АНАЛИЗ ДАННЫХ В ИММУНОЛОГИИ

Ст. преподаватель
ЖИВИЦКАЯ ЕЛЕНА ПЕТРОВНА

ЗАДАЧИ ИММУНОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

1. Определение связи между несколькими иммунологическими и/или иными показателями без предположения о том, что они вызывают друг друга (не рассматривая их как следствие друг друга).
2. Исследование связи между иммунологическим показателем и клиническими данными, рассматривая их как следствие друг друга. В данном случае иммунологический показатель относится к независимым признакам и рассматривается в качестве фактора риска, а в качестве зависимых признаков выступают клинические данные (исход заболевания, тяжесть течения, стадия патологического процесса).
3. Комплексное исследование, включающее два или более объекта, описанных выше.
4. Компьютерное конструирование иммунологических процессов, заключающееся в прогнозировании иммуногенных последовательностей микробного генома, идентификации регуляторных молекул иммунного ответа и т. д.

Этапы анализа данных

Ввод данных

Преобразование данных

Визуализация данных

Статистический анализ

Собственно выбор метода, анализ данных и интерпретация результатов

Представление результатов

ОСНОВНЫЕ ПОНЯТИЯ СТАТИСТИКИ

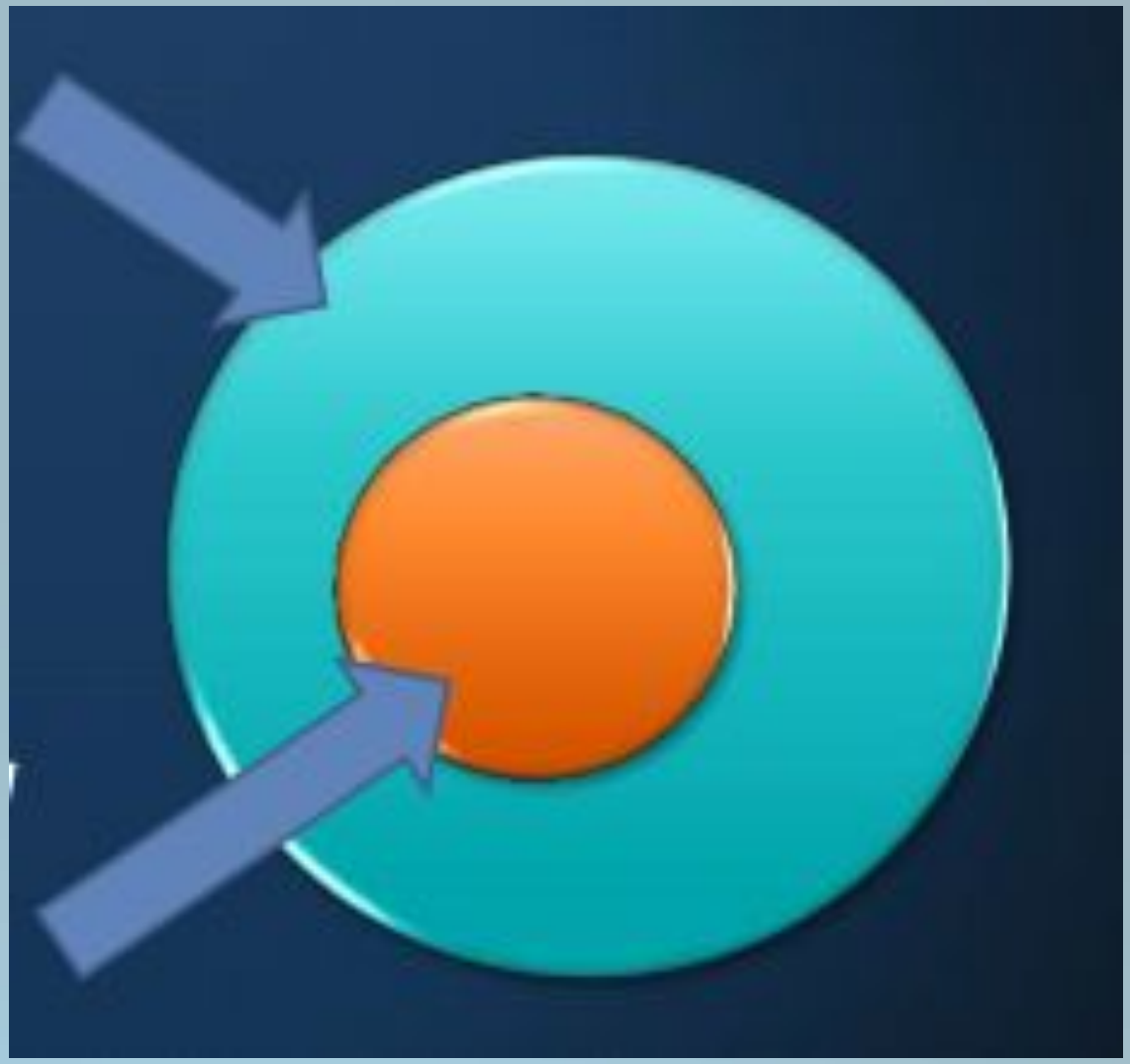
Совокупность – *это всякое множество отдельных объектов, отличающихся друг от друга и в то же время сходных по некоторым существенным признакам*.

Генеральная совокупность – *теоретически бесконечно большая совокупность всех единиц, которые могут быть к ней отнесены*.

Выборочная совокупность – *относительно небольшая выборка из генеральной совокупности, которая подвергается изучению*.

Объем совокупности – *число единиц совокупности*.

Генеральная совокупность

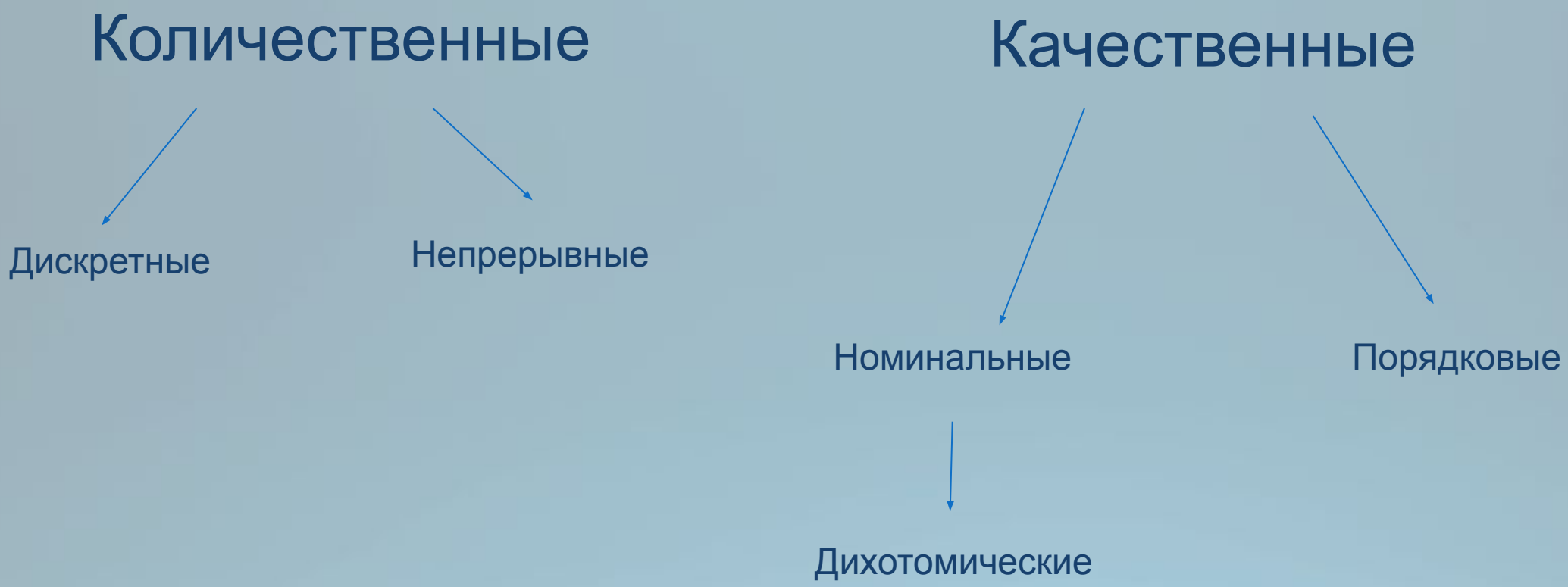


Выборочная совокупность

Репрезентативность - свойство выборочной совокупности отражать основные, важные для исследования, характеристики генеральной совокупности.

Репрезентативность определяет, насколько возможно обобщать результаты исследования с привлечением определённой выборки на всю генеральную совокупность, из которой она была собрана.

Типы данных


Количественные
Качественные
Дискретные
Непрерывные
Номинальные
Порядковые
Дихотомические

Типы данных

- Количественные
  - Различия равновелики
  - **Непрерывные** (напр., кровяное давление, масса тела, рост, возраст, биохимические показатели крови)
  - **Дискретные** (напр., кол-во беременностей, кол-во детей и др.; выражаются только целыми числами)

Типы данных

- Качественные
  Порядковые (отражают условную степень выраженности признака)
  - Можно ранжировать, но различия между категориями не обязательно равновелики
    - Напр., маленький/средний/большой, или состояние тяжести пациента

Типы данных

- Качественные
  Номинальные (отражают условные коды неизмеряемых категорий)
  - Коды диагнозов
  - Коды пола: мужской, женский
  - Раса: белая, черная, желтая
  - Семейное положение
  - **Дихотомические**: только 2 категории (да/нет, т.е. заболел/не заболел, умер/жив)

Для *различных переменных и шкал* применяются *разные методы* статистического анализа !!!

Виды статистических пакетов

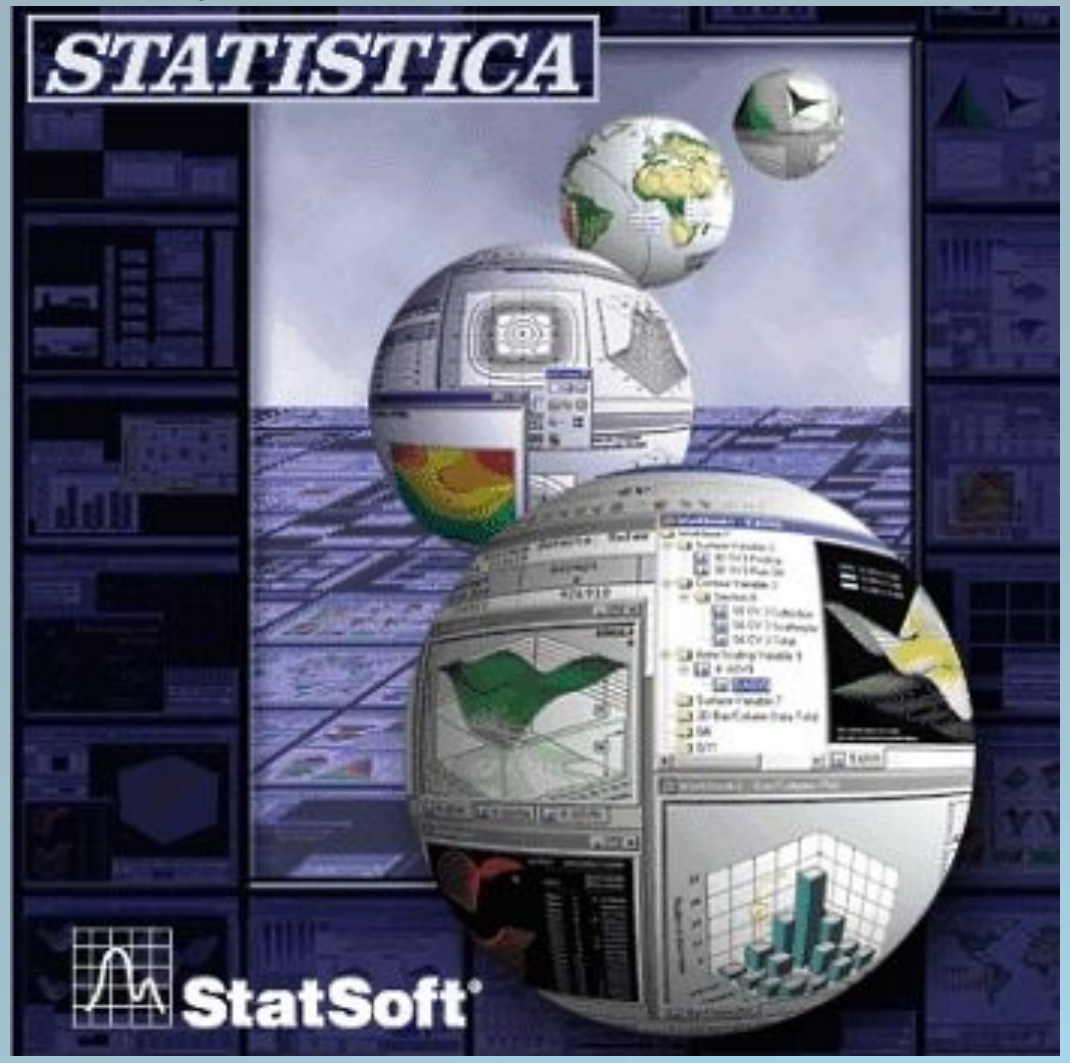
Универсальные пакеты

- отсутствие прямой ориентации на специфическую предметную область, предлагают широкий диапазон статистических методов (SPSS, Statistica, пакет анализа в Excel)

Специализированные пакеты

- обычно содержат методы из одного-двух разделов статистики или методы, используемые в конкретной предметной области (WinSTAT, Statit, STADIA)

STATISTICA - это универсальная интегрированная система, предназначенная для статистического анализа и визуализации данных, управления базами данных и разработки пользовательских приложений, содержащая широкий набор процедур анализа для применения в научных исследованиях .



Система обладает следующими общепризнанными достоинствами:

- содержит полный набор классических методов анализа данных;
- отвечает всем стандартам Windows;
- легка в освоении;
- данные системы STATISTICA легко конвертировать в различные базы данных и электронные таблицы;
- поддерживает высококачественную графику, позволяющую эффектно визуализировать данные и проводить графический анализ.

Основные формы представления выборки из генеральной совокупности

1. Представление выборки в несгруппированном виде, путём обычного перечисления вариант - **x**:

$$x_1, x_2, x_3, \ldots, x_n.$$

2. Представление выборки в упорядоченном виде: расположение вариант либо в порядке возрастания (чаще всего) либо в порядке убывания.

1 1 2 2 2 3 3 3 3 4 4 4 4 5 5 5 6 6 6 7 7

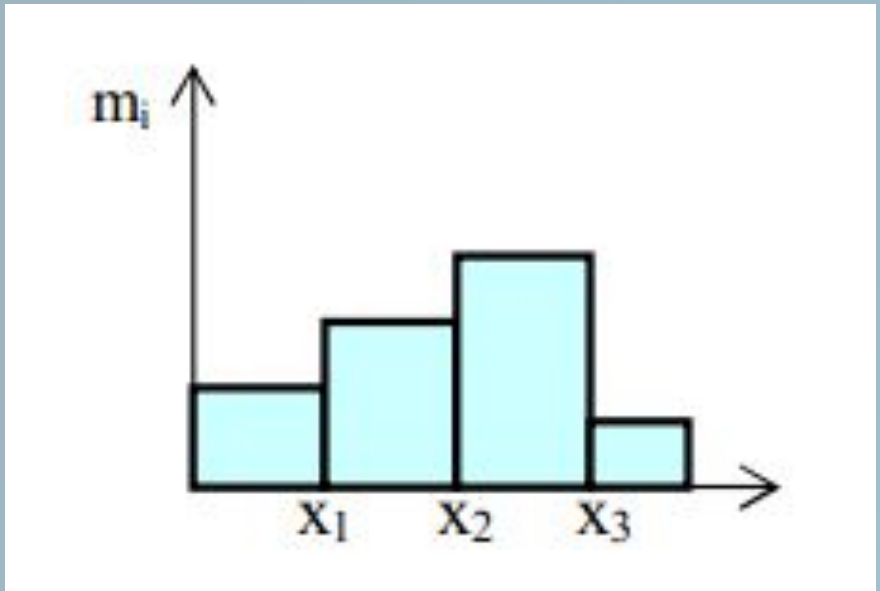
3. Представление выборки в сгруппированном виде, когда вместе с вариантами указываются числа (называемые частотами), равными числу повторений данной варианты в выборке.

Значения	x_1	x_2	…	x_n
Частоты	p_1	p_2	…	p_n
Относительные частоты	m_1	m_2	…	m_n

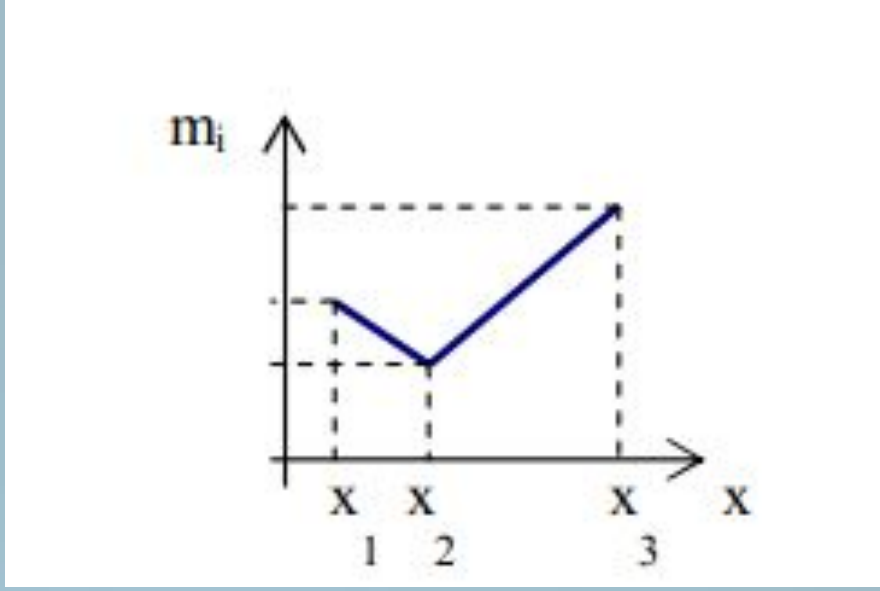
$m=p/n$, где n – объем выборки

Способы графического изображения данных

Гистограмма

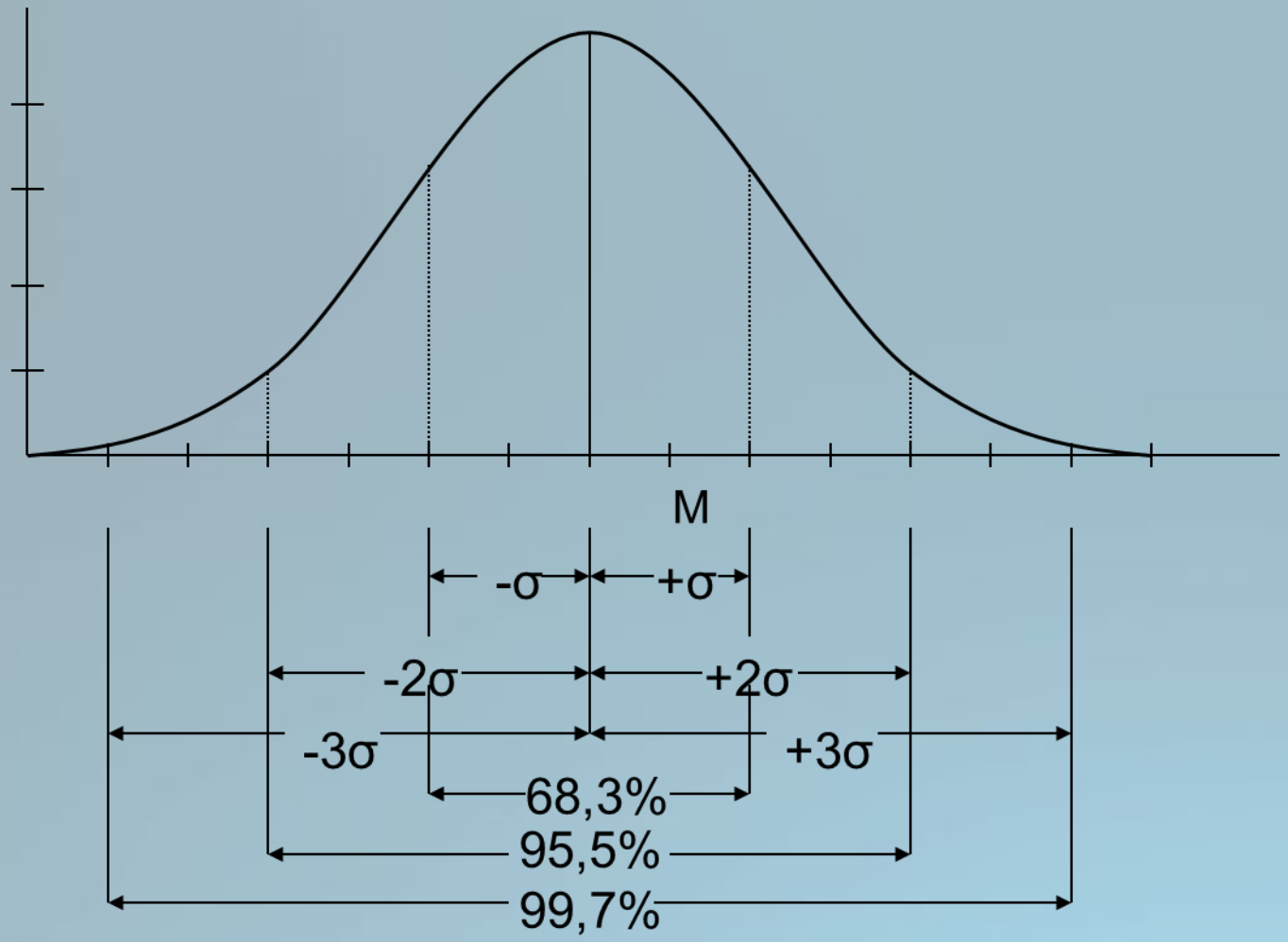


Полигон распределения



Первым этапом анализа количественных данных является анализ вида их распределения

Кривая нормального распределения



- 68% всех наблюдений лежат в диапазоне ±1 стандартное отклонение от среднего, а диапазон ±2 стандартных отклонения содержит 95% значений
- Числовые характеристики мода, медиана и среднее совпадают, распределение симметрично

Проверка соответствия распределения нормальному закону

- выборочные среднее, медиана и мода должны быть близки по значению и находиться примерно посередине между 25 и 75 процентилями;

- интервал среднее ± два стандартных отклонения должен включать примерно 95% значений выборки и не должен содержать значений, которых не может быть в данном распределении (например, отрицательных).

Статистические критерии для проверки нормальности распределения

- Критерий согласия χ^2 Пирсона (Pearson).
- Критерий Колмогорова-Смирнова (Kolmogorov-Smirnov). Применяется, если среднее значение и стандартное отклонение признака известны априори. *(для больших выборок)*
- Критерий Лиллиефорса (Lilliefors). Применяется, если среднее значение и стандартное отклонение признака неизвестны и вычисляются по выборке.
- Критерий Шапиро-Уилка (Shapiro–Wilk). Также применяется при априори неизвестных параметрах, является наиболее мощным, универсальным и строгим. *(для малых выборок)*

Как часто встречается нормальное распределение???

- Можно сказать, что из всех распределений в природе чаще всего встречается именно нормальное распределение – отсюда и произошло его название.
- Но для данных биомедицинских исследований это не всегда верно. Нормальное распределение встречается в биомедицинских признаках примерно в 20-25%.
- До тех пор пока выборка достаточно большая (например, 30 (100) или больше наблюдений), можно считать, что выборочное распределение нормально.

Статистические методы

- Описание данных
- Оценка статистической значимости результатов исследования (проверка гипотез)

Способы описания данных

Точечные характеристики
- Мода
- Медиана
- Средняя

Характеристики вариации
- Размах колебаний
- Дисперсия
- Стандартное отклонение

Точечные характеристики
(меры центральной тенденции)

- **Среднее арифметическое (среднее)**
- **Медиана (Me) - это средняя (центральная) варианта, делящая ряд распределения пополам, на две равные части.** *Применяется только для ранжированного (упорядоченного по убыванию или возрастанию) ряда значений признака.*
- **Мода (Mo) - наиболее часто встречающаяся в ряду распределения варианта**

Характеристики вариации (меры рассеяния)

Стандартное отклонение (σ) – величина, отражающая вариабельность данных относительно средней арифметической

Межквартильный размах (для медианы) – показывает значения 25-го и 75 процентилей, т.е. тот интервал, который включает в себя 50% данных в выборке
Пример описания: Me (25%÷75% процентили) = 70 (35÷89)

Интерпроцентильный размах – значения процентилей распределения данных (например, интервал между 10-м и 90-м процентилями)
Размах – разность максимального и минимального значений данных

Описание данных

Описание данных зависит от их типа (качественные или количественные) и способа их распределения !

Описание данных в зависимости от их типа

Количественные

Для описания используется **среднее или медиана**

Качественные (номинальные)

Для описания используется **мода**

Качественные (порядковые)

Для описания используется **медиана**

Какую среднюю величину использовать?

Нормальное

или

ненормальное распределение ?

Методы описания данных

- **Параметрический метод:** для нормально распределенных количественных данных
  - Для описания используется **среднее арифметическое и стандартное отклонение**

- **Непараметрический метод:** для не нормально распределенных количественных данных и качественных данных
  - Для описания используется **медиана и межквартильный размах**
  - Медиана менее чувствительна к асимметрии и «выскакивающим» значениям

Задача	Параметрические методы	Непараметрические методы
Выполнение описательной статистики	Вычисление среднего, стандартного отклонения	Вычисление медиан и интерквартильных интервалов, долей
Сравнение двух независимых групп по одному признаку	Критерий Стьюдента	Критерии Манна-Уитни, Колмогорова-Смирнова, Вальда-Вольфовица, χ^2, точный критерий Фишера
Сравнение двух зависимых групп по одному признаку	Критерий Стьюдента для зависимых выборок	Критерий Вилкоксона, критерий знаков, критерий МакНимара
Сравнение трех и более независимых групп по одному признаку	Дисперсионный анализ	Критерий Краскела-Уоллиса, медианный критерий, χ^2
Сравнение трех и более зависимых групп по одному признаку	Дисперсионный анализ для зависимых выборок	Критерий Фридмана, критерий Кохрена
Анализ взаимосвязи двух признаков	Коэффициент корреляции Пирсона, линейная регрессия	Коэффициенты корреляции Спирмена, Кендалла, гамма, точечно-бисериальный и рангово-бисериальный коэффициенты корреляции, коэффициент сопряженности, логистическая регрессия и др.