Статистика

Елена Игоревна Васенкова

Статистика

Существует более 200 определений

Статистика – искусство и наука сбора и анализа данных

Статистика - наука, разрабатывающая и систематизирующая понятия, приемы, методы и модели, предназначенные для сбора, стандартной записи, систематизации и обработки данных с целью их удобного представления, анализа и получения научных и практических выводов

Учебный план

- Описательная статистика
- Интервальное оценивание данных и проверка статистических гипотез
- Статистические методы исследования взаимосвязей
- Статистические методы исследования динамики и прогнозирования

Литература

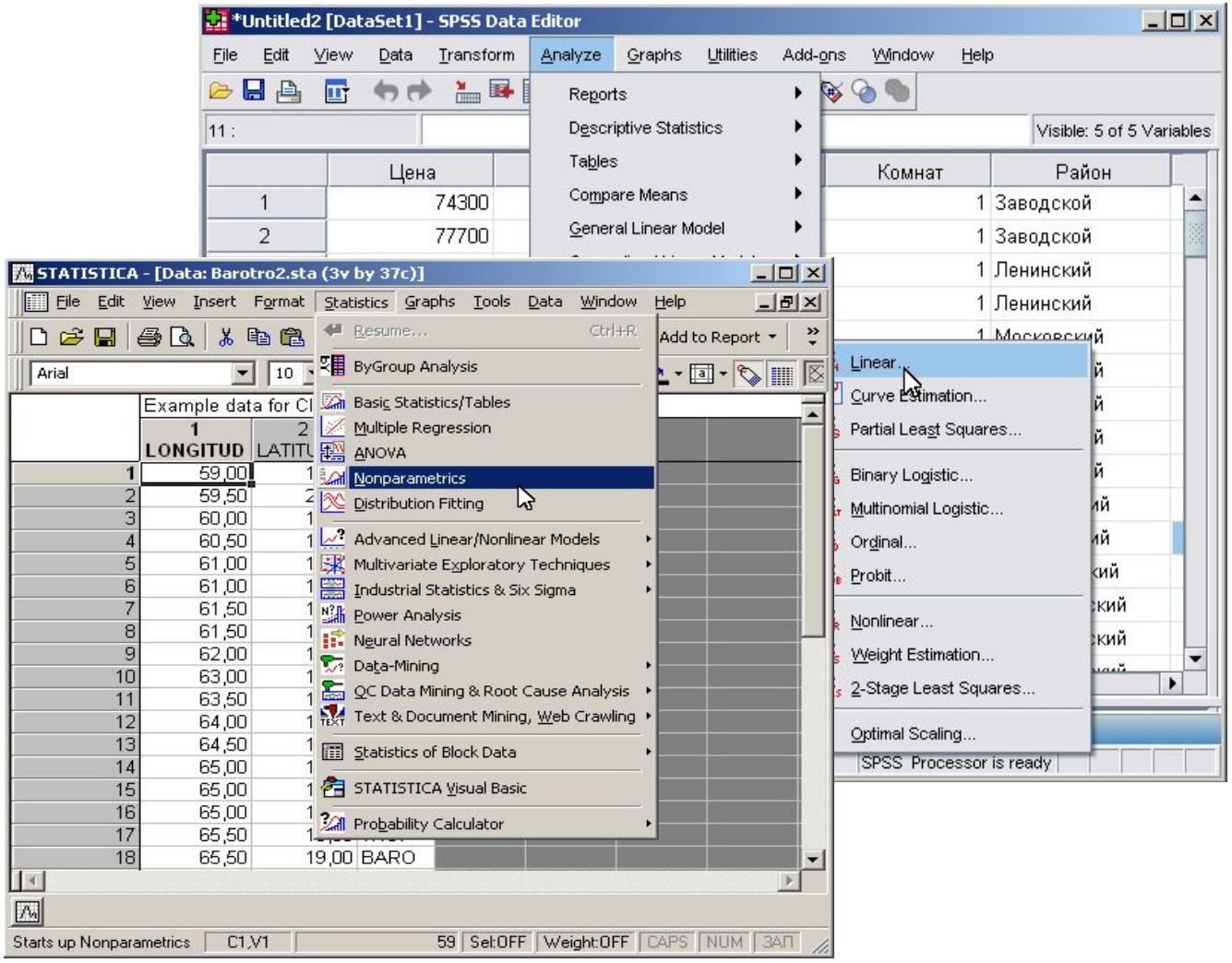
1. Сигел Э. Практическая бизнес-статистика, 2002
2. Статистика для менеджеров с использованием Microsoft Excel, 2005
3. Paul Newbold Statistics for business and economics, 2005
4. Васенкова Е.И. Статистика: конспект лекций для студентов программы переподготовки «финансы» http://www.elib.bsu.by

Использование Excel

- Распространенность
- Универсальность
- Большой набор статистических функций
- Наличие специализированных пакетов-расширений
- Недостатки: иногда сложно реализовать нестандартные расчетные методики

Статистические пакеты

- Statistica
- SPSS
- SAS



Решаемые в курсе задачи

- Описание данных
- Сравнение
- Изучение зависимостей
- Прогнозирование

Описание данных

Методы описательной статистики позволяют эффективно обработать большие массивы данных и представить их в виде удобном и пригодном для анализа.

Происходит своеобразное «сжатие» информации, получение небольшого количества наиболее важных характеристик, дающих возможность достаточно полно производить предварительный анализ и оценку статистических данных.

Сравнение

Интервальное оценивание и проверка гипотез позволяют сделать вывод о наличии либо отсутствии разницы между двумя ситуациями, проанализировать точность получаемых результатов и надежность сделанных предсказаний.

Эти инструменты оказываются полезными при исследовании эффективности новых методов работы или в изменяющихся внешних условиях, отвечая на вопрос: являются ли наблюдаемые изменения случайностью или же можно определенно говорить о влиянии?

Изучение зависимостей

Разные факторы практической деятельности неизбежно оказываются связанными друг с другом.

Корреляционный анализ оценивает связь на фоне неизбежных «шумов» и случайных выбросов.

Регрессионный анализ дает математическое выражение для обнаруженных зависимостей.

После этого можно производить подробное рассмотрение ситуации по схеме «что-если»: что произойдет при увеличении количества клиентов, изменении курса валют и т.д.

Прогнозирование

Статистические методы позволяют выделить основные составляющие изменяющегося во времени набора данных: долгосрочную тенденцию, периодические сезонные колебания, случайную составляющую.

После этого можно не только составить прогноз, но и оценить его точность и возможность долгосрочного прогнозирования в текущих условиях.

Почему это работает?

Статистика опирается на *универсальные* инструменты, практически не зависящие от конкретной области применения.

Используются *строгие* математические методы, в результате не все «очевидное» оказывается правильным.

Основные понятия

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

Генеральная совокупность – все возможные (реальные или гипотетические) значения случайной величины.

Выборочная совокупность (выборка) – реально наблюдаемая часть значений случайной величины.

Главная задача

По свойствам, полученным на основе данных выборка, определить свойства генеральной совокупности.

Пример: социологический опрос. По данным опроса 2000 человек в РБ делаются прогнозы результатов выборов.

Выборка – 1600 человек, генеральная совокупность – все избиратели.

Связь с теорией вероятностей

Теория вероятностей:

известны свойства генеральной совокупности -

–можно предсказать свойства выборки

Статистика:

измерено свойство выборки - можно судить о свойстве генеральной совокупности

Пример: подбрасывание монеты

Генеральная совокупность – всевозможные результаты бросания.

Теория вероятностей:
вероятность выпадения орлов и вероятность выпадения решки равна 0.5.

Статистика:
произведено 200 испытаний, орел выпал 105 раз, можно ли сделать вывод о равновероятности выпадения орла и решки.

Стадии статистического исследования

- Планирование и сбор данных
- Предварительное исследование
- Оценивание неизвестной величины
- Проверка статистических гипотез

Планирование и сбор данных

- Составление подробного плана исследования
- Определение необходимого (или доступного) количества данных
- Сбор данных, возможно, с использованием случайной выборки из генеральной совокупности

Предварительное исследование

- Оценка соответствия имеющихся данных предварительным прогнозам, фильтрация выбросов (цензурирование)
- Визуализация данных
- Оценка распределения данных (положение, разброс, …)
- Грубая проверка предположения о связи данных

Оценка неизвестной величины

- Предсказание значения неизвестной величины (победитель на выборах, объем продаж в следующем квартале, уровень брака, …)
- Оценка точности полученного значения (доверительного интервала)

Проверка статистических гипотез

- Использование данных для осуществления выбора одной из двух (или более) различных возможностей.
  - Использование нового метода работы с клиентами увеличивает (не увеличивает) объем продаж
  - В Вашем учреждении зарплата зависит (не зависит) от уровня образования сотрудники

Классификация статистических данных

- ***по количеству переменных, описывающих элементарную единицу данных**:*

- одномерные
- многомерные

Классификация статистических данных

- ***по типу измерения*** :

- количественные:
  - дискретные
  - непрерывные
- качественные:
  - порядковые
  - номинальные

Классификация статистических данных

- ***по отношению ко времени*:**

  - временные ряды
  - данные об одном временном срезе

Классификация статистических данных

- ***по способу получения данных**:*

  - первичные
  - вторичные

Описательная статистика

Методы описательной статистики – методы описания выборок с помощью различных показателей и графиков

Показатели описательной статистики

Показатели положения: среднее значение, медиана, мода, минимальной и максимальное значения, квартили

Показатели разброса: дисперсия, стандартное отклонение, размах, межквартильный размах

Показатели симметрии: асимметрии, положение медианы относительно среднего

Показатели формы: эксцесс

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Виды средних значений:

среднее арифметическое

среднее гармоническое

среднее геометрическое

среднее степенное

Среднее арифметическое

среднее $$\overline{x} = \frac{1}{n} \times \sum_{i=1}^{n} x_i$$

среднее для сгруппированных данных, $$\overline{x} = \frac{\sum_{i=1}^{K} x_i \times n_i}{\sum_{i=1}^{K} n_i}$$

$$\sum_{i=1}^{K} n_i = n$$

Определить среднее количество мячей, забитых за один матч

Число забитых мячей	Число матчей
0	21
1	41
2	45
3	37
4	19
5	10
6	6
7	1

Определить средний возраст сотрудников

Возраст сотрудников, лет	Число сотрудников
до 20	48
20 - 30	21
30 – 40	75
40 – 50	62
свыше 50	54

Среднее гармоническое

среднее

$$\overline{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

среднее для сгруппированных данных,

$$\overline{x} = \frac{\sum_{i=1}^{K} n_i}{\sum_{i=1}^{K} \frac{n_i}{x_i}}$$

Определить среднюю урожайность культур

Культура	Валовой сбор в ц.	Урожайность в ц/га
Пшеница	32500	25
Рожь	1620	18
Ячмень	13640	22
Овес	1650	15

Определить среднюю урожайность культур

Культура	Посевная площадь, га	Урожайность в ц/га
Пшеница	1300	25
Рожь	90	18
Ячмень	610	22
Овес	110	15

Среднее геометрическое

среднее $\bar{x} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$

среднее для сгруппированных данных, $\bar{x} = \sqrt[n]{(x_1)^{n_1} \times (x_2)^{n_2} \times \ldots \times (x_k)^{n_k}}$

Среднее степенное порядка p

$$\overline{x}_p = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}}$$

$$\ldots \overline{x}_{-1} < \overline{x}_0 \leq \overline{x}_1 \leq \overline{x}_2 \leq \ldots$$

Определение моды в интервальном ряду

$$Mo = x_{мо} + h \times \frac{n_{мо} - n_{мо-1}}{(n_{мо} - n_{мо-1}) + (n_{мо} - n_{мо+1})}$$

Определение медианы в интервальном ряду

$$Me = x_{Me} + h \times \frac{\frac{1}{2} \times \sum_{i=1}^{K} n_i - S_{Me-1}}{n_{Me}}$$

Показатели вариации

- Размах

$$R = X_{MAX} - X_{MIN}$$

- Среднелинейное отклонение

$$d = \frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right|}{n}$$

$$d = \frac{\sum_{i=1}^{K} \left| x_i - \overline{x} \right| \times n_i}{\sum_{i=1}^{k} n_i}$$

Показатели вариации

- Дисперсия

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

- Дисперсия для сгруппированных данных

$$\sigma^2 = \frac{\sum_{i=1}^{K} (x_i - \bar{x})^2 \times n_i}{\sum_{i=1}^{K} n_i}$$

Показатели вариации

- Среднеквадратическое (стандартное) отклонение

$$\sigma = \sqrt{\sigma^2}$$

- Коэффициент вариации

$$V = \frac{\sigma}{\overline{x}} \times 100\%$$

Табличное и графическое представление данных

Для описания количественных данных используют:

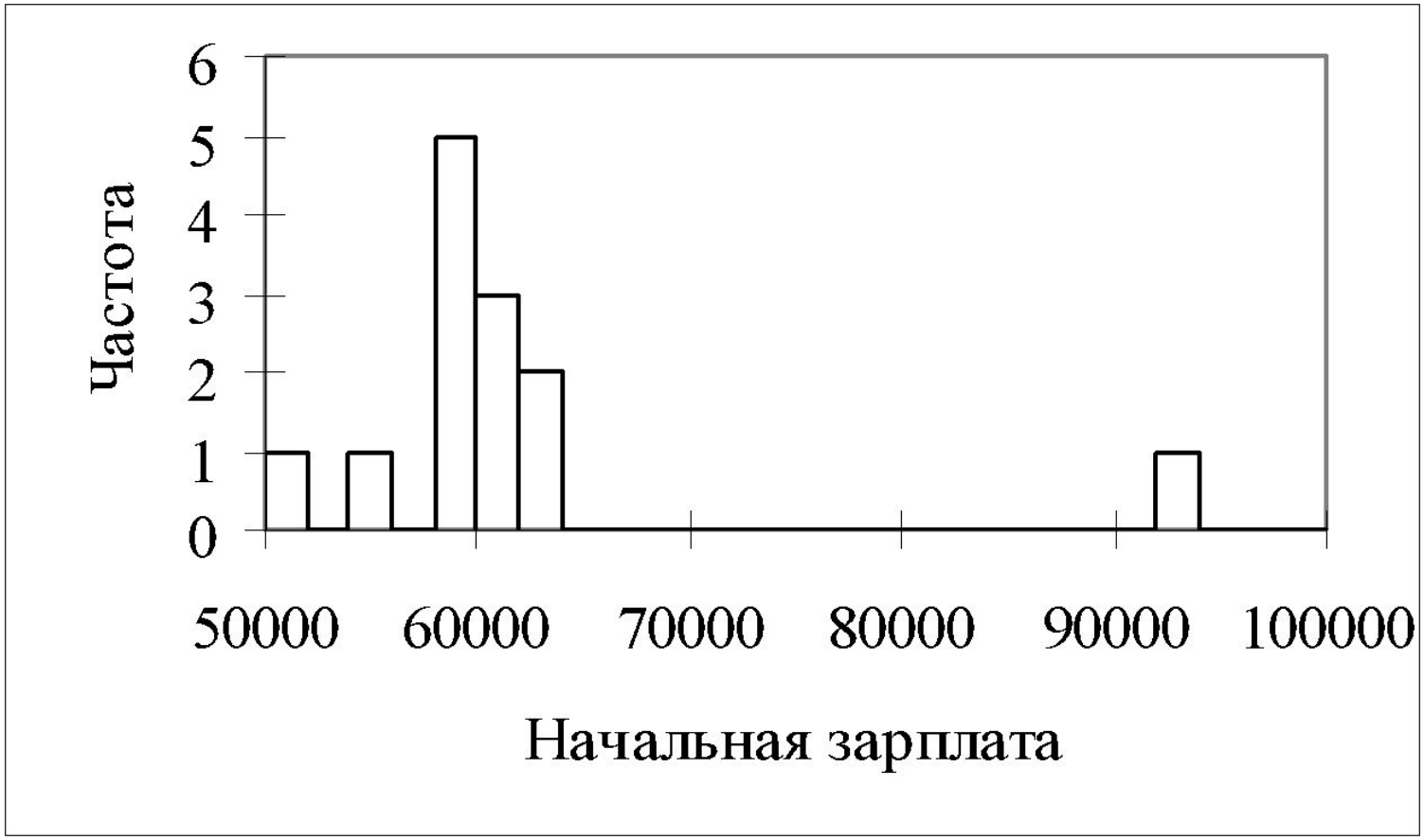
- распределение частот, распределение относительных частот, процентное распределение,
- распределение накопленных (кумулятивных) частот, распределение относительных накопленных (кумулятивных) частот,
- кростабуляцию,
- точечные и линейные диаграммы, гистограммы, интегральные (кумулятивные) кривые, диаграммы разброса, диаграмма «ствол и листья».

Табличное и графическое представление данных

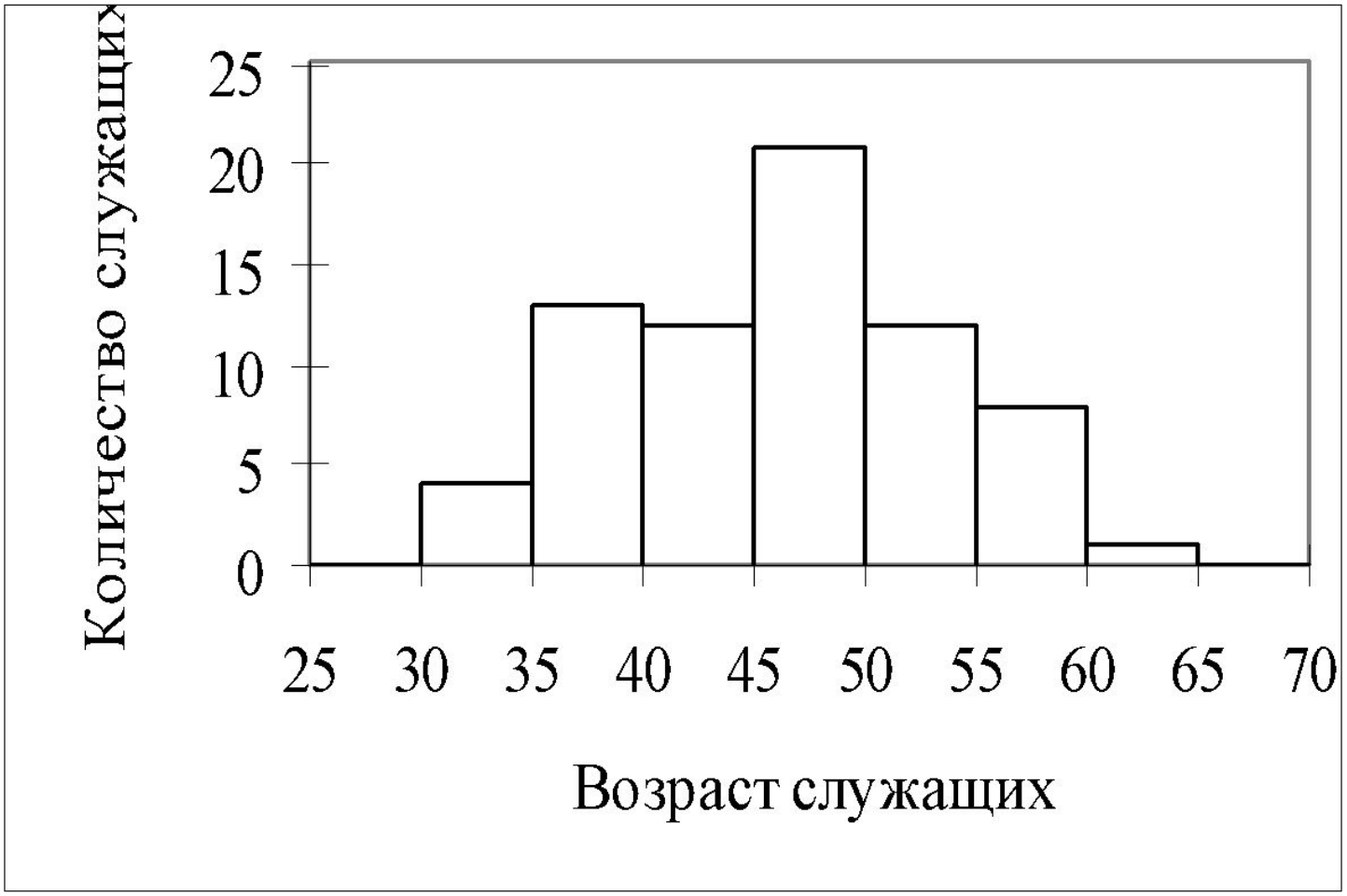
Для описания качественных данных используют:

- распределение частот, распределение относительных частот
- таблицы сопряженности
- линейчатые и секторные диаграммы.

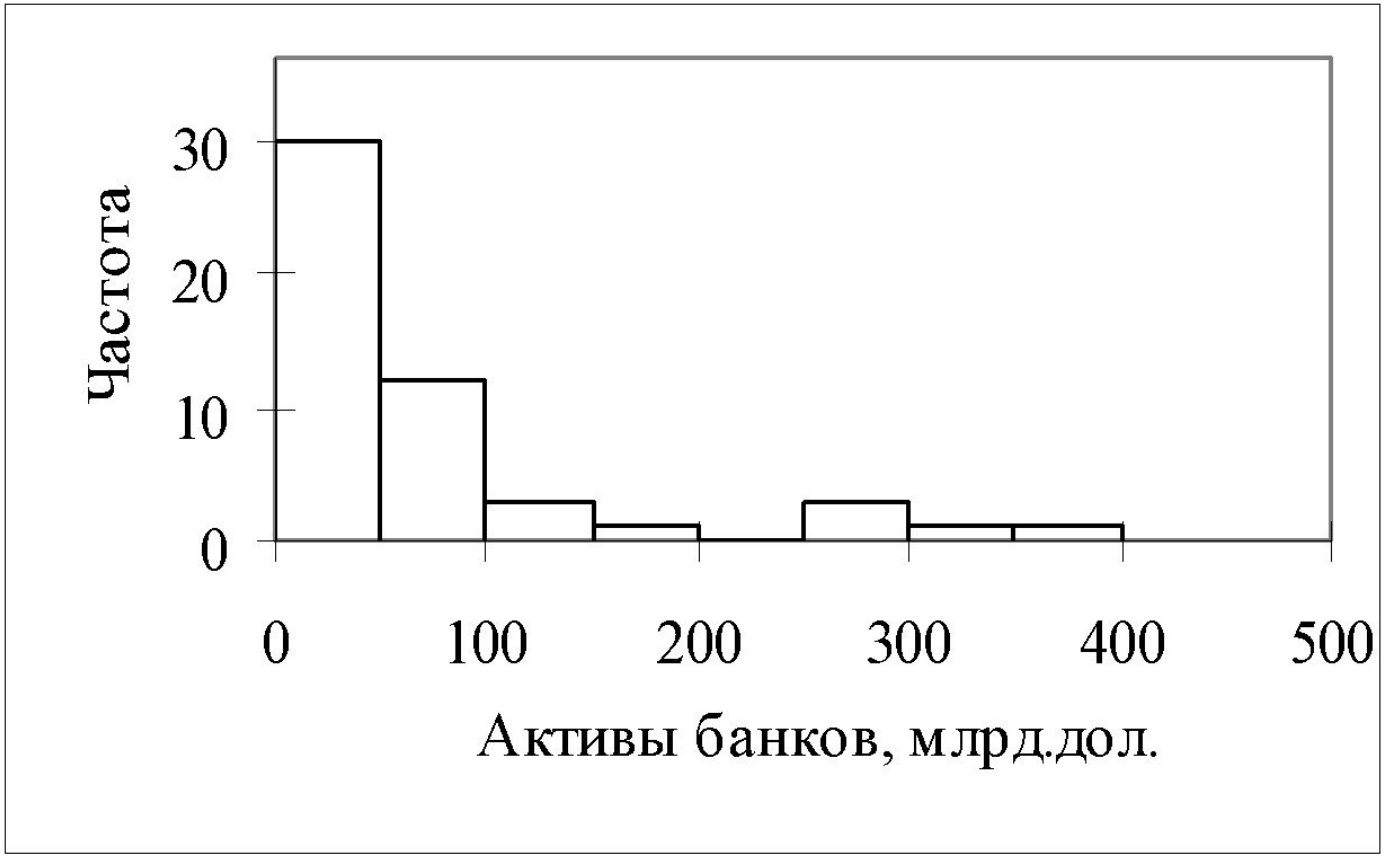
Гистограмма стартовой зарплаты выпускников с дипломом MBA



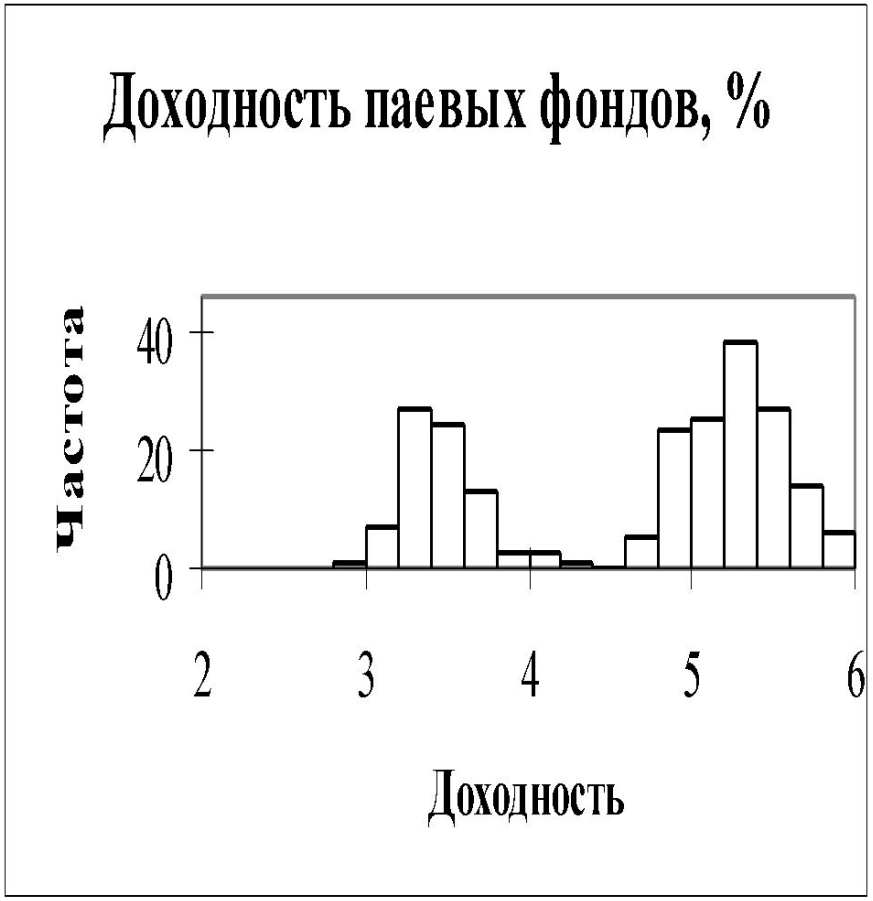
Гистограмма возраста служащих компании



Активы некоторых коммерческих банков

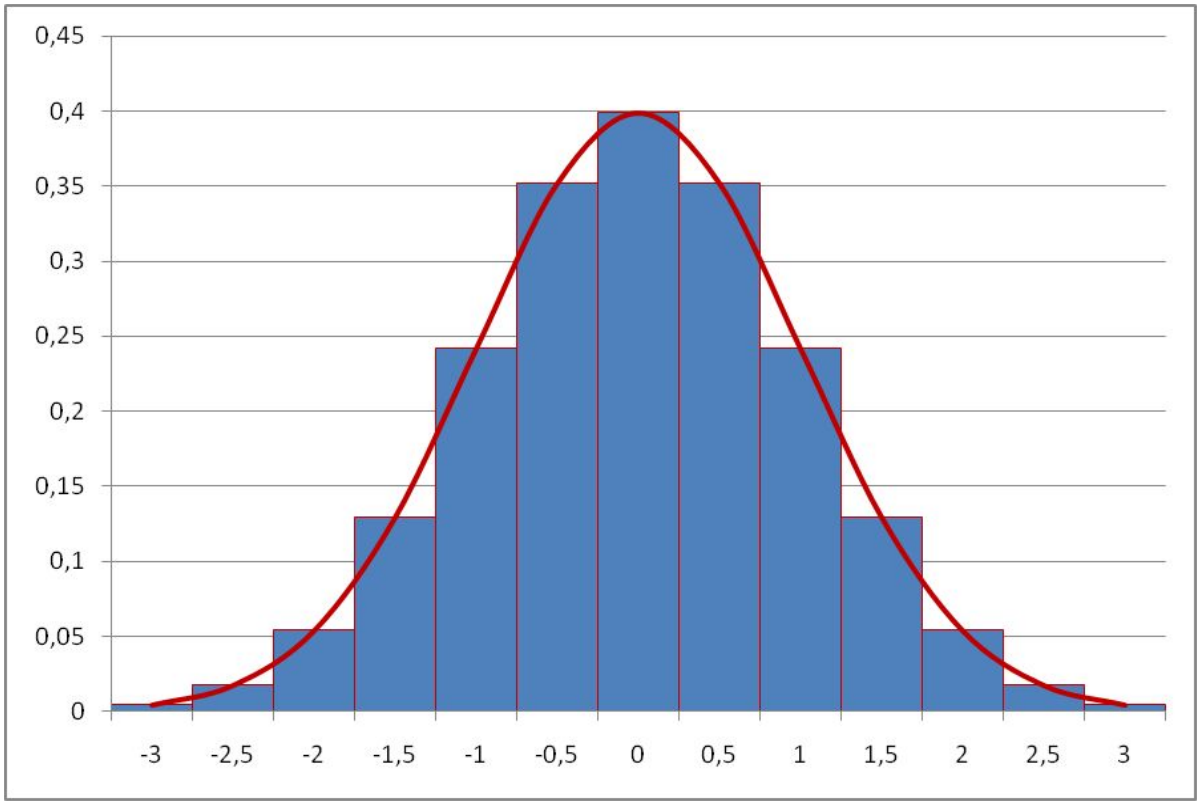


Гистограммы бимодальных распределений



Графическое представление данных

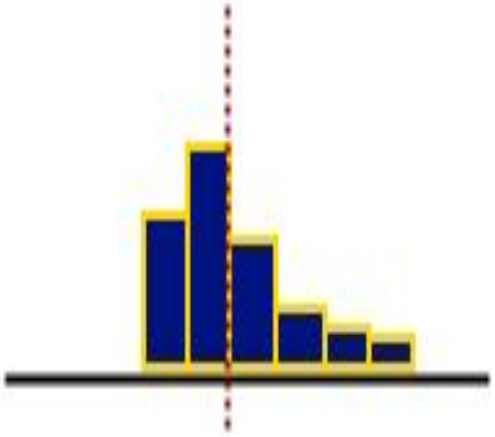
Гистограмма: данные разбиваются на интервалы последующим отображением на графике



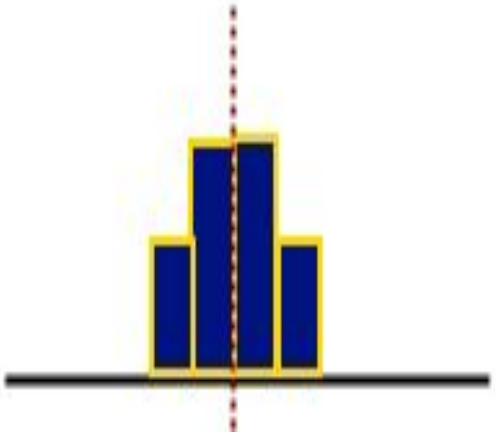
Асимметрия

Показывает, насколько симметрично расположены данные относительно среднего

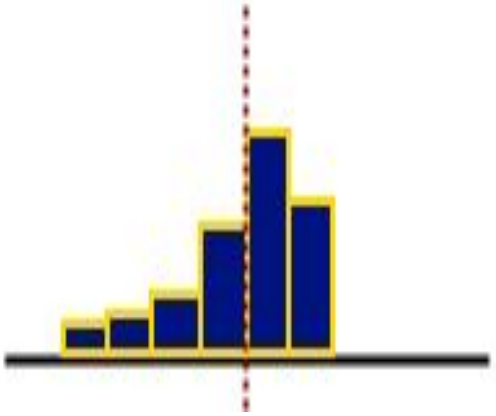
Асимметрия > 0



Асимметрия = 0



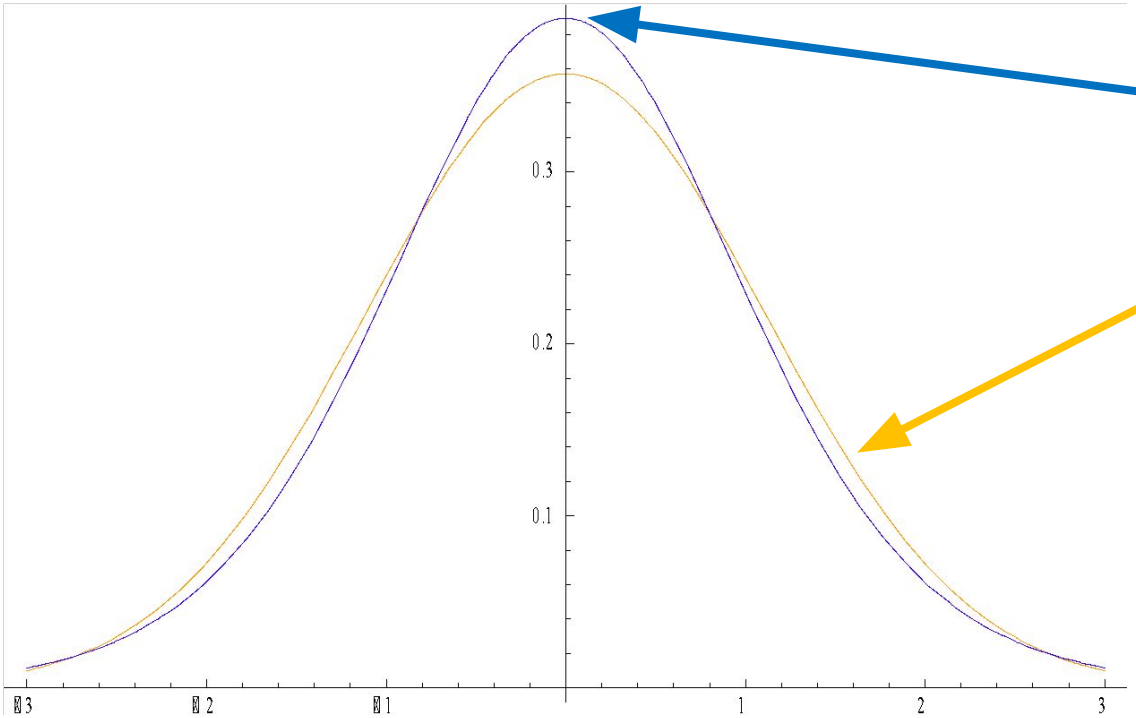
Асимметрия < 0



Эксцесс

Показатель «остроты» распределения.

Меньше эксцесс – «острее» распределение



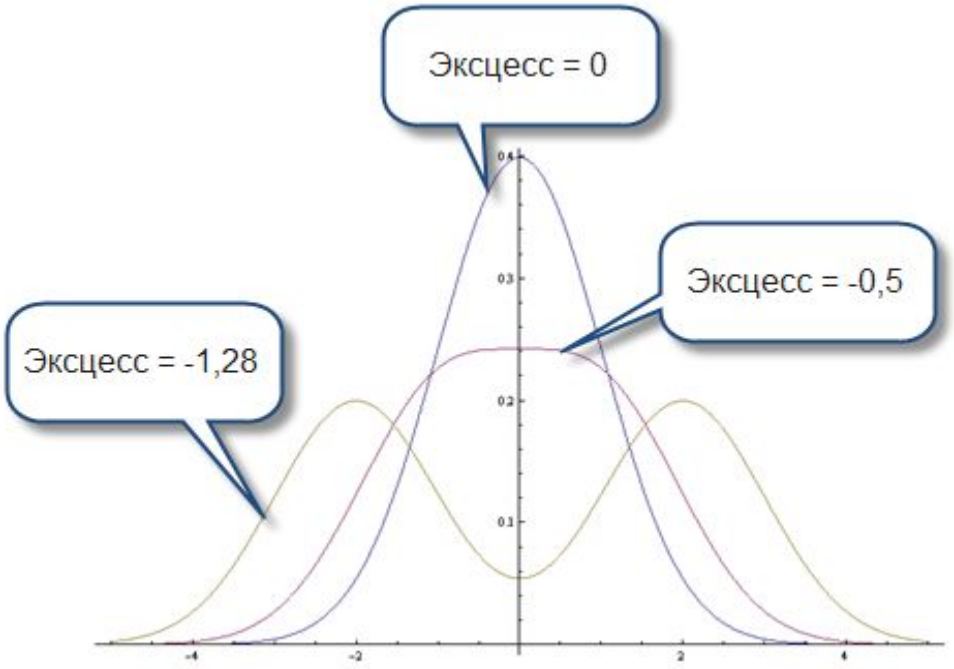
Эксцесс = 0

Эксцесс = 1

Эксцесс

Эталонным является нормальное распределение

Отрицательные значения эксцесса наблюдаются у бимодальных распределений



Нормальное распределение

Стандартизованное:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Общий вид:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

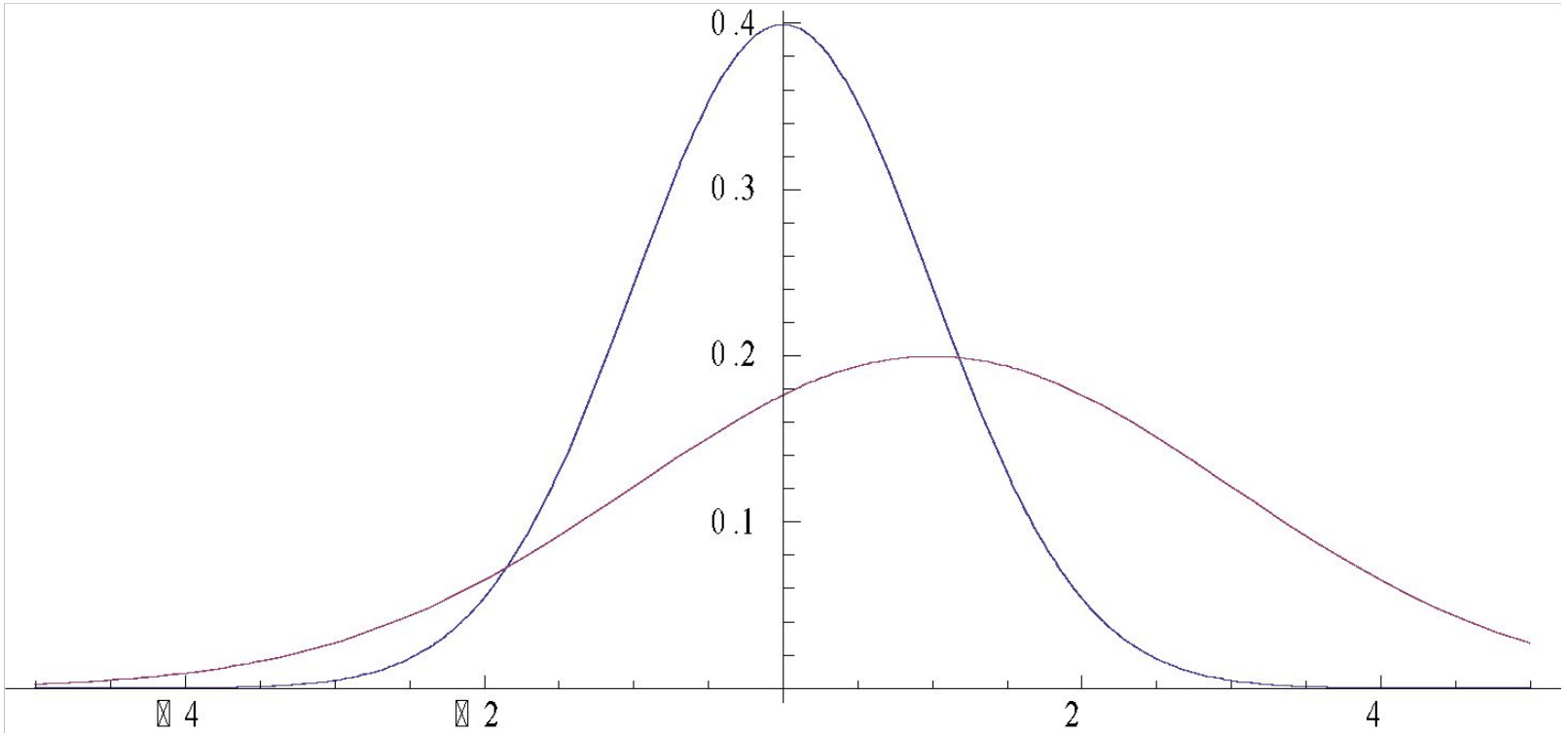
Среднее значение = μ

Среднеквадратичное отклонение = σ

Асимметрия = 0

Эксцесс = 0

Нормальное распределение



Некоторые свойства

68% значений отклоняются от среднего не более, чем на величину одного стандартного отклонения, 95% -- двух, 99,7% -- трех.

Распределение симметричное, эксцесс равен 0.