Анализ данных

Лекция 1

Основные понятия и категории анализа данных

Костромина Елена Валерьевна,
кафедра Информационных систем в экономике

Литература

1. Статистика : учебник для прикладного бакалавриата : / [М. В. Боченина и др.] ; под ред. И. И. Елисеевой ; С.-Петерб. гос. экон. ун-т. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2015. - 447 с.
2. Статистика: [учебник для студентов бакалавриата по направлению подготовки "Экономика"] / [Л. И. Ниворожкина и др.] ; под общ. ред. Л. И. Ниворожкиной. - 2-е изд., доп. и перераб. - Москва : Дашков и К : Наука-Спектр, 2013. - 414,
3. Статистика: учебник для бакалавров : [по направлению "Статистика" и другим экономическим специальностям] / [В. С. Мхитарян и др.] ; под ред. В. С. Мхитаряна. - Москва : Юрайт, 2015. - 590 с. : ил., табл. - (Учебник) (Бакалавр. Базовый курс). - Библиогр.: с. 589-590
4. Халафян, Алексан Альбертович. STATISTIKA 6: статистический анализ данных : [учебное пособие для студентов вузов по экономическим специальностям] / А. А. Халафян. - 2-е изд., перераб. и доп. - Москва : Бином, 2013. - 522 с.

Анализ данных

1. Совокупность действий, осуществляемых исследователем в процессе изучения полученных тем или иным образом данных в целях формирования определенных представлений о характере явления, описываемого этими данными.

Анализ данных

2. Процесс изучения стат. данных (поиска стат. закономерностей, закономерностей в среднем) с помощью математических методов, не предполагающих вероятностной модели изучаемого явления. Противостоит вероятностно-стат. подходу к обработке данных, опирающемуся на их вероятностную интерпретацию (как случайной выборки из генеральной совокупности) и использование вероятностных моделей для построения и выбора наилучших методов обработки

Анализ данных

3. Термин, отождествляемый с понятием «прикладная статистика», которая понимается как науч. дисциплина, разрабатывающая и систематизирующая понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации и обработки стат. данных в целях их удобного представления, интерпретации и получения научных и практических выводов.

Анализ данных

4. Процедуры поиска стат. закономерностей («свертки» информации), не сводящиеся к применению формальных алгоритмов. В основе лежит комплексное использование математико-статистических методов и методов А.д. с опорой на несколько методологических принципов.

Методологические принципы анализа данных:

Первый принцип

Вариация предпосылок, лежащих в основе выбираемых методов (любой метод опирается на определенную модель изучаемого явления, т.е. определенную систему предпосылок и постулатов): изменение таких предпосылок, рассмотрение последствий этого изменения, сравнение использования разных предпосылок и т.д.

Методологические принципы анализа данных:

Второй принцип

Системный подход. В процессе анализа данных изыскиваются различные приемы для наиб, полного использования и эндогенной информации (т.е. данных, описывающих изучаемый объект), и экзогенной (т.е. данных, описывающих «среду обитания» объекта).

Методологические принципы анализа данных:

Третий принцип - отказ от той точки зрения, что любое исследование имеет начало и конец. Готовность к постоянному возврату к одним и тем же данным. В непрерывном процессе анализа данных. предусматриваются разрывы, позволяющие извлекать накопленную информацию и принимать решения, связанные с управлением обработкой данных, с выбором дальнейших шагов анализа. Формальные операции перемежаются с неформальными процедурами принятия решения.

Основные задачи:

1. Классификация объектов:

   - Поиск однотипных групп объектов;

   - Создание типологии.

2. Сжатие информации:

   - Одномерный анализ – описательная статистика;

   - Многомерный анализ – связь между признаками;

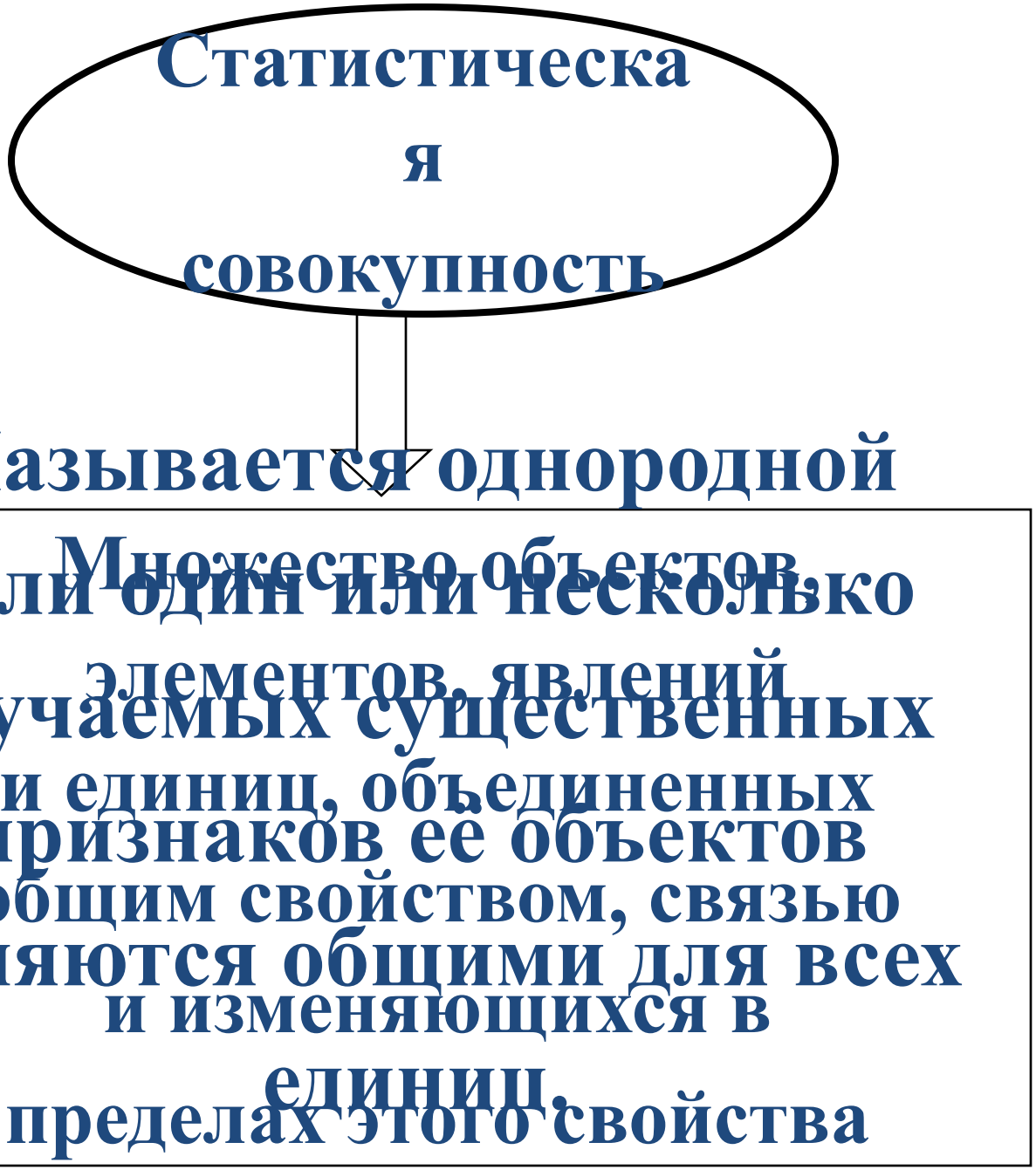
   - Поиск латентных переменных.

Этапы исследования

I. Статистическое наблюдение

II. Сводка и обработка информации, расчёт обобщающих показателей

III. Анализ, обобщение и интерпретация полученных результатов


Статистическая совокупность
Называется однородной
если один или несколько изучаемых существенных признаков ее объектов являются общими для всех
Множество объектов элементов, явлений и единиц, объединенных общим свойством, связью и изменяющихся в единиц пределах этого свойства

Статистическая совокупность

Статистический признак

Единица совокупности

Неделимый первичный элемент, носитель свойств изучаемого явления или процесса

Статистическая совокупность

Статистический показатель

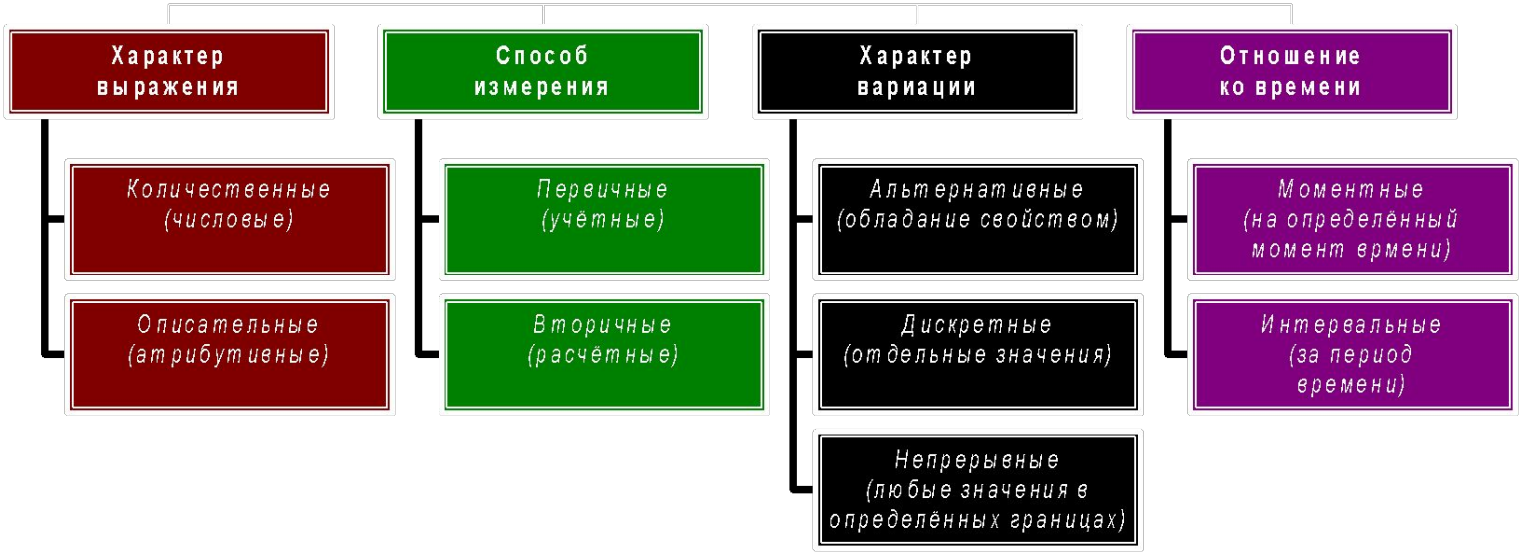
Группа единиц совокупности

Несколько элементов, единиц совокупности, объединенных общей связью, свойством

Статистический признак или показатель

Разли**ВАРИАЦИЯ**х одного и того же признака у разных единиц совокупности

Классификация признаков в статистике



Статистическое наблюдение

Достоверность и полнота информации

•Планомерность

подготовки

Срок наблюдения – время от начала до окончания сбора информации, сведений

и разработки … в плане

статистического наблюдения

Выбор времени и места

Время, в течение которого производится заполнение документов, статистического наблюдения, анкет, опросных листов


Статистик
Цели и задачи исследования
Для чего?
Инструкция
Инструментарий
формуляр,анкета и т.д.
образцы заполнения
Как?
Объект наблюдения
Выбор объекта
Кто?
Сбор данных
Первичный контроль
Что?


Арифметический
Логический
КОНТРОЛЬ

ПРИМЕР
арифметического контроля

Группа работников	Численность на начало года	Принято	Уволено	Численность на конец года
А	1	2	3	4
АУП	10	-	1	9
ПП	105	12	7	109
ВП	25	2	4	21
Итого:	140	14	12	142

ПРИМЕР
логического контроля

- Фамилия Ильин
- Имя Сергей
- Отчество Алексеевич
- Пол жен
- Возраст 10 лет
- Семейное положение вдовец
- Образование высшее
- Источник средств существования пенсия

Формы представления статистических данных

- Включения в текст;
- Занесение в таблицы;
- Графическое изображение.

Включения в текст

Во Владивостоке ветхим и аварийным жильем признан 571 дом общей площадью более 133 тыс. кв. м

Занесение в таблицы

Товары и услуги	цены		объём	
	2004	2005	2004	2005
Товары длительного пользования	62	60	540	640
Продукты	70	70	365	390
Транспортные расходы	110	100	215	240
Жильё	130	150	200	190
Медицинское обслуживание	330	390	160	165
Развлечения	430	430	141	142,5

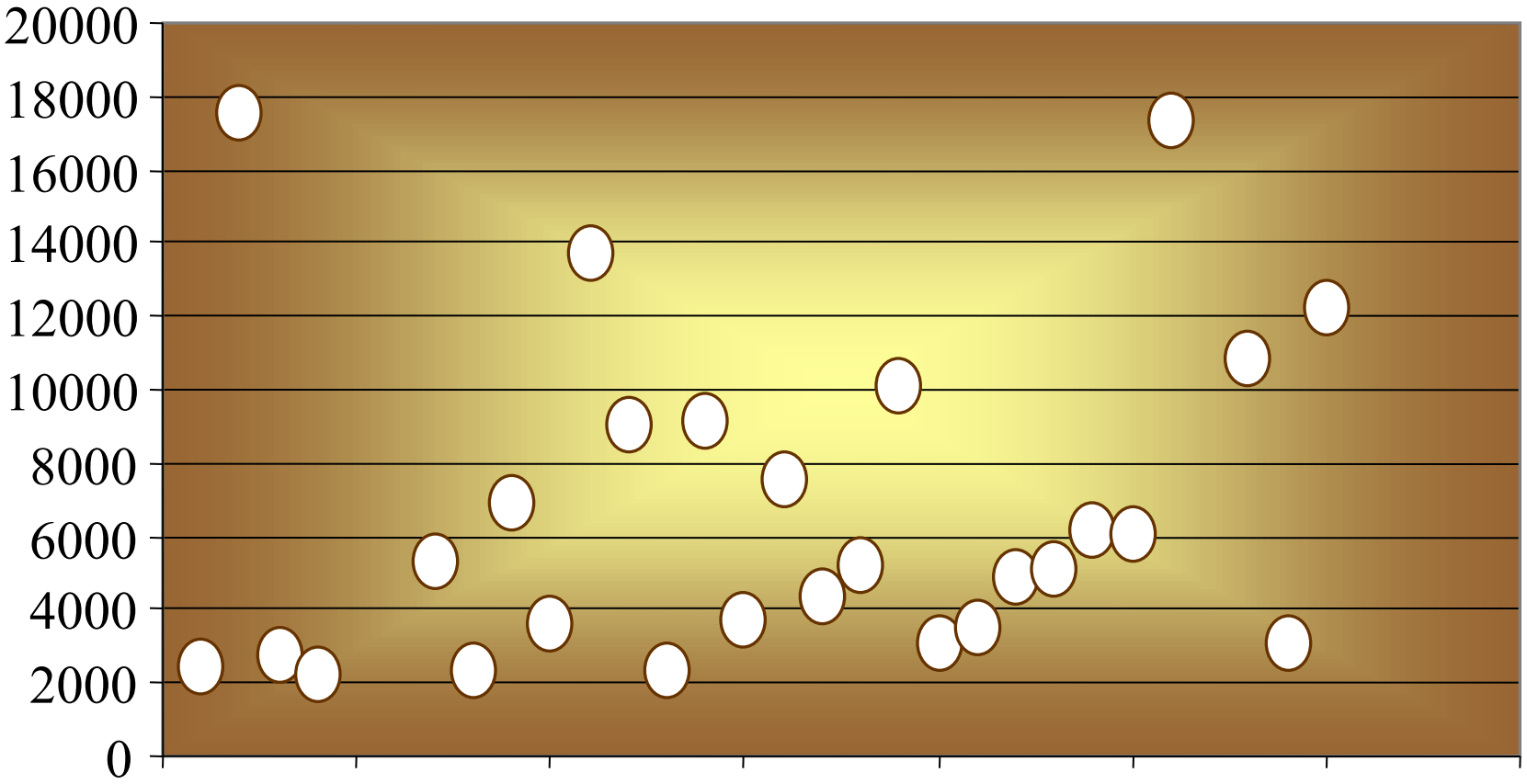
подлежащее

сказуемое

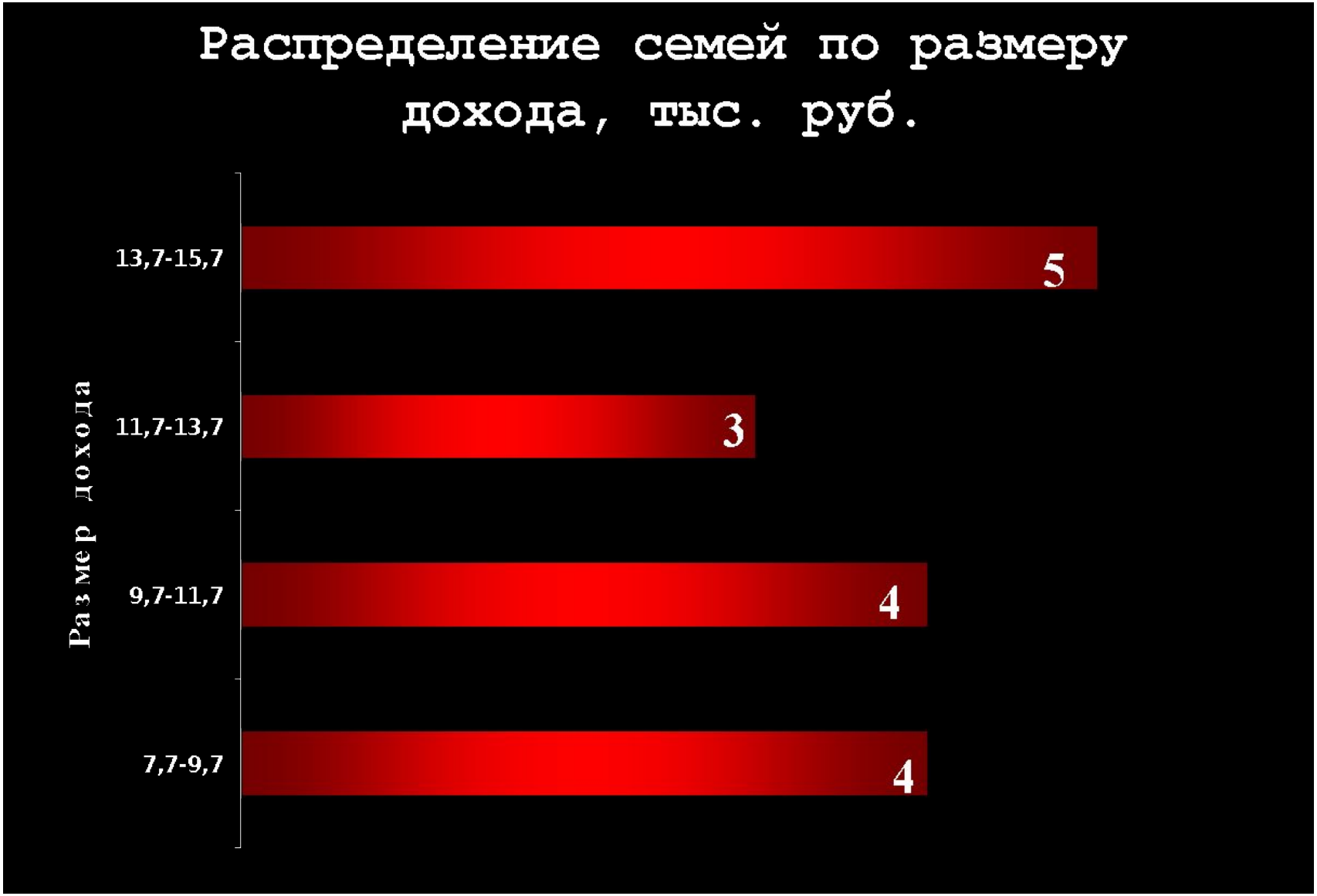
Виды графических изображений

Точечная диаграмма

Величина уставного капитала коммерческих банков региона, тыс. руб.



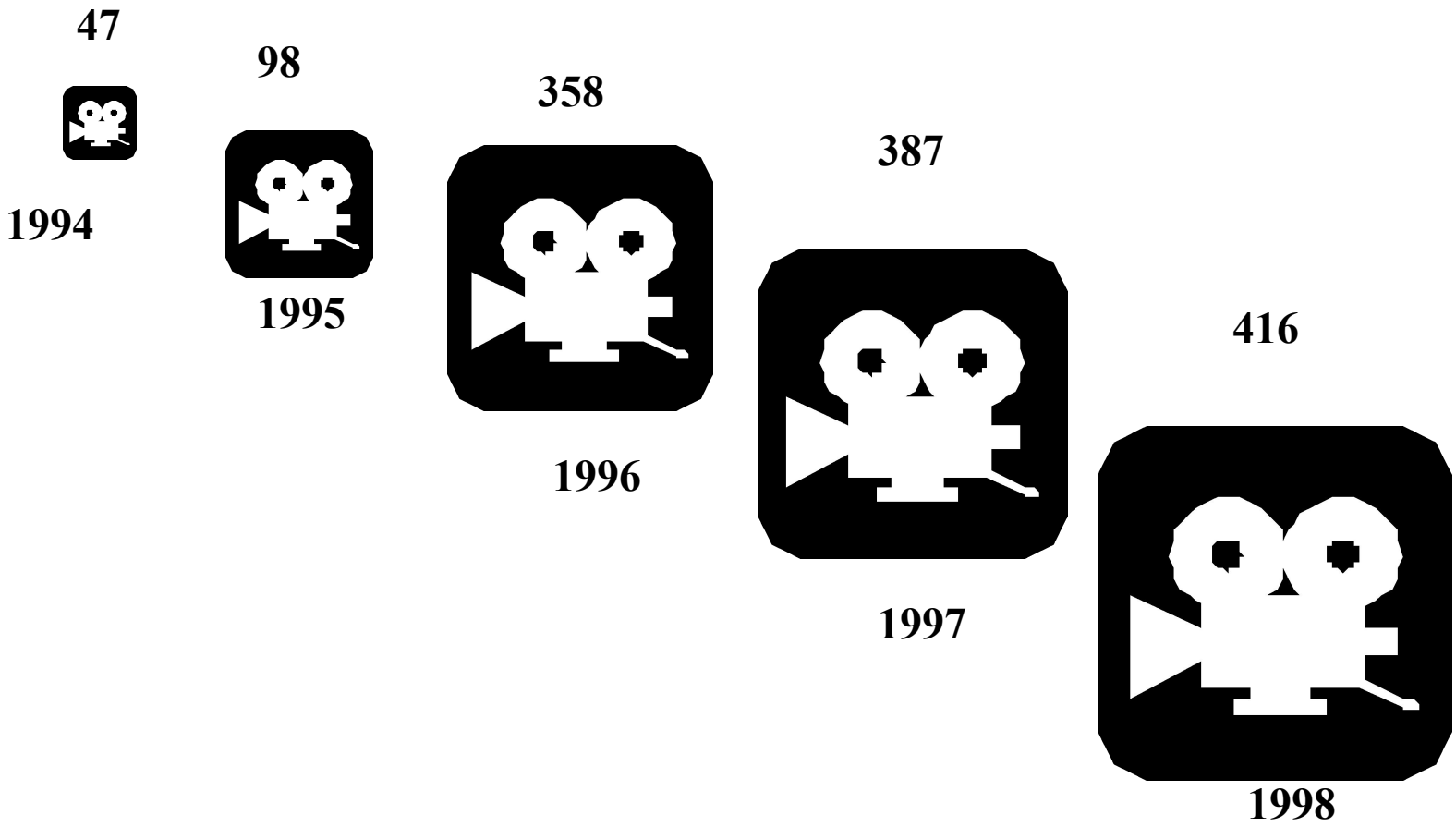
Линейчатая диаграмма



Плоскостная диаграмма (столбиковая)

Пример фигурной диаграммы

Выпуск документальных фильмов в России (шт.):



Сводка и группировка

Сводка - стадия, на которой осуществляется Систематизация первичных материалов статистического наблюдения

Группировка - объединение единиц совокупности в некоторые группы, имеющие свои характерные особенности, общие черты и сходные размеры изучаемого признака.

Виды группировок

- Типологическая
- Структурная
- Аналитическая

СТРУКТУРНАЯ ГРУППИРОВКА

Группы заводов по выручке от реализации		Число заводов (f_i)	Уд. веса заводов по группе
2,6	3,6	6	30,00%
3,6	4,6	9	45,00%
4,6	5,6	1	5,00%
5,6	6,6	1	5,00%
6,6	7,6	3	15,00%
Итого:		20	100,00%

АНАЛИТИЧЕСКАЯ ГРУППИРОВКА

Группы заводов по выручке от реализации, млн. руб.		**Прибыль предприятия в среднем по группе, тыс. руб.**
2,6	3,6	1335,3
3,6	4,6	1452,0
4,6	5,6	1402,0
5,6	6,6	1512,0
6,6	7,6	1448,6

ТИПОЛОГИЧЕСКАЯ ГРУППИРОВКА

Группы предприятий по формам хозяйствования	Объём промышленной продукции, млн. руб.
Государственные с традиционными формами управления	405,5
Арендные	19
Кооперативные	30

АНАЛИТИЧЕСКАЯ ГРУППИРОВКА НА ОСНОВЕ ТИПОЛОГИЧЕСКОЙ

Группы предприятий по формам хозяйствования	Средняя заработная плата на предприятии руб.
Государственные с традиционными формами управления	2405,5
Арендные	3319,8
Кооперативные	5630,6

ДАННЫЕ НЕ СГРУППИРОВАНЫ

№ предприятия	Выручка от реализации, млн.руб.	Прибыль предприятия, тыс.руб.
1	2,0	1270
2	2,0	1320
3	2,7	1250
4	2,8	1330
5	3,0	1410

Последовательность выполнения группировки по количественному признаку

1. Выбор группировочного признака
2. Расчёт числа групп
3. Расчёт шага или длины интервала
4. Построение интервалов
5. Подсчет численности групп
6. Расчёт удельных весов для структурных группировок или средних значений признака в группе для аналитических
7. Построение таблиц

Формула Стерджесса

$$k = 1 + (3{,}322 \times \lg N),$$

где N — количество наблюдений.

Высота интервала:

$$h = (X_{max} - X_{min})/k$$

Построение интервалов

Задача

Имеются данные по количеству работников, имеющих определенный стаж работы в организации.

Осуществить группировку по стажу, построив дискретный и интервальный ряды

Стаж работы, лет	Число работников, чел.
2	1
3	2
4	2
5	3
6	3
7	5
8	7
9	3
10	2
11	1
12	1
	30

$$k = [1 + 3{,}322 \cdot \lg 30] = 5$$

$$h = \frac{12 - 2}{5} = 2$$

Группы работников по стажу, лет		Число работников, чел.
2	4	5
4	6	6
6	8	12
8	10	5
10	12	2
		30

Спасибо за внимание!