Федеральное государственное образовательное учреждение
высшего профессионального образования
«Чувашский государственный университет имени И.Н. Ульянова»

Батыревский филиал

Кафедра экономических дисциплин

КОНСПЕКТ ЛЕКЦИЙ

по дисциплине
«Теория статистики»

Составитель: кандидат сельскохозяйственных наук, доцент
Баданов Геннадий Павлович

Батырево - 2009

Содержание

1. Понятие о статистике

1.1. Предмет и метод статистики

В научный обиход термин «статистика»[1] ввел немецкий ученый Готфрид Ахенваль в 1746 году, предложив заменить название курса «Государствоведение», преподававшегося в университетах Германии, на «Статистику», положив тем самым начало развитию статистики как науки и учебной дисциплины. Несмотря на это, статистический учет велся намного раньше: проводились переписи населения в Древнем Китае, осуществлялось сравнение военного потенциала государств, велся учет имущества граждан в Древнем Риме и пр.

У истоков статистической науки стояли 2 школы: *немецкая описательная* и *английская школа политических арифметиков*. Представители описательной школы (Герман Конринг, Готфрид Ахенваль, Август Людвиг Шленцер) своей задачей считали описание достопримечательностей государства: территории, населения, климата, политического устройства, вероисповедания, торговли и т.п. – без анализа закономерностей и связей между явлениями. Представители школы политических арифметиков (Уильям Петти, Джон Граунт, Эдмунд Галлей) своей главной задачей считали выявление на основе большого числа наблюдений различных закономерностей и взаимосвязей в изучаемых явлениях. Каждая школа развивалась своим путем, используя свои методы в исследованиях, но предмет изучения у них был общий – государство, общество и, в частности, массовые явления и процессы, происходящие в нем. Статистика сформировалась как наука в результате синтеза государствоведения и политической арифметики, причем от последней она взяла больше, поскольку статистика и в настоящее время призвана выявлять прежде всего различного рода закономерности в исследуемых явлениях.

Однако представители этих двух школ не дошли до теоретического обобщения практики учетно-статистических работ, до создания теории статистики. Эта задача была решена позднее, в XIX веке бельгийским ученым Адольфом Кетле, который дал определение предмета статистики, раскрыл суть ее методов. Под влиянием идей Кетле возникло третье направление статистической науки – *математико-статистическое*, которое получило свое развитие в работах таких ученых как: англичане Фрэнсис Гальтон, Фрэнсис Эджворт, Карл Пирсон, Одни Дж. Юл, Вильям Госсет, Рональд Фишер, Морис Дж.Кендэл, итальянец Коррадо Джини, русские – Пафнутий Львович Чебышёв, Андрей Андреевич Марков, Александр Михайлович Ляпунов, Александр Иванович и Александр Александрович Чупров и пр.

В настоящее время данный термин употребляется в 4 значениях:

1) *наука*, изучающая количественную сторону массовых явлений и процессов в неразрывной связи с их качественным содержанием – *учебный предмет* в высших и средних специальных учебных заведений;
2) *совокупность цифровых сведений*, характеризующих состояние массовых явлений и процессов общественной жизни; *статистические данные*, представляемые в отчетности предприятий, организаций, отраслей экономики, а также публикуемых в сборниках, справочниках, периодической печати и в сети Интернет, которые являются результатом статистической работы;
3) *отрасль практической деятельности* («статистический учет») по сбору, обработке, анализу и публикации массовых цифровых данных о самых различных явлениях и процессах общественной жизни[2];
4) некий *параметр* ряда случайных величин, получаемый по определенному алгоритму из результатов наблюдений, например, статистические критерии (критические статистики),

---

[1] От лат. *status* – состояние, положение вещей; первоначально термин употреблялся в значении «политическое состояние»

[2] Эту деятельность на профессиональном уровне осуществляет *государственная статистика* – Федеральная служба государственной статистики (ФСГС) и система ее учреждений, организованных по административно-территориальному признаку, а также *ведомственная статистика* (на предприятиях, ведомствах, министерствах и т.д.). Информация ФСГС публикуется в специадльных печатных изданиях, а также в сети Интернет: www.gks.ru (или www.fsgs.ru)

применяющиеся при проверке различных гипотез (предположительных утверждений) относительно природы или значений отдельных показателей исследуемых данных, особенностей их распределения и пр.[3]

Как и любая другая наука, статистика имеет свой предмет и метод исследования. Статистика изучает количественную сторону массовых общественных явлений в неразрывной связи с их качественной стороной или содержанием, а также исследует количественное выражение закономерностей общественного развития в конкретных условиях места и времени. Такое изучение основывается на системе категорий и понятий, отражающих наиболее общие и существенные свойства, признаки, связи и отношения предметов и явлений объективного мира.

Рассмотрим основные понятия, используемые в статистике.

1. *Статистическая совокупность* – множество социально-экономических объектов или явлений общественной жизни, объединенных качественной основой, но отличающихся друг от друга отдельными признаками, т.е. однородных в одном отношении, но разнородных в другом. Таковы, например, совокупность домохозяйств, семей, предприятий, фирм и т.п.
2. *Единица совокупности* – первичный элемент статистической совокупности, являющийся носителем признаков и основой ведущегося при обследовании счета.
3. *Признак единицы совокупности* – свойства единицы совокупности, которые различаются способами их измерения и другими особенностями, что дает основание для их классификации 1.

Таблица 1. Основная классификация признаков в статистике

Параметр классификации	*Вид признака*	*Пример признака*
По характеру выражения	Описательные (атрибутивные)	Цвет волос человека
	Количественные (числовые)	Рост человека
По способу измерения	Первичные (объемные)	Вес человека
	Вторичные (расчетные)	Производительность труда
По характеру вариации	Альтернативные	Пол человека
	Дискретные	Возраст человека
	Интервальные	Возраст группы людей
По отношению ко времени	Моментные	Количество денег в кармане человека
	Периодные	Заработная плата человека за месяц

4. *Статистический показатель* – понятие, отображающее количественные характеристики (размеры) или соотношения признаков общественных явлений. Статистические показатели можно подразделить на *первичные* (объемные) – характеризуют либо общее число единиц совокупности (объем совокупности), либо сумму значений какого-либо признака (объем признака) и выражаются абсолютными величинами и *вторичные* (расчетные) – задаются на единицу первичного показателя и выражаются относительными и средними величинами. Статистические показатели могут быть плановыми, отчетными и прогностическими.
5. *Система статистических показателей* – совокупность статистических показателей, отражающая взаимосвязи, которые объективно существуют между явлениями. Она охватывает все стороны общественной жизни как на макро-, так и на микроуровне. С изменением условий жизни общества меняются и системы статистических показателей, совершенствуется методология их расчета.

[3] Термин «статистика» как параметр, как статистический критерий употребляется преимущественно в математической статистике, некоторые из них (χ^2, t и др.) рассмотрены в соответствующих темах данного курса лекций

Совокупность приемов, пользуясь которыми статистика исследует свой предмет, составляет *метод* статистики. Можно выделить 3 группы статистических методов (этапов статистического исследования): 1) статистическое наблюдение; 2) сводка и 3) научный анализ исследуемых явлений.

Статистическое изучение тех или иных явлений предполагает как обязательное условие наличие информации, сведений об этих явлениях, поэтому первый этап, начало статистического исследования сводится к *сбору необходимой ин*формации. Научно организованный сбор сведений, заключающийся в регистрации тех или иных фактов, признаков, относящихся к каждой единице изучаемой совокупности, называется *статистическим наблюдением*.

В результате статистического наблюдения образуется масса первичной информации (сведений) о каждой единице совокупности. Чтобы получить характеристику всей исследуемой совокупности в целом, первичные данные должны быть подвергнуты обработке, обобщению. Обработка собранных первичных данных, включающая их группировку, обощение и оформление в таблицах, составляет второй этап статистического исследования, который называется *сводкой*.

На третьем этапе статистического исследования на основе итоговых данных сводки осуществляется *научный анализ исследуемых явлений*: рассчитываются различные обобщающие показатели в виде средних и относительных величин, выявляются определенные закономерности в распределениях, динамике показателей и т.п.

Таким образом, любое законченное статистическое исследование проходит в 3 этапа, между которыми, разумеется, могут быть перерывы во времени.

1.2. Статистическое наблюдение

Люди по-разному относятся к статистической информации: одни не воспринимают ее, другие безоговорочно верят, а третьи согласны с мнением английского политика Дизраэли: «Существует 3 типа лжи: ложь, наглая ложь и статистика»[4], однако ему же принадлежит следующее утверждение: «В жизни, как правило, преуспевает больше тот, кто располагает лучшей информацией»[5].

Статистическое наблюдение является начальным этапом статистического исследования, поэтому от того, насколько полными и качественными окажутся собранные первичные данные, зависят в значительной степени и конечные результаты работы, и выводы исследователей. В статистической практике используются разные формы, виды и способы наблюдения.

Различают 3 *формы* организации наблюдения: статистическая отчетность, специально организованные статистические обследования и регистры.

1. *Статистическая отчетность* – это особая форма организации сбора данных государственной статистикой о деятельности хозяйствующих субъектов, которые обязаны заполнять документы-бланки, называемые формами статистической отчетности. *Форма статистической отчетности* – это специальный документ-бланк, содержащий перечень определенных показателей, сведений, характеризующих ту или иную хозяйственую единицу и результаты ее деятельности, заполняемый на основе данных опертивного или бухгалтерского учета и представляемые в государственные статистические органы для дальнейшего обобщения. Перечень и содержание форм статистической отчетности утверждается органами государственной статистики и является обязательной для установленного круга предприятий и организаций. Каждая форма отчетности имеет шифр и название. В соответствии со сроками представления отчетность бывает *суточная* (ежедневная), *недельная*, *месячная*, *квартальная*, *полугодовая* и *годовая*. Все эти виды отчетности, кроме годовой, объединяют одним названием – *текущая* отчетность. Каждая форма отчетности должна прдставляться в установленные для нее сроки.

---

[4] «There are three types of lies - lies, damn lies, and statistics» (Benjamin Disraeli, 1804 – 1881)

[5] « As a general rule, the most successful man in life is the man who has the best information »

2. Круг явлений общественной жизни настолько велик, что полный охват их отчетностью невозможен. Во всех случаях, когда необходимо получить сведения, по которым отсутствует отчетность, когда требуется уточнить или дополнить данные той или иной отчетности либо провести разовое детальное, всестороннее обследование каких-либо объектов, применяют *специально организованные статистические наблюдения*, проводимые в виде переписей или специальных обследований (выборочных или сплошных). Такие обследования используются как органами статистики, так и отдельными хозяйствующими субъектами.

3. Наблюдение через *регистры* – сравнительно новая форма организации статистического наблюдения, основанная на применении компьютерных технологий. Регистр – это поименованный и постоянно уточняемый перечень тех или иных единиц наблюдения, созданный для непрерывного длительного статистического наблюдения за определенной совокупностью, в котором содержится информация о каждой единице совокупности (например, ЕГРПО – Единый государственный регистр предприятий и организаций).

Необходимо отметить, что все 3 организационные формы статистического наблюдения не противостоят, а дополняют друг друга, позволяя более глубоко, всесторонне изучать отдельные явления и процессы общественной жизни.

По времени регистрации фактов различают *текущее* (непрерывное) и *прерывное* наблюдение. Последнее, в свою очередь, подразделяется на *единовременное* и *периодическое*.

По охвату единиц наблюдения различают *сплошное*, когда наблюдению подлежат все единицы изучаемой совокупности, и *несплошное*. Несплошное наблюдение подразделяется на следующие виды: 1) *наблюдение основного массива* (исключаются из наблюдения малозначимые единицы); 2) *анкетное* (добровольное заполнение анкет приводит к несплошному виду наблюдения); 3) *выборочное* (случайный отбор единиц из изучаемой совокупности); 4) *монографическое* (детальное изучение какой-то одной единицы совокупности).

По источникам собираемых сведений различают следующие способы наблюдения: 1) *непосредственное* (осмотр, измерение, взвешивание); 2) *документальное* (на основе отчетности); 3) *опрос* (сведения регистрируются со слов опрашиваемой единицы наблюдения). Способы опроса: *экспидиционный*, *саморегистрация*, *корреспондентский* и *явочный*.

Любое статистическое исследование необходимо начинать с точной формулировки его цели и конкретных задач, а следовательно и тех сведений, которые могут быть получены в процессе наблюдения. После этого определяется объект и единица наблюдения, разрабатывается программа, выбирается вид и способ наблюдения.

Объект наблюдения – совокупность социально-экономических явлений и процессов, которые подлежат исследованию, или точные границы, в пределах которых будут регистрироваться статистические сведения. В ряде случаев пользуются цензом. *Ценз* – ограничительный признак, которому должны удовлетворять все единицы изучаемой совокупности. *Единицей наблюдения* называется составная часть объекта исследования, которая служит основой счета и обладает признаками, подлежащими регистрации при наблюдении. *Программа наблюдения* – перечень вопросов, по которым собираются сведения, либо перечень признаков или показателей, подлежащих регистрации. Она оформляется в виде бланка (анкеты, формуляра), в который заносятся первичные сведения. К нему прилагается инструкция (или указания на самих формулярах), разъясняющая смысл вопросов.

Организационные вопросы статистического наблюдения связаны с определением субъекта, места, времени, формы и способа наблюдения. *Субъект* наблюдения – орган, осуществляющий наблюдение. *Время* наблюдения – период, в течение которого будет проводиться наблюдение (срок наблюдения), либо время, к которому относятся регистрируемые сведения (критический момент наблюдения).

1.3. Сводка и группировка статистических данных

Сводка – научно организованная обработка материалов наблюдения (по заранее разработанной программе), включающая в себя кроме обязательного контроля собранных данных, систематизацию, группировку материалов, составление таблиц, получение итогов по группам и в целом. Программа сводки включает определение групп и подгрупп, системы показателей и видов таблиц. По технике и способу выполнения сводка может быть ручной либо механизированной.

Группировка – разбиение совокупности на группы, однородные по какому-либо признаку или объединение отдельных единиц совокупности в группы, однородные по каким-либо признакам. Устойчивое разграничение объектов называется классификацией или стандартом, в котором каждая атрибутивная запись может быть отнесена лишь к одной группе или подгруппе. Метод группировки основывается на двух категориях – группировочном признаке и интервале.

Группировочный признак – признак, по которому происходит объединение отдельных единиц совокупности в однородные группы. Он может носить как количественный, так и качественный характер. В ряде случаев группировка, которая представляется чисто качественной, в конечном итоге оказывается основанной на количественном признаке. Такова, например, классификация промышленных предприятий по отраслям. Поскольку одно и то же предприятие выпускает продукцию разных видов, статистика решает этот вопрос по количественному преобладанию того или иного вида.

Интервал очерчивает количественные границы групп и представляет собой промежуток между максимальным и минимальным значениями признака в группе. Интервалы бывают равные, неравные, закрытые (когда имеется верхняя и нижняя граница) и открытые (когда одна из границ отсутствует).

Статистические группировки и классификации преследуют цели выделения качественно однородных совокупностей, изучения структуры совокупности, исследования взаимосвязи факторных и результативных признаков. Каждой из этих целей соответствует особый вид группировки: типологическая, структурная и аналитическая.

В зависимости от числа положенных в основание группировки признаков различают простые и многомерные группировки. Простая группировка выполняется по одному признаку. Среди простых группировок особо выделяются ряды распределения. *Ряд распределения* – группировка, в которой для характеристики групп, упорядоченно расположенных по значению признака применяется один показатель – численность группы (более подробно об этом – тема 3 и 4).

Многомерная группировка производится по двум и более признакам. Частным случаем многомерной группировки является комбинационная группировка, базирующаяся на двух и более признаках, взятых во взаимосвязи.

По отношениям между признаками выделяют: *иерархические* группировки, выполняемые по двум и более признакам, при этом значения второго признака определяются областью значений первого (например, классификация отраслей промышленности по подотраслям); *неиерархические* группировки, когда строгой зависимости значений второго признака от первого не существует.

По очередности обработки информации группировки бывают *первичными*, составленные на основе первичных данных, и *вторичные*, являющиеся результатом перегруппировки ранее уже сгруппированного материала.

В соответствии со временным критерием различают *статические* группировки, дающие характеристику совокупности на определенный момент или за определенный период, и *динамические*, показывающие переходы единиц из одних групп в другие.

1.4. Формы представления статистических данных

Статистические данные должны быть представлены так, чтобы ими можно было пользоваться. Существует 3 основных формы представления статистических данных:

1) текстовая – включение данных в текст;
2) табличная – представление данных в таблицах;

3) графическая – выражение данных в виде графиков.

Текстовая форма применяется при малом количестве цифровых данных.

Табличная форма применяется чаще всего, так как является более эффективной формой представления статистических данных. В отличие от математических таблиц, которые по начальным условиям позволяют получить тот или иной результат, статистические таблицы рассказывают языком цифр об изучаемых объектах.

Статистическая таблица – это система строк и столбцов, в которых в определенной последовательности и связи излагается статистическая информация о социально-экономических явлениях.

Таблица 2. Внешняя торговля РФ за 2000 – 2006 годы, млрд.долл.

Показатель	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*
Внешнеторговый оборот	149,9	155,6	168,3	212	280,6	368,9	468,4
Экспорт	105	101,9	107,3	135,9	183,2	243,6	304,5
Импорт	44,9	53,8	61	76,1	97,4	125,3	163,9
Сальдо торгового баланса	60,1	48,1	46,3	59,9	85,8	118,3	140,7
в том числе:							
со странами дальнего зарубежья							
экспорт	90,8	86,6	90,9	114,6	153	210,1	261,1
импорт	31,4	40,7	48,8	61	77,5	103,5	138,6
сальдо торгового баланса	59,3	45,9	42,1	53,6	75,5	106,6	122,5

Например, в табл. 2 представлена информация о внешней торговле России, выражать которую в текстовой форме было бы неэффективным.

Различают *подлежащее* и *сказуемое* статистической таблицы. В подлежащем указывается характеризуемый объект – либо единицы совокупности, либо группы единиц, либо совокупность в целом. В сказуемом дается характеристика подлежащего, обычно в числовой форме. Обязателен *заголовок* таблицы, в котором указывается к какой категории и к какому времени относятся данные таблицы.

По характеру подлежащего статистические таблицы подразделяются на *простые*, *групповые* и *комбинационные*. В подлежащем простой таблицы объект изучения не подразделяется на группы, а дается либо перечень всех единиц совокупности, либо указывается совокупность в целом (например, табл. 11). В подлежащем групповой таблицы объект изучения подразделяется на группы по одному признаку, а в сказуемом указываются число единиц в группах (абсолютное или в процентах) и сводные показатели по группам (например, табл. 4). В подлежащем комбинационной таблицы совокупность подразделяется на группы не по одному, а по нескольким признакам (например, табл. 2).

При построении таблиц необходимо руководствоваться следующими *общими правилами*.

1. Подлежащее таблицы располагается в левой (реже – верхней) части, а сказуемое – в правой (реже – нижней).
2. Заголовки столбцов содержат названия показателей и их единицы измерения.
3. Итоговая строка завершает таблицу и располагается в ее конце, но иногда бывает первой: в этом случае во второй строке делается запись «в том числе», и последующие строки содержат составляющие итоговой строки.
4. Цифровые данные записываются с одной и той же степенью точности в пределах каждого столбца, при этом разряды чисел располагаются под разрядами, а целая часть отделяется от дробной запятой.
5. В таблице не должно быть пустых клеток: если данные равны нулю, то ставится знак «–» (прочерк); если данные не известны, то делается запись «сведений нет» или ставится знак «…» (троеточие). Если значение показателя не равно нулю, но первая значащая цифра появляется после принятой степени точности, то делается запись 0,0 (если, скажем, была принята степень точности 0,1).

Иногда статистические таблицы дополняются графиками, когда ставится цель подчеркнуть какую-то особенность данных, провести их сравнение. Графическая форма является самой эффективной формой представления данных с точки зрения их восприятия. С

помощью графиков достигается наглядность характеристики структуры, динамики, взаимосвязи явлений, их сравнения.

Статистические графики – это условные изображения числовых величин и их соотношений посредством линий, геометрических фигур, рисунков или географических карт-схем. Графическая форма облегчает рассмотрение статистических данных, делает их наглядными, выразительными, обозримыми. Однако графики имеют определенные ограничения: прежде всего, график не может включить столько данных, сколько может войти в таблицу; кроме того, на графике показываются всегда округленные данные – не точные, а приблизительные. Таким образом, график используется только для изображения общей ситуации, а не деталей. Последний недостаток – трудоемкость построения графиков. Он может быть преодолен использованием персонального компьютера (например, «Мастером диаграмм» из пакета *Microsoft Office Excel*).

По способу построения графики делятся на *диаграммы*, *картограммы* и *картодиаграммы*.

Наиболее распространенным способом графического изображения данных являются диаграммы, которые бывают следующих видов: линейные, радиальные, точечные, плоскостные, объемные, фигурные. Вид диаграмм зависит от вида представляемых данных и задачи построения. В любом случае график обязательно сопровождается заголовком – над или под полем графика. В заголовке указывается, какой показатель изображен, по какой территории и за какое время.

Линейные графики используются для представления количественных переменных: характеристики вариации их значений, динамики, взаимосвязи между переменными. Вариация данных анализируется с помощью *полигона распределения*, *кумуляты* (кривой «меньше, чем») и *огивы* (кривой «больше, чем»). Полигон распределения рассматривается в теме 4 (напр., рис. 5.). Для построения кумуляты значения варьирующего признака откладываются по оси абсцисс, а на оси ординат помещаются накопленные итоги частот или частостей (от f_1 до $\sum f$). Для построения огивы на оси ординат помещаются накопленные итоги частот в обратном порядке (от $\sum f$ до f_1). Кумуляту и огиву по данным табл. 4. изобразим на рис. 1.

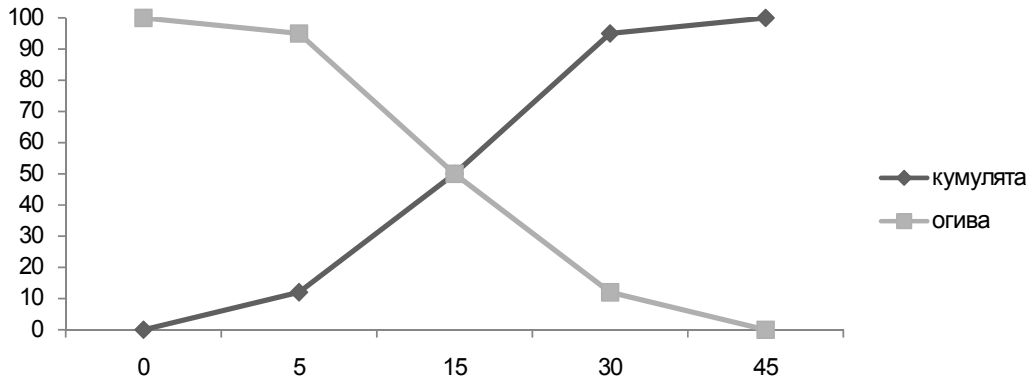


Рис. 1. Кумулята и огива распределения товаров по величине таможенной стоимости

Применение линейных графиков в анализе динамики рассматривается в теме 5 (напр., рис. 13), а использование их для анализа связей – в теме 6 (напр., рис.21). В теме 6 также рассмотрено использование точечных диаграмм (напр., рис. 20).

Линейные графики подразделяются на *одномерные*, используемые для представления данных по одной переменной, и *двумерные* – по двум переменным. Примером одномерного линейного графика является полигон распределения, а двумерного – линия регрессии (напр., рис. 21).

Иногда при больших изменениях показателя прибегают к логарифмической шкале. Например, если значения показателя изменяются от 1 до 1000, то это может вызвать затруднения при построении графика. В таких случаях переходят к логарифмам значений показателя, которые не будут столь сильно различаться: $lg\ 1 = 0$, $lg\ 1000 = 3$.

Среди *плоскостных* диаграмм по частоте использования выделяются столбиковые диаграммы (гистограммы), на которых показатель представляется в виде столбика, высота которого соответствует значению показателя (напр., рис. 4).

Пропорциональность площади той или иной геометрической фигуры величине показателя лежит в основе других видов плоскостных диаграмм: *треугольных*, *квадратных*, *прямоугольных*. Можно использовать и сравнение площадей круга – в этом случае задается радиус окружности.

Ленточная диаграмма представляет показатели в виде горизонтально вытянутых прямоугольников, а в остальном не отличается от столбиковой диаграммы.

Из плоскостных диаграмм часто используется *секторная диаграмма*, которая применяется для иллюстрации структуры изучаемой совокупности. Вся совокупность принимается за 100%, ей соответствует общая площадь круга, площади секторов соответствуют частям совокупности. Построим секторную диаграмму структуры внешней торговли РФ в 2006 году по данным табл. 2 (см. рис. 2). При использовании компьютерных программ секторные диаграммы строятся в объемном виде, то есть не в двух, а в трех плоскостях (см. рис. 3).

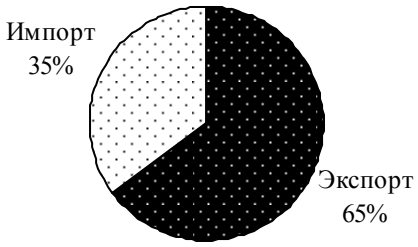


Рис. 2. Простая секторная диаграмма

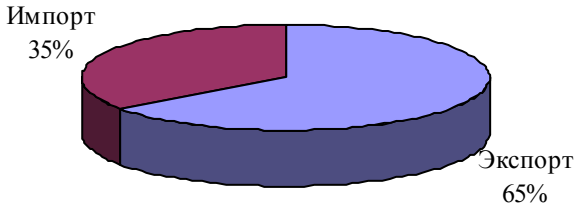


Рис. 3. Объемная секторная диаграмма

Фигурные (картинные) диаграммы усиливают наглядность изображения, так как включают рисунок изображаемого показателя, размер которого соответствует размеру показателя.

При построении графика одинаково важно все – правильный выбор графического изображения, пропорций, соблюдение правил оформления графиков. Подробнее эти вопросы освещаются в [8] и [5].

Картограммы и картодиаграммы применяются для изображения географической характеристики изучаемых явлений. Они показывают размещение изучаемого явления, его интенсивность на определенной территории – в республике, области, экономическом или административном округе и т.д.. Построение картограмм и картодиаграмм рассматривается в специальной литературе, например [3].

1.5. Контрольные задания

Выбрать какой-либо реальный объект наблюдения (например, студентов курса, факультета, преподавателей, родственников, друзей и т.п.). Спроектировать процесс наблюдения: сформулировать цель наблюдения; определить состав признаков, подлежащих регистрации; выбрать вид наблюдения; разработать инструментарий наблюдения. Провести спроектированное наблюдение, т.е. собрать сведения об объекте наблюдения, оформить результаты наблюдения и сдать преподавателю на проверку.

2. Обобщающие статистические показатели

2.1. Абсолютные величины

Для характеристики массовых явлений статистика использует *статистические величины (показатели)*, которые характеризуют группы единиц или совокупность (явление) в целом. Статистические величины (показатели) подразделяются на *абсолютные, относительные* и *средние*.

Результаты статистических наблюдений представляют собой ***абсолютные величины***, отражающие уровень развития какого-либо явления или процесса (например, величина экспорта/импорта *i*-го товара в *j*-ю страну). Абсолютные величины обозначаются *X*, а их общее количество в статистической совокупности *N*.

Абсолютные величины всегда имеют свою единицу измерения (размерность), присущую изучаемому явлению. Широко распространены следующие *виды единиц измерения*:

1) *натуральные*, подразделяющиеся на простые (например, штуки, тонны, метры) и сложные (составные), представляющие собой комбинацию двух разноименных величин (например, киловатт-час);
2) *условно-натуральные* (например, алкогольные напитки учитываются в дкл 100% спирта, а различные виды топлива соизмеряют по условному топливу с теплотворной способностью 7000 ккал/кг или 29,3 МДж/кг[6].);
3) *стоимостные*, позволяющие соизмерить в денежной форме товары, которые нельзя соизмерить в натуральной форме (доллары США, рубли и т.д.).

Количество единиц с одинаковым значением признака обозначается f и называется *частота*[7]. Очевидно, что суммируя число всех единиц с одинаковыми значениями признака[8], получаем *N*, то есть . (1):

$$\sum f = N . \qquad (1)$$

Анализируя абсолютные величины, например, статистические данные о торговле, необходимо сопоставлять эти данные во времени и пространстве, исследовать закономерности их изменения и развития, изучать структуру совокупностей. С помощью абсолютных величин эти задачи не выполнимы, в этом случае необходимо использовать относительные величины.

2.2. Относительные величины

Относительная величина – это результат деления (сравнения) двух абсолютных величин. В числителе дроби стоит величина, которую сравнивают, а в знаменателе – величина, с которой сравнивают (база сравнения). Например, если сопоставить величины экспорта США и России, которые в 2005 году составили 904,383 и 243,569 млрд. долл. соответственно, то относительная величина покажет, что величина экспорта США в 3,71 раза (904,383/243,569) больше экспорта России, при этом базой сравнения является величина экспорта России. Полученная относительная величина выражена в виде *коэффициента*, который показывает, во сколько раз сравниваемая абсолютная величина больше базисной. В данном примере база сравнения принята за единицу. В случае если основание принимается за 100, относительная величина выражается в *процентах* (%), если за 1000 – в *промилле* (‰). Выбор той или иной формы относительной величины зависит от ее абсолютного значения:

– если сравниваемая величина больше базы сравнения в 2 раза и более, то выбирают форму коэффициента (как в вышеприведенном примере);

---

[6] Аналогично общее количество школьных тетрадей измеряется в у.ш.т. (условные школьные тетради размером 12 листов), продукция консервного производства измеряется в у.к.б. (условные консервные банки емкостью 1/3 литра или 400 грамм); продукция моющих средств приводится к условной жирности 40%

[7] f – это начальная буква англ. слова *frequency* – частота

[8] В статистике, в отличие от математики, пределы суммирования не ставятся, а подразумеваются, так как абсолютные величины здесь не абстрактные, а смысловые (суммируются все величины совокупности – с первой по последнюю)

- если относительная величина близка к единице, то, как правило, ее выражают в процентах (например, сравнив величины экспорта России в 2006 и 2005 годах, которые составили 304,5 и 243,6 млрд. долл. соответственно, можно сказать, что экспорт в 2006 году составляет 125% от 2005 года [304,5/243,6*100%]);
- если относительная величина значительно меньше единицы (близка к нулю), ее выражают в промилле (например, в 2004 году Россия экспортировала в страны-СНГ всего 4142 тыс. т нефтепродуктов, в том числе в Грузию 10,7 тыс. т, что составляет 0,0026 [10,7/4142], или 2,6‰ от всего экспорта нефтепродуктов в страны СНГ).

Различают относительные величины динамики, структуры, координации, сравнения и интенсивности, для краткости именуемые в дальнейшем *индексами*.

Индекс динамики[9] характеризует изменение какого-либо явления во времени. Он представляет собой отношение значений одной и той же абсолютной величины в разные периоды времени. Данный индекс определяется по формуле , (2):

$$i_Д = \frac{X_1}{X_0}, \qquad (2)$$

где цифры означают: 1 – отчетный или анализируемый период, 0 – прошлый или базисный период.

Критериальным значением индекса динамики служит единица (или 100%), то есть если $i_Д$ >1, то имеет место рост (увеличение) явления во времени; если $i_Д$ =1 – стабильность; если $i_Д$ <1 – наблюдается спад (уменьшение) явления. Еще одно название индекса динамики – *индекс изменения*, вычитая из которого единицу (100%), получают *темп изменения (динамики)*[10] с критериальным значением 0, который определяется по формуле (3):

$$T = i_Д - 1. \qquad (3)$$

Если T>0, то имеет место рост явления; T=0 – стабильность, T<0 – спад.

В рассмотренном выше примере про экспорт России в 2006 и 2005 году был рассчитан именно индекс динамики по формуле , (2): $i_Д$ = 304,5/243,6*100% = 125%, что больше критериального значения 100%, что свидетельствует об увеличении экспорта. Используя формулу (3) получим темп изменения: T = 125% – 100% = 25%, который показывает, что экспорт увеличился на 25%.

Разновидностями индекса динамики являются индексы планового задания и выполнения плана, рассчитываемые для планирования различных величин и контроля их выполнения.

Индекс планового задания – это отношение планового значения признака к базисному. Он определяется по формуле (4):

$$i_{ПЗ} = \frac{X'_1}{X_0}, \qquad (4)$$

где X'_1 – планируемое значение; X_0 – базисное значение признака.

Например, таможенное управление перечислило в федеральный бюджет в 2006 году 160 млрд.руб., а на следующий год запланировали перечислить 200 млрд.руб., значит по формуле (4): $i_{пз}$ = 200/160 = 1,25, то есть плановое задание для таможенного управления на 2007 год составляет 125% от предыдущего года.

---

[9] Во многих учебниках по статистике встречается другое название индекса динамики – *темп роста*. Использование такого названия не совсем логично, так динамика может быть различна (не только рост, но и спад, а также стабильность), поэтому наиболее правильным является использование названия «индекс динамики» или «индекс изменения»

[10] Часто встречается и другое название темпа изменения – *темп прироста*, что не совсем логично (см. предыдущую сноску)

Для определения процента выполнения плана необходимо рассчитать *индекс выполнения плана*, то есть отношение наблюдаемого значения признака к плановому (оптимальному, максимально возможному) значению по формуле . (5):

$$i_{ВП} = \frac{X_1}{X_1'}. \qquad (5)$$

Например, на январь-ноябрь 2006 года таможенные органы запланировали перечислить в федеральный бюджет 1,955 трлн. руб., но фактически перечислили 2,59 трлн. руб., значит по формуле . (5): $i_{ВП}$ = 2,59/1,955 = 1,325, или 132,5%, то есть плановое задание выполнили на 132,5%.

Индекс структуры (доля) – это отношение какой-либо части объекта (совокупности) ко всему объекту. Он определяется по формуле (6):

$$i_{СТ} = d = \frac{f}{\sum f} \qquad (6)$$

В рассмотренном выше примере про экспорт нефтепродуктов в страны СНГ, была рассчитана доля этого экспорта в Грузию по формуле (6): d=10,7/4142 = 0,0026, или 2,6‰.

Индекс координации – это отношение какой-либо части объекта к другой его части, принятой за основу (базу сравнения). Он определяется по формуле . (7):

$$i_К = \frac{f}{f_б}. \qquad (7)$$

Например, импорт России в 2006 году составил 163,9 млрд.долл., тогда, сравнив его с экспортом (база сравнения), рассчитаем индекс координации по формуле . (7): $i_К$ = 163,9/304,5 = 0,538, который показывает соотношение между двумя составными частями внешнеторгового оборота, то есть величина импорта России в 2006 году составляет 53,8% от величины экспорта. Меняя базу сравнения на импорт, по той же формуле получим: $i_К$ = 304,5/163,9 = 1,858, то есть экспорт России в 2006 году в 1,858 раза больше импорта, или экспорт составляет 185,8% от импорта.

Индекс сравнения – это сравнение (соотношение) разных объектов по одинаковым признакам. Он определяется по формуле (8):

$$i_С = \frac{X_А}{X_Б}, \qquad (8)$$

где *А*, *Б* – сравниваемые объекты.

В рассмотренном выше примере, в котором сопоставлялись величины экспорта США и России, был рассчитан именно индекс сравнения по формуле (8): $i_с$ = 904,383/243,569 = 3,71. Меняя базу сравнения (то есть экспорт России – объект А, а экспорт США – объект Б), по той же формуле получим: $i_с$ = 243,569/904,383 = 0,27, то есть экспорт России составляет 27% от экспорта США.

Индекс интенсивности – это соотношение разных признаков одного объекта между собой. Он определяется по формуле (9):

$$i_{ИН} = \frac{X}{Y}. \qquad (9)$$

где X – один признак объекта; Y – другой признак этого же объекта

Например, показатели выработки продукции в единицу рабочего времени, затрат на единицу продукции, цены единицы продукции и т.д.

2.3. Средние величины

Как уже неоднократно было сказано ранее, статистика изучает массовые явления и процессы. Каждое из таких явлений обладает как общими для всей совокупности, так и

особенными, индивидуальными свойствами. Различие между индивидуальными явлениями называют *вариацией*, о ней подробно будет рассказано в теме 3. Здесь же рассмотрим другое свойство массовых явлений – присущую им близость характеристик отдельных явлений. В этом свойстве заключается причина широчайшего применения ***средних величин***. Главное значение средних величин состоит в их обобщающей функции, то есть замене множества различных индивидуальных значений признака средней величиной, характеризующей всю совокупность явлений.

Виды средних величин различаются прежде всего тем, какое свойство, какой параметр исходной варьирующей массы индивидуальных значений признака должен быть сохранен неизменным.

Средней арифметической величиной называется такое среднее значение признака, при вычислении которого общий объем признака в совокупности сохраняется неизменным. Иначе можно сказать, что средняя арифметическая величина – среднее слагаемое. При ее вычислении общий объем признака мысленно распределяется поровну между всеми единицами совокупности. Исходя из определения, формула средней арифметической величины имеет вид (10):

$$\overline{X} = \frac{X_1 + X_2 + ... + X_N}{N} = \frac{\sum X}{N}. \quad (10)$$

По формуле (10) вычисляются средние величины первичных признаков, если известны индивидуальные значения признака. Если изучаемая совокупность велика, исходная информация чаще представляет собой ряд распределения или группировку, как, например, табл. 3.

Таблица 3. Распределение студентов группы дневного отделения по возрасту

Возраст студентов, X	17	18	19	20	21
Число студентов, f	3	5	7	4	2

Средний возраст должен представлять собой результат равномерного распределения общего (суммарного) возраста всех студентов. Общий (суммарный) возраст всех студентов, согласно исходной информации табл. 3, можно получить как сумму произведений значений признака в каждой группе X_i, на число студентов с таким возрастом f_i (частоты). Получим формулу (11):

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i f_i}{\sum_{i=1}^{N} f_i}, \quad (11)$$

где i – число групп.

Такую форму средней арифметической величины называют *взвешенной арифметической средней*[11] в отличие от простой средней, рассчитанной по формуле (10). В качестве весов здесь выступают количество единиц совокупности в разных группах. Название «вес» выражает тот факт, что разные значения признака имеют неодинаковую «важность» при расчете средней величины. «Важнее», весомее возраст студентов 18, 19, 20 лет, а такие значения возраста как 17, 20 или 21 при расчете средней не играют большой роли – их «вес» мал.

По формуле (11) по данным табл. 3 имеем:

$$\overline{X} = \frac{17*3+18*5+19*7+20*4+21*2}{21} = 396/21 = 18{,}857 \text{ (лет)}.$$

Как видим, средняя арифметическая величина может быть дробным числом, если даже индивидуальные значения признака могут принимать только целые значения. Ничего

[11] Обычно (в т.ч. и в дальнейшем в данном пособии) в статистических формулах пределы суммирования не ставятся, а подразумеваются, т.е. подразумеваются именно такие пределы как формуле (11) – с 1-ой группы по N-ю (последнюю)

необычного для метода средних в этом не заключено, так как из сущности средней не следует, что она обязана быть реальным значением признака, которое могло бы встретиться у какой-либо единицы совокупности.

Если при группировке значения осредняемого признака заданы интервалами, то при расчете средней арифметической величины в качестве значения признака в группах принимают середины этих интервалов, то есть исходят из предположения о равномерном распределении единиц совокупности по интервалу значений признака. Для открытых интервалов в первой и последней группе, если таковые есть, значения признака надо определить экспертным путем исходя из сущности, свойств признака и совокупности. При отсутствии возможности экспертной оценки значения признака в открытых интервалах, для нахождения недостающей границы открытого интервала применяют размах (разность между значениями конца и начала интервала) соседнего интервала (*принцип «соседа»*).

Например, по данным табл. 4 можно минимальную и максимальную величину веса студентов определить затруднительно, поэтому воспользуемся принципом «соседа» – применим размах соседнего интервала, который у второго и предпоследнего составляет 10 кг, значит первый интервал будет от 50 до 60 кг, а последний – от 80 до 90 кг. Середины интервалов определяем как полусумму нижней и верхней границы интервалов.

Таблица 4. Распределение студентов по весу

Группы студентов по весу, кг	Количество студентов, чел.	Середина интервала X_i'	$X_i'f_i$
До 60	6	55	330
60 – 70	8	65	520
70 – 80	5	75	375
Более 80	2	85	170
Итого	21	66,429	1395

Средний вес студентов, рассчитанный по формуле (11) с заменой точных значений признака в группах серединами интервалов, составил:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i' f_i}{\sum_{i=1}^{N} f_i} = \frac{1395}{21} = 66{,}429 \text{ кг},$$

что и записано в итоговую строку в 3-м столбце табл. 4. Следует обратить внимание, что итог объемного показателя – это сумма, а итог по столбцам относительных показателей или средних групповых величин – средняя.

Средняя арифметическая величина обладает *свойствами*, знание которых полезно как при ее использовании, так и при ее расчете.

1. Сумма отклонений индивидуальных значений признака от его среднего значения равна нулю. Доказательство[12]:

$$\sum_{i=1}^{N}(X_i - \overline{X}) = (X_1 - \overline{X}) + (X_2 - \overline{X}) + ... + (X_N - \overline{X}) = X_1 + X_2 + ...X_N - N\overline{X} = \sum_{i=1}^{N} X_i - N\frac{\sum_{i=1}^{N} X_i}{N} = 0$$

2. Если каждое индивидуальное значение признака умножить или разделить на постоянное число, то и средняя увеличится или уменьшится во столько же раз. Доказательство:

$$\frac{\sum_{i=1}^{N}(X_i : c)}{N} = \frac{\frac{X_1}{c} + \frac{X_2}{c} + ... + \frac{X_N}{c}}{N} = \frac{\frac{X_1 + X_2 + ...X_N}{c}}{N} = \frac{X_1 + X_2 + ...X_N}{N} : c = \overline{X} : c$$

[12] Для взвешенной средней сумма взвешенных отклонений равна нулю – доказать самостоятельно

Вследствие этого свойства индивидуальные значения признака можно сократить в *c* раз, произвести расчет средней и результат умножить на *c*.

3. Если к каждому индивидуальному значению признака прибавить или из каждого значения вычесть постоянное число, то средняя величина возрастет или уменьшится на это же число. Доказательство:

$$\frac{\sum_{i=1}^{N}(X_i + c)}{N} = \frac{(X_1 + c) + (X_2 + c) + ... + (X_N + c)}{N} = \frac{\sum_{i=1}^{N} X_i + Nc}{N} = \overline{X} + c$$

Это свойство полезно использовать при расчете средней величины из многозначных и слабоварьирующих значений признака аналогично предыдущему свойству.

4. Если веса средней взвешенной умножить или разделить на постоянное число, средняя величина не изменится. Доказательство:

$$\frac{\sum_{i=1}^{N} X_i \frac{f_i}{c}}{\sum_{i=1}^{N} \frac{f_i}{c}} = \frac{\left(\sum_{i=1}^{N} X_i f_i\right) : c}{\left(\sum_{i=1}^{N} f_i\right) : c} = \overline{X}$$

Используя это свойство, при расчетах следует сокращать веса на их общий сомножитель либо выражать многозначные числа весов в более крупных единицах измерениях.

5. Сумма квадратов отклонений индивидуальных значений признака от средней арифметической меньше, чем от любого другого числа. Доказательство: составим сумму квадратов отклонений от переменной *a*: $f(a) = \sum_{i=1}^{N}(X_i - a)^2$, чтобы найти экстремум этой функции, найдем ее производную по *a* и приравняем ее нулю, т.е. $\frac{\partial f}{\partial a} = 2\sum_{i=1}^{N}(X_i - a)(-1) = 0$, отсюда получаем $\sum_{i=1}^{N}(X_i - a) = 0$; $a\sum_{i=1}^{N}(1) - \sum_{i=1}^{N} X_i = 0$; $aN = \sum_{i=1}^{N} X_i$; $a = \frac{\sum_{i=1}^{N} X_i}{N} = \overline{X}$. Таким образом, экстремум суммы квадратов отклонений достигает максимума при $a = \overline{X}$. Так как логически ясно, что максимума функция иметь не может, этот экстремум является минимумом.

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменную сумму квадратов исходных величин, то средняя будет являться *квадратической средней величиной*. Ее формула следующая:

$$\overline{X}_{кв} = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N}}. \qquad (12)$$

Главной сферой применения квадратической средней в силу пятого свойства средней арифметической величины является измерение вариации признака в совокупности.

Аналогично, если по условиям задачи необходимо сохранить неизменной сумму кубов индивидуальных значений признака при их замене на среднюю величину, мы приходим к *средней кубической величине*, имеющей вид:

$$\overline{X}_{кв} = \sqrt[3]{\frac{\sum_{i=1}^{N} X_i^3}{N}}. \qquad (13)$$

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменным произведение индивидуальных величин, то следует применить геометрическую среднюю величину, имеющую следующий вид:

$$\overline{X}_{геом} = \sqrt[N]{X_1 \cdot X_2 \cdot \ldots \cdot X_N} = \sqrt[N]{\prod X}. \qquad (14)$$

Основное применение средняя геометрическая находит при определении средних относительных изменений, о чем сказано в теме 6. Геометрическая средняя величина дает наиболее точный результат осреднения, если задача стоит в нахождении такого значения признака, который качественно был бы равноудален как от максимального, так и от минимального значения признака.

Когда статистическая информация не содержит частот f по отдельным вариантам X_i совокупности, а представлена как их произведение Xf, тогда применяется формула *средней гармонической взвешенной*, для получения которой обозначим $Xf=w$, откуда $f=w/X$, и, подставив эти обозначения в формулу (11), получим формулу (15):

$$\overline{X}_{гарм} = \frac{\sum w}{\sum \frac{w}{X}} = \frac{w_1 + w_2 + \ldots + w_N}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \ldots + \frac{w_N}{x_N}}. \qquad (15)$$

Таким образом, средняя гармоническая взвешенная применяется тогда, когда неизвестны действительные веса f, а известно $w=Xf$. В тех случаях, когда вес каждого варианта $w=1$, то есть индивидуальные значения X встречаются по 1 разу, применяется формула средней гармонической простой (16):

$$\overline{X}_{гарм} = \frac{1+1+\ldots+1}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N}} = \frac{N}{\sum \frac{1}{X}}. \qquad (16)$$

Все рассмотренные выше виды средних величин принадлежат к общему типу *степенных средних*, имеющему следующий вид:

$$\overline{X} = \sqrt[m]{\frac{\sum X^m}{N}}. \qquad (17)$$

При $m = 1$ получаем среднюю арифметическую; при $m = 2$ – среднюю квадратическую; при $m = 3$ – среднюю кубическую; при $m = 0$ – среднюю геометрическую; при $m = -1$ – среднюю гармоническую. Чем выше показатель степени m, тем больше значение средней величины (если индивидуальные значения признака варьируют). В итоге, можно построить следующее соотношение, которое называется *правилом мажорантности средних*:

$$\overline{X}_{ГМ} \le \overline{X}_{геом} \le \overline{X}_{арифм} \le \overline{X}_{КВ} \le \overline{X}_{куб}. \qquad (18)$$

2.4. Контрольные задания

Вариант 1. По данным об урожайности двух фермерских хозяйств, представленным в таблице 5, рассчитать среднюю урожайность и сравнить эти хозяйства по этой урожайности.

Таблица 5. Данные об урожайности двух фермерских хозяйств

Зерновая культура	Фермерское хозяйство №1		Фермерское хозяйство №2	
	Урожайность, ц/га	Посевная площадь, га	Урожайность, ц/га	Валовый сбор, ц
Пшено	16	100	18	1400
Рожь	20	250	19	5500
Ячмень	25	300	24	8000
Просо	22	200	23	4500

Вариант 2. В 2005 году импорт России составил 98,7 млрд.долл., а экспорт – 241 млрд.долл., а в 2006 году – 137 и 302 млрд.долл. соответственно. Рассчитать всевозможные индексы, построить диаграммы и сделать выводы.

Вариант 3. По условным данным табл. 6 рассчитать среднюю экспортную цену товара, применив при этом свойства средней арифметической.

Таблица 6. Распределение цены экспортируемого товара

Цена товара, долл./т.	До 500	500 – 600	600 – 700	Более 700
Физический объем, т.	25000	28000	21000	11000

Вариант 4. По данным о реализации товара по трем коммерческим магазинам представленным в таблице 7, рассчитать среднюю цену товара.

Таблица 7. Реализация товара по трем коммерческим магазинам

Номер магазина	Цена товара, руб./кг	Выручка от реализации, руб.
1	17	49020
2	20	17400
3	22	12320

Вариант 5. По официальным данным об индексах цен на вторичном рынке жилья в РФ за 2003 – 2006 гг., представленным в таблице 8, рассчитать среднегодовые индексы цен по федеральным округам и сравнить между собой и с РФ в целом.

Таблица 8. Индексы цен на вторичном рынке жилья в 2003 – 2006 гг. (на конец года, в % к предыдущему году)

Год	2003	2004	2005	2006
Российская Федерация	118,8	124,1	118,0	154,4
по федеральным округам:				
Приволжский	113,4	124,2	120,0	157,8
Центральный	123,9	122,9	115,0	170,6
Северо-Западный	130,8	127,2	108,0	156,3
Южный	119,6	117,8	118,6	124,7
Уральский	105,3	122,3	130,6	146,3
Сибирский	111,4	133,2	123,9	134,0
Дальневосточный	121,6	119,2	121,6	124,4

Вариант 6. В 1985 году в Китае было выработано 1544 млрд.кВт-ч электроэнергии, а в США – 2650 млрд.кВт-ч. Ежегодно производство электроэнергии в среднем в Китае увеличивается на 6,9%, а в США – на 4,5%. Когда Китай и США сравняются в производстве электроэнергии?

Вариант 7. В отделе заказов торговой фирмы заняты трое работников, имеющих 8-часовой рабочий день. Первый работник на оформление одного заказа в среднем затрачивает 14 мин., второй – 15 мин., третий – 19 мин. Определить средние затраты времени на 1 заказ в целом по отделу, а также после увеличения производительности третьего работника на 25%

Вариант 8. За два месяца по цехам завода имеются данные, представленным в таблице 9. Определить изменение средней месячной заработной платы на заводе.

Таблица 9. Данные о месячной заработной плате на заводе

№ цеха	Сентябрь		Октябрь	
	Средняя месячная заработная плата, руб./чел.	Численность работников, чел.	Средняя месячная заработная плата, руб./чел.	Фонд заработной платы, тыс. руб.
1	15000	150	16000	2240
2	15500	200	16200	3645
3	15900	220	17000	4165

Вариант 9. По данным об экспорте из таблицы 10 рассчитать всевозможные индексы, построить диаграмму и сделать выводы.

Таблица 10. Товарная структура экспорта и импорта РФ

Группа товаров	Экспорт		Импорт	
	2005	2006	2005	2006
Продовольственные товары и сырье (кроме текстильного)	4,5	5,5	17,4	21,6
Минеральные продукты	156	199	3,0	3,3
Продукция химической промышленности, каучук	14,4	16,9	16,3	21,8
Кожевенное сырье, пушнина и изделия из них	0,3	0,4	0,3	0,4
Продукция лесной и целлюлозно-бумажной промышленности	8,3	9,5	3,3	4,0
Текстиль, текстильные изделия и обувь	0,9	0,9	3,6	5,5
Металлы, драгоценные камни и изделия из них	40,9	49,5	7,6	10,6
Машины, оборудование и транспортные средства	13,5	17,5	43,4	65,6
Прочие	2,5	3,1	3,7	4,9

Вариант 10. По данным об импорте из таблицы 10 рассчитать всевозможные индексы, построить диаграмму и сделать выводы.

3. Вариационные ряды распределения

3.1. Построение ряда распределения

Признаки, изучаемые статистикой, варьируются (отличаются друг от друга) у различных единиц совокупности в один и тот же период или момент времени. Например, величина внешнеторгового оборота варьируется по подразделениям ФТС; величина экспорта (импорта) варьируется по направлениям экспорта (по разным странам-партнерам по внешней торговле), по видам товаров и т.п.

Причиной *вариации* являются разные условия существования разных единиц совокупности. Например, огромное число причин влияет на масштабы внешней торговли различных стран мира.

Для управления и изучения вариации статистикой разработаны специальные методы исследования вариации, система показателей, с помощью которой вариация измеряется, характеризуются ее свойства.

Первым этапом статистического изучения вариации является построение *ряда распределения* (или *вариационного ряда*) – упорядоченного распределения единиц совокупности по возрастающим (чаще) или по убывающим (реже) значениям признака и подсчет числа единиц с тем или иным значением признака.

Существует 3 *вида* ряда распределения:

1) *ранжированный ряд* – это перечень отдельных единиц совокупности в порядке возрастания изучаемого признака (например, таблица 11); если численность единиц совокупности достаточно велика ранжированный ряд становится громоздким, и в таких случаях ряд распределения строится с помощью группировки единиц совокупности по значениям изучаемого признака (ели признак принимает небольшое число значений, то строится дискретный ряд, а в противном случае – интервальный ряд);
2) *дискретный ряд* – это таблица, состоящая из двух столбцов (строк) – конкретных значений варьирующего признака X_i и числа единиц совокупности с данным значением признака f_i – частот; число групп в дискретном ряду определяется числом реально существующих значений варьирующего признака;
3) *интервальный ряд* – это таблица, состоящая из двух столбцов (строк) – интервалов варьирующего признака X_i и числа единиц совокупности, попадающих в данный интервал (частот), или долей этого числа в общей численности совокупностей (частостей).

Построим ряд распределения внешнеторгового оборота (ВО) по таможенным постам России, для чего необходимо провести статистическое наблюдение, то есть собрать первичный статистический материал, который представляет собой величину ВО по таможенным постам. Результаты наблюдения ВО по 35 таможенным постам региона за отчетный период представим в виде ранжированного по возрастанию величины ВО ряда распределения (таблица 11).

Таблица 11. Внешнеторговый оборот (ВО) по 35 таможенным постам, млн.долл.

№ поста	*ВО*	*№ поста*	*ВО*	*№ поста*	*ВО*
1	24,16	13	54,12	25	65,31
2	27,06	14	54,91	26	69,24
3	29,12	15	55,74	27	71,39
4	31,17	16	55,91	28	77,12
5	37,08	17	56,07	29	79,12
6	39,11	18	56,80	30	84,34
7	41,58	19	56,93	31	86,89
8	44,84	20	57,07	32	91,74
9	46,80	21	58,39	33	96,01
10	48,37	22	59,61	34	106,84
11	51,44	23	59,95	35	111,16
12	52,56	24	62,05	Итого	2100,00

Определим средний размер ВО по формуле (10), приняв за X величину ВО, а за N – численность постов:

$$\overline{X} = \frac{\sum X}{N} = 2100/35 = 60 \text{ (млн.долл.)}$$

Дисперсию (о ней будет рассказано чуть позднее – на 4-м этапе анализа вариации в этой теме) определим по формуле (28):

$$\widetilde{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{N} = \frac{(24{,}16 - 60)^2 + (27{,}06 - 60)^2 + ... + (111{,}16 - 60)^2}{35} = 445{,}778 \text{ (млн.долл.}^2\text{)}$$

Построим интервальный ряд распределения ВО по таможенным постам, для чего необходимо выбрать оптимальное число групп (интервалов признака) и установить длину (размах) интервала. Поскольку при анализе ряда распределения сравнивают частоты в разных интервалах, необходимо, чтобы длина интервалов была постоянной[13]. Оптимальное число групп выбирается так, чтобы достаточной мере отразилось разнообразие значений признака в совокупности и в то же время закономерность распределении, его форма не искажалась случайными колебаниями частот. Если групп будет слишком мало, не проявится закономерность вариации; если групп будет чрезмерно много, случайные скачки частот исказят форму распределения.

Чаще всего число групп в ряду распределения определяют по формуле Стерждесса (19) или (20):

$$k = 1 + 3{,}322 \lg N \quad (19) \qquad \text{или} \qquad k = 1 + 1{,}44 \ln N\,, \quad (20)$$

где k – число групп (округляемое до ближайшего целого числа); N – численность совокупности.

Из формулы Стерджесса видно, что число групп – функция объема данных (N).
Зная число групп, рассчитывают длину (размах) интервала[14] по формуле (21):

$$h = \frac{X_{\max} - X_{\min}}{k}, \quad (21)$$

где $X_{\max}$ и $X_{\min}$ — максимальное и минимальное значения в совокупности.

В нашем примере про ВО по формуле Стерждесса (19) определим число групп:
$k = 1 + 3{,}322 lg 35 = 1 + 3{,}322 * 1{,}544 = 6{,}129 \approx 6$.
Рассчитаем длину (размах) интервала по формуле (21):
$h = (111{,}16 - 24{,}16)/6 = 87/6 = 14{,}5$ (млн.долл.).

Теперь построим интервальный ряд с 6 группами с интервалом 14,5 млн.долл. (см. первые 3 столбца табл. 12).

Таблица 12. Интервальный ряд распределения ВО по таможенным постам, млн.долл.

i	Группы постов по величине ВО X_i	Число постов f_i	Середина интервала X_i'	$X_i' f_i$	Накопл. частота f_i'	$\lvert X_i' - \overline{X} \rvert f_i$	$(X_i' - \overline{X})^2 f_i$	$(X_i' - \overline{X})^3 f_i$	$(X_i' - \overline{X})^4 f_i$
1	24,16 – 38,66	5	31,41	157,05	5	147,071	4326,001	-127246,23	3742856,97
2	38,66 – 53,16	7	45,91	321,37	12	104,400	1557,051	-23222,31	346344,16
3	53,16 – 67,66	13	60,41	785,33	25	5,386	2,231	-0,92	0,38

[13] Если приходится иметь дело с интервальным рядом распределения с неравными интервалами, то для сопоставимости нужно частоты или частости привести к единице интервала, полученное значение называется *плотностью* ρ, то есть $\rho = f/h$

[14] Единицы совокупности, имеющие значение признака, равное границе интервала, включаются в тот интервал, где это точное значение впервые указывается

i	Группы постов по величине ВО X_i	Число постов f_i	Середина интервала X_i'	$X_i'f_i$	Накопл. частота f_i'	$\lvert X_i' - \overline{X} \rvert f_i$	$(X_i' - \overline{X})^2 f_i$	$(X_i' - \overline{X})^3 f_i$	$(X_i' - \overline{X})^4 f_i$
4	67,66 – 82,16	4	74,91	299,64	29	56,343	793,629	11178,84	157461,90
5	82,16 – 96,66	4	89,41	357,64	33	114,343	3268,572	93434,47	2670891,13
6	96,66 – 111,16	2	103,91	207,82	35	86,171	3712,758	159966,81	6892284,32
	Итого	35		2128,85		513,714	13660,243	114110,66	13809838,86

Существенную помощь в анализе ряда распределения и его свойств оказывает графическое изображение. Интервальный ряд изображается столбиковой диаграммой, в которой основания столбиков, расположенные по оси абсцисс, – это интервалы значений варьирующего признака, а высоты столбиков – частоты, соответствующие масштабу по оси ординат. Графическое изображение распределения таможенных постов в выборке по величине ВО приведено на рис. 4. Диаграмма такого типа называется *гистограммой*[15].

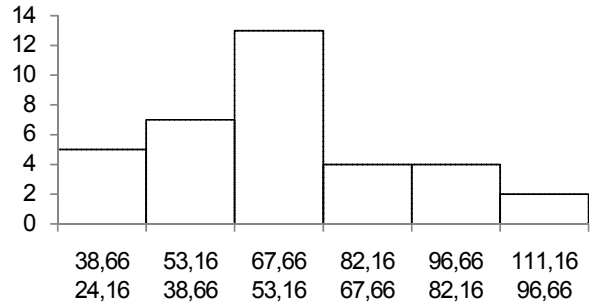


Рис. 4. Гистограмма распределения

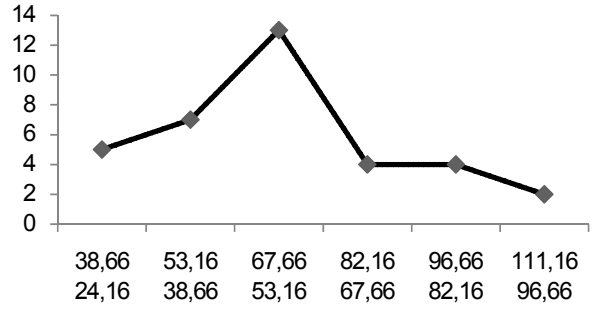


Рис. 5. Полигон распределения

Данные табл. 12 и рис. 4 показывают характерную для многих признаков форму распределения: чаще встречаются значения средних интервалов признака, реже – крайние (малые и большие) значения признака. Форма этого распределения близка к нормальному закону распределения, которое образуется, если на варьирующую переменную влияет большое число факторов, ни один из которых не имеет преобладающего значения.

Если имеется дискретный ряд распределения или используются середины интервалов (как в нашем примере про ВО – в таблице 12 в 4-м столбце рассчитаны середины интервалов как полусумма значений начала и конца интервала), то графическое изображение такого ряда называется *полигоном* (см. рис. 5)[16], которое получается соединением прямыми точек с координатами X_i и f_i.

3.2. Расчет структурных характеристик ряда распределения

При изучении вариации применяются такие характеристики ряда распределения, которые описывают количественно его структуру, строение. Такова, например, *медиана* – величина варьирующего признака, делящая совокупность на две равные части – со

[15] От греч. «гистос» – ткань, строение
[16] От греч. слов «поли» и «гонос» – многоугольник

значением признака меньше медианы и со значением признака больше медианы[17]. В нашем примере про ВО (табл. 11) медиана – это 18-й таможенный пост из 35 с величиной ВО 56,8 млн.долл. Из этого примера видно принципиальное различие между медианой и средней величиной: медиана не зависит от значений на краях ранжированного ряда. Даже если бы ВО 35-го таможенного поста был в 10 раз больше, величина медианы не изменилась бы. Поэтому медиану часто используют как более надежный показатель типичного значения признака, нежели средняя арифметическая, если ряд значений неоднороден, включает резкие отклонения от средней. В интервальном ряду распределения для нахождения медианы применяется формула:

$$Me = X_0 + h\frac{0{,}5\sum f - f'_{Me-1}}{f_{Me}}, \qquad (22)$$

где Me – медиана;
X_0 – нижняя граница интервала, в котором находится медиана;
h – величина (размах) интервала;
f'_{Me-1} – накопленная частота в интервале, предшествующем медианному;
f_{Me} – частота в медианном интервале.

В табл. 12 медианным является среднее из 35 значений, т.е. 18-е от начала значение ВО. Как видно из столбца накопленных частот (6-й столбец), оно находится в третьем интервале. Тогда по формуле (22):

$$Me = 53{,}16 + 14{,}5\frac{0{,}5*35-12}{13} = 59{,}30 \text{ (млн.долл.).}$$

Аналогично медиане вычисляются значения признака, делящие совокупность на 4 равные по численности части – *квартили*, которые обозначаются заглавной латинской буквой Q с подписным значком номера квартиля. Ясно, что Q_2 совпадает с Me. Для первого и третьего квартилей приводим формулы и расчет по данным табл. 12:

$$Q_1 = X_0 + h\frac{0{,}25\sum f - f'_{Q_1-1}}{f_{Q_1}} = 38{,}66 + 14{,}5\frac{0{,}25*35-5}{7} = 43{,}43 \text{ (млн.долл.)}$$

$$Q_3 = X_0 + h\frac{0{,}75\sum f - f'_{Q_3-1}}{f_{Q_3}} = 67{,}66 + 14{,}5\frac{0{,}75*35-25}{4} = 72{,}19 \text{ (млн.долл.)}$$

Так как $Q_2 = Me = 59{,}30$ млн.долл., видно, что различие между первым квартилем и медианой (–15,87) больше, чем между медианой и третьим квартилем (12,89). Этот факт свидетельствует о наличии некоторой несимметричности в средней области распределения, что заметно и на рис. 4.

Значения признака, делящие ряд на 5 равных частей, называются *квинтилями*, на 10 частей – *децилями*, на 100 частей – *перцентилями*. Эти характеристики применяются при необходимости подробного изучения структуры ряда распределения[18].

Безусловно, важное значение имеет такая величина признака, которая встречается в изучаемом ряду распределения чаще всего. Такую величину принято называть *модой*. В дискретном ряду мода определяется без вычисления как значение признака с наибольшей частотой. Обычно встречаются ряды с одним модальным значением признака. Если в ряду распределения встречаются 2 или несколько равных (и даже несколько различных, но больших чем соседние) значений признака, то он считается соответственно бимодальным или мультимодальным. Это свидетельствует о неоднородности совокупности, возможно, представляющей собой агрегат нескольких совокупностей с разными модами. В интервальном ряду распределения интервал с наибольшей частотой является модальным.

[17] При четном числе единиц совокупности за медиану принимают полусумму из двух центральных вариант

[18] Получите формулы и произведите их расчет (по аналогии с формулами для расчета квартилей) самостоятельно

Внутри этого интервала находят условное значение признака, вблизи которого *плотность распределения* (число единиц совокупности, приходящихся на единицу измерения варьирующего признака) достигает максимума. Это условное значение и считается *точечной модой*. Логично предположить, что такая точечная мода располагается ближе к той из границ интервала, за которой частота в соседнем интервале больше частоты в интервале за другой границей модального интервала. Отсюда получаем обычно применяемую формулу (23):

$$Mo = X_0 + h\frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})}, \qquad (23)$$

где Mo – мода;
X_0 – нижнее значение модального интервала;
f_{Mo} – частота в модальном интервале;
f_{Mo-1} – частота в предыдущем интервале;
f_{Mo+1} – частота в следующем интервале за модальным;
h – величина интервала.

По данным табл. 12 рассчитаем точечную моду по формуле (23):

$$Mo = 53{,}16 + 14{,}5\frac{13-7}{(13-7)+(13-4)} = 58{,}96 \text{ (млн.долл.)}.$$

К изучению структуры ряда распределения средняя арифметическая величина также имеет отношение, хотя основное значение этого обобщающего показателя другое. В интервальном ряду распределения ВО по таможенным постам средняя арифметическая рассчитывается как взвешенная по частоте середина интервалов X (расчет числителя – в 5-м столбце табл. **12**) по формуле (11):

$$\widetilde{X} = \frac{\sum X' f}{\sum f} = 2128{,}85/35 = 60{,}82 \text{ (млн.долл.)}.$$

Различие между средней арифметической величиной (60,82), медианой (59,30) и модой (58,96) в нашем примере невелико. Чем ближе распределение по форме к нормальному закону, тем ближе значения медианы, моды и средней величины между собой.

3.3. Расчет показателей размера и интенсивности вариации

Простейшим показателем является *размах вариации* – абсолютная разность между максимальным и минимальным значениями признака из имеющихся в изучаемой совокупности значений (24):

$$H = X_{\max} - X_{\min}. \qquad (24)$$

Поскольку величина размаха характеризует лишь максимальное различие значений признака, она не может измерять закономерную силу его вариации во всей совокупности. Предназначенный для данной цели показатель должен учитывать и обобщать все различия значений признака в совокупности без исключения. Число таких различий равно числу сочетаний по два из всех единиц совокупности (в нашем примере про ВО число сочетаний составит $C_n^k = \frac{n!}{(n-k)!k!} = \frac{35!}{(35-2)!2!} = 595$). Однако нет необходимости рассматривать, вычислять и осреднять все отклонения. Проще использовать среднюю из отклонений отдельных значений признака от среднего арифметического значения признака, а таковых в нашем примере про ВО всего 35. Но среднее отклонение значений признака от средней арифметической величины согласно первому свойству последней равно нулю. Поэтому показателем силы вариации выступает не арифметическая средняя отклонений, а средний модуль отклонений, или *среднее линейное отклонение* (25):

$$Л = \frac{\sum |X - \overline{X}|}{N}. \qquad (25)$$

В нашем примере про ВО по данным табл. 12 среднее линейное отклонение вычисляется как взвешенное по частоте отклонение по модулю середин интервалов от средней арифметической величины (расчет числителя произведен в 7-м столбце табл. 12), т.е. по формуле (26):

$$Л = \frac{\sum \left| X' - \overline{X} \right| f}{\sum f} = 513{,}714 / 35 = 14{,}678 \text{ (млн.долл.).} \qquad (26)$$

Это означает, что в среднем величина ВО в изучаемой совокупности таможенных постов отклонялась от средней величины ВО в РФ на 14,678 млн.долл.

Простота расчета и интерпретации составляют положительные стороны показателя *Л*, однако математические свойства модулей «плохие»: их нельзя поставить в соответствие с каким-либо вероятностным законом, в том числе и с нормальным распределением, параметром которого является не средний модуль отклонений, а *среднее квадратическое отклонение*, обозначаемое малой греческой буквой сигма (σ) или *s* и вычисляемое по формуле (27) – для ранжированного ряда и по формуле (28) – для интервального ряда:

$$\sigma = \sqrt{\frac{\sum (X - \overline{X})^2}{N}}; \qquad (27) \qquad \sigma = \sqrt{\frac{\sum (X' - \overline{X})^2 f}{\sum f}}. \qquad (28)$$

В нашем примере про ВО по данным табл. 12 среднее квадратическое отклонение величины ВО по формуле (28) составило (расчет числителя произведен в 8-м столбце табл. 12):

$$\sigma = \sqrt{\frac{13660{,}243}{35}} = \sqrt{390{,}293} = 19{,}756 \text{ (млн.долл.).}$$

Среднее квадратическое отклонение по величине в реальных совокупностях всегда больше среднего модуля отклонений. Разница между ними тем больше, чем больше в изучаемой совокупности резких, выделяющихся отклонений, что служит индикатором «засоренности» совокупности неоднородными с основной массой элементами. Для нормального закона распределения отношение $\sigma / Л \approx 1{,}25$. В нашем примере про ВО: $\sigma / Л \approx 19{,}756 / 14{,}678 = 1{,}35 > 1{,}25$, т.е. в изучаемой совокупности наблюдаются некоторое число таможенных постов с отличающимися от основной массы величинами ВО.

Квадрат среднего квадратического отклонения представляет собой *дисперсию* отклонений, на использовании которой основаны практически все методы математической статистики, ее формула имеет вид (29) – для несгруппированных данных (простая дисперсия) и (30) – для сгруппированных (взвешенная дисперсия):

$$\sigma^2 = \frac{\sum (X - \overline{X})^2}{n} = \overline{X^2} - \overline{X}^2; \qquad (29) \qquad \sigma^2 = \frac{\sum (X' - \overline{X})^2 f}{\sum f} = \overline{X^2} - \overline{X}^2. \qquad (30)$$

Еще одним показателем силы вариации, характеризующим ее не по всей совокупности, а лишь в ее центральной части, служит *среднее квартильное расстояние (отклонение)*, т.е. средняя величина разности между квартилями, определяемая по формуле (31):

$$q = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2} = \frac{Q_3 - Q_1}{2}. \qquad (31)$$

В нашем примере про ВО по формуле (31): $q = \frac{72{,}19 - 43{,}43}{2} = 14{,}38$ (млн.долл.).

Сила вариации в центральной части совокупности, как правило, меньше, чем в целом по всей совокупности. Соотношение между средним линейным отклонением и средним квартильным расстоянием служит для изучения структуры вариации: большое значение

такого соотношения свидетельствует о наличии слабоварьирующего «ядра» и сильно рассеянного вокруг него окружения в изучаемой совокупности. Для нашего примера про ВО соотношение *Л/q* = 1,021, что говорит о совсем незначительном различии силы вариации в центральной части совокупности и на ее периферии.

Для оценки интенсивности вариации и для сравнения ее в разных совокупностях и тем более для разных признаков необходимы *относительные показатели вариации*, которые вычисляются как отношение абсолютных показателей силы вариации, рассмотренных ранее, к средней арифметической величине признака, то есть показатели (32) – (35):

– *относительный размах вариации*: $\rho = \frac{H}{\overline{X}}$; (32)

– *линейный коэффициент вариации*: $\lambda = \frac{Л}{\overline{X}}$; (33)

– *квадратический коэффициент вариации*: $\nu = \frac{\sigma}{\overline{X}}$; (34)

– *относительное квартильное расстояние*: $d = \frac{q}{\overline{X}}$. (35)

В нашем примере про ВО эти показатели составляют:

ρ = 87/60,82 =1,43, или 143%; λ = 14,678/60,82 = 0,241, или 24,1%;

ν = 19,756/60,82 = 0,32, или 32%; d = 14,38/60,82 = 0,236, или 23,6%.

Оценка степени интенсивности вариации возможна только для каждого отдельного признака и совокупности определенного состава, она состоит в сравнении наблюдаемой вариации с некоторой обычной ее интенсивностью, принимаемой за норматив[19]. Так, для совокупности таможенных постов вариация величины ВО может быть определена как слабая, если $\nu <$ 25%, умеренная при 25% < ν < 50% и сильная при ν > 50%.

Различная сила, интенсивность вариации обусловлены объективными причинами, поэтому нельзя говорить о каком-либо универсальном критерии вариации (например, 33%), так как для разных явлений и признаков этот критерий различен[20].

3.4. Расчет моментов распределения и показателей его формы

Для дальнейшего изучения характера вариации используются средние значения разных степеней отклонений отдельных величин признака от его средней арифметической величины. Эти показатели называются *центральные моменты распределения* порядка, соответствующего степени, в которую возводятся отклонения (табл. 13) или просто моментов (нецентральные моменты в таможенной статистике практически не используются).

Таблица 13. Центральные моменты

Порядок момента	*Формула*	
	по несгруппированным данным	по сгруппированным данным
Первый μ_1	$\frac{\sum(X-\overline{X})}{N}=0$	$\frac{\sum(X'-\overline{X})f}{\sum f}=0$
Второй μ_2	$\frac{\sum(X-\overline{X})^2}{N}=\overline{X^2}-\overline{X}^2=\sigma^2$	$\frac{\sum(X'-\overline{X})^2 f}{\sum f}=\overline{X^2}-\overline{X}^2=\sigma^2$

[19] Максимально возможные значения показателей вариации: $Л_{max} = 2\overline{X} - 2\overline{X}/N$; $\sigma_{\max} = \bar{x}\sqrt{N-1}$; $\lambda_{\max} = 2 - 2/N$; $\nu_{\max} = \sqrt{N-1}$

[20] Например, цена продажи американского доллара в коммерческих банках Н.Новгорода 26 июля 2007 года варьировала от 25,45 до 26,00 при средней цене 25,595 руб., тогда по формуле (32) ρ = (26,00–25,45)/25,595 = 0,021, или 2,1%. Такая малая вариация вызвана тем, что при значительном различии курса доллара немедленно произошел бы отлив покупателей из «дорогого» банка в более «дешевые». Напротив, цена килограмма говядины в разных регионах России варьирует очень сильно – на десятки процентов и более. Это объясняется разными затратами на доставку товара из региона-производителя в регион потребитель.

Третий μ_3	$\frac{\sum(X-\overline{X})^3}{N}$	$\frac{\sum(X'-\overline{X})^3 f}{\sum f}$
Четвертый μ_4	$\frac{\sum(X-\overline{X})^4}{N}$	$\frac{\sum(X'-\overline{X})^4 f}{\sum f}$

Величина третьего момента μ^3 зависит, как и его знак, от преобладания положительных кубов отклонений над отрицательными кубами либо наоборот. При нормальном и любом другом строго симметричном распределении сумма положительных кубов строго равна сумме отрицательных кубов, поэтому на основе третьего момента строится показатель, характеризующий степень асимметричности распределения – *коэффициент асимметрии* (36):

$$As = \frac{\mu_3}{\sigma^3}. \qquad (36)$$

В нашем примере про ВО показатель асимметрии по формуле (36) составил (расчет числителя произведен в 9-м столбце табл. 12):

$As = \frac{114110,66}{19,756^3 * 35} = 0,423 > 0$, т.е. асимметрия значительна.

Английский статистик К.Пирсон на основе разности между средней арифметической величиной и модой предложил другой показатель асимметрии (37):

$$As_{П} = \frac{\overline{X} - Mo}{\sigma}. \qquad (37)$$

В нашем примере по данным табл. 12 показатель асимметрии по формуле (37) составил: $As = \frac{60,82 - 58,96}{19,756} = 0,09.$

Показатель асимметрии Пирсона (37) зависит от степени асимметричности в средней части ряда распределения, а показатель асимметрии (36) – от крайних значений признака. Таким образом, в нашем примере про ВО в средней части распределения наблюдается меньшая асимметрия, чем по краям, что видно и по графику (рис. 5). Распределения с сильной правосторонней и левосторонней асимметрией показаны на рис. 6.

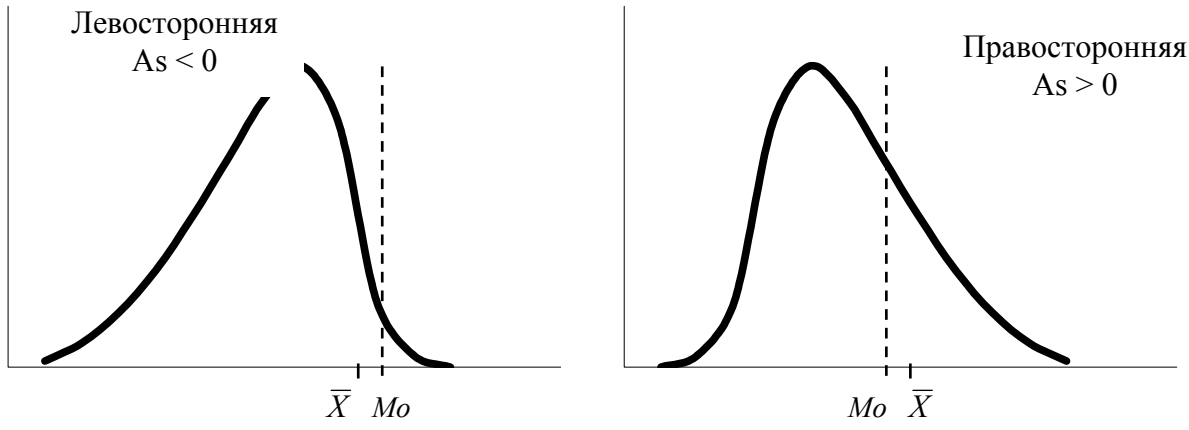


Рис. 6. Асимметрия распределения

С помощью момента четвертого порядка характеризуется еще более сложное свойство рядов распределения – *эксцесс* (от англ. «излишество»). Показатель эксцесса рассчитывается по формуле (38):

$$Ex = \frac{\mu_4}{\sigma^4} - 3. \qquad (38)$$

Чаще всего эксцесс интерпретируется как «крутизна» распределения, что не совсем верно. График распределения может выглядеть сколь угодно крутым в зависимости от силы

вариации признака: чем слабее вариация, тем круче кривая распределения при данном масштабе. Не говоря уже о том, что, изменяя масштабы по осям абсцисс и ординат, любое распределение можно искусственно сделать «крутым» и «пологим». Чтобы показать, в чем состоит эксцесс распределения, и правильно его интерпретировать, нужно сравнить ряды с одинаковой силой вариации (одной и той же величиной σ) и разными показателями эксцесса. Чтобы не смешать эксцесс с асимметрией, все сравниваемые ряды должны быть симметричными. Такое сравнение изображено на рис. 7.

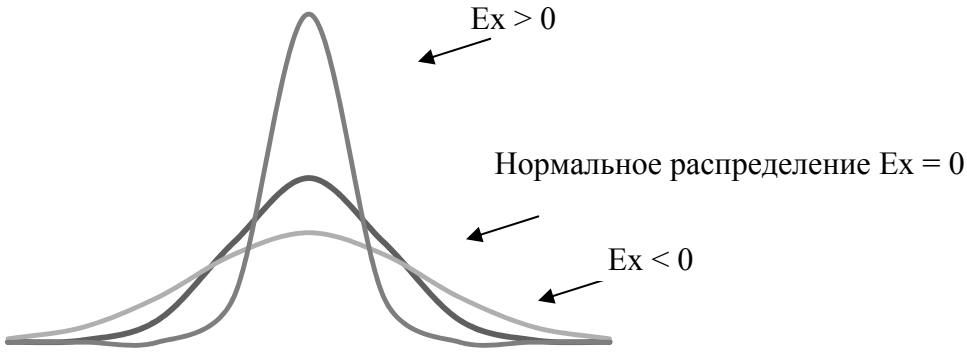


Рис. 7. Эксцесс распределения

Наличие положительного эксцесса означает наличие слабоварьирующего «ядра» и сильно рассеянного вокруг него окружения в изучаемой совокупности. Отрицательный эксцесс означает отсутствие такого «ядра».

В нашем примере по формуле (38) эксцесс составил (расчет числителя произведен в 10-м столбце табл. 12): $Ex = \frac{13809838,86}{19,756^4 * 35} - 3 = -0,41$, т.е. величина ВО по таможенным постам варьирует сильнее, чем при нормальном распределении.

По значениям показателей асимметрии и эксцесса распределения можно судить о близости распределения к нормальному: показатели асимметрии и эксцесса не должны превышать своих двукратных средних квадратических отклонений, т.е. $|As| < 2\sigma_{As}$ и $|Ex| < 2\sigma_{Ex}$. Эти средние квадратические отклонения вычисляются по формулам (39) и (40):

$$\sigma_{As} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}; \qquad (39)$$

$$\sigma_{Ex} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}. \qquad (40)$$

В нашем примере по формулам (39) и (40):

$$\sigma_{As} = \sqrt{\frac{6 * 35 * (35-1)}{(35-2)(35+1)(35+3)}} = 0,40;$$

$$\sigma_{Ex} = \sqrt{\frac{24 * 35(35-1)^2}{(35-3)(35-2)(35+3)(35+5)}} = 0,78.$$

Так как показатели асимметрии и эксцесса не превышают своих двукратных средних квадратических отклонений (As = |0,423| < 0,4*2; Ex = |–0,41| < 0,78*2), можно говорить о сходстве анализируемого распределения с нормальным.

3.5. Проверка соответствия ряда распределения нормальному

Под теоретической кривой распределения понимается графическое изображение ряда в виде непрерывной линии изменения частот в вариационном ряду, функционально связанного с изменением вариантов, другими словами, теоретическое распределение может быть выражено аналитически – формулой, которая связывает частоты и соответствующие

значения признака. Такие алгебраические формулы носят название *законов распределения*. Большое познавательное значение имеет сопоставление фактических кривых распределения с теоретическими.

Как уже неоднократно отмечалось, часто пользуются типом распределения, которое называется *нормальным*. Формула функции плотности нормального распределения имеет следующий вид (41):

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\overline{X})^2}{2\sigma^2}} \text{ или } \varphi(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}} \qquad (41)$$

где X – значение изучаемого признака;

$\overline{X}$ – средняя арифметическая ряда;

σ – среднее квадратическое отклонение;

$t = \frac{X - \overline{X}}{\sigma}$ – нормированное отклонение;

π = 3,1415 – постоянное число (отношение длины окружности к ее диаметру);

e = 2,7182 – основание натурального логарифма.

Следовательно, кривая нормального распределения может быть построена по двум параметрам – средней арифметической и среднему квадратическому отклонению. Поэтому важно выяснить, как эти параметры влияют на вид нормальной кривой.

Если $\overline{X}$ не меняется, а изменяется только σ, то чем меньше σ, тем более вытянута вверх кривая и наоборот, чем больше σ, тем более плоской и растянутой вдоль оси абсцисс становится кривая нормального распределения (см. рис. 8).

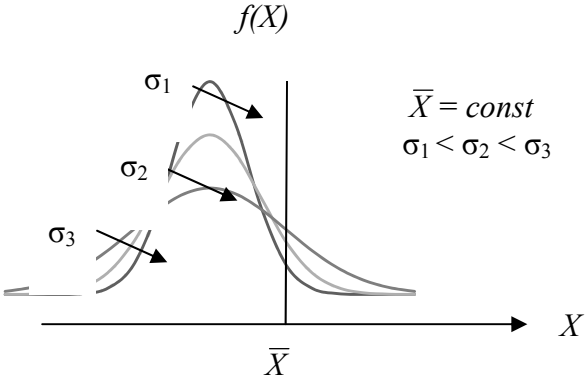


Рис. 8. Влияние величины σ на кривую нормального распределения

Если σ остается неизменной, а $\overline{X}$ изменяется, то кривые нормального распределения имеют одинаковую форму, но отличаются друг от друга положением максимальной ординаты (вершины) (см. рис. 9).

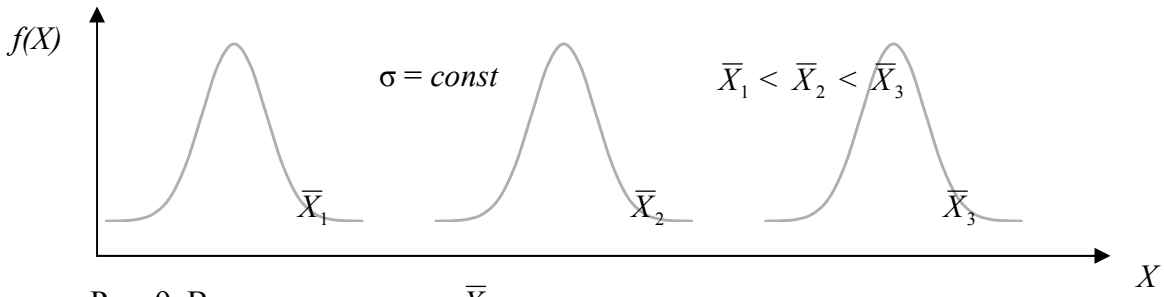


Рис. 9. Влияние величины $\overline{X}$ на кривую нормального распределения

Итак, выделим *особенности кривой нормального распределения*:

1) кривая симметрична и имеет максимум в точке, соответствующей значению $\overline{X} = Me = Mo$;

2) кривая асимптотически приближается к оси абсцисс, продолжаясь в обе стороны до бесконечности (чем больше отдельные значения X отклоняются от $\overline{X}$, тем реже они встречаются);
3) кривая имеет две точки перегиба на расстоянии ± σ от $\overline{X}$;
4) коэффициенты асимметрии и эксцесса равны нулю.

Гипотезы о распределениях заключаются в том, что выдвигается предположение о том, что распределение в изучаемой совокупности подчиняется какому-то определенному закону. Проверка гипотезы состоит в том, чтобы на основании сравнения фактических (эмпирических) частот с предполагаемыми (теоретическими) частотами сделать вывод о соответствии фактического распределения гипотетическому распределению.

Под гипотетическим распределением необязательно понимается нормальное распределение. Может быть выдвинута гипотеза о логнормальном, биномиальном распределениях, распределении Пуассона и пр.[21] Причина частого обращения к нормальному распределению состоит в том, что, как уже было замечено ранее, в этом типе распределения выражается закономерность, возникающая при взаимодействии множества случайных причин, когда ни одна из не имеет преобладающего влияния.

В нашем примере про ВО близость значений средней арифметической величины (60,82), медианы (59,30) и моды (58,96) указывает на вероятное соответствие изучаемого распределения нормальному закону.

Проверка гипотезы о соответствии теоретическому распределению предполагает расчет теоретических частот этого распределения.

Для нормального распределения порядок расчета этих частот следующий:

1) по эмпирическим данным рассчитывают среднюю арифметическую ряда $\overline{X}$ и среднее квадратическое отклонение σ;
2) находят нормированное (выраженное в σ) отклонение каждого эмпирического значения от средней арифметической:

$$t = \frac{X - \overline{X}}{\sigma}; \qquad (42)$$

3) по формуле (41) или с помощью таблиц интеграла вероятностей Лапласа находят значение $\varphi(t)$[22];
4) вычисляют теоретические частоты m по формуле:

$$m_i = N h_i \varphi(t), \qquad (43)$$

где N – объем совокупности, h_i – длина (размах) i-го интервала.

Определим теоретические частоты нормального распределения в нашем примере про ВО по данным табл. 12, для чего построим вспомогательную таблицу 14. Средняя арифметическая величина и среднее квадратическое отклонение нами уже найдены ранее ($\widetilde{X} = 60{,}82;\ \sigma = 19{,}756$); значения нормированных отклонений t рассчитаны в 5-м столбце таблицы 14, а значения плотностей $\varphi(t)$ – в 8-м столбце (в 6-м и 7-м столбцах приведены промежуточные расчеты по формуле (41)); в последнем столбце – теоретические частоты нормального распределения.

Таблица 14. Расчет теоретических частот нормального распределения

[21] Прочие виды распределений изучаются дисциплиной «Теория вероятностей»

[22] Простой расчет возможен при наличии *Excel* из пакета *Microsoft Office*, где имеется функция, вычисляющая плотность (или интеграл) функции нормального распределения =НОРМРАСП(А;Б;В;Г), где параметры: А – значение X; Б – средняя арифметическая $\overline{X}$; В – среднее квадратическое отклонение σ; Г – «0» для вычисления плотности (или «1» для вычисления интеграла) распределения

i	X_i	f_i	X_i'	$t=\frac{X_i-\widetilde{X}}{\sigma}$	$-\frac{t^2}{2}$	$e^{-\frac{t^2}{2}}$	$\varphi(t)$	m_i
1	24,16 – 38,66	5	31,41	-1,4889	-1,1084	0,3301	0,0067	3,383
2	38,66 – 53,16	7	45,91	-0,7549	-0,2850	0,7520	0,0152	7,707
3	53,16 – 67,66	13	60,41	-0,0210	-0,0002	0,9998	0,0202	10,246
4	67,66 – 82,16	4	74,91	0,7130	-0,2542	0,7756	0,0157	7,948
5	82,16 – 96,66	4	89,41	1,4470	-1,0468	0,3510	0,0071	3,598
6	96,66 – 111,16	2	103,91	2,1809	-2,3782	0,0927	0,0019	0,950
	Итого	35						33,832

Сравним на графике эмпирические f (ВО по таможенным постам) и теоретические m (нормальное распределение) частоты, полученные на основе данных табл. 14 (рис. 10). Близость этих частот очевидна[23], но объективная оценка их соответствия может быть получена только с помощью критериев согласия.

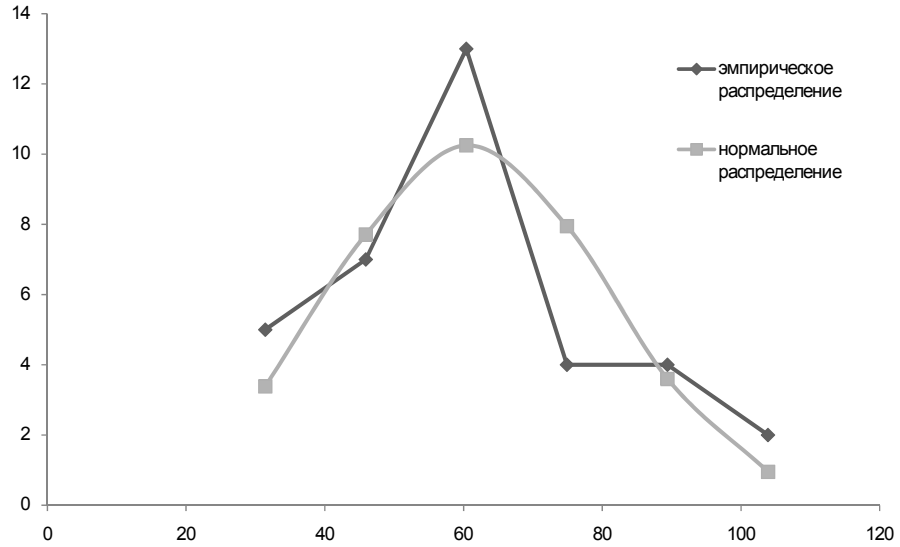


Рис. 10. Распределение ВО по таможенным постам (эмпирическое) и нормальное

Критерии согласия, опираясь на установленный закон распределения, дают возможность установить, когда расхождения между теоретическими и эмпирическими частотами следует признать несущественными (случайными), а когда – существенными (неслучайными). Таким образом, критерии согласия позволяют отвергнуть или подтвердить правильность выдвинутой гипотезы о характере распределения в эмпирическом ряду и дать ответ, можно ли принять для данного эмпирического распределения модель, выраженную некоторым теоретическим законом распределения.

Существует ряд критериев согласия, но чаще всего применяют критерии Пирсона χ^2, Колмогорова и Романовского.

Критерий согласия Пирсона χ^2 (хи-квадрат) – один из основных критериев согласия, рассчитываемый по формуле (44):

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-m_i)^2}{m_i}, \qquad (44)$$

где k – число интервалов;
f_i – эмпирическая частота i-го интервала;
m_i – теоретическая частота.

[23] Иногда за счет округлений при расчетах (использование функции плотности распределения вместо интеграла) может быть нарушено равенство сумм эмпирических и теоретических частот, что и произошло в нашем примере про ВО ($\sum f$=35, $\sum m$=33,832)

Для распределения χ^2 составлены таблицы, где указано критическое значение критерия согласия χ^2 для выбранного уровня значимости α и данного числа степеней свободы ν (см. Приложение 3).

Уровень значимости α – это вероятность ошибочного отклонения выдвинутой гипотезы, т.е. вероятность (P) того, что будет отвергнута правильная гипотеза. В статистических исследованиях в зависимости от важности и ответственности решаемых задач пользуются следующими тремя уровнями значимости:

1) $\alpha = 0{,}10$, тогда P = 0,90;
2) $\alpha = 0{,}05$, тогда P = 0,95 [24];
3) $\alpha = 0{,}01$, тогда P = 0,99.

Число степеней свободы v определяется по формуле:

$$v = k - z - 1, \tag{45}$$

где k – число интервалов;

z – число параметров, задающих теоретический закон распределения.

Для нормального распределения $z = 2$, так как нормальное распределение зависит от двух параметров – средней арифметической ($\overline{X}$) и среднего квадратического отклонения (σ).

Для оценки существенности расхождений расчетное значение χ^2 сравнивают с табличным $\chi^2_{табл}$. Расчетное значения критерия должно быть *меньше* табличного, т.е. $\chi^2 < \chi^2_{табл}$, в противном случае расхождения между теоретическим и эмпирическим распределением не случайны, а теоретическое распределение не может служить моделью для изучаемого эмпирического распределения.

Использование критерия χ^2 рекомендуется для достаточно больших совокупностей (N>50), при этом частота каждой группы не должна быть менее 5, в противном случае повышается вероятность получения ошибочных выводов.

В нашем примере про ВО для расчета критерия χ^2 построим вспомогательную таблицу 15.

Таблица 15. Вспомогательные расчеты критериев согласия

i	X_i	f_i	m_i	$\frac{(f_i - m_i)^2}{m_i}$	f_i'	m_i'	$\lvert f_i' - m_i' \rvert$
1	24,16 – 38,66	5	3,383	0,773	5	3,383	1,617
2	38,66 – 53,16	7	7,707	0,065	12	11,090	0,910
3	53,16 – 67,66	13	10,246	0,740	25	21,336	3,664
4	67,66 – 82,16	4	7,948	1,961	29	29,284	0,284
5	82,16 – 96,66	4	3,598	0,045	33	32,882	0,118
6	96,66 – 111,16	2	0,950	1,160	35	33,832	1,168
	Итого	35	33,832	4,744			

Теперь по формуле (44): χ^2 =4,744, что меньше табличного (Приложение 3) значения $\chi^2_{табл}$=7,8147 при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы v=6–2–1=3, значит с вероятностью 0,95 можно говорить, что в основе эмпирического распределения величины ВО по таможенным постам лежит закон нормального распределения, т.е. выдвинутая гипотеза не отвергается, а расхождения объясняются случайными факторами.

Критерий Романовского K_P основан на использовании критерия Пирсона χ^2, т.е. уже найденных значений χ^2 и числа степеней свободы v, рассчитывается по формуле (46):

$$K_P = \frac{\left|\chi^2 - v\right|}{\sqrt{2v}}. \tag{46}$$

Он используется в том случае, когда отсутствует таблица значений χ^2. Если $K_P < 3$, то расхождения между теоретическим и эмпирическим распределением случайны, если $K_P > 3$,

[24] Практически приемлемая вероятность в экономических исследованиях, означающая, что в 5 случаях из 100 может быть отвергнута правильная гипотеза

то не случайны, и теоретическое распределение не может служить моделью для изучаемого эмпирического распределения.

В нашем примере про ВО по формуле (46): $K_P = \frac{|4{,}744 - 3|}{\sqrt{2*3}} = 0{,}712 < 3$, что подтверждает несущественность расхождений между эмпирическими и теоретическими частотами.

Критерий Колмогорова λ основан на определении максимального расхождения между накопленными частотами эмпирического и теоретического распределений (D), рассчитывается по формуле (47) [25]:

$$\lambda = D / \sqrt{N} . \qquad (47)$$

Рассчитав значение λ, по таблице $P(\lambda)$ (см. Приложение 6) определяют вероятность, с которой можно утверждать, что отклонения эмпирических частот от теоретических случайны. Вероятность $P(\lambda)$ может изменяться от 0 до 1. При $P(\lambda) = 1$ (т.е. при $\lambda < 0{,}3$) происходит полное совпадение частот, при $P(\lambda) = 0$ – полное расхождение.

В нашем примере про ВО в последних трех столбцах таблицы 15 приведены расчеты накопленных частот и разностей между ними, откуда видно, что в 3-ей группе наблюдается максимальное расхождение (разность) $D = 3{,}664$. Тогда по формуле (47): $\lambda = 3{,}664 / \sqrt{35} = 0{,}619$. По таблице Приложения 6 находим значение вероятности при $\lambda = 0{,}6$: $P = 0{,}86$ (наиболее близкое значение к 0,619), т.е. с вероятностью, близкой к 0,86, можно говорить, что в основе эмпирического распределения величины ВО по таможенным постам лежит закон нормального распределения, а расхождения эмпирического и теоретического распределений носят случайный характер.

Итак, подтвердив правильность выдвинутой гипотезы с помощью известных критериев согласия, можно использовать результаты распределения для практической деятельности. Какое же практическое значение может иметь произведенная проверка гипотезы? Во-первых, соответствие нормальному закону позволяет прогнозировать, какое число таможенных постов (или их доля) попадет в тот или иной интервал значений величины ВО. Во-вторых, нормальное распределение возникает при действии на вариацию изучаемого показателя множества независимых факторов. Из чего следует, что нельзя существенно снизить вариацию величины ВО, воздействуя только на один-два управляемых фактора, скажем число работников таможенного поста или степень технической оснащенности.

3.6. Проверка соответствия ряда распределения закону Пуассона

Таможенная инспекция провела проверку после выпуска товаров. В результате получен следующий дискретный ряд распределения числа нарушений, выявленных в каждой проверке (табл. 16).

Таблица 16. Ряд распределения числа нарушений, выявленных таможенной инспекцией

Число нарушений	0	1	2	3
Число проверок	24	4	2	1

Проведем анализ этого ряда распределения. Сначала рассчитаем среднее число нарушений в выборке, а также его дисперсию, для чего построим вспомогательную таблицу 17.

Таблица 17. Ряд распределения числа нарушений, выявленных таможенной инспекцией

Число нарушений X	Число проверок f	Xf	$(X - \bar{X})^2 f$	m	$\frac{(f-m)^2}{m}$	f'	m'	$\lvert f' - m' \rvert$

[25] Основное условие для использования критерия Колмогорова – достаточно большое число наблюдений ($N > 50$)

0	24	0	3,022	21,7	0,244	24	21,7	2,3
1	4	4	1,665	7,7	1,778	28	29,4	1,4
2	2	4	5,413	1,4	0,257	30	30,8	0,8
3	1	3	6,997	0,2	3,200	31	31	0
Итого	31	11	17,097	31	5,479			

Среднее число нарушений в выборке по формуле (11): $\overline{X} = 11/31 = 0{,}355$ (нарушений).

Дисперсию определим по формуле (28): $\sigma^2 = \dfrac{\sum (X_i - \overline{X})^2}{N} = \dfrac{17{,}097}{31} = 0{,}552$ (нарушений^2).

Построив график этого распределения (полигон) – рис. 11, видно, что данное распределение не похоже на нормальное.

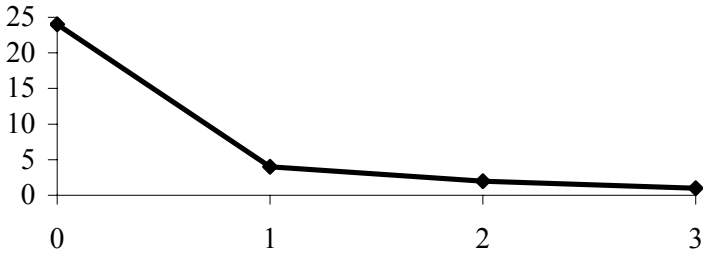


Рис. 11. Кривая распределения числа нарушений, выявленных таможенной инспекцией

Из структурных характеристик ряда распределения можно определить только моду: *Mo* = 0, так как по данным табл. 17 такое число нарушений чаще всего встречается (*f*=24).

По формуле (24) определим размах вариации: $H = 3 - 0 = 3$, что характеризует вариацию в 3 нарушения.

По формуле (26) найдем среднее линейное отклонение:

$$Л = \frac{\sum |X - \overline{X}| f}{\sum f} = \frac{|0 - 0{,}355|24 + |1 - 0{,}355|4 + |2 - 0{,}355|2 + |3 - 0{,}355|1}{31} = 17{,}035 / 31 = 0{,}550 .$$

Это означает, что в среднем число нарушений отклоняется от среднего их числа на 0,55.

Среднее квадратическое отклонение рассчитаем не по формуле (28), а как корень из дисперсии, которая уже была рассчитана нами выше: $\sigma = \sqrt{0{,}552} = 0{,}743$, тогда $\sigma / Л \approx 0{,}743 / 0{,}550 = 1{,}35 > 1{,}25$, т.е. в изучаемом распределении наблюдается некоторое число выделяющихся нарушений (с большим числом нарушений, выявленных в одной проверке).

Поскольку квартили на предыдущем этапе не определялись, на данном этапе расчет среднего квартильного расстояния пропускаем.

Теперь рассчитаем *относительные показатели вариации*:

- относительный размах вариации по формуле (32): $\rho = 3/0{,}355 = 8{,}45$;
- линейный коэффициент вариации по формуле (33): $\lambda = 0{,}550/0{,}355 = 1{,}55$;
- квадратический коэффициент вариации по формуле (34): $\nu = 0{,}743/0{,}355 = 2{,}09$.

Все расчеты на данном этапе свидетельствуют о значительных размере и интенсивности вариации нарушений, выявленных таможенной инспекцией.

Не имеет практического смысла расчет моментов распределения, так как видно из рис. 11, что в изучаемом распределении симметрия отсутствует вовсе, поэтому и расчет эксцесса также бесполезен.

Выдвинем гипотезу о соответствии изучаемого распределения распределению Пуассона[26], которое описывается формулой (48):

$$P(x) = \frac{a^X e^{-a}}{X!}, \qquad (48)$$

где $P(X)$ – вероятность того, что признак примет то или иное значение X;

$e = 2{,}7182$ – основание натурального логарифма;

$X!$ – факториал числа X (т.е. произведение всех целых чисел от 1 до X включительно);

$a = \overline{X}$ – средняя арифметическая ряда распределения.

Из формулы (48) видно, что единственным параметром распределения Пуассона является средняя арифметическая величина. Порядок определения теоретических частот этого распределения следующий:

1) рассчитать среднюю арифметическую ряда, т.е. = a;
2) рассчитать e^{-a};
3) для каждого значения X рассчитать теоретическую частоту по формуле (49):

$$m = N\frac{a^x e^{-a}}{x!} = N * P(X). \qquad (49)$$

Поскольку $a = \overline{X} = 0{,}355$ найдем значение $e^{-0{,}355} = 0{,}7012$. Затем, подставив в формулу (49) значения X от 0 до 3, вычислим теоретические частоты:

$$m_0 = 31\frac{0{,}355^0 * 0{,}7012}{0!} = 21{,}7 \text{ (т.к. 0! = 1);} \qquad m_1 = 31\frac{0{,}355^1 * 0{,}7012}{1!} = 7{,}7;$$

$$m_2 = 31\frac{0{,}355^2 * 0{,}7012}{2!} = 1{,}4; \qquad m_3 = 31\frac{0{,}355^3 * 0{,}7012}{3!} = 0{,}2.$$

Полученные теоретические частоты занесем в 5-й столбец табл. 17 и построим график эмпирического и теоретического распределений (рис. 12), из которого видна близость эмпирического и теоретического распределений.

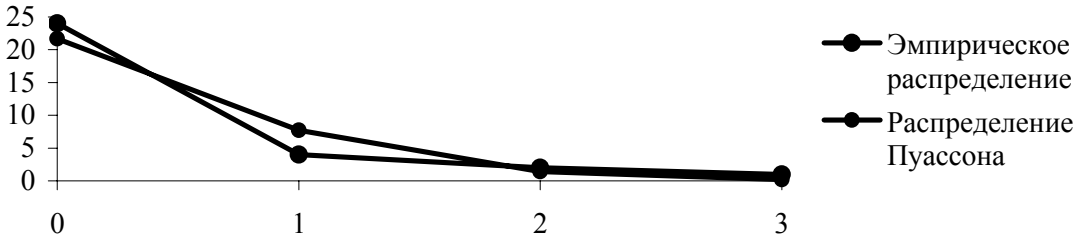


Рис. 12. Эмпирическая и теоретическая (распределение Пуассона) кривые распределения

Проверим выдвинутую гипотезу о соответствии изучаемого распределения закону Пуассона с помощью критериев согласия.

Рассчитаем значение критерия Пирсона χ^2 по формуле (44) в 6-м столбце табл. 17: $\chi^2 = 5{,}479$, что меньше табличного (Приложение 3) значения $\chi^2_{табл} = 5{,}9915$ при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $v = 4–1–1 = 2$, значит с вероятностью 0,95 можно говорить, что в основе эмпирического распределения лежит закон распределения Пуассона, т.е. выдвинутая гипотеза не отвергается, а расхождения объясняются случайными факторами.

Определим значение критерия Романовского по формуле (46):

---

[26] Названо по имени французского математика Симеона Пуассона (1781 – 1840), еще называют законом распределения редких явлений; возникает, когда значения признака выражены дискретно и являются результатом какого-либо редко возникающего события среди наблюдаемых единиц, причем с увеличением значений признака вероятность наступления события падает

$K_P = \frac{|5{,}479 - 2|}{\sqrt{2 * 2}} =$ 1,74 < 3, что подтверждает несущественность расхождений между эмпирическими и теоретическими частотами.

Для расчета критерия Колмогорова в последних трех столбцах таблицы 17 приведены расчеты накопленных частот и разностей между ними, откуда видно, что в 1-ой группе наблюдается максимальное расхождение (разность) D = 2,3. Тогда по формуле (47): $\lambda = 2{,}3 / \sqrt{31} = 0{,}413$. По таблице Приложения 6 находим значение вероятности при λ = 0,4: P = 0,9972 (наиболее близкое значение к 0,413), т.е. с вероятностью, близкой к единице, можно говорить, что в основе эмпирического распределения величины нарушений, выявленных таможенной инспекцией, лежит закон распределения Пуассона, а расхождения эмпирического и теоретического распределений носят случайный характер.

3.7. Контрольные задания

На основе условных ранжированных данных таблицы 18 провести анализ вариации величины налоговых сборов (тыс. руб.) с предприятий района, собранных налоговыми органами.

Таблица 18. Распределение вариантов для выполнения контрольного задания

№ п/п	Вариант									
	1	2	3	4	5	6	7	8	9	10
1	107	109	118	155	104	101	142	123	128	158
2	139	111	165	178	107	163	143	124	180	177
3	142	199	168	182	113	200	169	184	208	292
4	144	226	247	223	133	230	169	247	247	317
5	150	239	249	227	186	308	223	295	259	327
6	207	289	293	269	186	314	233	303	262	380
7	207	318	299	272	195	320	236	312	325	433
8	217	319	302	286	230	328	290	332	341	449
9	233	346	339	294	232	367	292	335	344	458
10	244	390	361	301	243	405	292	351	353	490
11	271	390	364	306	264	410	338	378	362	505
12	273	405	405	361	356	420	359	379	366	506
13	275	428	410	362	368	427	363	388	377	526
14	300	436	429	392	372	440	367	389	387	553
15	302	438	439	428	387	458	368	393	389	567
16	305	450	458	454	403	464	411	420	429	586
17	312	451	462	462	467	465	436	422	466	604
18	320	496	492	466	482	482	449	425	485	618
19	359	497	498	482	491	495	460	461	491	624
20	369	502	543	487	494	497	480	465	515	627
21	370	513	550	490	510	545	488	495	523	633
22	372	517	566	493	511	549	493	498	534	653
23	382	531	581	501	512	582	500	526	546	656
24	411	545	588	508	533	590	500	528	550	657
25	414	558	590	511	540	602	513	531	573	673

№ п/п	Вариант									
	1	2	3	4	5	6	7	8	9	10
26	416	560	593	519	576	603	515	531	574	677
27	426	571	609	533	577	614	523	544	604	689
28	428	573	610	539	579	621	526	563	618	702
29	436	580	612	550	579	633	533	576	624	709
30	451	593	622	555	589	643	553	584	653	723
31	496	597	658	555	590	664	559	585	657	734
32	497	615	680	561	591	666	560	597	673	752
33	513	649	706	597	598	676	564	602	685	755
34	517	661	716	600	604	691	580	604	701	756
35	545	668	726	621	630	692	585	631	702	779
36	558	680	737	643	687	708	592	639	706	785
37	571	693	751	674	703	717	595	647	723	802
38	580	801	795	676	705	726	604	665	734	819
39	593	813	812	683	729	743	653	671	755	822
40	597	816	825	689	738	744	671	699	756	829
41	615	825	849	712	740	753	676	716	785	842
42	649	675	855	735	776	758	698	719	802	848
43	661	842	858	766	786	772	700	720	842	864
44	680	845	861	799	792	793	717	764	864	886
45	801	650	865	818	825	808	761	803	886	888
46	816	858	866	824	851	861	808	873	888	926
47	825	878	867	858	854	867	818	879	926	930
48	845	958	938	861	895	880	838	898	930	945
49	961	972	939	898	896	897	869	922	945	951
50	972	994	989	937	949	929	888	991	961	961

4. Статистическое изучение структуры совокупности

4.1. Абсолютные и относительные показатели изменения структуры

Развитие статистической совокупности проявляется не только в количественном росте или уменьшении элементов системы, но также и в изменении ее структуры. *Структура* – это строение совокупности, состоящее из отдельных элементов и связей между ними. Например, экспорт страны (совокупность) состоит из различных видов товаров (элементов), стоимость которых различается по видам и по странам. Кроме того, происходит постоянное изменение структуры экспорта в динамике. Соответственно возникает задача изучения структуры совокупностей и их динамики, для чего разработаны специальные методы, которые будут рассмотрены далее.

В теме 2 был рассмотрен индекс структуры, рассчитываемый по формуле (6), который характеризует долю отдельных элементов в итоге абсолютного признака совокупности. В теме 3 рассмотрена система показателей и методика анализа распределения совокупности по значениям какого-либо отдельного признака (вариационный ряд распределения). Здесь излагаются показатели, характеризующие изменение структуры в целом, т.е. «структурный сдвиг» [27]. Практическое применение этих показателей рассмотрим на двух примерах, представленных в таблицах 19 и 20 (первые 4 столбца, выделенные полужирным шрифтом, – исходные данные, а остальные – вспомогательные расчеты).

Таблица 19. Распределение населения России по величине среднедушевых денежных доходов (СДД)

№ группы (*j*)	**СДД, руб./чел. в месяц**	**Доли населения**		$\lvert d_1 - d_0 \rvert$	d_0^2	d_1^2	$(d_1 - d_0)^2$	$(d_1 + d_0)^2$	$\left(\frac{d_{1j} - d_{0j}}{d_{1j} + d_{0j}}\right)^2$
		2005 год (d_0)	**2006 год (d_1)**						
1	**до 1500**	**0,032**	**0,018**	0,014	0,0010	0,0003	0,0002	0,0025	0,0784
2	**1500-2500**	**0,088**	**0,058**	0,030	0,0077	0,0034	0,0009	0,0213	0,0422
3	**2500-3500**	**0,113**	**0,085**	0,028	0,0128	0,0072	0,0008	0,0392	0,0200
4	**3500-4500**	**0,114**	**0,094**	0,020	0,0130	0,0088	0,0004	0,0433	0,0092
5	**4500-6000**	**0,149**	**0,135**	0,014	0,0222	0,0182	0,0002	0,0807	0,0024
6	**6000-8000**	**0,149**	**0,149**	0,000	0,0222	0,0222	0,0000	0,0888	0,0000
7	**8000-12000**	**0,174**	**0,197**	0,023	0,0303	0,0388	0,0005	0,1376	0,0038
8	**более 12000**	**0,181**	**0,264**	0,083	0,0328	0,0697	0,0069	0,1980	0,0348
	Итого	**1,000**	**1,000**	0,212	0,1420	0,1687	0,0099	0,6114	0,1909

Таблица 20. Распределение численности безработных России по уровню образования в 2006 г.

№ группы (*j*)	**Имеют образование**	**Мужчины (d_0)**	**Женщины (d_1)**	$\lvert d_1 - d_0 \rvert$	d_0^2	d_1^2	$(d_1 - d_0)^2$	$(d_1 + d_0)^2$	$\left(\frac{d_{1j} - d_{0j}}{d_{1j} + d_{0j}}\right)^2$
1	**Высшее профессиональное**	**0,087**	**0,130**	0,043	0,0076	0,0169	0,0018	0,0471	0,0393
2	**Неполное высшее профессиональное**	**0,019**	**0,023**	0,004	0,0004	0,0005	0,0000	0,0018	0,0091
3	**Среднее профессиональное**	**0,130**	**0,221**	0,091	0,0169	0,0488	0,0083	0,1232	0,0672

[27] Важно не путать понятие «структурный сдвиг», оцениваемый в теме 8, где он представляет не величину самого изменения структуры, а его влияние на результативный показатель

4	**Начальное профессиональное**	**0,200**	**0,149**	0,051	0,0400	0,0222	0,0026	0,1218	0,0214
5	**Среднее (полное) общее**	**0,398**	**0,338**	0,060	0,1584	0,1142	0,0036	0,5417	0,0066
6	**Основное общее**	**0,148**	**0,121**	0,027	0,0219	0,0146	0,0007	0,0724	0,0101
7	**Начальное общее, не имеют образ-я**	**0,018**	**0,018**	0,000	0,0003	0,0003	0,0000	0,0013	0,0000
	Итого	**1,000**	**1,000**	0,276	0,2455	0,2177	0,0171	0,9092	0,1536

Обобщающим абсолютным показателем изменения структуры может служить *сумма модулей абсолютных изменений долей*, определяемая по формуле (50):

$$\Delta d = \sum_{j=1}^{k} \left| d_{1j} - d_{0j} \right|, \qquad (50)$$

где d_{1j} – доля j-ой группы элементов в отчетном периоде; d_{0j} – доля j-ой группы элементов в базисном периоде.

По данным таблицы 19 в 5-м столбце произведен расчет по формуле (50): Δd =0,212, то есть суммарное изменение долей в распределении россиян по доходам составило 21,2%. Аналогично по той же формуле по данным таблицы 20: Δd =0,276, то есть различие структуры безработных среди женщин и мужчин по уровню образованию составляет 27,6%.

Расчет среднего абсолютного изменения, приходящегося на одну долю (группу, элемент совокупности) не дает никакой дополнительной информации. Зато можно определить, насколько сильно произошедшее изменение структуры в сравнении с предельно возможной величиной суммы модулей, которая равна 2. Для этого используется показатель *степени интенсивности абсолютного сдвига* (или *индекс Лузмора-Хэнби*), который определяется по формуле . (51):

$$K_{\Delta d} = 0{,}5 \sum_{j=1}^{k} \left| d_{1j} - d_{0j} \right|. \qquad (51)$$

По данным таблицы 19 по формуле . (51): $K_{\Delta d}$ =0,106, то есть интенсивность изменения долей в распределении россиян по доходам составила 10,6% от максимально возможного. Аналогично по той же формуле по данным таблицы 20: $K_{\Delta d}$ =0,138, то есть различие структуры безработных среди женщин и мужчин по уровню образованию составляет 13,8% от максимально возможного.

Обобщенная оценка степени структуризации явления в целом обычно выполняется по формуле уровня концентрации (или *коэффициент Герфиндаля*), который более чувствителен к изменению долей групп с наибольшим удельным весом в итоге, определяемый по формуле (52):

$$H = \sum_{j=1}^{k} d_j^2 \qquad (52)$$

где d_i – доля i-го объекта в общем итоге изучаемого показателя; k – количество объектов.

По данным таблицы 19 в 6-м и 7-м столбцах произведен расчет коэффициента Герфиндаля по формуле (52): H_{2005}=0,142 и H_{2006}=0,1687, то есть уровень концентрации в распределении россиян по доходам увеличился в 2006 году по сравнению с 2005 годом. Аналогично по той же формуле по данным таблицы 20: $H_{муж}$=0,2455 и $H_{жен}$=0,2177, то есть уровень концентрации в распределении безработных по уровню образованию среди мужчин выше, чем среди женщин (влияние уровня образования на статус безработного среди мужчин выше, чем среди женщин).

Обратная индексу Герфиндаля величина – это *эффективное число групп* в структуре, которое показывает количество групп без учета групп, имеющих ничтожно малые доли, определяется по формуле (53):

$$E = 1/H. \quad (53)$$

По данным таблицы 19 эффективное число групп по формуле (53): E_{2005}=1/0,142=7,0 и E_{2006}=5,9, то есть эффективное число групп в распределении россиян по доходам уменьшилось с 7 в 2005 году до 6 в 2005 году, что свидетельствует о необходимости пересмотра интервалов распределения россиян по доходам в будущем году. Аналогично по той же формуле по данным таблицы 20: $E_{муж}$=1/0,2455=4,07 и $E_{жен}$=1/0,2177=4,59, то эффективное число групп в распределении безработных по уровню образованию среди мужчин выше и среди женщин – 4 у мужчин и 5 у женщин.

Еще один вариант оценки степени структуризации явления в целом – *индекс Грофмана* . (54), который представляет собой сумму модулей абсолютных изменений долей, приходящихся на одну эффективную группу:

$$I_{Grofman} = \frac{\Delta d}{E_0} = H_0 \Delta d . \quad (54)$$

По данным таблицы 19 в по формуле . (54): $I_{Grofman}$=0,212*0,142=0,030, то есть изменение долей, приходящихся на одну эффективную группу в распределении россиян по доходам незначительно (3,0%). Аналогично по той же формуле по данным таблицы 20: $I_{Grofman}$=0,2455*0,276=0,068, то есть различие структуры в расчете на одну эффектиную группу среди безработных женщин и мужчин по уровню образованию слабое (6,8%).

Для оценки изменений двух наибольших долей (доминантные доли) применяется *индекс Липхарта* . (55):

$$I_{Lijphart} = 0{,}5\sum_{m=1}^{2} |d_{1m} - d_{0m}| . \quad (55)$$

где d_{1m} и d_{0m} – доля m-ой группы элементов в отчетном периоде и базисном периодах; m – максимальная доля в совокупности.

По данным таблицы 19 по формуле . (55): $I_{Lijphart}$=0,5*(0,083+0,023)=0,053, то есть среднее изменение долей в двух доминантных группах распределения россиян по доходам составило 5,3%. Аналогично по той же формуле по данным таблицы 20: $I_{Lijphart}$=0,5*(0,060+0,051)=0,056, то есть различие структуры в двух доминантных группах среди безработных женщин и мужчин по уровню образованию составляет 5,6%.

Рассмотренные показатели основаны на средней арифметической в различных вариантах, и из-за их линейности по отклонениям они одинаково учитывают большие и малые отклонения. *Квадратические индексы* позволяют сравнивать различные структуры, неразличимые с точки зрения суммы изменений.

Квадратический индекс структурных сдвигов *Казинца* . (56):

$$I_{Kazinets} = \sqrt{\frac{\sum \left(d_{1j} - d_{0j}\right)^2}{k}} . \quad (56)$$

По данным таблицы 19 по формуле . (56): $I_{Kazinets} = \sqrt{0{,}0099/8}$ =0,035, то есть среднее измененение долей в группе в распределении россиян по доходам составило 3,5% (незначительно). Аналогично по той же формуле по данным таблицы 20: $I_{Kazinets} = \sqrt{0{,}0171/7}$ =0,049, то есть различие в группах в структуре безработных среди женщин и мужчин по уровню образованию составляет 4,9% (несущественно).

Аналогичен индексу Казинца *индекс наименьших квдратов* (или *индекс Галлахера*), при расчете которого, в отличие от формулы . (51), малые разности долей слабее влияют на индекс, чем большие, определяется по формуле . (57)[28]:

$$I_{LSQ} = \sqrt{0{,}5\sum_{j=1}^{k}\left(d_{1j} - d_{0j}\right)^2} \ . \qquad (57)$$

По данным таблицы 19 по формуле . (57): $I_{LSQ} = \sqrt{0{,}5 * 0{,}0099} = 0{,}070$, то есть интенсивность изменения долей в распределении россиян по доходам составила 7,0%. Аналогично по той же формуле по данным таблицы 20: $I_{LSQ} = \sqrt{0{,}5 * 0{,}0171} = 0{,}092$, то есть различие структуры безработных среди женщин и мужчин по уровню образованию составляет 9,2%.

Незначительную модификацию индекса наименьших квадратов представляет *индекс Монро* . (58):

$$I_{Monroe} = \sqrt{\frac{\sum_{j=1}^{k}\left(d_{1j} - d_{0j}\right)^2}{1 + H_0}} \ . \qquad (58)$$

По данным таблицы 19 по формуле . (58): $I_{Monroe} = \sqrt{0{,}0099/(1 + 0{,}142)} = 0{,}093$, то есть интенсивность изменения долей в распределении россиян по доходам по формуле Монро составила 9,3%. Аналогично по той же формуле по данным таблицы 20: $I_{Monroe} = \sqrt{0{,}0171/(1 + 0{,}2455)} = 0{,}117$, то есть различие структуры безработных среди женщин и мужчин по уровню образованию по формуле Монро составляет 11,7%.

Интегральный коэффициент структурных сдвигов Гатева (59), который различает структуры с равными суммами квадратов отклонений (принимает более высокие значения, когда группы имеют примерно одинаковые доли)**:**

$$I_{Gatev} = \sqrt{\frac{\sum\left(d_{1j} - d_{0j}\right)^2}{\sum\left(d_{1j}^2 + d_{0j}^2\right)}} \ . \qquad (59)$$

По данным таблицы 19 по формуле **(**59**)**: $I_{Gatev} = \sqrt{0{,}0099/(0{,}142 + 0{,}1687)} = 0{,}179$, то есть интенсивность изменения долей в распределении россиян по доходам по методике Гатева составила 17,9% (незначительно). Аналогично по той же формуле по данным таблицы 20: $I_{Gatev} = \sqrt{0{,}0171/(0{,}2455 + 0{,}2177)} = 0{,}192$, то есть различие структуры безработных среди женщин и мужчин по уровню образованию по методике Гатева составляет 19,2% (незначительно).

Индекс *Рябцева,* отличающийся от (59) только знаменателем, принимает обычно более низкие значения, рассчитывается по формуле (60)**:**

$$I_{Ryabtsev} = \sqrt{\frac{\sum\left(d_{1j} - d_{0j}\right)^2}{\sum\left(d_{1j} + d_{0j}\right)^2}} \ . \qquad (60)$$

По данным таблицы 19 по формуле **(**60**)**: $I_{Ryabtsev} = \sqrt{0{,}0099/0{,}6114} = 0{,}127$, то есть интенсивность изменения долей в распределении

[28] Индекс не удовлетворяет свойству независимости от раскола совокупности

россиян по доходам по методике Рябцева составила 12,7% (незначительно). Аналогично по той же формуле по данным таблицы 20: $I_{Ryabtsev}=\sqrt{0{,}0171/0{,}9092}=0{,}137$, то есть различие структуры безработных среди женщин и мужчин по уровню образованию по методике Рябцева составляет 13,7% (достаточно значительно).

Индекс структурных различий *Салаи* (61), особенностью которого является то, что чем больше доля *j*-ой группы, тем большее значение будет принимать $(d_{1j}+d_{0j})^2$, что ведет к уменьшению вклада *j*-ой группы в общей сумме, тем самым увеличивая значимость изменения долей малых групп:

$$I_{Szalai}=\sqrt{\frac{\sum\left(\left(d_{1j}-d_{0j}\right)/\left(d_{1j}+d_{0j}\right)\right)^2}{k}}. \qquad (61)$$

По данным таблицы 19 по формуле (61): $I_{Szalai}=\sqrt{0{,}1909/8}=0{,}154$, то есть средняя интенсивность изменения долей в распределении россиян по доходам по методике Салаи составила 15,4%. Аналогично по той же формуле по данным таблицы 20: $I_{Szalai}=\sqrt{0{,}1536/7}=0{,}148$, то есть среднее различие долей в группах безработных среди женщин и мужчин по уровню образованию по методике Салаи составляет 14,8%.

Для оценки структуры распределения доходов применяются специфические индексы: *индекс Джини, индекс Аткинсона*, *индекс обобщенной энтропии*, которые будут рассмотрены в курсе социально-экономической статистики в теме «Статистика уровня жизни».

4.2. Ранговые показатели изменения структуры

Для измерения различий структуры часто используют менее точные, но более простые по расчету показатели, которые основаны на оценки различий не самих значений долей, а их рангов, то есть порядковых номеров. Для этого чаще всего используются 2 показателя[29] – линейный и квадратический коэффициенты изменения (различия) рангов долей. Эти показатели как правило применяются для анализа структуры распределения описательных (атрибутивных) признаков (например, таблица 20), а также для оценки вотумов (голосований).

В 5-м и 6-м столбцах таблицы 21 определены ранги по данным таблицы 20, а в последующих приведены вспомогательные расчеты, необходимые в дальнейшем.

Таблица 21. Вспомогательные расчеты для определения ранговых показателей изменения структуры

№ группы (*j*)	Имеют образование	d_0	d_1	Ранг мужчин R_0	Ранг женщин R_1	$\lvert R_1-R_0\rvert$	$(R_1-R_0)^2$
1	Высшее профессиональное	0,087	0,130	5	4	1	1
2	Неполное высшее профессиональное	0,019	0,023	6	6	0	0
3	Среднее профессиональное	0,130	0,221	4	2	2	4
4	Начальное профессиональное	0,200	0,149	2	3	1	1
5	Среднее (полное) общее	0,398	0,338	1	1	0	0
6	Основное общее	0,148	0,121	3	5	2	4
7	Начальное общее, не имеют образ-я	0,018	0,018	7	7	0	0
	Итого	1,000	1,000			6	10

[29] Существуют и другие показатели, о которых можно прочитать в специальной литературе

Линейный коэффициент различия рангов долей ($ЛК_R$) – это отношение фактической суммы модулей изменения рангов к предельно возможной сумме модулей при k элементах структуры. Для четного k определяется по формуле , (62), а для нечетного k – по формуле , (63):

$$ЛК_R = \frac{\sum \left| R_{1j} - R_{0j} \right|}{k^2 / 2}, \quad (62) \qquad ЛК_R = \frac{\sum \left| R_{1j} - R_{0j} \right|}{(k^2 - 1)/2}, \quad (63)$$

где R_{1j} и R_{0j} – ранги доли j-го элемента структуры (группы) в сравниваемых совокупностях.

Так по данным таблицы 21, где в предпоследнем столбце рассчитана сумма модулей различий рангов, по формуле , (63):

$ЛК_R = \frac{6}{(7^2 - 1)/2}$ = 6/24 = 0,25, то есть различие структуры безработных среди женщин и мужчин по уровню образованию ощутимо и составляет 25% от максимально возможного.

Квадратический коэффициент различия рангов долей ($КК_R$) основан на коэффициенте корреляции рангов Спирмена, особенностью которого является то, что он позволяет определить корреляцию по таким признакам, которые нельзя выразить численно, но можно проранжировать (об этом будет подробно рассказано позднее – в теме 7.4). При полном совпадении рангов долей в базисном и отчетном периодах коэффициент Спирмена равен +1, а при максимальном различии рангов (первый становится последним, порядок рангов «переворачивается») коэффициент Спирмена составит –1, следовательно максимальное значение изменения коэффициента Спирмена равно 2. Чтобы получить показатель степени (существенности) различия рангов элементов структуры, следует отклонение фактического коэффициента Спирмена от единицы разделить на 2:

$$KK_R = \frac{1 - r_{спирмен}}{2} = \frac{1 - \left(1 - \frac{6 \sum \left(R_{1j} - R_{0j} \right)^2}{k^3 - k} \right)}{2} = \frac{3 \sum \left(R_{1j} - R_{0j} \right)^2}{k^3 - k}. \quad (64)$$

Для расчета квадратического коэффициента различия рангов долей необходима сумма квадратов различий рангов, которая рассчитана в последнем столбце таблицы 21, тогда по формуле (64):

$КК_R = \frac{3*10}{7^3 - 7}$ = 30/336 = 0,089, то есть различие структуры безработных среди женщин и мужчин по уровню образованию составляет 8,9% от максимально возможного.

4.3. Контрольные задания

Вариант 1. По данным ФСГС о распределении численности занятых в экономике России по уровню образования, представленным в таблице 22, проанализировать различия в структурах распределения среди мужчин и женщин.

Таблица 22. Варианты выполнения контрольного задания

Год (**вариант**)	Имеют образ-е / Доля	Высшее профес-сиональное	Неполное высшее профес-сиональное	Среднее профес-сиональное	Начальное профес-сиональное	Среднее (полное) общее	Основное общее	Начальное общее, не имеют образ-я
1995 (**1**)	мужчин	0,160	0,017	0,276	…	0,377	0,149	0,021
	женщин	0,192	0,014	0,387	…	0,299	0,096	0,012
1997 (**2**)	мужчин	0,184	0,019	0,280	0,073	0,292	0,124	0,028
	женщин	0,220	0,018	0,377	0,049	0,238	0,081	0,017
1998 (**3**)	мужчин	0,189	0,019	0,290	0,088	0,279	0,113	0,022
	женщин	0,226	0,019	0,384	0,060	0,225	0,072	0,014
1999 (**4**)	мужчин	0,184	0,022	0,290	0,107	0,268	0,101	0,028
	женщин	0,222	0,023	0,377	0,068	0,214	0,071	0,025
2000 (**5**)	мужчин	0,186	0,041	0,247	0,128	0,267	0,107	0,024
	женщин	0,228	0,048	0,317	0,095	0,219	0,076	0,017
2001 (**6**)	мужчин	0,205	0,024	0,266	0,146	0,258	0,090	0,011
	женщин	0,250	0,027	0,349	0,090	0,216	0,060	0,008
2002 (**7**)	мужчин	0,198	0,023	0,280	0,139	0,265	0,087	0,008
	женщин	0,249	0,026	0,353	0,087	0,220	0,057	0,008
2003 (**8**)	мужчин	0,205	0,020	0,211	0,198	0,264	0,093	0,009
	женщин	0,248	0,022	0,317	0,130	0,210	0,064	0,009
2004 (**9**)	мужчин	0,215	0,019	0,203	0,219	0,255	0,083	0,006
	женщин	0,262	0,022	0,312	0,133	0,213	0,052	0,006
2006 (**10**)	мужчин	0,235	0,017	0,198	0,218	0,255	0,072	0,005
	женщин	0,279	0,018	0,315	0,142	0,196	0,045	0,005

5. Выборочное наблюдение

5.1. Понятие выборочного наблюдения

Выборочный метод используется, когда применение сплошного наблюдения физически невозможно из-за огромного массива данных или экономически нецелесообразно. Физическая невозможность имеет место, например, при изучении пассажиропотоков, рыночных цен, семейных бюджетов. Экономическая нецелесообразность имеет место при оценке качества товаров, связанной с их уничтожением. Например, дегустация, испытание кирпичей на прочность и т.п. Выборочное наблюдение используется также для проверки результатов сплошного.

Статистические единицы, отобранные для наблюдения, составляют *выборочную* совокупность или *выборку,* а весь их массив - *генеральную* совокупность (ГС). При этом число единиц в выборке обозначают *n,* во всей ГС – *N.* Отношение *n/N* называется относительный размер или *доля выборки*.

Качество результатов выборочного наблюдения зависит от *репрезентативности* выборки, т.е. от того, насколько она представительна в ГС. Для обеспечения репрезентативности выборки необходимо соблюдать принцип случайности отбора единиц, который предполагает, что на включение единицы ГС в выборку не может повлиять какой-либо иной фактор кроме случая..

5.2. Способы формирования выборки

1. *Собственно случайный* отбор: все единицы ГС нумеруются, а выпавшие в результате жеребьевки номера соответствуют единицам, попавшим в выборку, причем число номеров равно запланированному объему выборки. На практике вместо жеребьевки используют генераторы случайных чисел. Данный способ отбора может быть *повторным* (когда каждая единица, отобранная в выборку, после проведения наблюдения возвращается в ГС и может быть вновь подвергнута обследованию) и *бесповторным* (когда обследованные единицы в ГС не возвращаются и не могут быть обследованы повторно). При повторном отборе вероятность попадания в выборку для каждой единицы ГС остается неизменной, а при бесповторном отборе она меняется (увеличивается), но для оставшихся в ГС после отбора из нее нескольких единиц, вероятность попадания в выборку одинакова.

2. Механический отбор: отбираются единицы генеральной совокупности с постоянным шагом *N/n*. Так, если она генеральная совокупность содержит 100 тыс.ед., а требуется выбрать 1 тыс.ед., то в выборку попадет каждая сотая единица.

3. *Стратифицированный* (расслоенным) отбор осуществляется из неоднородной генеральной совокупности, когда ее предварительно разбивают на однородные группы, после чего производят отбор единиц из каждой группы в выборочную совокупность случайный или механическим способом пропорционально их численности в генеральной совокупности.

4. *Серийный* (гнездовой) отбор: случайным или механическим способом выбирают не отдельные единицы, а определенные серии (гнезда), внутри которых производится сплошное наблюдение.

5.3. Средняя ошибка выборки

После завершения отбора необходимого числа единиц в выборку и регистрации предусмотренных программой наблюдения изучаемых признаков этих единиц, переходят к расчету обобщающих показателей. К ним относят среднюю величину изучаемого признака и долю единиц, обладающих каким-либо значением этого признака. Однако, если ГС произвести несколько выборок, определив при этом их обобщающие характеристики, то можно установить, что их значения будут различными, кроме того, они будут отличаться и от реального их значения в ГС, если такое определить с помощью сплошного наблюдения. Другими словами, обобщающие характеристики, рассчитанные по данным выборки, будут отличаться от их реальных значений в ГС, поэтому введем следующие условные обозначения (табл. 23).

Таблица 23. Условные обозначения

Показатель	Совокупность

	генеральная	выборочная
Число единиц совокупности	N	n
Среднее значение	$\overline{X}$	$\widetilde{X}$
Доля единиц, обладающих каким-либо значением признака	d	$\widetilde{d}$
Доля единиц, не обладающих каким-либо значением признака	$1-d$	$1-\widetilde{d}$
Дисперсия	σ^2	$\widetilde{\sigma}^2$

Разность между значением обобщающих характеристик выборочной и генеральной совокупностей называется *ошибкой выборки,* которая подразделяется на ошибку *регистрации* и ошибку *репрезентативности*. Первая возникает из-за неправильных или неточных сведений по причинам непонимания существа вопроса, невнимательности регистратора при заполнении анкет, формуляров и т.п. Она достаточно легко обнаруживается и устраняется. Вторая возникает из-за несоблюдения принципа случайности отбора единиц в выборку. Ее сложнее обнаружить и устранить, она гораздо больше первой и потому ее измерение является основной задачей выборочного наблюдения.

Для измерения ошибки выборки определяется ее средняя ошибка по формуле (65) для повторного отбора и по формуле (66) – для бесповторного:

$$\mu = \sqrt{\frac{\widetilde{\sigma}^2}{n}}; \qquad (65)$$

$$\mu = \sqrt{\frac{\widetilde{\sigma}^2}{n}\left(1-\frac{n}{N}\right)}. \qquad (66)$$

Из формул (65) и (66) видно, что средняя ошибка меньше у бесповторной выборки, что и обусловливает ее более широкое применение.

5.4. Предельная ошибка выборки

Учитывая, что на основе выборочного обследования нельзя точно оценить обобщающую характеристику ГС, необходимо найти пределы, в которых он находится. В конкретной выборке разность $\left|\widetilde{X}_i - \overline{X}\right|$ может быть больше, меньше или равна μ. Каждое из отклонений $\left|\widetilde{X}_i - \overline{X}\right|$ от μ имеет определенную вероятность. При выборочном обследовании реальное значение $\overline{X}$ в ГС неизвестно. Зная среднюю ошибку выборки, с определенной вероятностью можно оценить отклонение выборочной средней от генеральной и установить пределы, в которых находится изучаемый параметр (в данном случае среднее значение) в генеральной совокупности. Отклонение выборочной характеристики от генеральной называется *предельной ошибкой выборки* Δ. Она определяется в долях средней ошибки с заданной вероятностью, т.е.

$$\Delta = t\mu, \qquad (67)$$

где t – *коэффициент доверия*, зависящий от вероятности, с которой определяется предельная ошибка выборки.

Вероятность появления определенной ошибки выборки находят с помощью теорем теории вероятностей. Согласно теореме Чебышёва, *при достаточно большом объеме выборки и ограниченной дисперсии генеральной ГС вероятность того, что разность между выборочной средней и генеральной средней будет сколь угодно мала, близка к единице*:

$$P\left(\left|\widetilde{X} - \overline{X}\right| \le \xi\right) \to 1 \text{ при } n \to \infty. \qquad (68)$$

А. М. Ляпунов доказал, что *независимо от характера распределения генеральной ГС при увеличении объема выборки распределение вероятностей появления того или иного значения выборочной средней приближается к нормальному распределению* (центральная предельная теорема). Следовательно, вероятность отклонения выборочной средней от генеральной средней, т.е. вероятность появления заданной предельной ошибки, также подчиняется указанному закону и может быть найдена как функция от t с помощью интеграла вероятностей Лапласа:

$$P\left(\left|\widetilde{X}-\overline{X}\right|\le t\mu\right)=\frac{1}{\sqrt{2\pi}}\int_{-t}^{+t}e^{-\frac{t^2}{2}}dt\,, \qquad (69)$$

где $t=\dfrac{\widetilde{X}-\overline{X}}{\mu}$ – нормированное отклонение выборочной средней от генеральной средней.

Значения P (интеграла Лапласа) для разных t рассчитаны и имеются в специальной таблице, которая приведена в Приложении 1.

Вероятность, которая принимается при расчете выборочной характеристики, называется *доверительной*. Чаще всего принимают вероятность P = 0,950, которая означает, что только в 5 случаях из 100 ошибка может выйти за установленные границы. Задавшись конкретным уровнем вероятности, выбирают величину нормированного отклонения t по Приложению 1 и рассчитывают предельную ошибку выборки по формуле (67).

После расчета предельной ошибки находят *доверительный интервал* обобщающей характеристики ГС совокупности по формуле (70) – для среднего значения, и по формуле (71) – для доли единиц, обладающих каким-либо значением признака:

$$\overline{X}=\widetilde{X}\pm\Delta \quad \text{или} \quad (\widetilde{X}-\Delta)\le\overline{X}\le(\widetilde{X}+\Delta) \qquad (70)$$

$$d=\widetilde{d}\pm\Delta \quad \text{или} \quad (\widetilde{d}-\Delta)\le d\le(\widetilde{d}+\Delta) \qquad (71)$$

Следовательно, при выборочном наблюдении определяется не одно, точное значение обобщающей характеристики ГС, а лишь ее доверительный интервал с заданным уровнем вероятности. И это серьезный недостаток выборочного метода статистики.

5.5. Необходимая численность выборки

Разрабатывая программу выборочного наблюдения, задаются конкретным значением предельной ошибки и уровнем вероятности. Неизвестной остается минимальная численность выборки, обеспечивающая заданную точность. Ее можно получить из формул средней и предельной ошибок в зависимости от типа выборки. Так, подставляя формулы сначала (65) и затем (66) в формулу (67) и решая ее относительно численности выборки, получим следующие формулы:

для повторной выборки $n=\dfrac{\widetilde{\sigma}^2t^2}{\Delta^2}$; (72) для бесповторной выборки $n=\dfrac{\widetilde{\sigma}^2t^2}{\Delta^2+\widetilde{\sigma}^2t^2/N}$. (73)

Вариация ($\widetilde{\sigma}^2$) значений признака к началу выборочного наблюдения как правило неизвестна, поэтому ее берут приближенно одним из способов:

1) берется из предыдущих выборочных наблюдений;
2) по правилу «трех сигм», согласно которому в размахе вариации укладывается примерно 6 стандартных отклонений σ (H/σ = 6, отсюда $\widetilde{\sigma}^2 = H^2/36$);
3) если приблизительно известна средняя величина изучаемого признака, то $\widetilde{\sigma}^2 = \overline{X}^2/9$;
4) если неизвестна дисперсия доли единиц, обладающих каким-либо значением признака, то используется ее максимально возможная величина $\widetilde{\sigma}^2 = 0{,}25$.

5.6. Методические указания

*Задача***.** На предприятии в порядке случайной бесповторной выборки было опрошено 100 рабочих из 1000 и получены следующие данные об их доходе за месяц (таблица 24):

Таблица 24. Результаты бесповторного выборочного наблюдения на предприятии

Доход, у.е.	до 300	300-500	500-700	700-1000	более 1000
Число рабочих	8	28	44	17	3

С вероятностью 0,950 определить:

1) среднемесячный размер дохода работников данного предприятия;
2) долю рабочих предприятия, имеющих месячный доход более 700 у.е.;
3) необходимую численность выборки при определении среднемесячного дохода работников предприятия, чтобы не ошибиться более чем на 50 у.е.;
4) необходимую численность выборки при определении доли рабочих с размером месячного дохода более 700 у.е., чтобы при этом не ошибиться более чем на 5%.

Решение. Для расчета обобщающих характеристик выборки построим вспомогательную таблицу 25.

Таблица 25. Вспомогательные расчеты для решения задачи

X	f	X'	$X'f$	$(X' - \widetilde{X})^2$	$(X' - \widetilde{X})^2 f$
до 300	8	200	1600	137641	1101128
300 - 500	28	400	11200	29241	818748
500 - 700	44	600	26400	841	37004
700 - 1000	17	850	14450	77841	1323297
более 1000	3	1150	3450	335241	1005723
Итого	100		57100		4285900

По формуле (11) рассчитаем средний доход в выборке: $\widetilde{X}$ = 57100/100 = 571 (у.е.). Применив формулу (28) и рассчитав ее числитель в последнем столбце таблицы, получим дисперсию среднего выборочного дохода: $\widetilde{\sigma}^2$ = 4285900/100 = 42859.

Теперь можно определить среднюю ошибку выборки по формуле (66): $\mu = \sqrt{\frac{42859}{100}\left(1-\frac{100}{1000}\right)} =$ 19,640 (у.е.).

В нашей задаче β = 0,950, значит t = 1,96. Тогда предельная ошибка выборки по формуле (67):

Δ = 1,96*19,64 = 38,494 (у.е.).

Для определения средней ошибки выборки при определении доли рабочих с доходами более 700 у.е. в ГС необходимо определить их долю: w = 20/100 = 0,2 или 20%, а затем ее дисперсию по формуле $\widetilde{\sigma}^2 = w(1-w)$ = 0,2*(1–0,2) = 0,16. Тогда можно рассчитать среднюю ошибку выборки по формуле (66): $\mu = \sqrt{\frac{0,16}{100}\left(1-\frac{100}{1000}\right)}$ = 0,038 или 3,8%. А затем и предельную ошибку выборки по формуле (67):

Δ = 1,96*0,038 = 0,075 или 7,5%.

Доверительный интервал среднего дохода находим по формуле (70):

571-38,494 $\le \overline{X} \le$ 571+38,494 или 532,506 у.е. $\le \overline{X} \le$ 609,494 у.е., то есть средний доход всех рабочих предприятия с вероятностью 95% будет лежать в пределах от 532,5 до 609,5 у.е.

Аналогично определяем доверительный интервал для доли по формуле (71):

0,2-0,075 $\le p \le$ 0,2+0,075 или 0,125 $\le p \le$ 0,275, то есть доля рабочих с доходами более 700 у.е. на всем предприятии с вероятностью 95% будет лежать в пределах от 12,5% до 27,5%.

В нашей задаче выборка бесповторная, значит, воспользуемся формулой (73), в которую подставим уже рассчитанные дисперсии среднего выборочного дохода рабочих ($\widetilde{\sigma}^2$ = 42859) и доли рабочих с доходами более 700 у.е. ($\widetilde{\sigma}^2$ = 0,16):

$n_{б/повт} = \frac{42859*1,96^2}{50^2 + 42859*1,96^2/1000}$ = 62 (чел.), $n_{б/повт} = \frac{0,16*1,96^2}{0,05^2 + 0,16*1,96^2/1000}$ = 197 (чел.).

Таким образом, необходимо включить в выборку не менее 62 рабочих при определении среднего месячного дохода работников предприятия, чтобы не ошибиться более чем на 50 у.е., и не менее 197 рабочих при определении доли рабочих с размером месячного дохода более 700 у.е., чтобы при этом не ошибиться более чем на 5%.

5.7. Контрольные задания

Для изучения вкладов населения в коммерческом банке города была проведена 5%-я случайная бесповторная выборка лицевых счетов, в результате которой в таблице 26 получено распределение клиентов по размеру вкладов.

Таблица 26. Варианты выполнения контрольного задания

Размер вклада, у.е.	Число вкладчиков, чел.									
	Вариант									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
до 5000	10	80	100	50	60	30	90	20	70	40
5 000 – 15 000	40	60	150	30	40	110	75	65	90	80
15 000 – 30 000	25	35	70	90	120	90	130	140	60	95
30 000 – 50 000	30	45	40	5	80	30	60	75	20	115
свыше 50 000	15	10	30	25	50	15	25	5	10	5

С вероятностью 0,954 определить:

1) средний размер вклада во всем банке;
2) долю вкладчиков во всем банке с размером вклада свыше 15000 у.е.;
3) необходимую численность выборки при определении среднего размера вклада, чтобы не ошибиться более чем на 500 у.е.;
4) необходимую численность выборки при определении доли вкладчиков во всем банке с размером вклада свыше 30 000 у.е., чтобы не ошибиться более чем на 10%.

6. Ряды динамики

6.1. Понятие о рядах динамики

Одной из важнейших задач статистики является изучение изменений анализируемых показателей во времени, то есть их динамика. Эта задача решается при помощи анализа рядов динамики (временных рядов).

Ряд динамики – это числовые значения определенного статистического показателя в последовательные моменты или периоды времени (т.е. расположенные в хронологическом порядке).

Числовые значения того или иного статистического показателя, составляющего ряд динамики, называют *уровнями* ряда и обычно обозначают через y. Первый член ряда y_1 называют начальным (базисным) уровнем, а последний y_n – конечным. Моменты или периоды времени, к которым относятся уровни, обозначают через t.

Ряды динамики, как правило, представляют в виде таблицы (см. табл. 27) или графически (см. рис. 13), причем по оси абсцисс строится шкала времени t, а по оси ординат – шкала уровней ряда y.

Таблица 27. Внешнеторговый оборот (ВО) России за период 2000-2006 гг.

Год	2000	2001	2002	2003	2004	2005	2006
Млрд. долл. США	149,9	155,6	168,3	212,0	280,6	368,9	468,4

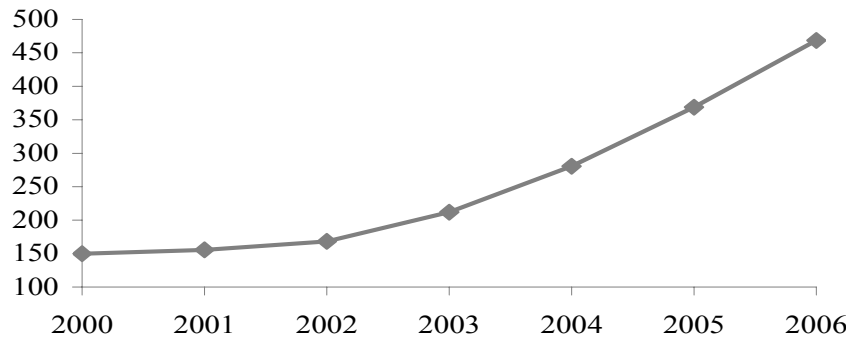


Рис. 13. Внешнеторговый оборот (ВО) России за период 2000-2006 гг.

Данные табл. 27 и рис. 13 наглядно иллюстрируют ежегодный рост внешнеторгового оборота (ВО) в России за период 2000-2006 гг.

6.2. Показатели изменения уровней ряда динамики

Анализ рядов динамики начинается с определения того, как именно изменяются уровни ряда (увеличиваются, уменьшаются или остаются неизменными) в абсолютном и относительном выражении. Чтобы проследить за направлением и размером изменений уровней во времени, для рядов динамики рассчитывают *показатели изменения уровней ряда динамики*:

- абсолютное изменение (абсолютный прирост);
- относительное изменение (темп роста или индекс динамики);
- темп изменения (темп прироста).

Все эти показатели могут определяться *базисным* способом, когда уровень данного периода сравнивается с первым (базисным) периодом, либо *цепным* способом – когда сравниваются два уровня соседних периодов.

Абсолютное изменение (абсолютный прирост) уровней рассчитывается как разность между двумя уровнями ряда по формуле (74) – для базисного способа сравнения или по формуле (75) – для цепного. Оно показывает, на сколько (в единицах показателей ряда) уровень одного (i-того) периода больше или меньше уровня какого-либо предшествующего

периода, и, следовательно, может иметь знак «+» (при увеличении уровней) или «–» (при уменьшении уровней).

$$\Delta y_i^{Б} = y_i - y_1; \qquad (74) \qquad \Delta y_i^{Ц} = y_i - y_{i-1}. \qquad (75)$$

В табл. 28 в столбце 3 рассчитаны базисные абсолютные изменения по формуле (74), а в столбце 4 – цепные абсолютные изменения по формуле (75).

Таблица 28. Анализ динамики ВО России

Год	*y*	$\Delta y_i^{Б}$	$\Delta y_i^{Ц}$	$i_i^{Б}$	$i_i^{Ц}$	$T_i^{Б}$, %	$T_i^{Ц}$,%
2000	149,9						
2001	155,6	5,7	5,7	1,038	1,038	3,8	3,8
2002	168,3	18,4	12,7	1,123	1,082	12,3	8,2
2003	212,0	62,1	43,7	1,414	1,260	41,4	26,0
2004	280,6	130,7	68,6	1,872	1,324	87,2	32,4
2005	368,9	219,0	88,3	2,461	1,315	146,1	31,5
2006	468,4	318,5	99,5	3,125	1,270	212,5	27,0
Итого	1803,7		318,5		3,125		

Между базисными и цепными абсолютными изменениями существует взаимосвязь: сумма цепных абсолютных изменений равна последнему базисному изменению, то есть

$$\sum_{i=1}^{n} \Delta y_i^{Ц} = \Delta y_n^{Б}. \qquad (76)$$

В нашем примере про ВО подтверждается правильность расчета абсолютных изменений по формуле (76): $\sum_{i=1}^{n} \Delta y_i^{Ц}$ = 318,5 рассчитана в итоговой строке 4-го столбца, а $\Delta y_n^{Б}$ = 318,5 – в предпоследней строке 3-го столбца табл. 28.

Относительное изменение (темп роста или индекс динамики) уровней рассчитывается как отношение (деление) двух уровней ряда по формуле (77) – для базисного способа сравнения или по формуле (78) – для цепного.

$$i_i^{Б} = y_i / y_1; \qquad (77) \qquad i_i^{Ц} = y_i / y_{i-1}. \qquad (78)$$

Относительное изменение показывает во сколько раз уровень данного периода больше уровня какого-либо предшествующего периода (при i_i>1) или какую его часть составляет (при i_i<1). Относительное изменение может выражаться в виде *коэффициентов*, то есть простого кратного отношения (если база сравнения принимается за единицу), и в *процентах* (если база сравнения принимается за 100 единиц) путем домножения относительного изменения на 100%.

В табл. 28 в столбце 5 рассчитаны базисные относительные изменения по формуле (77), а в столбце 6 – цепные относительные изменения по формуле (78).

Между базисными и цепными относительными изменениями существует взаимосвязь: произведение цепных относительных изменений равно последнему базисному изменению, то есть

$$\prod_{i=1}^{n} i_i^{Ц} = i_n^{Б}. \qquad (79)$$

В нашем примере про ВО подтверждается правильность расчета относительных изменений по формуле (79): $\prod_{i=1}^{n} i_i^{Ц} =$ 1,038*1,082*1,260*1,324*1,315*1,270 = 3,125 рассчитано по данным 6-го столбца, а $i_n^{Б} = 3{,}125$ – в предпоследней строке 5-го столбца табл. 28.

Темп изменения (темп прироста) уровней – относительный показатель, показывающий, на сколько процентов данный уровень больше (или меньше) другого, принимаемого за базу сравнения. Он рассчитывается путем вычитания из относительного изменения 100%, то есть по формуле (80):

$$T_i = i_i - 100\%, \tag{80}$$

или как процентное отношение абсолютного изменения к тому уровню, по сравнению с которым рассчитано абсолютное изменение (базисный уровень), то есть по формуле (81):

$$T_i = \frac{\Delta y_i}{y_{баз}} 100\%. \tag{81}$$

В табл. 28 в столбце 7 рассчитаны базисные темпы изменения ВО по формуле (80), а в столбце 8 – цепные темпы изменения по формуле (81). Все расчеты в табл. 28 свидетельствуют о ежегодном росте ВО России за период 2000-2006 гг.

6.3. Средние показатели ряда динамики

Каждый ряд динамики можно рассматривать как некую совокупность n меняющихся во времени показателей, которые можно обобщить в виде средних величин. Такие обобщенные (средние) показатели особенно необходимы при сравнении динамики изменений того или иного показателя ВЭД в разные периоды, в разных странах и т.д.

Обобщенной характеристикой ряда динамики служит прежде всего средний уровень ряда $\bar{y}$. Для разных видов рядов динамики он рассчитывается неодинаково. Ряды динамики бывают равномерные (с равными интервалами времени между уровнями), для которых средний уровень определяется по простой формуле средней величины, и неравномерные (с неравными интервалами), для которых используются формулы средних взвешенных (по интервалам времени) величин. В интервальном ряду динамики (в котором время задано в виде промежутков времени, к которым относятся уровни) $\bar{y}$ определяется по формуле средней арифметической, а в моментном ряду (в котором время задано в виде конкретных моментов времени или дат, к которым относятся уровни) – по формуле средней хронологической. В табл. 29 приводятся виды рядов динамики и соответствующие формулы для расчета их среднего уровня $\bar{y}$.

Таблица 29. Виды средних величин, применяемых при расчете среднего уровня

Вид ряда динамики	*Название средней величины*	*Формула средней величины*	*Номер формулы*
Равномерный интервальный	Арифметическая простая	$\bar{y} = \frac{\sum y}{n}$	(82)
Равномерный моментный	Хронологическая простая	$\bar{y} = \frac{\frac{y_1}{2} + y_2 + y_3 + \ldots + y_{n-1} + \frac{y_n}{2}}{n-1} = \frac{\frac{y_1 + y_n}{2} + \sum_{i=2}^{n-1} y_i}{n-1}$	(83)
Неравномерный интервальный	Арифметическая взвешенная	$\bar{y} = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i}$	(84)

Вид ряда динамики	*Название средней величины*	*Формула средней величины*	*Номер формулы*
Неравномерный моментный	Хронологическая взвешенная	$\overline{y} = \frac{\sum_{i=1}^{n-1}(y_i + y_{i+1})t_i}{2\sum_{i=1}^{n-1} t_i}$	(85)

В нашем примере про ВО России за период 2000-2006 гг. имеем равномерный интервальный ряд динамики, поэтому его средний уровень определяем по формуле (82): $\overline{y}$ = 1803,7/7 = 257,671, то есть ВО России в период 2000-2006 гг. составлял ежегодно в среднем 257,671 млрд. долл. США.

Кроме среднего уровня ряда рассчитываются и другие средние показатели:
- среднее абсолютное изменение (средний абсолютный прирост);
- среднее относительное изменение (средний темп роста);
- средний темп изменения (средний темп прироста).

Каждый из этих показателей может рассчитываться базисным и цепным способом.

Базисное среднее абсолютное изменение – это частное от деления последнего базисного абсолютного изменения на количество изменений уровней (86); *цепное среднее абсолютное изменение* уровней ряда – это частное от деления суммы всех цепных абсолютных изменений на количество изменений (87):

$$\overline{\Delta Y}^{Б} = \frac{\Delta Y_n^{Б}}{n-1}; \qquad (86) \qquad\qquad \overline{\Delta Y}^{Ц} = \frac{\sum \Delta Y^{Ц}}{n-1}. \qquad (87)$$

По знаку средних абсолютных изменений также судят о характере изменения явления в среднем: рост, спад или стабильность. Очевидно, что числители формулы (86) и (87) равны между собой по формуле (76), значит, среднее абсолютное изменение не зависит от способа расчета (базисный или цепной), так как результат получится одинаковый. В нашей задаче по формуле (86) или (87):

$\overline{\Delta Y}$ = 318,5/6 = 53,083, то есть ежегодно в среднем ВО растет на 53,083 млрд. долл.

Наряду со средним абсолютным изменением рассчитывается и среднее относительное. *Базисное среднее относительное изменение* определяется по формуле (88), а *цепное среднее относительное изменение* – по формуле (89):

$$\overline{i}^{Б} = \sqrt[n-1]{i_n^{Б}} = \sqrt[n-1]{Y_n / Y_1}, \qquad (88) \qquad\qquad \overline{i}^{Ц} = \sqrt[n-1]{\prod i_n^{Ц}}. \qquad (89)$$

Естественно, базисное и цепное среднее относительное изменения должны быть одинаковыми и сравнением их с критериальным значением 1 делается вывод о характере изменения явления в среднем: рост, спад или стабильность. В нашем примере про ВО: $\overline{i}$ = $\sqrt[6]{3{,}125}$ = 1,209, то есть ежегодно в среднем в период 2000-2006 гг. ВО России растет в 1,209 раза.

Вычитанием 100% из среднего относительного изменения образуется соответствующий *средний темп изменения*, по знаку которого также можно судить о характере изменения изучаемого явления, отраженного данным рядом динамики. В нашем примере про ВО: $\overline{T}$ = 1,209 – 1 = 0,209, то есть ежегодно в среднем в период 2000-2006 гг. ВО России растет на 20,9%.

6.4. Методы выявления основной тенденции (тренда) в рядах динамики

Одна из основных задач изучения рядов динамики – выявить основную тенденцию (закономерность) в изменении уровней ряда, именуемую *трендом*. Закономерность в изменении уровней ряда в одних случаях проявляется наглядно, в других – может маскироваться колебаниями случайного или неслучайного характера. Поэтому, чтобы

сделать правильные выводы о закономерностях развития того или иного показателя, надо суметь отделить тренд от колебаний, вызванных случайными кратковременными причинами. На основании выделенного тренда можно экстраполировать (прогнозировать) развитие явления в будущем. С этой целью (устранить колебания, вызванные случайными причинами) ряды динамики подвергают *обработке*.

Существует несколько методов обработки рядов динамики, помогающих выявить основную тенденцию изменения уровней ряда, а именно: метод укрупнения интервалов, метод скользящей средней и аналитическое выравнивание. Во всех методах вместо фактических уровней при обработке ряда рассчитываются иные (расчетные) уровни, в которых тем или иным способом взаимопогашается действие случайных факторов и тем самым уменьшается колеблемость уровней. Последние в результате становятся как бы «выравненными», «сглаженными» по отношению к исходным фактическим данным. Такие методы обработки рядов динамики называются *сглаживанием* или *выравниванием* рядов динамики.

Простейший метод сглаживания уровней ряда – *укрупнения интервалов,* для определяется итоговое значение или средняя величина исследуемого показателя. Этот метод особенно эффективен, если первоначальные уровни ряда относятся к коротким промежуткам времени. Например, если имеются данные о ежесуточном производстве мороженого на предприятии за месяц, то, естественно, в таком ряду возможны значительные колебания уровней, так как чем меньше период, за который приводятся данные, тем больше влияние случайных факторов. Чтобы устранить это влияние, рекомендуется укрупнить интервалы времени, например до 5 или 10 дней, и для этих укрупненных интервалов рассчитать общий или среднесуточный объем производства (соответственно по пятидневкам или декадам). В ряду с укрупненными интервалами времени закономерность изменения уровней будет более наглядной. Или, например, имеются ежемесячные данные о производстве мороженого – табл.32, еще более сильно укрупним интервалы – до трех месяцев (см. табл.33).

По своей сути метод *скользящей средней* похож на метод укрупнения интервалов, но в данном случае фактические уровни заменяются средними уровнями, рассчитанными для последовательно подвижных (скользящих) укрупненных интервалов, охватывающих m уровней ряда. Например, если принять m=3, то сначала рассчитывается средняя величина из первых трех уровней, затем находится средняя величина из 2-го, 3-го и 4-го уровней, потом из 3-го, 4-го и 5-го и т.д., т.е. каждый раз в сумме трех уровней появляется новый уровень, а два остаются прежними, что и обусловливает взаимопогашение случайных колебаний в средних уровнях. Рассчитанные из m членов скользящие средние относятся к середине (центру) каждого рассматриваемого интервала.

Сглаживание методом скользящей средней можно проводить по любому числу членов m, но удобнее, если m – нечетное число, так как в этом случае скользящая средняя сразу относится к конкретной временнОй точке – середине (центру) интервала. Если же m – четное, то скользящая средняя относится к промежутку между временнЫми точками: например, при сглаживании по четырем членам (m=4) средняя из первых четырех уровней будет находиться между второй и третьей временной точкой, следующая – между третьей и четвертой и т.д. Тогда, чтобы сглаженные уровни относились непосредственно к конкретным временнЫм точкам, из каждой пары смежных промежуточных значений скользящих средних находят среднюю арифметическую, которую относят к временной точке, находящейся между смежными. Такой прием двойного расчета сглаженных уровней называется *центрированием*.

Недостатком метода скользящей средней является то, что сглаженный ряд укорачивается по сравнению с фактическим с двух концов: при нечетном m на (m-1)/2, а при четном m – на m/2 с каждого конца. Применяя этот метод, надо помнить, что он сглаживает (устраняет) лишь случайные колебания. Если же, например, ряд содержит сезонную волну (см. 6.6), она сохранится и после сглаживания методом скользящей средней. Кроме того, этот

метод сглаживания, как и метод укрупнения интервалов не позволяет выражать общую тенденцию изменения уровней в виде математической модели.

Наиболее совершенным методом обработки рядов динамики в целях устранения случайных колебаний и выявления тренда является *выравнивание уровней ряда по аналитическим формулам* (или *аналитическое выравнивание*). Суть аналитического выравнивания заключается в замене эмпирических (фактических, исходных) уровней y_i теоретическими $\hat{y}_t$, которые рассчитаны по определенному уравнению, принятому за математическую модель тренда, где теоретические уровни рассматриваются как функция времени: $\hat{y}_t = f(t)$.

При этом каждый фактический уровень y_i рассматривается обычно как сумма двух составляющих:

$$y_i = f(t) + \varepsilon_t, \tag{90}$$

где $f(t) = \hat{y}_t$ - систематическая составляющая, отражающая тренд и выраженная определенным уравнением; ε_t - случайная величина, вызывающая колебания уровней вокруг тренда.

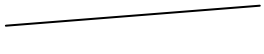
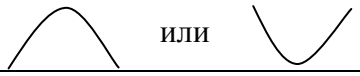
Задача аналитического выравнивания сводится к следующему:

1) определение на основе фактических данных формы (вида) гипотетической функции $\hat{y}_t = f(t)$, способной наиболее адекватно отразить тенденцию развития исследуемого показателя;
2) нахождение по эмпирическим данным параметров указанной функции (уравнения);
3) расчет по найденному уравнению теоретических (выравненных) уровней.

В аналитическом выравнивании наиболее часто используются простейшие функции, представленные в табл. 30, где обозначено $\hat{y}_t$ - теоретические (выравненные) уровни (читается как «игрек, выравненный по *t*»); *t* – условное обозначение времени (1, 2, 3 …); a_0, a_1, a_2, ... – параметры аналитической функции; *k* – число гармоник (при выравнивании по ряду Фурье).

Выбор той или иной функции для выравнивания ряда динамики осуществляется на основании графического изображения эмпирических данных. Если по тем или иным причинам уровни эмпирического ряда трудно описать одной функцией, следует разбить анализируемый период на отдельные части и затем выровнять каждую часть по соответствующей кривой.

Таблица 30. Виды математических функций[30], используемые при выравнивании

Название функции	*Вид функции*	*Формула*
Прямая линия		$\hat{y}_t = a_0 + a_1 t$ (91)
Парабола 2-го порядка		$\hat{y}_t = a_0 + a_1 t + a_2 t^2$ (92)
Парабола 3-го порядка		$\hat{y}_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3$ (93)
Гипербола		$\hat{y}_t = a_0 + \frac{a_1}{t}$ (94)
Показательная		$\hat{y}_t = a_0 a_1^t$ (95)
Степенная		$\hat{y}_t = a_0 t^{a_1}$ (96)

[30] Приведены наиболее простые функции, более сложные виды, такие как логарифмическая, логистическая и др. описаны в специальной литературе, например – [2]

Название функции	*Вид функции*	*Формула*
Ряд Фурье		$\hat{y}_t = a_0 + \sum_{k=1}^{m}(a_k \cos kt + b_k \sin kt)$ (97)

Нередко один и тот же ряд можно выровнять по разным аналитическим функциям и получить довольно близкие результаты. В нашем примере про ВО России можно произвести выравнивание и по прямой линии, и по параболе. Чтобы решить вопрос о том, использование какой кривой дает лучший результат, обычно сопоставляют суммы квадратов отклонений эмпирических уровней от теоретических (*остатки*), рассчитанным по разным функциям, то есть:

$$\sum(\hat{y}_t - y)^2 . \qquad (98)$$

Та функция, при которой эта сумма минимальна, считается наиболее адекватной, приемлемой. Однако сравнивать непосредственно суммы квадратов отклонений можно в том случае, если сравниваемые уравнения имеют одинаковое число параметров. Если же число параметров *k* разное, то каждую сумму квадратов делят на разность (*n* – *k*), выступающую в роли числа степеней свободы, и сравнивают уже квадраты отклонений уровней, рассчитанные на одну степень свободы (т.е. остаточные дисперсии на одну степень свободы).

Параметры искомых уравнений ($a_0, a_1, a_2, ...$) при аналитическом выравнивании могут быть определены по-разному, но наиболее распространенным методом является *метод наименьших квадратов* (МНК). При этом методе учитываются все эмпирические уровни и должна обеспечиваться минимальная сумма квадратов отклонений эмпирических значений уровней *y* от теоретических уровней $\hat{y}_t$:

$$\sum(\hat{y}_t - y)^2 \to \min . \qquad (99)$$

В частности, при выравнивании по прямой вида (91) параметры a_0 и a_1 отыскиваются по МНК следующим образом. В формуле (99) вместо $\hat{y}_t$ записываем его конкретное выражение $a_0 + a_1 t$. Тогда $S = \sum(a_0 + a_1 t - y)^2 \to \min$. Дальнейшее решение сводится к задаче на экстремум, т.е. к определению того, при каком значении a_0 и a_1 функция двух переменных *S* может достигнуть минимума. Как известно, для этого надо найти частные производные *S* по a_0 и a_1, приравнять их к нулю и после элементарных преобразований решить систему двух уравнений с двумя неизвестными.

В соответствии с вышеизложенным найдем частные производные:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 2\sum(a_0 + a_1 t - y) = 0 \\ \dfrac{\partial S}{\partial a_1} = 2\sum(a_0 + a_1 t - y)t = 0 \end{cases}$$

Сократив каждое уравнение на 2, раскрыв скобки и перенеся члены с *y* в правую сторону, а остальные – оставив в левой, получим систему нормальных уравнений:

$$\begin{cases} na_0 + a_1 \sum t = \sum y \\ a_0 \sum t + a_1 \sum t^2 = \sum yt \end{cases} \qquad (100)$$

где *n* – количество уровней ряда; *t* – порядковый номер в условном обозначении периода или момента времени; *y* – уровни эмпирического ряда.

Эта система и, соответственно, расчет параметров a_0 и a_1 упрощаются, если отсчет времени ведется от середины ряда[31]. Например, при *нечетном* числе уровней (как в нашем примере про ВО России – 7 уровней) серединная точка времени (год, месяц) принимается за нуль, тогда предшествующие периоды обозначаются соответственно –1, –2, –3 и т.д., а следующие за средним (центральным) – соответственно 1, 2, 3 и т.д. (см. 3-й столбец табл. 31). При *четном* числе уровней два серединных момента (периода) времени обозначают –1 и +1, а все последующие и предыдущие, соответственно, через два интервала: ± 3, ± 5, ± 7 и т.д.

При таком порядке отсчета времени (от середины ряда) $\sum t = 0$, поэтому, система нормальных уравнений (100) упрощается до следующих двух уравнений, каждое из которых решается самостоятельно:

$$\begin{cases} na_0 = \sum y \Rightarrow a_0 = \dfrac{\sum y}{n} \\ a_1 \sum t^2 = \sum yt \Rightarrow a_1 = \dfrac{\sum yt}{\sum t^2} \end{cases} \qquad (101)$$

Как видим, при такой нумерации периодов параметр a_0 представляет собой средний уровень равномерного интервального ряда, то есть формулу (82). Определим по формуле (101) параметры уравнения прямой для нашего примера про ВО России, для чего исходные данные и все расчеты необходимых сумм представим в табл. 31.

Таблица 31. Вспомогательные расчеты для линейного тренда

Год	y	t	t^2	yt	$\hat{y}_t$	$(\hat{y}_t - y)^2$	$(\hat{y}_t - \bar{y})^2$	$(y - \bar{y})^2$
2000	149,9	-3	9	-449,7	97,557	2739,775	25636,584	11614,681
2001	155,6	-2	4	-311,2	150,929	21,822	11394,038	10418,577
2002	168,3	-1	1	-168,3	204,300	1296,000	2848,509	7987,252
2003	212	0	0	0	257,671	2085,879	0,000	2085,879
2004	280,6	1	1	280,6	311,043	926,768	2848,509	525,719
2005	368,9	2	4	737,8	364,414	20,122	11394,038	12371,795
2006	468,4	3	9	1405,2	417,786	2561,806	25636,584	44406,531
Итого	1803,7	0	28	1494,4	1803,700	9652,171	79758,263	89410,434

Из табл. 31 получаем, что: a_0 = 1803,7/7 = 257,671 и a_1 = 1494,4/28 = 53,371. Отсюда искомое уравнение тренда: $\hat{y}_t = 257{,}671 + 53{,}371t$. В 6-м столбце табл. 31 приведены теоретические (трендовые) уровни, рассчитанные по этому уравнению, а в итоге 7-го столбца – остатки по формуле (98). Для иллюстрации построим график эмпирических и трендовых уровней – рис. 14.

[31] При расчете параметров уравнения тренда на ЭВМ необходимость вести отсчет от середины ряда динамики отпадает. Например, для получения уравнения тренда в *Microsoft Office Excel* необходимо построить его график с помощью «Мастера диаграмм», после чего вызвать контекстное меню, нажав на правую кнопку мыши на построенном графике, и выбрать пункт «Добавить линию тренда», в появившемся окне выбрать подходящую математическую функцию и установить галочку «показывать уравнение на диаграмме»

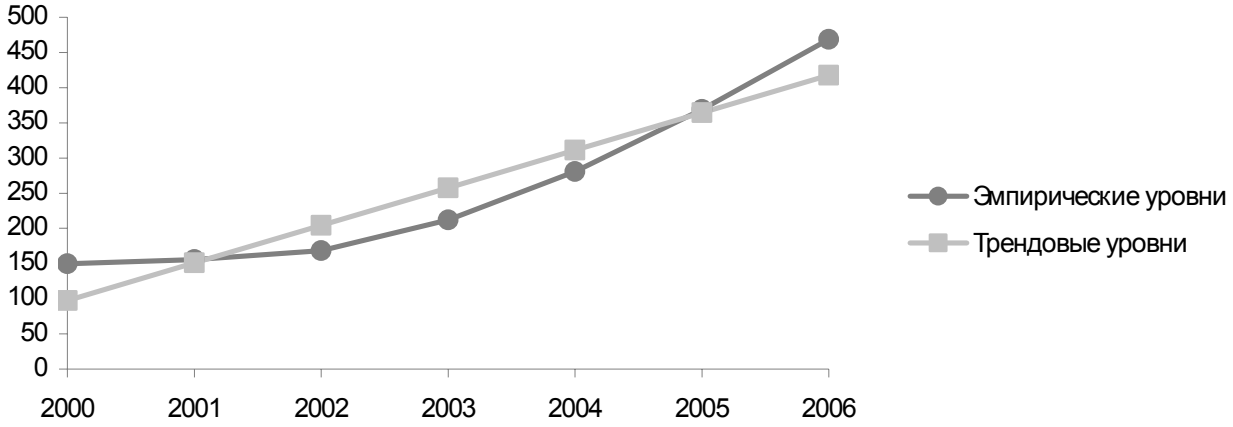


Рис. 14. Эмпирические и трендовые уровни ряда динамики ВО России

6.5. Оценка адекватности тренда и прогнозирование

Для найденного уравнения тренда необходимо провести оценку его *надежности (адекватности),* что осуществляется обычно с помощью критерия Фишера, сравнивая его расчетное значение F_p с теоретическим (табличным) значением F_T (Приложение 4). При этом расчетный критерий Фишера определяется по формуле (102):

$$F_P = \frac{(n-k)\sum(\hat{y}_t - \bar{y})^2}{(k-1)\sum(\hat{y}_t - y)^2}, \qquad (102)$$

где k – число параметров (членов) выбранного уравнения тренда.

Для проверки правильности расчета сумм в формуле (102) можно использовать следующее равенство (103):

$$\sum(y - \bar{y})^2 = \sum(\hat{y}_t - y)^2 + \sum(\hat{y}_t - \bar{y}). \qquad (103)$$

В нашем примере про ВО равенство (103) соблюдается (необходимые суммы рассчитаны в трех последних столбцах табл. 31): 89410,434 = 9652,171 + 79758,263.

Сравнение расчетного и теоретического значений критерия Фишера ведется при заданном уровне значимости[32] с учетом степеней свободы: $\nu_1 = k - 1$ и $\nu_2 = n - k$. При условии $F_p > F_T$ считается, что выбранная математическая модель ряда динамики адекватно отражает обнаруженный в нем тренд.

Проверим тренд на адекватность в нашем примере про ВО по формуле (102):
F_P = 79758,263*5/(9652,171*1) = 41,32 > F_T, значит, модель адекватна и ее можно использовать для прогнозирования (F_T = 6,61 находим по Приложению 4 в 1-ом столбце [$\nu_1 = k - 1 = 2 - 1 = 1$] и 5-й строке [$\nu_2 = n - k = 5$]).

Как уже было отмечено ранее, в нашем примере про ВО России можно произвести выравнивание не только по прямой линии, но и по параболе, чего делать не будем, так как уже найденный линейный тренд адекватно описывает тенденцию[33].

При составлении прогнозов уровней социально-экономических явлений обычно оперируют не точечной, а интервальной оценкой, рассчитывая так называемые

[32] Понятие «уровень значимости» описано ранее на стр. 29

[33] Выравнивание по параболе рассмотрено в методических указаниях к теме на другом примере

доверительные интервалы прогноза. Границы интервалов определяются по формуле , (104):

$$\hat{y}_t \pm t_\alpha \sigma_{\hat{y}}, \qquad (104)$$

где $\hat{y}_t$ – точечный прогноз, рассчитанный по модели тренда; t_α – *коэффициент доверия по распределению Стьюдента* при уровне значимости α и числе степеней свободы $\nu = n-1$ (Приложение 2)[34]; $\sigma_{\hat{y}}$ – *ошибка аппроксимации*, определяемая по формуле . (105):

$$\sigma_{\hat{y}} = \sqrt{\frac{\sum(\hat{y}_t - y)^2}{n-k}}. \qquad (105)$$

Спрогнозируем ВО России на 2007 и 2008 годы с вероятностью 0,95 (значимостью 0,05), для чего найдем ошибку аппроксимации по формуле . (105): $\sigma_{\hat{y}} = \sqrt{9652{,}171/(7-2)} = 43{,}937$ и найдем коэффициент доверия по распределению Стьюдента по Приложению 2: $t_\alpha = 2{,}4469$ при $\nu = 7 - 1 = 6$.

Прогноз на 2007 и 2008 годы с вероятностью 0,95 по формуле , (104):

$Y_{2007} = (257{,}671 + 53{,}371*4) \pm 2{,}4469*43{,}937$ или $363{,}6 < Y_{2007} < 578{,}7$ (млрд. долл.);
$Y_{2008} = (257{,}671 + 53{,}371*5) \pm 2{,}4469*43{,}937$ или $417{,}0 < Y_{2008} < 632{,}0$ (млрд. долл.).

Как видно из полученных прогнозов, доверительный интервал достаточно широк (из-за достаточно большой величины ошибки аппроксимации). Более точный прогноз можно получить при выравнивании по параболе 2-го порядка[35].

6.6. Анализ сезонных колебаний

В рядах динамики, уровни которых являются месячными или квартальными показателями, наряду со случайными колебаниями часто наблюдаются *сезонные колебания*, под которыми понимаются периодически повторяющиеся из года в год повышение и снижение уровней в отдельные месяцы или кварталы.

Сезонным колебаниям подвержены внутригодовые уровни многих показателей. Например, расход электроэнергии в летние месяцы значительно меньше, чем в зимние; или рыночные цены на овощи в отдельные месяцы далеко не одинаковы.

При графическом изображении таких рядов сезонные колебания проявляются в повышении и снижении уровней в определенные месяцы (кварталы). В качестве иллюстрации рядов с сезонными колебаниями могут служить данные, представленные в табл. 32 и их графическое изображение (рис. 15).

Таблица 32. Динамика производства мороженого предприятием по месяцам, тонн

Номер строки	Год	*Месяц t*											
		январь	февраль	март	апрель	май	июнь	июль	август	сентябрь	октябрь	ноябрь	декабрь
1	2004	30	35	45	55	58	64	69	52	42	35	33	31
2	2005	37	40	44	52	46	70	60	48	46	38	36	35
3	2006	33	39	42	56	62	73	65	56	39	35	31	28
4	Итого	100	114	131	163	166	207	194	156	127	108	100	94
5	$\overline{Y_t}$	33,333	38,000	43,667	54,333	55,333	69,000	64,667	52,000	42,333	36,000	33,333	31,333
6	$\overline{I}_{сезт}$	0,723	0,824	0,947	1,178	1,200	1,496	1,402	1,128	0,918	0,781	0,723	0,680

[34] Используется при малом количестве уровней (n<30), в противном случае (n>30) вместо t_α используют коэффициент доверия t нормального закона распределения (Приложение 1)

[35] Попробуйте проделать данное задание самостоятельно (в случае затруднений обратитесь к методическим указаниям по данной теме)

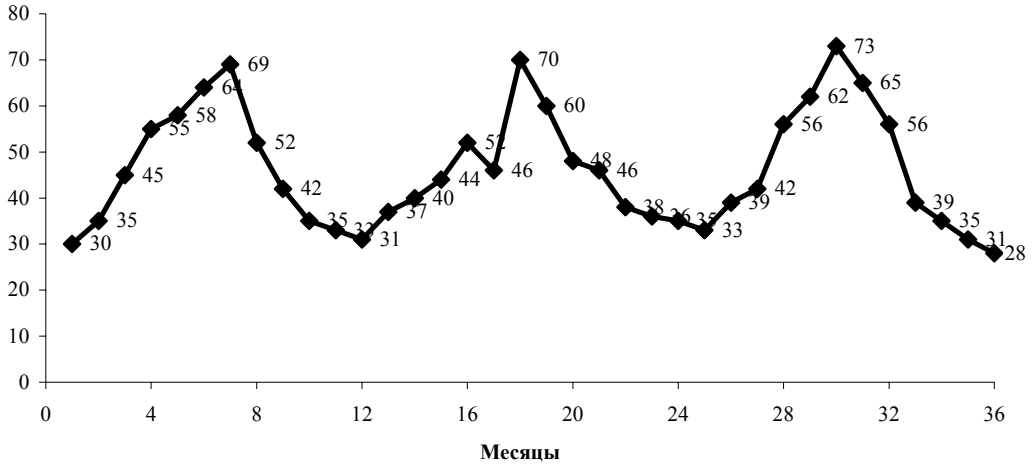


Рис. 15. Динамика производства мороженого предприятием по месяцам, тонн

Вместо месячных показателей могут быть квартальные. Если колебания не случайны, то они сохраняются и в квартальных уровнях, как это показано в табл. 33 и на рис. 16, где месячные данные из табл. 32 преобразованы в квартальные.

Таблица 33. Динамика производства мороженого предприятием по кварталам, тонн

Год	Кварталы				Итого
	1	2	3	4	
2004	110	177	163	99	549
2005	121	168	154	109	552
2006	114	191	160	94	559
Итого	345	536	477	302	1660

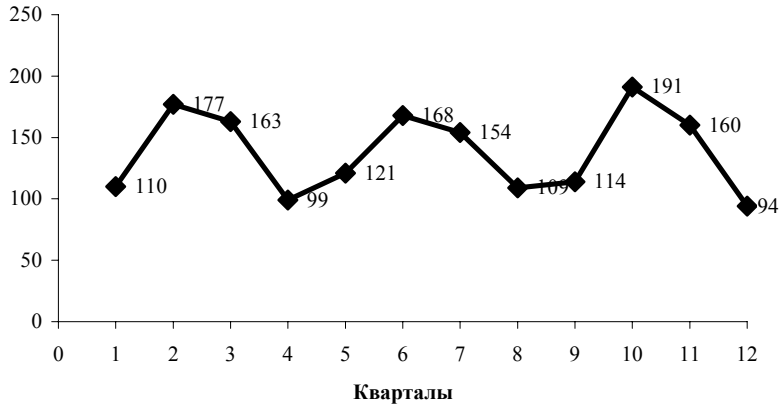


Рис. 16. Динамика производства мороженого предприятием по кварталам, тонн

Наблюдение за сезонными колебаниями позволяет устранить их там, где они нежелательны, а также решить ряд практических задач, например, определить потребности в сырье, рабочей силе в тех отраслях, где влияние сезонности велико.

При изучении рядов динамики, содержащих «сезонную волну», ее выделяют из общей колеблемости уровней и измеряют. Существует 2 основных метода для решения этой задачи: расчет *индексов сезонности* и *гармонический анализ*.

Индексы сезонности показывают, во сколько раз фактический уровень ряда в определенный момент или интервал времени t больше среднего уровня, либо уровня, вычисляемого по уравнению тренда ($\hat{y}_t$). Способы расчета индексов сезонности зависят от наличия или отсутствия тренда. Если тренда нет или от незначителен, то для каждого месяца (квартала) индекс сезонности определяется по формуле (106):

$$i_{сезт} = \frac{Y_t}{\overline{Y}}, \qquad (106)$$

где Y_t – уровень ряда динамики за месяц (квартал) t;

$\overline{Y}$ – средний уровень всего ряда динамики.

Индексы сезонности желательно рассчитывать для рядов динамики, длиной в несколько лет, тогда формула индекса сезонности примет следующий вид:

$$\overline{I}_{сезт} = \frac{\overline{Y_t}}{\overline{Y}}, \qquad (107)$$

где $\overline{Y_t} = \frac{\sum Y_t}{T}$ – средний уровень ряда динамики по одноименным месяцам t за T лет.

Например, по данным таблицы 32, представляющим ряд динамики за 3 года, индексы сезонности будем рассчитывать по формуле (107), для чего сначала рассчитаем $\sum Y_t$ (4-я строка таблицы 32), а затем, разделив полученные значение на T=3, получим средние уровни за каждый месяц $\overline{Y_t}$ (5-я строка таблицы 32). Средний уровень всего ряда определяем по формуле средней арифметической простой: $\overline{Y} = \frac{1660}{32} = 46{,}111$. В 6-й строке таблицы 32 определены индексы сезонности для каждого месяца по формуле (107), то есть делением значений в 5-й строке на 46,111.

При наличии тренда индексы сезонности определяются определяются аналогично по формулам (106) – (107) с учетом замены $\overline{Y}$ на выравненные по уравнению тренда уровни $\hat{Y}_t$. На основе найденных индексов сезонности и тренда можно спрогнозировать (экстраполировать) ряд динамики по формуле:

$$\hat{Y}_{t\,прогн} = \hat{Y}_t \overline{I}_{сезт}. \qquad (108)$$

Особое место при анализе сезонных колебаний занимает *гармонический анализ* сезонных колебаний, в котором осуществляется выравнивание ряда динамики с помощью ряда Фурье, уровни которого можно выразить как функцию времени следующим уравнением:

$$\hat{Y}_t = a_0 + \sum_{k=1}^{m} (a_k \cos kt + b_k \sin kt). \qquad (109)$$

То есть сезонные колебания уровней динамического ряда можно представить в виде синусоидальных колебаний. Поскольку последние представляют собой гармонические колебания, то синусоиды, полученные при выравнивании по ряду Фурье, называют *гармониками* различных порядков (показатель k в этом уравнении определяет число гармоник). Обычно при выравнивании по ряду Фурье рассчитывают несколько гармоник (чаще не более 4) и затем уже определяют, с каким числом гармоник ряд Фурье наилучшим образом отражает изменения уровней ряда.

При выравнивании по ряду Фурье периодические колебания уровней динамического ряда представлены в виде суммы нескольких синусоид (гармоник), наложенных друг на друга.

Так, при $k=1$ ряд Фурье будет иметь вид

$$\hat{Y}_t = a_0 + a_1 \cos t + b_1 \sin t, \qquad (110)$$

а при $k=2$, соответственно,

$$\hat{Y}_t = a_0 + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t \qquad (111)$$

и так далее.

Параметры уравнения теоретических уровней, определяемого рядом Фурье, находят, как и в других случаях, методом наименьших квадратов. Приведем без вывода формулы[36], используемые для исчисления параметров ряда Фурье:

$$a_0 = \frac{\sum Y}{n};\ a_k = \frac{2\sum Y\cos kt}{n};\ b_k = \frac{2\sum Y\sin kt}{n}. \qquad (112)$$

Последовательные значения t обычно определяются от 0 с увеличением (приростом), равным $\frac{2\pi}{n}$, где n – число уровней эмпирического ряда.

Например, при $n=10$ временнЫе точки t можно записать следующим образом:

$$0;\ \frac{2\pi}{10}\cdot 1;\ \frac{2\pi}{10}\cdot 2;\ \frac{2\pi}{10}\cdot 3;\ \frac{2\pi}{10}\cdot 4;\ \frac{2\pi}{10}\cdot 5;\ \frac{2\pi}{10}\cdot 6;\ \frac{2\pi}{10}\cdot 7;\ \frac{2\pi}{10}\cdot 8;\ \frac{2\pi}{10}\cdot 9,$$

или (после сокращения): $0;\ \frac{\pi}{5};\ \frac{2\pi}{5};\ \frac{3\pi}{5};\ \frac{4\pi}{5};\ \pi;\ \frac{6\pi}{5};\ \frac{7\pi}{5};\ \frac{8\pi}{5};\ \frac{9\pi}{5}$.

При $n=12$ значения t приведены в первой строке таблицы 34, а во второй и третьей строках определены значения sinkt и coskt для первой гармоники.

Таблица 34. Значения sinkt и coskt для первой гармоники 12-ти уровнего ряда динамики

t	0	π/6	π/3	π/2	2π/3	5π/6	π	7π/6	4π/3	3π/2	5π/3	11π/6
cost	1	$\sqrt{3}/2$	1/2	0	$-1/2$	$-\sqrt{3}/2$	–1	$-\sqrt{3}/2$	$-1/2$	0	1/2	$\sqrt{3}/2$
sint	0	1/2	$\sqrt{3}/2$	1	$\sqrt{3}/2$	1/2	0	$-1/2$	$-\sqrt{3}/2$	–1	$-\sqrt{3}/2$	$-1/2$

В таблице 35 приведены исходные данные (графы 1 и 2) и расчет показателей, необходимых для получения уравнений первой гармоники ($k=1$) по формуле . (112).

Таблица 35. Вспомогательные расчеты параметров ряда Фурье

Год		Месяц (t)												Итого
		январь (0)	февраль (π/6)	март (π/3)	апрель (π/2)	май (2π/3)	июнь (5π/6)	июль (π)	август (7π/6)	сентябрь (4π/3)	октябрь (3π/2)	ноябрь (5π/3)	декабрь (11π/6)	
2004	y	30	35	45	55	58	64	69	52	42	35	33	31	
	ycost	30	30,31	22,5	0	-29	-55,4	-69	-45	-21	-0	16,5	26,85	
	ysint	0	17,5	38,97	55	50,23	32	0	-26	-36,4	-35	-28,6	-15,5	
	$\hat{y}_t$	31,71	37,84	46,18	54,51	60,58	62,78	60,51	54,39	46,04	37,72	31,64	29,44	
2005	y	37	40	44	52	46	70	60	48	46	38	36	35	
	ycost	37	34,64	22	0	-23	-60,6	-60	-41,6	-23	-0	18	30,31	
	ysint	0	20	38,11	52	39,84	35	0	-24	-39,8	-38	-31,2	-17,5	
	$\hat{y}_t$	31,71	37,84	46,18	54,51	60,58	62,78	60,51	54,39	46,04	37,72	31,64	29,44	
2006	y	33	39	42	56	62	73	65	56	39	35	31	28	1660
	ycost	33	33,77	21	0	-31	-63,2	-65	-48,5	-19,5	-0	15,5	24,25	-259,234
	ysint	0	19,5	36,37	56	53,69	36,5	0	-28	-33,8	-35	-26,8	-14	151,122
	$\hat{y}_t$	31,71	37,84	46,18	54,51	60,58	62,78	60,51	54,39	46,04	37,72	31,64	29,44	1660

Искомое уравнение первой гармоники имеет вид: $\hat{y}_t = 46{,}111 - 14{,}402\cos t + 8{,}396\sin t$, подстановкой в которое значений t в последней строке табл.35 получены теоретические значения объема производства мороженого $\hat{y}_t$ по месяцам, а на рис.17 приведено графическое изображение, из которого видно, что различия эмпирических и теоретических уровней незначительны.

[36] Выполните это задание дома самостоятельно (подсказка: продифференцировав и приравняв нулю уравнение учтите, что $\sum\cos t = \sum\sin t = 0$ и $\cos^2 t = \frac{1+\cos 2t}{2}$)

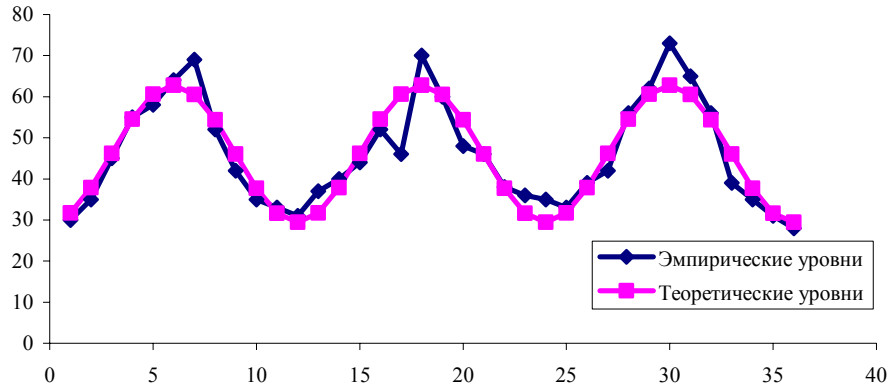


Рис. 17. Динамика производства мороженого предприятием, тонн

Аналогично рассчитываются параметры уравнения с применением второй, третьей и т.д. гармоник[37], а затем выбирается наиболее адекватное уравнение, то есть с минимальной ошибкой аппроксимации.

На основе подобранного уравнения по ряду Фурье можно прогнозировать (экстраполировать) развитие уровней ряда в будущем по формуле , (104). Например, определим доверительные интервалы производства мороженого на январь 2007 года с вероятностью 0,95, для чего найдем ошибку аппроксимации по формуле . (105): $\sigma_{£} = \sqrt{737{,}332/(36-3)} =$ 4,727 и определим коэффициент доверия по нормальному распределению (так как число уровней n>30) по Приложению 1: t = 1,96. Тогда прогноз на январь 2007 года с вероятностью 0,95 по формуле , (104): $Y_{янв07} = 31{,}71 \pm 1{,}99*4{,}727$ или $22{,}44 < Y_{2007} < 40{,}974$ (т).

6.7. Методические указания

По данным ФСГС сальдо внешней торговли (СВТ) России за период 2000-2006 гг. характеризуется рядом динамики, представленным в табл. 36.

Таблица 36. Сальдо внешней торговли (СВТ) России за период 2000-2006 гг.

Год	2000	2001	2002	2003	2004	2005	2006
Млрд. долл. США	60,1	48,1	46,3	59,9	85,8	118,3	140,7

Проанализируем данный ряд динамики: выявим тенденцию и сделаем прогноз на 2007 и 2008 годы с вероятностью 0,95.

Для большей наглядности представим данные табл. 36 на графике – рис. 18.

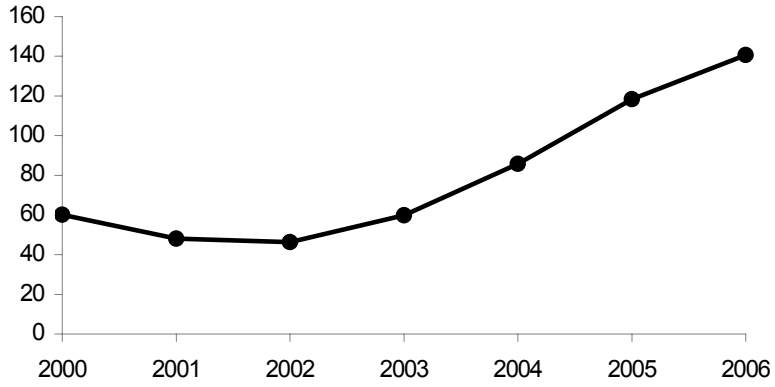


Рис. 18. Сальдо внешней торговли (СВТ) России за период 2000-2006 гг.

Данные табл. 36 и рис. 18 наглядно иллюстрируют постепенное уменьшение и последующий рост СВТ России за период 2000-2006 гг.. Очевидно, что такую динамику не следует

[37] Подобрать уравнение второй гармоники ряда Фурье по данным табл. 32 самостоятельно

описывать линейной функцией тренда. Попробуем описать эту динамику с помощью тренда по параболе 2-го порядка по формуле (92). Параметры параболы (a_0, a_1, a_2) определим методом МНК, для чего в формуле (99) вместо $\hat{y}_t$ записываем выражение параболы $a_0 + a_1 t + a_2 t^2$. Тогда $S = \sum (a_0 + a_1 t + a_2 t^2 - y)^2 \to \min$. Дальнейшее решение сводится к задаче на экстремум, т.е. к определению того, при каком значении a_0, a_1, a_2 функция трех переменных S может достигнуть минимума. Как известно, для этого надо найти частные производные S по a_0, a_1, a_2 и приравнять их к нулю и после элементарных преобразований решить систему трех уравнений с тремя неизвестными.

В соответствии с вышеизложенным найдем частные производные:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 2\sum (a_0 + a_1 t + a_2 t^2 - y) = 0 \\ \dfrac{\partial S}{\partial a_1} = 2\sum (a_0 + a_1 t + a_2 t^2 - y)t = 0 \\ \dfrac{\partial S}{\partial a_2} = 2\sum (a_0 + a_1 t + a_2 t^2 - y)t^2 = 0 \end{cases}$$

Сократив каждое уравнение на 2, раскрыв скобки и перенеся члены с y в правую сторону, а остальные – оставив в левой, получим систему нормальных уравнений:

$$\begin{cases} na_0 + a_1 \sum t + a_2 \sum t^2 = \sum y \\ a_0 \sum t + a_1 \sum t^2 + a_2 \sum t^3 = \sum yt \\ a_0 \sum t^2 + a_1 \sum t^3 + a_2 \sum t^4 = \sum yt^2 \end{cases} \qquad (113)$$

Упростим систему (113), введя условную нумерацию t от середины ряда. Тогда $\sum t = 0$ и $\sum t^3 = 0$, а система (113) упростится до следующего вида:

$$\begin{cases} na_0 + a_2 \sum t^2 = \sum y \\ a_1 \sum t^2 = \sum yt \\ a_0 \sum t^2 + a_2 \sum t^4 = \sum yt^2 \end{cases} \qquad (114)$$

Решая систему (114)[38], находим параметры a_0, a_1, a_2:

$$a_0 = \frac{\sum y \sum t^4 - \sum t^2 \sum yt^2}{n\sum t^4 - \left(\sum t^2\right)^2} \quad (115) \qquad a_1 = \frac{\sum yt}{\sum t^2} \quad (116) \qquad a_2 = \frac{n\sum yt^2 - \sum y \sum t^2}{n\sum t^4 - \left(\sum t^2\right)^2} \quad (117)$$

Определим по формулам (115) – (117) параметры уравнения параболы для нашего примера про СВТ России, для чего исходные данные и все расчеты необходимых сумм представим в табл. 37.

Таблица 37. Вспомогательные расчеты для параболического тренда

Год	y	t	t^2	t^4	yt	yt^2	$\hat{y}_t$	$(\hat{y}_t - y)^2$	$(\hat{y}_t - \bar{y})^2$	$(y - \bar{y})^2$
2000	60,1	-3	9	81	-180,3	540,9	56,614	12,150	1338,514	1095,610
2001	48,1	-2	4	16	-96,2	192,4	49,764	2,770	1886,661	2034,010
2002	46,3	-1	1	1	-46,3	46,3	51,679	28,929	1724,029	2199,610
2003	59,9	0	0	0	0,0	0,0	62,357	6,038	951,282	1108,890
2004	85,8	1	1	1	85,8	85,8	81,800	16,000	129,960	54,760
2005	118,3	2	4	16	236,6	473,2	110,007	68,771	282,480	630,010
2006	140,7	3	9	81	422,1	1266,3	146,979	39,420	2892,135	2256,250

[38] Проделайте данное задание самостоятельно

Год	y	t	t^2	t^4	yt	yt^2	$\hat{y}_t$	$(\hat{y}_t - y)^2$	$(\hat{y}_t - \bar{y})^2$	$(y - \bar{y})^2$
Итого	559,2	0	28	196	421,7	2604,9	559,200	174,079	9205,061	9379,140

Из табл. 37 получаем по формулам (115) – (117): $a_0 = 62{,}357$, $a_1 = 15{,}061$ и $a_2 = 4{,}382$. Отсюда искомое уравнение тренда $\hat{y}_t = 62{,}357 + 15{,}061t + 4{,}382t^2$. В 8-м столбце табл. 37 приведены теоретические (трендовые) уровни, рассчитанные по этому уравнению, а в итоге 9-го столбца – остатки по формуле (98). Для иллюстрации построим график эмпирических и трендовых уровней – рис. 19.

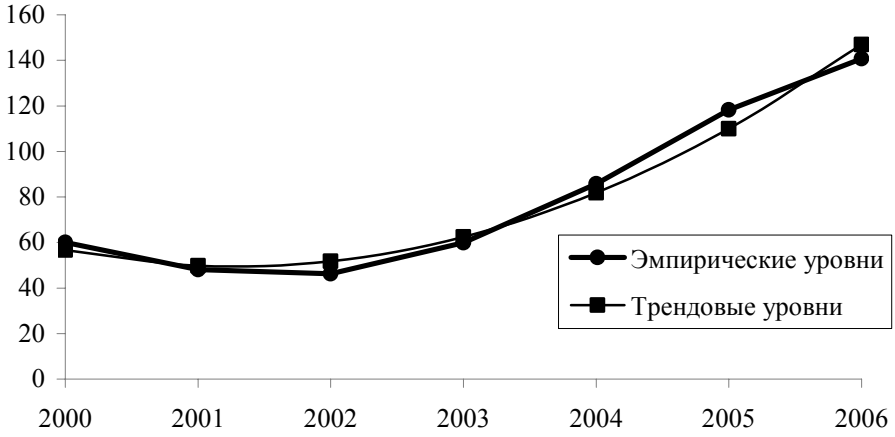


Рис. 19. Эмпирические и трендовые уровни СВТ России

Анализируя рис. 19, то есть сравнивая эмпирические и теоретические уровни, отмечаем, что они почти полностью совпадают, значит парабола 2-го порядка – вполне адекватная функция для отражения основной тенденции (тренда) СВТ России за 2000-2006 годы.

Равенство (103) соблюдается (необходимые суммы рассчитаны в трех последних столбцах табл. 37): 9379,140 = 174,079 + 9205,061. Теперь проверим тренд на адекватность по формуле (102): F_P = 9205,061*4/(174,079*2) = 105,76 > F_T, значит модель адекватна и ее можно использовать для прогнозирования (F_T = 6,94 находим по Приложению 4 в 2-ом столбце [$\nu_1 = k - 1 = 3 - 1 = 2$] и 4-й строке [$\nu_2 = n - k = 4$]).

Спрогнозируем СВТ России на 2007 и 2008 годы с вероятностью 0,95, для чего найдем ошибку аппроксимации по формуле . (105): $\sigma_{\hat{y}} = \sqrt{174{,}079 / (7 - 3)} = 6{,}597$ и найдем коэффициент доверия по распределению Стьюдента по Приложению 2: $t_\alpha = 2{,}4469$ при $\nu = 7 - 1 = 6$.

Прогноз СВТ России на 2007 и 2008 годы с вероятностью 0,95 по формуле , (104):

$Y_{2007} = (62{,}357 + 15{,}061*4 + 4{,}382*4^2) \pm 2{,}4469*6{,}597$ или $176{,}6 < Y_{2007} < 208{,}9$ (млрд. долл.);

$Y_{2008} = (62{,}357 + 15{,}061*5 + 4{,}382*5^2) \pm 2{,}4469*6{,}597$ или $231{,}1 < Y_{2007} < 263{,}4$ (млрд. долл.).

Как видно из полученных прогнозов, доверительный интервал достаточно узок, значит получен достаточно точный прогноз СВТ России на 2006 и 2007 годы. Его надежная оценка имеет принципиальное значение для макроэкономического анализа и прогнозирования, поскольку его величина влияет на общую картину платежного баланса. Так, недооценка положительного сальдо означает недооценку отрицательного сальдо потоков капитала, и наоборот. В то же время потоки капитала увязаны с динамикой внутренних сбережений, что имеет принципиально важное значение для анализа инвестиционного потенциала и прогнозирования инвестиционной активности.

6.8. Контрольные задания

Проанализировать ряд динамики, приведенный в таблице 38 (по данным ФСГС), сделать прогноз на 2007 год.

Таблица 38. Варианты выполнения контрольного задания

Год	**Вариант**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Число заключенных браков, тыс.	Число разводов, тыс.	Среднедушевые денежные доходы населения (в месяц), руб.	Численность студентов, тыс.чел (на начало учеб.года)	Численность профессорско-преподавательского персонала в ВУЗах, тыс.чел. (на начало учеб.года)	Численность лиц, впервые признанных инвалидами, тыс. чел.	Численность осужденных за преступления, , тыс. чел.	Численность населения, тыс.чел. (на начало года)	Число кредитных организаций, зарегистрированных Банком России (на конец года)	Индекс потребительских цен, % (на конец года)
2000	897,3	627,7	2281	4727	307,4	1109	1184	146890	2126	120,2
2001	1001,6	763,5	3062	5427	319,6	1200	1244	146304	2003	118,6
2002	1019,8	853,6	3947	5948	339,6	1184	859	145649	1828	115,1
2003	1091,8	798,8	5170	6456	354,1	1092	767	144964	1668	112,0
2004	979,7	635,8	6410	6884	364,3	1463	794	144168	1518	111,7
2005	1066,4	604,9	8023	7064	387,3	1799	879	143474	1409	110,9
2006	1113,7	640,9	9947	7310	409,0	1443	910	142754	1345	109,0

7. Статистическое изучение взаимосвязей

7.1. Понятие корреляционной зависимости

Один из наиболее общих законов объективного мира – закон всеобщей связи и зависимости между явлениями. Естественно, что, исследуя явления в самых различных областях, статистика неизбежно сталкивается с зависимостями как между количественными, так и между качественными показателями, признаками. Ее задача – обнаружить (выявить) такие зависимости и дать им количественную характеристику.

Среди взаимосвязанных признаков (показателей) одни могут рассматриваться как определенные факторы, влияющие на изменение других (*факторные*), а вторые (*результативные*) – как следствие, результат влияния первых.

Существует 2 вида связи между отдельными признаками: функциональная и стохастическая (статистическая), частным случаем которой является корреляционная.

Связь между двумя переменными x и y называется *функциональной*, если определенному значению переменной x строго соответствует одно или несколько значений другой переменной y, и с изменением значения x значение y меняется строго определенно. Такие связи обычно встречаются в точных науках. Например, известно, что площадь квадрата равна квадрату его стороны ($S = a^2$). Это соотношение характерно для каждого единичного случая (квадрата), это так называемая *жестко детерминированная* связь. Такие связи можно встретить и в области экономических явлений. Например, при простой сдельной оплате труда связь между оплатой труда y и количеством изготовленных изделий x при фиксированной расценке за одну деталь, например 5 руб., легко выразить формулой $y = 5x$.. Для изучения функциональных связей применяется *индексный метод,* который рассматривается в теме 7.

Существуют и иного рода связи, где взаимно действуют многие факторы, комбинация которых приводит к вариации значений результативного признака (показателя) при одинаковом значении факторного признака. Например, при изучении зависимости величины таможенных платежей, поступающих в федеральный бюджет, от количества товаров, перемещаемых через таможенную границу государства, (или от стоимостного товарооборота) последние будут рассматриваться как факторный признак, а величина таможенных платежей – как результативный. Между ними нет жестко детерминированной связи, т.е. при одном и том же количестве перемещенных через таможенную границу товаров (или стоимости товарооборота) величина таможенных платежей, перечисленных разными таможнями будет различной, так как кроме количества товаров, перемещаемых через таможенную границу государства, (или стоимость товарооборота) на величину таможенных платежей влияет много других факторов (различная номенклатура товаров, для которых применяются различные таможенные пошлины, сборы и льготы; различные таможенные режимы перемещения товаров через таможенную границу и др.), комбинация которых вызывает вариацию величины таможенных платежей.

Там, где взаимодействует множество факторов, в том числе и случайных, выявить зависимости, рассматривая единичный случай, невозможно. Такие связи можно обнаружить только при массовом наблюдении как статистические закономерности[39]. Выявленная таким образом связь именуется *стохастической*[40].

Корреляционная связь[41] – понятие более узкое, чем стохастическая связь, это ее частный случай. Именно корреляционные связи являются предметом изучения статистики.

---

[39] Проявление стохастических связей подвержено *действию закона больших чисел*: лишь в достаточно большом числе единиц индивидуальные особенности сгладятся, случайности взаимопогасятся и зависимость, если она имеет существенную силу, проявится достаточно отчетливо

[40] Термин «стохастический» происходит от греч. «stochos» – мишень. Стреляя в мишень, даже хороший стрелок редко попадает в ее центр, выстрелы ложатся в некоторой близости от него. Другими словами стохастическая связь означает приблизительный характер значений признака

[41] Термин «корреляция» ввел в статистику английский биолог и статистик Ф. Гальтон в конце XIX в., под которым понималась «как бы связь», т.е. связь в форме, отличающейся от функциональной. Еще ранее этот термин применил

Корреляционная связь – это связь, проявляющаяся при большом числе наблюдений в виде определенной зависимости между средним значением результативного признака и признаками-факторами. Другими словами, корреляционную связь условно можно рассматривать как своего рода функциональную связь средней величины одного признака (результативного) со значением другого (или других). При этом, если рассматривается связь средней величины результативного показателя y с одним признаком-фактором x, корреляция называется *парной*, а если факторных признаков 2 и более ($x_1, x_2, \ldots, x_m$) – *множественной*[42].

По характеру изменений x и y в парной корреляции различают *прямую* и *обратную* связь. При прямой связи значения обоих признаков изменяются в одном направлении, т.е. с увеличением (уменьшением) значений x увеличиваются (уменьшаются) и значения y. При обратной связи значения факторного и результативного признаков изменяются в разных направлениях.

Изучение корреляционных связей сводится в основном к решению следующих задач:

1) выявление наличия (отсутствия) корреляционной связи между изучаемыми признаками;
2) измерение тесноты связи между двумя (и более) признаками с помощью специальных коэффициентов (эта часть исследования именуется корреляционным анализом);
3) определение уравнения регрессии – математической модели, в которой среднее значение результативного признака y рассматривается как функция одной или нескольких переменных – факторных признаков (эта часть исследования именуется регрессионным анализом).

Общий термин «*корреляционно-регрессионный анализ*» подразумевает всестороннее исследование корреляционных связей (т.е. решение всех трех задач).

Корреляционно-регрессионный анализ находит широкое применение в статистике. Рассмотрим его практическое применение на примере данных таможенной статистики внешней торговли России в 2006 году – таблица 39.

Таблица 39. Величина внешнеторгового оборота и таможенных платежей

Месяц	**Оборот, млрд.долл.**	**Платеж, млрд.руб.**
Январь	27,068	172,17
Февраль	29,889	200,90
Март	34,444	231,83
Апрель	33,158	232,10
Май	37,755	233,40
Июнь	37,554	236,99
Июль	37,299	246,53
Август	40,370	253,62
Сентябрь	37,909	256,43
Октябрь	38,348	261,89
Ноябрь	39,137	259,36
Декабрь	46,298	278,87

В качестве факторного признака x примем стоимостной внешнеторговый товарооборот в млрд. долл. США, а в качестве результативного признака y – величину таможенных платежей в федеральный бюджет в млрд. руб.

француз Ж.Кювье в палеонтологии, где под законом корреляции частей животных он понимал возможность восстановить по найденным в раскопках частям облик всего животного

[42] Множественная корреляция изучается в курсе эконометрики на основе применения компьютерных программ (напр., специальная надстройка к *Excel*, *SPSS* и др.), в курсе статистики изучается только парная корреляция

7.2. Методы выявления и оценки корреляционной связи

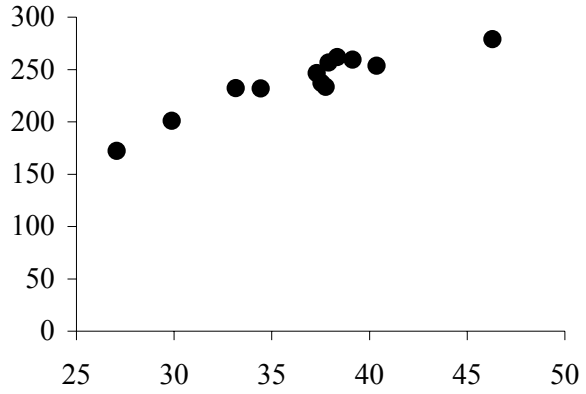
x	*y*
27,068	172,17
29,889	200,90
33,158	232,10
34,444	231,83
37,299	246,53
37,554	236,99
37,755	233,40
37,909	256,43
38,348	261,89
39,137	259,36
40,370	253,62
46,298	278,87

Для выявления наличия и характера корреляционной связи между *двумя* признаками в статистике используется ряд методов.

1. Рассмотрение параллельных данных (значений *x* и *y* в каждой из *n* единиц). Единицы наблюдения необходимо расположить по возрастанию значений факторного признака *x* (как в таблице справа) и затем сравнить с ним (визуально) поведение результативного признака *y*.

В нашей задаче в 6 случаях по мере увеличения значений *x* увеличиваются и значения *y*, а в 5 случаях этого не происходит, поэтому затруднительно говорить о прямой связи между *x* и *y*.

2. Графический метод – это графическое изображение корреляционной зависимости. Для этого, имея *n* взаимосвязанных пар значений *x* и *y* и пользуясь прямоугольной системой координат, каждую такую пару изображают в виде точки на плоскости с координатами *x* и *y*. Совокупность полученных точек представляет собой *корреляционное поле* (рис. 20), а соединяя последовательно нанесенные точки отрезками, получают ломаную линию, именуемую *эмпирической линией регрессии* (рис. 21).



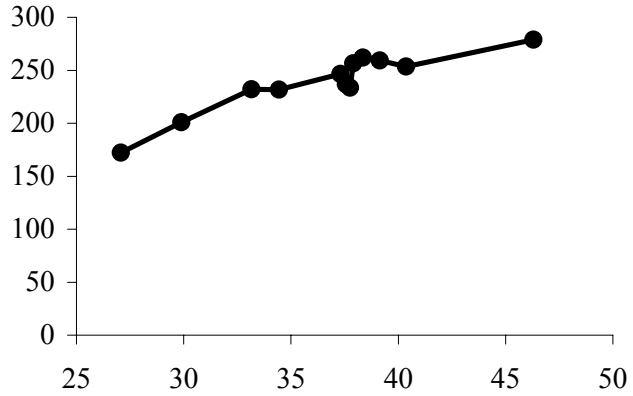


Рис. 20. Корреляционное поле Рис. 21. Эмпирическая линия регрессии

Визуально анализируя график, можно предположить характер зависимости между признаками *x* и *y*. В нашей задаче эмпирическая линия регрессии (рис.21) похожа на восходящую прямую, что позволяет выдвинуть гипотезу о наличии прямой зависимости между величиной стоимостного внешнеторгового товарооборота и величиной таможенных платежей в федеральный бюджет.

3. Метод аналитических группировок используется при большом числе наблюдений для выявления корреляционной связи между двумя количественными признаками. Чтобы выявить наличие корреляционной связи между двумя признаками, проводится группировка

единиц совокупности по факторному признаку x и для каждой выделенной группы рассчитывается среднее значение результативного признака $\overline{y}_j$. Если результативный признак y зависит от факторного x, то в изменении среднего значения $\overline{y}_j$ будет прослеживаться определенная закономерность. Примером такой группировки могут служить данные об издержках обращения предприятий оптовой торговли с различным товарооборотом (см. табл. 40).

Таблица 40. Условные пример аналитической группировки

Оптовый товарооборот, млн.руб.	Количество предприятий	Издержки обращения, % к оптовому товарообороту
менее 25	9362	46,0
26-50	3633	26,5
51-100	3618	24,4
101-200	3261	23,0
201-500	3031	17,6
более 501	3100	16,9

В последнем столбце табл. 40 приведены средние величины, рассчитанные на основе индивидуальных данных об издержках отдельных предприятий каждой группы. Данные таблицы 40 свидетельствуют, что чем крупнее товарооборот, тем меньше издержки обращения. Таким образом, с помощью простой аналитической группировки можно выявить наличие зависимости между рассматриваемыми показателями: объемом товарооборота как показателем размера предприятий и средним уровнем издержек обращения.

4. Метод корреляционных таблиц предполагает комбинационное распределение единиц совокупности по двум количественным признакам. Такая таблица строится по типу «шахматной», т.е. в подлежащем (строках) таблицы выделяются группы по факторному признаку x, а в сказуемом (столбцах) – по результативному y (или наоборот), а в клетках таблицы на пересечении x и y показано число случаев совпадения каждого значения x с соответствующим значением y. Общий вид такой таблицы показан на условном распределении 40 единиц по признакам x и y, где x – стаж работы, y – производительность труда (число изделий, вырабатываемых в час одним рабочим) – таблица 41. Среднее значение по группам определяется по средней арифметической взвешенной по серединам группировочных интервалов.

Таблица 41. Условные корреляционной таблицы

Значение признака x_j	Значение признака y_i				Итого	Среднее значение по группам $\overline{y}_j$
	менее 7,5	7,5-12,5	12,5-17,5	более 17,5		
менее 2	1	3	–	–	4	8,75
2 – 4	2	3	7	–	12	12,08
4 – 6	–	3	9	4	16	15,31
6 – 8	–	–	5	3	8	16,87
Итого	3	9	21	7	40	14,00

Как видно из таблицы 41, по мере увеличения значений x итоговые групповые средние $\overline{y}_j$ тоже увеличиваются от группы к группе, что свидетельствует о том, что между x и y существует корреляционная связь. О наличии и направлении связи можно судить и по «внешнему виду» таблицы, т.е. по расположению в ней частот: если частоты расположены в клетках таблицы беспорядочно, то это чаще всего свидетельствует об отсутствии связи между группировочными признаками (или о незначительной зависимости); если частоты сконцентрированы ближе к одной из диагоналей и центру таблицы, образуя своего рода эллипс, то это почти всегда свидетельствует о наличии зависимости между x и y, близкой к

линейной. Расположение по диагонали из верхнего левого угла в нижний правый свидетельствует о прямой линейной связи, а из нижнего левого угла в верхний правый – об обратной.

На основе аналитических группировок и корреляционных таблиц можно не только выявить наличие зависимости между двумя коррелируемыми показателями, но и измерить тесноту этой связи, в частности, с помощью *эмпирического корреляционного отношения*.

$$\eta_{ЭМП} = \sqrt{\frac{Д_м}{Д_{общ}}}, \qquad (118)$$

$$Д_м = \frac{\sum_{1}^{m}(y_j - \bar{y})^2 f_j}{\sum f_j}, \qquad (119)$$

$$Д_{общ} = \frac{\sum_{1}^{k}(y_i - \bar{y})^2 f_i}{\sum f_i}. \qquad (120)$$

где m – число групп по факторному признаку x;

k – число групп по результативному признаку y;

$\bar{y}_j$ – средние значения результативного признака по группам;

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

y_i – индивидуальные значения результативного признака;

$f_j = f_x$ – частота в j-й группе x;

$f_i = f_y$ – частота в i-й группе y.

Рассчитаем это отношение для нашего примера (таблица 41):

$$\bar{y} = \frac{\sum y_i f_i}{\sum f_i} = (5*3+10*9+15*21+20*7)/40 = 14$$

$$Д_м = \frac{(8{,}75-14)^2*4 + (12{,}08-14)^2*12 + (15{,}31-14)^2*16 + (16{,}87-14)^2*8}{40} = 6{,}19599;$$

$$Д_{общ} = \frac{(5-14)^2*3 + (10-14)^2*9 + (15-14)^2*21 + (20-14)^2*7}{40} = 16{,}5;$$

$$\eta_{ЭМП} = \sqrt{\frac{6{,}19599}{16{,}5}} = 0{,}613.$$

Полученное значение $\eta = 0{,}613$ позволяет утверждать, что существует заметная связь между стажем работы и производительностью труда.

5. Коэффициент корреляции знаков (Фехнера) – простейший показатель тесноты связи, основанный на сравнении поведения отклонений индивидуальных значений каждого признака (x и y) от своей средней величины. При этом во внимание принимаются не величины отклонений ($x_i - \bar{x}$) и ($y_i - \bar{y}$), а их знаки («+» или «–»). Определив знаки отклонений от средней величины в каждом ряду, рассматривают все пары знаков и подсчитывают число их совпадений (C) и несовпадений (H). Тогда коэффициент Фехнера рассчитывается как отношение разности чисел пар совпадений и несовпадений знаков к их сумме, т.е. к общему числу наблюдаемых единиц:

$$K_Ф = \frac{\sum C - \sum H}{\sum C + \sum H}. \qquad (121)$$

Очевидно, что если знаки всех отклонений по каждому признаку совпадут, то $K_Ф$=1, что характеризует наличие прямой связи. Если все знаки не совпадут, то $K_Ф$=–1 (обратная связь). Если же $\sum C = \sum H$, то $K_Ф$=0. Итак, как и любой показатель тесноты связи, коэффициент Фехнера может принимать значения от 0 до ± 1. Однако, если $K_Ф$=1, то это ни в коей мере нельзя воспринимать как свидетельство функциональной зависимости между x и y.

Средние значения факторного и результативного признаков определяем по формуле средней арифметической простой (10):

$$\bar{x} = \frac{\sum x}{n} = \frac{439{,}229}{12} = 36{,}602; \quad \bar{y} = \frac{\sum y}{n} = \frac{2864{,}09}{12} = 238{,}674.$$

В двух последних столбцах таблицы 42 приведены знаки отклонений каждого x и y от своей средней величины. Число совпадений знаков – 10, а несовпадений – 2, тогда определяем коэффициент корреляции знаков (Фехнера) по формуле (121):

$$K_{\Phi}=\frac{10-2}{10+2}=\frac{8}{12}=\frac{2}{3}=0{,}667.$$

Таблица 42. Вспомогательная таблица для расчета коэффициента Фехнера

№ п/п	x	y	$x-\bar{x}$	$y-\bar{y}$
1	27,068	172,17	–	–
2	29,889	200,90	–	–
3	33,158	232,10	–	–
4	34,444	231,83	–	–
5	37,299	246,53	+	+
6	37,554	236,99	+	–
7	37,755	233,40	+	–
8	37,909	256,43	+	+
9	38,348	261,89	+	+
10	39,137	259,36	+	+
11	40,370	253,62	+	+
12	46,298	278,87	+	+
Итого	439,229	2864,09		

Обычно такое значение показателя тесноты связи характеризует заметную прямую зависимость между x и y, однако, следует иметь в виду, что поскольку K_{Φ} зависит только от знаков и не учитывает величину самих отклонений x и y от их средних величин, то он практически характеризует не столько тесноту связи, сколько ее наличие и направление.

6. Линейный коэффициент корреляции – самый популярный измеритель тесноты линейной связи между двумя количественными признаками x и y. Он основан на предположении, что при *полной независимости*[43] *признаков* x и y отклонения значений факторного признака от средней ($x-\bar{x}$) носят случайный характер и должны случайно сочетаться с различными отклонениями ($y-\bar{y}$). При наличии значительного перевеса совпадений или несовпадений таких отклонений делается предположение о наличии связи между x и y.

В отличие от K_{Φ} в линейном коэффициенте корреляции учитываются не только знаки отклонений от средних величин, но и значения самих отклонений, выраженные для сопоставимости в единицах среднего квадратического отклонения t:

$$t_x=\frac{x-\bar{x}}{\sigma_x} \qquad \text{и} \qquad t_y=\frac{y-\bar{y}}{\sigma_y}.$$

Линейный коэффициент корреляции r представляет собой среднюю величину из произведений нормированных отклонений для x и y:

$$r=\frac{\sum\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right)}{n}=\frac{\sum t_x t_y}{n}, \quad (122) \qquad \text{или} \qquad r=\frac{\sum(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}. \quad (123)$$

Числитель формулы (123), деленный на n, представляющий собой среднее произведение отклонений значений двух признаков от их средних значений, называется

[43] При измерении тесноты связи между *рядами динамики* это равнозначно отсутствию автокорреляции между уровнями ряда, т.е. прежде чем оценивать тесноту связи между рядами динамики, необходимо проверить каждый ряд на автокорреляцию – см. методические указания

коэффициентом ковариации – это мера совместной вариации факторного x и результативного y признаков:

$$\mathrm{cov}(x,y)=\frac{\sum(x-\bar{x})(y-\bar{y})}{n}=\overline{(x-\bar{y})(y-\bar{y})} \qquad (124)$$

Недостатком коэффициента ковариации является то, что он не нормирован, в отличие от линейного коэффициента корреляции. Очевидно, что линейный коэффициент корреляции представляет собой частное от деления ковариации между x и y на произведение их средних квадратических отклонений:

$$r=\frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y}. \qquad (125)$$

Путем несложных математических преобразований[44] можно получить и другие модификации формулы линейного коэффициента корреляции, например:

$$r=\frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_x\sigma_y}, \qquad (126)$$

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}, \qquad (127)$$

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{\left[n\sum x^2-(\sum x)^2\right]\left[n\sum y^2-(\sum y)^2\right]}}, \qquad (128)$$

$$r=\frac{\sum xy-\sum x\frac{\sum y}{n}}{\sqrt{\left[\sum x^2-\frac{(\sum x)^2}{n}\right]\left[\sum y^2-\frac{(\sum y)^2}{n}\right]}}. \qquad (129)$$

Линейный коэффициент корреляции может принимать значения от –1 до +1, причем знак определяется в ходе решения. Например, если $\overline{xy}>\bar{x}\bar{y}$, то r по формуле (126) будет положительным, что характеризует прямую зависимость между x и y, в противном случае ($r<0$) – обратную связь. Если $\overline{xy}=\bar{x}\bar{y}$, то $r=0$, что означает отсутствие линейной зависимости между x и y, а при $r=1$ – функциональная зависимость между x и y. Следовательно, всякое промежуточное значение r от 0 до 1 характеризует степень приближения корреляционной связи между x и y к функциональной. Существует эмпирическое правило (шкала Чэддока) для оценки тесноты связи, представленное в таблице 43.

Таблица 43. Шкала Чэддока

$\lvert r\rvert$	*Теснота связи*
менее 0,1	отсутствует линейная связь
0,1 ÷ 0,3	слабая
0,3 ÷ 0,5	умеренная
0,5 ÷ 0,7	заметная
более 0,7	сильная (тесная)

Таким образом, коэффициент корреляции при линейной зависимости служит как мерой тесноты связи, так и показателем, характеризующим степень приближения корреляционной зависимости между x и y к линейной. Поэтому близость значения r к 0 в одних случаях может означать отсутствие связи между x и y, а в других свидетельствовать о том, что зависимость не линейная.

В нашей задаче для расчета r построим вспомогательную таблицу 44.

Таблица 44. Вспомогательные расчеты линейного коэффициента корреляции

№ п/п	x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	t_x	t_y	$t_x t_y$	$(x-\bar{x})(y-\bar{y})$	xy
1	27,068	172,17	90,905	4422,804	-1,993	-2,408	4,799	634,078	4660,298

[44] Проделать это самостоятельно

№ п/п	x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	t_x	t_y	$t_x t_y$	$(x-\bar{x})(y-\bar{y})$	xy
2	29,889	200,90	45,070	1426,888	-1,403	-1,368	1,919	253,594	6004,700
3	33,158	232,10	11,864	43,220	-0,720	-0,238	0,171	22,644	7695,972
4	34,444	231,83	4,659	46,843	-0,451	-0,248	0,112	14,773	7985,153
5	37,299	246,53	0,485	61,714	0,146	0,284	0,041	5,472	9195,322
6	37,554	236,99	0,906	2,836	0,199	-0,061	-0,012	-1,603	8899,922
7	37,755	233,40	1,328	27,817	0,241	-0,191	-0,046	-6,079	8812,017
8	37,909	256,43	1,707	315,270	0,273	0,643	0,176	23,199	9721,005
9	38,348	261,89	3,047	538,975	0,365	0,841	0,307	40,525	10042,958
10	39,137	259,36	6,424	427,904	0,530	0,749	0,397	52,430	10150,572
11	40,37	253,62	14,195	223,378	0,788	0,541	0,426	56,310	10238,639
12	46,298	278,87	94,004	1615,705	2,027	1,455	2,950	389,722	12911,123
Итого	439,229	2864,09	274,594	9153,353			11,241	1485,066	106317,681

В нашей задаче: $\sigma_x = \sqrt{274{,}594/12} = 4{,}784$; $\sigma_y = \sqrt{9153{,}353/12} = 27{,}618$.

Тогда линейный коэффициент корреляции по формуле (122): r = 11,241/12 = 0,937.
Аналогичный результат получаем по формуле (123): r = 1485,066/(12*4,784*27,618) = 0,937
Или по формуле (126): r = (106317,681/12 – 36,602*238,674) / (4,784*27,618) = 0,937,
Найденное значение свидетельствует о том, что связь между величиной стоимостного внешнеторгового товарооборота и величиной таможенных платежей в федеральный бюджет очень близка к функциональной (сильная по шкале Чэддока).

Проверка коэффициента корреляции на значимость (существенность). Интерпретируя значение коэффициента корреляции, следует иметь в виду, что он рассчитан для ограниченного числа наблюдений и подвержен случайным колебаниям, как и сами значения x и y, на основе которых он рассчитан. Другими словами, как любой выборочный показатель, он содержит случайную ошибку и не всегда однозначно отражает действительно реальную связь между изучаемыми показателями. Для того, чтобы оценить существенность (значимость) самого r и, соответственно, реальность измеряемой связи между x и y, необходимо рассчитать среднюю квадратическую ошибку коэффициента корреляции σ_r. Оценка существенности (значимости) r основана на сопоставлении значения r с его средней квадратической ошибкой: $\frac{|r|}{\sigma_r}$.

Существуют некоторые особенности расчета σ_r в зависимости от числа наблюдений (объема выборки) – n.

1. Если число наблюдений достаточно велико (n>30), то σ_r рассчитывается по формуле . (130):

$$\sigma_r = \frac{1-r^2}{\sqrt{n}}. \qquad (130)$$

Обычно, если $\frac{|r|}{\sigma_r}$>3, то r считается значимым (существенным), а связь – реальной. Задавшись определенной вероятностью, можно определить *доверительные пределы (границы)* $r = (r \pm t\sigma_r)$, где t – коэффициент доверия, рассчитываемый по интегралу Лапласа (см. Приложение 1).

2. Если число наблюдений небольшое (n<30), то σ_r рассчитывается по формуле (131):

$$\sigma_r = \frac{\sqrt{1-r^2}}{\sqrt{n-2}}, \qquad (131)$$

а значимость r проверяется на основе t-критерия Стьюдента, для чего определяется расчетное значение критерия по формуле . (132) и сопоставляется с $t_{ТАБЛ}$.

$$t_{РАСЧ} = \frac{|r|}{\sigma_r} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (132)$$

Табличное значение $t_{ТАБЛ}$ находится по таблице распределения *t*-критерия Стьюдента (см. Приложение 2) при уровне значимости $\alpha=1-\beta$ и числе степеней свободы $v=n-2$. Если $t_{РАСЧ} > t_{ТАБЛ}$, то *r* считается значимым, а связь между *x* и *y* – реальной. В противном случае ($t_{РАСЧ} < t_{ТАБЛ}$) считается, что связь между *x* и *y* отсутствует, и значение *r*, отличное от нуля, получено случайно.

В нашей задаче число наблюдений небольшое, значит, оценивать существенность (значимость) линейного коэффициента корреляции будем по формулам (131) и . (132):

$$\sigma_r = \frac{\sqrt{1-0{,}937^2}}{\sqrt{12-2}} = 0{,}349/3{,}162 = 0{,}110;$$

$$t_{РАСЧ} = \frac{|r|}{\sigma_r} = 0{,}937/0{,}110 = 8{,}482.$$

Из приложения 2 видно, что при числе степеней свободы $v = 12 - 2 = 10$ (в 10-й строке) и вероятности $\beta = 95\%$ (уровень значимости $\alpha = 1 - \beta = 0{,}05$) $t_{табл}=2{,}2281$, а при вероятности 99% ($\alpha=0{,}01$) $t_{табл}=3{,}169$, значит, $t_{РАСЧ} > t_{ТАБЛ}$, что дает возможность считать линейный коэффициент корреляции $r = 0{,}937$ значимым.

7. *Подбор уравнения регрессии*[45] представляет собой математическое описание изменения взаимно коррелируемых величин по эмпирическим (фактическим) данным. Уравнение регрессии должно определить, каким будет среднее значение результативного признака *y* при том или ином значении факторного признака *x,* если остальные факторы, влияющие на *y* и не связанные с *x,* не учитывать, т.е. абстрагироваться от них. Другими словами, уравнение регрессии можно рассматривать как вероятностную гипотетическую функциональную связь величины результативного признака *y* со значениями факторного признака *x*.

Уравнение регрессии можно также назвать *теоретической линией регрессии.* Рассчитанные по уравнению регрессии значения результативного признака называются *теоретическими*. Они обычно обозначаются $\hat{y}_x$ или $\bar{y}_x$ (читается: «игрек, выравненный по *x»)* и рассматриваются как функция от *x*, т.е. $\hat{y}_x = f(x)$.

Найти в каждом конкретном случае тип функции, с помощью которой можно наиболее адекватно отразить ту или иную зависимость между признаками *x* и *y,* — одна из основных задач регрессионного анализа. Выбор теоретической линии регрессии часто обусловлен формой эмпирической линии регрессии; теоретическая линия как бы сглаживает изломы эмпирической линии регрессии. Кроме того, необходимо учитывать природу изучаемых показателей и специфику их взаимосвязей.

Для аналитической связи между *x* и *y* могут использоваться виды уравнений, приведенные в таблице 30 (при условии замены *t* на *x*). Обычно зависимость, выражаемую уравнением прямой, называют *линейной* (или *прямолинейной),* а все остальные — *криволинейными зависимостями*.

Выбрав тип функции (таблица 30), по эмпирическим данным определяют параметры уравнения. При этом отыскиваемые параметры должны быть такими, при которых рассчитанные по уравнению теоретические значения результативного признака $\hat{y}_x$ были бы максимально близки к эмпирическим данным.

---

[45] Термин «регрессия» ввел в статистику Ф. Гальтон, который изучив большое число семей, установил, что в группе семей высокорослыми отцами сыновья в среднем ниже ростом, чем их отцы, а в группе семей с низкорослыми отцами сыновья в среднем выше отцов, т.е. отклонение роста от среднего в следующем поколении уменьшается – регрессирует

Существует несколько методов нахождения параметров уравнения регрессии. Наиболее часто используется *метод наименьших квадратов* (МНК). Его суть заключается в следующем требовании: искомые теоретические значения результативного признака $\hat{y}_x$ должны быть такими, при которых бы обеспечивалась минимальная сумма квадратов их отклонений от эмпирических значений, т.е.

$$S = \sum (y - \hat{y}_x)^2 \to \min .$$

Поставив данное условие, легко определить, при каких значениях a_0, a_1 и т.д. для каждой аналитической кривой эта сумма квадратов отклонений будет минимальной. Данный метод уже использовался нами в теме 6 «Статистическое изучение динамики ВЭД», поэтому, воспользуемся формулой (100) для нахождения параметров теоретической линии регрессии, заменив параметр t на x:

$$\begin{cases} na_0 + a_1 \sum x = \sum y \\ a_0 \sum x + a_1 \sum x^2 = \sum xy \end{cases} \qquad (133)$$

Выразив из первого уравнения системы (133) a_0, получим[46]:

$$a_0 = \frac{\sum y}{n} - a_1 \frac{\sum x}{n} = \bar{y} - a_1 \bar{x} . \qquad (134)$$

Подставив . (134) во второе уравнение системы (133), затем разделив обе его части на n, получим:

$$\left(\bar{y} - a_1 \bar{x}\right) \frac{\sum x}{n} + a_1 \frac{\sum x^2}{n} = \frac{\sum xy}{n} . \qquad (135)$$

Применяя 3 раза формулу средней арифметической, получим:

$$\left(\bar{y} - a_1 \bar{x}\right) \bar{x} + a_1 \overline{x^2} = \overline{xy} . \qquad (136)$$

Раскрыв скобки и перенеся члены без a_1 в правую часть уравнения, выразим a_1:

$$a_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x^2} . \qquad (137)$$

Параметр a_1 в уравнении линейной регрессии называется *коэффициентом регрессии*, который показывает на сколько изменяется значение результативного признака y при изменении факторного признака x на единицу.
Исходные данные и расчеты для нашего примера представим в таблице 45.

Таблица 45. Вспомогательные расчеты для нахождения уравнения регрессии

№ п/п	x	y	x^2	xy	$\hat{y}_x$	$(y - \hat{y}_x)^2$	$(\hat{y}_x - \bar{y})^2$
1	27,068	172,17	732,677	4660,298	187,124	223,612	2657,453
2	29,889	200,90	893,352	6004,700	202,377	2,181	1317,497
3	33,158	232,10	1099,453	7695,972	220,052	145,147	346,774
4	34,444	231,83	1186,389	7985,153	227,006	23,274	136,153
5	37,299	246,53	1391,215	9195,322	242,443	16,706	14,202
6	37,554	236,99	1410,303	8899,922	243,821	46,669	26,495
7	37,755	233,40	1425,440	8812,017	244,908	132,441	38,864

[46] Параметры a_0 и a_1 можно получить не только методом подстановки как приводится далее, но и методом определителей 2-го порядка (проделать данное задание самостоятельно)

№ п/п	x	y	x^2	xy	$\hat{y}_x$	$(y-\hat{y}_x)^2$	$(\hat{y}_x-\bar{y})^2$
8	37,909	256,43	1437,092	9721,005	245,741	114,256	49,940
9	38,348	261,89	1470,569	10042,958	248,115	189,761	89,122
10	39,137	259,36	1531,705	10150,572	252,381	48,710	187,871
11	40,370	253,62	1629,737	10238,639	259,048	29,459	415,076
12	46,298	278,87	2143,505	12911,123	291,100	149,580	2748,498
Итого	439,229	2864,09	16351,437	106317,681	2864,115	1121,795	8027,945

По формуле . (137): $a_1 = \dfrac{106317,681/12 - 36,602 * 238,674}{4,784^2} = 5,407.$

По формуле . (134): $a_0 = 238,674 - 5,407*36,602 = 40,767.$

Отсюда получаем уравнение регрессии: $\hat{y}_x = 40,767 + 5,407x$, подставляя в которое вместо x эмпирические значения факторного признака (2-й столбец таблицы 45), получаем выравненные по прямой линии теоретические значения результативного признака $\hat{y}_x$ (6-й столбец таблицы 45)[47]. Для иллюстрации различий между эмпирическими и теоретическими линиями регрессии построим график (рисунок **Ошибка! Источник ссылки не найден.**).

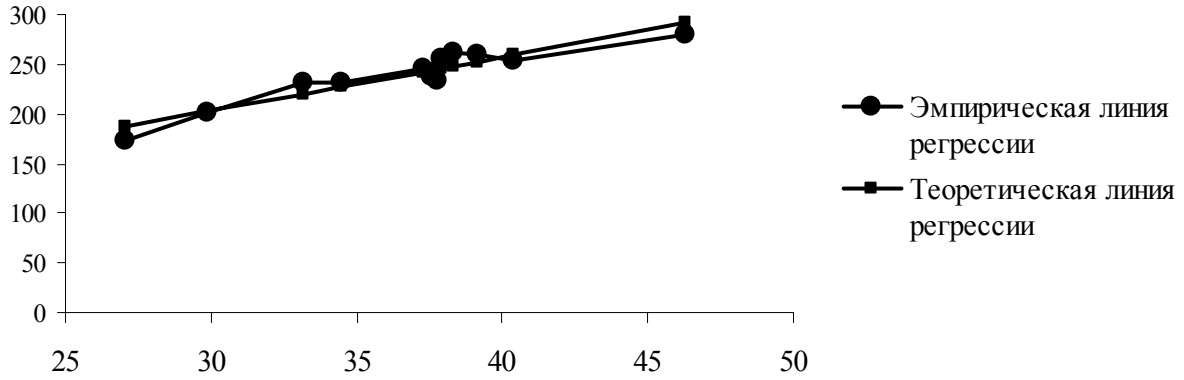


Рис. 22. График эмпирической и теоретической линий регрессии

Из рисунка 22 видно, что небольшие различия между эмпирической и теоретической линиями регрессии существуют, поэтому необходимо *оценить существенность* коэффициента регрессии и уравнения связи, для чего определяют среднюю ошибку параметров уравнения регрессии и сравнивают их с этой ошибкой.

Расчет ошибок параметров уравнения регрессии основан на использовании остаточной дисперсии, характеризующей расхождение (отклонение) между эмпирическими и теоретическими значениями результативного признака. Для линейного уравнения регрессии ($\hat{y}_x = a_0 + a_1 x$) средние ошибки параметров a_1 и a_2 определяются по формулам (138) и (139) соответственно:

$$\mu_{a_0} = \frac{\sigma_{OCT}}{\sqrt{n-2}}, \qquad (138) \qquad \mu_{a_0} = \frac{\sigma_{OCT}}{\sigma_x\sqrt{n-2}}, \qquad (139) \qquad \sigma_{OCT} = \sqrt{\frac{\sum(y-\hat{y}_x)^2}{n}}. \qquad (140)$$

Значимость параметров проверяется путем сопоставления его значения со средней ошибкой. Обозначим это соотношение как t:

$$t_{a_i} = \frac{a_i}{\mu_{a_i}}, \qquad (141)$$

[47] Сумма эмпирических (2864,09) и выравненных по прямой линии (2864,115) значений должна совпадать, но в нашем случае этого не происходит из-за округлений расчетов до 3-х знаков после запятой

При большом числе наблюдений (n>30) параметр a_i считается значимым, если t_{a_i} >3.

Если выборка малая (n<30), то значимость параметра a_i проверяется путем сравнения с табличным значения t-критерия Стьюдента при числе степеней свободы $v=n$-2 и заданном уровне значимости α (Приложение 2). Если рассчитанное по формуле , (141) значение больше табличного, то параметр считается значимым.

В нашем примере по формуле (140): $\sigma_{ОСТ} = \sqrt{\frac{1121{,}795}{12}} = 9{,}669.$

Находим среднюю ошибку параметра a_0 по формуле (138): $\mu_{a_0} = \frac{9{,}669}{\sqrt{12-2}} = 3{,}06.$

Теперь находим среднюю ошибку параметра a_1 по формуле (139): $\mu_{a_0} = \frac{9{,}669}{4{,}784\sqrt{12-2}} = 0{,}639.$

Теперь по формуле , (141) для параметра a_0: $t_{a_0} = \frac{40{,}767}{3{,}06} = 13{,}3.$

И по той же формуле для параметра a_1: $t_{a_0} = \frac{5{,}407}{0{,}639} = 8{,}46.$

Так как выборка малая, то задавшись стандартной значимостью α=0,05 находим в 10-й строке Приложения 2 табличное значение t_α=2,23, которое значительно меньше полученных значений 13,3 и 8,46, что свидетельствует о значимости обоих параметров уравнения регрессии.

Наряду с проверкой значимости отдельных параметров осуществляется *проверка значимости уравнения регрессии* в целом или, что то же самое, проверка адекватности модели с помощью критерия Фишера по Приложению 4. Данный метод уже использовался нами для проверки адекватности уравнения тренда в предыдущей теме, поэтому воспользовавшись формулой (102) в нашем примере получим[48]:

$$F_P = \frac{(12-2)8027{,}945}{(2-1)1121{,}795} = 71{,}56$$

Сравнивая расчетное значение критерия Фишера F_p = 71,56 с табличным F_m = 4,96, определяемое по Приложению 4 при числе степеней свободы $v_1 = k - 1 = 2 - 1 = 1$ и $v_2 = n - k = 12 - 2 = 10$ (т.е. 1-й столбец и 10-я строка) и стандартном уровне значимости α=0,05, можно сделать вывод, что уравнение регрессии значимо.

8. Коэффициент эластичности показывает, на сколько процентов изменяется в среднем результативный признак y при изменении факторного признака x на 1%. Он рассчитывается на основе уравнения регрессии:

$$Э = \frac{\partial \hat{y}_x}{\partial x}\frac{x}{\hat{y}_x}, \qquad (142)$$

где $\frac{\partial \hat{y}_x}{\partial x}$ – первая производная уравнения регрессии y по x.

Коэффициент эластичности – величина переменная, т.е. изменяется с изменением значений фактора x. Так, для линейной зависимости $\hat{y}_x = a_0 + a_1 x$:

$$Э = a_1 \frac{x}{a_0 + a_1 x}. \qquad (143)$$

Применительно к рассмотренному уравнению регрессии, выражающему зависимость величины таможенных платежей в федеральный бюджет от величины стоимостного

[48] В числителе – сумма последнего столбца, а в знаменателе – сумма предпоследнего столбца таблицы 45

внешнеторгового оборота ($\hat{y}_x = 40{,}767 + 5{,}407x$), коэффициент эластичности по формуле (143): $Э = \frac{5{,}407x}{40{,}767 + 5{,}407x}$.

Подставляя в данное выражение разные значения x, получаем и разные значения Э. Так, например, при $x = 40$ коэффициент эластичности $Э = \frac{5{,}407 * 40}{40{,}767 + 5{,}407 * 40} = 0{,}84$, а при $x =$ 50 соответственно $Э = \frac{5{,}407 * 50}{40{,}767 + 5{,}407 * 50} = 0{,}87$ и т.д. Это значит, что при увеличении внешнеторгового товарооборота x с 40 до 40,4 млрд.долл. (т.е. на 1%), величина таможенных платежей возрастет в среднем на 0,84% прежнего уровня; при увеличении x с 50 до 50,5 млрд.долл. (т.е. на 1%) y возрастет на 0,87% и т.д.

9. Теоретическое корреляционное отношение *как универсальный показатель тесноты связи.* Измерить тесноту связи между коррелируемыми величинами – значит определить, насколько вариация результативного признака обусловлена вариацией факторного (факторных) признака. Ранее были рассмотрены показатели, с помощью которых можно выявить наличие корреляционной связи между двумя признаками x и y и измерить тесноту этой связи. Наряду с ними существует универсальный показатель – *корреляционное отношение* (или коэффициент корреляции по Пирсону), применимое ко всем случаям корреляционной зависимости независимо от формы этой связи. Следует различать эмпирическое и теоретическое корреляционное отношение. *Эмпирическое корреляционное отношение* рассчитывается на основе правила сложения дисперсий как корень квадратный из отношения межгрупповой дисперсии к общей дисперсии, т.е.

$$\eta_{эмп} = \sqrt{\frac{\delta^2}{\sigma^2}}. \qquad (144)$$

Теоретическое корреляционное отношение $\eta_{теор}$ определяется на основе выравненных (теоретических) значений результативного признака $\bar{y}_x$, рассчитанных по уравнению регрессии. $\eta_{теор}$ представляет собой относительную величину, получаемую в результате сравнения среднего квадратического отклонения в ряду теоретических значений результативного признака со средним квадратическим отклонением в ряду эмпирических значений. Если обозначить дисперсию эмпирического ряда игреков через σ_y^2, а теоретического ряда – δ^2, то каждая из них выразится формулами

$$\sigma_y^2 = \frac{\sum (y_i - \bar{y})^2}{n}, \qquad \delta^2 = \frac{\sum (\bar{y}_x - \bar{y})^2}{n}.$$

Сравнивая вторую дисперсию с первой, получим *теоретический коэффициент детерминации*:

$$\eta^2_{теор} = \frac{\delta^2}{\sigma_y^2} = \frac{\sum (\bar{y}_x - \bar{y})^2}{\sum (y_i - \bar{y})^2}, \qquad (145)$$

который показывает, какую долю в общей дисперсии результативного признака занимает дисперсия, выражающая влияние вариации фактора x на вариацию y. Извлекая корень квадратный из коэффициента детерминации, получаем *теоретическое корреляционное отношение*

$$\eta_{теор} = \sqrt{\frac{\delta^2}{\sigma_y^2}} = \sqrt{\frac{\sum (\bar{y}_x - \bar{y})^2}{\sum (y_i - \bar{y})^2}}. \qquad (146)$$

Оно может находиться в пределах от 0 до 1, чем ближе его значение к 1, тем теснее связь между вариацией y и x. Для оценки тесноты связи обычно применяется шкала Чэддока (таблица 43). Корреляционное отношение применимо как для парной, так и для

множественной корреляции независимо от формы связи. В этом смысле его можно назвать универсальным показателем тесноты связи. При линейной зависимости $\eta_{теор} \equiv r$.

Покажем расчет $\eta_{теор}$ на условном примере. Исходные данные и расчет дополнительных показателей приведен в таблице 46.

Таблица 46. Исходные данные и вспомогательные расчеты для нахождения теоретического корреляционного отношения

Внесено удобрений, т/га	Урожайность		$y-\bar{y}$	$(y-\bar{y})^2$	$\bar{y}_x-\bar{y}$	$(\bar{y}_x-\bar{y})^2$	$y-\bar{y}_x$	$(y-\bar{y}_x)^2$
	фактическая	рассчитанная по уравнению регрессии						
x	y	$\bar{y}_x$						
1	16	16,2	-4	16	-3,8	14,44	-0,2	0,04
2	19	18,5	-1	1	-1,5	2,25	0,5	0,25
3	20	20,4	0	0	0,4	0,16	-0,4	0,16
4	22	21,9	2	4	1,9	3,61	0,1	0,01
5	23	23	3	9	3	9	0	0
15	100	100		30		29,46		0,46

В данном примере общая средняя урожайность: $\bar{y} = \frac{\sum y}{n} = \frac{100}{5} = 20$ (ц/га).

Общая дисперсия: $\sigma_y^2 = \frac{\sum (y_i - \bar{y})^2}{n}$=30/5=6, факторная дисперсия: $\delta^2 = \frac{\sum (\bar{y}_x - \bar{y})^2}{n}$=29,46/5=5,892.

Отсюда теоретическое корреляционное отношение: $\eta_{теор} = \sqrt{\frac{\delta^2}{\sigma_y^2}} = \sqrt{\frac{5,892}{6}}$=0,99. Данное значение характеризует очень тесную зависимость изменения урожайности от изменения количества внесенных удобрений. В нашем примере незначительные расхождения ($30 \neq 29,46+0,46$ – это правило сложения дисперсий) объясняются округлением значений параметров уравнения регрессии и самих $\bar{y}_x$.

7.3. Коэффициенты корреляции рангов

Коэффициенты корреляции рангов – это менее точные, но более простые по расчету непараметрические показатели для измерения тесноты связи между двумя коррелируемыми признаками. К ним относятся коэффициенты Спирмэна (ρ) и Кендэла (τ), основанные на корреляции не самих значений коррелируемых признаков, а их *рангов* – порядковых номеров, присваиваемых каждому индивидуальному значению *x* и *y* (отдельно) в ранжированном ряду. Оба признака необходимо ранжировать (нумеровать) в одном и том же порядке: от меньших значений к большим и наоборот. Если встречается несколько значений *x* (или *y*), то каждому из них присваивается ранг, равный частному от деления суммы рангов (мест в ряду), приходящихся на эти значения, на число равных значений. Ранги признаков *x* и *y* обозначают символами R_x и R_y (иногда N_x и N_y). Суждение о связи между изменениями значений *x* и *y* основано на сравнении поведения рангов по двум признакам параллельно. Если у каждой пары *x* и *y* ранги совпадают, это характеризует максимально тесную связь. Если же наблюдается полная противоположность рангов, т.е. в одном ряду ранги возрастают от 1 до *n*, а в другом – убывают от *n* до 1, это максимально возможная обратная связь. Подходы для оценки тесноты связи у Спирмэна и Кендэла несколько различаются. Для расчета *коэффициента Спирмэна* значения признаков *x* и *y* нумеруют (отдельно) в порядке возрастания от 1 до *n*, т.е. им присваивают определенный ранг (R_x и R_y) – порядковый номер

в ранжированном ряду. Затем для каждой пары рангов находят их разность (обозначается как $d = R_x - R_y$), и квадраты этой разности суммируют.

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\sum d^2}{n^3 - n}, \qquad (147)$$

где d – разность рангов x и y;

n – число наблюдаемых пар значений x и y.

Коэффициент ρ может принимать значения от 0 до ±1. Следует иметь в виду, что, поскольку коэффициент Спирмэна учитывает разность только рангов, а не самих значений x и y, он менее точен по сравнению с линейным коэффициентом. Поэтому его крайние значения (1 или 0) нельзя безоговорочно расценивать как свидетельство функциональной связи или полного отсутствия зависимости между x и y. Во всех других случаях, т.е. когда ρ не принимает крайних значений, он довольно близок к r.

Формула , (147) применима строго теоретически только тогда, когда отдельные значения x (и y), а следовательно, и их ранги не повторяются. Для случая повторяющихся (связанных) рангов есть другая, более сложная формула, скорректированная на число повторяющихся рангов. Однако опыт показывает, что результаты расчетов по скорректированной формуле для связанных рангов мало отличаются от результатов, полученных по формуле для неповторяющихся рангов. Поэтому на практике формула , (147) успешно применяется как для неповторяющихся, так и для повторяющихся рангов.

Коэффициент корреляции рангов Кендэла τ строится несколько по-другому, хотя его расчет также начинается с ранжирования значений признаков x и y. Ранги x (R_x) располагают строго *в порядке возрастания* и параллельно записывают соответствующее каждому R_x значение R_y. Поскольку R_x записаны строго по возрастанию, то ставится задача определить меру соответствия последовательности R_y «правильному» следованию R_x. При этом для каждого R_y последовательно определяют число следующих за ним рангов, превышающих его значение, и число рангов, меньших по значению. Первые («правильное» следование) учитываются как баллы со знаком «+», и их сумма обозначается буквой P. Вторые («неправильное» следование) учитываются как баллы со знаком «–», и их сумма обозначается буквой Q. Очевидно, что максимальное значение P достигается в том случае, если ранги y (R_y) совпадают с рангами x (R_x) и в каждом ряду представляют ряд натуральных чисел от 1 до n. Тогда после первой пары значений $R_x = 1$ и $R_y = 1$ число превышения данных значений рангов составит $(n - 1)$, после второй пары, где $R_x = 2$ и $R_y = 2$, соответственно $(n - 2)$ и т.д. Таким образом, если ранги x и y совпадают и число пар рангов равно n, то

$$P_{\max} = (n-1) + (n-2) + \ldots + 3 + 2 + 1 = \frac{n(n-1)}{2}.$$

Если же последовательность рангов x и y имеет обратную тенденцию по отношению к последовательности рангов x, то Q будет такое же максимальное значение по модулю:

$$|Q_{\max}| = \frac{n(n-1)}{2}.$$

Если же ранги y не совпадают с рангами x, то суммируются все положительные и отрицательные баллы ($S=P+Q$); отношение этой суммы S к максимальному значению одного из слагаемых и представляет собой коэффициент корреляции рангов Кендэла τ, т.е.:

$$\tau = \frac{S}{n(n-1)/2} = \frac{2S}{n(n-1)}. \qquad (148)$$

Формула коэффициента корреляции рангов Кендэла . (148) применяется для случаев, когда отдельные значения признака (как x, так и y) не повторяются и, следовательно, их ранги не объединены. Если же встречается несколько одинаковых значений x (или y), т.е. ранги повторяются, становятся *связанными*, коэффициент корреляции рангов Кендэла определяется по формуле:

$$\tau = \frac{S}{\sqrt{\left(\frac{n(n-1)}{2} - U_x\right)\left(\frac{n(n-1)}{2} - U_y\right)}}, \qquad (149)$$

где S – фактическая общая сумма баллов при оценке +1 каждой пары рангов с одинаковым порядком изменения и –1 каждой пары рангов с обратным порядком изменения;

$U_x = U_y = 0{,}5\sum t(t-1)$ – число баллов, корректирующих (уменьшающих) максимальную сумму баллов за счет повторений (объединений) t рангов в каждом ряду.

Отметим, что случаи следования одинаковых повторяющихся рангов (в любом ряду) оцениваются баллом 0, т.е. они не учитываются при расчете ни со знаком «+», ни со знаком «–».

Преимущества ранговых коэффициентов корреляции Спирмэна и Кендэла: они легко вычисляются, с их помощью можно изучать и измерять связь не только между количественными, но и между качественными (описательными) признаками, ранжированными определенным образом. Кроме того, при использовании ранговых коэффициентов корреляции не требуется знать форму связи изучаемых явлений.

Если число ранжируемых признаков (факторов) больше двух, то для измерения тесноты связи между ними можно использовать предложенный М. Кендэлом и Б. Смитом *коэффициент конкордации* (множественный коэффициент ранговой корреляции):

$$W = \frac{12S}{m^2(n^3 - n)}, \qquad (150)$$

где S — сумма квадратов отклонений суммы m рангов от их средней величины;

m — число ранжируемых признаков;

n — число ранжируемых единиц (число наблюдений).

Формула , (150) применяется для случая, кода ранги по каждому признаку не повторяются. Если же есть связанные ранги, то коэффициент конкордации рассчитывается с учетом числа таких повторяющихся (связанных) рангов по каждому фактору:

$$W = \frac{12S}{m^2(n^3 - n) - m\sum_1^m (t^3 - t)}, \qquad (151)$$

где t – число одинаковых рангов по каждому признаку.

Коэффициент конкордации W может принимать значения от 0 до 1. Однако, необходимо проверить его на существенность (значимость) с помощью критерия χ^2 при отсутствии связанных рангов по формуле (152), а при их наличии – по формуле (153):

$$\chi^2 = \frac{12S}{mn(n-1)}, \qquad (152)$$

$$\chi^2 = 12S \Big/ \left(mn(n-1) - \frac{\sum_1^m (t^3 - t)}{n-1} \right). \qquad (153)$$

Фактическое значение χ^2 сравнивается с табличным, соответствующим принятому уровню значимости α (0,05 или 0,01) и числу степеней свободы $\nu = n - 1$. Если $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$, то W – существенен (значим).

Коэффициент конкордации особенно часто используется в экспертных оценках, например, для того, чтобы определить степень согласованности мнений экспертов о важности того или иного оцениваемого показателя или составить рейтинг отдельных единиц по какому-либо признаку. В формуле , ***(150)*** в этих случаях m означает число экспертов, а n — число ранжируемых единиц (или признаков).

7.4. Особенности коррелирования рядов динамики

Во многих исследованиях приходится изучать динамику нескольких показателей одновременно, т.е. рассматривать параллельно несколько рядов динамики. В этом случае возникает необходимость измерить зависимость между ними, вернее, определить, насколько изменения уровней одного ряда зависят от изменения уровней другого ряда. Эта задача решается путем коррелирования рядов динамики.

Однако при этом возникает следующая проблема: если показатели ряда x и ряда y рассматривать как функцию времени, т.е. $x = f(t)$ и $y = f(t)$, то при однонаправленности их трендов можно получить большое значение коэффициента корреляции между x и y даже тогда, когда они независимы, именно в силу однонаправленности их изменения.

Поэтому, прежде чем коррелировать ряды динамики, необходимо установить путем логического (качественного) анализа, возможна ли связь между исследуемыми показателями x и y. Кроме того, одно из условий корреляции – независимость отдельных значений переменных множества x, так же как и множества y. Для рядов динамики это равнозначно отсутствию *автокорреляции* между уровнями ряда, т.е. отсутствию зависимости между последовательными (соседними) уровнями ряда динамики. Другими словами, прежде чем коррелировать ряды динамики, необходимо проверить каждый ряд на автокорреляцию.

Если исходные фактические уровни ряда, относящиеся к определенному моменту (периоду) времени t, обозначить через y_t, то сдвинутые на один момент (период) уровни обозначают y_{t-1}. Тогда, подставив в формулу коэффициента корреляции (126) значения y_t и y_{t-1}, получим формулу:

$$r_a = \frac{\overline{y_t y_{t-1}} - \bar{y}_t \bar{y}_{t-1}}{\sigma_{y_t} \sigma_{y_{t-1}}}, \qquad (154)$$

а поскольку $\bar{y}_t \cong \bar{y}_{t-1}$ и $\sigma_{y_t} \cong \sigma_{y_{t-1}}$, получим следующие формулы[49] для расчета коэффициента автокорреляции:

$$r_a = \frac{\overline{y_t y_{t-1}} - (\bar{y}_t)^2}{\sigma_{y_t}^2}, \qquad (155) \quad \text{или} \quad r_a = \frac{\sum y_t y_{t-1} - n(\bar{y}_t)^2}{\sum y_t^2 - n(\bar{y}_t)^2}. \qquad (156)$$

Сдвинутый (укороченный) ряд условно дополняют, принимая $y_1 = y_n$ (чтобы сдвинутый ряд не укорачивался и чтобы средний уровень и дисперсия исходного и сдвинутого рядов были одинаковы).

Найденное по формуле (155) или (156)[50] значение коэффициента автокорреляции само по себе еще не говорит о наличии или отсутствии автокорреляции. Его нужно сравнить с *критическим*.

Существуют специальные таблицы, в которых для разного числа членов ряда n и разных уровней значимости α определено критическое значение коэффициента автокорреляции: если найденное по формуле (155) или (156) значение окажется меньше критического, то автокорреляция отсутствует. Одна из таких таблиц, составленная Р. Андерсоном, приведена в Приложении 5.

В нашем примере про внешнеторговый оборот и таможенные платежи проверим оба эти ряда динамики на автокорреляцию с помощью формулы (155), для чего построим таблицу 47.

Таблица 47. Вспомогательные расчеты для проверки на автокорреляцию

Месяц	x_t	x_{t-1}	$x_t x_{t-1}$	x_t^2	y_t	y_{t-1}	$y_t y_{t-1}$	y_t^2
1	27,068	46,298	1253,194	732,677	172,170	278,870	48013,048	29642,509
2	29,889	27,068	809,035	893,352	200,900	172,170	34588,953	40360,810
3	34,444	29,889	1029,497	1186,389	231,830	200,900	46574,647	53745,149

[49] Коэффициент автокорреляции можно рассчитывать либо между соседними уровнями, либо между уровнями, сдвинутыми на другое число единиц времени (временной лаг) m; приведенные формулы с временным лагом m=1 (между соседними уровнями) являются самыми распространенными

[50] Формула (156) является тождественной формуле (155)

Месяц	x_t	x_{t-1}	$x_t x_{t-1}$	x_t^2	y_t	y_{t-1}	$y_t y_{t-1}$	y_t^2
4	33,158	34,444	1142,094	1099,453	232,100	231,830	53807,743	53870,410
5	37,755	33,158	1251,880	1425,440	233,400	232,100	54172,140	54475,560
6	37,554	37,755	1417,851	1410,303	236,990	233,400	55313,466	56164,260
7	37,299	37,554	1400,727	1391,215	246,530	236,990	58425,145	60777,041
8	40,370	37,299	1505,761	1629,737	253,620	246,530	62524,939	64323,104
9	37,909	40,370	1530,386	1437,092	256,430	253,620	65035,777	65756,345
10	38,348	37,909	1453,734	1470,569	261,890	256,430	67156,453	68586,372
11	39,137	38,348	1500,826	1531,705	259,360	261,890	67923,790	67267,610
12	46,298	39,137	1811,965	2143,505	278,870	259,360	72327,723	77768,477
Итого	439,229	439,229	16106,951	16351,437	2864,090	2864,090	685863,823	692737,647

Теперь по формуле (155) для ряда x: $r_a = \dfrac{16106{,}951 - 12 * 36{,}602^2}{16351{,}437 - 12 * 36{,}602^2} = 0{,}111.$

Аналогично по формуле (155) для ряда y: $r_a = \dfrac{685863{,}823 - 12 * 238{,}674^2}{692737{,}647 - 12 * 238{,}674^2} = 0{,}249.$

По таблице Приложения 5 определяем критическое (предельное) значение коэффициента корреляции для числа уровней n = 12 и уровне значимости α = 0,05. Оно равно 0,348. Оба рассчитанных значения оказались меньше критического, значит автокорреляция между уровнями в обоих рядах динамики отсутствует, следовательно, можно коррелировать уровни x и y.

Исключение автокорреляции в рядах динамики. Если между уровнями ряда (при коррелировании рядов динамики) существует автокорреляция, она должна быть устранена.

Есть несколько способов исключения автокорреляции в рядах динамики. Наиболее простой – *коррелирование отклонений от выравненных уровней*. Для этого каждый ряд динамики выравнивают по определенной для него аналитической формуле (т.е. находят $\hat{x}_t$ и $\hat{y}_t$)[51], затем из эмпирических уровней вычитают выравненные (т.е. находят остаточные величины[52], не описываемые уравнением тренда: $d_x = x - \hat{x}_t$ и $d_y = y - \hat{y}_t$). Так как остаточные величины могут содержать автокорреляцию (например, в случае недостаточно точно подобранного уравнения тренда), необходимо убедиться, что между ними автокорреляция отсутствует. Лишь после этого можно определять тесноту связи между d_x и d_y. Формулу коэффициента корреляции между остаточными величинами можно записать в следующем виде:

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}. \qquad (157)$$

7.5. Показатели тесноты связи между качественными признаками

Метод корреляционных таблиц применим не только к количественным, но и к описательным (качественным) признакам, взаимосвязи между которыми часто приходится изучать при проведении различных социологических исследований путем опросов или анкетирования. В этом случае такие таблицы называют *таблицами сопряженности*. Они могут иметь различную размерность. Простейшая размерность – 2x2 (таблица «четырех полей»), когда по альтернативному признаку («да» – «нет», «хорошо» – «плохо» и т.д.) выделяются 2 группы. В таблице 48 приведены условные данные о распределении 500 опрошенных человек по двум показателям: наличие (отсутствии) у них прививки против гриппа и факт заболевания (незаболевания) гриппом во время его эпидемии.

---

[51] См. тему 5 «Ряды динамики», метод аналитического выравнивания

[52] Остаточные величины обычно обозначают ε_t, но для того, чтобы различать их для разных рядов динамики x и y, приняты обозначения d_x и d_y

Таблица 48. Распределение 500 опрошенных человек

Группа лиц	Число лиц		
	заболевших гриппом	не заболевших гриппом	Итого
Сделавших прививку	30 (*a*)	270 (*b*)	300
Не сделавших прививку	120 (*c*)	80 (*d*)	200
Итого	150	350	500

Нетрудно заметить, что среди сделавших прививку подавляющее большинство (270 из 300, или 90%) не заболели гриппом, а среди не сделавших большая часть заболела (120 из 200, или 60%). Таким образом, можно предположить, что прививка положительно влияет на предупреждение заболевания; другими словами, можно предположить, что распределение в таблице (*a, b, c, d*) не случайно и существует стохастическая зависимость между группировочными признаками. Однако выводы о зависимости, сделанные «на глаз», часто могут быть ненадежными (ошибочными), поэтому они должны подкрепляться определенными статистическими критериями, например ***критерием Пирсона*** χ^2. Он позволяет судить о случайности (или неслучайности) распределения в таблицах взаимной сопряженности, а следовательно, и об отсутствии или наличии зависимости между признаками группировки в таблице. Чтобы воспользоваться критерием Пирсона χ^2, в таблице взаимной сопряженности наряду с эмпирическими частотами записывают теоретические частоты, рассчитываемые исходя из предположения, что распределение внутри таблицы случайно и, следовательно, зависимость между признаками группировки отсутствует. То есть считается, что распределение частот в каждой строке (столбце) таблицы пропорционально распределению частот в итоговой строке (столбце). Поэтому теоретические частоты по строкам (столбцам) рассчитывают пропорционально распределению единиц в итоговой строке (столбце).

Так, в нашем примере в итоговой строке число заболевших 150 из 500, т.е. их доля – 30%, а доля не заболевших – 70%. Следовательно, теоретические частоты в первой строке для заболевших составят 30% от 300, т.е. 0,3*300=90, а для не заболевших – 0,7*300=210. По второй строке произведем аналогичные расчеты и их результаты занесем в таблицу в скобках.

Таблица 49. Эмпирические и теоретические частоты

Группа	I (да)	II (нет)	∑
I (да)	30 (90)	270 (210)	300
II (нет)	120 (60)	80 (140)	200
∑	150	350	500

На сопоставлении эмпирических и теоретических частот и основан *критерий Пирсона* χ^2, рассчитываемый по формуле (44):

$$\chi^2 = \frac{(30-90)^2}{90} + \frac{(270-210)^2}{210} + \frac{(120-60)^2}{60} + \frac{(80-140)^2}{140} = 142{,}85 .$$

Рассчитанное (фактическое) значение χ^2 сопоставляют с табличным (критическом), определяемым по таблице Приложения 3 для заданного уровня значимости α и числа степеней свободы $\nu = (k_1 - 1)(k_2 - 1)$, где k_1 и k_2 – число групп по одному и второму признакам группировки (число строк и число столбцов в таблице).

В рассматриваемом примере ν=(2-1)(2-1)=1, а приняв уровень значимости α=0,01, по таблице Приложения 3 находим $\chi^2_{табл}$=6,63. Поскольку рассчитанное значение $\chi^2 > \chi^2_{табл}$, значит существует стохастическая зависимость между рассматриваемыми показателями. При независимости признаков част**о**ты теоретического и эмпирического распределений совпадают, а значит χ^2=0. Чем больше различия между теоретическими и эмпирическими частотами, тем больше значение χ^2 и вероятность того, что оно превысит критическое табличное значение, допустимое для случайных расхождений. Аналогично рассчитываются теоретические частоты и χ^2 в таблицах большей размерности.

В корреляционном анализе недостаточно лишь выявить тем или иным методом наличие связи между исследуемыми показателями. Теснота такой связи может быть различной, поэтому весьма важно ее измерить, т.е. определить меру связи в каждом конкретном случае. В статистике для этой цели разработан ряд показателей (коэффициентов), используемых как для количественных, так и для качественных признаков.

Для измерения тесноты связи между группировочными признаками в таблицах взаимной сопряженности могут быть использованы такие показатели, как коэффициент ассоциации и контингенции (для «четырехклеточных таблиц»), а также коэффициенты взаимной сопряженности Пирсона и Чупрова (для таблиц любой размерности).

Применительно к таблице «*четырех полей»,* частоты которых можно обозначить через *a, b, c, d*, *коэффициент ассоциации* (Д. Юла) выражается формулой . (158):

$$K_{AC} = \frac{ad - bc}{ad + bc}. \qquad (158)$$

Его существенный недостаток: если в одной из четырех клеток отсутствует частота (т.е. равна 0), то $|K_{AC}| = 1$, и тем самым преувеличена мера действительной связи.

Чтобы этого избежать, предлагается (К. Пирсоном) другой показатель – *коэффициент контингенции*[53]:

$$K_{КОНТ} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \qquad (159)$$

Рассчитаем коэффициенты . (158) и . (159) для нашего примера (таблица 48):

$$K_{AC} = \frac{30*80 - 270*120}{30*80 + 270*120} = -0{,}862;$$

$$K_{КОНТ} = \frac{30*80 - 270*120}{\sqrt{300*200*150*350}} = -0{,}534$$

Связь считается достаточно значительной и подтвержденной, если $|K_{AC}|$>0,5 или $|K_{КОНТ}|$>0,3.

Поэтому в нашем примере оба коэффициента характеризуют достаточно большую обратную зависимость между исследуемыми признаками.

Теснота связи между 2 и более признаками измеряется с помощью *коэффициентов взаимной сопряженности Пирсона* (160) или *Чупрова* (161), рассчитываемых на основе показателя χ^2 :

$$K_{П} = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \qquad (160) \qquad K_{Ч} = \sqrt{\frac{\chi^2}{n\sqrt{(k_1 - 1)(k_2 - 1)}}} \qquad (161)$$

В нашем примере $K_{П} = \sqrt{\frac{142{,}85}{142{,}85 + 500}} = 0{,}47$. Рассчитывать коэффициент Чупрова для таблицы «четырех полей» не рекомендуется, так как при числе степеней свободы *v*=(2-1)(2-1)=1 он будет больше коэффициента Пирсона (в нашем примере $К_Ч$=0,54). Для таблиц же большей размерности всегда $К_Ч < К_П$.

7.6. Множественная корреляция

При решении практических задач исследователи сталкиваются с тем, что корреляционные связи не ограничиваются связями между двумя признаками: результативным *y* и факторным *x*. В действительности результативный признак зависит от нескольких факторных. Например, инфляция тесно связана с динамикой потребительских цен, розничным товарооборотом, численностью безработных, объемами экспорта и импорта,

[53] По значению коэффициент контингенции всегда меньше коэффициента ассоциации

курсом доллара, количеством денег в обращении, объемом промышленного производства и другими факторами.

В условиях действия множества факторов показатели парной корреляции оказываются условными и неточными. Количественно оценить влияние различных факторов на результат, определить форму и тесноту связи между результативным признаком y и факторными признаками x_1, x_2, …, x_k можно методами множественной (многофакторной) корреляции.

Математически задача сводится к *нахождению аналитического выражения, наилучшим образом описывающего связь факторных признаков с результативным*, т.е. к отысканию функции $\overline{y}_{1,2,...,k} = f(x_1, x_2, ..., x_k)$. Выбрать форму связи довольно сложно. Эта задача на практике основывается на априорном теоретическом анализе изучаемого явления и подборе известных типов математических моделей.

Среди многофакторных регрессионных моделей выделяют *линейные* (относительно независимых переменных) и *нелинейные*. Наиболее простыми для построения, анализа и экономической интерпретации являются многофакторные линейные модели, которые содержат независимые переменные только в первой степени:

$$\overline{y}_x = a_0 + a_1 x_1 + a_2 x_2 + ... + a_k x_k, \qquad (162)$$

где a_0 – свободный член;

$a_1, a_2, ..., a_k$ – коэффициенты регрессии;

$x_1, x_2, ..., x_k$ – факторные признаки.

Если связь между результативным признаком и анализируемыми факторами нелинейна, то выбранная для ее описания нелинейная многофакторная модель (степенная, показательная и т.д.) может быть сведена к линейной путем линеаризации.

Параметры уравнения множественной регрессии, как и парной, рассчитываются методом наименьших квадратов, при этом решается система нормальных уравнений с *(k+1)* неизвестным:

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} x_{i1} + a_2 \sum_{i=1}^{n} x_{i2} + ... + a_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i, \\ a_0 \sum_{i=1}^{n} x_{i1} + a_1 \sum_{i=1}^{n} x_{i1}^2 + a_2 \sum_{i=1}^{n} x_{i1} x_{i2} + ... + a_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} y_i x_{i1}, \\ ..... \\ a_0 \sum_{i=1}^{n} x_{ik} + a_1 \sum_{i=1}^{n} x_{i1} x_{ik} + a_2 \sum_{i=1}^{n} x_{i2} x_{ik} + ... + a_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} y_i x_{ik} m, \end{cases} \qquad (163)$$

где x_{ij} – значение j-го факторного признака в i-м наблюдении;

y_i – значение результативного признака в i-м наблюдении.

Как правило, прежде чем найти параметры уравнения множественной регрессии, определяют и анализируют парные коэффициенты корреляции. При этом систему нормальных уравнений можно видоизменить таким образом, чтобы при вычислении параметров регрессии использовать уже найденные парные коэффициенты корреляции. Для этого в уравнении регрессии заменим переменные *y, x_1, x_2, …, x_k* переменными t_j, полученными следующим образом:

$$t_{iy} = \frac{y_i - \overline{y}}{\sigma_y}, \qquad t_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_{x_j}}. \qquad (i = \overline{1,n},\ \ j = \overline{1,k}).$$

Эта процедура называется *стандартизацией переменных*. В результате осуществляется переход от натурального масштаба переменных x_{ij} к центрированным и нормированным отклонениям t_{ij}. В стандартизированном масштабе среднее значение признака равно 0, а среднее квадратическое отклонение равно 1, т.е. $\overline{t}_j = 0$, $\sigma_{t_j} = 1$. При переходе к

стандартизированному масштабу переменных уравнение множественной регрессии принимает вид

$$t_y = \beta_1 t_1 + \beta_2 t_2 + ... + \beta_k t_k , \qquad (164)$$

где β_j ($j = \overline{1,k}$) – коэффициенты регрессии.

Параметры уравнения множественной регрессии в натуральном масштабе и уравнения регрессии в стандартизированном виде взаимосвязаны:

$$a_j = \frac{\sigma_y}{\sigma_{x_j}} \beta_j \qquad (j = \overline{1,k}) \qquad (165)$$

Нетрудно заметить, что это обычная формула коэффициента регрессии, выраженного через линейный коэффициент корреляции.

Стандартизированные коэффициенты множественной регрессии β_j также вычисляют методом наименьших квадратов, который приводит к системе нормальных уравнений

$$\begin{cases} r_{y1} = \beta_1 + r_{12}\beta_2 + ... + r_{1k}\beta_k , \\ r_{y2} = r_{21}\beta_1 + \beta_2 + ... + r_{2k}\beta_k , \\ ..... \\ r_{yk} = r_{k1}\beta_1 + r_{k2}\beta_2 + ... + \beta_k , \end{cases} \qquad (166)$$

где $r_{yi} = \frac{1}{n}\sum_{i=1}^{n} t_{iy} t_{ij}$ – парный коэффициент корреляции результативного признака *y* с *j*-м факторным;

$r_{jl} = \frac{1}{n}\sum_{i=1}^{n} t_{ij} t_{il}$ – парный коэффициент корреляции *j*-го факторного признака с *l*-м факторным.

После того как получено уравнение множественной регрессии (в стандартизированном или натуральном масштабе), необходимо *измерить тесноту связи между результативным признаком и факторными признаками*. Для измерения степени совокупного влияния отобранных факторов на результативный признак рассчитывается совокупный коэффициент детерминации R^2 и совокупный коэффициент множественной корреляции R – общие показатели тесноты связи многих признаков независимо от формы связи. Приведем несколько формул для их расчета.

1. При линейной форме связи расчет совокупного коэффициента детерминации можно выполнить, используя парные коэффициенты корреляции:

$$R^2_{y,x_1,x_2,...,x_k} = \frac{a_1 r_{y1}\sigma_{x_1} + a_2 r_{y2}\sigma_{x_2} + ... + a_k r_{yk}\sigma_{x_k}}{\sigma_y} , \qquad (167)$$

где $a_1, a_2, ..., a_k$ – параметры уравнения множественной регрессии в натуральном масштабе.

2. Еще легче вычислить совокупный коэффициент детерминации, используя уравнение регрессии в стандартизированном виде:

$$R^2_{y,x_1,x_2,...,x_k} = \beta_1 r_{y1} + \beta_2 r_{y2} + ... + \beta_k r_{yk} . \qquad (168)$$

3. Через соотношение факторной и общей дисперсий (или остаточной и общей дисперсий):

$$R^2_{y,x_1,x_2,...,x_k} = \frac{\delta^2_{фатор}}{\sigma^2_y} , \text{ или } R^2_{y,x_1,x_2,...,x_k} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_y} , \qquad (169)$$

где $\delta^2_{фатор} = \frac{1}{n}\sum_{i=1}^{n} ((\overline{y}_x)_i - \overline{y})^2$ – факторная дисперсия, характеризующая вариацию результативного признака, обусловленную вариацией включенных в анализ факторов; σ^2_y – общая дисперсия результативного признака; $\sigma^2_{ост} = \frac{1}{n}\sum_{i=1}^{n} (y_i - (\overline{y}_x)_i)^2 = \sigma^2_y - \delta^2_{фатор}$ – остаточная

дисперсия, характеризующая отклонения фактических уровней результативного признака y_i от рассчитанных по уравнению множественной регрессии $(\overline{y}_x)_i$.

Совокупный коэффициент множественной корреляции R представляет собой корень квадратный из совокупного коэффициента детерминации R^2. Пределы его изменения: $0 \le R \le 1$. Чем ближе его значение к 1, тем точнее уравнение множественной линейной регрессии отражает реальную связь. Иначе говоря, среди отобранных факторов присутствуют те, которые решающим образом влияют на результативный. Малое значение R можно объяснить тем либо тем, что в уравнение множественной регрессии не включены существенно влияющие на результат факторы, либо тем, что установленная линейная форма зависимости не отражает реальной взаимосвязи признаков. Добиться адекватности модели множественной регрессии эмпирическим данным возможно, соответственно, либо включением в уравнение регрессии дополнительных, ранее не учитываемых факторов, либо построением нелинейной модели множественной регрессии.

Для более глубокого знакомства с темой «Множественная корреляция» необходимо воспользоваться литературой курса «Эконометрика».

7.7. Контрольные задания

На основе исходных данных контрольных заданий по теме 6 (таблица 38) с использованием таблицы 50 проанализировать взаимосвязь между признаками x и y всеми возможными методами, изложенными в теме 7.

Таблица 50. Распределение вариантов для выполнения контрольного задания

Признак	**Вариант**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x (№ варианта темы 6)	1	4	3	6	7	3	3	3	3	2
y (№ варианта темы 6)	2	5	9	8	8	1	2	4	7	10

8. Индексы

8.1. Назначение и виды индексов

Индекс – относительная величина, показывающая во сколько раз уровень изучаемого явления в данных условиях отличается от уровня того же явления в других условиях. Различие условий может проявляться во времени (тогда получается индекс динамики), в пространстве (территориальный индекс), в выборе в качестве базы сравнения планового показателя (индекс выполнения плана) и т.п.

Каждый индекс включает 2 вида данных: оцениваемые данные, которые принято называть *отчетными* и обозначать значком «1», и данные, которые используются в качестве базы сравнения – *базисные*, обозначаемые значком «0».

Индекс, который строится как сравнение обобщенных величин, называется *общим* (*сводным*) и обозначается *I*. Если же сравниваются необобщенные величины, то индекс называется *индивидуальным* и обозначается *i*. Как правило, подстрочно ставится значок, показывающий для оценки какой величины построе индекс. Например, I_q и i_q – это общий и индивидуальный индекс для величины *q*.

В статистическе индексы используются не только для сопоставления уровней изучаемого явления, но и для определения экономической значимости факторов, объясняющих абсолютное различие сравниваемых уровней.

В зависимости от сложности сравниваемых уровней принято выделять 2 типа индексов: индивидуальные и общие.

8.2. Индивидуальные индексы

Относительная величина, получаемая при сравнении уровней, называется *индивидуальным индексом*, если не имеет значения структура изучаемого явления. Индивидуальные индексы обозначаются *i*. Расчет индивидуальных индексов прост: их определяют вычислением отношения двух индексируемых величин, то есть по формуле ,(2).

Например, если уровень товарооборота (Q) в виде суммы выручки от продажи товара в условиях отчетного периода сравнивается с аналогичным показателем базисного периода, то в итоге получаем индивидуальный индекс выручки (170), показывающий во сколько раз изменилась (или сколько процентов составляет) выручка в отчетном периоде по сравнению с базисным:

$$i_Q=Q_1/Q_0. \qquad (170)$$

Разность между числителем и знаменателем формулы (170) представляет собой *абсолютное изменение* выручки (171), показывающее на сколько в денежных единицах (например, рублях) изменилась выручка в отчетном периоде по сравнению с базисным:

$$\Delta Q = Q_1 - Q_0. \qquad (171)$$

Аналогично определяются индивидуальные индексы можно для любого интересующего показателя (производительности, заработной платы, себестоимости и т.д.).

В частности, поскольку сумма выручки определяется ценой товара *p* (от англ. «*price*») и количеством (физическим объемом, или объемом продаж в натуральном выражении) *q* (от англ. «*quantity*») т.е. $Q = qp$, можно определить соответствующие индивидуальные индексы – цены (172) и количества (173):

$$i_p=p_1/p_0, \qquad (172) \qquad i_q=q_1/q_0. \qquad (173)$$

Очевидно, что произведение индивидуальных индексов цены и количества дает индивидуальный индекс выручки (174):

$$i_Q=i_q i_p. \qquad (174)$$

Например, вчера бабушка торговала семечками по 5 руб. за кулёк и всего продала 50 кульков, а сегодня – по 7 руб. и продала 20 кульков. Определим индивидуальный индекс

цены по формуле (172): $i_p = 7/5 = 1{,}4$, то есть бабушка увеличила цену семечек в 1,4 раза, или на 40%. Рассчитаем индивидуальный индекс количества по формуле (173): $i_q = 20/50 = 0{,}4$, то есть количество проданных семечек сегодня составило 40% от вчерашнего, то есть уменьшилось на 60%. Найдем индивидуальный индекс выручки по формуле (174): $i_Q = 0{,}4*1{,}4 = 0{,}56$, то есть выручка сегодня составила 56% от вчерашней, то есть она уменьшилась на 44%. Рассчитав выручку вчера $Q_0 = 50*5 = 250$ (руб.) и сегодня $Q_1 = 20*7 = 140$ (руб.), можно получить аналогичный результат по формуле (170): $i_Q = 140/250 = 0{,}56$. Очевидно, что абсолютное изменение выручки по формуле (171) составило: $\Delta Q = 140 - 250 = -110$ (руб.), то есть выручка уменьшилась на 110 руб. (или на 44%), что объясняется изменением количества проданных семечек в 0,4 раза (уменьшением на 60%) и изменением их цены в 1,4 раза (повышением цены на 40%).

Подставим формулу (170) в формулу (174) и выразим выручку отчетного периода:

$$Q_1 = i_q i_p Q_0. \qquad (175)$$

Формула (175) представляет собой двухфакторную *мультипликативную индексную модель* итогового показателя, в данном случае – выручки, посредством которой находят изменение этого показателя под влиянием каждого фактора (цены и количества) в отдельности (факторный анализ), то есть:

$$\Delta Q = \Delta Q_q + \Delta Q_p, \qquad (176)$$

где ΔQ_q – изменение выручки под влиянием изменения количества товара q (экстенсивный фактор);

ΔQ_p – изменение выручки под влиянием изменения цены p товара (интенсивный фактор).

Для проведения факторного анализа по формуле (176) необходимо определить *очередность влияния факторов* на результативный показатель (выручку), которая может быть следующей:

1) сначала менялось количество q, а затем цена p (то есть количество – это 1-ый фактор, а цена – 2-ой)[54];

2) сначала менялась цена p, а потом количество q (то есть цена – это 1-ый фактор, а количество – 2-ой).

В соответствии с этой очередностью влияния факторов запись факторов в мультипликатиавной модели: то есть формула (175) – это ее запись для количества как 1-го фактора и цены как 2-го. В случае, когда цена является 1-ым фактором, а количество – 2-ым, необходимо мультипликативную модель записывать в виде (177), то есть меняя факторы местами:

$$Q_1 = i_p i_q Q_0. \qquad (177)$$

Чтобы найти изменение результативного показателя на основе мультипликативной модели за счет 1-го фактора, необходимо исключить влияние остальных факторов. Тогда при использовании формулы (175) влияние 1-го определяем по формуле (178), а при использовании формулы (177) – по формуле (179):

$$\Delta Q_q = i_q Q_0 - Q_0 = (i_q - 1)Q_0, \qquad (178) \qquad \Delta Q_p = i_p Q_0 - Q_0 = (i_p - 1)Q_0. \qquad (179)$$

В нашем примере про бабушку сначала изменилась цена, то есть цена – это 1-ый фактор, а количество – 2-ой, значит необходимо использовать формулу (177) и, как следствие, влияние 1-го фактора – цены определяем по формуле (179): $\Delta Qp = (1{,}4 - 1)*250 = 100$ (руб.), то есть повышение цены семечек с 5 до 7 руб. за кулёк должно было увеличить сегодняшнюю выручку на 100 руб., однако выручка уменьшилась на 110 руб., значит – это отрицательное влияние 2-го фактора – изменение количества.

---

[54] Такая очередность изменения факторов (то есть 1-ый – экстенсивный, а 2-ой – интенсивный) применяется по умолчанию тогда, когда ее затруднительно точно установить

Чтобы найти изменение результативного показателя на основе мультипликативной модели за счет 2-го фактора, необходимо из общего изменения результативного показателя вычесть его изменение под влиянием только 1-го фактора. Тогда, подставляя формулы (171) и (178) в формулу (176), можно выразить влияние второго фактора – цена:

$\Delta Q_p = \Delta Q - \Delta Q_q = (Q_1 - Q_0) - (i_q Q_0 - Q_0) = i_q i_p Q_0 - Q_0 - i_q Q_0 + Q_0 = (i_q i_p - 1 - i_q + 1)Q_0 = i_q (i_p - 1)Q_0$.

В итоге получим формулу для расчета влияния второго фактора – цена (180):

$$\Delta Q_p = i_q (i_p - 1)Q_0. \qquad (180)$$

Аналогично, подставляя формулы (171) и (177) в формулу (176) выводится формула для определения влияния второго фактора – количества:

$\Delta Q_q = \Delta Q - \Delta Q_p = (Q_1 - Q_0) - (i_p Q_0 - Q_0) = i_p i_q Q_0 - Q_0 - i_p Q_0 + Q_0 = (i_p i_q - 1 - i_p + 1)Q_0 = i_p (i_q - 1)Q_0$.

В итоге получим формулу для расчета влияния второго фактора – количества (181):

$$\Delta Q_q = i_p (i_q - 1)Q_0. \qquad (181)$$

В нашем примере про бабушку изменение выручки под влиянием второго фактора – количества определим по формуле (181): $\Delta Q_q = 1{,}4*(0{,}4-1)*250 = -210$ (руб.), то есть снижение количества проданных семечек с 50 кульков до 20 уменьшило выручку на 210 руб. Проверка правильности расчета влияния факторов осуществляется по формуле (176): $\Delta Q = 100 + (-210) = -110$, что совпадает с общим изменением выручки, рассчитанным ранее по формуле (171).

В статистике нередки случаи использования индексных моделей с тремя и более факторными индексами[55]. В случае необходимости проведения факторного анализа таких моделей применяется *метод Чалиева*: для определения влияния i-го фактора на результативный показатель необходимо его базисную величину умножить на индексы факторов, влиявших на него с 1-го до i-го фактора и на темп изменения самого i-го фактора. Темп изменения определяется по формуле (80), то есть надо из индекса вычесть единицу (100%).

Например, общая сумма материальных затрат (M) зависит от объема производства продукции (q), от расхода данного материала на единицу продукции – удельного расхода (m) и от цены единицы данного материала (p) т.е. $M = qmp$. Сравнивая сумму материальных затрат в отчетном периоде с суммой материальных затрат базисного периода получаем (если q - 1-ый фактор, m – 2-ой и p – 3-ий):

$$i_M = \frac{M_1}{M_0} = \frac{q_1 m_1 p_1}{q_0 m_0 p_0} = i_q i_m i_p \text{ или } M_1 = i_q i_m i_p M_0. \qquad (182)$$

Тогда, применяя метод Чалиева, изменение общей суммы материальных затрат $\Delta M = M_1 - M_0$ объясняется:

1) изменением объема продукции $\Delta M_q = T_q M_0 = (i_q - 1)M_0$;
2) изменением удельного расхода материала $\Delta M_m = i_q T_m M_0 = i_q (i_m - 1)M_0$;
3) изменением цены на материал $\Delta M_p = iq i_m T_p M_0 = i_q i_m (i_p - 1)M_0$.

8.3. Общие индексы

Если изучаемое явление неоднородно и сравнение уровней можно провести только после приведения их к общей мере, экономический анализ выполняют посредством общих

[55] В случае построения многофакторных мультипликативных индексных моделей бывает сложно точно определить очередность влияния факторов на результативный показатель, поэтому можно рекомендовать ставить на 1-ое место индекс того фактора, который сильнее всего изменился, на 2-ое место – индекс того фактора, который изменился слабее первого, но сильнее остальных и так далее в порядке убывания изменений индексов

индексов. Индекс становится *общим*, когда в его расчетной формуле показывается неоднородность изучаемой совокупности. Примером неоднородной совокупности является общая масса проданных товаров всех или нескольких видов. Действительно нельзя, например, складывать непосредственно килограммы мяса и рыбы, так как полученный результат в прямом смысле не являлся бы «ни рыбой, ни мясом».

Любые общие индексы могут быть построены 2-мя способами: как агрегатные и как средние из индивидуальных.

Агрегатный индекс является основной и наиболее распространенной формой индекса, если числитель и знаменатель представляют собой набор – «агрегат» (от лат. *aggregatus* – складываемый, суммируемый) непосредственно несоизмеримых и не поддающихся суммированию элементов – сумму произведений двух величин, одна из которых меняется (индексируется), а другая остается неизменной в числителе и знаменателе (вес индекса). Вес индекса служит для целей соизмерения индексируемых величин.

Например, общую сумму выручки можно записать в виде *агрегата* (суммы произведений объемного показателя q на взвешивающий – p), т.е.

$$\sum Q = \sum qp. \qquad (183)$$

Отношение агрегатов, построенных для разных условий, дает общий индекс показателя в агрегатной форме. Так получают *индекс общего объема товарооборота* (*выручки*), показывающий во сколько раз он изменился (или сколько процентов составляет) в отчетном периоде по сравнению с базисным:

$$I_Q = \frac{\sum Q_1}{\sum Q_0} = \frac{\sum q_1 p_1}{\sum q_0 p_0}. \qquad (184)$$

Разность между числителем и знаментелем формулы (184) представляет собой *абсолютное изменение общего товарооборота* (*выручки*) (185), показывающее на сколько в денежных единицах (например, рублях) он изменился в отчетном периоде по сравнению с базисным:

$$\Delta\sum Q = \sum Q_1 - \sum Q_0 = \sum q_1 p_1 - \sum q_0 p_0. \qquad (185)$$

Например, дедушка торговал яблоками двух сортов: «антоновкой» и «белым наливом», результаты торговли за 2 дня представлены в таблице 51:

Таблица 51. Условные данные о торговле яблоками дедушкой за 2 дня

Сорт яблок	Цена за кг, руб.		Объем продаж, кг	
	вчера (p_0)	сегодня (p_1)	вчера (q_0)	сегодня (q_1)
Антоновка	20	18	100	160
Белый налив	22	25	150	120

Рассчитаем выручку дедушки по формуле . (183):

– в отчетном периоде: $\sum Q_1$= 18*160+25*120 = 5880 (руб.) – это выручка от продажи яблок сегодня;

– в базисном периоде: $\sum Q_0$= 20*100+22*150 = 5300 (руб.) – это выручка от продажи яблок вчера.

Теперь определим изменение общей выручки дедушки:

– по формуле (184): I_Q= 5880/5300 = 1,1094, то есть выручка увеличилась в 1,1094 раза, или на 10,94%.

– по формуле (185): $\Delta\sum Q$ = 5880 – 5300 = 580, то есть выручка увеличилась на 580 руб.

При анализе изменения общего объема товарооборота (выручки) это изменение также объясняется изменением уровня цен и количества проданных товаров. Влияние этих факторов выражается агрегатными индексами физического объема (количества) и цен.

Если уровни взвешивающего показателя взяты по данным базисного периода, то получают *агрегатный индекс Ласпейреса*:

$$I_q^{Л} = \frac{\sum q_1 p_0}{\sum q_0 p_0}; \qquad (186) \qquad I_p^{Л} = \frac{\sum p_1 q_0}{\sum p_0 q_0}. \qquad (187)$$

Формула (186) применяется, когда количество – это 1-ый фактор, а формула (187) – когда цена является 1-ым фактором.

Если уровни взвешивающего показателя взяты по данным отчетного периода, то получают *агрегатный индекс Пааше*:

$$I_q^{П} = \frac{\sum p_1 q_1}{\sum p_1 q_0}; \qquad (188) \qquad I_p^{П} = \frac{\sum q_1 p_1}{\sum q_1 p_0}. \qquad (189)$$

Формула (188) применяется, когда количество – это 2-ой фактор, а формула (189) – когда цена является 2-ым фактором.

Произведение агрегатных индексов Ласпейреса и Пааше дает общий индекс выручки:

$$I_Q = I_q^{Л} \; I_p^{П}; \qquad (190) \qquad I_Q = I_p^{Л} \; I_q^{П}. \qquad (191)$$

Для облегчения запоминания студентами формул Ласпейреса и Пааше предлагаю обратить внимание на букву «Ш» в слове «Пааше», которая напоминает «111» - так обозначены отчетные периоды в общей формуле (две единицы – в числителе и одна – в знаменателе). В формуле Ласпейреса нет буквы «Ш», значит в ней не будет трех единиц, а будут три нуля (два нуля – в знаменателе и один – в числителе).

В нашем примере про дедушку (как и в примере про бабушку) цена яблок – это 1-ый фактор, а количество – 2-ой. Поэтому для определения агрегатного индекса цен применяем формулу (187):

$I_p^{Л} = \frac{18*100 + 25*150}{5300} = 5550/5300 =$ 1,0472, то есть цена на яблоки увеличилась в 1,0472 раза (на 4,72%).

Определим агрегатный индекс количества проданных яблок по формуле (188):

$I_q^{П} = \frac{5880}{18*100 + 25*150} =$ 1,0594, то есть количество проданных яблок выросло в 1,0594 раза (на 5,94%).

Контроль правильности расчетов производим по формуле (191): I_Q= 1,0472*1,0594 = 1,1094, то есть изменение общей выручки дедушки в 1,1094 раза (на 10,94%) объясняется изменением цены в 1,0472 раза (на 4,72%) и изменением количества продаж в 1,0594 раза (на 5,94%).

Из формул (186) – (189) видно, что индексы Ласпейреса и Пааше по одному и тому же фактору не равны между собой, то есть $I_q^{Л} \neq I_q^{П}$ и $I_p^{Л} \neq I_p^{П}$. Американский экономист Гершенкрон обширными расчетами установил, что по одному и тому же фактору индекс Ласпейреса обычно больше индекса Пааше, и это открытие названо *эффектом Гершенкрона*[56], то есть $I_q^{Л} > I_q^{П}$ и $I_p^{Л} > I_p^{П}$.

Когда нет возможности определить очередность влияния факторов на результативный показатель (какой из факторов 1-ый – цена или количество) проблематично выбрать одну из формул (186) или (187) и (188) или (189). В таких случаях рекомендуется применить все формулы (186) – (189) и рассчитать среднюю геометрическую величину из однофакторных индексов – *индексы Фишера*:

$$I_q^{Ф} = \sqrt{I_q^{Л} I_q^{П}}; \qquad (192) \qquad I_p^{Ф} = \sqrt{I_p^{Л} I_p^{П}}. \qquad (193)$$

[56] Самостоятельно догадайтесь и придумайте пример, когда эффект Гершенкрона выполняться не будет (подсказка – «эффект картошки»)

Сравнивая значения индексов Фишера, которые показывают среднее изменение цен (193) и количества (192), решается вопрос об очередности влияния факторов: какой из индексов показывает большее изменение, тот фактор и считают 1-ым.

Из формул (190) и (191) легко получить двухфакторные мультипликативные индексные модели общей выручки, подставив в них формулу (184) и выразив $\sum Q_1$:

$$\sum Q_1 = I_q^{Л}\, I_p^{П}\, Q_0, \qquad (194) \qquad\qquad \sum Q_1 = I_q^{П}\, I_p^{Л}\, Q_0. \qquad (195)$$

Формула (194) применяется, когда количество товара – 1-ый фактор, а цена 2-ой, а формула (195) – наоборот, цена – 1-ый фактор, а количество – 2-ой. Тогда, применяя метод Чалиева, можно выполнить факторный анализ, то есть объяснить изменение результативного показателя (общей выручки) изменением каждого фактора (цен и количества) в отдельности в абсолютных (денежных) единицах. Более детальный анализ изменения итогового показателя возможен при изучении так называемых структурных сдвигов.

В нашем примере про дедушку мы применяли формулу (187), значит должны производить факторный анализ по модели (195). Тогда, применяя метод Чалиева, изменение общей выручки $\Delta\sum Q = \sum Q_1 - \sum Q_0$ объясняется изменением:

1) количества проданных проданных яблок $\Delta\sum Q_q = (I_q^{П} - 1)\ \sum Q_0 = (1{,}0594 - 1) * 5300 \approx 315$ (руб.)

2) цены яблок $\Delta\sum Q_p = I_q^{П}\,(I_p^{Л} - 1)\ \sum Q_0 = 1{,}0594 * (1{,}0472 - 1) * 5300 \approx 265$ (руб.)

Проверка правильности расчета влияния факторов: $\Delta\sum Q = 265 + 315 = 580$, что совпадает с общим изменением общей выручки, рассчитанным ранее по формуле (185).

Помимо записи общих индексов в агрегатной форме на практике часто используют формулы их расчета как величин, средних из соответствующих индивидуальных индексов. Так, общий индекс выручки может быть записан как средняя арифметическая взвешенная (196) или средняя гармоническая взвешенная (197) из индивидуальных индексов выручки по отдельным товарным группам:

$$I_Q = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum q_1 p_1}{\sum (q_1 / i_q)(p_1 / i_p)} = \frac{\sum Q_1}{\sum Q_1 / i_Q}; \qquad (196)$$

$$I_Q = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum i_q q_0 i_p p_0}{\sum q_0 p_0} = \frac{\sum i_Q Q_0}{\sum Q_0}. \qquad (197)$$

В формуле (196) весами являются показатели объема товарооборота отдельных товарных групп в отчетном периоде, в формуле (197) – в базисном.

Аналогично через индивидуальных индексы количества товара и цены могут быть выражены общие агрегатные индексы Ласпейреса и Пааше:

$$I_q^{Л} = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0} = \frac{\sum i_q Q_0}{\sum Q_0}; \qquad (198)$$

$$I_p^{Л} = \frac{\sum i_p p_0 q_0}{\sum p_0 q_0} = \frac{\sum i_p Q_0}{\sum Q_0}; \qquad (199)$$

$$I_q^{П} = \frac{\sum p_1 q_1}{\sum p_1 q_1 / i_q} = \frac{\sum Q_1}{\sum Q_1 / i_q}; \qquad (200)$$

$$I_p^{\Pi} = \frac{\sum q_1 p_1}{\sum q_1\, p_1 / i_p} = \frac{\sum Q_1}{\sum Q_1 / i_p}. \qquad (201)$$

8.4. Индексы средних величин

При изучении качественных показателей часто приходится рассматривать изменение во времени (или пространстве) *средней* величины индексируемого показателя для определенной однородной совокупности. Например, в статистических сборниках публикуются данные о динамике средних цен, средней номинальной заработной плате в отдельных отраслях и т.д.

Средняя величина является обобщающей характеристикой качественного показателя и складывается как под влиянием значений показателя у индивидуальных элементов (единиц), из которых состоит объект, так и под влиянием соотношения их весов («структуры» объекта).

Если любой качественный индексируемый показатель обозначить через *x*, а его веса – через *f*, то динамику среднего показателя можно отразить как за счет изменения обоих факторов (*x* и *f*), так и за счет каждого фактора отдельно. В результате получим 3 различных индекса: индекс переменного состава, индекс фиксированного состава и индекс структурных сдвигов.

Индекс переменного состава отражает динамику среднего показателя (для однородной совокупности) за счет изменения *индексируемой величины x* у отдельных элементов (частей целого) и за счет изменения *весов f*, по которым взвешиваются отдельные значения *x*. Любой индекс переменного состава – это отношение двух средних величин для однородной совокупности (за два периода или по двум территориям) (202):

$$I_{n.c.} = \frac{\bar{x}_1}{\bar{x}_0} = \frac{\sum x_1 f_1}{\sum f_1} : \frac{\sum x_0 f_0}{\sum f_0}. \qquad (202)$$

Свое название этот индекс получил потому, что он характеризует динамику средних величин не только за счет изменения индексируемой величины у отдельных элементов (частей целого), но и за счет изменения удельного веса этих частей в общей совокупности, т.е. изменения состава совокупности.

Индекс фиксированного состава отражает динамику среднего показателя лишь за счет изменения *индексируемой величины x*, при фиксировании весов. Если фиксировать веса на уровне отчетного периода f_1, то получим формулу самую распространенную[57] формулу индекса фиксированного состава (203):

$$I_{ф.c.} = \frac{\sum x_1 f_1}{\sum f_1} : \frac{\sum x_0 f_1}{\sum f_1}\ . \qquad (203)$$

Другими словами, индекс фиксированного состава исключает влияние структуры (состава) совокупности на динамику средних величин, рассчитанных для двух периодов по одной и той же фиксированной структуре весов (на уровне отчетного или базисного периода).

По аналогии можно показать динамику среднего показателя лишь за счет изменения только *весов f* при фиксировании индексируемой величины *x*. Такой индекс условно назван *индексом структурных сдвигов*, который определеятся при фиксировании индексируемой

[57] Если зафиксировать веса на уровне базисного периода f_0, то получим менее распространенную формулу индекса фиксированного состава: $I'_{ф.c.} = \frac{\sum x_1 f_0}{\sum f_0} : \frac{\sum x_0 f_0}{\sum f_0}$ или $I'_{ф.c.} = \frac{\sum x_1 d_0}{\sum x_0 d_0}$.

величины на уровне базисного периода x_0 по самой распространенной[58] формуле (204):

$$I_{стр} = \frac{\sum x_0 f_1}{\sum f_1} : \frac{\sum x_0 f_0}{\sum f_0}, \qquad (204)$$

Формулы (203) – (204) обычно применяются в тех случаях, когда влияние изменения структуры совокупности на динамику среднего показателя сильнее (1-ый фактор) влияния изменения только самой индексируемой величины (2-ой фактор)[59].

Если от абсолютных весов f перейти к относительным весам (долям) по формуле (6), то формулы (202) – (204) примут следующий вид:

$$I_{п.с.} = \frac{\sum x_1 d_1}{\sum x_0 d_0}; \qquad (205) \qquad I_{ф.с.} = \frac{\sum x_1 d_1}{\sum x_0 d_1}; \quad (206) \qquad I_{стр} = \frac{\sum x_0 d_1}{\sum x_0 d_0}. \quad (207)$$

В формулах (202) – (207) при анализе конкретных качественных индексируемых показателей (например, цены товара, себестоимости, производительности труда, урожайности и т.п.) вместо обозначений x и f должны использоваться другие общепринятые обозначения.

Например, при анализе такого качественного показателя как цена формулы (202) – (207) примут следующий вид:

$$I_{п.с.} = \frac{\overline{p}_1}{\overline{p}_0} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0} = \frac{\sum p_1 d_1}{\sum p_0 d_0}; \qquad (208)$$

$$I_{ф.с.} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_1}{\sum q_1} = \frac{\sum p_1 d_1}{\sum p_0 d_1}; \quad (209) \qquad I_{стр} = \frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0} = \frac{\sum p_0 d_1}{\sum p_0 d_0}. \quad (210)$$

Нетрудно заметить, что индекс переменного состава есть произведение индекса фиксированного состава на индекс структурных сдвигов:

$$I_{п.с.} = I_{ф.с.} I_{стр}. \qquad (211)$$

Из формулы (211) видно, что, например, индекс структурных сдвигов можно рассчитать путем деления индекса переменного состава на индекс фиксированного состава.

В нашем примере про дедушку определяем индекс переменного состава по формуле (208):

$I_{п.с.} = \frac{5880}{160+120} : \frac{5300}{100+150} = \frac{21{,}0}{21{,}2} = 0{,}9906$, то есть средняя цена яблок сегодня составляет 99,06% от вчерашней, то есть средняя цена снизилась с 21,2 руб. до 21,0 руб. за кг, что составило 0,94%.

Чтобы исключить влияние изменения структуры продаж яблок на динамику средней цены, рассчитаем индекс цены фиксированного состава по формуле (209)[60]:

---

[58] При фиксировании индексируемой величины на уровне отчетного периода x_0 получается менее распространенная формула индекса структурных сдвигов: $I'_{стр} = \frac{\sum x_1 f_1}{\sum f_1} : \frac{\sum x_1 f_0}{\sum f_0}$ или $I'_{стр} = \frac{\sum x_1 d_1}{\sum x_1 d_0}$.

[59] В противном случае применяются формулы, приведенные в сносках к этим формулам. Для определения очередности влияния факторов рассчитываются и те, и другие формулы, а затем рассчитывается их средняя геометрическая величина (индексы Фишера). Сравнивая значения этих индексов Фишера, решается вопрос об очередности влияния факторов: какой из индексов показывает большее изменение, тот фактор и считают 1-ым.

[60] Выбор этой формулы вызван тем, что изменение структуры – это 1-ый фактор, и изменение самих цен – 2-ой (доказать это самостоятельно, воспользовавшись предыдущей сноской)

$$I_{ф.с.} = \frac{5880}{160+120} : \frac{20*160+22*120}{160+120} = \frac{21{,}0}{20{,}857} = 1{,}0069 .$$

Влияние изменения структуры продаж (доля продаж яблок сорта «антоновка» увеличилась, а сорта «белый налив» – уменьшилась) на динамику средней цены яблок отразим с помощью индекса структурных сдвигов, расчитав его по формуле . (210):

$$I_{стр} = \frac{20*160+22*120}{160+120} : \frac{5300}{100+150} = \frac{20{,}587}{21{,}2} = 0{,}9838.$$

Проверку правильности расчетов выполним по формуле (211): 1,0069*0,9838 = 0,9906.

8.5. Территориальные индексы

Территориальные индексы применяются для пространственных, межрегиональных сопоставлений различных показателей. Их расчет более сложен, чем расчет традиционных (динамических) индексов, рассмотренных ранее, по следующим причинам:

1) различия в структуре цен и количества товаров между странами гораздо значительнее, чем между периодами в рамках одной страны, что обусловлено особенностями экономики разных стран.
2) территориальные (международные) сопоставления нередко осуществляются одновременно для группы стран (например, стран ЕС или СНГ), поэтому необходимо согласовывать индексы, исчисленные для всей группы стран.

Для исчисления территориальных индексов применяются особые формулы, которые разработаны на основе положений двух теорий индексов: аксиоматической и экономической.

В *аксиоматической теории индексов* сформулирован ряд требований к индексам с точки зрения формальной логики (например, требования факторной пробы, обратимости во времени, тождественности и др.) Так, требование тождественности означает, что если цены в отчетном периоде не изменились по сравнению с ценами в базисном периоде, то общий индекс цен должен быть равен единице независимо от изменения физического объема. Другое требование этой теории – пропорциональность индексов…

В экономической теории индексов содержится концептуальная основа для поиска «истинного» индекса. Так, истинный индекс цен можно получить, сопоставив расходы потребителей в текущем и базисном периодах при условии, что они обеспечивают равную полезность потребителям при разных ценах, т.е. фактические расходы потребителей сравниваются с условными, гипотетическими, которые при разных ценах в двух периодах обеспечивают одинаковую полезность. Это сравнение и должно обеспечить отыскание «истинного» индекса цен. Заметим, что экономическая теория индексов достаточно абстрактна, поскольку статистики не оперируют категорией полезности, а имеют дело с конкретными товарами и услугами. Тем не менее, теория выражает некий общий теоретический подход к разработке индексов.

В специальной литературе не прекращается дискуссия об обоснованности аксиоматической и экономической теорий индексов и о возможности применения положений этих теорий в статистической практике. Аксиоматическую теорию критикуют за то, что в ней предполагается отсутствие связи между изменением цен и изменением физического объема. Экономическую теорию критикуют за абстрактный характер, то есть за то, что невозможно использовать ее выводы в практической деятельности.

Основные требования к территориальным индексам:

1. *Характерность весов*, то есть для показателей двух стран А и Б в качестве весов должны использоваться цены (физический объем) этих стран А и Б (или средние из них), а не цены (физический объем) какой-либо третьей страны.

2. *Независимость от выбора базисной страны* (требование обратимости индексов во времени, адаптированное к территориальным сопоставлениям), то есть

$$I^{А/Б} \, I^{Б/А} = 1, \qquad (212)$$

где $I^{А/Б}$ – индекс цен (физического объема) страны А по отношению к стране Б;
$I^{Б/А}$ – индекс цен (физического объема) страны Б по отношению к стране А.

3. *Транзитивность*, то есть

$$I^{А/Б} = I^{А/В} : I^{Б/В}, \qquad (213)$$

где $I^{А/В}$ – индекс цен (физического объема) страны А по отношению к стране В;
$I^{Б/В}$ – индекс цен (физического объема) страны Б по отношению к стране В.

Суть требования транзитивности состоит в том, что индекс, полученный для некоторой пары стран А и Б путем прямого сопоставления их цен (физического объема), должен быть равен этому же индексу, полученному косвенным путем, то есть делением индекса $I^{А/В}$ на индекс $I^{Б/В}$.

4. *Аддитивность*, то есть индексы цен (физического объема), рассчитанные для всей совокупности товаров и услуг, должны быть четко согласованы с индексами, исчисленными для всех групп этой совокупности.

5. *Требование факторной пробы*, то есть произведение индекса цен и индекса физического объема должно быть равно индексу стоимости:

$$I_p^{А/Б} I_q^{А/Б} = I_Q^{А/Б}. \qquad (214)$$

В теории и практике международных сопоставлений различают прямые парные и многосторонние сопоставления. Каждые имеют свою специфику, поэтому для их проведения используют различные формулы индексов.

Прямые парные сопоставления проводятся для какой-либо изолированной пары стран (например, для России и США), на которые не влияют показатели третьих стран. Для таких сопоставлений важным является требование характерности весов, факторной пробы и независимости от выбора базисной страны.

Многосторонние сопоставления проводятся одновременно для группы стран, поэтому особое значение приобретает требование транзитивности индексов.

Например, прямые парные сопоставления ВВП и паритетов покупательной способности (ППС) валют проводят в 4 этапа:

1) ВВП сопоставляемых стран А и Б подразделяется на однородные товарные группы (как правило, около 300 групп);
2) для каждой товарной группы подбирается несколько идентичных товаров-представителей с ценами, что дает возможность вычислить индивидуальные индексы цен для всех отобранных товаров-представителей (i_1, i_2, i_3, …, i_n);
3) для каждой товарной группы по индивидуальным индексам цен на товары-представители исчисляется средний индекс цен по формуле средней геометрической простой:

$$\bar{i} = \sqrt[n]{\prod_{i=1}^{n} i_i}, \qquad (215)$$

что связано с необходимостью обеспечить независимость индексов от выбора базисной страны (формула средней арифметической не обеспечивает этого требования) и с практическим отсутствием информации о весах товаров-представителей;

4) рассчитываются средние индексы цен (физического объема) для ВВП в целом по формулам Ласпейреса (199) и Пааше (201), в которых в качестве весов Q используются доли отдельных товарных групп в ВВП:

$$I_p^{Л;А/Б} = \frac{\sum \bar{i} d_Б}{\sum d_Б}, \qquad (216) \qquad I_p^{П;А/Б} = \frac{\sum d_A}{\sum \frac{d_A}{\bar{i}}}; \qquad (217)$$

5) рассчитывается средний индекс цен (физического объема) по формуле Фишера (193);
6) определяется индекс физического объема ВВП стран А и Б путем делением индекса стоимости ВВП этих стран на средний индекс цен Фишера.

Еще один способ прямого парного сопоставления ВВП – это последовательно сопоставить ВВП двух стран А и Б соответственно в ценах стран А и Б, при этом получим 2 индекса физического объема по формулам Ласпейреса (186) и Пааше (188) и исчислить средний индекс физического объема по формуле Фишера (192).

Для проведения многосторонних сопоставлений ВВП и ППС разработаны формулы индексов, которые удовлетворяют требованию транзитивности: формулы ЭКШ, Гири-Камиса, Уолша и Джерарди.

Чаще других используется формула *ЭКШ*[61], которая представляет собой среднюю геометрическую из индексов Фишера для любой пары сравниваемых стран А и Б, исчисленных косвенным путем, т.е. через третью страну *j*:

$$I_{ЭКШ}^{А/Б} = \sqrt[n]{(F^{А/Б})^2 (F^{А/j} F^{j/Б})}, \qquad (218)$$

где $F^{А/Б}$ – индекс Фишера для стран А и Б; $F^{А/j}$ – индекс Фишера для стран А и *j*; $F^{j/Б}$ – индекс Фишера для стран *j* и Б; *n* – число стран, участвующих в сопоставлении.

Недостатком формулы (218) является то, что она не удовлетворяет требованию аддитивности. Этого недостатка нет у формулы Гири-Камиса.

Формула *Гири-Камиса* позволяет исчислять средние международные цены на различные группы товаров, выраженные в единицах условной международной валюты, а также ППС валют всех стран, участвующих в многосторонних сопоставлениях, по отношению друг к другу и к условной международной валюте:

$$I_{ГК} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}\overline{p}_i}{\sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}p_{ij}}, \qquad (219)$$

где q_{ij}– количество *i*-го товара в *j*-ой стране; p_{ij}– цена *i*-го товара в *j*-ой стране; $\overline{p}_i$ – международная цена *i*-го товара.

Недостатком формулы (219) является то, что она не удовлетворяет требованию характерности весов.

Еще один метод территориальных сопоставлений, для которого разработана особая форма индекса, носит название *метода Уолша*, формула которого имеет следующий вид:

$$I_{У}^{А/Б} = \prod (i_p^{А/Б})^{\overline{d}}, \qquad (220)$$

где $i_p^{А/Б}$ – средний индекс цен для *i*-ой товарной группы в стране А по сравнению со страной Б; $\overline{d}$ – средняя доля *i*-ой товарной группы для всей совокупности стран, принимающих участие в сопоставлении.

По формуле (220) рассчитывается средний геометрический индекс, взвешенный по средним весам для группы стран, участвующих в сопоставлении; в качестве этих средних весов выступают средние (для всей совокупности стран) доли товарных групп в соответствующих показателях (например, в ВВП). Формула (220) удовлетворяет требованиям транзитивности и независимости от выбора базисной страны, но не удовлетворяет требованию аддитивности, а также в меньшей мере, чем индексы ЭКШ, удовлетворяет требованию характерности весов.

В практике международных сопоставлений ВВП, проводимых в рамках ЕС, в течение нескольких лет применялся *метод Джирарди*, в основе которого лежит исчисление индексов физического объема ВВП различных стран с помощью оценки ВВП в средних международных ценах, получаемых по формуле средней геометрической простой. Этот метод похож на метод Гири-Камиса, однако, в отличие от него средние международные цены исчисляются здесь по формуле средней геометрической простой (а не по формуле средней арифметической взвешенной, как в методе Гири-Камиса).

При территориальном сопоставлении макроэкономических показателей широко применяется также *метод цепных индексов*, когда в рамках некоторой группы стран интересующий показатель (например, ВВП) сравнивается с этим показателем какой-либо

[61] В названии использованы начальные буквы фамилий трех статистиков, предложивших этот индекс: венгров Элтетэ и Кэвеша и поляка Шульца

одной базисной страны, тогда анализируемые показатели каждой из этой группы стран, кроме базисной, сравниваются с помощью цепных индексов, то есть по отнишению к базисной стране.

8.6. Контрольные задания

Имеются данные (табл. 52) о продажах минимаркетом 3-х видов однородных товаров (А, В и С).

Таблица 52. Варианты выполнения контрольного задания

Вид товара	Цена за единицу товара, руб.		Объем продаж, тыс. штук		Вид товара	Цена за единицу товара, руб.		Объем продаж, тыс. штук	
	1 квартал	2 квартал	1 квартал	2 квартал		1 квартал	2 квартал	1 квартал	2 квартал
1 вариант					**6 вариант**				
A	102	105	205	195	A	130	125	138	198
B	56	51	380	423	B	50	56	339	264
C	26	30	510	490	C	20	21	613	511
2 вариант					**7 вариант**				
A	112	109	202	260	A	107	110	220	189
B	51	48	365	420	B	46	44	490	550
C	22	26	477	316	C	18	20	720	680
3 вариант					**8 вариант**				
A	99	103	198	182	A	95	98	264	197
B	55	59	370	361	B	48	50	360	294
C	20	18	502	456	C	26	25	448	640
4 вариант					**9 вариант**				
A	99	109	188	182	A	89	92	360	294
B	55	56	380	385	B	58	56	410	482
C	20	21	508	444	C	24	25	558	593
5 вариант					**10 вариант**				
A	120	110	170	220	A	120	125	150	108
B	60	58	350	390	B	44	46	513	461
C	19	20	550	490	C	16	19	891	550

Рассчитать индивидуальные, общие и средние индексы, выполнить факторный анализ общей выручки от продажи товаров. По итогам расчетов сделать аргументированные выводы.

Список литературы

1 Агапова Т.Н. Методы статистического изучения структуры сложных систем и ее изменения. – М.: Финансы и статистика, 1996

2 Анализ временных рядов и прогнозирование: Учебник / Афанасьев В.Н., Юзбашев М.М. – М.: Финансы и статистика, 2001. – 228 с.

3 Герчук Я. П. Графические методы в статистике. – М.: Статистика, 1968

4 Общая теория статистики: Учебник/ Под ред. И.И. Елисеевой. – 4-е изд., перераб. и доп. – М.: Финансы и статистика, 2002. – 480 с.

5 Практикум по теории статистики: Учеб. пособие / Под ред. Р.А. Шмойловой. – М.: Финансы и статистика, 2003. – 416 с.

6 Статистика: Учеб. пособие / Под ред. В.Г. Ионина. – Изд. 2-е, перераб. и доп. – М.: ИНФРА-М, 2006. – 384 с.

7 Теория статистики: Учебник / Под ред. Г.Л. Громыко. – Изд. 2-е, перераб. и доп. – М.: ИНФРА-М, 2005. – 476 с.

8 Теория статистики: Учебник для вузов (под ред. Шмойловой Р.А.). – Изд. 4-е, доп., перераб. – М.: Финансы и статистика, 2007. – 656 с.

9 Чалиев А.А., Овчаров А.О. СТАТИСТИКА. Учебно-методическое пособие. Часть 1. – Нижний Новгород: Издательство Нижегородского госуниверситета, 2007.– 87 с.

10 http://www.gks.ru – официальный сайт ФСГС России

11 http://www.chaliev.narod.ru – персональный сайт автора этого конспекта лекций

Приложения – статистические таблицы

Приложение 1. Значения интеграла Лапласа

$$p(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt$$

t	Сотые доли									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,00	0,0000	0,0080	0,0160	0,0239	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,10	0,0797	0,0876	0,0955	0,1034	0,1113	0,1192	0,1271	0,1350	0,1428	0,1507
0,20	0,1585	0,1663	0,1741	0,1819	0,1897	0,1974	0,2051	0,2128	0,2205	0,2282
0,30	0,2358	0,2434	0,2510	0,2586	0,2661	0,2737	0,2812	0,2886	0,2961	0,3035
0,40	0,3108	0,3182	0,3255	0,3328	0,3401	0,3473	0,3545	0,3616	0,3688	0,3759
0,50	0,3829	0,3899	0,3969	0,4039	0,4108	0,4177	0,4245	0,4313	0,4381	0,4448
0,60	0,4515	0,4581	0,4647	0,4713	0,4778	0,4843	0,4907	0,4971	0,5035	0,5098
0,70	0,5161	0,5223	0,5285	0,5346	0,5407	0,5467	0,5527	0,5587	0,5646	0,5705
0,80	0,5763	0,5821	0,5878	0,5935	0,5991	0,6047	0,6102	0,6157	0,6211	0,6265
0,90	0,6319	0,6372	0,6424	0,6476	0,6528	0,6579	0,6629	0,6680	0,6729	0,6778
1,00	0,6827	0,6875	0,6923	0,6970	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243
1,10	0,7287	0,7330	0,7373	0,7415	0,7457	0,7499	0,7540	0,7580	0,7620	0,7660
1,20	0,7699	0,7737	0,7775	0,7813	0,7850	0,7887	0,7923	0,7959	0,7995	0,8029
1,30	0,8064	0,8098	0,8132	0,8165	0,8198	0,8230	0,8262	0,8293	0,8324	0,8355
1,40	0,8385	0,8415	0,8444	0,8473	0,8501	0,8529	0,8557	0,8584	0,8611	0,8638
1,50	0,8664	0,8690	0,8715	0,8740	0,8764	0,8789	0,8812	0,8836	0,8859	0,8882
1,60	0,8904	0,8926	0,8948	0,8969	0,8990	0,9011	0,9031	0,9051	0,9070	0,9090
1,70	0,9109	0,9127	0,9146	0,9164	0,9181	0,9199	0,9216	0,9233	0,9249	0,9265
1,80	0,9281	0,9297	0,9312	0,9328	0,9342	0,9357	0,9371	0,9385	0,9399	0,9412
1,90	0,9426	0,9439	0,9451	0,9464	0,9476	0,9488	0,9500	0,9512	0,9523	0,9534
2,00	0,9545	0,9556	0,9566	0,9576	0,9586	0,9596	0,9606	0,9615	0,9625	0,9634
2,10	0,9643	0,9651	0,9660	0,9668	0,9676	0,9684	0,9692	0,9700	0,9707	0,9715
2,20	0,9722	0,9729	0,9736	0,9743	0,9749	0,9756	0,9762	0,9768	0,9774	0,9780
2,30	0,9786	0,9791	0,9797	0,9802	0,9807	0,9812	0,9817	0,9822	0,9827	0,9832
2,40	0,9836	0,9840	0,9845	0,9849	0,9853	0,9857	0,9861	0,9865	0,9869	0,9872
2,50	0,9876	0,9879	0,9883	0,9886	0,9889	0,9892	0,9895	0,9898	0,9901	0,9904
2,60	0,9907	0,9909	0,9912	0,9915	0,9917	0,9920	0,9922	0,9924	0,9926	0,9929
2,70	0,9931	0,9933	0,9935	0,9937	0,9939	0,9940	0,9942	0,9944	0,9946	0,9947
2,80	0,9949	0,9950	0,9952	0,9953	0,9955	0,9956	0,9958	0,9959	0,9960	0,9961
2,90	0,9963	0,9964	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972
3,00	0,9973	0,9974	0,9975	0,9976	0,9976	0,9977	0,9978	0,9979	0,9979	0,9980
3,10	0,9981	0,9981	0,9982	0,9983	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986
3,20	0,9986	0,9987	0,9987	0,9988	0,9988	0,9988	0,9989	0,9989	0,9990	0,9990
3,30	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,40	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995	0,9995
3,50	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997	0,9997

Приложение 2. Значения t-критерия Стьюдента

при уровне значимости α: 0,10, 0,05, 0,01

Число степеней свободы ν	α			Число степеней свободы ν	α		
	0,1	0,05	0,01		0,1	0,05	0,01
1	6,314	12,706	63,66	**18**	1,734	2,101	2,878
2	2,92	4,3027	9,925	**19**	1,729	2,093	2,861
3	2,353	3,1825	5,841	**20**	1,725	2,086	2,845
4	2,132	2,7764	4,604	**21**	1,721	2,08	2,831
5	2,015	2,5706	4,032	**22**	1,717	2,074	2,819
6	1,943	2,4469	3,707	**23**	1,714	2,069	2,807
7	1,895	2,3646	3,5	**24**	1,711	2,064	2,797
8	1,86	2,306	3,355	**25**	1,708	2,06	2,787
9	1,833	2,2622	3,25	**26**	1,706	2,056	2,779
10	1,813	2,2281	3,169	**27**	1,703	2,052	2,771
11	1,796	2,201	3,106	**28**	1,701	2,048	2,763
12	1,782	2,1788	3,055	**29**	1,699	2,045	2,756
13	1,771	2,1604	3,012	**30**	1,697	2,042	2,75
14	1,761	2,1448	2,977	**40**	1,684	2,021	2,705
15	1,753	2,1315	2,947	**60**	1,671	2	2,66
16	1,746	2,1199,	2,921	**120**	1,658	1,98	2,617
17	1,74	2,1098	2,898	∞	1,645	1,96	2,576

Приложение 3. Значения χ^2-критерия Пирсона

ν \ α	0,10	0,05	0,025	0,01	0,005
1	2,7055	3,8415	5,0239	6,6349	7,8794
2	4,6052	5,9915	7,3778	9,2103	10,5966
3	6,2514	7,8147	9,3484	11,3449	12,8382
4	7,7794	9,4877	11,1433	13,2767	14,8603
5	9,2364	11,0705	12,8325	15,0863	16,7496
6	10,6446	12,5916	14,4494	16,8119	18,5476
7	12,0170	14,0671	16,0128	18,4753	20,2777
8	13,3616	15,5073	17,5346	20,0902	21,9550
9	14,6837	16,9190	19,0228	21,6660	23,5894
10	15,9872	18,3070	20,4832	23,2093	25,1882
11	17,2750	19,6751	21,9201	24,7250	26,7569
12	18,5494	21,0261	23,3367	26,2170	28,2995
13	19,8119	22,3620	24,7356	27,6883	29,8195
14	21,0641	23,6848	26,1190	29,1412	31,3194
15	22,3071	24,9958	27,4884	30,5779	32,8013
16	23,5418	26,2962	28,8454	31,9999	34,2672
17	24,7690	27,5871	30,1910	33,4087	35,7185
18	25,9894	28,8693	31,5264	34,8053	37,1565
19	27,2036	30,1435	32,8523	36,1909	38,5823
20	28,4120	31,4104	34,1696	37,5662	39,9969
21	29,6151	32,6706	35,4789	38,9322	41,4011
22	30,8133	33,9244	36,7807	40,2894	42,7957
23	32,0069	35,1725	38,0756	41,6384	44,1813
24	33,1962	36,4150	39,3641	42,9798	45,5585
25	34,3816	37,6525	40,6465	44,3141	46,9279
26	35,5632	38,8851	41,9232	45,6417	48,2899
27	36,7412	40,1133	43,1945	46,9629	49,6449
28	37,9159	41,3371	44,4608	48,2782	50,9934
29	39,0875	42,5570	45,7223	49,5879	52,3356
30	40,2560	43,7730	46,9792	50,8922	53,6720

Приложение 4. Значения F-критерия Фишера

при уровне значимости $\alpha = 0,05$

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	161,5	200	215,7	224,6	230,2	234	238,9	243,9	249	254,3
2	18,5	19	19,16	19,25	19,3	19,33	19,37	19,41	19,45	19,5
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,9	2,71
10	4,96	4,1	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,2	3,09	2,95	2,79	2,61	2,4
12	4,75	3,88	3,49	3,26	3,11	3	2,85	2,69	2,5	2,3
13	4,67	3,8	3,41	3,18	3,02	2,92	2,77	2,6	2,42	2,21
14	4,6	3,74	3,34	3,11	2,96	2,85	2,7	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,9	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,2	2,96	2,81	2,7	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,9	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,1	2,87	2,71	2,6	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,3	3,44	3,05	2,82	2,66	2,55	2,4	2,23	2,03	1,78
23	4,28	3,42	3,03	2,8	2,64	2,53	2,38	2,2	2	1,76
24	4,26	3,4	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,6	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,3	2,13	1,93	1,67
28	4,2	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,7	2,54	2,43	2,28	2,1	1,9	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2	1,79	1,52
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,4	2,29	2,13	1,95	1,72	1,44
60	4	3,15	2,76	2,52	2,37	2,25	2,1	1,92	1,7	1,39
70	3,98	3,13	2,74	2,5	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,1	2,71	2,47	2,32	2,2	2,04	1,86	1,64	1,28
100	3,94	3,09	2,7	2,46	2,3	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,6	1,21
150	3,9	3,06	2,66	2,43	2,27	2,16	2	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,8	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79.	1,55	1,1
400	3,86	3,02	2,63	2,4	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3	2,61	2,38	2,22	2,1	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,6	2,37	2,21	2,09	1,94	1,75	1,52	

Приложение 5. Критические значения коэффициента автокорреляции

при уровне значимости α: 0,05 и 0,01

Объем выборки *n*	**Положительные значения**		**Отрицательные значения**	
	α = 0,05	**α = 0,01**	**α = 0,05**	**α = 0,01**
5	0,253	0,297	-0,753	-0,798
6	0,345	0,447	-0,708	-0,863
7	0,370	0,510	-0,674	-0,799
8	0,371	0,531	0,625	-0,764
9	0,366	0,533	-0,593	-0,737
10	0,360	0,525	-0,564	-0,705
11	0,353	0,515	-0,539	-0,679
12	0,348	0,505	-0,516	-0,655
13	0,341	0,495	-0,497	-0,634
14	0,335	0,485	-0,479	-0,615
15	0,328	0,475	-0,462	-0,597
20	0,299	0,432	-0,399	-0,524

Приложение 6. Значения критерия Колмогорова P(λ)

λ	*P*	*λ*	*P*
0,30	1	0,80	0,5441
0,35	0,9997	0,85	0,4653
0,40	0,9972	0,90	0,3927
0,45	0,9874	0,95	0,3275
0,50	0,9639	1,0	0,2700
0,55	0,9228	1,1	0,1777
0,60	0,8643	1,2	0,1122
0,65	0,7920	1,3	0,0681
0,70	0,7112	1,4	0,0397
0,75	0,6272	1,5	0,0222