x \cdot \sigma_y};\quad r = \frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sigma_x\sigma_y};\quad r = \frac{\sum(x-\bar{x})\cdot(y-\bar{y})}{n\sigma_x\sigma_y};$$

$$r = \frac{n\cdot\sum xy-\sum x\sum y}{\sqrt{\left[n\sum x^2-\left(\sum x\right)^2\right]\left[n\sum y^2-\left(\sum y\right)^2\right]}};\quad r = \frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{\left[\sum x^2-\frac{\left(\sum x\right)^2}{n}\right]\left[\sum y^2-\frac{\left(\sum y\right)^2}{n}\right]}},\ \text{где}$$

где $\left(x-\bar{x}\right)\cdot\left(y-\bar{y}\right)$ – ковариация факторного и результативного признаков;

σ_x, σ_y – среднее квадратическое (стандартное) отклонение соответственно факторного и результативного признака;

n – число наблюдений.

Коэффициент корреляции обладает следующими свойствами:

1) r симметричен относительно случайных величин X и Y, то есть $r_{xy} = r_{yx}$;

2) абсолютная величина r не превышает единицу: $-1 \le r \le 1$. При отрицательных значениях r говорят об отрицательной корреляции, характеризующейся обратным направлением связи между признаками, при положительных значениях r об положительной корреляции, указывающей на прямую связь между признаками.

3) если случайные величины X и Y независимы, то r =0;

4) равенство $|r| = 1$ имеет место тогда и только тогда, когда X и Y связаны линейной функциональной зависимостью;

5) коэффициент корреляции случайных величин X и Y не зависит от выбора начала отчета и единиц измерения этих случайных величин.

При словесном описании линейной корреляционной зависимости используются следующие градации:

Величина коэффициента корреляции	Характер связи
до |0,2|	очень слабая
|0,2|-|0,4|	слабая
|04|-|0,7|	средняя, умеренная
|0,7|-|0,9|	сильная
|0,9|-|1|	очень сильная

Случайные величины, коэффициент корреляции которых равен нулю, называются *линейно некоррелированными*. Независимые случайные величины всегда некоррелированы. Обратное утверждение в общем случае неверно – некоррелированные случайные величины могут быть связаны не только вероятностной, но и функциональной связью. Некоррелированные случайные величины независимы только в том случае, когда они имеют двумерное нормальное распределение.

Квадрат коэффициента корреляции называется *коэффициентом детерминации*. Он показывает долю общей дисперсии результативного признака, которая объясняется вариацией признака-фактора.

Коэффициент корреляции достаточно точно оценивает степень тесноты связи лишь в случае наличия линейной зависимости между признаками. При наличии криволинейной зависимости он недооценивает степень тесноты связи и даже может быть равен нулю.

Поэтому в качестве математической меры нелинейной_корреляционной зависимости между случайными величинами X и Y используется корреляционное отношение, являющее более универсальным показателем. Значения корреляционного отношения зависят только от силы корреляционной зависимости и совсем не зависят от того, линейна или нелинейна эта зависимость. В этом важное преимущество корреляционного отношения перед коэффициентом корреляции.

Расчет эмпирического корреляционного отношения может быть осуществлен по данным корреляционной таблицы, а также по сгруппированным данным наблюдения. Второй способ, основанный на построении групповой таблицы по результатам аналитической группировки, наиболее распространен в практике статистических расчетов и основан на использовании теоремы (правила) сложения дисперсий:

$$\sigma^2 общ = \sigma^2 межгр + \overline{\sigma}^2 внутргр$$

Эмпирическое корреляционное отношение определяется по формуле:

$$\eta = \sqrt{\frac{\sigma^2 межгр}{\sigma^2 общ.}} = \sqrt{1 - \frac{\overline{\sigma}^2 внутргр}{\sigma^2 общ.}}$$

Межгрупповая дисперсия характеризует ту часть колеблемости результативного признака, которая складывается под влиянием изменения факторного признака, положенного в основание группировки:

$$\sigma^2 межгр = \frac{\sum (\overline{y}_i - \overline{y})^2 \cdot n_i}{\sum n_i}$$

Средняя из внутригрупповых дисперсий оценивает ту часть вариации результативного признака, которая обусловлена действием других, прочих, «случайных» причин:

$$\bar{\sigma}^2 вн = \frac{\sum \sigma_i^2 вн. \cdot n_i}{\sum n_i}$$, где

$\sigma_i^2 внутргр$-дисперсия результативного признака в соответствующей группе.

$$\sigma_i^2 = \frac{\sum (y_i - \overline{y_i})^2}{n_i};$$

Общая дисперсия характеризует вариацию результативного признака, обусловленную влиянием всех факторов:

$$\sigma_{общ.}^2 = \frac{\sum (y_i - \overline{y})^2}{n}$$

Величина η будет равна нулю, когда межгрупповая дисперсия равна нулю, то есть нет колоблемости средних по выделенным группам. В тех случаях, когда внутригрупповая дисперсия близка к нулю (то есть практически вся вариация результативного признака обусловлена действием признака *x*), величина η близка к единице. То есть для определения степени тесноты связи применяются те же шкалы, что и для коэффициента корреляции.

Направление связи устанавливается по данным корреляционной или групповой таблиц.

Свойства корреляционного отношения:

1) корреляционное отношение несимметрично относительно *X* и *У*, то есть в общем случае $\eta_{yx} \neq \eta_{xy}$.

2) корреляционное отношение неотрицательно и не превосходят единицу;

3) равенство $\eta_{yx} = 0$ означает, что случайная величина *У* не коррелированна со случайной величиной *X*. Однако некоррелированность *У* с *X* не влечет за собой некоррелированности *X* с *У*. Возможны случаи, когда одно из корреляционных отношений равно нулю, тогда как другое равно единице.

4) если случайные величины независимы, то $\eta_{yx} = \eta_{xy} = 0$. Обратное утверждение в общем случае неверно, то есть из некоррелированности случайных величин не следует их независимость.

5) равенство $\eta_{yx} = \eta_{xy} = 1$справедливо тогда и только тогда, когда случайные величины X и Y связаны функциональной зависимостью. При этом если $|r| = \eta_{yx} = \eta_{xy} = 1$-то функциональная зависимость линейна. Если же $|r| \prec \eta_{yx} = \eta_{xy} = 1$, то функциональная связь между X и Y не линейна.

6) если корреляция нелинейна, то $|r| < \min(\eta_{yx}, \eta_{xy})$. При этом, чем меньше разность между η^2 и r^2, тем ближе к линейной корреляционная зависимость между исследуемыми случайными величинами. Считается возможным говорить от линейной зависимости при $\eta^2 - r^2 < 0{,}1$.

Следует учитывать, что сопоставление коэффициента корреляции и корреляционного отношения имеет смысл, если эти показатели вычислены одинаковым образом: либо по данным корреляционной таблицы, либо по первичным данным и групповой таблицы, что предпочтительнее.

Для измерения тесноты связи между результативным признаком и несколькими факторами, используется показатель *совокупный индекс корреляции или индекс множественной корреляции:*

$$J_{y123\ldots m} = \sqrt{\frac{\sigma^2_{y123\ldots m}}{\sigma^2_y}} = \sqrt{1 - \frac{\delta^2_{y(123\ldots m)}}{\sigma^2_y}}$$, где

$\sigma^2_{y123\ldots m}$ - факторная дисперсия, характеризует вариацию результативного признака, которая при данной форме связи объясняется влиянием факторов $x_1, x_2, x_3, \ldots, x_m$, включенными в исследование:

$$\sigma^2_{y123\ldots m} = \frac{\Sigma(\tilde{y}_i - \bar{y})^2}{n-1}$$

$\delta^2_{y(123\ldots m)}$ - остаточная (случайная) дисперсия, характеризует вариацию результативного признака, обусловленную влиянием прочих, не включенных в исследование:

$$\delta^2_{y(123\ldots m)} = \frac{\Sigma(y_i - \tilde{y}_i)^2}{n-1}$$

$\delta^2_y = \frac{\Sigma(y_i - \bar{y}_i)^2}{n-1}$ - общая дисперсия результативного признака;

$\tilde{y}_i$ - теоретические (расчетные по уравнению регрессии) значения результативного признака;

y_i – эмпирическое значение результативного признака;

$\bar{y}$ - среднее значение результативного признака.

Квадрат индекса множественной корреляции называется *совокупным индексом детерминации:* $J^2_{y123...m} = \frac{\sigma^2_{y123...m}}{\sigma^2_y}$

Он показывает, какая часть общей вариации результативного признака объясняется вариацией факторов, включенных в исследование.

Индексы множественной корреляции и совокупный индекс детерминации могут быть использованы для измерения тесноты связи как при линейной, так и при криволинейной форме связи. При линейной форме связи показатели тесноты связи называются коэффициентами и имеют тот же смысл, что и индексы множественной корреляции и детерминации.

Кроме выше указанной формулы множественный коэффициент корреляции можно рассчитать, используя парные коэффициенты корреляции и коэффициенты регрессии в стандартизированном масштабе:

$$R_{y123...m} = \sqrt{\beta_1 r_{yx_1} + \beta_2 r_{yx_2} + ... + \beta_m r_{yx_m}}$$

а так как $\beta_i = a_i \frac{\sigma_i}{\sigma_y}$, то

$$R_{y123...m} = \sqrt{\frac{a_1\sigma_1 r_{yx_1} + a_2\sigma_2 r_{yx_2} + ... + a_m\sigma_m r_{yx_m}}{\sigma_y}},$$

где R – коэффициент множественной корреляции;

a_i - коэффициент уравнения регрессии при факторе x_i;

σ_i - среднее квадратическое отклонение фактора x_i;

r_{yx_i} - парный коэффициент корреляции между результативным признаком и фактором x_i.

Множественный коэффициент корреляции изменяется в пределах от 0 до 1; равенство его нулю говорит об отсутствии линейной связи; приближение к единице свидетельствует о сильной зависимости между признаками.

При небольшом числе наблюдений величина коэффициента множественной корреляции, как правило, завышается и подлежит корректировке на основании следующего выражения:

$$\hat{R}_{y123...m} = \sqrt{1-(1-R^2)\cdot\frac{n-1}{n-m-1}}$$

$\hat{R}_{y123...m}$ - скорректированное значение коэффициента множественной корреляции;

n - число наблюдений;

m - число факторных признаков.

Корректировка R не производится при условии, если $\frac{n-m}{m} \geq 20$, m – число факторных признаков.

В отношении корреляционной зависимости проверяются следующие гипотезы:

– гипотеза о некоррелированности двух нормальных случайных величин;

– гипотеза о значении коэффициента корреляции двух нормальных случайных величин;

-гипотеза о равенстве коэффициентов корреляции двух двумерных нормальных совокупностей;

- гипотеза о равенстве нескольких коэффициентов корреляции.

Проверка первой гипотезы связана с тем, что показатели корреляционной связи, вычисляемые по ограниченной совокупности (по выборке), являются лишь оценками статистической закономерности в генеральной совокупности. Отклонение от нуля полученной величины выборочного показателя тесноты связи может быть обусловлено случайными колебаниями тех выборочных данных, на основании которых он вычислен. В этом случае заключение в отношении действительного наличия корреляционной связи в генеральной совокупности, из которой была произведена выборка, будет не правомерна.

Поэтому необходима статистическая оценка надежности показателей корреляции. Под надежностью понимается вероятность того, что значение проверяемого показателя не равно нулю, не включает в себя величины противоположных знаков.

В случае линейной корреляционной зависимости некоррелированность двух случайных величин X и Y проверяется на основе гипотезы $H_0 : r = 0$ о

равенстве коэффициента корреляции в генеральной совокупности нулю (это означает, что в действительности связь между изучаемыми признаками отсутствует, а эмпирическое значение выборочного коэффициента корреляции обусловлено только случайными совпадениями x и y в выборке).

При альтернативе $H_1 : r \neq 0$ критическая область $\Omega_{кр.}(\alpha) = |T| \geq t(\alpha/2; n-2)$, а область принятия гипотезы $\Omega_{пр.} = (-t(\alpha/2; n-2); +t(\alpha/1; n-2))$, где $t(\alpha/2; n-2)$ - критическое значение распределения Стьюдента значимости α/2 (0,025) и с (n-2) степенями свободы, определяемое по таблице распределения Стьюдента. Фактическое (расчетное) значение t - критерия рассчитывается по формуле подстановкой в нее значения выборочного коэффициента корреляции:

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

Если расчётное значение t – критерия больше критического, то гипотеза $H_0 : r = 0$ о том, что линейный коэффициент корреляции в генеральной совокупности равен нулю и лишь в силу случайных обстоятельств оказался равен проверяемому значению, отклоняется, то есть коэффициент корреляции признаётся значимым, а связь между признаками – статистически существенной.

Если расчётное значение t – критерия меньше критического, то нулевая гипотеза принимается, что означает, что коэффициент корреляции в генеральной совокупности в действительности равен нулю и соответственно выборочный коэффициент корреляции существенно не отличается от нуля (или вероятность того, что нулевое значение коэффициента входит в возможный интервал его оценок значительно больше α (5%), соответственно нулевая гипотеза не может быть отклонена).

Если проверяемый фактор потенциально мог влиять на результативный признак, но $t_{расч} > t_{крит.}$, то вывод следует формулировать не в терминах отсутствия связи, а в том, что по изучаемой информации связь надежно не установлена.

Для оценки значимости эмпирического корреляционного отношения η используется F – критерий Фишера–Снедекора, вычисленный по формуле:

$$F = \frac{\eta^2}{1-\eta^2} \cdot \frac{n-m}{m-1},$$

где n - число наблюдений; m – число интервалов группировки.

При этом проверяется гипотеза $H_0 : \eta = 0$ об отсутствии корреляционной зависимости между изучаемыми признаками. Проверяемая гипотеза отклоняется на уровне значимости α, если расчётное значение F – критерия превышает его критическое значение для принятого уровня значимости α и чисел степеней свободы $k_1=m-1$ и $k_2=n-m$. В этом случае величина корреляционного отношения признаётся значимой, а связь между признаками существенной.

При проверке гипотезы используются специальные таблицы F – распределения. В них указывается предельные (критические) значения F – критерия для различных степеней свободы k_1 и k_2, которые могут быть превзойдены с вероятностью α = 0,05.

Для проверки гипотезы $H_0 : r = r_0$ о том, что коэффициент корреляции r нормальных случайных величин X и Y равен заданному числу $r_0 \neq 0$. Используется статистика $U = \frac{Z - \bar{z}}{\sigma_z}$, где

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r}; \bar{z} = \frac{1}{2}\ln\frac{1+r_0}{1-r_0} + \frac{r_0}{2(n-1)}; \sigma_z = \frac{1}{\sqrt{n-3}}.$$

Если проверяемая гипотеза H_0 верна, то при $n \geq 20$ статистика U имеет распределение, близкое к стандартному нормальному распределению.

а) При альтернативе $H_1 : r > r_0$, критическая область размера α имеет вид $\Omega_{кр.}(\alpha) = \lfloor u_{(\alpha)}, \infty)$, где $u_{(\alpha)}$ - критическое значение порядка α стандартного нормального распределения.

б) при альтернативе $H_1 : r < r_0$ критическая область порядка α. $\Omega_{кр.}(\alpha) = (-\infty; -u_{(\alpha)})$

в) при альтернативе $H_1: r \neq r_0$ $\Omega_{кр.}(\alpha) = |U| \geq u_{(\alpha/2)}$

Рассматриваемый критерий основан на z –преобразовании Фишера $z = \frac{1}{2} \ln \frac{1+r}{1-r}$.

Гипотеза $H_0: r_1 = r_2$ о равенстве коэффициентов корреляции двух двумерных нормальных совокупностей проверяется по данным двух независимых двумерных выборок объёмов n_1 и n_2. При этом используется статистика:

$$U = \frac{z_1 - z_2}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}},$$

где $z_i = 0{,}5 \ln((1 + R_i)/(1 - R_i))$; R_i – выборочный коэффициент корреляции, найденный по i-й выборке (i =1,2).

а) при альтернативе $H_1: r_1 > r_2$ критическая область $\Omega_{кр.}(\alpha) = \lfloor u_{(\alpha)}, \infty)$, где $u_{(\alpha)}$ - критическое значение порядка α стандартного нормального распределения.

б) при альтернативе $H_1: r_1 < r_2$ критическая область порядка α. $\Omega_{кр.}(\alpha) = (-\infty; -u_{(\alpha)})$;

в) при альтернативе $H_1: r_1 \neq r_2$ $\Omega_{кр.}(\alpha) = |U| \geq u_{(\alpha/2)}$.

5. ОСНОВЫ РЕГРЕССИОННОГО АНАЛИЗА

5.1 Этапы и показатели регрессионного анализа

Регрессионным анализом называется совокупность методов исследования формы корреляционной зависимости между случайными величинами по статистическим данным, полученным в ходе наблюдения.

Форма корреляционной зависимости аналитически выражается в виде статистико-математических моделей связи – *уравнений регрессии*, описывающих зависимость результативного признака y (зависимой переменной) от определяющих его факторов x_1, x_2, …x_n.

В терминологии регрессионного анализа результативный признак также принято называть *откликом*, факторные признаки – *регрессорами*, *предикторами*.

В зависимости от количества факторов, включенных в исследование, различают *парную* (между двумя признаками x и y) и *множественную регрессию* (описывает зависимость y от нескольких факторов).

Таким образом, регрессионный анализ заключается в построении уравнений регрессии или корреляционно – регрессионных моделей (КРМ).

Между данными понятиями существует различие. Не всякое уравнение регрессии можно считать моделью, так как определение КРМ включает достаточно строгие условия.

Корреляционно- регрессионной моделью системы взаимосвязанных признаков принято считать такое уравнение регрессии, которое включает основные факторы, влияющие на вариацию результативного признака, обладает высоким (не ниже 0,5) коэффициентом детерминации и коэффициентами регрессии, интерпретируемыми в соответствии с теоретическим знанием о природе связей в изучаемой системе.

Кроме того теория и практика статистики выработали ряд требований для построения регрессионной модели, выполнение которых способствует адекватному отражению в ней моделируемых явлений и процессов.

Во-первых, вводятся следующие требования к признакам, включаемым в исследование:

–признаки-факторы должны находиться в причинной связи с результативным признаком (следствием);

–признаки-факторы не должны быть составными частями результативного признака или его функциями;

–признаки-факторы не должны дублировать друг друга, т.е. быть коллинеарными (с коэффициентом корреляции более 0,8);

–не следует включать в модель факторы разных уровней иерархии, т.е. фактор ближайшего порядка и его субфакторы;

–желательно, чтобы для результативного признака и факторов соблюдалось единство единицы совокупности, к которой они отнесены;

–математическая форма уравнения регрессии должна соответствовать логике связи факторов с результатом в реальном объекте;

–принцип простоты: предпочтительнее модель с меньшим (даже несущественно меньшим) числом факторов при том же коэффициенте детерминации.

Во-вторых, требуется обеспечение достаточного объема совокупности для получения несмещенных оценок параметров. Исследователь должен стремиться к увеличению числа наблюдений.

Невыполнение этих условий ставит под сомнение адекватность построенных уравнений регрессий и правомерность их использования в качестве аппроксимирующих статистических моделей.

Регрессионный анализ включает в себя следующие *этапы*:

1) выбор формы связи (уравнения или модели регрессии);

2) отбор факторных признаков (определение размерности модели связи);

3) оценка (вычисление) параметров выбранной модели регрессии;

4) проверка адекватности построенной модели регрессии и ее интерпретация.

1 этап. Выбор формы связи означает выдвижение и принятие некоторой теоретически обоснованной или практически приемлемой рабочей гипотезы о механизме взаимодействия изучаемых признаков.

То есть выбор функции регрессии должен быть основан на теоретическом, логическом анализе, качественном исследовании сущности рассматриваемого явления или профессиональных соображениях. Иначе регрессионный анализ будет лишь формальным математическим упражнением (на основе исходной информации о явлении мы построим его математическую модель, позволяющее оценить величину результативного признака при определенных значениях факторных признаков).

Приблизительное представление о функции регрессии при выборе аналитической формы связи можно получить также на основе:

- графического изображения зависимости в виде эмпирической линии регрессии;
- опыта предыдущих аналогичных исследований, где выбранные формы связи давали удовлетворительные результаты или на базе анализа подобных работ в смежных отраслях знаний.

Наиболее приемлемым способом определения вида исходного уравнения регрессии является метод перебора различных уравнений: строятся уравнения регрессии с различными формами связи, а затем с помощью специальных статистико-математических критериев оценивается их адекватность, и выбирается та форма связи, которая обеспечивает наилучшую аппроксимацию (приближение) и достаточную статистическую достоверность и надёжность.

В качестве наиболее употребляемых критериев подбора функции регрессии используется величина F –критерия Фишера- Снедекора и относительная ошибка аппроксимации. В первом случае предпочтение отдаётся той модели, у которой величина F –критерия Фишера наибольшая, во втором случае – где величина ошибки наименьшая.

Наиболее разработанной в теории статистики является методология парной регрессии, рассматривающая влияние вариации факторного признака x

на вариацию результативного признака *y*. При этом для изучения связи применяются различного вида уравнения (прямо)линейной и (криво)линейной зависимостей.

При анализе линейной связи применяется уравнение прямой линии: $y_x=b_0+bx$.

При анализе нелинейных связей используются следующие функции:

параболическая $y_x= b_0+bx+cx^2$

гиперболическая $y_x=b_0+\frac{b}{x}$

показательная $y_x=b_0\, b^x$

степенная $y_x=b_0\, x^b$

логарифмическая $y_x=a+b\lg x$

логистическая $y_x=\frac{d}{1+e^{a+bx}}$ и др.

Уравнение линейной множественной регрессии имеет вид:

$$\tilde{y}_{123...m} = b_0 + b_1 x_1 + d_2 x_2 + ... + d_m x_m$$

Степенной множественной регрессии:

$$\tilde{y}_{12...m} = b_0 x_1^{b_1} \cdot x_2^{b_2} ... x_m^{b_m}$$

Параболической регрессии:

$$\tilde{y}_{12...m} = b_0 + b_1 x_1^2 + b_2 x_2^2 + ... + b_m x_m^2$$

Показательной регрессии:

$$\tilde{y}_{12...m} = e^{b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m}$$

Гиперболической регрессии:

$$\tilde{y}_{12...m} = b_0 + \frac{b_1}{x_1} + \frac{b_2}{x_2} + ... + \frac{b_m}{x_m}$$

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

$b_0, b_1, b_2, \ldots, b_m$ – выборочные параметры уравнения регрессии.

Основное значение имеют линейные модели в силу простоты и логичности их интерпретации. Нелинейные формы зависимости приводятся к линейным путем линеаризации (например, логарифмирования).

2 этап. Важным этапом построения уже выбранного уравнения множественной регрессии является отбор и последующее включение факторных признаков.

Сложность формирования уравнения множественной регрессии заключается в том, что почти все факторные признаки находятся в зависимости друг от друга.

Определение размерности модели связи, то есть определении оптимального числа факторных признаков, является одной из основных проблем построения множественного уравнения регрессии. Модель большой размерности (100 и более факторных признаков) сложно реализуема и интерпретируема. Сокращение размерности модели за счет исключения второстепенных и статистически несущественных факторов способствует простоте и качеству ее реализации. В то же время чем больше факторных признаков включено в уравнение, тем оно лучше описывает явление. Построение модели малой размерности может привести к тому, что она будет недостаточно адекватна исследуемым явлениям и процессам.

Как уже отмечалось, отбор факторов производится на основе качественного, теоретического анализа с одновременным использованием статистико-математических критериев.

Общепринятым является трех-стадийный отбор факторов. На первой стадии осуществляется априорный анализ и на факторы, включаемые в предварительный их перечень, не накладывается особых ограничений. На второй стадии производится сравнительная оценка и отсев части факторов. Это достигается анализом парных коэффициентов и индексов корреляции, измеряющих тесноту связи каждого из факторов-признаков с результативным признаком и между собой, и оценкой их значимости. Для этого составляется матрица парных коэффициентов корреляции.

Матрица парных коэффициентов корреляции множественной модели регрессии

	y	x_1	x_2	…	x_j	…	x_m
y	1	r_{y1}	r_{y2}	…	r_{yj}	…	r_{ym}
x_1	r_{1y}	1	r_{12}	…	r_{1j}	…	r_{1m}

x_2	r_{2y}	r_{21}	1	…	r_{2j}	…	r_{2m}
…	…	…	…	…	…	…	…
x_i	r_{iy}	r_{i1}	r_{i2}	…	1	…	r_{im}
…	…	…	…	…	…	…	…
x_m	r_{my}	r_{m1}	r_{m2}	…	r_{mj}	…	1

Корреляционная матрица позволяет выявить факторы, которые находятся между собой в тесной линейной корреляционной взаимосвязи. Подобное явление, называемое *мультиколлинеарностью*, искажает величину коэффициентов регрессии (завышает их значения), затрудняет их экономическую интерпретацию.

Анализ таблицы ведется с использованием следующих критериев (индикаторов мультиколлинеарности):

$$r_{yi}>r_{ij} \text{ и } r_{yj}>r_{ij} \text{ ; } r_{ij}<0{,}8$$

r_{ij} – парный коэффициент корреляции между x_i и x_j.

При отсутствии мультиколлинеарности коэффициент корреляции, выражающий связь данного факторного признака с результативным, должен быть больше коэффициента корреляции, выражающего его связь с другими факторными признаками (т.е. связь данного фактора с результативным признаком должна быть теснее, чем с другими факторами).

Парный коэффициент корреляции двух факторных признаков не должен превышать величину 0,8.

Устранение мультиколлениарности происходит путем исключения из регрессионной модели одного или нескольких линейно-связанных факторных признаков. Вопрос о том, какой фактор исключить, решается на основании качественного анализа и с учетом тесноты связи с результативным признаком. Предпочтение отдается факторам, имеющим наиболее сильную по тесноте связь.

На третьей стадии производится окончательный отбор факторов путем анализа значимости (существенности) параметров различных вариантов уравнений множественной регрессии с использованием критерия Стьюдента.

При этом наиболее приемлемым способом завершающего отбора факторов является *шаговый регрессионный анализ*. Он заключается в том, что

после решения модели и оценки всех коэффициентов регрессии из модели исключается тот фактор, коэффициент при котором незначим или имеет наименьший расчетный *t*-критерий Стьюдента. После этого модель решается заново и снова производится оценка значимости оставшихся коэффициентов регрессии. Процесс исключения факторов продолжается до тех пор, пока не будет получено уравнение регрессии, все коэффициенты в котором значимы.

Однако если задачи исследования предусматривают главным образом использование модели для получения теоретических значений результативного признака y_{xi}, то недостаточный уровень значимости коэффициента регрессии не является решающим аргументом для исключения из модели соответствующего фактора, особенно если он важен экономически. Поэтому нередко из модели исключаются лишь факторы, без которых существенно не увеличивается скорректированная остаточная дисперсия.

Существует также так называемый *«прямой метод»* шаговой регрессии – когда факторы не исключаются из модели, а поочередно вводятся в нее с последующей проверкой их значимости.

При проверке значимости введенного фактора определяется, на сколько уменьшается сумма квадратов остатков и увеличивается величина множественного коэффициента корреляции.

Фактор является незначимым, если его включение в модель только изменяет значения коэффициентов регрессии, не уменьшая и не увеличивая суммы квадратов остатков.

Фактор признается существенным, а его включение в уравнение регрессии целесообразным, если величина множественного коэффициента корреляции увеличивается, а коэффициент регрессии не изменяется (или меняется несущественно).

Если же при включении в модель факторного признака коэффициенты регрессии меняют не только величину, но и знак, а множественный коэффициент корреляции не возрастает, то данный факторный признак признается нецелесообразным для включения в модель связи.

В пакетах прикладных программ обычно есть программа пошаговой регрессии. В частности в пакете Statistica предусмотрены две процедуры «включение» новой переменной X_{k+1} в набор независимых переменных Y, $X_1,\ldots, X_k$ и «исключение» переменной X_k из набора независимых переменных Y, $X_1,\ldots, X_k$. Для проверки процедуры служит соответственно статистика F-включение и F- исключение для проверки гипотезы о том, что включение X_{k+1} в набор (исключение X_k из набора) значимо улучшает предсказание Y.

3 этап. Определение числовых значений (оценка) параметров уравнения как парной регрессии, так и множественной регрессии производится методом наименьших квадратов. В основу этого метода положено требование минимальности суммы разности квадрата отклонений эмпирических значений результативного признака от его выровненных (теоретических) значений y_{xi}, полученных по выбранному уравнению регрессии: $\Sigma(y_i - y_{xi})^2 \to \min$.

В случае парной регрессии расчет параметров можно осуществлять также по сгруппированным данным (обычно при большом числе наблюдений), то есть по данным корреляционной или групповой таблицы.

Коэффициенты множественной линейной регрессии отражают абсолютный размер влияния в изменении результативного признака и показывают, на сколько единиц своего измерения изменится результативный признак при изменении соответствующего факторного признака на единицу своего измерения. Если знак при коэффициенте регрессии «+», то с увеличением факторного признака результативный признак возрастает. Если знак при коэффициенте регрессии «–», то с увеличением факторного признака результативный признак уменьшается.

При оценке многофакторных моделей следует особое значение обращать на соответствие знаков коэффициентов регрессии теоретическим и логическим представлениям о направлении влиянии факторов на результативный показатель. В случаях, когда наблюдается несоответствие знаков априорным предложениям, необходимо выяснить причины этого явления.

Причинами такого несоответствия могут быть недостатки исходной информации (малое число наблюдений, неоднородность совокупности) и ошибки при построении модели (при отборе факторов, при выборе формы связи). По этим причинам уравнение регрессии окажется недоброкачественным: знаки параметров будут искажать действительное направление влияния факторов.

Однако ошибочными могут оказаться как знаки параметров, так и априорные логические предположения относительно направления связи. Необоснованность этих предположений бывает вызвана тем, что в них упущены и не учтены взаимодействия факторов и тенденции. Например, за факторами, направление влияния которых не подтвердилось, скрывается влияние других корреляционно связанных с ними факторов, не включенных в исследование. Также характер влияния факторов может меняться в контексте совокупного влияния того набора факторов, который включен в исследование. Поэтому необходима тщательная проверка решения данного уравнения.

Коэффициенты множественной регрессии зависят от единиц измерения факторов и поэтому непосредственно несопоставимы между собой. Чтобы сделать коэффициенты регрессии сопоставимыми их выражают в стандартизированном масштабе: $\beta_{x_i} = b_i \cdot \frac{\sigma_i}{\sigma_y}$;

β_{x_i} - стандартизированный коэффициент регрессии при соответствующем факторном признаке;

b_i - коэффициент регрессии при соответствующем факторном признаке;

σ_i - среднее квадратическое отклонение соответствующего факторного признака;

σ_y - среднее квадратическое отклонение результативного признака.

Величина *стандартизированных бета-коэффициентов* показывает, на сколько средних квадратических отклонений σ_y изменяется y при изменении соответствующего фактора x_i на одно среднее квадратическое отклонение σ_i при неизменности остальных факторов, входящих в уравнение регрессии. Для парной линейной регрессии $\beta = r$.

С целью расширения возможностей анализа используются *частные коэффициенты эластичности*: $Э_{x_i} = b_i \cdot \frac{\overline{x_i}}{\overline{y}}$;

$\overline{x_i}$ -среднее значение соответствующего факторного признака;

$\overline{y}$ - средне значение результативного признака;

b_i - коэффициент регрессии при соответствующем факторном признаке.

Коэффициент эластичности показывает, на сколько процентов в среднем измениться значение результативного признака при измерении факторного признака на 1%.

Другим показателем анализа является *коэффициент раздельной детерминации*: и $d^2_{x_i} = r_{x_i} \cdot \beta_{x_i}$; $\sum d_{x_i} = R^2$;

r_{x_i} – парный коэффициент корреляции между результативным и *i*-м факторным признаком;

β_{x_i} - стандартизированный коэффициент уравнения регрессии при соответствующем факторном признаке;

R^2 – некорректируемый коэффициент множественной детерминации.

Коэффициент раздельной детерминации показывает, на сколько процентов вариация результативного признака объясняется вариацией *i*-ого факторного признака, входящего в множественное уравнение регрессии.

Долю влияния каждого фактора в суммарном влиянии факторов, включенных в уравнение регрессии, позволяют оценить *дельта – коэффициенты*: $\Delta_{x_i} = \frac{d^2_{x_i}}{R^2}$

4 этап. Проверка адекватности построенных моделей регрессии и их интерпретация.

Выборочные параметры регрессии и выборочный коэффициент детерминации, вычисленные по ограниченному числу единиц наблюдения, всегда содержат элемент случайности, в связи, с чем возникает необходимость проверки значимости этих выборочных характеристик.

Адекватность уравнения регрессии означает, что заложенные в модели связи соответствуют реально существующим связям. Она проверяется в два этапа.

1) Проверка адекватности всей модели с помощью F-критерия и величины относительной ошибки аппроксимации.

На данном этапе проверяется значимость коэффициента детерминации: выдвигается гипотеза $H_0 : R^2 = 0$ о том, что коэффициент детерминации генеральной совокупности, из которой извлечена исследуемая выборка, равен нулю. Эта гипотеза равносильна гипотезе о том, что ни один из факторов, включённых в регрессию, не оказывает существенного влияния на результативный признак ($H_0 : b_1 = b_2 = ... = b_m = 0$). Поэтому проверка значимости коэффициента детерминации и является проверкой адекватности (соответствия) выбранной модели регрессии реальным данным наблюдения.

В качестве альтернативы рассматривается гипотеза H_1 : хотя бы один коэффициент регрессии $b_i \neq 0$.

Оценка значимости коэффициента детерминации осуществляется с помощью *F*-критерия. Если $F_{расч..} > F_{крит.}$, то гипотеза о равенстве коэффициента детерминации нулю и несоответствии заложенных в модели связей реально существующим отклоняется на уровне значимости α, то есть коэффициент детерминации признаётся статистически значимым, а модель регрессии – адекватной. Отклонение проверяемой гипотеза H_0 в пользу альтернативы H_1 означает, что, по крайней мере, один из *m* факторов, включенных в регрессию, оказывает существенное (значимое) влияние на результативный признак. И наоборот.

2) Если гипотеза $H_0 : b_1 = b_2 = ... = b_m = 0$ отклоняется в пользу альтернативы H_1 : хотя бы один коэффициент регрессии $b_i \neq 0$, то переходят ко второму этапу – выясняют, какой из коэффициентов регрессии привёл к отклонению гипотезы H_0.

На этом этапе последовательно одна за другой проверяются гипотезы $H_{0(i)}: b_i = 0$ о том, что фактор x_i не оказывает заметного влияния на результативный признак. В качестве альтернативы рассматривается гипотеза $H_{1(i)}: b_i \neq 0$.

То есть на данном этапе производится оценка значимости параметров уравнения регрессии.

Значимость параметров регрессии проверяется на основе *t* – критерия Стьюдента:

$t_{\text{расч}} = \frac{|b_i|}{\sigma_{b_i}}$, где $\sigma_{b_i} = \sqrt{\frac{\sigma_y^2}{m}}$ - стандартное отклонение (стандартная ошибка) выборочного параметра b_i;

Параметр признаётся статистически значимым, если расчётное значение *t* – критерия Стьюдента превосходит его критическое значение при заданном уровне значимости α =0,05 и числе степеней свободы $v = n - m - 1$.

При анализе адекватности уравнения регрессии исследуемому процессу возможны следующие варианты:

- построенное уравнение на основе её проверки по *F*- критерию Фишера в целом адекватно, и все коэффициенты регрессии значимы. Такое уравнение может быть использовано для принятия решений и осуществления прогнозов;
- уравнение по *F*- критерию Фишера адекватно, но часть коэффициентов регрессии незначима. В этом случае уравнение пригодно для принятия некоторых решений, но не для производства прогнозов;
- уравнение по *F*- критерию Фишера адекватно, но все коэффициенты регрессии незначимы. В этом случае уравнение полностью считается неадекватным. На его основе не принимаются решения и не осуществляются прогнозы;
- уравнение по *F*- критерию Фишера неадекватно, и все коэффициенты регрессии незначимы. В этом случае уравнение полностью считается

неадекватным. На его основе не принимаются решения и не осуществляются прогнозы.

В качестве меры адекватности модели регрессии используется также процентное отношение стандартной ошибки (S_ℓ) к среднему уровню результативного признака $(\bar{y})$ – *относительная ошибка аппроксимации*:

$$\frac{S_\ell}{\bar{y}} \times 100\%; \quad S_\ell = \sqrt{\frac{\sum (y - y_x)^2}{n - m}} \text{ , где}$$

Если $\frac{S_\ell}{\bar{y}} \times 100\% \le 10\%$, то точность модели регрессии высокая, если 10-20% – точность модели регрессии хорошая (то есть уравнение достаточно хорошо описывает взаимосвязь между изучаемыми признаками), если 20-50% – точность модели регрессии удовлетворительная.

Процедурой, завершающей регрессионный анализ, является *интерпретация уравнения***.** Она осуществляется методами той отрасли знаний, к которой относятся исследуемые явления.

При изучении экономических явлений осуществляется перевод параметров уравнения с языка статистики и математики на язык экономики.

5.2 Основные направления применения регрессионного анализа в экономических исследованиях

Корреляционно-регрессионные модели могут быть использованы для решения различных задач:

1) выявление важнейших факторов, влияющих на результативный признак. Содержательный анализ моделей в целях уточнения приоритетности факторов опирается на сравнении значений бета- коэффициентов, частных коэффициентов эластичности и детерминации, дельта-коэффициентов. В этих целях производиться ранжирование факторов по величине данных коэффициентов.

Результаты могут быть оформлены в виде таблицы:

Факторы	Значения коэффициентов		Ранг факторов по величине коэффициентов		Средний ранг
	β_{x_i}	$Э_{x_i}$	β_{x_i}	$Э_{x_i}$	
X1					
X2					
X3					
…					

2) оценка хозяйственной деятельности субъектов экономики и определение резервов с целью эффективностью управления теми или иными экономическими системами.

При обычных, традиционных методах анализа деятельности хозяйствующих субъектов показатели их деятельности сравниваются со среднеотраслевыми и среднерегиональными уровнями. Такие сравнения основаны на допущении, что все предприятия отрасли, региона работают примерно в одинаковых условиях и располагают боле или менее одинаковыми возможностями. Однако на деле это далеко не всегда так. Уровень анализируемого результативного показателя на отдельных предприятиях и его средний уровень отличаются один от другого за счет очень многих факторов.

Поэтому более обоснованно сравнивать фактический уровень результативного показателя y на данном предприятии не со среднеотраслевым и среднерегиональным, а с расчетным уровнем $\tilde{y}$, который вычислен по уравнению регрессии путем подстановки в него значений факторов на данном предприятии. В этом случае сопоставляемые уровни y и $\tilde{y}$ будут отличаться друг от друга за счет факторов, не входящих в модель. Следующим этапом анализа будет сравнение расчетного уровня $\tilde{y}$ и среднего уровня $\bar{y}$, различие которых обусловлено только факторами, включенными в модель.

Расчетные уровни для отдельных предприятий выражают такие уровни результативного признака, которые были бы достигнуты при фактических значениях факторов, входящих в модель, и при средней по всей совокупности эффективности их использования. Положительная величина абсолютного отклонения y - $\tilde{y}$ свидетельствует о том, что эффективность использования этих факторов выше средней эффективности их использования в данной

совокупности. Такие предприятия могут быть исследованы с целью выявления передового опыта (или каких-либо благоприятных обстоятельств, способствующей высокой эффективности). Отрицательная величина отклонения $y - \tilde{y}$ указывает на то, что эти предприятия использовали факторы с более низкой эффективностью, чем в среднем по совокупности. Исследование таких предприятий позволит выявить причины этого и определить неиспользованные резервы.

3) прогнозирование возможных значений результативных экономических показателей при заданных значениях факторных признаков. Эта задача решается путем подстановки в уравнение регрессии ожидаемых в ближайшем будущем, планируемых или оптимальных значений факторных признаков. В первом случае предполагается, что форма взаимосвязи (значения коэффициентов регрессии) будет сохраняться неизменной на весь период прогнозирования. Ожидаемые значения факторных признаков определяются либо экспертным путем, на основании заключений специалистов в данной области или на основе экстраполяции трендов.

Для вычисления доверительного интервала прогноза точечный прогноз нужно скорректировать на величину стандартной ошибки уравнения регрессии, умноженную на t –критерий Стьюдента:

$$\tilde{y}' \pm t \cdot \sqrt{\frac{(y-\tilde{y})^2}{n-m}}.$$

Для повышения достоверности прогноза целесообразно получить прогнозное значение с помощью другого метода и сравнить полученные оценки, которые в этом случае взаимно «проверяют» друг друга.

6. НЕПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА

6.1 Понятие непараметрического тестирования

Значительная часть критериев, применяемая в статистических процедурах, предполагает, что исследуемая генеральная совокупность, из которой взята изучаемая выборка, имеет закон распределения определенного типа (чаще всего нормальный). Между тем в подавляющем большинстве практических задач тип распределения исследуемых случайных величин изначально не известен. В связи с этим актуальным становится применение критериев, не требующих предварительных предположений относительно типа распределения случайных величин, по данным наблюдения над которыми проверяются гипотезы. Такие критерии называются *свободными (независимыми) от распределения* или *непараметрическими*.

То есть в параметрических постановках на данные накладывается требование – функции распределения исследуемых величин должны принадлежать определенному параметрическому семейству; в непараметрических постановках – такие требования отсутствуют, требуется лишь, чтобы функции распределения были непрерывны.

Параметрические критерии - это группа статистических критериев, которые включают в свой расчет параметры вероятностного распределения (среднюю величину, дисперсии, стандартное отклонение). Они позволяют прямо оценить уровень основных параметров генеральных совокупностей, разности средних, различия в дисперсиях. Они используются в задачах проверки параметрических гипотез. К ним относятся: t-критерий Стьюдента, *F-к*ритерий Фишера, критерий отношения правдоподобия, критерий Романовского.

Непараметрические критерии – это группа статистических критериев, которые не включают в свой расчёт параметры вероятностного распределения и основаны на оперировании другими данными: знаками, частотами, рангами.

По сравнению с параметрическими тестами непараметрическое тестирование имеет ряд преимуществ, что способствует росту актуальности его использования:

1) непараметрические тесты не включают никаких предположений относительно характера распределения генеральной совокупности, из которой взята выборка; они менее чувствительны к «выбросам» - наличию в выборке единиц наблюдения, имеющих принципиальное иное распределение;

2) непараметрические методы наиболее приемлемы, когда объем выборок мал. Если выборка большая ($n > 100$), то не имеет смысла использовать непараметрические статистики. В этом случае считается, что выборочные средние подчиняются нормальному закону, даже если исходная переменная не является нормальной или измерена с погрешностью. Если выборка мала, параметрические критерии следует использовать только при наличии уверенности, что переменная действительно имеет нормальное распределение. Однако нет способа проверить это предположение на малой выборке.

3) непараметрические методы могут использоваться данные, представленные в разных шкалах измерения. Большинство параметрических тестов требуют, чтобы данные были представлены в интервальной шкале или шкале отношений, в то время как многие непараметрические тесты не содержат таких требований: они могут работать и в случае, когда данные измерены в номинальной и порядковой шкале.

Для анализа малых выборок и для данных, измеренных в бедных шкалах, применяют непараметрические методы.

4) простота вычислений (что связано с малым числом выборки, отсутствием предварительных расчетов по выявлению типа распределения, более простым формулам)

Недостатки:

1) в непараметрических критериях по сравнению с параметрическими тестами информация, имеющаяся в данных, используется менее эффективно, их мощность ниже, чем параметрических – они с меньшей вероятностью

отвергают нулевую гипотезу, если последняя неверна. По этой причине параметрические тесты предпочтительнее, когда требуемые предположения относительно генеральной совокупности могут быть сделаны.

2) непараметрические тесты не позволяют осуществить прямую оценку уровня таких важных параметров, как среднее или дисперсия.

3) непараметрическое тестирование больше зависит от статистических таблиц, если не используется специальный пакет прикладных программ.

Среди непараметрических критериев можно выделить *два подвида*:

1) классический пример критериев, свободных от распределения – критерии согласия и однородности, основанные на эмпирических функциях распределения;

2) критерии однородности, случайности, симметрии и независимости, свободные от распределения, основанные на порядковых статистиках и рангах.

1 группа. а) *Критерии согласия* используются для проверки соответствия эмпирического (фактического) и теоретического (гипотетического) распределений. Задача в этом случае формулируется следующим образом: имеются данные наблюдения $x_1, x_2 ..... x_j, ... x_n$ над случайной величиной X, функция распределения которой неизвестна. Выдвигается гипотеза о том, что истинной функцией распределения исследуемой случайной величины *X* является некоторая заданная функция F(x).

Если гипотеза верна, то найденная по данным наблюдения эмпирическая функция распределения $F^*(x)$ не должна сильно отличаться от гипотетической функции распределения, и с увеличением объема *n* выборки различие между ними должно уменьшаться. В связи с этим вопрос о принятии или отклонении проверяемой гипотезы решается в зависимости от того насколько хорошо согласуются эмпирическая $F^*(x)$ и гипотетическая F(x) функции распределения. Статистические критерии, базирующиеся на таком подходе, называются критериями согласия. В основе этих критериев лежит выбранная соответствующим образом статистика, которая может служить *мерой*

расхождения между эмпирическим и гипотетическим законами распределения исследуемой случайной величины.

б) *критерии однородности*. Известно, что чем больше данных наблюдения используется в ходе статистического исследования какого-либо объекта, тем точнее полученные при этом эмпирические оценки параметров данного объекта. Для того, чтобы увеличить количество данных наблюдения и обеспечить требуемую точность и надежность выборочных оценок, на практике широко используют следующий прием: объединяют несколько выборок, полученных в процессе наблюдения за однотипными объектами в одну общую выборку (например, поставщиков разных форм собственности, предприятия разных организационно-правовых форм, работников разных видов экономической деятельности и т.п.). Такое объединение выборок допустимо только в тех случаях, когда они однородны. Поэтому перед объединением выборок надо обязательно проверить их однородность. Такая проверка особенно необходима при объединении данных наблюдения, полученных из разных источников.

Гипотеза об однородности двух выборок $x_1, x_2 \ldots\ldots x_j, \ldots\ldots x_n$ и $y_1, y_2 \ldots\ldots x_{yj}, \ldots\ldots y_n$, полученных при наблюдении над независимыми случайными величинами X и $У$, есть ни что иное как предположение о том, что случайные величины X и $У$ подчиняются одному и тому же закону распределения. Используется и другая формулировка гипотезы об однородности двух выборок: выборки $x_1, x_2 \ldots\ldots x_j, \ldots\ldots x_n$ и $y_1, y_2 \ldots\ldots x_{yj}, \ldots\ldots y_n$ извлечены из одной и той же генеральной совокупности.

Критерии однородности: Н.В. Смирнова, Андерсона, хи-квадрат Пирсона.

2 группа. а**)** *критерии однородности, случайности, симметрии*: критерий знаков, одновыборочный критерий серий, двухвыборочный критерий серий Вальда-Вольфовица, медианные критерии, критерии знаков Вилкоксона (для

одной выборки, для сравнения двух выборок, знаковых рангов), критерий Манна – Уитни, Краскала- Уоллиса, Фридмена и др.

б) *ранговые коэффициенты независимости*: коэффициенты корреляции рангов Спирмена, Кендалла , коэффициент гамма; коэффициент конкордации.

Тестированию подлежат все процедуры по статистической обработке случайных выборок. Причем почти для каждого параметрического критерия имеется, по крайней мере, один непараметрический аналог.

Статистические процедуры	Параметрические критерии	Непараметрические критерии
Определение меры соответствия выборки (эмпирического, фактического распределения) **какому-либо теоретическому** (гипотетическому) **распределению**	–	Критерии согласия: критерий хи- квадрат Пирсона; критерий согласия Колмогорова; критерии согласия омега –квадрат: критерий Крамера-Мизеса-Смирнова, критерий Андерсона – Дарлинга и их модификации
Выявление различий **–между двумя независимыми выборками** – **между двумя зависимыми выборками** *(проверяется эффект обработки, «до» и «после»)* Если данные категориальны или являются категоризованными переменными (*представлены в виде частот попавших в определенные категории*)	*t*-критерий Стьюдента для независимых выборок *t*-критерий Стьюдента для зависимых выборок	критерий серий Вальда-Вольфовица, *U* критерий Манна-Уитни; двухвыборочный критерий Колмогорова-Смирнова критерий знаков; критерий Вилкоксона критерий *хи-квадрат* Макнемара
Выявление различий между несколькими выборками: **–независимыми** **–зависимыми**	*F* –критерий Фишера (классический одно-, двух-и многофакторный дисперсионный анализ)	критерий Краскела-Уоллиса (*ранговый однофакторный дисперсионный анализ*) критерий Фридмана (*ранговый двухфакторный дисперсионный анализ*) *Q* критерий *Кохрена*
Оценка зависимости между выборками (переменными) → корреляционный анализ	коэффициент корреляции Пирсона; корреляционные отношения	Ранговые коэффициенты корреляции Спирмена, Кендалла, коэффициент гамма; коэффициент конкордации; коэффициенты взаимной сопряженности
Установление формы зависимости между выборками (зависимой и независимыми переменными)→ регрессионный анализ	коэффициенты регрессии и эластичности	–

6.2 Непараметрические методы корреляции

Методы корреляционного анализа различаются в зависимости от шкалы измерения переменных, зависимость между которыми изучается.

Для переменных, измеренных в интервальной шкале и шкале отношений, применяются так называемые *собственно-корреляционные методы,* основанные на специальных предпосылках, основной из которых является нормальность распределения изучаемых признаков.

При использовании этих методов нельзя обойтись без вычисления основных параметров распределения (средней величины, дисперсии), поэтому они получили название параметрических методов.

Между тем в статистической практике приходится сталкиваться с задачами измерения связи между переменными, к которым параметрические методы анализа в их обычном виде неприменимы. В связи с этим статистической наукой разработаны методы, с помощью которых можно измерить связь, не используя количественные значения признака, а значит и параметры распределения. Такие методы получили название *непараметрических методов корреляции*. Они применяются:

1) в случаях, когда изучается связь между количественными переменными, измеренными в интервальной шкале и шкале отношений, форма распределения которых существенно отличается от нормальной (то есть форма распределения которых подчиняется различным законам распределения);

2) когда изучается связь между неколичественными переменными (качественными признаками), измеренными в номинальной или порядковой шкале.

При этом действует правило: меры связей, разработанные для переменных более низкого уровня измерений (для более бедных шкал) могут использоваться для измерения связей между переменными более высокого уровня измерения. Но при этом происходит потеря части информации.

При изучении взаимосвязи *между порядковыми переменными* единицам наблюдения присваиваются ранги – порядковые номера этих единиц

наблюдения в ранжированном ряду. Мерой связи являются ранговые коэффициенты корреляции.

Коэффициент корреляции рангов Спирмена определяется по формуле:

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}, \text{ где}$$

d – разность между рангами соответствующих величин двух признаков;

n – число единиц в ряду (число пар рангов).

Коэффициент корреляции рангов принимает любые значения от -1 до +1. Если все ранги строго изменяются в одном и том же порядке, то d=0, а ρ=1. Если же ранги изменяются строго в противоположных направлениях, то ρ= -1. Значение ρ=0 характеризует отсутствие связи. Применяются те же шкалы оценивания.

Значимость коэффициентов корреляции рангов для совокупностей небольшого объёма (n≤30) проверяется по таблице предельных значений коэффициента корреляции рангов Спирмена при заданном уровне значимости α и определённом объёме совокупности.

Значимость ρ может быть проверена также на основе t – критерия Стьюдента. Расчётное значение критерия определяется по формуле: $t_{расч} = \rho \cdot \sqrt{\frac{n-2}{1-\rho^2}}$

Значение коэффициента корреляции считается статистически существенным, если $t_{расч} > t_{крит.}$ при заданном уровне значимости α и числе степеней свободы k=n-2.

При наличии связанных рангов коэффициент корреляции рангов Спирмена вычисляется по формуле:

$$\rho = 1 - \frac{6\Sigma d^2 - A - B}{\sqrt{\left(n^3 - n - 12A\right)\cdot\left(n^3 - n - 12B\right)}},$$

где d^2 – квадрат разности рангов, n – число наблюдений,

$A = \frac{1}{12}\sum_j \left(A_j^{\,3} - A_j\right)$; $B = \frac{1}{12}\sum_k \left(B_k^3 - B_k\right)$;

j – номера связок по порядку для первого признака;

A_j – число одинаковых рангов в j -й связке по первому признаку;

k – номера связок по порядку для второго признака;

B_k – число одинаковых рангов в k -й связке по второму признаку.

Ранговый коэффициент корреляции Кендалла рассчитывается по формуле:

$$(\text{тау})\ \tau = \frac{2S}{n(n-1)}, \quad S=P+Q$$

n – число наблюдений;

S – сумма разностей между числом последовательностей и числом инверсий по результативному признаку.

Расчёт данного коэффициента выполняется в следующей последовательности:

1) ранги факторного признака располагаются в порядке возрастания;
2) ранги результативного признака располагаются в порядке, соответствующем рангам признака x;
3) для каждого ранга результативного признака определяется сколько чисел, находящихся справа от него (следующих за ним) имеют величину ранга, превышающую его величину. Суммируя полученные таким образом числа, получаем слагаемое P, которое можно рассматривать как меру соответствия последовательностей рангов по x и y, и которое учитывается со знаком «+»;
4) для каждого ранга y определяется число, следующих за ним рангов, меньших его величины. Суммарная величина обозначается через Q и фиксируется со знаком «-»;
5) определяется сумма баллов S=P+Q

Коэффициент Кендалла также изменяется в пределах от -1 до +1.

При достаточно большом числе наблюдений между коэффициентами корреляции рангов Спирмена и Кендалла существует следующее соотношение: $\rho \approx \frac{3}{2}\tau$.

Значимость коэффициента корреляции рангов Кендалла проверяется при выбранном уровне значимости α при больших n по формуле:

$$\tau > t_a \cdot \sqrt{\frac{2(2n+5)}{9n(n-1)}}, \text{ где}$$

t_a – коэффициент, определяемый по таблице нормального распределения.

Для определения тесноты связи между несколькими порядковыми переменными применяется *множественный коэффициент ранговой корреляции* (коэффициент конкордации):

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

где m– количество признаков (переменных); n– число наблюдений; S– отклонение суммы квадратов рангов от средней квадратов рангов:

$$S = \sum_{1}^{n}(\sum_{1}^{m} R_{ij})^2 - \frac{(\sum_{1}^{n}\sum_{1}^{m} R_{ij})^2}{n}.$$

В социологических, маркетинговых исследованиях данный показатель также называется также *коэффициентом согласованности***.** В этом случае $S = \sum_{i-1}^{n}(\sum_{j=1}^{m} R_{ij} - \frac{m(n+1)}{2})^2$

.

Значимость множественного коэффициента ранговой корреляции проверяется на основе χ^2 - критерия Пирсона: $\chi^2 = \frac{12S}{m \cdot n(n-1)}$.

Если расчётное значение χ^2 больше критического значения $\chi^2_{кр.}$ (α =0,05; $d.f. = n-1$), то множественный коэффициент ранговой корреляции признается значимым.

В случае наличия связанных рангов формула множественного коэффициента ранговой корреляции имеет вид:

$$W = \frac{S}{\frac{1}{12}\left[m^2\left(n^3 - n\right) - m\sum_{j=1}^{m} T_j\right]},$$

где $m-$ количество признаков (переменных); $n-$ число наблюдений; $S-$ отклонение суммы квадратов рангов от средней квадратов рангов:

$$S = \sum_{1}^{n}(\sum_{1}^{m} R_{ij})^{2} - \frac{(\sum_{1}^{n}\sum_{1}^{m} R_{ij})^{2}}{n};$$

T_j – характеристика связанности рангов по *j*-й переменной,

$$T_j = \frac{1}{12}\sum_{j=1}^{m}\left(t_j^3 - t_j\right),$$

t_j – количество связанных рангов по по *j*-й переменной.

$$\chi_{расч.}^{2} = \frac{12S}{m \cdot n(n-1) - \frac{1}{n-1}\sum_{j=1}^{m} T_j}$$

При измерении связи *между номинальными переменными* значения переменных не участвуют в расчетах: меры связей основаны на частоте совместного появления определенной категории одной переменной и определенной категории другой переменной.

Для исследования степени тесноты связи *между дихотомическими переменными* (признаками, каждый из которых принимает два значения), используют коэффициенты *ассоциации* (Д.Юла) и *контингенции* (К.Пирсона). Для их вычисления строится таблица сопряженности или «таблица четырех полей», частоты которой обозначаются a,b,c,d.

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

Коэффициент ассоциации определяется по формуле:

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

В случаях, когда хотя бы один из четырех показателей в таблице «четырех полей» отсутствует, величина K_A=1, что дает преувеличенную оценку степени тесноты связи между признаками, и предпочтение следует отдать *коэффициенту контингенции*, который определяется по формуле:

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

Измерение связи *между категориальными переменными* (измеренными в номинальной шкале) производится с помощью специальных мер связи. Исходные данные представляются в виде таблиц размерности *m* (число категорий, выделяемых по одной категории) на *p* (число категорий другой категории). Для данных таблиц используют, прежде всего, коэффициенты взаимной сопряженности: Пирсона, Чупрова, Крамера. Все эти меры основаны на критерии хи-квадрат.

При изучении номинальных переменных применяются также *теоретико-информационные меры связи*, основанные на величине количества информации. С этой целью оценивается неопределенность распределения переменной *y* (без учета знания переменной *x*), то есть вычисляется полная энтропия распределения переменной *y*.

Разработано целое семейство теоретико-информационных коэффициентов связи. Наиболее популярен коэффициент нормированной информации. Существуют также другие меры связи между номинальными переменными : лямда- Гутмана и тау-Гудмена и Краскала.

Первые непараметрические методы, основанные на коэффициентах ранговой корреляции, появились в первой трети XX века в работах Спирмена и Кендалла. Заметной частью статистики непараметрика стала лишь со второй трети XX века. В 30-е годы появились работы А.Н.Колмогорова и Н.В.Смирнова, предложивших и изучивших статистические критерии. После второй мировой войны развитие непараметрической статистики пошло быстрыми темпами. Большую роль сыграли работы Вилкоксона и его школы. К настоящему времени с помощью непараметрических методов можно решать практически тот же круг статистических задач, что и с помощью параметрических.

В нашей стране непараметрические методы получили достаточно большую известность после выхода в 1965 г. первого издания сборника статистических таблиц Л.Н.Большева и Н.В.Смирнова, содержащего подробные таблицы для основных непараметрических критериев. Тем не менее параметрические методы все еще популярнее непараметрических.

7. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

7.1 Элементы, показатели и компоненты временного ряда

В отечественных учебниках по общей теории статистики преобладают термины *ряды динамики, динамические ряды*. Однако неточность этих терминов состоит в том, что не каждый ряд уровней за последовательные моменты или периоды времени содержит (отражает) на самом деле динамику какого-либо показателя.

Термин *динамика* правильнее относить к изменениям, направленному развитию, наличию тенденции рассматриваемых во времени показателей. То есть *ряды динамики* – это ряды уровней статистических показателей, в которых содержится тенденция изменения. Однако существуют ряды уровней, содержащие лишь колебания, но не имеющие надежно установленной тенденции, характеризующие явления или процессы, находящиеся в статическом состоянии.

В связи с этим более правильным будет использовать термин, обозначающий более общее понятие, включающее как динамические, так и статические последовательности уровней какого-либо показателя. В зарубежной статистической литературе принят термин в*ременные ряды*.

Временные ряды – это случайная последовательность упорядоченных во времени числовых показателей, характеризующих уровень состояния и изменения изучаемого явления.

Временной ряд состоит из двух элементов:

1) уровней ряда y_i – числовых значений статистических показателей, характеризующих величину изучаемого явления;

2) времени – периодов (или моментов) времени, к которым относятся данные уровни.

Временной ряд имеет два существенных отличия от случайной выборки:

1) элементы $x_1, x_2 \ldots x_n$ случайной выборки взаимно независимы, тогда как значение y_i временного ряда, зафиксированное в момент времени t_i, может

существенно зависеть от одного или нескольких значений ряда, зафиксированных до этого момента;

2) элементы $x_1, x_2 \ldots x_n$ случайной выборки имеют один и тот же закон распределения, в то время как закон распределения i –ого уровня временного ряда может изменяться при изменении его номера i.

Различают несколько видов рядов динамики:

При практическом изучении временных рядов исследователь на основании наблюденного отрезка временного ряда (конечной длины) должен сделать выводы о свойствах этого ряда и о вероятностном механизме, порождающем этот ряд. Чаще всего при изучении временных рядов ставятся следующие цели:

- анализ скорости и интенсивности развития явления во времени;
- описание характерных особенностей (закономерностей) ряда;
- подбор статистической модели, описывающей временной ряд;
- предсказание будущих значений на основе прошлых наблюдений;
- управление процессом, порождающим временной ряд.

На практике эти и подобные цели достижимы далеко не всегда и далеко не в полной мере. Часто этому препятствует недостаточный объем наблюдений (недостаточная длительность); еще чаще — изменяющаяся с течением времени статистическая структура временного ряда. Из-за этих изменений значение прошлых наблюдений обесценивается, и они уже не помогают предвидеть будущее.

Анализ скорости и интенсивности развития явления во времени осуществляется с помощью аналитических показателей, которые получаются в результате сравнения уровней временного ряда между собой (абсолютный прирост, коэффициенты роста и прироста, темпы роста и прироста, значение 1% прироста).

Эти показатели могут быть:

1) *базисными*, когда каждый уровень ряда сравнивается с одним и тем же уровнем, принятым за базу сравнения;

В качестве базисного уровня выбирается либо начальный (первый) уровень динамического ряда или уровень, с которого начинается какой-то новый этап развития явления.

2) *цепными*, когда каждый последующий уровень ряда сравнивается с предшествующим уровнем.

Абсолютный прирост характеризует размер увеличения (уменьшения) уровня ряда за определенный промежуток времени, т.е. абсолютную скорость роста (снижения).

$$A^{б} = y_i - y_0$$

$$A^{ц} = y_i - y_{i-1}$$

где y_0 - уровень базисного периода

Коэффициент роста показывает, во сколько раз данный уровень больше базисного или предшествующего уровня (в случае $К_p \succ 1$) или какую часть базисного или предшествующего уровня составляет уровень текущего периода (в случае $К_p \prec 1$).

$$K_p^{б} = \frac{y_i}{y_0}$$

$$K_p^{ц} = \frac{y_i}{y_{i-1}}$$

Темп роста показывает, сколько процентов в уровне базисного или предшествующего уровня составляет уровень текущего периода.

$$T_p^{б} = \frac{y_i}{y_0} \times 100\% = K_p^{б} \times 100\%$$

$$T_p^{ц} = \frac{y_i}{y_{i-1}} \times 100\% = K_p^{ц} \times 100\%$$

Коэффициент прироста определяется по формуле:

$$K_{пр}^{б} = K_p^{б} - 1 \quad \text{или} \quad \frac{y_i - y_0}{y_0} = \frac{A^{б}}{y_0}$$

$$K_{пр}^{ц} = K_p^{ц} - 1 \quad \text{или} \quad \frac{y_i - y_{i-1}}{y_{i-1}} = \frac{A^{ц}}{y_{i-1}}$$

Темп прироста показывает, на сколько процентов увеличился (уменьшился) уровень ряда по сравнению с базисным или предшествующим уровнем.

$$T_{пр}^{б} = T_p^{б} - 100\%; \quad T_{пр}^{ц} = T_p^{ц} - 100\%$$

Абсолютное значение 1% прироста:

$$З = \frac{A^{ц}}{T_{пр}^{ц}} \quad или \quad З = 0{,}01\, y_{i-1}$$

Между базисными и цепными показателями существует взаимосвязь:

- сумма цепных абсолютных приростов равна базисному абсолютному приросту за изучаемый период: $\sum A^{ц} = A_n^{б}$

- произведение цепных коэффициентов роста равно базисному коэффициенту роста за изучаемый период: $Кр_1^{ц} \times Кр_2^{ц} \times \ldots \times Кр_n^{ц} = Кр_n^{б}$

Для обобщающей характеристики динамики исследуемого явления за период времени определяют средние показатели.

Средний абсолютный прирост показывает на сколько единиц увеличился или уменьшился уровень по сравнению с предыдущим в среднем за единицу времени и определяется по формуле:

$$\overline{A} = \frac{\sum A^{ц}}{n-1} = \frac{A_n^{б}}{n-1}$$, где n - число уровней ряда.

Средний коэффициент роста показывает во сколько раз в среднем за единицу времени изменился уровень ряда по сравнению с предыдущим за определенный промежуток времени и исчисляется по формуле средней геометрической:

$$\overline{K}_p = \sqrt[n]{K_{p1}^{ц} \cdot K_{p2}^{ц} \cdot \ldots \cdot K_{pn-1}^{ц} \cdot K_{pn}^{ц}} = \sqrt[n-1]{K_{pn}^{б}}$$ –для равноотстоящих рядов;

$$\overline{K}_p = \sqrt[\sum t]{K_1^{t_1} \times K_2^{t_2} \times \ldots \times K_n^{t_n}}$$ – для неравноотстоящих рядов.

Средний темп роста определяется следующим образом:

$$\overline{T}_p = \overline{K}_p \cdot 100\%$$

Средний темп прироста показывает на сколько процентов увеличился или уменьшился уровень по сравнению с предыдущим в среднем за единицу времени и рассчитывается по формуле:

$$\overline{T}_{np} = \overline{T}_p - 100\%$$

Средний размер абсолютного значения 1% прироста вычисляется по формуле: $\overline{З} = \frac{\overline{A}}{\overline{T}_{np.}}$

Характеристика закономерностей изменений во времени – сложная и трудоемкая процедура исследования, так как первоначальные значения временного ряда формируются под совокупным влиянием множества факторов, действующих в разных направлениях. Эти факторы сгруппировать следующим образом:

1) факторы эволюционного характера;

2) факторы осциллятивного характера;

3) факторы нерегулярного характера.

Факторы эволюционного характера вызывают изменения, определяющие некое общее направление развития явления, как бы его многолетнюю эволюцию. Такие долговременные, устойчивые изменения динамического ряда называются *тенденцией развития* или *трендом*.

Влияние факторов осциллятивного характера вызывает циклические и сезонные колебания.

Циклические (периодические) колебания состоят в том, что значение изучаемого показателя в течение какого-то времени возрастает, достигает определённого максимума, затем понижается, достигает определённого минимума, вновь возрастает до прежнего значения и т.д. Схематически циклические колебания можно представить в виде синосоиды.

Сезонные колебания – это колебания, периодически повторяющиеся в некоторое определённое время каждого года, дня месяца или часа дня.

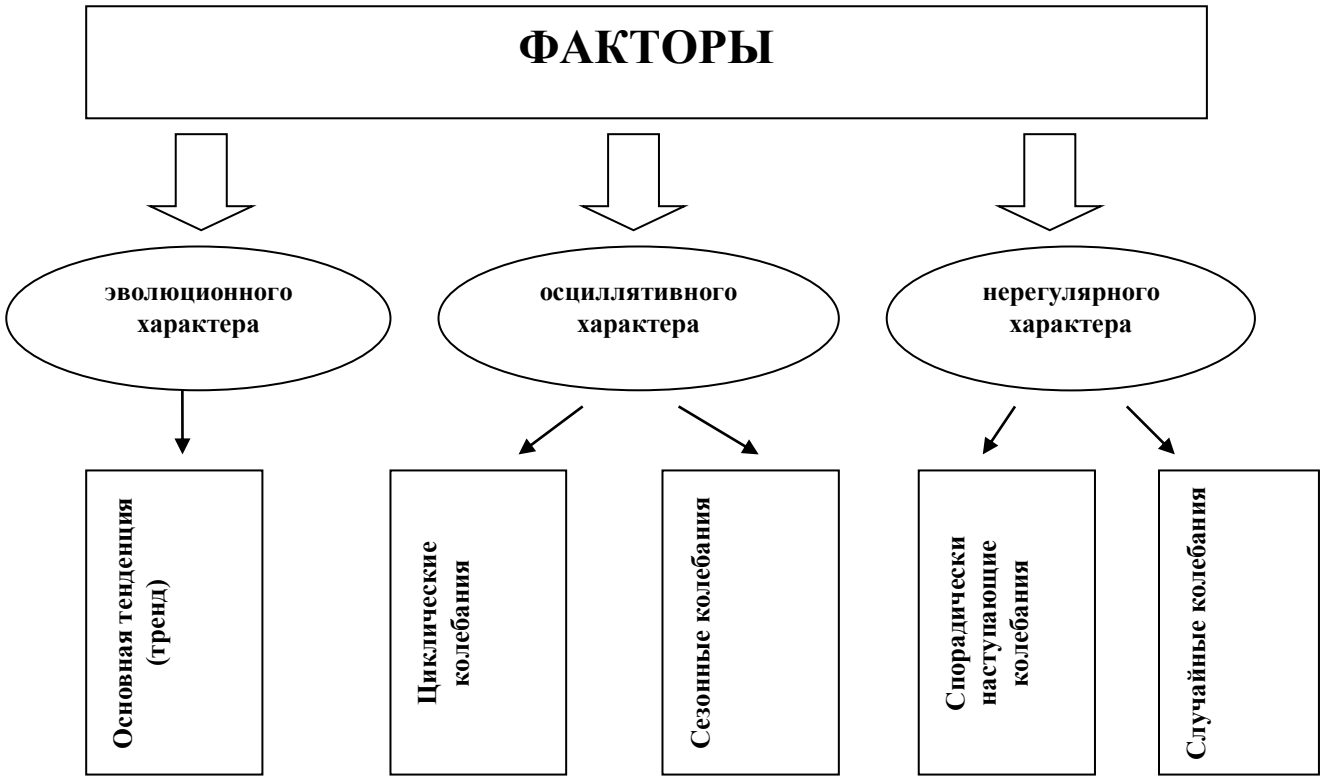
Нерегулярные факторы применительно к социально-экономическим явлениям формируют спорадически наступающие и случайные колебания.

К спорадически наступающим колебаниям (*интервенциям*) относятся изменения, вызванные существенным кратковременным воздействием на временной ряд, например, экологической катастрофой, резким ростом курс доллара

Случайные колебания являются результатом действия большого количества относительно слабых второстепенных факторов.

Таким образом, можно выделить четыре основные компоненты временного ряда:

1) основную тенденцию (тренд) (*T*);
2) циклическую (*C*);
3) сезонную (*S*);
4) случайную (*E*).



Наличие в структуре временного ряда четырех компонент привело к формированию двух основных направлений анализа временных рядов.

В рамках первого направления выделяют детерминированную (систематическую, закономерную) и случайную составляющие временного ряда.

Под *детерминированной составляющей временного ряда* $y_1,...,y_n$ понимается числовая последовательность $d_1,...,d_n$, элементы которой d_t вычисляются по определенному правилу как функция времени t. При этом мы получаем закономерные, предсказуемые изменения уровней временного ряда. Эта составляющая y_t может быть вычислена при каждом t как некоторая функция от текущего момента t.

В экономических (и многих других) приложениях в детерминированную составляющую временного ряда d_t обычно включают три первые компоненты: тренд, сезонную компоненту и циклическую компоненту.

Если мы полностью выявим детерминированную составляющую в поведении временного ряда, то оставшаяся часть выглядит хаотично и

непредсказуемо. Ее называют иррегулярной, или *случайной компонентой* временного ряда. Для описания и анализа случайной компоненты временных рядов обычно используют понятия и методы теории вероятностей. В частности, *случайной компонентой* называют случайную последовательность $\varepsilon_1,\ldots, \varepsilon_t,\ldots, \varepsilon_n$, имеющую некоторое распределение вероятностей.

Формы разложения (декомпозиции) временного ряда на детерминированную и случайную компоненты могут различаться. Наиболее простыми из них являются аддитивная и мультипликативная модели.

Аддитивной моделью временного ряда называется представление ряда в виде суммы детерминированной и случайной компонент, а именно:

$$y_t = d_t + \varepsilon_t \quad при \quad t=1,\ldots,n \quad или \quad y = D + E = T + C + S + E.$$

Мультипликативной моделью временного ряда называется представление ряда в виде произведения детерминированной и случайных компонент, а именно:

$$y_t = d_t \times \varepsilon_t \quad при \quad t=1,\ldots,n \quad или \quad y = D \times E = T \cdot C \cdot S \cdot E$$

Мультипликативные модели часто применяются при анализе экономических временных рядов.

В основе второго направления анализа временного ряда лежит его деление на два основных элемента: тенденцию (тренд) и колеблемость (отклонения уровней отдельных периодов или моментов времени от тренда). То есть второй элемент при данном аналитическом подходе представлен циклической, сезонной и случайной компонентами.

В разных реальных временных рядах эти два элемента находятся в неодинаковом соотношении, а в крайних случаях остается один элемент: ряд без колеблемости уровней представляет собой тренд в чистом виде, а ряд без тенденции динамики, но с колебаниями уровней около постоянной средней величины – это *стационарный* временной ряд.

Оба крайних случая крайне редки на практике. Обычно тенденция и колеблемость сочетаются в исходном ряду.

Задача статистического анализа в этом случае может формулироваться в двух аспектах:

1) «очистить» тенденцию от колебаний, измерить ее параметры. Колемлемость в этом случае выступает как помеха, «шум», мешающий выделить и интерпретировать «сигнал», то есть параметры тренда.

Нередко в учебной литературе взгляд на колеблемость как на помеху в изучении тенденции преобладает или является единственным.

Однако сама колеблемость также представляет собой важный предмет статистического исследования временных рядов.

2) выявление и измерение различных типов колебаний временного ряда.

Значение колеблемости многогранно:

– она позволяет выдвинуть гипотезы о причинах колебаний, о путях влияния на них;

– на основе параметров колеблемости можно ее прогнозировать или учитывать как фактор ошибки прогноза, то есть сделать прогноз более надежным и (или) точным;

На основе параметров и прогнозов колемлемости можно рассчитать резервы, страховой запас, необходимый для преодоления вредных последствий колебаний уровней.

7.2 Методы анализа основной тенденции во временных рядах

Анализ временного ряда обычно начинается с выделения трендовой компоненты.

Трендом называют плавно изменяющееся, устойчивое изменение временного ряда, происходящее в течение длительного времени под влиянием долговременных факторов, эффект которых сказывается постепенно. Он характеризует основную тенденцию развития явления. *Основная тенденция* – общее, направление к росту, снижению или стабилизации уровня явления, достаточно устойчивое на протяжении изучаемого этапа.

Во временных рядах экономических данных можно наблюдать тенденцию трех видов:

–среднего уровня;

–дисперсии;

–автокорреляции.

Тенденция среднего уровня аналитически выражается с помощью математической функции, вокруг которой варьируют фактические уровни изучаемого явления.

Тенденция дисперсии представляет собой тенденцию изменения отклонений между эмпирическими уровнями и тенденцией среднего уровня.

Тенденция автокорреляции характеризует изменения связи между отдельными уровнями временного ряда.

Прежде чем приступать к выделению тренда, следует проверить гипотезу о том, существует ли он вообще. Отсутствие тренда означает неизменность среднего уровня ряда во времени.

В настоящее время для проверки наличия тренда известно около десятка критериев, различающихся как по мощности, так и по сложности математического аппарата.

Одними из наиболее известных является метод, основанный на проверке разности средних двух разных частей одного и того же ряда, и метод Фостера – Стьюарта.

После того, как установлено наличие тенденции во временном ряду, производится ее описание посредством методов *выравнивания (сглаживания)*. Сущность выравнивания сводится к замене фактических уровней ряда расчетными значениями, имеющими значительно меньшую флуктуацию (величину отклонения, колебаний), чем исходные, эмпирические уровни. При этом основная тенденция проявляется более ясно, более наглядно. Различают методы механического и аналитического выравнивания.

Наиболее распространенными методами механического выравнивания являются метод скользящего среднего и метод экспоненциального сглаживания.

При сглаживании временного ряда *методом скользящей средней* выбирается размер интервала сглаживания *p.* Чаще всего его выбирают равным нечетному числу 3, 5 или 7. Чем больше интервал сглаживания, тем более гладкий вид имеет график скользящего среднего, тем более четко проявляется тенденция. Интервал сглаживания, равный или кратный периоду сезонного колебания, полностью устраняет сезонную компоненту.

Скользящие средние уровни рассчитываются сначала из p первых по порядку уровней временного ряда $\bar{y}_{m+1} = \frac{y_1 + y_2 + ... + y_p}{p}$, затем из p уровней, начиная со второго $\bar{y}_{m+2} = \frac{y_2 + y_3 + ... + y_{p+1}}{p}$ и так до тех пор, пока не вычислим среднее $\bar{y}_{n-m} = \frac{y_{n-p+1} + y_{n-p+2} + ... + y_n}{p}$.

В процессе расчета интервал сглаживания как бы скользит по исследуемому временному ряду с шагом, равным единице (отсюда и название метода). Полученная средняя относится к середине интервала скольжения.

Сглаженный ряд короче исходного на *p-1* число уровней. Для восстановления «потерянных» уровней в начале и конце сглаживаемого ряда используются следующие формулы:

- для $p = 3$ ($m=1$):

$$\bar{y}_1 = (7y_1 + 4y_2 - 2y_3)/9; \qquad \bar{y}_n = (7y_n + 4y_{n-1} - 2y_{n-2})/9$$

- для $p = 5$ ($m=2$):

$$\bar{y}_1 = (8y_1 + 2y_2 - y_3)/9; \qquad \bar{y}_n = (8y_n + 2y_{n-1} - y_{n-2})/9;$$

$$\bar{y}_2 = (7y_1 + 4y_2 + y_3 + 3y_4)/15; \qquad \bar{y}_{n-1} = (7y_n + 4y_{n-1} + y_{n-2} + 3y_{n-3})/15$$

При сглаживании временного ряда с четным числом уровней нужна дополнительная процедура *центрирования средних*.

Экспоненциальное среднее имеет вид:

$$\tilde{y}_t = (1-\alpha)\cdot\tilde{y}_{t-1} + \alpha\cdot y_t = \tilde{y}_{t-1} + \alpha\cdot(y_t - \tilde{y}_{t-1})$$

где $\tilde{y}_t$ – значение экспоненциального среднего в момент *t*;

y_t - фактическое значение временного ряда в момент *t*;

α – параметр сглаживания.

Параметр α характеризует скорость реакции экспоненциального среднего на изменение текущего значения временного ряда и одновременно определяет его способность сглаживать случайные флуктуации. Чем больше α, тем быстрее реакция экспоненциального среднего на изменение временного ряда и тем меньше его сглаживающие возможности. В качестве приемлемого рекомендуется брать α в пределах от 0,1 до 0,3.

Метод экспоненциального сглаживания учитывает устаревание данных наблюдения. При его реализации веса уровней временного ряда уменьшаются экспоненциально, в зависимости от «возраста» (давности) наблюдения – текущее наблюдение имеет вес α, а веса предшествующих ему значений равны, соответственно $\alpha(1-\alpha)$, $\alpha(1-\alpha)^2$ и т.д. (при использовании метода простого скользящего среднего все усредняемые уровни имеют одинаковый вес, равный 1/*p)*.

Аналитическое выравнивание – это подбор количественной (математической) модели, в аналитической форме и параметрах которой сконцентрировалась бы вся существенная информация о тенденции развития временного ряда;

– описание основной тенденции y_t как функция времени в форме уравнения определенной математической функции. В этом случае фактические (эмпирические) уровни заменяются теоретическими, вычисленными по соответствующему аналитическому уравнению.

Аналитическое выравнивание временного ряда выполняется в следующем порядке: сначала выбирается тип сглаживающей функции, затем определяются выборочные оценки параметров этой функции.

Выбор формы аналитического выражения тренда (вида аппроксимирующей функции) может быть осуществлен на основе:

- качественного, содержательного анализа сущности развития данного явления;
- результатов предыдущих исследований в данной области;

- графического изображения эмпирических или скользящих средних уровней ряда (в них колебания уже в некоторой степени оказываются погашенными);
- анализа изменения показателей временного ряда;
- статистико-математических критериев адекватности.

Наиболее часто используемыми при анализе экономических временных рядов являются следующие модели трендов:

– *линейная модель* $y_t = a_0 + a_1 t$, где

a_0 и a_1 – параметры уравнения;

a_0 – начальный уровень тренда в момент или период, принятый за начало отсчёта времени;

a_1 – среднее абсолютное изменение за единицу времени;

t – обозначение времени.

Параметр a_1 определяет направление развития: если $a_1 \succ 0$, то уровни ряда равномерно возрастают в среднем за единицу времени на величину a_1, если $a_1 \prec 0$, то происходит их равномерное снижение.

– *полиномиальные модели* (значение степени полинома n в практических задачах редко превышает 5). Это связано с тем, что при ограниченной длине временного ряда сложно получить надежные оценки параметров. На практике чаще всего применяется парабола второго порядка: $y_t = a_0 + a_1 t + a_2 t^2$ – описывает равномерно возрастающие ли равномерно убывающие изменения уровней.

Значение параметров a_0 и a_1 идентично предыдущему уравнению.

Параметр a_2 характеризует изменение интенсивности развития в единицу времени. При $a_2 \succ 0$ происходит ускорение развития, при $a_2 \prec 0$ – замедление развития.

Соответственно при параболической форме тренда возможны следующие варианты развития:

- если $a_1 \succ 0$; $a_2 \succ 0$ – ускорение роста;
- если $a_1 \succ 0$; $a_2 \prec 0$ – замедление роста;
- если $a_1 \prec 0$; $a_2 \prec 0$ – ускорение замедления;

- если $a_1 \prec 0;\ a_2 \succ 0$ – снижение замедления.

Параболы третьего порядка $y_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3$ описывает развитие с переменным ускорением.. Параметр a_3 отображает изменение ускорения (замедления);

- если $a_1 \succ 0;\ a_2 \succ 0;\ a_3 \succ 0$ – возрастающее ускорение роста;
- если $a_1 \succ 0;\ a_2 \succ 0; a_3 \prec 0$ – замедляющееся ускорение роста;
- если $a_1 \succ 0;\ a_2 \prec 0;\ a_3 \prec 0$ – возрастающее замедление роста;
- если $a_1 \succ 0;\ a_2 \prec 0;\ a_3 \succ 0$ – снижающееся замедление роста;
- если $a_1 \prec 0;\ a_2 \prec 0; a_3 \prec 0$ – возрастающее ускорение замедления;
- если $a_1 \prec 0;\ a_2 \prec 0; a_3 \succ 0$ – замедляющееся ускорение замедления;
- если $a_1 \prec 0;\ a_2 \succ 0; a_3 \prec 0$ – уменьшающееся снижение замедления
- если $a_1 \prec 0;\ a_2 \succ 0; a_3 \succ 0$ – возрастающее снижение замедления

– *экспоненциальный тренд* $y_t = a_0 a_1^{\,t}$, где a_0 – константа ряда, a_1 –темп изменения уровней в разах.

При a_1>1 экспоненциальный тренд выражает тенденцию ускоренного и всё более ускоряющегося возрастания уровней, при a_1<1 экспоненциальный тренд означает всё более замедляющегося снижения уровней динамического ряда.

– *гиперболический тренд* $y_t = a_0 + a_1 \frac{1}{t}$. Если $a_1 > 0$, то данный тренд выражает тенденцию замедляющегося снижения уровней (например, при изучении снижения себестоимости). Если $a_1 < 0$, то уровни тренда с течением времени возрастают, но не могут превысить определенного предела. (например, показатели использования оборудования, степень изношенности фондов). В целом, описывает тенденцию процессов, показатели которого затухают, то есть происходит переход от движения к застою.

–*степенной тренд* $y_t = a_0 t^{a_1}$, применим для отображения тенденции явлений с разной мерой пропорциональности изменений во времени;

–*логарифмический тренд* $y_t = a_0 + a_1 \log t$. Логарифмическая форма тренда применяется для отображения тенденции замедляющегося роста уровней при отсутствии предельно возможного значения, например, роста

спортивных достижений, производительности агрегата, продуктивности скота.

– *логистический тренд* $y_t = \frac{1}{1 + a_1 e^{a_0 + a_1 t}}$;

– *кривая Гомперца*: $y_t = K \cdot a_0^{a_1 t}$

Две последние модели задают кривые тренда *S*-образной формы. Они соответствуют процессам с постепенно возрастающими темпами роста в начальной стадии и постепенно затухающими темпами роста в конце. Необходимость подобных моделей обусловлена невозможностью многих экономических процессов продолжительное время развиваться с постоянными темпами роста или по полиномиальным моделям, в связи с их довольно быстром ростом (или уменьшением).

Расчет параметров аналитических моделей, как правило, осуществляется методом наименьших квадратов. МНК дает оценки параметров, отвечающие принципу максимального правдоподобия: сумма квадратов отклонений фактический уровней от тренда должна быть минимальной для данного типа уравнения.

Наиболее точным способом выбора формы тренда является *применение статистико-математических критериев*, в качестве которых могут выступать остаточное среднее квадратическое отклонение, средняя ошибка аппроксимации (ε_t), стандартизированная ошибка аппроксимации (σ_{y_t}), относительная ошибка аппроксимации (модифицированный коэффициент вариации V):

$$\varepsilon_t = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y - y_t}{y_t}\right|;\ \sigma_{\tilde{y}_t} = \sqrt{\frac{\sum (y - \tilde{y}_t)^2}{n - m}};\ V = \frac{\sigma_{\tilde{y}_t}}{\bar{y}} \cdot 100\ \%\ ,\text{где}$$

y и $\tilde{y}_t$ - соответственно фактические и теоретические значения ряда динамики;

n – число уровней ряда;

m – количество параметров в уравнении тренда.

Предпочтение отдаётся той функции, которая имеет наименьшую величину ошибки аппроксимации.

7.3 Методы распознавания типа колебаний и оценки параметров колеблемости

Если при изучении и измерении тенденции динамики колебания уровней играли лишь роль помех, «информационного шума», от которого следовало по возможности абстрагироваться, то на следующих этапах статистического анализа сама колеблемость становиться предметом исследования.

Для измерения силы колебаний уровней существуют показатели, аналогичные по содержанию с показателями пространственной вариации. Однако вариация в пространстве и колеблемость во времени принципиально различны по ряду позиций (одной из которых является то, что в основу методики расчета показателей колеблемости положено отклонение от тренда, а не от среднего уровня). Поэтому при обозначении этих понятий следует придерживаться соответствующей терминологии.

Абсолютные показатели:

- амплитуда отклонений или размах колебаний:

$$R_t = E_{\max} - E_{\min}, \text{ где}$$

$E_{\max}$ – максимальное положительное отклонение уровней от тренда;

$E_{\min}$ – максимальное отрицательное отклонение уровней от тренда.

- среднее абсолютное отклонение:

$$d_t = \frac{\sum_{t=1}^{n} |y - y_t|}{n - m}, \text{ где}$$

y – фактический уровень ряда динамики;

y_t –теоретический уровень ряда динамики;

n – число уровней;

m – число параметров в модели тренда.

- среднее квадратическое отклонение:

$$\sigma_t\left(S_{y_t}\right) = \sqrt{\frac{\sum (y - y_t)^2}{n - m}}.$$

Относительные показатели определяются делением абсолютных показателей на средний уровень ряда:

- относительное линейное отклонение:

$$V_d = \frac{d_t}{\bar{y}} \times 100\%$$

- коэффициент колеблемости:

$$V_{t(\sigma_t)} = \frac{\sigma_t}{\bar{y}} \times 100\% .$$

В качестве категории, противоположной колеблемости рассматривается понятие «устойчивости». Для её измерения используется *показатель устойчивости*, который бывает только относительным, изменяется от 0 до 1 (100%) и представляет собой разность между 1(100%) и коэффициентом колеблемости. Также данный показатель носит название *коэффициента алиенации:* $A = 1 - V_t$.

Дальнейшая задача состоит в распознавании типа колебаний и их измерении.

Типы колебаний статистических показателей разнообразны, но среди них в качестве основных выделяют следующие:

1) периодические колебания:

а–пилообразная или маятниковая колеблемость;

б–долгопериодическая циклическая колеблемость;

2) случайно распределенная во времени колеблемость.

К *способам распознавания* типов колебаний относятся:

1) графическое изображение временного ряда;

2) особенности отклонения фактических уровней от тренда;

3) подсчет числа локальных экстремумов в ряду отклонений от тренда. *Локальный экстремум* – поворотная точка – отклонение, большее или меньшее по алгебраической величине двух соседних.

4) по знаку и величине коэффициента автокорреляции отклонений от тренда разных порядков (I порядка – со сдвигом (лагом) на один год; II порядка – со сдвигом на два года; III порядка – со сдвигом на три года и т.д.):

Коэффициент автокорреляции отклонений I порядка:

$$r_u = \frac{\sum_{i=1}^{n-1} u_i u_{i+1}}{\frac{u_1^2}{2} + \sum_{i=2}^{n-1} u_i^2 + \frac{u_n^2}{2}};$$

$\sum_{i=1}^{n-1} u_i u_{i+1}$ - сумма произведений каждого отклонения на следующее, кроме последнего в ряду отклонений.

Типы колебаний	Особенности отклонения фактических уровней от тренда	Подсчет числа локальных экстремумов в ряду отклонений от тренда	По знаку и величине коэффициента автокорреляции отклонений от тренда разных порядков
Пилообразная колеблемостиь	регулярное чередование отклонений от тренда «вверх» и «вниз», то есть положительных и отрицательных по знаку через одно.	Общее количество локальных экстремумов равно *n-2* (все отклонения, кроме двух крайних являются поворотными точками)	Коэффициент автокорреляции I порядка близок к -1. При $r_u > -0,3$ - пилообразная составляющая несущественна или отсутствует при длине ряда не больше 20 уровней.
Долгосрочная периодическая колеблемость	Наличие нескольких подряд отклонений одного знака, сменяющихся таким же количеством отклонений противоположного знака подряд	На цикл приходится один минимум и один максимум; общее число поворотных точек *–2(n:l)*, где *l* – длительность цикла.	Коэффициент автокорреляции – величина положительная. При $r_u > +0,3$ - можно считать, что в общей колеблемости есть существенная циклическая составляющая; при $r_u > 0,6 - 0,7$ - циклическая составляющая является главной.
Случайно распределенная колеблемость	хаотичность последовательности отклонений	Среднее число локальных экстремумов равняется *2/3 (n-2)* при среднем квадратическом отклонении равном $\sqrt{\frac{16n-29}{90}}$	Коэффициент автокорреляции стремится к нулю. При $r_u \le \lvert 0,3 \rvert$ - случайная компонента преобладает в общем комплексе колебаний (в случае, если ряд состоит менее чем из19-22 уровне

Характерной особенностью *пилообразной колеблемости* является правильное, регулярное чередование отклонений от тренда «вверх» и «вниз», то есть положительных и отрицательных по знаку через одно. График такого временного ряда похож на зубья пилы – отсюда и название. Это похоже на колебания маятника вправо-влево, поэтому данный тип колеблемости называют также «маятниковой».

При пилообразной колеблемости из-за частой смены знаков отклонений от тренда не происходит аккумуляции ни положительных, ни отрицательных отклонений. Число положительных отклонений при достаточно большой длине ряда равно числу отрицательных отклонений.

Циклическая компонента c_t временного ряда описывает длительные периоды относительного подъема и спада. Она состоит из циклов, которые меняются по амплитуде и протяженности.

Для нахождения *длины циклы* нужно последовательно вычислить коэффициенты автокорреляции отклонений от тренда разных порядков, то есть с лагом 1,2, 3 и т.д. периодов времени. Наибольший по алгебраической величине коэффициент корреляции отметит длину цикла.

Как правило, длина циклов одинакова или хотя бы примерно равная. Если равенство отдельных циклов существенно нарушается, говорят о *квазициклической колеблемости*, то есть будто бы циклической.

Случайно распределенной во времени колеблемости свойственна хаотичность последовательности отклонений: после отрицательного отклонения от тренда может следовать снова отрицательное или даже два-три отрицательных, а может и положительное (два-три).

Сезонные колебания – это разновидность периодических колебаний. Для них характерны внутригодовые, повторяющиеся устойчиво из месяца в месяц (из квартала в квартал) изменения в уровнях, то есть это регулярно повторяющиеся подъемы и снижения уровней временного ряда внутри года на протяжении ряда лет.

Сезонные колебания характеризуются специальными показателями, которые называются *индексами сезонности*. Совокупность этих показателей отражает *сезонную волну*.

Для выявления сезонных колебаний обычно берут данные за несколько лет, распределенные по месяцам (кварталам). Данные за несколько лет (не

менее трех) используются для того, чтобы исключить случайные условия одного года.

Для вычисления индексов сезонности применяют различные методы.

Если временной ряд не содержит ярко выраженной тенденции в развитии, то индексы сезонности вычисляют непосредственно по эмпирическим данным без их предварительного выравнивания.

Для каждого месяца (квартала) рассчитывается средняя величина уровня (например, за три года), затем из них вычисляется среднемесячный уровень для всего ряда. Процентное отношение средних для каждого месяца к общему среднемесячному уровню ряда и будет индексом сезонности.

Если временной ряд содержит тенденцию, то предварительно производят выравнивание временного ряда.

Определяют индивидуальные индексы сезонности – отношение фактических месячных (квартальных) данных к соответствующим выровненным данным.

Далее находят средние арифметические из индивидуальных индексов сезонности для одноименных периодов.

Уровни тренда умножают на средние индексы сезонности соответствующих месяцев (кварталов) и получают уровень *модели – тренд-сезонность.*

Отклонение уровней тренд-сезонность от уровней тренда показывает отклонения за счет сезонности.

На базе суммы квадратов этих отклонений рассчитывается коэффициент сезонной колеблемости.

Отклонения эмпирических уровней временного ряда от уровней тренд - сезонность отражает за счет случайной колеблемости.

На базе суммы квадратов этих отклонений рассчитывается коэффициент случайной колеблемости.

Периодические колебания измеряются посредством *гармонического* (или спектрального) анализа, который представляет собой операцию по выражению заданной периодической функции в виде ряда Фурье по гармоникам разных порядков. Каждый член ряда представляет собой слагаемое постоянной величины с функциями косинусов и синусов определенного периода.

Список рекомендуемой литературы

1. Анализ данных на компьютере : учеб. пособие / Ю.Н. Тюрин, А.Н. Макаров; ред.: В.Э. Фигурнов. - 4-е изд., перераб. - М. : ИД ФОРУМ, 2013
2. Вадзинский Р. Статистические вычисления в среде Excel. –СПб.: Питер,2008.
3. Вуколов Э. А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL: учеб. пособие по специальности "Менеджмент организации". – 2-е изд., испр. и доп. – М. : ФОРУМ-ИНФРА-М, 2013. – 463 с.
4. Елисеева И.И., Юзбашев М.М. Общая теория статистики: Учебник. 5-е изд., перераб. и доп. М: Финансы и статистика, 2005 - 656 с.
5. Ефимова М.Р., Петрова Е.В., Румянцев В.Н. Общая теория статистики: Учебник.2-е изд., перераб. и доп. - М.: ИНФРА- М, 2006 - 416 с.
6. Кильдишев, Г.С. Многомерные группировки /Г.С. Кильдишев, Ю.И. Аболенцев.–М.: «Статистика», 1978.–160 с.
7. Конрад Карлберг. Бизнес-анализ с использованием Excel. – М.: Вильямс, 2012 – 576 с.
8. Мостеллер, Ф. Анализ данных и регрессия: В2-х вып. Вып.1 /Ф.Мостеллер, Д.Тьюки.–М.: Финансы и статистика, 1982.–317 с.
9. Мостеллер, Ф. Анализ данных и регрессия: В2-х вып. Вып.2 /Ф.Мостеллер, Д.Тьюки.–М.: Финансы и статистика, 1982.–239 с.
10. Ниворожкина Л.И., Арженовский С.В., Рудяга А.А. и др. Статистика: Учебник/ Под общей ред. Л.И. Ниворожкиной. – М.: Издательско- торговая корпорация «Дашков и К», 2012 .- 416 с.
11. Сиськов В.И. Корреляционный анализ в экономических исследованиях.–М.: Статистика, 1975.–168 с.
12. Смагин Б.И. Экономико-математические методы: учеб. пособие. М.: КолосС, 2012. 271 с.
13. Статистика: учебник / А.М. Годин. - 11-е изд., перераб. и доп. - 2014
14. Статистика: Учебник/ Под ред. В.С. Мхитаряна.- М: Экономисть, 2010.-671 с.
15. Статистика: Учебник/ Под ред. И.И. Елисеевой.-М: Издательство Юрайт, Высшее образование, 2011. – 565с.
16. Теория статистики: Учебник / Под ред. Р.А. Шмойловой .4-е изд., доп. и перераб. - М.: Финансы и статистика, 2005.- 656 с.
17. Шмойлова Р.А. и др. Практикум по теории статистики: Учебное пособие. 2-е изд., перераб. и доп.- М. : Финансы и статистика, 2005.- 656с.