МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования

Томский государственный университет систем управления и радиоэлектроники

Факультет систем управления

Кафедра автоматизированных систем (АСУ)

**Е.Б. Грибанова**

# Статистика

**Учебное пособие**

**2016**

Грибанова, Е. Б. Статистика: Учебное пособие / Грибанова Е. Б. — Томск: ТУСУР, 2016. — 101 с.

В пособии представлены следующие разделы статистики: основные понятия, сводка и группировка, статистические показатели, выборочное наблюдение, статистические индексы. Пособие подготовлено для студентов, обучающихся по направлению 09.03.03 – «Прикладная информатика» (профиль прикладная информатика в экономике)».

# Содержание

# 1. Статистика как наука

Статистика как наука имеет свой предмет исследования. В каждый момент времени массовые социально-экономические явления имеют определенные размеры, уровни, между ними существуют определенные количественные соотношения. Например, численность населения страны на определенную дату, темпы роста валового внутреннего продукта, изменение уровня заработной платы и т.д.

Предметом статистики являются размеры и количественные соотношения качественно определённых массовых социально-экономических явлений, закономерности их связи и развития в конкретных условиях места и времени.

Рассмотрим основные признаки статистики.

Первая особенность статистики как науки заключается в исследовании не отдельных фактов, а массовых социально-экономических явлений и процессов, выступающих как множества отдельных фактов, обладающих как индивидуальными, так и общими признаками.

Вторая особенность статистики как науки в том, что она изучает прежде всего количественную сторону общественных явлений и процессов в конкретных условиях места и времени, т.е. предметом статистики являются размеры и количественные соотношения социально-экономических явлений, закономерности их связи и развития.

Третья особенность статистики как науки заключается в том, что она характеризует структуру общественных явлений. Структура - это внутреннее строение массовых явлений, т.е. внутреннее строение статистического множества. Статистика должна эту структуру обнаружить, выразить и отразить с помощью статистических показателей.

Каждому общественному явлению свойственны изменения в пространстве и времени. Изменения в пространстве, те. в статике, выявляются анализом структуры общественного явления, а изменения во времени, т.е. в динамике, -

исследованием уровня и структуры явления. Такова четвертая особенность статистики как науки.

Анализ динамики включает:

1. установление уровня общественного явления на определенный момент или промежуток времени и определение среднего уровня;
2. выявление характера изменений за каждый промежуток времени и в целом;
3. определение величины и темпов изменения;
4. установление основной тенденции изменений, их закономерностей и составление статистического прогноза.

Исходя из характера и основных черт предмета определим следующие познавательные задачи статистики как науки. Это изучение следующих характеристик:

1. уровня и структуры массовых социально-экономических явлений;
2. взаимосвязи массовых социально-экономических явлений и процессов;
3. динамики массовых социально-экономических явлений.

Теоретическую основу любой науки, в том числе и статистики, составляют понятия и категории, в совокупности которых выражаются основные принципы данной науки. В статистике к важнейшим категориям и понятиям относятся: совокупность, вариация, признак, закономерность.

**Статистическая совокупность** – это множество единиц, обладающих массовостью, однородностью, определенной целостностью, взаимозависимостью состояний отдельных единиц и наличием вариаций.

Каждый отдельно взятый объект данного множества называется **единицей статистической совокупности.**

**Вариация** – количественные изменения значений признака при переходе от одной единицы совокупности к другой. Например, для студентов отличия могут быть в следующих характеристиках: успеваемость, пол, посещаемость и др.

**Статистическая закономерность** – количественная закономерность изменения массовых явлений в пространстве и во времени. Например, соотношение рождаемости, разводов.

**Признак** – конкретное свойство единицы совокупности. Признаки бывают качественными и количественными.

*Качественные* – признаки, выраженные смысловыми понятиями: пол человека, специализация магазинов и др.

*Количественные* – признаки, выраженные числовыми значениями.

Атрибутивные (качественные) признаки не поддаются прямому количественному (числовому) выражению. Отличие количественных признаков от качественных состоит в том, что первые можно выразить итоговыми значениями, вторые - только числом единиц в совокупности. Количественные признаки делятся на дискретные (прерывные) и непрерывные.

Дискретные данные представляют собой отдельные значения признака, общее число которых конечно, либо если бесконечно, то является счетным, т.е. может быть подсчитано натуральными числами от одного до бесконечности (итоговая оценка за экзамен, варианты: 5,4,3,2).

Непрерывные данные могут принимать любое значение в некотором диапазоне (время сдачи экзамена).

Статистическое исследование состоит из трех основных стадий:

1. статистического наблюдения;
2. первичной обработки, сводки и группировки результатов наблюдения;
3. анализа полученных сводных материалов.

Первой стадией статистических исследований является статистическое наблюдение - научно организованный сбор сведений об изучаемых социально-экономических процессах или явлениях. Полученные в результате статистического наблюдения данные являются исходным материалом для выполнения последующих этапов статистического исследования. Характерным для этой стадии является метод массовых наблюдений. Это объясняется тем, что

статистика изучает закономерности, которые выделяются через исследование многочисленных массовых явлений под действием закона больших чисел.

Данные можно разделить на три типа:

- ***cross-sectional data*** – пространственные данные – набор сведений по разным экономическим объектам в один и тот же момент времени
- ***time-series data*** – временные ряды – наблюдение одного экономического параметра в разные периоды или моменты времени. Эти данные естественным образом упорядочены во времени.
- ***panel data*** – панельные данные – набор сведений по разным экономическим объектам за несколько периодов времени (данные переписи населения).

Для измерения информации используются различные виды шкал (табл.1.1). Числовые шкалы, используемые для получения социально-экономической информации, классифицируются по способу преобразования информации в числа. Тип шкалы определяется соответствующим этой шкале множеством допустимых преобразований. Из всего множества теоретически возможных шкал на практике чаще всего используются четыре типа шкал: ***номинальная, порядковая, интервальная и метрическая (относительная)*** шкалы.

Каждая из шкал определяется наличием или отсутствием четырех характеристик:

- Описание
- Порядок
- Расстояние
- Начальная точка.

Таблица 1.1 Виды шкал

| Шкала | Особенности | Пример |
|---|---|---|
| 1. Номинальная | Содержит только категории, данные не могут упорядочиваться | Хобби |

| | | |
|---|---|---|
| 2. Порядковая | Категории могут упорядочиваться, но разности не имеют смысла | Место на соревнованиях |
| 3. Интервальная | Разности могут быть вычислены, но нет отношений | Температура |
| 4. Относительная | Имеется точка отсчета, возможны отношения между значениями | Рост |
| 5. Дихотомическая | Содержит две категории | Пол |

Описание шкалы предполагает использование единого способа записи информации, то есть характеризует составляющие шкалу элементы, например, степень согласия, или способ согласия («да», «нет», «не знаю»). При этом между данными элементами не вводится какая-либо характеристика сравнений – осуществляется только идентификация информации.

Порядок характеризует наличие отношений в способах записи информации, наличия крайних точек зрения («не согласен», «не совсем согласен», «согласен», и т.п.). При этом предусматриваются некоторые сравнительные характеристики, позволяющие, например, упорядочить отношение к предмету исследования.

Расстояние шкалы может быть измерено. Это значит, что оно существует только в тех шкалах, в которых элементы шкалы определены количественно, а между этими элементами шкалы имеются интервалы, расстояние между которыми имеет смысловое значение.

Начальная точка задает тот или иной уровень соотношений между элементами шкалы. Например, начальная точка шкалы измерения массы тела в килограммах, говорит о том, что нулевое значение на этой шкале (нуль килограммов) свидетельствует об отсутствии массы тела вообще.

Вторая стадия статистического исследования представляет собой комплекс последовательных действий по обобщению конкретных единичных фактов, образующих совокупность в целях выявления типичных черт и закономерностей, присущих изучаемому явлению в целом. Важнейшим специфическим методом на этой стадии является метод группировок.

Статистическая сводка включает в себя распределение исходных данных по группам, качественно однородным по одному или нескольким признакам, и получение групповых итогов.

На третьей стадии исследования проводится анализ статистической информации на основе применения обобщающих показателей: абсолютных, относительных. Анализ статистической информации позволяет раскрывать причинные связи изучаемых явлений, определять взаимодействие различных факторов, оценивать эффективность принимаемых управленческих решений, возможные экономические и социальные последствия складывающихся ситуаций.

# 2. Статистическая группировка

## 2.1 Основные понятия

На основе информации, собранной в ходе статистического наблюдения, как правило, нельзя непосредственно выявить и охарактеризовать закономерности социально-экономических явлений. Это связано с тем, что наблюдение дает сведения по каждой единице исследуемого объекта. Полученные данные не являются обобщающими показателями. С их помощью нельзя сделать выводы в целом об объекте без предварительной обработки данных. Поэтому необходимо выполнить систематизацию первичных данных и получение на этой основе сводной характеристики всего объекта при помощи обобщающих статистических показателей.

**Сводка** представляет собой комплекс последовательных операций по обобщению конкретных единичных фактов, образующих совокупность, для выявления типичных черт и закономерностей, присущих изучаемому явлению в целом.

По глубине обработки данных сводка бывает простая и сложная. Простой сводкой называется операция по подсчету общих итогов по совокупности единиц наблюдения или общего объема изучаемого показателя. Например,

чтобы получить общую численность студентов вузов в России, достаточно сложить данные по всем высшим учебным заведениям страны. Сложная сводка представляет собой комплекс операций, включающих группировку единиц наблюдения, подсчет итогов по каждой группе и по всему объекту и представление результатов группировки и сводки в виде статистических таблиц.

Отдельные единицы статистической совокупности объединяются в группы при помощи метода группировки. Это позволяет «сжать» информацию, полученную в ходе наблюдения, и на этой основе выявить закономерности, присущие изучаемому явлению. Группировкой называется разделение множества единиц изучаемой совокупности на группы по определенным существенным для них признакам.

Группировки делятся на типологические, структурные и аналитические. *Типологическая группировка* - это разделение исследуемой качественно разнородной совокупности на классы, социально-экономические типы, однородные группы единиц.

Типологические группировки позволяют проследить зарождение, развитие и отмирание различных типов явлений (развитие различных форм собственности, формирование новых слоев населения). Примером типологической группировки по атрибутивному при знаку является группировка предприятий и организаций по формам собственности (табл.2.1).

Таблица 2.1. Группировка предприятий и организаций по формам собственности в России (на январь 2001)

| № | Группы предприятий по формам собственности | Число предприятий | |
|---|---|---|---|
| | | всего, тыс. | % к итогу |
| 1 | Государственная | 151 | 4,5 |
| 2 | Муниципальная | 217 | 6,5 |
| 3 | Частная | 2510 | 75 |
| 4 | Собственность общественных и религиозных организаций | 223 | 6,7 |
| 5 | Прочие формы собственности (иностранная) | 247 | 7,3 |
| | Всего | 3346 | 100 |

Анализ данных табл.2.1 показал, что подавляющее большинство предприятий находится в частной собственности (75%); предприятия государственной собственности составили 4,5%; предприятия муниципальной собственности - 6,5%; в собственности общественных объединений и прочих форм собственности находится 14% пред приятий.

*Структурная группировка* (табл.2.2) разделяет однородную в качественном отношении совокупность единиц по определенным, существенным признакам на группы, характеризующие ее состав и структуру. Структурные группировки применяются в изучении практически всех социально-экономических явлений и процессов. С их помощью исследуется состав населения по полу, возрасту, месту проживания; состав коммерческих банков по капиталу, численности работников и т.д.

Таблица 2.2 Группировка населения по размеру среднедушевого денежного дохода

| №п/п | Размер среднедушевого денежного дохода, тыс.руб. в месяц | Численность населения | |
|---|---|---|---|
| | | Всего, млн. человек | В % к итогу |
| 1 | До 40 | 2,4 | 1,6 |
| 2 | 40-80 | 23,4 | 15,8 |
| 3 | 80-120 | 34,8 | 23,5 |
| 4 | 120-60 | 29,4 | 19,8 |
| 5 | 160-200 | 20,7 | 13,9 |
| 6 | 200-240 | 13,5 | 9,1 |
| 7 | 240-280 | 8,7 | 5,9 |
| 8 | 280 и более | 15,5 | 10,4 |
| Всего | | 148,4 | 100 |

*Аналитическая группировка* выявляет взаимосвязи и взаимозависимости между изучаемыми социально-экономическими явлениями и признаками, их характеризующими. В статистике признаки делятся на факторные и результативные. Факторными называются признаки, под воздействием которых изменяются другие, результативные, признаки. Особенностью аналитической группировки является то, что в основание группировки кладется факторный

признак, затем подсчитывается количество единиц совокупности и общее суммарное значение результативного признака по каждой выделенной группе и даже производится расчет среднего значения результативного признака по выделенным группам. Взаимосвязь проявляется в том, что с возрастанием (убыванием) значения факторного признака систематически возрастает (убывает) среднее значение результативного признака. Результаты группировки излагаются в статистической таблице.

Данные табл. 2.3 показывают, что с ростом процентной ставки, под которую выдается кредит, средняя сумма кредита, выдаваемая одним банком, уменьшается. Это говорит о том, что между исследуемыми признаками существует обратная связь.

Таблица 2.3 Группировка банков по величине процентной ставки

| Величина процентной ставки | Число банков | Сумма выданных кредитов, млн руб. | |
|---|---|---|---|
| | | всего | в среднем на один банк |
| 11-15 | 7 | 168,1 | 24 |
| 15-19 | 13 | 200,5 | 15,4 |
| 19-23 | 7 | 54,4 | 7,8 |
| 23-27 | 3 | 6,8 | 2,3 |
| Всего | 30 | 429,8 | 14,3 |

Деление рассмотренных группировок, в зависимости от цели и решаемых задач, на три вида носит условный характер, так как группировка может быть универсальной, т.е. одновременно выделяя типы, показывать структуру совокупности и отражать закономерности изменения значений признака в зависимости от другого.

## 2.2 Принципы построения статистических группировок и классификаций

Построение статистических группировок предполагает решение ряда основных задач. Прежде всего, необходимо выбрать группировочный признак,

затем определить число групп, на которые нужно разбить изучаемую совокупность, и зафиксировать границы интервалов группировки. На завершающей стадии необходимо для каждой группировки найти конкретные показатели или их систему, которые должны характеризовать выделенные группы. Выбор группировочного признака является одним из самых важных и сложных вопросов теории статистической группировки. Группировочным признаком называется признак, по которому единицы совокупности разбиваются на отдельные группы. Его часто называют основанием группировки.

В основание группировки могут быть положены как количественные, так и качественные признаки. Первые имеют числовое выражение (объем торгов, курс доллара в рублях, возраст человека, де нежный доход семьи и т.д.), а вторые отражают состояние единицы совокупности: пол человека, его национальность, семейное положение, отраслевую принадлежность предприятия, его форму собственности и организационно-правовую форму и т.д.

После определения основания группировки следует решить вопрос о количестве групп, на которые надо разбить исследуемую совокупность. Число групп зависит от задач исследования и вида признака, положенного в основание группировки, численности совокупности, степени вариации признака.

При построении группировки по качественному (атрибутивному) признаку групп, как правило, будет столько, сколько имеется градаций, видов, состояний у этого признака. Например, в случае проведения группировки населения по полу можно образовать только две группы- мужчины и женщины.

От группировок следует отличать классификацию. Классификацией называется систематизированное распределение явлений и объектов на определенные группы, классы, разряды на основании их сходства и различия. Отличительные черты классификации:

- в основе классификации лежит качественный признак;
- классификации стандартны: они устанавливаются органами государственной и международной статистики. Если в каждом конкретном

исследовании строится своя группировка, то классификация едина для любого исследования независимо от того, проводят его органы государственной статистики или другие учреждения и ведомства (министерства, налоговые органы и т.п.);

- классификации устойчивы. Они остаются неизменными в течение длительного времени. Однако если появляются новые группы единиц, их классы, разряды, то в классификации вносятся соответствующие изменения и дополнения.

Если группировка проводится по количественному признаку, то необходимо тщательно изучить экономическую (социальную) сущность изучаемого явления. Лишь после этого в соответствии с задачами исследования можно решать вопрос о числе групп, близких

При небольшом объеме совокупности не следует образовывать большое число групп, так как группы будут малочисленными. Поэтому показатели, рассчитанные для таких групп, не будут представительными и не позволят получить адекватную характеристику исследуемого явления.

Часто группировка по количественному признаку имеет задачу отразить распределение единиц совокупности по этому признаку. В данном случае количество групп зависит в первую очередь от степени колеблемости группировочного признака: чем больше его колеблемость, тем больше следует образовать групп. Чем больше групп, тем точнее будет воспроизведен характер исследуемого объекта. Однако слишком большое число групп затрудняет выявление закономерностей при исследовании социально-экономических явлений и процессов. Поэтому в каждом конкретном случае при определении числа групп следует исходить не только из степени колеблемости признака, но еще учитывать и особенности объекта, и цель исследования. При использовании персональных компьютеров для обработки статистических данных группировка единиц объекта проводится с помощью стандартных процедур. Одна из таких процедур основана на использовании следующей формулы Стерджесса для определения оптимального числа групп:

$$n = 1 + 3{,}322 \lg N ,$$

где $n$ - число групп;

$N$-объем совокупности.

Пусть даны следующие наблюдения: 2, 4, 6, 1, 5, 9, 11, 3, 5, 6. Тогда число групп будет равно $n = 1 + 3{,}322 \lg 10 = 4{,}322$.

Согласно формуле выбор числа групп зависит от объема совокупности. Недостаток формулы состоит в том, что ее применение дает хорошие результаты, если совокупность состоит из большого числа единиц и распределение единиц по признаку, положенному в основание группировки, близко к нормальному.

После определения числа групп решается задача определения интервалов группировки. Интервал группировки - это интервал значений варьирующего признака, лежащих в пределах определенной группы. Каждый интервал имеет свою ширину, верхнюю и нижнюю границы или хотя бы одну, из них. Нижней границей интервала называется наименьшее значение признака в интервале, а верхней границей - наибольшее значение признака в нем. Ширина интервала (ее еще часто называют интервальной разностью) представляет собой разность между верхней и нижней границами интервала. Интервалы группировки, в зависимости от их величины, бывают равные и неравные. Последние делятся на прогрессивно возрастающие, прогрессивно убывающие, произвольные и специализированные. Если вариация признака проявляется в сравнительно узких границах и распределение носит более или менее равномерный характер, то строят группировку с равными интервалами. Величина равного интервала определяется по следующей формуле:

$$h = (X \max - X \min) / n,$$

где $h$ - величина равного интервала (шаг интервала);

$X_{\max,} X_{\min}$ - максимальное и минимальное значение признака в совокупности

$n$ -число групп.

Прежде чем определять размах вариации, из совокупности рекомендуется исключить аномальные наблюдения; это значит, если максимальные или минимальные значения сильно отличаются от смежных с ними значений вариантов в упорядоченном ряду значений группировочного признака, то для определения величины интервала следует использовать не максимальное и минимальное значения, а значения, несколько превышающие минимум и несколько меньшие, чем максимум.

Существуют следующие определения шага интервала. Если шаг интервала, рассчитанный по формуле представляет собой величину, имеющую один знак до запятой (например, 0,66; 1,372; 5,8), то в этом случае полученные значения целесообразно округлить до десятых долей (0,7; 1,4; 5,8). Когда рассчитанный шаг интервала имеет две значащие цифры до запятой и несколько знаков после запятой, то это значение надо округлить до целого числа. Пусть величина интервала, исчисленная равна 12,785. Тогда это значение следует округлить до целого числа, т.е. до 13. В случае, когда рассчитанный шаг интервала представляет собой трехзначное, четырехзначное и так далее число, эту величину необходимо округлить до ближайшего числа, кратного 100 или 50. Например, 248 следует округлить до 250.

Пример. Пусть требуется произвести группировку с равными интервалами предприятий по стоимости основных фондов; при этом максимальное значение признака равно 2040 млн руб., а минимальное его значение – 290 млн руб. Совокупность включает 80 единиц. Согласно формуле Стерджесса она должна быть разбита на 7 групп. Сначала найдем размах вариации ($R$): $R$ = 2040 - 290 = 1750 млн руб. Затем определим величину интервала: h = 1750 / 7 = 250 млн руб. После этого построим интервалы групп (табл. 2.4). Чтобы не писать каждый раз от ... до, границы групп обозначают следующим образом: 290-540, 540-790 и т.д. Особенностью 1-го варианта построения групп является то, что у всех групп имеются закрытые интервалы. Во 2-м варианте первая и последняя группы — это группы с открытыми интервалами. Открытые - это те интервалы, у которых указана только одна граница: верхняя - у первого, нижняя - у последнего.

Ширина открытого интервала принимается равной ширине смежного с ним интервала. Закрытыми называются интервалы, у которых обозначены обе границы.

Таблица 2.4 Пример построения групп

| № групп | I вариант | II вариант |
|---|---|---|
| I | От 290 до 540 | До 540 |
| II | От 540 до 790 | 540-790 |
| III | От 790 до 1040 | 790-1040 |
| IV | От 1040 до 1290 | 1040-1290 |
| V | От 1290 до 1540 | 1290-1540 |
| VI | От 1540 до 1790 | 1540-1790 |
| VII | От 1790 до 2040 | 1790 и более |

При группировке по количественному признаку границы интервалов могут быть обозначены по-разному. Если основанием группировки служит непрерывный признак, то одно и то же значение при знака выступает и верхней, и нижней границами у двух смежных интервалов. Таким образом, верхняя граница *i*-го интервала равна нижней границе *i*+1-го интервала. Примером такой группировки служат интервалы, приведенные в табл.2.4. При таком обозначении границ может возникнуть вопрос, в какую группу включать единицы объекта, значения признака у которых со впадают с границами интервалов. Например, во вторую или третью группу должно войти предприятие со стоимостью фондов 790 млн руб. Если нижняя граница формируется по принципу «включительно», а верхняя - по принципу «исключительно», то предприятие должно быть отнесено к третьей группе, в противном случае - ко второй. Для того чтобы правильно отнести к той или иной группе единицу объекта, у которой значение признака совпадает с границами интервалов, можно использовать открытые интервалы. Если в основании группировки лежит дискретный признак, то нижняя граница *i*-го интервала равна верхней границе *i*-1-го интервала, увеличенной на 1.

Пусть совокупность состоит из 80 предприятий и ее надо разделить на группы по численности занятых. Минимальное и максимальное значения

группировочного признака соответственно равны 290 и 2040 человек. В этом случае возможны следующие варианты построения групп (табл.2.5).

Таблица 2.5 Варианты построения групп

| № групп | I вариант | II вариант |
|---|---|---|
| I | 290 - 540 | До 540 |
| II | 541 - 790 | 541-790 |
| III | 791 - 1040 | 791-1040 |
| IV | 1041 - 1290 | 1041-1290 |
| V | 1291 - 1540 | 1291-1540 |
| VI | 1541 - 1790 | 1541-1790 |
| VII | 1791 - 2040 | 1791 и более |

Неравные интервалы применяются в статистике, когда значения признака варьируют неравномерно и в значительных размерах, что характерно для большинства социально-экономических явлений, особенно при анализе макроэкономических показателей. Неравные интервалы могут быть прогрессивно возрастающие или убывающие в арифметической или геометрической прогрессии. Величина интервалов, изменяющихся в арифметической прогрессии, определяется следующим образом:

$$h_{i+1} = h_i + a\,.$$

В геометрической

$$h_{i+1} = h_i \cdot q\,.$$

где $a$-константа - число, которое будет положительным при прогрессивно возрастающих интервалах и отрицательным при прогрессивно убывающих интервалах;

$q$-константа - положительное число, которое при прогрессивно возрастающих интервалах будет больше 1, а при прогрессивно убывающих - меньше 1.

Если необходимо построить группировку предприятий отрасли по показателю выручки от реализации продукции, который варьирует от 500 млн руб. до 4000 млн руб., то строить группировку с равными интервалами нецелесообразно, потому что, как правило, совокупность предприятий любой

отрасли промышленности, торговли включает большое число малых предприятий, имеющих небольшую выручку. С ростом выручки от реализации продукции значительно снижается число предприятий. Таким образом, распределение числа предприятий по величине выручки является неравномерным. Поэтому следует построить группировку с неравными интервалами (табл. 2.6).

Таблица 2.6 Группировка с неравными интервалами

| № групп | Интервалы |
|---|---|
| I | 500-800 |
| II | 800-1300 |
| III | 1300-2000 |
| IV | 2000-2900 |
| V | 2900-4000 |

Величина каждого последующего интервала у этой группировки больше предыдущего на 200 млн руб., т.е. увеличивается в арифметической прогрессии.

## 2.3 Ряды распределения и группировки

Рядом распределения в статистике называется ряд цифровых показателей, представляющих распределение единиц совокупности по одному существенному признаку, разновидности которого расположены в определенной последовательности.

По своей конструкции ряд распределения состоит из двух элементов: вариантов (групп по выделенному признаку) и частот (численности групп). Частоты, выраженные в виде относительных величин (доли единиц, процентов), называются частостями. Сумма всех частот называется объемом распределения, или его численностью. Сумма частостей равна 1, если они выражены в долях единицы, и 100%, если они выражены в процентах. Он оформляется в виде статистической таблицы. Общая схема ряда распределения такова: в совокупности, состоящей из $N$ единиц, некоторая переменная величина $x$ (т.е.

какой-либо варьирующий признак) принимает различные значения $x_1, x_2, ..., x_n$. Каждое из этих значений имеет частоту $f_1, f_2, ..., f_n$. Исходя из этого вариационный ряд распределения можно пред ставить в следующем виде (табл. 2.7).

Таблица 2.7 Ряд распределения

| Вариант, $x_i$ | Частота $f_i$ |
|---|---|
| $x_1$ | $f_1$ |
| $x_2$ | $f_2$ |
| … | … |
| $x_n$ | $f_n$ |
| Итого | $\sum_i f_i$ (или N) |

Ряды распределения, являясь группировкой, могут быть образованы по качественному (атрибутивному) и количественному (прерывному или непрерывному) признакам. В первом случае они называются атрибутивными, во втором – вариационными.

В дискретном вариационном ряду распределения группы составлены по признаку, изменяющемуся дискретно и принимающему только целые значения.

В интервальном вариационном ряду распределения группировочный признак, составляющий основание группировки, может принимать в определенном интервале любые значения. Данный ряд распределения целесообразно строить прежде всего при не прерывной вариации признака, а также если дискретная вариация проявляется в широких пределах, т.е. число вариантов дискретного признака достаточно велико.

Если вариационный ряд распределения имеет группы с неравными интервалами, то частоты в отдельных интервалах непосредственно несопоставимы, так как зависят от ширины интервала. Для того чтобы частоты можно было бы сравнивать, исчисляют плотность распределения. Можно рассчитать как абсолютную, так и относительную плотность распределения. Абсолютная плотность распределения - это частота, приходящаяся на единицу

длины интервала, т.е. $f_i / h_i$, а относительная плотность распределения - частость, приходящаяся на единицу длины интервала, т.е. $w_i / h_i$.

Для различных целей бывает уместным осуществлять еще одно преобразование ряда распределения, заключающееся в построении ряда накопленных частот (кумулятивного ряда). Этот ряд показывает число случаев ниже или выше определенного уровня. Отсюда и возникают два варианта в построении ряда накопленных частот: один показывает число случаев, менее определенного значения варьирующего признака, а другой - число случаев, превышающее определенное значение варьирующего признака.

**Графическое изображение рядов распределения**

Графическое изображение облегчает анализ ряда распределения и позволяет судить о форме распределений единиц совокупности по значениям группировочного признака.

Полигон используется при изображении дискретных вариационных рядов. Он представляет собой замкнутый многоугольник, абсциссами вершин которого являются значения варьирующего признака, а ординатами - соответствующие им частоты или частости. Так в таблице 2.8 представлен дискретный ряд. Полигон изображен на рис.2.1.

Таблица 2.8 Дискретный ряд распределения

| №п/п | Группы квартир по числу комнат | Число квартир, тыс.ед. |
|---|---|---|
| 1 | 1 | 10 |
| 2 | 2 | 35 |
| 3 | 3 | 30 |
| 4 | 4 | 15 |
| 5 | 5 | 5 |

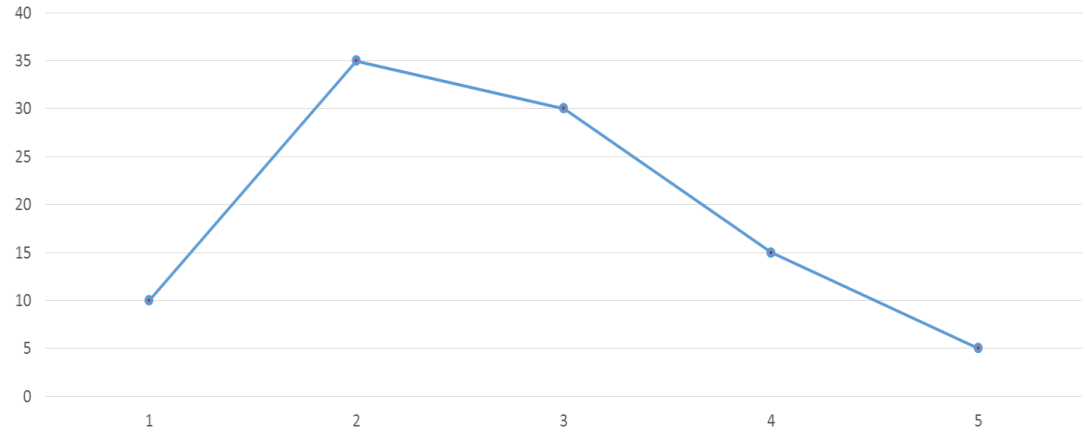



Рис.2.1 Полигон

Гистограмма (гр. histos - ткань, строение) применяется для изображения интервального вариационного ряда, который представляют столбики с основаниями, равными ширине интервалов, и высотой, соответствующей частоте.

Для таблицы 2.9 гистограмма приведена на рис.2.2.

Таблица 2.9 Интервальный ряд распределения

| №п/п | Размер жилой площади, приходящейся на одного человека | Число семей с данным размером жилой площади | Накопленное число семей |
|---|---|---|---|
| 1 | 3-5 | 10 | 10 |
| 2 | 5-7 | 20 | 30 |
| 3 | 7-9 | 40 | 70 |
| 4 | 9-11 | 30 | 100 |
| 5 | 11-13 | 15 | 115 |
| Всего | | 115 | - |

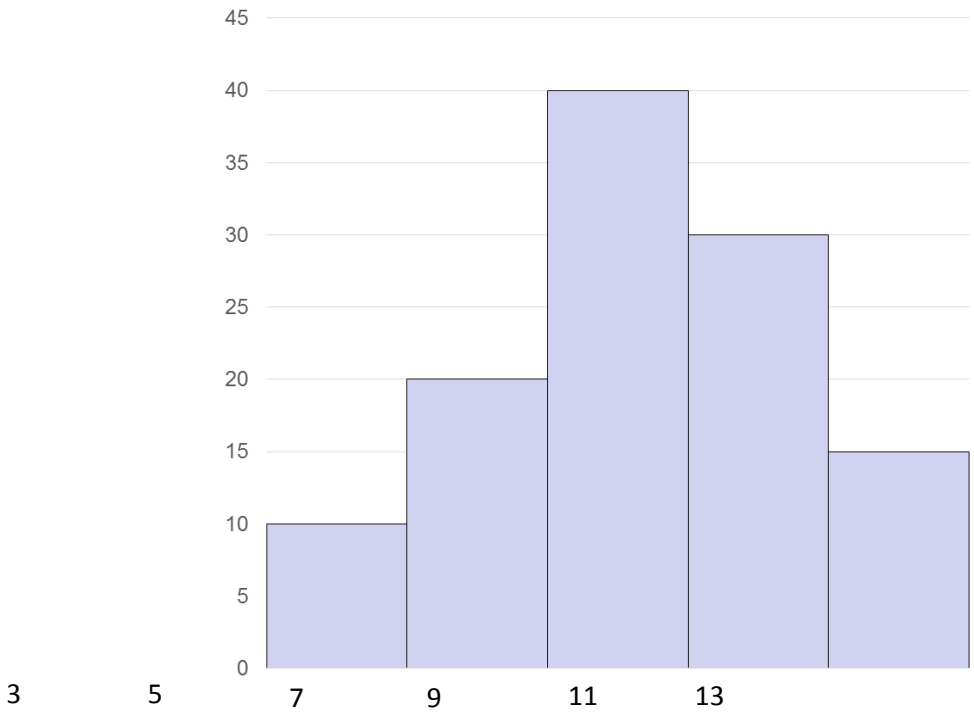



Рис.2.2 Гистограмма распределения

Гистограмма может быть преобразована в полигон распределения, если найти середины сторон прямоугольников и затем эти точки соединить прямыми линиями. При построении гистограммы распределения вариационного ряда с неравными интервалами по оси ординат наносят не частоты, а плотность распределения признака в соответствующих интервалах.

Для графического изображения вариационных рядов может так же использоваться кумулятивная кривая. При помощи кумуляты изображается ряд накопленных частот. При построении кумуляты интервального вариационного ряда по оси абсцисс откладываются варианты ряда, а по оси ординат накопленные частоты, которые наносят на поле графика в виде перпендикуляров к оси абсцисс в верхних границах интервалов. Затем эти перпендикуляры соединяют и получают ломаную линию, т.е. кумуляту (рис.2.3).

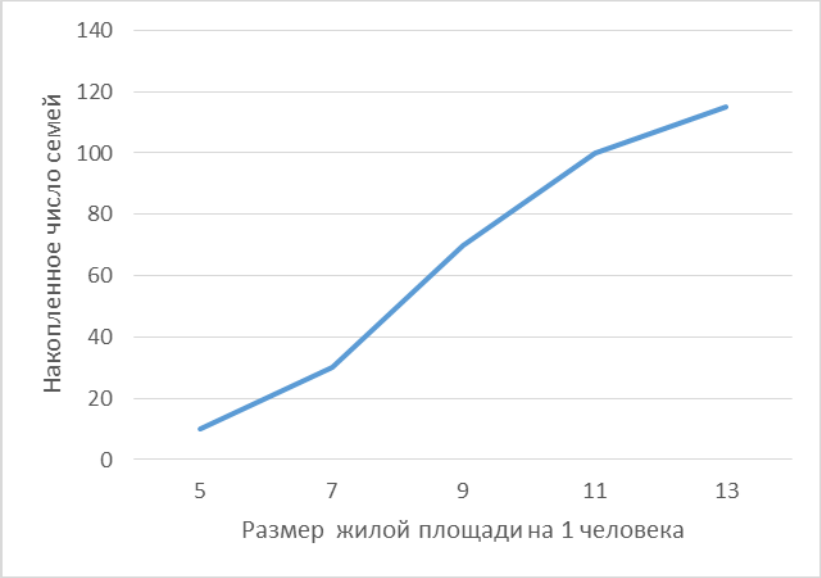



Рис.2.3 Кумулята

## 2.4 Сравнимость статистических группировок

Группировки, построенные за один и тот же период времени, но для разных регионов или, наоборот, для одного региона, но за два разных периода времени, могут оказаться несопоставимыми из-за различного числа выделенных групп или неодинаковости границ интервалов. Для того чтобы привести такие группировки к сопоставимому виду (это позволяет провести их сравнительный анализ), используется метод вторичной группировки. Суть метода состоит в перегруппировке единиц объекта без обращения к первичным данным. Вторичная группировка - операция по образованию новых групп на основе ранее построенной группировки. Применяют два способа образования новых групп. Первым, наиболее простым и распространенным способом является *объединение первоначальных интервалов*. Способ используется, когда нужен переход от мелких интервалов к более крупным интервалам, а также когда границы новых и старых интервалов совпадают. Второй способ получил название *долевой перегруппировки*; он состоит в образовании новых групп на основе закрепления за каждой группой определенной доли единиц совокупности. Этот способ употребляется, когда необходимо в ходе перегруппировки данных определить, какая часть (доля) единиц совокупности перейдет из старых групп в новые.

Рассмотрим первый способ проведения вторичной группировки объединением первоначальных интервалов. Возьмем две группировки кредитов по сроку выдачи за ноябрь и декабрь (табл. 2.10 и 2.11).

Таблица 2.10 Группировка кредитов коммерческих банков по сроку выдачи в ноябре

| №п\п | Группы кредитов по сроку выдачи, мес. | Число заключенных договоров, в % от их общего количества | Сумма выданных кредитов, в % от общей суммы |
|---|---|---|---|
| 1 | 1-3 | 87,05 | 66,87 |
| 2 | 3-6 | 10,43 | 24,86 |
| 3 | 6-12 | 1,8 | 8,17 |
| 4 | Более 12 | 0,72 | 0,1 |
| ИТОГО | | 100 | 100 |

Таблица 2.11 Группировка кредитов коммерческих банков по сроку выдачи в декабре

| №п/п | Группы кредитов по сроку выдачи, мес. | Число заключенных договоров, в % от их общего количества | Суммы выданных кредитов, в % от общей суммы |
|---|---|---|---|
| 1 | 1-6 | 86,54 | 97,91 |
| 2 | 6-12 | 1,92 | 1,7 |
| 3 | Более 12 | 11,54 | 0,39 |
| ИТОГО | | 100 | 100 |

При анализе двух группировок прежде всего их результаты необходимо привести к сопоставимому виду, перегруппировав данные первой группировки. Для этого данные (табл. 2.10) 1-й и 2-й групп объединяют вместе, образуя одну группу краткосрочных кредитов. В эту группу включают все кредиты, выданные в ноябре на срок от 1 до 6 месяцев. Данные 3-й группы (среднесрочные кредиты) и 4-й группы (долгосрочные кредиты) полностью переносятся в табл. 2.12, в которой представлены результаты вторичной группировки кредитов коммерческих банков, выданных в ноябре и декабре.

Таблица 2.12 Результат вторичной группировки

| № п/п | Группы кредитов по сроку выдачи, мес. | Число заключенных договоров, в % от их общего количества | | Сумма выданных кредитов, в % от общей суммы | |
|---|---|---|---|---|---|
| | | ноябрь | декабрь | ноябрь | декабрь |
| 1 | 1-6 | 97,48 | 86,54 | 91,73 | 97,91 |
| 2 | 6-12 | 1,80 | 1,92 | 8,17 | 1,70 |
| 3 | Более 12 | 0,72 | 11,54 | 0,10 | 0,39 |
| ИТОГО | | 100 | 100 | 100 | 100 |

В табл. 2.13 приведены данные о распределении семей по размеру площади, приходящейся на одного человека по двум регионам. Как видно из табл. 2.13, семьи первого региона разбиты на семь групп, а второго - на пять. Чтобы привести данные к сопоставимому виду, произведем перегруппировку семей второго региона. Для этого придется раздробить группы. Так как границы 1-й группы одинаковы у двух группировок, то проведение каких-либо изменений нецелесообразно. 2-ю группу (5-10) необходимо разделить на три группы: семьи, в которых на одного человека приходится 5 и 6 м2, должны образовать 2-ю группу; семьи, где на человека приходится 7 и 8 м2, - 3-ю группу, а где 9-10 м2 следует включить в 4-ю группу. Таким образом, 2-ю группу в группировке семей второго региона следует разбить на три равные по величине интервала группы. При разбивке семей по группам полагают, что их распределение внутри группы 5-10 равномерное. Тогда 1/3 семей группы 5-10 войдет в группу 5-6; 1/3 - в группу 7-8, а оставшаяся часть должна быть включена в группу 9-12. Кроме того, в эту группу следует включить и часть семей из следующей 3-й группы (11–15), т.е. семьи, в которых приходится 11 и 12 м2 жилой площади на одного человека. Поэтому 40% семей 3-й группы надо включить в группу 9-12. Для составления группы 13-14 необходимо взять 40% семей группы 11-15. В группу 15-19 войдут оставшиеся 20% семей группы 11-15, т.е. семьи, в которых приходится на одного человека 15 м2, и все семьи группы 16-19. Перегруппировка последней группы, как и первой, не нужна. Результаты перегруппировки представлены в табл.2.14.

Таблица 2.13 Данные о распределении семей по размеру площади

| Первый регион | Второй регион |
|---|---|

| № групп | Группы семей по размеру жилой площади, приходящейся на одного человека, м2 | Доля семей в % к итогу | № групп | Группы семей по размеру жилой площади, приходящейся на одного человека, м2 | Доля семей в % к итогу |
|---|---|---|---|---|---|
| 1 | До 5 | 3,6 | 1 | До 5 | 6,2 |
| 2 | 5-6 | 11,4 | 2 | 5-10 | 46,3 |
| 3 | 7-8 | 19,4 | 3 | 11-15 | 28,5 |
| 4 | 9-12 | 37,8 | 4 | 16-19 | 10,8 |
| 5 | 13-14 | 11,1 | 5 | 20 и более | 8,2 |
| 6 | 15-19 | 13 | 6 | | |
| 7 | 20 и более | 3,7 | 7 | | |
| ИТОГО | | 100 | ИТОГО | | 100 |

Таблица 2.14 Результаты перегруппировки

| № группы | Группы семей по размеру жилой площади, приходящейся на одного человека, м2 | Доля семей в процентах к итогу | |
|---|---|---|---|
| | | Первый регион | Второй регион |
| 1 | До 5 | 3,6 | 6,2 |
| 2 | 5-6 | 11,4 | 1/3·46,3=15,43 |
| 3 | 7-8 | 19,4 | 1/3·46,3=15,43 |
| 4 | 9-12 | 37,8 | (46,3-2·15,43)+ (0,4·28,5)=26,84 |
| 5 | 13-14 | 11,1 | 0,4·28,5=11,4 |
| 6 | 15-19 | 13 | 0,2·28,5+10,8=16,5 |
| 7 | 20 и более | 3,7 | 8,2 |
| Итого | | 100 | 100 |

## 2.5 Метод группировок и многомерные классификации

Метод группировок позволяет получить общее представление о различных сторонах изучаемого объекта или процесса, выявить закономерности изменения основных показателей в совокупности, установить взаимосвязи и зависимости различных сторон изучаемых явлений, определить влияние факторов на изменение результативного признака. Но часто встречаются задачи, когда группировку нужно выполнить не по одному, а по двум и более факторам. В качестве примеров такой группировки можно назвать сегментацию рынка (объединение товаров в группы по каким-либо признакам), логистику (разбиение

множества точек доставки в группы по их местоположению), анализ социальных сетей (объединение участников в группы по каким –либо признакам и т.д.). На рис.2.4 представлен пример группировки по двум признакам: вес и рост. Предприятие хочет произвести рубашки трех размеров, при этом оно располагает сведениями о параметрах людей, проживающих в данном регионе. Имеющийся набор людей нужно разделить на три группы.

Для исследования таких многофакторных связей используются различные методы многомерной классификации: метод ближайшего соседа, метод k-средних. Рассмотрим метод k-средних.

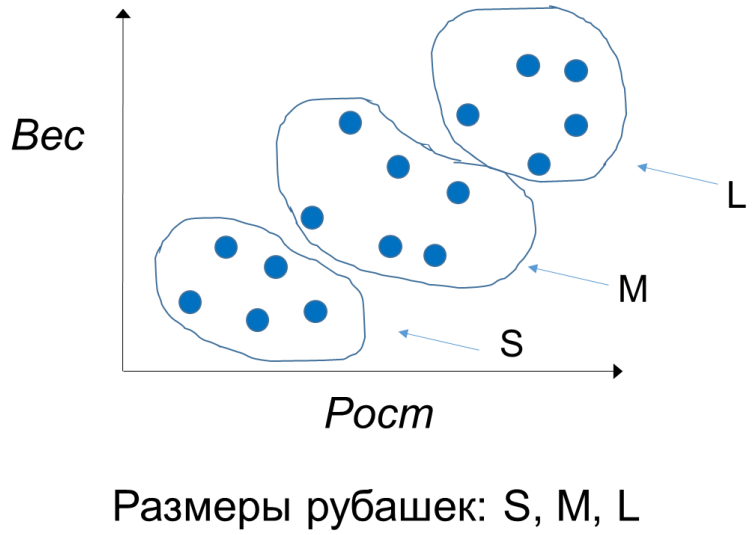



Рис.2.4 Определение размеров рубашек

**Метод k-средних**

Пусть имеется набор данных: $x^{(1)}, x^{(2)}, x^{(3)} \ldots x^{(n)}$ (рис.2.5 а), каждый элемент которого характеризуется двумя показателями: $x_1, x_2$. Необходимо объединить данные в две группы (кластера). Для этого нужно выполнить следующие шаги:

1. случайно сгенерировать центры кластеров (рис.2.5 б);

2. произвести обход всех элементов. В зависимости от того, к какому центру ближе точка, она относится к первому или второму кластеру (рис.2.5 в).

3. определяются новые центры кластеров как среднее значение всех элементов, относящихся к конкретному кластеру (рис.2.5 г). Возврат на шаг 2.

Данные шаги повторяются до тех пор, пока центры кластеров не будут значительно изменяться.

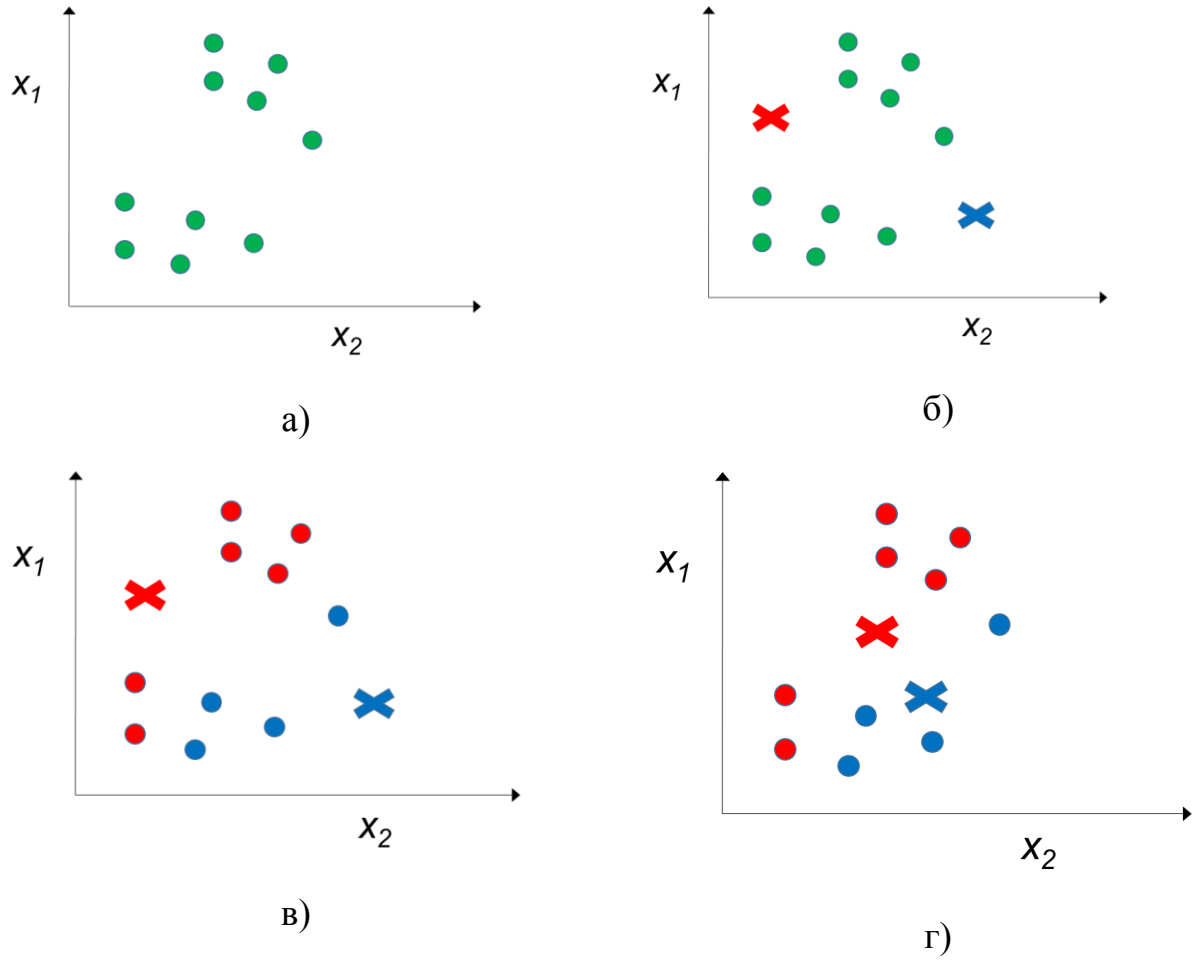



Рис.2.5 Шаги метода k-средних: а) исходный набор данных; б) определение центров кластеров; в) отнесение элементов к конкретному кластеру; г) определение новых центров и отнесение элементов к новым центрам

Таким образом, алгоритм метода можно представить в следующем виде:

Входные данные:

- K (число кластеров);

- Набор данных $x^{(1)}, x^{(2)}, x^{(3)} ... x^{(n)}$;

Случайно генерируются K центров кластеров: $m_1,\ldots, m_k$

Цикл {

for i=1 to n

$c^{(i)}$=индекс (от 1 до K) центра кластера, ближайшего к $x^{(i)}$

for j=1 to K

$m_j$=average(среднее) точек закрепленных за кластером j

}

В качестве центров кластеров на начальном этапе можно выбрать наиболее удаленные друг от друга точки.

В случае необходимости сравнения нескольких вариантов построения кластеров может быть использована следующая целевая функция:

$$J(c^{(1)}, c^{(2)}, ..., c^{(n)}, m_1, ..., m_k) = \frac{1}{n}\sum_{i=1}^{n}\left\| x^{(i)} - m_{c^{(i)}} \right\|^2,$$

где $c^{(i)}$=номер(от 1 до $K$) кластера, к которому текущий элемент $x^{(i)}$ относится

$m_k$- центр кластера $k$;

$m_c^{(i)}$- центр кластера, к которому относится элемент $x^{(i)}$

Также может стоять задача определения наилучшего числа кластеров. Выбор количества кластеров может быть выполнен с помощью метода «локтя». На рис.2.6 показано изменение целевой функции в зависимости от числа кластеров, при этом можно увидеть точку перегиба, после которой уменьшение функции происходит не так значительно. Эта точка и может быть выбрана в качестве решения задачи. В случае, если такой точки нет, то решение о количестве кластеров принимается исследователем на основе его представления о наилучшем значении целевой функции и максимальном количестве кластере. При этом нужно помнить о конечной цели исследования. Так, возможно для задачи на рис.2.4 разбиение нужно провести на пять кластеров, т.к. в этом случае покупатели будут более расположены покупать рубашки.

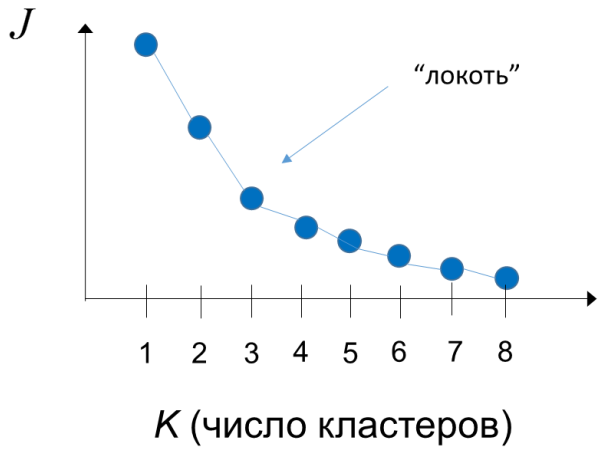



Рис. 2.6 Зависимость целевой функции от числа кластеров

# 3. Статистические показатели

Статистическое исследование независимо от его масштабов и целей всегда завершается расчетом и анализом различных по виду и форме выражения статистических показателей. Статистический показатель представляет собой количественную характеристику социально-экономических явлений и процессов.

Статистические показатели в форме абсолютных величин характеризуют абсолютные размеры изучаемых статистикой процессов и явлений: массу, площадь, объем, протяженность; отражают временные характеристики, а так же могут представлять объем совокупности, т.е. число составляющих ее единиц.

Абсолютные статистические показатели всегда являются именованными числами. В зависимости от социально-экономической сущности исследуемых явлений, их физических свойств, они выражаются в натуральных, стоимостных или трудовых единицах измерения.

Относительный показатель представляет собой результат деления одного абсолютного показателя на другой и выражает соотношение между количественными характеристиками социально-экономических процессов и явлений. Поэтому по отношению к абсолютным показателям относительные показатели или показатели в форме относительных величин являются производными (вторичными). Без относительных показателей невозможно измерить интенсивность развития изучаемого явления во времени, оценить уровень развития одного явления на фоне других взаимосвязанных с ним явлений, осуществить пространственно-территориальные сравнения. При расчете относительного показателя абсолютный показатель, находящийся в числителе получаемого отношения, называется текущим или сравниваемым. Показатель же, с которым производится сравнение и который находится в знаменателе, называется основанием, или базой сравнения. Таким образом рассчитываемый относительный показатель указывает, во сколько раз

сравниваемый абсолютный показатель больше базисного, или какую он составляет от него долю.

Относительные показатели могут выражаться в коэффициентах, процентах, промилле, продецимилле или быть именованными числами. Если база сравнения принимается за 1, то относительный показатель выражается в коэффициентах, если база принимается за 100,1000 или 10 000, то относительный показатель соответственно выражается в процентах (%), промилле и продецимилле.

## 3.1 Средние значения

Наиболее распространенной формой статистических показателей, используемых в социально-экономических исследованиях, является средняя величина, представляющая собой обобщенную количественную характеристику признака в статистической совокупности в конкретных условиях места и времени.

Определить среднюю во многих случаях можно через исходное соотношение средней (ИСС) или ее логическую формулу:

$$Среднее\ \ значение = \frac{Суммарное\ \ значение}{Число\ \ единиц}.$$

При расчете средних величин отдельные значения осредняемого признака могут повторяться, встречаться по нескольку раз. В подобных случаях расчет средней производится по сгруппированным данным или вариационным рядам, которые могут быть дискретными или интервальными.

$$\overline{x} = \sum_{i=1}^{n} x_i \cdot f_i / \sum_{i=1}^{n} f_i ,$$

где $f_i$ - частота *i*-го признака.

Для интервального ряда среднее значение вычисляется с использованием середин интервалов.

Пусть по данным таблицы 3.1 нужно определить стаж работы. В данном случае используем среднюю арифметическую невзвешенную

$$\overline{x} = \frac{x_1 + x_2 + ...x_7}{7} = \frac{10+3+5+12+11+7+9}{7} = 8{,}1$$

Таблица 3.1 Сведения о стаже работников

| Табельный номер рабочего | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Стаж работы | 10 | 3 | 5 | 12 | 11 | 7 | 9 |

По данным таблицы 3.2 определим средний курс продажи. Используем формулу средней арифметической взвешенной:

$$\overline{x} = \frac{\sum_i x_i \cdot f_i}{\sum f_i} = \frac{1080 \cdot 500 + 1050 \cdot 300 + 1145 \cdot 1100}{500 + 300 + 1100} = \frac{2114500}{1900} = 1112{,}9\,.$$

Таблица 3.2 Сведения о продажах акций

| Сделка | Количество проданных акций, шт. | Курс продажи, руб. |
|---|---|---|
| 1 | 500 | 1080 |
| 2 | 300 | 1050 |
| 3 | 1100 | 1145 |

Также рассмотрим определение среднего значения по интервальному ряду (табл.3.3).

Таблица 3.3 Группировка работников по возрасту

| Возраст, лет | Число работников, человек |
|---|---|
| До 25 | 7 |
| 25-30 | 13 |
| 30-40 | 38 |
| 40-50 | 42 |
| 50-60 | 16 |
| 60 и более | 5 |
| ИТОГО | 121 |

Вычислим середины интервалов: 22,5; 27,5; 35; 45; 55; 65. Получим:

$$\bar{x} = \frac{22{,}5 \cdot 7 + 27{,}5 \cdot 13 + 35 \cdot 38 + 45 \cdot 42 + 55 \cdot 16 + 65 \cdot 5}{7 + 13 + 38 + 42 + 16 + 5} = 41.$$

## 3.2 Показатели вариации

При изучении социально-экономических явлений и процессов статистика встречается с разнообразной вариацией признаков, характеризующих отдельные единицы совокупности. Величины признаков колеблются, варьируют под действием различных причин и условий, которые в статистике называются факторами.

Показатели вариации делятся на две группы: абсолютные и относительные. К абсолютным показателям относятся: размах вариации, среднее линейное отклонение, дисперсия и среднее квадратическое отклонение. К относительным показателям вариации относятся: коэффициенты осцилляции, вариации, относительное линейное отклонение и др. Относительные показатели вычисляются как отношение абсолютных показателей вариации к средней арифметической (или медиане).

Вариационный размах ($R$) (или, как еще говорят, амплитуда колебаний) показывает, насколько велико различие между единицами совокупности, имеющими самое маленькое и самое большое значение признака. Размах рассчитывают как разность между наибольшим ($X_{\max}$) и наименьшим ($X_{\min}$) значениями варьирующего признака, т.е.:

$$R = X_{\max} - X_{\min}.$$

Рассмотрим возраст студентов какого-нибудь ВУЗа: самому молодому студенту - 17 лет, самому старшему - 25 лет. Разность составляет 8 лет.

Для анализа вариации необходим показатель, который бы отражал все колебания варьирующего признака и давал обобщенную его характеристику. Такая средняя называется средним линейным отклонением ($d$). Эта величина вычисляется как средняя арифметическая из абсолютных значений отклонений вариант $x_i$ и $\bar{x}$ (простая или взвешенная, в зависимости от исходных условий).

Простая:

$$\bar{d} = \frac{\sum_i |x_i - \bar{x}|}{n}$$

Взвешенная:

$$\bar{d} = \frac{\sum_i |x_i - \bar{x}| \cdot f_i}{\sum_i f_i}.$$

Покажем расчет среднего линейного отклонения по данным табл.3.4.

Таблица 3.4 Группировка конкурсантов по опыту работы

| Группы конкурсантов по опыту работы | Число конкурсантов |
|---|---|
| До 4 лет | 10 |
| 4-6 лет | 10 |
| 6-8 лет | 50 |
| 8-10 лет | 20 |
| 10 и более | 10 |
| Итого | 100 |

Промежуточные результаты расчетов представлены в таблице 3.5.

Таблица 3.5 Промежуточные результаты расчетов

| Группы конкурсантов по опыту работы | Число конкурсантов, $f$ | Середина интервалов, $x'$ | $x'f$ | $\lvert x' - \bar{x} \rvert$ | $\lvert x' - \bar{x} \rvert f_i$ |
|---|---|---|---|---|---|
| До 4 лет | 10 | 3 | 30 | 4,2 | 42 |
| 4-6 лет | 10 | 5 | 50 | 2,2 | 22 |
| 6-8 лет | 50 | 7 | 350 | 0,2 | 10 |
| 8-10 лет | 20 | 9 | 180 | 1,8 | 36 |
| 10 и более | 10 | 11 | 110 | 3,8 | 38 |
| Итого | 100 | | 720 | | 148 |

Алгоритм расчета среднего линейного отклонения следующий:

1. Найдем середину интервалов ($x'$) по исходным данным (столбец 1) и запишем в таблицу (столбец 3).
2. Определим произведения значений середины интервалов ($x'$) на соответствующие им веса ($f$) (столбец 4). В итоге получим 720.

Рассчитаем среднюю величину по формуле средней арифметической взвешенной:

$$\bar{x} = \frac{720}{100} = 7,2 .$$

3. Для расчета линейного отклонения найдем абсолютные отклонения середины интервалов, принятых нами в качестве вариантов признака ( $x'$ ) от средней величины ( $\bar{x}$ ) (столбец 5).
4. Наконец, вычислим произведения отклонений $|x' - \bar{x}|$ на их веса ( $f$ ) и подсчитаем сумму их произведений. Она равна 148. Результаты записываем в столбец 6. Делим эту сумму на сумму весов, чтобы получить искомую величину:

$$\bar{d} = \frac{148}{100} = 1,48 .$$

Дисперсия представляет собой средний квадрат отклонений индивидуальных значений признака от их средней величины и в зависимости от исходных данных вычисляется по формулам простой дисперсии и взвешенной дисперсии:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n},$$

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}.$$

Среднее квадратическое отклонение равно корню квадратному из дисперсии. Среднее квадратическое отклонение, как и среднее линейное отклонение, показывает, на сколько в среднем отклоняются конкретные варианты признака от среднего значения. Они выражаются в тех же единицах измерения, что и признак (в метрах, тоннах, рублях и т.д.).

Рассмотрим расчет среднего квадратического отклонения по данным таблицы 3.6.

Таблица 3.6 Данные о прибыли предприятия

| Предприятие | Прибыль, тыс.руб., х |
|---|---|
| 1 | 600 |
| 2 | 520 |
| 3 | 400 |
| 4 | 600 |
| 5 | 500 |
| 6 | 380 |
| ИТОГО | 3000 |

В таблице 3.7 представлены рассчитанные значения.

Алгоритм расчета следующий:

1. Определим среднюю величину по исходным данным (столбец 2) по формуле средней арифметической простой:

$$\overline{x} = \frac{3000}{6} = 500.$$

2. Найдем отклонения ( $x_i - \overline{x}$ ) и запишем их в столбец 3.

3. Возведем отклонения во вторую степень и запишем в столбец 4. Определим их сумму. Она равна 44800.

4. Разделив эту сумму на число единиц совокупности и взяв квадратный корень из полученного значения определим среднее квадратическое отклонение:

$$\sigma^2 = \frac{44800}{6} = 7466,67$$

$$\sigma = \sqrt{7466,67} = 86,41.$$

Таблица 3.7 Данные о прибыли предприятия

| Предприятие | Прибыль, тыс.руб., х | $x_i - \overline{x}$ | $(x_i - \overline{x})^2$ |
|---|---|---|---|
| 1 | 600 | 100 | 10000 |
| 2 | 520 | 20 | 400 |
| 3 | 400 | -100 | 10000 |
| 4 | 600 | 100 | 10000 |
| 5 | 500 | 0 | 0 |
| 6 | 380 | -120 | 14400 |
| ИТОГО | 3000 | 0 | 44800 |

Рассмотрим также расчет взвешенного средне квадратического отклонения (табл.3.8).

Таблица 3.8 Данные о разрядах работников

| Тариф, разряд $x_i$ | Число работников, $f_i$ |
|---|---|
| 12 | 1 |
| 13 | 5 |
| 14 | 30 |
| 15 | 60 |
| 16 | 30 |
| 17 | 5 |
| 18 | 1 |
| ИТОГО | 132 |

Вычисленные показатели представлены в таблице 3.9.

Таблица 3.9 Данные о разрядах работников

| Тариф, разряд xi | Число работников, fi | $(x_i - \overline{x})^2$ | $(x_i - \overline{x})^2 f_i$ |
|---|---|---|---|
| 12 | 1 | -3 | 9 |
| 13 | 5 | -2 | 4 |
| 14 | 30 | -1 | 1 |
| 15 | 60 | 0 | 0 |
| 16 | 30 | 1 | 1 |
| 17 | 5 | 2 | 4 |
| 18 | 1 | 3 | 9 |
| ИТОГО | 132 | | |

Получим, что среднее значение, дисперсия и среднее квадратическое отклонение равны:

$$\overline{x} = 15$$

$$\sigma^2 = \frac{118}{132} = 0,89, \quad \sigma = \sqrt{0,89} = 0,941$$

Но для целей сравнения колеблемости различных признаков в одной и той же совокупности или же при сравнении колеблемости одного и того же признака в нескольких совокупностях представляют интерес показатели вариации, приведенные в относительных величинах. Базой для сравнения должна служить средняя арифметическая. Эти показатели вычисляются как отношение размаха вариации, среднего линейного отклонения или среднего квадратического отклонения к средней арифметической или медиане. Чаще всего они выражаются в процентах и определяют не только сравнительную оценку вариации, но и дают характеристику однородности совокупности. Совокупность

считается однородной, если коэффициент вариации не превышает 33% (для распределений, близких к нормальному). Различают следующие относительные показатели вариации.

Коэффициент осцилляции:

$$V_R = \frac{R}{x} \cdot 100\%$$

Линейный коэффициент вариации:

$$V_{\bar{d}} = \frac{\bar{d}}{\bar{x}} \cdot 100\%$$

Коэффициент вариации

$$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100\% .$$

В ряде случаев возникает необходимость в измерении дисперсии так называемых альтернативных признаков, тех, которыми обладают одни единицы совокупности и не обладают другие. Примером таких признаков являются: бракованная продукция, ученая степень преподавателя вуза, работа по полученной специальности и т.д. Вариация альтернативного признака количественно проявляется в значении нуля у единицы, которая этим признаком не обладает, или единицы у той, которая данный признак имеет. Пусть $p$ - доля единиц в совокупности, обладающих данным признаком ( $p = \frac{m}{n}$); $q$ - доля единиц, не обладающих данным признаком, причем $p$+$q$=1. Альтернативный признак принимает всего два значения - 0 и 1 с весами соответственно $q$ и $p$. Исчислим среднее значение альтернативного признака по формуле средней арифметической:

$$\bar{x} = \frac{1 \cdot p - 0 \cdot q}{p + q} = p .$$

Дисперсия альтернативного признака определяется по формуле:

$$\sigma^2 = \frac{(1-p)^2 \cdot p + (0-p)^2 \cdot q}{p+q} = \frac{q^2 \cdot p + p^2 \cdot q}{p+q} = p \cdot q .$$

Например, из 40 студентов 10 сдали сессию. Найдем среднее значение и среднее кваратическое отклонение.

Исходные данные: $n = 40$, $m = 10$,

Тогда доля единиц, обладающих данным признаком равна:

$$p = \frac{10}{40} = 0,25.$$

Доля единиц, не обладающих данным признаком:

$$q = 1 - p = 1 - 0,25 = 0,75.$$

Среднее значение и дисперсия равны:

$$\overline{x} = p = 0,25$$

$$\sigma = \sqrt{p \cdot q} = \sqrt{0,25 \cdot 0,75} = \sqrt{0,1875} = 0,43$$

**Виды дисперсий**

Можно определить три показателя колеблемости признака в совокупности: дисперсию общую, межгрупповую и среднюю из внутригрупповых дисперсий. Общая дисперсия ($\sigma^2$) измеряет вариацию признака во всей совокупности под влиянием всех факторов, обусловивших эту вариацию:

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2 \cdot f_i}{\sum f_i}.$$

Межгрупповая дисперсия ($\delta_x^2$) характеризует систематическую вариацию, т.е. различия в величине изучаемого признака, возникающие под влиянием признака-фактора, положенного в основание группировки. Она рассчитывается по формуле:

$$\delta_x^2 = \frac{\sum_{j=1}^{k} (\overline{x}_j - \overline{x}_o)^2 \cdot n_j}{\sum_{j=1}^{k} n_j},$$

где $k$ - число групп;

$n_j$ - число единиц в $j$-й группе;

$\bar{x}_j$ - частная средняя по *j*-й группе;

$\bar{x}_o$ - общая средняя по совокупности единиц.

Внутригрупповая дисперсия ($\sigma_j^2$) отражает случайную вариацию, т.е. часть вариации, происходящую под влиянием неучтенных факторов и не зависящую от признака-фактора, положенного в основание группировки. Она исчисляется следующим образом:

$$\sigma_j^2 = \frac{\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{\sum n_j}.$$

По совокупности в целом вариация значений признака под влиянием прочих факторов характеризуется средней из внутригрупповых дисперсий:

$$\bar{\sigma}^2 = \frac{\sum_{j=1}^{k}\sigma_j^2 \cdot n_j}{\sum_{j=1}^{k} n_j}.$$

Между общей дисперсией, средней из внутригрупповых дисперсий и межгрупповой дисперсией существует соотношение, определяемое правилом сложения дисперсий. Согласно этому правилу общая дисперсия равна сумме средней из внутригрупповых и межгрупповой дисперсий:

$$\sigma^2 = \bar{\sigma}^2 + \delta_x^2.$$

Рассмотрим пример расчета дисперсий (табл.3.10).

Таблица 3.10 Исходные данные о прохождении технического обучения

| Производительность труда рабочих | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| прошедших техническое обучение (деталей за смену) | | | | | не прошедших техническое обучение (деталей за смену) | | | | |
| 84 | 93 | 95 | 101 | 102 | 62 | 68 | 82 | 88 | 105 |

Вычислим средние значения первой и второй группы, а также общее среднее:

$$\bar{x}_1 = \frac{84+93+95+101+102}{5} = \frac{475}{5} = 95$$

$$\bar{x}_2 = \frac{62+68+82+88+105}{5} = \frac{405}{5} = 81$$

$$\bar{x} = \frac{475+405}{10} = 88.$$

Рассчитаем внутригрупповые и общую дисперсии.

Внутригрупповые:

$$\sigma_1^2 = \frac{\sum (x_i - \bar{x}_1)^2}{n_1} = \frac{(84-95)^2 + (93-95)^2 + ... + (102-95)^2}{5} = 42$$

$$\sigma_2^2 = \frac{\sum (x_i - \bar{x}_2)^2}{n_2} = \frac{(62-81)^2 + (68-81)^2 + ... + (102-81)^2}{5} = 231,2$$

Общая:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{(84-88)^2 + (93-88)^2 + ... + (105-88)^2}{10} = 185,6$$

Найдем среднюю из внутригрупповых дисперсий:

$$\bar{\sigma}^2 = \frac{42+231,2}{2} = 136,6.$$

Наконец, вычислим межгрупповую дисперсию:

$$\delta^2 = \frac{(95-88)^2 \cdot 5 + (81-88)^2 \cdot 5}{10} = 49$$

Сумма средней из внутригрупповых дисперсий и межгрупповой дисперсии равна общей дисперсии: $136,6+49=185,6$ (правило сложения дисперсий выполняется).

*Правило сложения дисперсии для доли признака*

Рассмотренное правило сложения дисперсий распространяется и на дисперсии доли признака, т.е. доли единиц с определенным признаком в совокупности, разбитой на группы. При этом изучение вариации происходит непосредственно при вычислении и анализе видов дисперсий для доли признака. Внутригрупповая дисперсия доли определяется по формуле

$$\sigma_{pi}^2 = p_i \cdot (1 - p_i),$$

где $p_i$ - доля изучаемого признака в отдельных группах.

Средняя из внутригрупповых дисперсий имеет следующий вид:

$$\bar{\sigma}_{pi}^2 = \frac{\sum p_i \cdot (1 - p_i) \cdot n_i}{\sum n_i}.$$

Формула межгрупповой дисперсии имеет следующий вид:

$$\delta_{pi}^2 = \frac{\sum (p_i - \bar{p})^2 \cdot n_i}{\sum n_i}.$$

где $n_i$ - численность единиц в отдельных группах;

$\bar{p}$ - доля изучаемого признака во всей совокупности.

Доля признака в совокупности определяется по формуле средней арифметической взвешенной:

$$\bar{p} = \frac{\sum p_i \cdot n_i}{\sum n_i}.$$

Общая дисперсия определяется по формуле

$$\sigma_{\bar{p}}^2 = \bar{p} \cdot (1 - \bar{p}).$$

Три вида рассмотренных дисперсий связаны между собой следующим образом:

$$\sigma_{\bar{p}}^2 = \bar{\sigma}_{pi}^2 + \delta_{pi}^2.$$

Это соотношение дисперсий называется правилом сложения дисперсий доли признака.

Рассмотрим пример вычисления дисперсий. Имеются следующие данные удельного веса основных рабочих в трех цехах фирмы (табл. 3.11).

Таблица 3.11 Исходные данные об основных рабочих

| Цех | Удельный вес основных рабочих, в %, $p_i$ | Численность всех рабочих, человек, $n_i$ |
|---|---|---|
| 1 | 80 | 100 |
| 2 | 75 | 200 |
| 3 | 90 | 150 |
| Итого | | 450 |

Определим долю основных рабочих в целом по фирме:

$$\bar{p} = \frac{0,80 \cdot 100 + 0,75 \cdot 200 + 0,9 \cdot 150}{450} = \frac{365}{450} = 0,81.$$

Общая дисперсия доли основных рабочих по всей фирме в целом равна

$$\sigma_{\bar{p}}^2 = 0,81 \cdot (1 - 0,81) = 0,154.$$

Рассчитаем внутрицеховые дисперсии:

$$\sigma_{p1}^2 = 0,8 \cdot (1 - 0,8) = 0,16$$

$$\sigma_{p1}^2 = 0,75 \cdot (1 - 0,75) = 0,19$$

$$\sigma_{p3}^2 = 0,9 \cdot (1 - 0,9) = 0,09.$$

Средняя из внутригрупповых дисперсий будет равна

$$\bar{\sigma}_{pi}^2 = \frac{0,16 \cdot 100 + 0,19 \cdot 200 + 0,09 \cdot 150}{450} = \frac{675}{450} = 0,15.$$

Межгрупповая дисперсия:

$$\delta_{pi}^2 = \frac{(0,8 - 0,81)^2 \cdot 100 + (0,75 - 0,81)^2 \cdot 200 + (0,9 - 0,81)^2 \cdot 150}{450} = \frac{365}{450} = 0,004$$

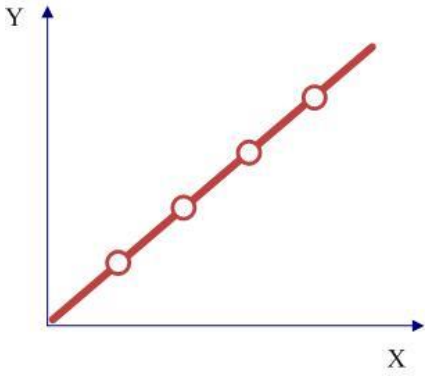
Проверка вычислений показывает: 0,154 = 0,15 + 0,004.

### 3.3 Показатели связи величин

В процессе статистического исследования зависимостей вскрываются причинно-следственные отношения между явлениями, что позволяет выявлять факторы (признаки), оказывающие существенное влияние на вариацию изучаемых явлений и процессов. Причинно-следственные отношения - это связь явлений и процессов, при которой изменение одного из них - причины - ведет к изменению другого – следствия.

В статистике различают функциональную связь и стохастическую зависимость. Функциональной называют такую связь, при которой определенному значению факторного признака соответствует одно и только

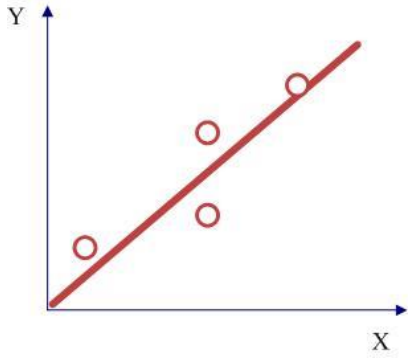
одно значение результативного признака. Функциональная связь проявляется во всех случаях наблюдения и для каждой конкретной единицы исследуемой совокупности. Если причинная зависимость проявляется не в каждом отдельном случае, а в общем, среднем при большом числе наблюдений, то такая зависимость называется стохастической. На рис. 3.1 представлен пример функциональной и стохастической связи. Выручка рассчитывается как произведение количества и цены, поэтому связь между величинами функциональная (рис.3.1 а). Количество проданных товаров невозможно однозначно определить в зависимости от продолжительности рабочего дня, поэтому связь стохастическая.




а)

X – Количество проданного товара

Y – Выручка




б)

X – Продолжительность рабочего дня

Y – Число проданных товаров

Рис.3.1 Связь: а) функциональная; б) стохастическая

По направлению выделяют связь прямую и обратную. При прямой связи с увеличением или уменьшением значений факторного признака происходит увеличение или уменьшение значений результативного показателя (рис.3.2). Так, например, рост производительности труда способствует увеличению уровня рентабельности производства; чем больше студент времени потратит на изучение материала, тем выше его балл. В случае обратной связи значения результативного признака изменяются под воздействием факторного, но в противоположном направлении по сравнению с изменением факторного

признака (рис.3.3). Например, чем больше занятий студент пропустил, тем меньше его балл за семестр. На рис.3.4 представлен пример отсутствия связи: балл студента не зависит того, сколько времени он добирается до университета.

| Студент | Время сам.работы | Балл |
|---|---|---|
| Иванов | 6 | 85 |
| Петров | 2 | 75 |
| Сидоров | 3 | 90 |
| Кротов | 10 | 98 |
| Алиев | 9 | 95 |
| Мухов | 4 | 89 |
| Сонин | 7 | 93 |

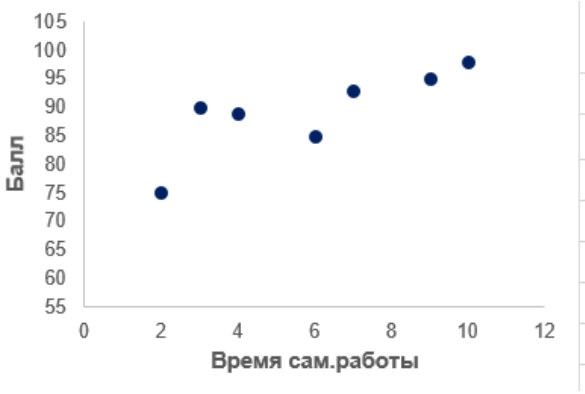



Рис.3.2 Прямая связь

| Студент | Число пропусков | Балл |
|---|---|---|
| Иванов | 6 | 80 |
| Петров | 2 | 95 |
| Сидоров | 3 | 97 |
| Кротов | 10 | 62 |
| Алиев | 9 | 60 |
| Мухов | 4 | 92 |
| Сонин | 7 | 83 |

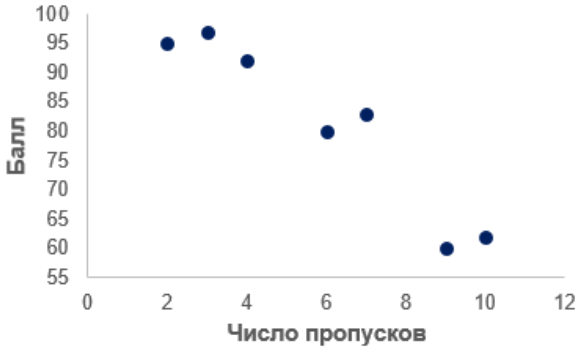



Рис.3.3 Обратная связь

| Студент | Время в пути от дома до университета | Балл |
|---|---|---|
| Иванов | 6 | 80 |
| Петров | 2 | 60 |
| Сидоров | 3 | 94 |
| Кротов | 10 | 67 |
| Алиев | 9 | 91 |
| Мухов | 4 | 74 |
| Сонин | 7 | 71 |

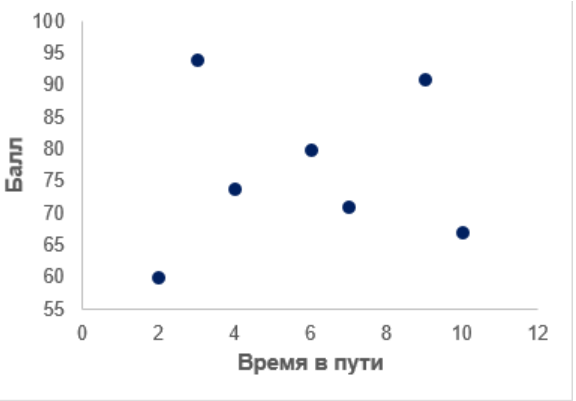



Рис.3.4 Отсутствие связи

По аналитическому выражению выделяют связи линейные и нелинейные. Если статистическая связь между явлениями может быть приближенно

выражена уравнением прямой линии, то ее называют линейной связью; если же она выражается уравнением какой-либо кривой линии (параболы, гиперболы, степен ной, показательной, экспоненциальной и т. д.), то такую связь называют нелинейной, или криволинейной.

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

Корреляционный анализ как общее понятие включает в себя измерение тесноты, направления связи и установление аналитического выражения (формы) связи.

*Примеры исследований:*

1. Менеджер интересуется, зависит ли объем продаж в этом месяце от объема рекламы в этом же периоде?

2. Врач исследует, влияет ли кофеин на сердечные болезни и существует ли связь между возрастом человека и его кровяным давлением?

3. Социолог исследует, какова связь между уровнем преступности и уровнем безработицы в регионе? Связаны ли доход от профессиональной деятельности и продолжительность образования?

Простая связь означает наличие двух переменных (рис.3.5).

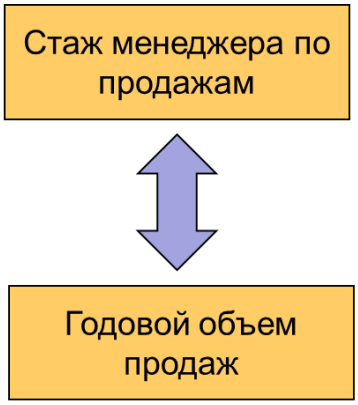



Рис.3.5 Простая связь

Множественная связь означает наличие нескольких переменных (рис.3.6).

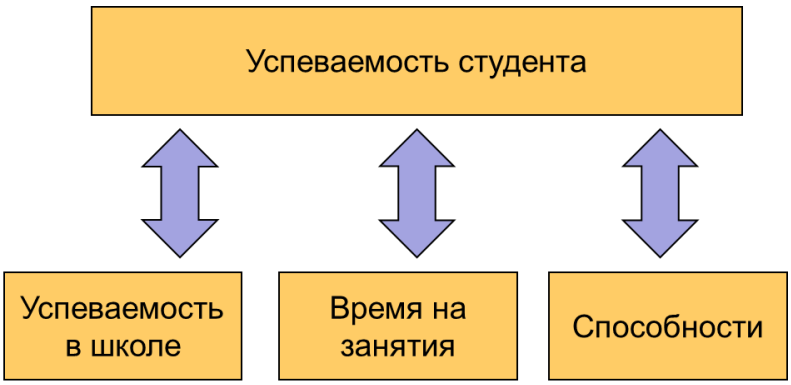



Рис.3.6 Множественная связь

Коэффициент корреляции Пирсона вычисляется по следующим формулам:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}.$$

Коэффициент корреляции изменяется на отрезке от –1 до +1. Если между переменными существует сильная положительная связь, то значение *r* будет близко к +1 (рис.3.7 в). Если между переменными существует сильная отрицательная связь, то значение *r* будет близко к –1 (рис.3.7 а). Когда между переменными нет линейной связи или она очень слабая, значение *r* будет близко к 0.

Интерпретация коэффициента корреляции может быть выполнена в соответствии со следующей шкалой (табл.3.12).

Таблица 3.12 Шкала коэффициента корреляции

| Значение *r* | Уровень связи между переменными |
|---|---|
| 0,75 – 1,00 | Очень высокая положительная |
| 0,50 – 0,74 | Высокая положительная |

| | |
|---|---|
| 0,25 – 0,49 | Средняя положительная |
| 0,00 – 0,24 | Слабая положительная |
| 0,00 – -0,24 | Слабая отрицательная |
| -0,25 – -0,49 | Средняя отрицательная |
| -0,50 – -0,74 | Высокая отрицательная |
| -0,75 – -1,00 | Очень высокая отрицательная |

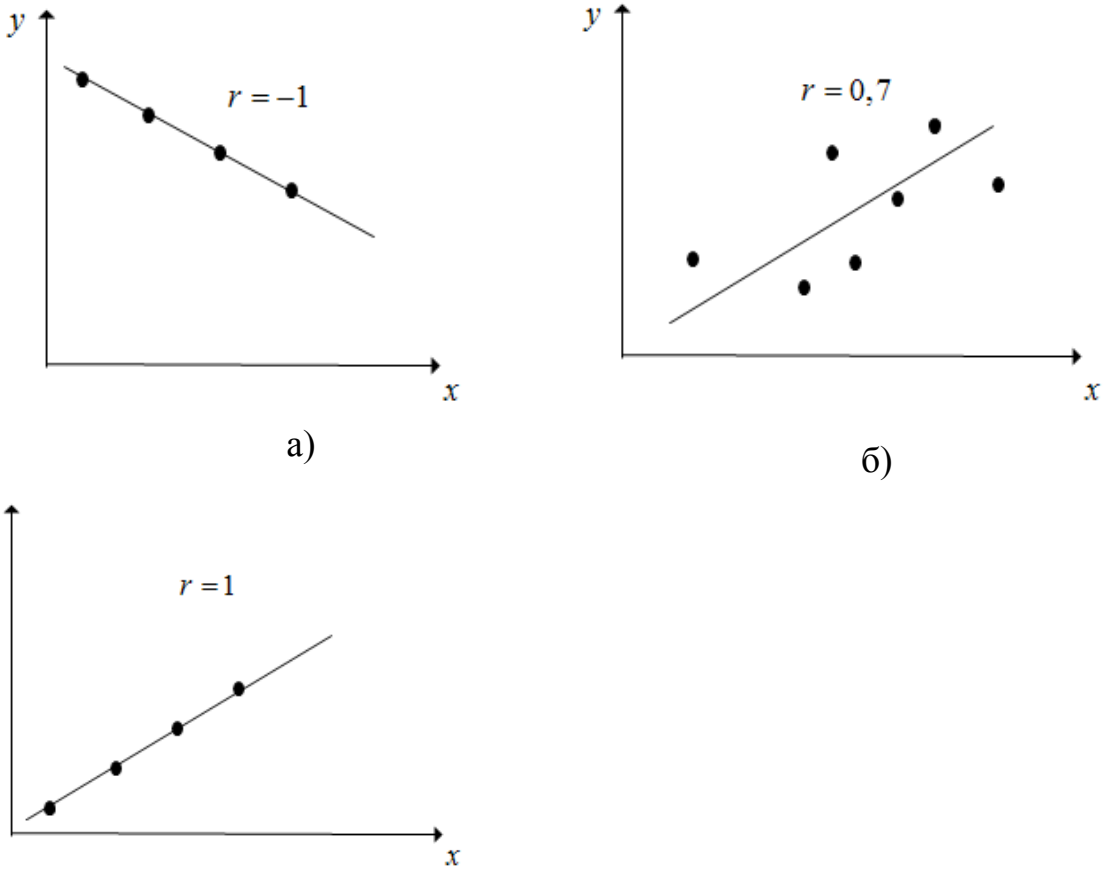



Рис.3.7 Вид зависимости при различных значениях коэффициента корреляции

Рассмотрим пример вычисления коэффициента по данным таблицы 3.13.

Таблица 3.13 Информация об успеваемости студентов

| Студент | Часы изучения | Балл за экзамен |
|---|---|---|
| Иванов А.А. | 3 | 86 |
| Петров А.В. | 5 | 95 |
| Сидоров С.С. | 4 | 92 |
| Ермаков А.Д. | 4 | 83 |
| Нагайцева Е.И. | 2 | 78 |
| Минина К.С. | 3 | 82 |

Таблица 3.14 Промежуточные вычисления

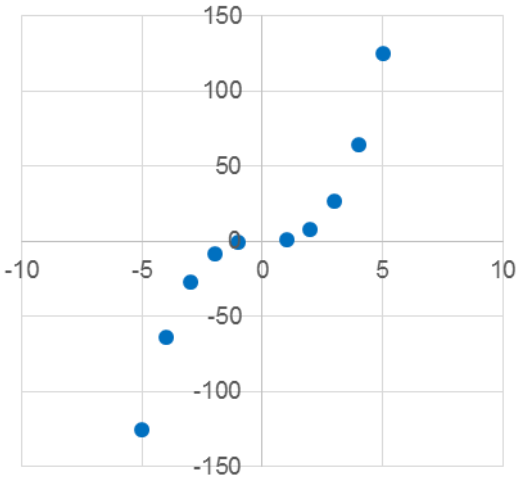
| Студент | Часы изучения, $x$ | Балл за экзамен, $y$ | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| Иванов А.А. | 3 | 86 | 258 | 9 | 7396 |
| Петров А.В. | 5 | 95 | 475 | 25 | 9025 |
| Сидоров С.С. | 4 | 92 | 368 | 16 | 8464 |
| Ермаков А.Д. | 4 | 83 | 332 | 16 | 6889 |
| Нагайцева Е.И. | 2 | 78 | 156 | 4 | 6084 |
| Минина К.С. | 3 | 82 | 246 | 9 | 6724 |
| $\sum$ | 21 | 516 | 1835 | 79 | 44582 |

Таким, образом коэффициент корреляции будет равен

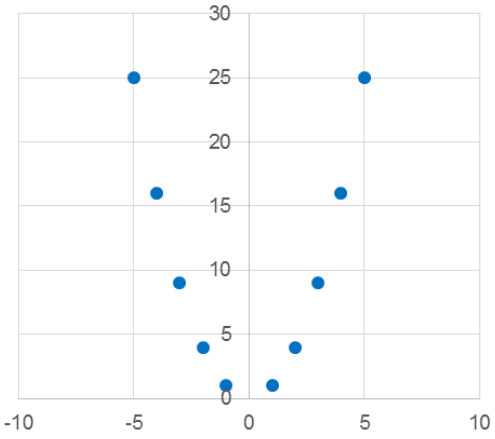
$$r = \frac{6 \cdot 1835 - 21 \cdot 516}{\sqrt{6 \cdot 79 - 21^2}\sqrt{6 \cdot 44582 - 516^2}} = 0{,}862 .$$

Следовательно, между количеством часов обучения и баллом за экзамен существует сильная положительная связь.

В случае нелинейной связи коэффициент корреляции может быть, как равен нулю, так и быть близким к единице (рис.3.8).




а)




б)

Рис.3.8 Нелинейные связи: а) коэффициент корреляции равен 0,93; б) коэффициент корреляции равен 0

Если переменные – порядковые, например, это места, занятые *n* объектами, то обе переменные принимают значения от 1 до *n*. В этом случае формулу для коэффициента корреляции можно упростить:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}.$$

где *d* - разность между значениями величин: *d=x-y*;

*n* - число наблюдений.

Данный коэффициент называют коэффициентом корреляции Спирмена.

Рассмотрим его вычисление для следующего примера. В таблице 3.15 представлены сведения о рейтинге (от 1 до 5) разных профессий по размеру заработной платы и пользе обществу.

Таблица 3.15 Информация о профессиях

| Профессия | Место (оплата) | Место (польза обществу) |
|---|---|---|
| Звезда эстрады | 1 | 5 |
| IT-специалист | 2 | 3 |
| Аудитор | 3 | 4 |
| Шеф-повар | 4 | 2 |
| Врач | 5 | 1 |

Вычислим разность между величинами мест (столбец 4, табл.3.16) и возведем полученное значение в квадрат (столбец 5, табл. 3.16)

Таблица 3.16 Промежуточные вычисления

| Профессия | Место (оплата) | Место (польза обществу) | $d$ | $d^2$ |
|---|---|---|---|---|
| Звезда эстрады | 1 | 5 | -4 | 16 |
| IT-специалист | 2 | 3 | -1 | 1 |
| Аудитор | 3 | 4 | -1 | 1 |
| Шеф-повар | 4 | 2 | 2 | 4 |
| Врач | 5 | 1 | 4 | 16 |
| Итого | | | | 38 |

Полученное значение коэффициента Спирмена равно:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 38}{5(25-1)} = -0,9$$

Т.е. между величинами существует сильная отрицательная связь.

Также для порядковых переменных используется коэффициент корреляции Кендалла.

Алгоритм его расчета следующий:

Шаг 1. Упорядочиваем значения по возрастанию первой переменной.

Шаг 2. Находим число проверсий ($P$) и инверсий ($I$).

Шаг 3. Вычисляем коэффициент корреляции по формуле:

$$\tau = \frac{P - I}{P + I}.$$

Проверсией называется согласованная пара, например, пара IT-специалист и Аудитор.

Таблица 3.17 Проверсия

| Профессия | Место (оплата) | Место (польза обществу) |
|---|---|---|
| IT-специалист | 2 | 3 |
| Аудитор | 3 | 4 |

IT-специалист занимает более высокое место, чем аудитор, и по оплате и по пользе обществу.

Инверсия-несогласованная пара, например, шеф-повар и врач. Шеф-повар занимает более высокое место по оплате, чем врач, но более низкое по пользе обществу (табл.3.18).

Таблица 3.18 Инверсия

| Профессия | Место (оплата) | Место (польза обществу) |
|---|---|---|
| Шеф-повар | 4 | 2 |
| Врач | 5 | 1 |

Рассмотрим вычисление коэффициента Кендалла.

Упорядочиваем профессии по возрастанию первой переменной (табл.3.19).

Таблица 3.19 Упорядочивание по первой переменной

| Профессия | Место (оплата) | Место (польза обществу) | Проверсий | Инверсий | Всего |
|---|---|---|---|---|---|
| Звезда эстрады | 1 | 5 | | | |
| IT-специалист | 2 | 3 | | | |
| Аудитор | 3 | 4 | | | |
| Шеф-повар | 4 | 2 | | | |
| Врач | 5 | 1 | | | |

На первом месте профессия «Звезда эстрады». Её место для пользы обществу– пятое. Считаем количество профессий в последующих строках, у которых место в пользе обществу >5 (получаем 0 проверсии). Те, у которых место <5, – это инверсии (их 4) (табл. 3.20).

Таблица 3.20 Расчет проверсий и инверсий для первой строки

| Профессия | Место (оплата) | Место (польза обществу) | Проверсий | Инверсий | Всего |
|---|---|---|---|---|---|
| Звезда эстрады | 1 | 5 | 0 | 4 | 4 |
| IT-специалист | 2 | 3 | | | |
| Аудитор | 3 | 4 | | | |
| Шеф-повар | 4 | 2 | | | |
| Врач | 5 | 1 | | | |

Повторяем подсчет для остальных строк. Сравниваем место только с последующими строками, так как с предыдущими уже сравнили раньше. Получаем следующие значения (табл.3.21).

Таблица 3.21 Рассчитанные значения

| Профессия | Место (оплата) | Место (польза обществу) | Проверсий | Инверсий | Всего |
|---|---|---|---|---|---|
| Звезда эстрады | 1 | 5 | 0 | 4 | 4 |
| IT-специалист | 2 | 3 | 1 | 2 | 3 |
| Аудитор | 3 | 4 | 0 | 2 | 2 |
| Шеф-повар | 4 | 2 | 0 | 1 | 1 |
| Врач | 5 | 1 | 0 | 0 | 0 |
| Итого | | | 1 | 9 | 10 |

Получили, что число проверсий равно 1, а инверсий - 9. Коэффициент корреляции Кендалла будет равен:

$$\tau = \frac{P - I}{P + I} = \frac{1 - 9}{1 + 9} = -0{,}8.$$

## 3.4 Структурные средние

Наряду со средними величинами в качестве статистических характеристик вариационных рядов распределения рассчитываются структурные средние – мода и медиана.

**Модой** распределения (Мо) называется такая величина изучаемого признака, которая в данной совокупности встречается наиболее часто, т.е. один из вариантов признака повторяется чаще, чем все другие.

Например, в таблице 3.22 представлены сведения об оценках учащихся.

Табл.3.22 Сведения об оценках

| № студента | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Оценка | 4 | 3 | 4 | 5 | 3 | 3 | 5 | 5 | 5 |

Если данные сгруппировать, то получим таблицу 3.23.

Таблица 3.23 Сгруппированные данные

| Оценка | Количество |
|---|---|
| 5 | 4 |
| 4 | 2 |
| 3 | 3 |

Больше всего студенты получили оценок «5», следовательно, мода будет принимать значение, равное пяти.

**Медиана** – значение признака, приходящееся на середину ранжированной совокупности.

Определим медиану по данным таблицы 3.22. Для этого упорядочим значения оценок (табл. 3.24).

Таблица 3.24 Упорядоченные данные

| № студента | 2 | 5 | 6 | 1 | 3 | 4 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Оценка | 3 | 3 | 3 | 4 | 4 | 5 | 5 | 5 | 5 |

Значение оценки, приходящееся на середину упорядоченной совокупности, равно 4. При четном количестве находится среднее значение из двух величин, расположенных в середине совокупности.

Если мода отражает типичный, наиболее распространенный вариант значения признака, то медиана практически выполняет функцию средней величины для неоднородной совокупности, не подчиняющейся нормальному закону распределения. Проиллюстрируем ее познавательное значение. Допустим, нам необходимо дать характеристику среднего дохода группы людей из 100 человек, 99 из которых имеют доход в интервале от 100 до 200 долл. в месяц, а месячный доход последнего человека из группы составляют 50000 долл. (табл.3.25). Если мы воспользуемся формулой средней арифметической, то получим средний доход, равный примерно 600-700 долл., который не только в несколько раз меньше дохода 100 -го человека, но и имеет мало общего с доходами остальных членов группы. Медиана же, равная в данном случае 163 долл., позволит дать объективную характеристику уровня дохода 99% данной группы людей.

Таблица 3.25 Месячные доходы исследуемой группы людей

| №п/п | 1 | 2 | 3 | 4 | … | 50 | 51 | … | 99 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Доход | 10 | 10,4 | 10,4 | 10,7 | … | 16,2 | 16,4 | … | 20,0 | 500,0 |

Главное свойство медианы в том, что сумма абсолютных отклонений значений признака от медианы меньше, чем от любой другой величины

$$\sum |x_i - Me| = \min .$$

Рассмотрим определение моды и медианы по дискретному ряду распределения. В таблице 3.26 представлена информация об успеваемости студентов.

Таблица 3.26 Сведения об успеваемости студентов

| Оценка | Численность студентов |
|---|---|
| 2 | 12 |
| 3 | 48 |
| 4 | 56 |
| 5 | 60 |
| Всего | 176 |

По данным табл.3.26 наибольшую частоту (60 чел.) имеет оценка «5», следовательно, это значение и является модальным (Mo = 5).

Положение медианы в ряду распределения определяется ее номером:

$$N_{Me} = \frac{n+1}{2},$$

где $n$ - число единиц совокупности.

Для примера таблицы 3.26:

$$N_{Me} = \frac{n+1}{2} = \frac{176+1}{2} = 88{,}5 .$$

Полученное значение указывает, что середина ряда приходится на 88 и 89 студентов. Необходимо определить, какую оценку получили студенты с данными номерами. Это можно сделать, рассчитав накопленные частоты (табл.3.27).

Таблица 3.27 Накопленные частоты

| Оценка | Численность студентов | Накопленные частоты |
|---|---|---|
| 2 | 12 | 12 |
| 3 | 48 | 60 |
| 4 | 56 | 116 |
| 5 | 60 | 176 |
| Всего | 176 | |

Очевидно, что студентов с таким номером нет в первой группе, где всего 12 человек, нет их и во второй группе (12 + 48). 88 и 89-й номера студентов находятся в третьей группе (12 + 48 + 56 = 116), следовательно, медианным значением является оценка «4».

Модальный интервал (т.е. содержащий моду) в случае интервального распределения с равными интервалами определяется по наибольшей частоте; с неравными интервалами - по наибольшей плотности, а определение моды требует проведения расчетов на основе следующих формул:

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

где $x_0$ - нижняя граница модального интервала;

$i$ - величина модального интервала;

$f_{Mo}$ -частота модального интервала;

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

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

Медиана в случае интервального распределения вычисляется по формуле:

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

где $x_0$ - нижняя граница медианного интервала;

$i$ - величина медианного интервала;

$f_{Me}$ - частота медианного интервала;

$S_{Me-1}$ - накопленная частота интервала, предшествующего медианному.

Для установления медианного интервала необходимо определить накопленную частоту каждого последующего интервала до тех пор, пока она не превысит половины суммы накопленных частот.

Вычислим моду и медиану по данным таблицы 3.28.

Таблица 3.28 Данные о весе людей

| Вес (фунты) | Количество человек | Удельный вес, % |
|---|---|---|
| 100-110 | 3 | 3,75 |
| 110-120 | 3 | 3,75 |
| 120-130 | 7 | 8,75 |
| 130-140 | 12 | 15 |
| 140-150 | 23 | 28,75 |
| 150-160 | 13 | 16,25 |
| 160-170 | 10 | 12,5 |
| 170-180 | 4 | 5 |
| 180-190 | 2 | 2,5 |
| 190-200 | 3 | 3,75 |
| Всего | 80 | 100 |

Модальным является интервал 140-150 (частота его наибольшая), величина интервала равна 10, удельный вес равен 28,75, удельный вес предыдущего интервала – 15, а последующего – 16,25. Следовательно, мода будет вычислена по формуле:

$$Mo = 140 + 10 \cdot \frac{28,75 - 15}{(28,75 - 15) + (28,75 - 16,25)} = 145,238.$$

Для определения медианного интервала необходимо рассчитать накопленные частоты (табл.3.29).

Таблица 3.29 Накопленные частоты

| Вес (фунты) | Количество человек | Удельный вес, % | Накопленная частота |
|---|---|---|---|
| 100-110 | 3 | 3,75 | 3,75 |
| 110-120 | 3 | 3,75 | 7,5 |
| 120-130 | 7 | 8,75 | 16,25 |
| 130-140 | 12 | 15 | 31,25 |
| 140-150 | 23 | 28,75 | 60 |
| 150-160 | 13 | 16,25 | 76,25 |
| 160-170 | 10 | 12,5 | 88,75 |
| 170-180 | 4 | 5 | 93,75 |
| 180-190 | 2 | 2,5 | 96,25 |
| 190-200 | 3 | 3,75 | 100 |
| Всего | 80 | 100 | |

Половина накопленных частот равна 50, следовательно, медианным интервалом будет интервал 140-150 (т.к. его накопленная частота первая превышает 50). Величина интервала равна 10, половина суммы частот - 50,

накопленная частота предыдущего интервала (130-140) – 31,25, частота интервала 140-150 равна 28,75. Вычислим медиану:

$$Me = 140 + 10 \cdot \frac{50 - 31,25}{28,75} = 146,52 .$$

Моду и медиану в интервальном ряду распределения можно определить графически. Мода определяется по гистограмме распределения. Для этого выбирается самый высокий прямоугольник (интервал с наибольше частотой), который в данном случае является модальным. Затем правую вершину модального прямоугольника соединяют с правым верхним углом предыдущего прямоугольника. А левую вершину модального прямоугольника - с левым верхним углом последующего прямоугольника. Далее из точки их пересечения опускают перпендикуляр на ось абсцисс. Абсцисса точки пересечения этих прямых и будет модой распределения

Для примера таблицы 3.29 графическое определение моды будет выглядеть следующим образом (рис.3.9).

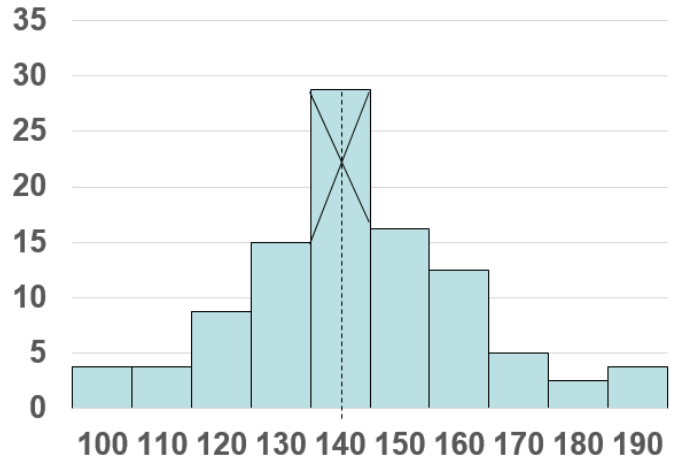



Рис.3.9 Графическое определение моды

Медиана рассчитывается по кумуляте. Для ее определения из точки на шкале накопленных частот (частостей), соответствующей 50%, проводится прямая, параллельная оси абсцисс, до пересечения с кумулятой. Затем из точки пересечения указанной прямой с кумулятой опускается перпендикуляр на ось абсцисс. Абсцисса точки пересечения является медианой.

Для примера (табл.3.29) медиана будет определена следующим образом (рис.3.10).

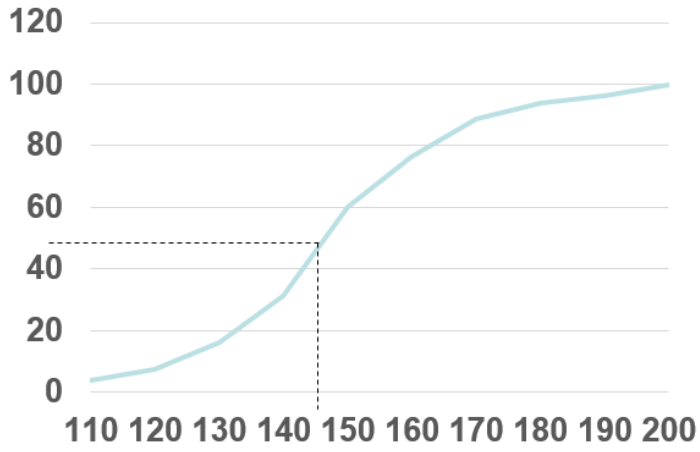



Рис.3.10 Графическое определение медианы

**Квартили** представляют собой значение признака, делящее ранжированную совокупность на четыре равные части. Различают квартиль нижний ($Q_1$), отделяющий 1/4 часть совокупности с наименьшими значениями признака, и квартиль верхний ($Q_3$), отсекающий 1/4 часть с наибольшими значениями признака (рис.3.11). Это означает, что 25% единиц совокупности будут меньше по величине $Q_1$; 25% единиц будут заключены между $Q_1$ и $Q_2$; 25% - между $Q_2$ и $Q_3$ и остальные 25% превзойдут $Q_3$. Вторая квартиль $Q_2$ является медианой. Вычисление квартилей аналогично вычислению медианы. Для расчета квартилей по интервальному вариационному ряду используются формулы:

$$Q_1 = x_{Q_1} + i\frac{\frac{1}{4}\sum f - S_{Q_1-1}}{f_{Q_1}},$$

$$Q_3 = x_{Q_3} + i\frac{\frac{3}{4}\sum f - S_{Q_3-1}}{f_{Q_3}}\frac{1}{2},$$

где $x_{Q_1}$ - нижняя граница интервала, содержащего нижний квартиль (интервал определяется по накопленной частоте, первой превышающей 25%);

$x_{Q_3}$ - нижняя граница интервала, содержащего верхний квартиль (интервал определяется по накопленной частоте, первой превышающей 75%);

$i$ - величина интервала;

$S_{Q_1-1}$ - накопленная частота интервала, предшествующего интервалу, содержащему нижний квартиль;

$S_{Q_3-1}$ - то же для верхнего квартиля;

$f_{Q_1}$ - частота интервала, содержащего нижний квартиль;

$f_{Q_3}$ - то же для верхнего квартиля.

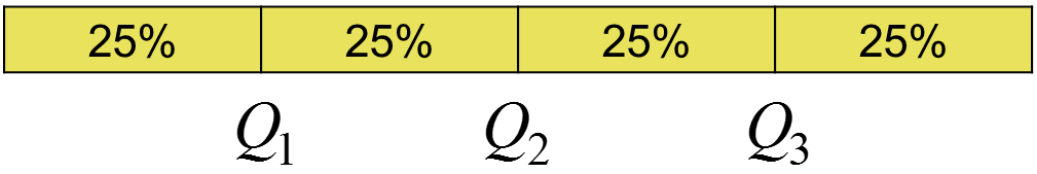



Рис.3.11 Квартили

Вычислим квартили по данным таблицы 3.29:

$$Q_1 = 130 + 10\frac{25 - 16,25}{15} = 135,83$$

$$Q_3 = 150 + 10\frac{75 - 60}{16,25} = 159,23.$$

## 3.5 Основные характеристики социальных и экономических сетей

Ранее мы рассматривали объекты (студенты, предприятия), не принимая во внимание связь этих элементов. Сетевые структуры обладают дополнительными характеристиками, основные из которых будут рассмотрены далее.

Исследование сетей обусловлено тем, что многие экономические, политические и социальные взаимодействия определяются структурой отношений элементов:

- торговля товарами и услугами;
- обмен информацией
- устройство на работу и т.д.

При этом можно выделить следующие задачи, решаемые в области исследования связанных элементов:

- нахождение кратчайшего пути между элементами;
- вычисление скорости распространения информации, развития сети;
- поиск элемента с наибольшим влиянием;
- выделение кластеров, определение структуры сети.

Сеть состоит из следующих элементов:

- N={1,…,n} – узлы, агенты, вершины, игроки;
- Ребра, связи между вершинами

При описании сети ребра могут принимать значение 0 (связь отсутствует) или 1 (связь между вершинами есть): если исследователи написали статью вместе, то связь есть и значение ребра принимается равным 1; если люди являются друзьями в социальных сетях, то это также говорит о наличии связи, значение ребра принимается равным 1.

Ребра могут иметь интенсивность, которая может представлять оценку степени взаимодействия объектов: время, которое люди проводят друг другом, оборот между странами и т.д.

Также связь может быть ориентированной или неориентированной. Например, при рассмотрении соавторов, друзей, родственников мы имеем дело с неориентированными ребрами (если человек А является родственником В, то и В в свою очередь является родственником А). В качестве примера ориентированной связи можно привести переход по ссылке, цитату.

Сеть может быть представлена в виде матрицы. На рис.3.12 представлена неориентированная, невзвешенная сеть. Ноль означает отсутствие связи, единица- её наличие. Список связей: g={12,14,24,34}.

$$g = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

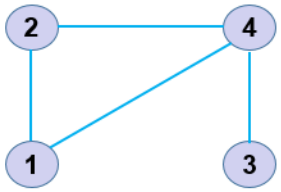



Рис.3.12 Неориентированная невзвешенная сеть

На рис.3.1.3 представлена ориентированная невзвешенная сеть. Список связей: g={12,14, 24, 41, 43}. Здесь порядок значений в паре играет роль.

$$g = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

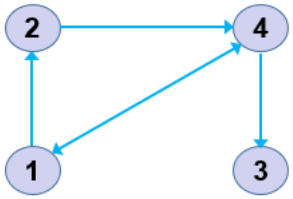



Рис.3.13 Ориентированная невзвешенная сеть

На рис.3.14 представлена взвешенная ориентированная сеть.

$$g = \begin{pmatrix} 0 & 7 & 2 \\ 5 & 0 & 0 \\ 0 & 4 & 0 \end{pmatrix}$$

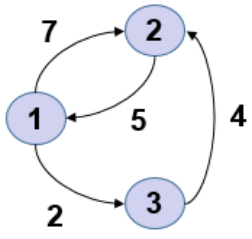



Рис.3.14 Ориентированная взвешенная сеть

Путь из $i_1$ в $i_k$ – это последовательность вершин ($i_1$,$i_2$,…,$i_k$) и ребер ($i_1i_2$, $i_2i_3$,…,$i_{k-1}i_k$) . Если $i_1$=$i_k$, то такой путь называют циклом. На рис. 3.15 а) путь от 1 до 7 включает следующие вершины: 1, 2, 3, 4, 5, 6, 7 , на рис. 3.15 б) можно увидеть цикл: 1, 2, 3, 1.


а) б)

Рис. 3.15 Путь сети

Рассмотрим основные характеристики сети и вершин.

**Диаметр**

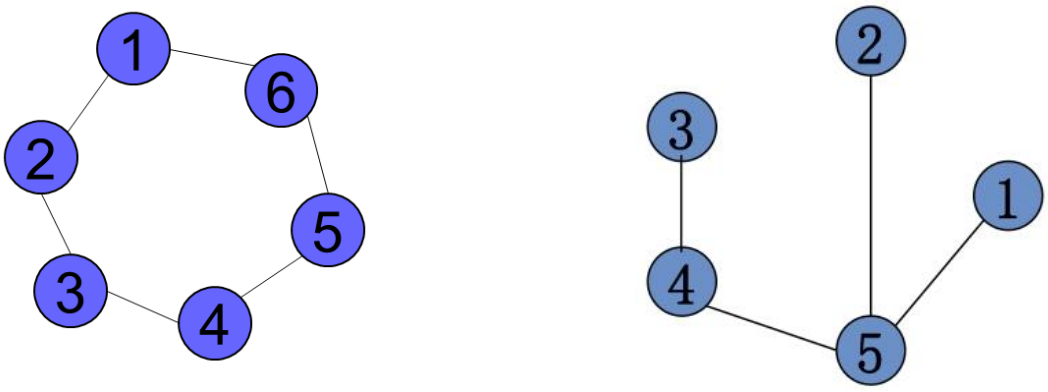
Диаметром называется наибольший кратчайший путь. Например, на рис.3.16 диаметр в случае 3.16 а) будет равен 3, т.к. это наибольший кратчайший путь (например, (1,2,3,4), (1,6,5,4)), в случае 3.16 б) диаметр также будет равен 3, т.к. существует два наибольших кратчайших пути длиною три: (1,5,4,3), (2,5,4,3).

**Степень**

Степень (*d(i)*)– число соседних вершин вершины *i*.

Данный показатель отражает количество связей вершины и является важным, например, при определении количества людей, с которыми вы можете делиться информацией, оценке влияния в науке (расчет числа цитирований) и т.д.

По данным рисунка 3.16 а) степень четвертой вершины будет равна двум, на рис. 3.16 б) степень пятой вершины будет равна трем.




а) б)

Рис.3.16 Графы

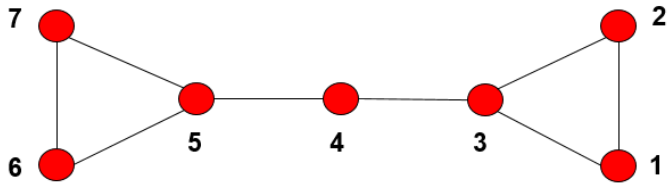
Также может быть вычислено относительное значение степени:

$$d_{rel}(i) = \frac{d(i)}{N-1},$$

где $N$ - число вершин в графе.

Т.е. степень узла $i$ делится на максимально возможное значение, т.е. на $N$-1.

На рис.3.17 представлена сеть и рассчитаны значения относительной степени узлов.




| | Узел 1 | Узел 3 | Узел 4 |
|---|---|---|---|
| Степень | 0,33 | 0,5 | 0,33 |

Рис.3.17 Пример расчета относительной степени узлов

Для первого узла значение вычислено по формуле:

$$d_{rel} = \frac{d}{N-1} = \frac{2}{6} = 0,33.$$

Для третьего

$$d_{rel} = \frac{d}{N-1} = \frac{3}{6} = 0,5$$

Для четвертого

$$d_{rel} = \frac{d}{N-1} = \frac{2}{6} = 0,33.$$

**Централизация**

Для оценки степени централизации сети используется формула Фримана:

$$d_f = \frac{\sum_{i=1}^{g}[d_{\max} - d(i)]}{[(N-1)(N-2)]},$$

где $d_{\max}$ - максимальное значение степени в сети.

На пример, для графа на рис.3.18 степень централизации будет равна:

$$d_f = \frac{\sum_{i=1}^{g}[d_{\max} - d(i)]}{[(N-1)(N-2)]} = \frac{(7-7)+7\cdot(7-1)}{7\cdot 6} = \frac{42}{42} = 1$$

А для графа на рис.3.16 а) степень централизации будет равна 0.

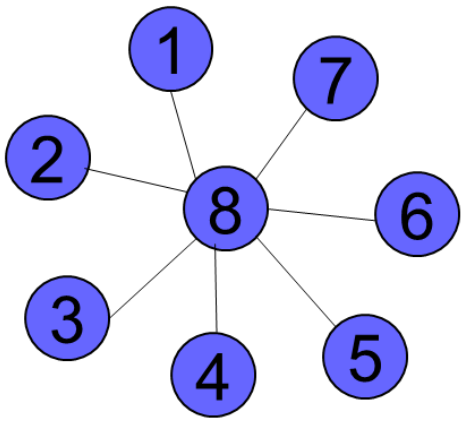



Рис.3.18 Граф

**Кластеризация**

Коэффициент кластеризации является показателем процентного соотношения соседних вершин связанных между собой (например, с помощью него можно определить, какой процент друзей конкретного человека в социальных сетях дружат между собой, рис.3.19). Формула расчета:

$$Cl(i) = \frac{k_{link}}{k},$$

где $k$-число пар соседних узлов.

$k_{link}$-число пар соседних узлов, связанных между собой.

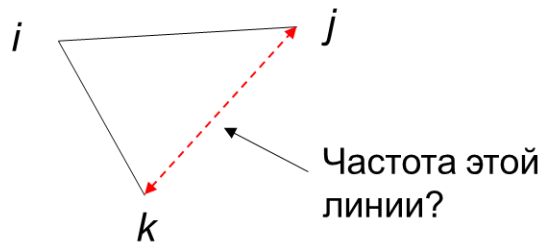



Рис.3.19 Кластеризация

Средняя кластеризация рассчитывается по формуле:

$$Cl_{avg} = \frac{\sum_{i} Cl(i)}{N}.$$

Например, коэффициент кластеризации вершины с номером пять (рис.3.20 а) будет равна 1/3. Узел «5» имеет три возможных пары соседей: 13, 12, 23. Из них только одна пара 13 связана между собой

**Скорость доступа к другим узлам**

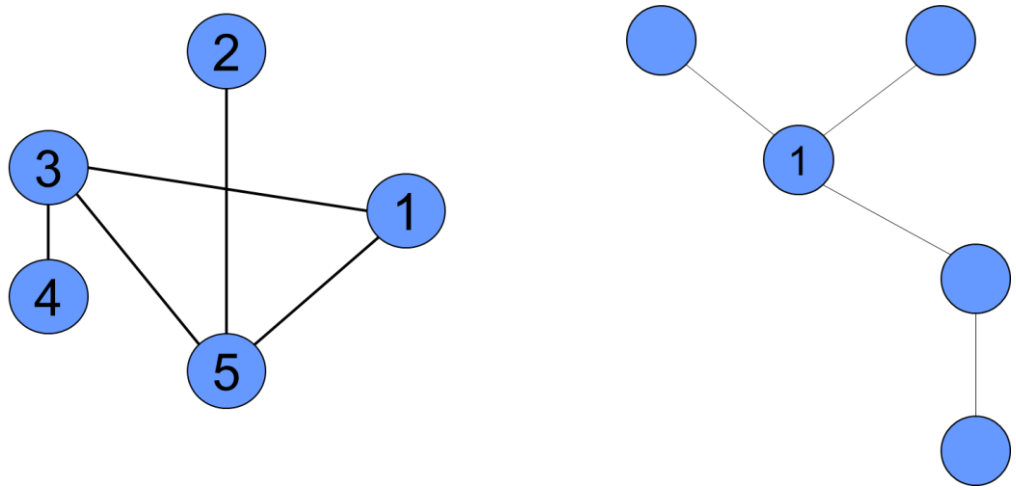
Относительная скорость доступа (closeness) определяется как величина обратная расстоянию до других узлов:

$$C_c(i) = \frac{N-1}{\sum_{j=1}^{N} l(i,j)},$$

где $l(i,j)$ -кратчайший путь из $i$ в $j$ .

Для графа на рис.3.20 б) относительная скорость доступа вершины с номером 1 к другим узлам равна:

$$C_c(1) = \frac{N-1}{\sum_{j=1}^{N} l(A,j)} = \frac{5-1}{1+1+1+2} = 0{,}8.$$




а) б)

Рис. 3.20 Пример графа

Относительная скорость доступа к другим узлам также может быть определена с помощью альтернативного показателя (decay):

$$C_d(i) = \sum_{j=1,\, j\neq i}^{N} \delta^{l(i,j)},\ 0 < \delta < 1$$

При расчете показателя берется некоторое значение на интервале от 0 до 1 и возводится в степень, равную длине от рассматриваемой вершины до другой вершины графа. Данное значение определяется для всех вершин графа (относительно рассматриваемой) и находится их сумма. Если дельта стремится к единице, то показатель равен размеру сети.

Для графа на рис.3.20 б) данный показатель для вершины с номером один (при $\delta = 0{,}5$) будет равен:

$$C_d(1) = \sum_{j=1}^{N} \delta^{l(1,j)} = 3 \cdot 0{,}5^1 + 0{,}5^2 = 1{,}75.$$

**Отношение «между»**

Характеристика соединяющей роли вершины (betweenness) показывает, у скольких пар узлов есть необходимость пройти через текущую вершину, чтобы путь между ними был кратчайшим. Формула расчета:

$$C_B(i) = \sum_{j<k} nl_i(j,k) / nl(j,k),$$

где $nl_i(j,k)$ - число кратчайших путей между $j$ и $k$, проходящих через $i$ ;

$nl(j,k)$ - число кратчайших путей между $j$ и $k$;

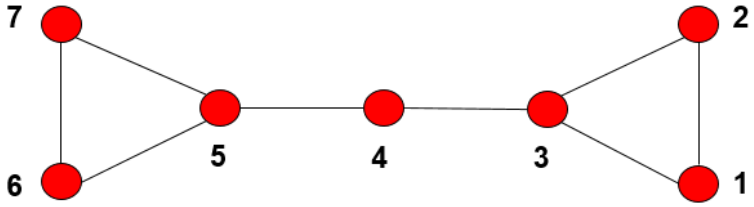
Нормированное значение данного показателя вычисляется по формуле:

$$C'_B(i) = C_B(i\ ) / [(N-1)(N-2)/2]$$

На рис.3.21 представлены значения $nl_i(j,k)$ - число кратчайших путей между $j$ и $k$, проходящих через $i$.


Рис.3.21 Значения $nl_i(j,k)$

На рис.3.22 представлен граф и нормированные значения показателя.




| | Узел 1 | Узел 3 | Узел 4 |
|---|---|---|---|
| betweenness | 0,00 | 0,53 | 0,6 |

Нормированное значение показателя для первой вершины равно:

$$C'_B(1) = \sum_{j<k} l_1(j,k)/l(j,k)/[(N-1)(N-2)/2] = \left(\frac{0}{1}\right)/15 = 0.$$

Нормированное значение показателя для третьей вершины равно:

$$C'_B(3) = \left(\frac{8}{1}\right)/15 = 0{,}53.$$

Всего 15 кратчайших путей (между вершинами за исключением третьей), из них 8 проходят через вершину 3: 14,15,16,17,24,25,26,27.

Нормированное значение показателя для четвертой вершины равно:

$$C'_B(4) = \left(\frac{9}{1}\right)/15 = 0{,}6.$$

Всего 15 кратчайших путей, из них 9 проходят через вершину 4: 35,36,37,15,16,17,25,26,27.

# 4. Выборочное наблюдение

Статистическая методология исследования массовых явлений различает два способа наблюдения в зависимости от полноты охвата объекта: сплошное и несплошное. Разновидностью несплошного наблюдения является выборочное.

Выборочным называется такое несплошное наблюдение, при котором признаки регистрируются у отдельных единиц изучаемой статистической совокупности, отобранных с использованием специальных методов, а полученные в процессе обследования результаты с определенным уровнем вероятности распространяются на всю исходную совокупность.

Реализация выборочного метода базируется на понятиях генеральной и выборочной совокупностей. Генеральной совокупностью называется вся исходная изучаемая статистическая совокупность, из которой на основе отбора единиц или групп единиц формируется совокупность выборочная. Поэтому генеральную совокупность также называют основой выборки. Отбор единиц в выборочную совокупность может быть повторным или бесповторным. При повторном отборе попавшая в выборку единица подвергается обследованию, т.е. регистрации значений ее признаков, возвращается в генеральную совокупность и наравне с другими единицами участвует в дальнейшей процедуре отбора. Таким образом некоторые единицы могут попадать в выборку дважды, трижды или даже большее число раз. И при изучении выборочной совокупности они будут рассматриваться как отдельные независимые наблюдения.

При бесповторном отборе попавшая в выборку единица подвергается обследованию и в дальнейшей процедуре отбора не участвует. Такой отбор целесообразен и практически возможен в тех случаях, когда объем генеральной совокупности четко определен. Получаемые при этом результаты, как правило, являются более точными по сравнению с результатами, основанными на повторной выборке. Необходимо отметить, что в выборочную совокупность могут отбираться не только отдельные единицы, но и группы единиц. В первом случае отбор называется индивидуальным, во втором случае - групповым.

Выборочное наблюдение всегда связано с определенными ошибками получаемых характеристик. *Ошибки регистрации* являются следствием неправильного установления значения наблюдаемого признака или неправильной записи. Они свойственны не только выборочному, но и сплошному наблюдению.

*Ошибки репрезентативности* обусловлены тем, что выборочная совокупность не может по всем параметрам в точности воспроизвести генеральную совокупность. Получаемые расхождения называются ошибками репрезентативности, или представительности, так как они отражают, в какой степени попавшие в выборку единицы могут представлять всю генеральную совокупность. При этом следует различать систематические и случайные ошибки репрезентативности.

Систематические ошибки репрезентативности связаны с нарушением принципов формирования выборочной совокупности. Например, вследствие каких-либо причин, связанных с организацией отбора, в выборку попали единицы, характеризующиеся несколько большими или, наоборот, несколько меньшими по сравнению с другими единицами значениями наблюдаемых признаков. В этом случае и рассчитанные выборочные характеристики будут завышенными или заниженными. Случайные ошибки репрезентативности обусловлены действием случайных факторов, не содержащих каких-либо элементов системности в направлении воздействия на рассчитываемые выборочные характеристики. Но даже при строгом соблюдении всех принципов формирования выборочной совокупности выборочные и генеральные характеристики будут несколько различаться. Получаемые случайные ошибки могут быть статистически оценены и учтены при распространении результатов выборочного наблюдения на всю генеральную совокупность. Оценка ошибок выборочного наблюдения основана на теоремах теории вероятностей.

Средняя ошибка выборки вычисляется по формуле:

$$\mu = \sqrt{\frac{\sigma_{\bar{x}}^2}{n}},$$

где $\sigma_{\tilde{x}}^2$ - генеральная дисперсия;

$n$ - объем выборки.

При определении возможных границ значений характеристик генеральной совокупности рассчитывается предельная ошибка выборки, которая зависит от величины ее средней ошибки и уровня вероятности, с которым гарантируется, что генеральная средняя не выйдет за указанные границы. Согласно теореме А.М. Ляпунова, вероятность той или иной величины предельной ошибки, при достаточно большом объеме выборочной совокупности, подчиняется нормальному закону распределения и может быть определена на основе интеграла Лапласа:

$$P\left\{\left|\overline{x}-\tilde{x}\right| \le \Delta_{\tilde{x}}\right\} = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{\frac{-t^2}{2}} dt = \Phi(t)$$

где $\Delta_{\tilde{x}}$ -предельная ошибка выборки;

$\overline{x}$ - генеральное среднее;

$\tilde{x}$ - выборочное среднее.

При обобщении результатов выборочного наблюдения наиболее часто используют следующие уровни вероятности и соответствующие им значения *t*:

$t = 1,\ \Phi(t) = 0,683$

$t = 1,96,\ \Phi(t) = 0,95$

$t = 2,\ \Phi(t) = 0,954$

$t = 3,\ \Phi(t) = 0,997$ .

Например, если при определении предельной ошибки выборки мы используем t = 2, то с вероятностью P = 0,954 можно утверждать, что расхождение между выборочной и генеральной средними не превысит двухкратной величины рассчитанной средней ошибки выборки.

Различают следующие виды выборки:

- собственно-случайная;
- механическая;

- типическая;
- серийная;
- комбинированная.

## 4.1 Случайная выборка

Собственно-случайная выборка заключается в отборе единиц из генеральной совокупности в целом, без разделения ее на группы, подгруппы или серии отдельных единиц. При этом единицы отбираются в случайном порядке, не зависящем ни от последовательности расположения единиц в совокупности, ни от значений их признаков.

После проведения отбора с использованием какого-либо алгоритма, реализующего принцип случайности, или на основе таблицы случайных чисел, необходимо определить границы генеральных характеристик. Для этого рассчитываются средняя и предельная ошибки выборки.

Средняя ошибка повторной собственно-случайной выборки определяется по формуле:

$$\mu_{\tilde{x}} = \frac{\sigma_{\bar{x}}}{\sqrt{n}}.$$

С учетом выбранного уровня вероятности и соответствующего ему значения t предельная ошибка выборки составит:

$$\Delta_{\tilde{x}} = t\mu.$$

Тогда можно утверждать, что при заданной вероятности генеральная средняя будет находиться в следующих границах:

$$\tilde{x} - \Delta_{\tilde{x}} \le \bar{x} \le \tilde{x} + \Delta_{\tilde{x}}.$$

Пусть исходные данные приведены в табл.4.1 Выборка включает всего два элемента: Иван, Павел. Необходимо с вероятностью 0,954 определить границы изменения генеральной средней в случае повторной выборки (т.е. возможна ситуация, когда мы два раза обследуем одного и того же человека).

Таблица 4.1 Данные о затратах времени на чтение

| Имя студента | Затраты времени на чтение в среднем за день, *мин* |
|---|---|
| Иван | 10 |
| Петр | 20 |
| Александр | 40 |
| Николай | 50 |
| Павел | 80 |

Для этого вычислим отклонения признака от среднего значения (табл.4.2).

Таблица 4.2 Вычисление квадрата отклонения признака от среднего значения

| Выборки | Затраты времени на чтение в среднем за день, *мин* | $(x-\overline{x})^2$ |
|---|---|---|
| Иван | 10 | 900 |
| Петр | 20 | 400 |
| Александр | 40 | 0 |
| Николай | 50 | 100 |
| Павел | 80 | 1600 |
| *Итого* | 200 | 3000 |

Генеральное среднее значение равно:

$$\overline{x} = \frac{200}{5} = 40.$$

Генеральная дисперсия:

$$\sigma_{ген} = \sqrt{\frac{3000}{5}} = 24,5.$$

Средняя ошибка равна:

$$\mu_{\tilde{x}} = \frac{\sigma_{ген}}{\sqrt{n}} = \frac{24,5}{\sqrt{2}} = 17,3$$

Предельная ошибка составит:

$$\Delta_{\tilde{x}} = 2 \cdot 17,3 = 34,6.$$

Выборочное среднее для двух наблюдений (Иван, Павел) равно:

$$\tilde{x} = \frac{10+80}{2} = 45$$

Следовательно, границы изменения генеральной средней равны:

$45-34,6 \le \overline{x} \le 45+34,6$

$10,4 \le \overline{x} \le 79,6$.

Рассчитанное выше генеральное среднее ($\overline{x}=40$) попадает в полученный интервал.

Для бесповторной выборки средняя ошибка вычисляется по формуле:

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

Рассмотрим тот же пример (табл.4.1) в случае бесповторной выборки.

Средняя ошибка будет равна:

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma^2_{ген}}{n}\cdot\left(1-\frac{n}{N}\right)} = \frac{24,5}{\sqrt{2}}\cdot\sqrt{1-\frac{2}{5}} = 13,4$$

Предельная ошибка составит:

$\Delta_{\tilde{x}} = 2\cdot 13,4 = 26,8$.

Следовательно, границы изменения генеральной средней (в случае если в выборку попали Иван и Павел) равны:

$45-26,8 \le \overline{x} \le 45+26,8$

$18,2 \le \overline{x} \le 71,8$.

**Альтернативный признак**

Мы рассмотрели определение границ генеральной средней. Рассмотрим теперь, как определяются границы генеральной доли. Для повторной выборки нужно выполнить расчет средней ошибки, используя формулу

$$\mu = \sqrt{\frac{w(1-w)}{n}},$$

где $w$ - выборочная доля;

$n$ - объем выборки.

Для бесповторной формула средней ошибки имеет вид:

$$\mu = \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}.$$

Предельная ошибка и границы изменения доля определяются следующим образом:

$$\Delta_w = t\mu$$

$$w - \Delta_w \le p \le w + \Delta_w .$$

Пусть по данным таблицы необходимо с вероятностью 0,683 определить границы доли единиц, которые тратят на чтение менее 25 минут в день (повторная выборка). Пусть объем выборки равен трем и в выборку попали: Иван, Павел, Павел.

Выборочная доля будет равна (из трех наблюдений только один раз было зафиксировано, что человек тратит на чтение менее 25 минут в день)

$$w = \frac{n_p}{n} = \frac{1}{3} = 0,33 .$$

Дисперсия будет равна

$$\sigma_w^2 = w(1-w) = 0,33 \cdot (1-0,33) = 0,22 .$$

Средняя ошибка

$$\mu = \sqrt{\frac{w(1-w)}{n}} = \sqrt{\frac{0,22}{3}} = 0,27 .$$

Предельная ошибка:

$$\Delta_w = t\mu = 1 \cdot 0,27 = 0,27 .$$

Границы генеральной доли:

$$w - \Delta_w \le p \le w + \Delta_w$$

$$0,33 - 0,27 \le p \le 0,33 + 0,27$$

$$0,06 \le p \le 0,6 .$$

Вычислим реальное значение генеральной доли:

$p = \frac{2}{5} = 0,4$ (из пяти человек только двое тратят на чтение менее 25 минут).

Значение генеральной доли попало в полученный интервал.

**Определение объема выборки**

Чем больше объем выборки, тем меньше значения средней и предельной ошибок выборочного наблюдения и, следовательно, тем уже границы генеральной средней и генеральной доли. В то же время необходимо учитывать, что большой объем выборки приводит к удорожанию обследования, увеличению сроков сбора и обработки материалов, требует привлечения дополнительного персонала и соответствующего материально-технического обеспечения. Поэтому при подготовке выборочного наблюдения необходимо определить тот минимально необходимый объем выборки, который обеспечит требуемую точность полученных статистических характеристик при заданном уровне вероятности. Представим формулу предельной ошибки при повторном отборе следующим образом:

$$\Delta = t \cdot \sqrt{\frac{\sigma^2}{n}},$$

Отсюда можно вывести формулу для определения необходимого объема собственно-случайной повторной выборки:

$$n = \frac{t^2 \sigma^2}{\Delta^2}.$$

Рассмотрим пример. Для определения средней длины детали следует провести выборочное обследование методом случайного повторного отбора. Какое количество деталей надо отобрать, чтобы ошибка выборки не превышала 3 мм с вероятностью 0,997 (t=3) при среднем квадратическом отклонении 6 мм.

Необходимый объем выборки (число деталей):

$$n = \frac{t^2 \sigma^2}{\Delta^2} = \frac{3^2 \cdot 6^2}{3^2} = 36.$$

Необходимый объем выборки будет тем больше, чем выше заданный уровень вероятности и чем сильнее варьирует наблюдаемый признак. В то же время повышение допусти мой предельной ошибки выборки приводит к снижению необходимого ее объема.

Необходимый объем собственно-случайной бесповторной выборки может быть определен по следующей формуле:

$$n = \frac{t^2\sigma^2 N}{\Delta^2 N + t^2\sigma^2}.$$

Пусть в микрорайоне проживает 5000 семей. В порядке случайной бесповторной выборки предполагается определить средний размер семьи при условии, что ошибка выборочной средней не должна превышать 0,8 человека с вероятностью P=0,954 ($t$=2) при среднем квадратическом отклонении 3 человека.

В качестве решения задачи получим:

$$n = \frac{t^2\sigma^2 N}{\Delta^2 N + t^2\sigma^2} = \frac{2^2 \cdot 3^2 \cdot 5000}{5000 \cdot 0,64 + 2^2 \cdot 3^2} = \frac{180000}{3236} = 56$$

## 4.2 Механическая выборка

Механическая выборка может быть применена в тех случаях, когда генеральная совокупность каким-либо образом упорядочена, т.е. имеется определенная последовательность в расположении единиц (табельные номера работников, списки избирателей, телефонные номера респондентов, номера домов и квартир и т.п.). Для проведения отбора желательно, чтобы все единицы также имели порядковые номера от 1 до $N$. Для проведения механической выборки устанавливается пропорция отбора, которая определяется соотнесением объемов выборочной и генеральной совокупностей. Так, если из совокупности в 500000 ед. предполагается отобрать 10000 ед., то пропорция отбора составит 1/50 = 1/(500 000:10 000). Отбор единиц осуществляется в соответствии с установленной пропорцией через равные интервалы. Например, при пропорции 1:50 (2%-я выборка) отбирается каждая 50-я единица, при пропорции 1:20 (5%-я выборка) - каждая 20-я единица и т.д.

Для определения средней ошибки механической выборки, а также необходимой ее численности используются соответствующие формулы, применяемые при собственно-случайном бесповторном отборе.

Пусть у нас содержится упорядоченный по именам список групп студентов. Выборка предполагает, что будет опрошен каждый второй студент (рис.4.1).

| Единицы генеральной совокупности | Затраты времени на чтение в среднем за день, *мин* |
|---|---|
| Александр | 40 |
| Иван | 10 |
| Николай | 50 |
| Павел | 80 |
| Петр | 20 |

Рис.4.1 Выборка студентов

Тогда для представленных данных средняя ошибка будет равна (дисперсия была рассчитана выше):

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma^2_{ген}}{n} \cdot \left(1 - \frac{n}{N}\right)} = \sqrt{\frac{24,5^2}{3} \cdot \left(1 - \frac{3}{5}\right)} = 8,9 .$$

Предельная ошибка:

$\Delta_{\tilde{x}} = 2 \cdot 8,9 = 17,9 .$

Выборочное среднее:

$$\tilde{x} = \frac{40 + 50 + 20}{3} = 36,7$$

Границы генеральной средней:

$36,7 - 17,9 \le \bar{x} \le 36,7 + 17,9$

$18,8 \le \bar{x} \le 54,6 .$

## 4.3 Типический отбор

Типический отбор целесообразно использовать в тех случаях, когда все единицы генеральной совокупности объединены в несколько крупных типических групп. Такие группы также называют стратами, или слоями, в связи с чем типический отбор также называют стратифицированным, или

расслоенным. При обследовании населения в качестве типических групп могут быть выбраны области, районы, социальные, возрастные или образовательные группы, при обследовании предприятий - отрасли или подотрасли, формы собственности и т.п.

Отбор единиц в выборочную совокупность из каждой типической группы осуществляется собственно-случайным или механическим способом.

Отбор единиц в типическую выборку может быть организован либо пропорционально объему типических групп, либо пропорционально внутригрупповой вариации (дифференциации) признака.

При типической выборке, *пропорциональной объему типических групп*, число единиц, подлежащих отбору из каждой группы, определяется следующим образом:

$$n_i = n\frac{N_i}{N},$$

где $N_i$ - объем $i$ -й группы;

$n_i$ - объем выборки из $i$ -й группы.

Средняя ошибка типической выборки определяется по формулам:

$\mu = \sqrt{\frac{\bar{\sigma}^2}{n}}$ (повторный отбор),

$\mu = \sqrt{\frac{\bar{\sigma}^2}{n}\left(1-\frac{n}{N}\right)}$ (бесповторный отбор),

где $\bar{\sigma}^2$ - средняя из внутригрупповых дисперсий.

Рассмотрим пример. Пусть необходимо обследовать две группы студентов (рис.4.2). Предполагается, что выборка будет включать 10 студентов (выборка бесповторная).

Вычислим дисперсию для первой группы:

$$\bar{x} = \frac{4+3,5+3+4+3,7+2,6+3}{7} = 3,4$$

$$\sigma_1^2 = \frac{\sum_{i=1}^{7}(x-\bar{x})^2}{7} = \frac{0,36+0,01+0,16+...+0,16}{7} = 0,25.$$

Группа 1

| № студента | Средний балл |
|---|---|
| 1 | 4 |
| 2 | 3,5 |
| 3 | 3 |
| 4 | 4 |
| 5 | 3,7 |
| 6 | 2,6 |
| 7 | 3 |

Группа 2

| № студента | Средний балл |
|---|---|
| 1 | 4 |
| 2 | 3,5 |
| 3 | 5 |
| 4 | 4 |
| 5 | 3,7 |
| 6 | 5 |
| 7 | 3,8 |
| 8 | 3 |
| 9 | 5 |
| 10 | 3 |

Рис.4.2 Исходные данные об успеваемости студентов

Аналогично вычислим дисперсию для второй группы:

$$\bar{x} = \frac{4+3,5+5+4+3,7+...+3}{10} = 4,$$

$$\sigma_2^2 = \frac{\sum_{i=1}^{10}(x-\bar{x})^2}{10} = \frac{0+0,25+1+...+1}{10} = 0,54.$$

Определим, сколько студентов нужно обследовать из первой группы:

$$n_1 = 10 \cdot \frac{7}{17} = 4.$$

Из второй группы:

$$n_2 = 10 \cdot \frac{10}{17} = 6$$

Пусть случайным образом выбранные номера студентов из первой группы: 1,4,6,7, из второй – 1,3,6,8,9,10. Вычислим выборочные средние значения:

$$\tilde{x}_1 = \frac{4+4+2,6+3}{4} = 3,4.$$

$$\tilde{x}_2 = \frac{4+5+5+3+5+3}{6} = 4,2.$$

Сводные данные представлены в таблице 4.3.

Таблица 4.3 Сводные данные

| Группа | Всего студентов | Опрошено | Средний балл за последнюю сессию | |
|---|---|---|---|---|
| | | | средняя | дисперсия |
| I | 7 | 4 | 3,4 | 0,25 |
| II | 10 | 6 | 4,2 | 0,54 |
| Всего | 17 | 10 | | |

Вычислим среднюю из внутригрупповых дисперсий:

$$\overline{\sigma}^2 = \frac{\sum \sigma_i^2 n_i}{\sum n_i} = \frac{0,25 \cdot 4 + 0,54 \cdot 6}{4+6} = 0,42.$$

Средняя ошибка равна:

$$\mu_{\tilde{x}} = \sqrt{\frac{0,42}{10}\left(1 - \frac{10}{17}\right)} = 0,13.$$

Предельная ошибка (для определения интервала с вероятностью 0,954):

$$\Delta_{\tilde{x}} = 2 \cdot 0,13 = 0,26$$

Рассчитаем выборочную среднюю:

$$\tilde{x} = \frac{\sum x_i n_i}{\sum n_i} = \frac{3,4 \cdot 4 + 4,2 \cdot 6}{10} = 3,88.$$

Границы изменения генеральной средней:

$$3,88 - 0,26 \le \overline{x} \le 3,88 + 0,26$$

$$3,62 \le \overline{x} \le 4,14.$$

Второй вариант формирования типической выборки заключается в отборе единиц, *пропорциональном вариации при знака в типических группах*. Логика такого отбора заключается в следующем: если внутри какой-либо типической группы наблюдаемый признак варьирует слабо, то для определения границ генеральных характеристик из данной группы достаточно обследовать относительно небольшое число единиц; при сильной же вариации признака

объем выборки должен быть соответственно увеличен. При выборке, пропорциональной вариации признака, число наблюдений по каждой группе рассчитывается по формуле:

$$n_i = n\frac{\sigma_i N_i}{\sum \sigma_i N_i}.$$

где $\sigma_i$ - среднее квадратическое отклонение признака в *i*-й группе.

Средняя ошибка такого отбора определяется следующим образом:

$$\mu = \frac{1}{N}\sqrt{\sum \frac{\sigma_i^2 N_i^2}{n_i}}$$ (повторный отбор)

$$\mu = \frac{1}{N}\sqrt{\sum \frac{\sigma_i^2 N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)}$$ (бесповторный отбор).

Отбор, пропорциональный вариации признака, дает лучшие результаты, однако на практике его применение затруднено из-за трудности получения сведений о вариации до проведения выборочного наблюдения. Воспользуемся данными о размере генеральной совокупности и дисперсии, приведенными в табл.4.3, для иллюстрации этого способа выборочного наблюдения (повторный отбор).

$$\sum \sigma_i N_i = \sqrt{0,25}\cdot 7 + \sqrt{0,54}\cdot 10 = 10,86$$

$$n_1 = 10\cdot\frac{\sqrt{0,25}\cdot 7}{10,86} = 3$$

$$n_2 = 10\cdot\frac{\sqrt{0,54}\cdot 10}{10,86} = 7.$$

Пусть случайным образом выбранные номера студентов из первой группы: 1,2,7, из второй – 1,2,5,6,8,9,10. Вычислим выборочные средние значения:

$$\tilde{x}_1 = \frac{4+3,5+3}{3} = 3,5.$$

$$\tilde{x}_2 = \frac{4+3,5+3,7+5+3+5+3}{7} = 3,9.$$

Получим таблицу 4.4.

Таблица 4.4

| Группа | Всего студентов | Опрошено | Средний балл за последнюю сессию | |
|---|---|---|---|---|
| | | | средняя | дисперсия |
| I | 7 | 3 | 3,5 | 0,25 |
| II | 10 | 7 | 3,9 | 0,54 |
| Всего | 17 | 10 | | |

Вычислим среднюю ошибку:

$$\mu_{\tilde{x}} = \frac{1}{17}\sqrt{\frac{0,25 \cdot 7^2}{3} + \frac{0,54 \cdot 10^2}{7}} = 0,2 .$$

Предельная ошибка (для определения интервала с вероятностью 0,954):

$\Delta_{\tilde{x}} = 2 \cdot 0,2 = 0,4 .$

Рассчитаем выборочную среднюю:

$$\tilde{x} = \frac{\sum x_i n_i}{\sum n_i} = \frac{3,5 \cdot 3 + 3,9 \cdot 7}{10} = 3,78$$

Границы изменения генеральной средней:

$3,78 - 0,4 \le \bar{x} \le 3,78 + 0,4$

$3,38 \le \bar{x} \le 4,18.$

## 4.4 Серийная выборка

Сущность серийной выборки заключается в собственно-случайном либо механическом отборе групп единиц (серий), внутри которых производится сплошное обследование. Единицей отбора при этой выборке является группа или серия, а не отдельная единица генеральной совокупности, как это имело место в рассматриваемых ранее выборках. Данный способ отбора удобен в тех случаях, когда единицы генеральной совокупности изначально объединены в небольшие более или менее равновеликие группы или серии. В качестве таких серий могут

выступать упаковки с определенным количеством готовой продукции, партии товара, студенческие группы, бригады и другие подобные объединения. Например, в Великобритании серийный отбор используется в обследованиях населения, когда серией являются домохозяйства, объединенные общим почтовым индексом. В случайном порядке про изводится выборка индексов, и под обследование попадают все домохозяйства, имеющие индекс попавших в выборочную совокупность почтовых отделений.

В отдельных случаях серийная выборка имеет не столько методологические, сколько организационные преимущества перед другими способами формирования выборочной совокупности. Например, Управление маркетинга и регионального развития Московского государственного университета экономики, статистики и информатики периодически проводит обследования школьников Москвы. С организационной точки зрения достаточно сложно опрашивать отдельных учеников из разных классов. Значительно проще из общего списка всех классов всех школ округа сформировать выборку классов, а внутри отобранных классов провести 100%-е обследование учащихся.

В связи с тем, что при серийном отборе внутри отобранных групп обследуются все без исключения единицы, внутригрупповая вариация признака не отразится на ошибках выборочного наблюдения. В то же время обследуются не все группы, а только попавшие в выборку. Следовательно, на ошибках получаемых характеристик отразятся различия между группами, которые определяются межгрупповой дисперсией. Поэтому средняя ошибка серийной выборки рассчитывается по формулам

$$\mu = \sqrt{\frac{\delta^2}{r}} \text{ (повторный отбор)}$$

$$\mu = \sqrt{\frac{\delta^2}{r}\left(1 - \frac{r}{R}\right)} \text{ (бесповторный отбор) .}$$

Межгрупповую дисперсию при равновеликих группах вычисляют следующим образом:

$$\delta^2 = \frac{\sum(\tilde{x}_i - \tilde{x})^2}{r},$$

где $\tilde{x}_i$ - средняя *i*-й серии;

$\tilde{x}$ - общая средняя по всей выборочной совокупности.

Предположим, партия готовой продукции предприятия упакована в 200 коробок по 50 изделий в каждой. В целях контроля соблюдения параметров технологического процесса проведена 5%-я серийная выборка, в ходе которой отбиралась каждая 20-я коробка. Все изделия, находящиеся в отобранных упаковках, были подвергнуты сплошному обследованию, заключавшемуся в определении их точного веса. Полученные результаты представлены в табл.4.5.

Таблица 4.5 Данные об изделиях

| Номер коробки | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Средний вес изделия в коробке, г. | 997 | 1001 | 1003 | 998 | 1000 | 1000 | 998 | 999 | 1000 | 1002 |

С вероятностью 0,954 требуется определить границы среднего веса изделия во всей партии. На основе приведенных в таблице внутригрупповых средних определим средний вес изделия по выборочной совокупности:

$$\tilde{x} = \frac{997 + 1001 + .. + 1002}{10} = 999{,}8 \text{ г.}$$

С учетом полученной средней рассчитаем межгрупповую дисперсию:

$$\delta^2 = \frac{(997 - 999{,}8)^2 + (1001 - 999{,}8)^2 + ... + (1002 - 999{,}8)^2}{10} = 3{,}16.$$

Рассчитаем среднюю и предельную ошибки выборки:

$$\mu = \sqrt{\frac{3{,}16}{10}\left(1 - \frac{10}{200}\right)} = 0{,}55,$$

$$\Delta_{\tilde{x}} = 2 \cdot 0{,}55 = 1{,}1.$$

Границы изменения генеральной средней:

$$999,8-1,1 \le \overline{x} \le 999,8+1,1$$

$$998,7 \le \overline{x} \le 1000,9$$

На основе результатов проведенных расчетов с вероятностью 0,954 можно утверждать, что средний вес изделия в целом по всей партии продукции находится в пределах от 998,7 г до 1000,9 г.

## 4.5 Определение вариации

В приведенных расчетах дисперсия была вычислена, либо уже дана в условии, на практике же часто возникает трудность с определением вариации. Источниками данных для оценки вариации могут служить:

- результаты обследования данного объекта в предшествующие периоды;
- результаты обследования аналогичных объектов (жителей других населенных пунктов, предприятий других регионов и т.п.);
- специально проведенное небольшое по объему выборочное обследование данного объекта, ставящее целью лишь изучение вариации наблюдаемых признаков.

Вариацию признака можно вычислить приближенно по величине предполагаемого размаха или среднего линейного отклонения по формулам:

$$\sigma = \frac{R}{6},$$

$$\sigma = 1,25 \cdot \overline{d},$$

где $R$ - размах вариации;

$\overline{d}$ -среднее линейное отклонение;

$\sigma$ - среднее квадратическое отклонение.

Важным условием практического применения этих формул является близость фактического распределения нормальному.

Если расчет проводится по качественному альтернативному признаку и неизвестна его доля в генеральной совокупности, рекомендуется принять её равной 0,5, т.к. дисперсия доли достигает максимума $\sigma_w^2 = 0,25$ при $w = 0,5$.

# 5. Статистические индексы

**Индекс** – обобщенный показатель, сконструированный специальным образом и применяемый для наблюдения за количественными характеристиками социальных объектов, явлений или событий (табл. 5.1).

Таблица 5.1 Показатели деятельности предприятия

| Показатель | Предприятие 1 | Предприятие 2 |
|---|---|---|
| Прибыль | 5 | 10 |
| Объем продаж | 3 | 5 |
| Число сотрудников | 7 | 2 |

Подобную систему формирования интегрального показателя на основе нескольких показателей и ранжирования полученных значения ещё называют рейтинговой (рис.5.1).

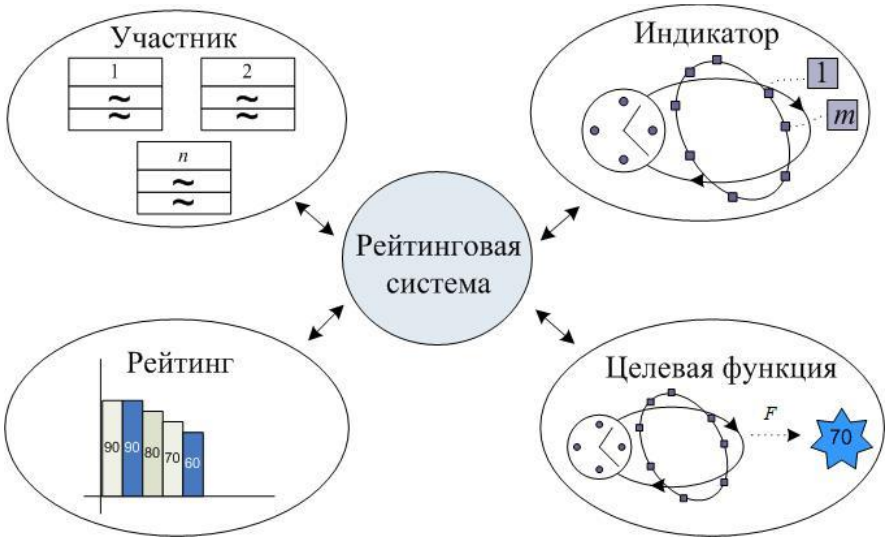



Рис.5.1 – Рейтинговая система

***Участник системы*** – это исследуемый объект, для которого выполняется расчет рейтинга. В качестве такого объекта может выступать предприятие, регион страны, высшее учебное заведение и т.д.

***Индикатор*** - показатель, характеризующий определенное свойство участника (прибыль, средний доход населения и т.д.)

***Целевая функция*** – правило преобразования индикаторов в интегральную характеристику с целью её сравнения с другими показателями.

***Рейтинг*** – число, полученное путем преобразования индикаторов в единый показатель.

Формирования интегрального показателя осуществляется с помощью следующих этапов:

- нормирование индикаторов;
- расчет интегрального показателя.

Нормирование может осуществляться различными способами.

В методе эталонного значения в исходных данных по каждому показателю определяется максимальный элемент, который принимается за единицу. Затем все показатели $(a_{ij})$ делятся на максимальный элемент предприятия-эталона $(\max a_{ij})$. В результате создается матрица стандартизованных коэффициентов:

$$x_{ij} = \frac{a_{ij}}{\max a_{ij}}$$

Эталонное предприятие формируется обычно из совокупности однородных объектов, принадлежащих к одной отрасли. Однако это не исключает возможности выбора предприятия-эталона из совокупности предприятий, принадлежащих к различным отраслям деятельности, так как многие финансовые показатели сопоставимы и для разнородных субъектов хозяйствования.

Если с экономической стороны лучшим является минимальное значение показателя (например, затраты на рубль продукции), то надо изменить шкалу расчета так, чтобы наименьшему результату соответствовала наибольшая сумма показателя:

$$x_{ij} = \frac{\min a_{ij}}{a_{ij}}$$

На рис.5.2 представлен пример нормирования показателей.

| Показатель | Предприятие 1 | Предприятие 2 |
|---|---|---|
| Прибыль, тыс.руб. | 500 | 250 |
| Число бракованных изделий | 60 | 45 |

| Нормирован. показатель | Предприятие 1 | Предприятие 2 |
|---|---|---|
| Прибыль, тыс.руб. | 1 | 0,5 |
| Число бракованных изделий | 0,75 | 1 |

Рис.5.2 Нормирование показателей методом эталонного предприятия

С экономической точки зрения максимальное значение прибыли является лучшим показателем, поэтому первую строчку мы делим на 500 (максимальное значение). Для числа бракованных изделий лучшим является наименьшее значение, поэтому во второй строке минимальное значение (45) делится на значения второй строки. В итоге полученные значения лежат в интервале от 0 до 1 (1 соответствует наилучшему значению показателя).

Также возможен вариант нормирования, когда от значения элемента отнимается среднее значения и делится на среднее квадратическое отклонение (способ предлагает, что увеличение каждого показателя – это положительная тенденция):

$$x_{ij} = \frac{a_{ij} - \bar{a}_j}{s_{a_j}}$$

где $s_{a_j}$ среднее квадратическое отклонение индикатора;

$\bar{a}_j$ - среднее значение индикатора;

$x_{ij}$ - нормированное значение индикатора.

На рис.5.3 представлен пример нормирования показателей путем вычитания среднего значения и деления на среднее квадратическое отклонение (СКО).

| Показатель | Предприятие 1 | Предприятие 2 | Среднее | СКО |
|---|---|---|---|---|
| Прибыль, тыс.руб. | 500 | 250 | 375 | 125 |
| Число сотрудников | 60 | 45 | 52,5 | 7,5 |

| Нормирован. показатель | Предприятие 1 | Предприятие 2 |
|---|---|---|
| Прибыль, тыс.руб. | 1 | -1 |
| Число сотрудников | 1 | -1 |

Рис. 5.3 Нормирование значение путем вычитания среднего значения и деления на среднее квадратическое отклонение

Для расчета интегрального показателя также используются различные методики. Так, например, в случае использования метода нормирования с помощью эталонного значения, определяется сумма квадратов всех элементов объекта. Если задача решается с учетом разного веса показателей, полученные квадраты умножаются на величину соответствующих весовых коэффициентов

$(K)$, установленных экспертным путем, после чего результаты складываются по строкам (рис. 5.4):

$$R_j = K_1 x_{j1}^2 + K_2 x_{j2}^2 + \ldots + K_n x_{jn}^2 .$$

$R$ – интегральный показатель $j$-го объекта

$K$ – коэффициент важности.

| Нормирован. показатель | Предприятие 1 | Предприятие 2 | Коэффициент важности |
|---|---|---|---|
| Прибыль, тыс.руб. | 1 | 0,5 | 0,6 |
| Число бракованных изделий | 0,75 | 1 | 0,4 |
| Интегральный показатель | 0,825 | 0,55 | |
| Место | 1 | 2 | |

Рис. 5.4 Определение интегрального показателя

На рис.5.4 значения первого и второго интегрального показателя вычислены по формулам:

$$R_1 = 0,6 \cdot 1^2 + 0,4 \cdot 0,75^2 = 0,825$$

$$R_2 = 0,6 \cdot 0,5^2 + 0,4 \cdot 1^2 = 0,55$$

При нормировании показателей путем вычитания среднего значения и деления на среднее квадратическое отклонение интегральный показатель может быть вычислен по формуле:

$$R_j = \sqrt{(1-x_{j1})^2 + \ldots + (1-x_{jn})^2} .$$

При определении коэффициентов важности используют различные процедуры оценки мнений экспертов. Так, например, в процедуре Саймона эксперту предоставляются карточки с названиями показателей, которые нужно разместить снизу-вверх: от наименее важного критерия к наиболее важному. Затем он получает белые карточки, которые нужно разместить между карточками с показателями, чем больше разница в их важности, тем больше должно быть белых карточек. Веса рассчитываются путем деления ранга характеристики на сумму рангов. На рис.5.5 эксперту предоставлено 4 показателя для оценки. Он размести их снизу-вверх, таким, образом, наиболее важным показателем является прибыль, а наименее важным – число сотрудников. Кроме того, он отметил двойную разницу между прибылью и объемом продаж и одинарную – между уровнем качества и числом сотрудников. Сумма рангов получилась равна 7.


Рис.5.5 Первый этап процедуры Саймона

Выполняя деление ранга каждого показателя на сумму рангов получим таблицу (рис.5.6.)

| | | |
|---|---|---|
| 4 | Прибыль | 1 |
| 3 | Объем продаж | 0,57 |
| 2 | Уровень качества | 0,43 |
| 1 | Число сотрудников | 0,14 |

Рис.5.6 Расчет значений показателя

Для числа сотрудников это отношение равно $\frac{1}{7}=0{,}14$, для объема продаж - $\frac{4}{7}=0{,}57$, для уровня качеств - $\frac{3}{7}=0{,}43$. Сумма полученных значений равна 2,14. Наконец, выполняем нормирование показателей, путем деления каждого значения на общую сумму (рис.5.7).

| | | |
|---|---|---|
| Прибыль | 0,47 | =1/2,14 |
| Объем продаж | 0,27 | =0,57/2,14 |
| Уровень качества | 0,2 | =0,43/2,14 |
| Число сотрудников | 0,06 | =0,14/2,14 |

Рис.5.7 Результаты нормирования

В случае, когда рассматриваются не различные показатели объекта, а значение одного показателя в разные периоды, используется динамический рейтинг. Расчет рейтинговых оценок проводится путем суммирования по всем отчетным датам с линейным пропорциональным забыванием более старых значений и с нормированием весов до единицы:

$$D(P^k)=\frac{2}{13}\sum_{t=1}^{12}\frac{t}{12}\frac{a_t^k}{\sum_{j=1}^{N}a_t^j},$$

где $D(P^k)$ - долевой динамический рейтинг по данному показателю;

$a^k$ - один из выбранных показателей для *k*-го участника, k=1..*N*;

$a_t^k$ - один из выбранных показателей для *k*-го участника в момент времени *t*;

*N*- общее число участников;

*t*- время ($t=1,...,T;T=12$).

На рис.5.8 представлен пример расчета динамического рейтинга. Нужно сравнить два предприятия по показателю прибыли за последний год.

| Месяц, *t* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Прибыль, тыс.руб., предприятие 1 | 100 | 120 | 140 | 90 | 170 | 100 | 105 | 110 | 115 | 101 | 118 | 120 | |
| Прибыль, тыс.руб., предприятие 2 | 50 | 70 | 30 | 40 | 45 | 50 | 60 | 65 | 68 | 68 | 70 | 71 | |
| Сумма | 150 | 190 | 170 | 130 | 215 | 150 | 165 | 175 | 183 | 169 | 188 | 191 | |
| $t/12\times P_t^1/(P_t^1+P_t^2)$ | 0,06 | 0,11 | 0,21 | 0,23 | 0,33 | 0,33 | 0,37 | 0,42 | 0,47 | 0,50 | 0,58 | 0,63 | **0,65** |
| $t/12\times P_t^2/(P_t^1+P_t^2)$ | 0,03 | 0,06 | 0,04 | 0,10 | 0,09 | 0,17 | 0,21 | 0,25 | 0,28 | 0,34 | 0,34 | 0,37 | **0,35** |

Рис.5.8 Результаты расчета динамического рейтинга

Так, рейтинг первого предприятия получается путем умножения суммы $\sum 0,06+0,11+0,21+0,23+...+0,63=4,22$ на 2/13 и получаем $D(P^1)=\frac{2}{13}\cdot 4,22=0,65$

.

**Глобальные индексы** создаются исследователями и исследовательскими центрами для сопоставления стран мира по определенным характеристикам: степени развитости экономики, уровню конкурентоспособности, благосостоянию и другим.

*Глобальный индекс миролюбия (Global Peace Index (GPI))*

Создан The Institute for Economics and Peace (Sydney), при его построении используются следующие группы признаков:

- Внутренние и внешние конфликты (5 индикаторов);
- Безопасность внутри страны (10 индикаторов);
- Милитаризация (8 индикаторов).

Результаты расчетов индекса для различных стран представлены на сайте http://economicsandpeace.org/.

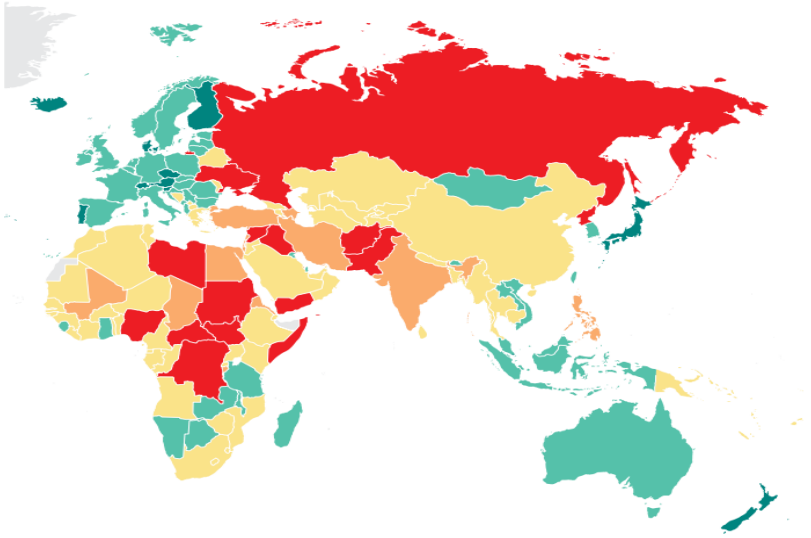


Рис. 5.9 Индекс миролюбия для разных стран (зеленым цветом обозначены наиболее миролюбивые страны, красным – наименее миролюбивые)

*Индекс экономической свободы (Index of economic freedom)*

Создан Heritage Foundation, The Wall Street Journal, при его построении используются следующие признаки:

- свобода бизнеса;
- свобода торговли;
- участие правительства;
- свобода от коррупции;
- свобода труда;
- и др. (всего 10 индикаторов).

Результаты расчетов индекса для различных стран представлены на сайте http://www.heritage.org/.

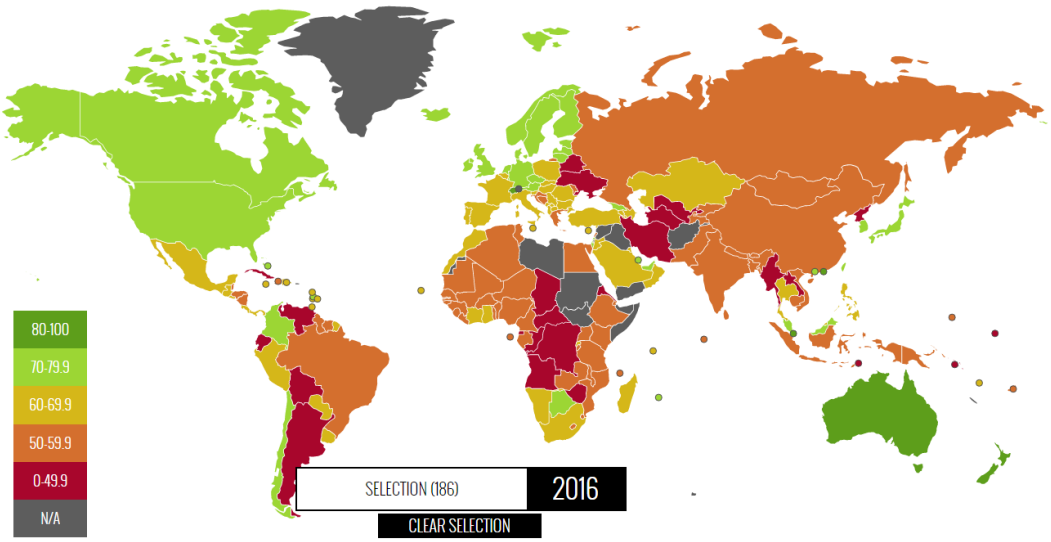



Рис. 5.10 Индекс экономической свободы для разных стран (зеленым цветом обозначены наиболее экономически свободные страны, красным – наименее экономически свободные)

## Список литературы


1. Шмойлова Р. А., Минашкиг В. Г., Садовникова Н. А., Шувалов Е. Б. Теория статистики. – М.: Финансы и статистика, 2004. – 656 с.
2. Лузина Л. И. Статистика: Учебное пособие. – Томск: ТМЦДО, 2009. – 141 с.
3. Светуньков С. Г., Светуньков И. С. Методы социально-экономического прогнозирования: Учебник для вузов. – СПб.: Изд-во СПбГУЭФ, 2009. – 147 с.
4. Иванов О. В. Статистика. Учебный курс для социологов и менеджеров. Часть 1. Описательная статистика. Теоретико-вероятностные основания статистического вывода. – М. 2005. – 187 с.
5. Jackson M. O. Social and economic networks: models and analysis. [Электронный ресурс]. – Режим доступа - https://www.coursera.org/learn/social-economic-networks/home/welcome.
6. Andrew Ng. Machine learning. [Электронный ресурс]. – Режим доступа - https://www.coursera.org/learn/machine-learning/home/welcome
7. Online social networks and media. [Электронный ресурс]. – Режим доступа - http://www.cs.uoi.gr/~tsap/teaching/2013-cs-l14/slides/L14-lecture5.pdf.
8. Грибанова Е.Б. Информационная система рейтинговой оценки объектов экономики / Е.Б. Грибанова, А.Н. Алимханова, П.Э. Тугар-оол // Доклады Том. гос. ун-та систем управления и радиоэлектроники. – 2016. – № 2(19). – С. 51–55.
9. Карминский А.М., Полозов А.А. Энциклопедия рейтингов: экономика, общество, спорт. – М.: ИД «Форум», Инфра-М, 2016. – 36 с.