Введение в биостатистику

План лекции

- Что такое биостатистика
- Основные понятия
- Эпидемиология и биостатистика
- Статистика на этапах научного исследования

Что такое статистика?

Статистика – область деятельности людей, направленная на сбор информации и ее анализ с целью изучения массовых явлений в природе и обществе.

Статистика (Statistics) – наука, изучающая количественные характеристики массовых явлений и процессов в неразрывной связи с учетом их качественного своеобразия.

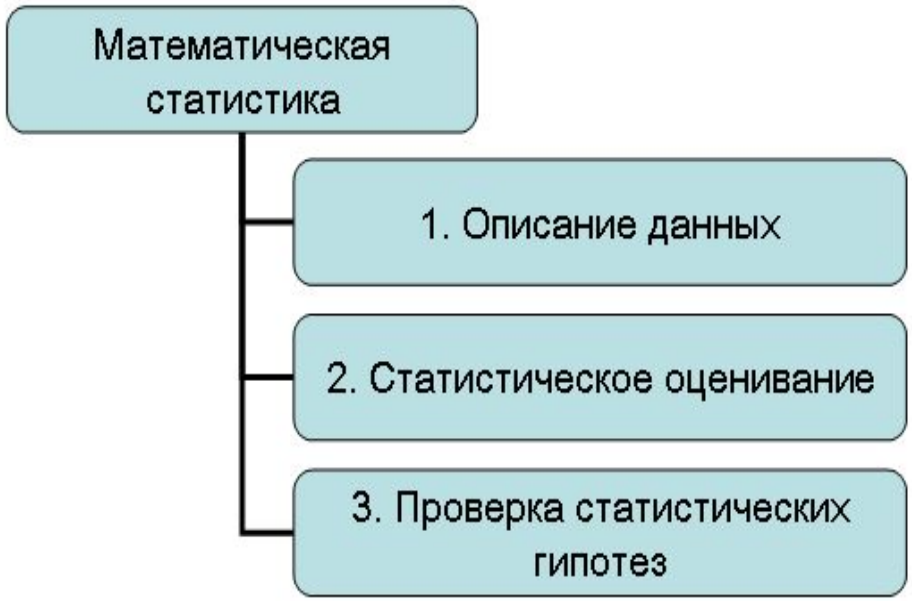
Статистика – наука о сборе, предоставлении и анализе данных (*Oxford Dictionary of Statistics*, 2002)

Статистика – функция от элементов выборки, которая используется для проверки статистических гипотез в качестве критерия.

Математическая статистика

- область науки, разрабатывающая математические методы для

изучения количественных характеристик массовых явлений.



Биостатистика -

- приложение общей теории статистики для решения научно-практических проблем в области биологии, медицины и здравоохранения

- Биостатистика – статистическая наука в приложении к живому миру. Включает в себя демографию, эпидемиологию и организацию клинических испытаний. Синоним – биометрия

- Медицинская (отраслевая) статистика:
  - статистика здоровья населения
  - статистика системы здравоохранения

Биостатистика

- Биостатистика и математическая статистика – родственные, но не одинаковые дисциплины. МС относится к точным наукам, биостатистика – к социальным.
- Биостатистика развивается.
- Она не «вскрывает истину», а лишь помогает интерпретации данных, как результаты лабораторных исследований помогают интерпретировать клиническую картину.
- Биостатистика не делает выводы, как врач-лаборант не ставит диагноз.
- Квинтэссенция статистики – это описание степени неопределенности наших заключений.

На чем основаны количественные методы

- Сбор, обработка и представление данных
- Анализ, Алгебра, Дифференциальные Уравнения, Ряды, Дискретная метаматика
- Описательные методы, Статистическое оценивание и Проверка гипотез
- Типы шкал, Шкала Лайкерта, Семантический дифференциал, Многомерное шкалирование
- Дисперсионный анализ, Кластерный анализ, Факторный анализ
- Вероятностные методы, Непрерывные и дискретные модели

Объект исследования

Объектами количественного исследования являются единицы, которые исследователь наблюдает, подсчитывает, описывает, измеряет для того, чтобы получить выводы относительно их свойств и наблюдаемых закономерностей.

Примеры: пациенты, организации, системы…

Переменные, признаки (variable)

Переменная, признак – это некоторая общая для всех изучаемых объектов характеристика или свойство, конкретные проявления которого могут меняться от объекта к объекту. Различные проявления признака называют **значениями, альтернативами, градациями.**

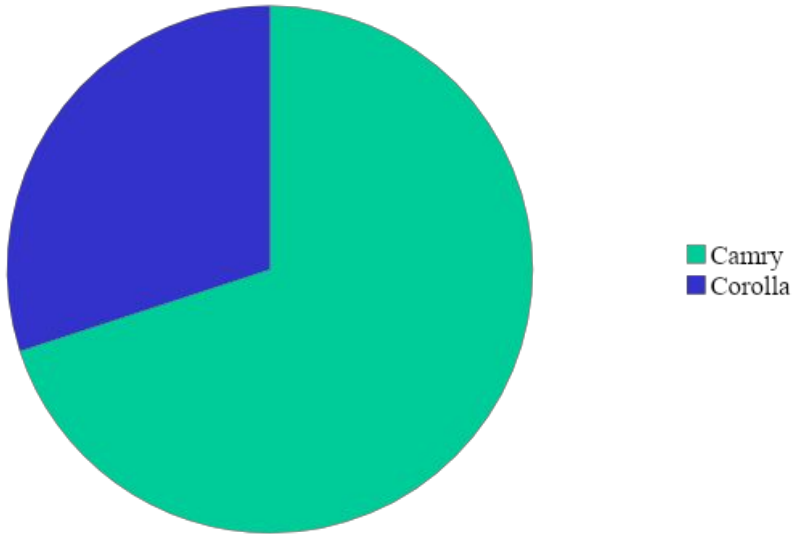
Умение «мыслить признаками», правильно определять переменные для достижения исследовательских целей является одним из важнейших качеств специалиста.

Примеры переменных

Переменная	**Возможные значения**
«Пол»	два значения: «мужчина» и «женщина»
«Профессия»	большое число значений, например, «политолог», «социолог», «менеджер»
«Рост»	от «очень низкий» до «очень высокий» или от 150 см до 210 см

Распределения переменных (distribution)

Значения переменной, которые она принимает для различных изучаемых объектов, приводят нас к необходимости рассматривать **распределение переменной.**



Пример распределения переменных

Изучаем сообщество из 5 000 жителей одного района:

Переменная	**Распределение**
«Пол»	55% женщин и 45% мужчин
«Возраст»	список возрастов 5 000 жителей
«Профессия» …	
«Годовой доход» …	

Распределение указанных переменных в изучаемом сообществе может отличиться от распределения этой же переменной, измеренной в другом сообществе.

Данные

Да́нные - представление фактов и идей в формализованном виде, пригодном для передачи и обработки в некотором информационном процессе.

Данные, являющиеся результатом фиксации некоторой информации, сами могут выступать как источник информации. Данные, извлекаемые из информации, могут подвергаться обработке, и результаты обработки фиксируются в виде новых данных.

Данные могут рассматриваться как записанные наблюдения, которые не используются, а пока хранятся.

Обработка данных включает операции:

ввод (сбор) данных — накопление данных с целью обеспечения достаточной полноты для принятия решений;

формализация данных — приведение данных, поступающих из разных источников, к одинаковой форме, для повышения их доступности;

фильтрация данных — это отсеивание «лишних» данных, в которых нет необходимости для повышения достоверности и адекватности;

сортировка данных — это упорядочивание данных по заданному признаку с целью удобства их использования;

архивация — это организация хранения данных в удобной и легкодоступной форме;

защита данных — включает меры, направленные на предотвращение утраты, воспроизведения и модификации данных;

транспортировка данных — приём и передача данных между участниками информационного процесса;

преобразование данных — это перевод данных из одной формы в другую или из одной структуры в другую.

Статистические данные

- числовые или нечисловые значения контролируемых параметров (признаков) исследуемых объектов, которые получены в результате наблюдений (измерений, анализов, испытаний, опытов и т.д.) определенного числа признаков, у каждой единицы, вошедшей в исследование.

Генеральная совокупность и выборка

- 150 тыс. человек
- Генеральная совокупность

- 250 человек
- Выборка

Какова доля одиноких людей?

Генеральная совокупность и выборка

Генеральная совокупность (population) – вся интересующая исследователя совокупность изучаемых объектов.

Выборка (sample) – некоторая, обычно небольшая, часть генеральной совокупности, отбираемая специальным образом и исследуемая с целью получения выводов о свойствах генеральной совокупности.

Репрезентативная выборка

(от франц. representatif — показательный, характерный), представительность, мера возможности восстановить, воспроизвести представление о целом по его части или мера возможности распространить представление о части на включающее эту часть целое.

Репрезентативная выборка хорошо представляет генеральную совокупность.

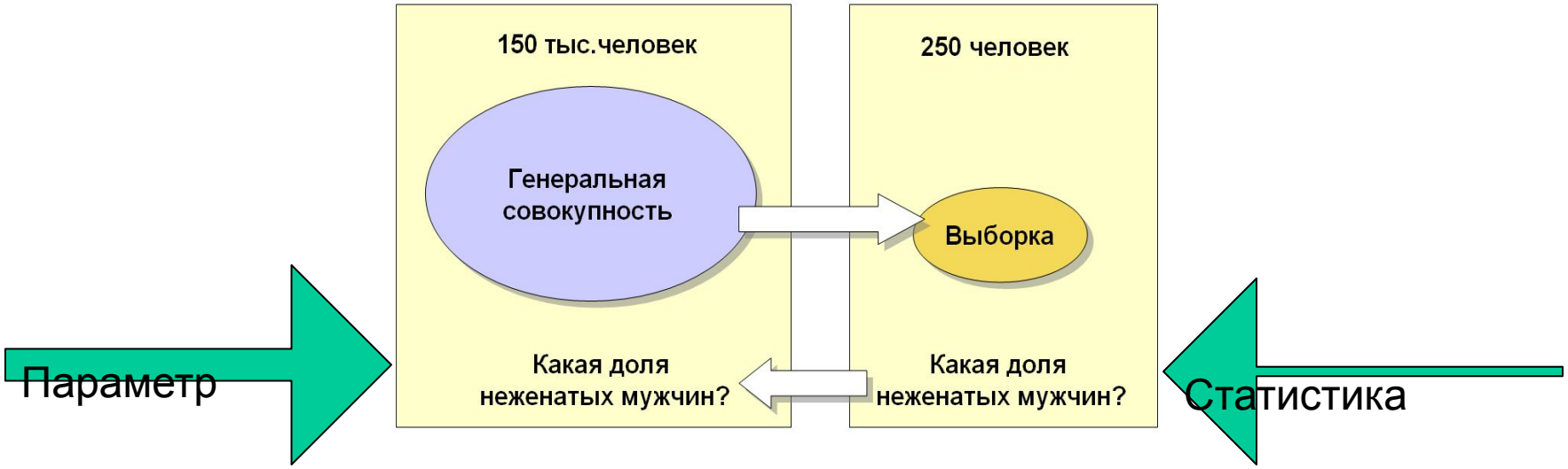
Это означает, что каждое свойство (или комбинация свойств) наблюдается в выборке с той же частотой, что и в генеральной совокупности.

Параметры и статистики

Параметры - *характеристики генеральной совокупности.*

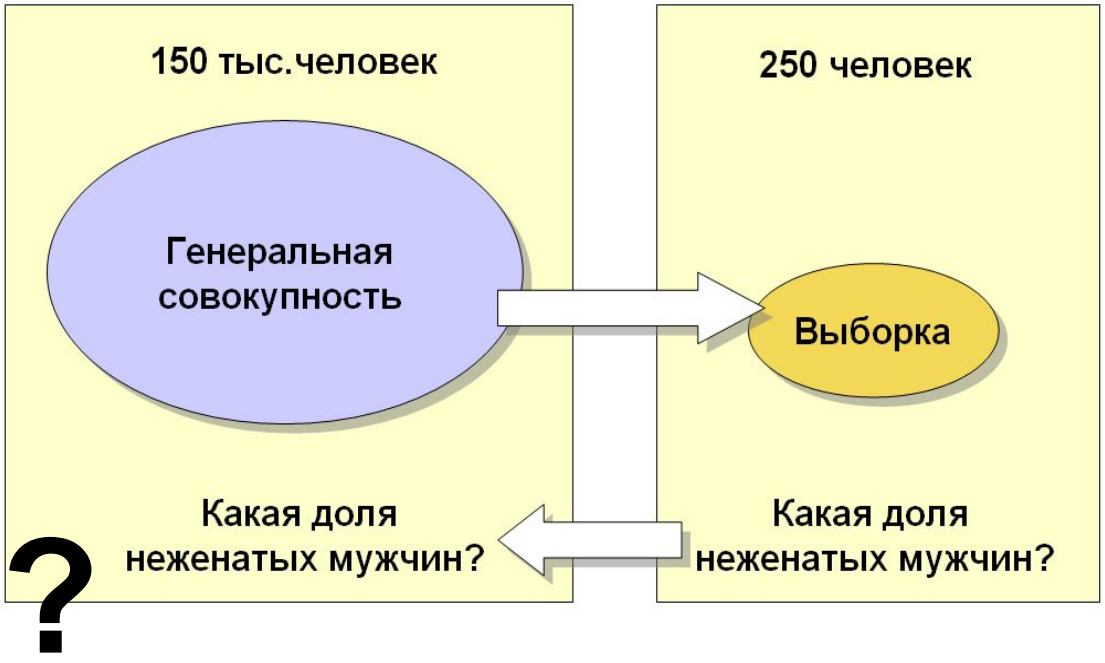
Статистики - характеристики выборки.

Мы будем использовать статистики для оценки тех параметров генеральной совокупности, которым они соответствуют.



Гипотеза (hypothesis) –

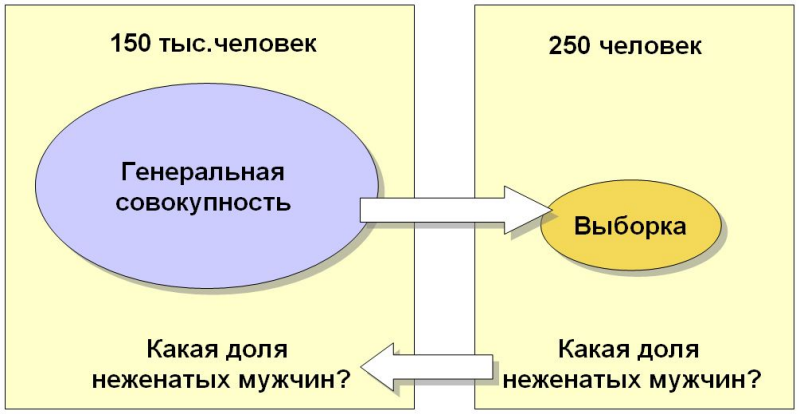
предположение относительно значения параметров генеральной совокупности, которое подлежит проверке на основе анализа выборки.



Описательная и аналитическая статистика

Описательная статистика (descriptive statistics) состоит из статистических методов, которые позволяют проводить сбор, упорядочение, обобщение и визуализацию данных.

Аналитическая статистика (inferential statistics) состоит из методов, которые на основе изучения статистик выборки позволяют получать выводы о параметрах генеральной совокупности.



Эпидемиология и биостатистика

Эпидемиология - дисциплина, изучающая особенности болезней с точки зрения их распространения и способов борьбы с ними.

Эпидемиология является пропедевтической дисциплиной медицины и содержит основы понимания медицинской реальности, эффективности вмешательств и методов исследований в медицине

Эпидемиология (ἐπιδημία — имеющая всенародное распространение) — общемедицинская наука, изучающая закономерности возникновения и распространения заболеваний различной этиологии с целью разработки профилактических мероприятий (преморбидная, первичная, вторичная и третичная профилактика)

Графство Освего, США, 1940

- 19 апреля 1940 года эпидемиологи были вызваны в деревню Ликоминг, графства Освего, расследовать вспышку желудочно-кишечного заболевания
- Все заболевшие присутствовали на церковном ужине 18 апреля
- Члены семей, которые не присутствовали на ужине, не заболели
- Эпидемиолог опросил 75 из 80 присутствовавших на ужине, из которых 46 сообщили о наличии симптомов (тошнота, рвота, боли в животе)

Расследование

- Ужин проходил в подвале церкви с 18 до 23 часов. Еду принесли прихожане. Еда была разложена на столах и прихожане брали ее и ели.

Еда

- Ветчина
- Шпинат
- Картофельное пюре
- Салат
- Черный хлеб
- Молоко
- Кофе
- Вода
- Пироги
- Ванильное мороженное
- Шоколадное мороженное
- Фруктовый салат

Что явилось причиной?

Подходы

- Построить таблицы, где съеденная еда будет фактором риска, а результатом – наличие или отсутствие заболевания

Таблицы

```
> table(baked.ham,ill)
          ill
baked.ham  N  Y
        N 12 17
        Y 17 29
> table(spinach,ill)
       ill
spinach  N  Y
      N 12 20
      Y 17 26
> table(mashed.potato,ill)
             ill
mashed.potato  N  Y
            N 14 23
            Y 14 23
> table(cabbage.salad,ill)
             ill
cabbage.salad  N  Y
            N 19 28
            Y 10 18
```

Что можно сказать?

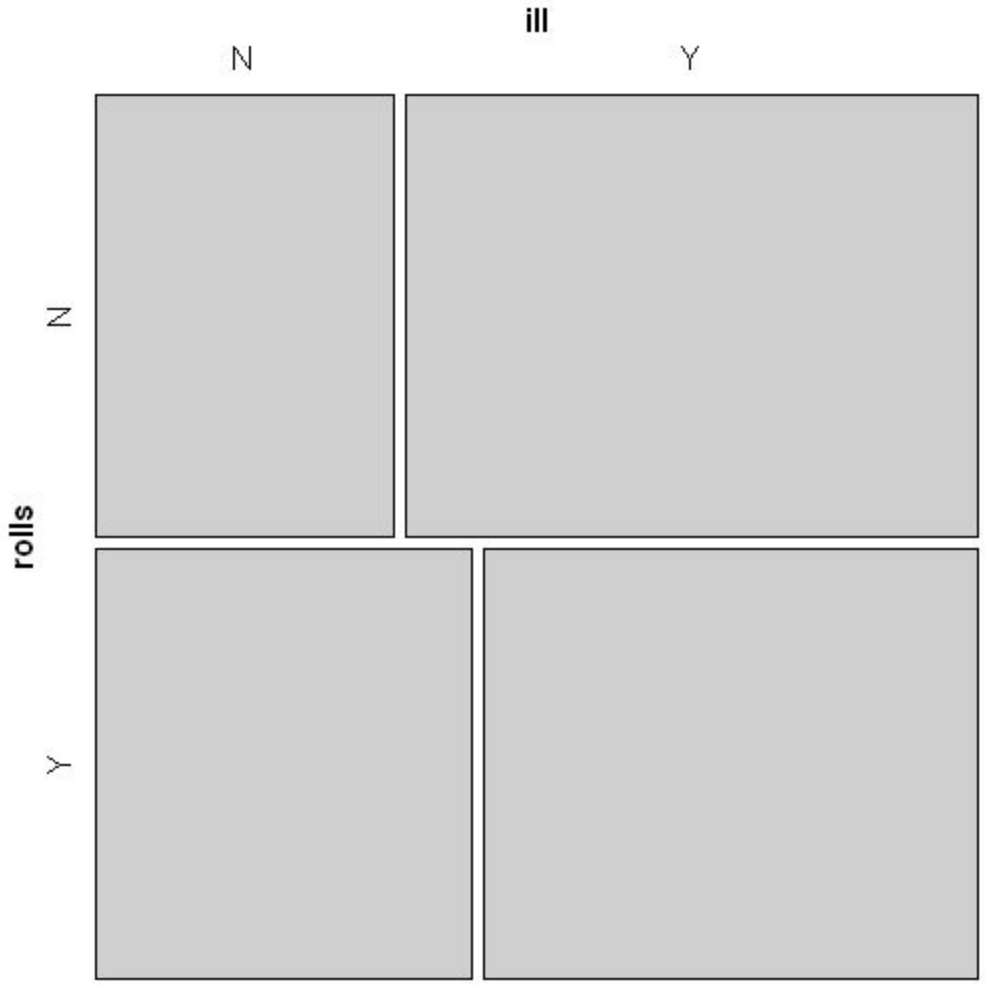
Таблицы

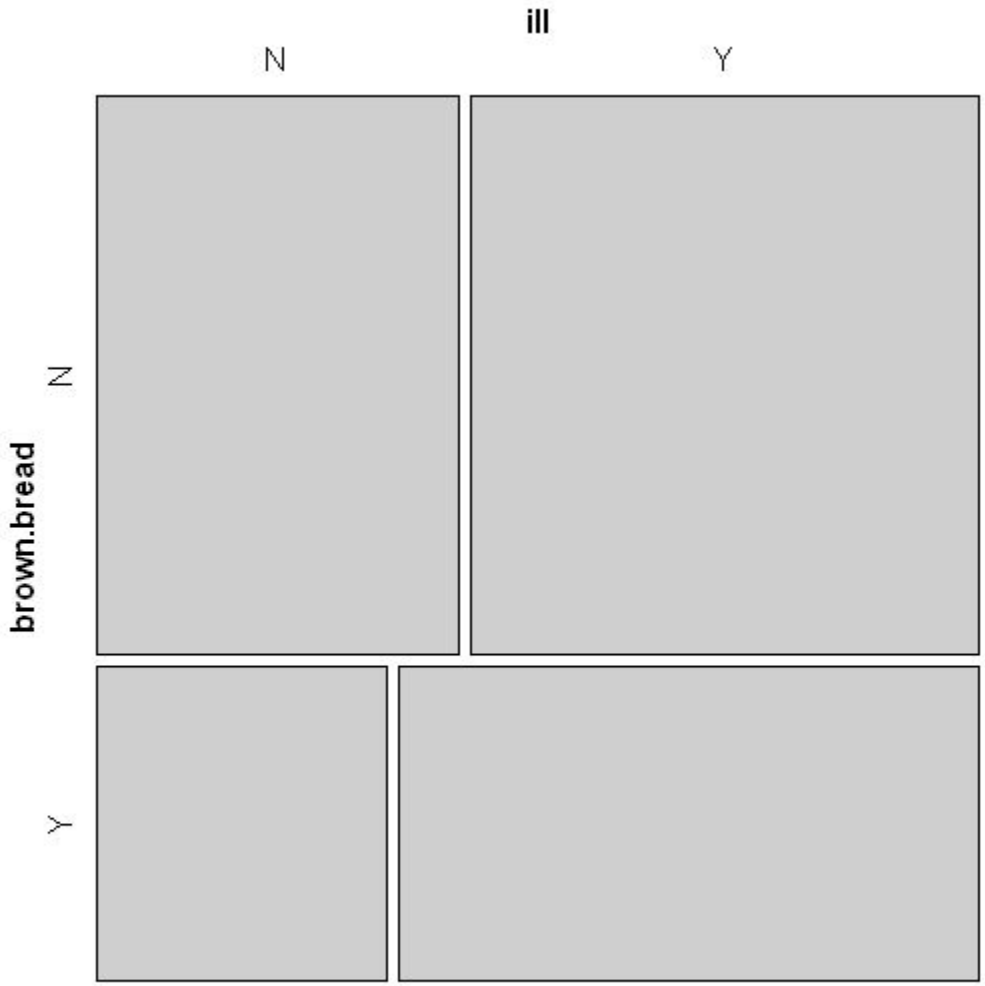
- Оценить таблицу сложно
- Однако можно попытаться суммировать ее при помощи одного числа (отношение шансов)

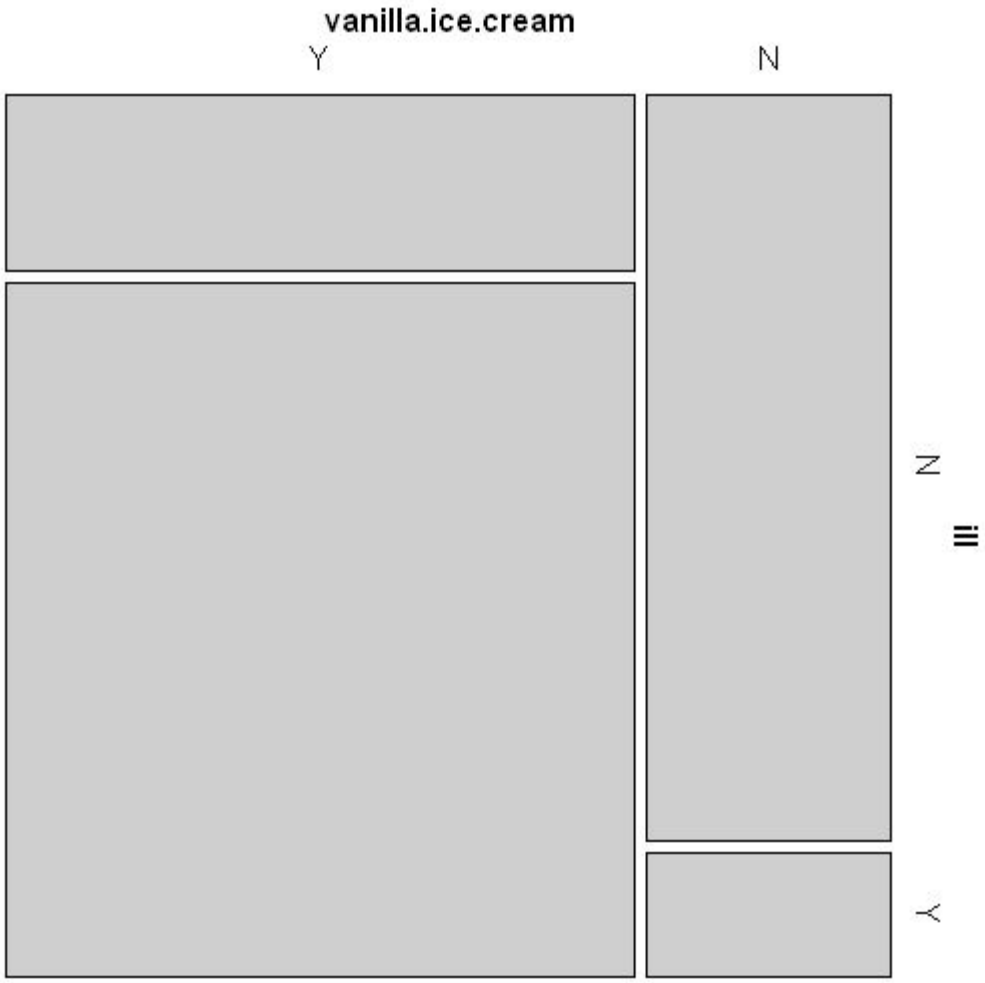
$$ОШ = (a \times d) / (b \times c)$$

! Первая задача статистики - представить данные в сконденсированном виде для облегчения анализа

Графики




ill
N
Y
brown.bread
N
Y


vanilla.ice.cream
Y
N
N
Y

Отношение шансов (ОШ)

- Это отношение шансов развития заболевания среди подвергшейся воздействию популяции к шансам развития заболевания в не подвергавшейся воздействию популяции

Для одномоментных исследований случай-контроль:

$$ОШ = (a \times d) / (b \times c)$$

Отношения шансов

- Ветчина 1,20
- Шпинат 0,92
- Картофельное пюре 0,38
- Салат 1,21
- Черный хлеб 1,41
- Молоко 0,61
- Кофе 1.00
- Вода 0,65
- Пироги 1,73
- Ванильное мороженное 21,4
- Шоколадное мороженное 0,41
- Фруктовый салат 1,24
- Рулеты 0,69

Кто виноват?

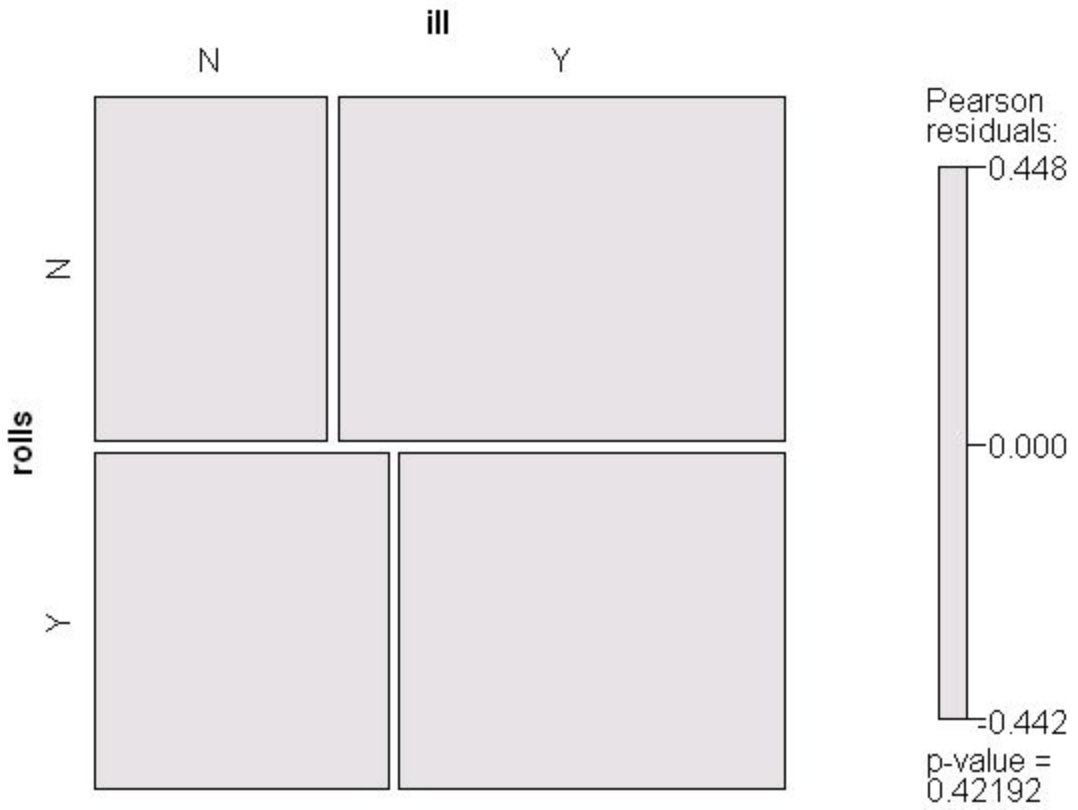
Итак

- Мы просуммировали значения, но ряд отношений шансов больше нуля, а ряд – меньше. Только для кофе ответ очевиден (ОШ=1)
- Но, может, это случайные колебания?

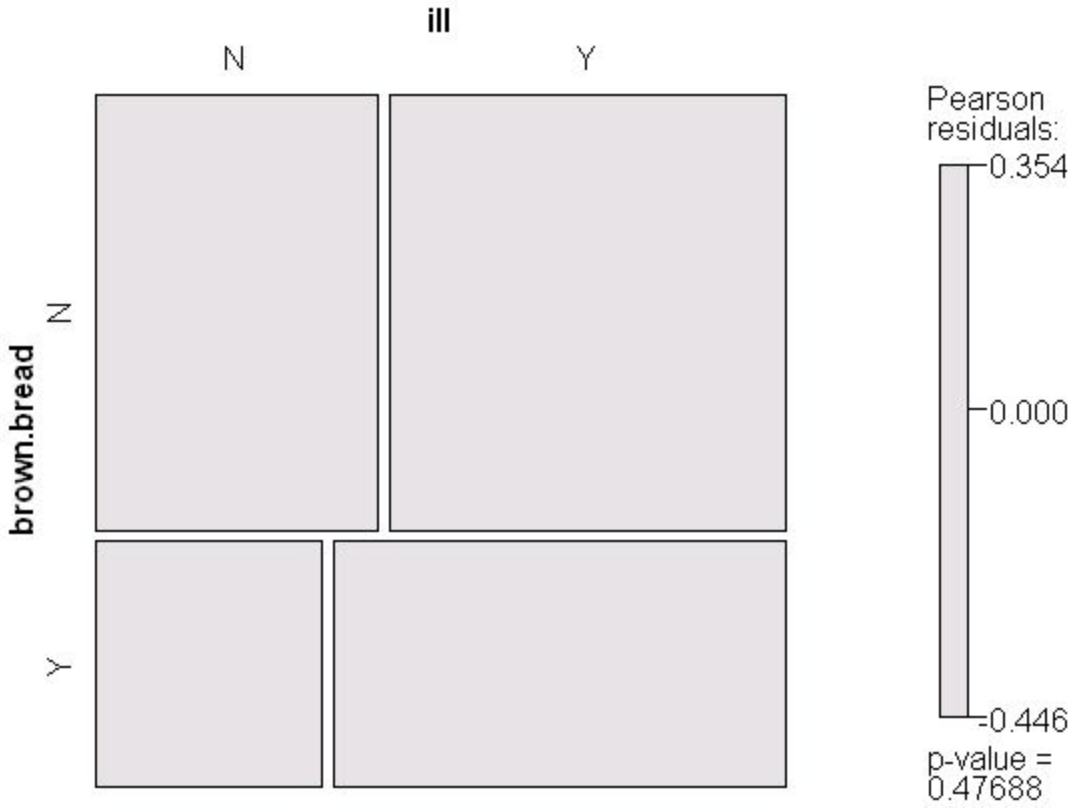
  !Вторая задача статистики

  – Определить роль случайных колебаний в полученных результатах

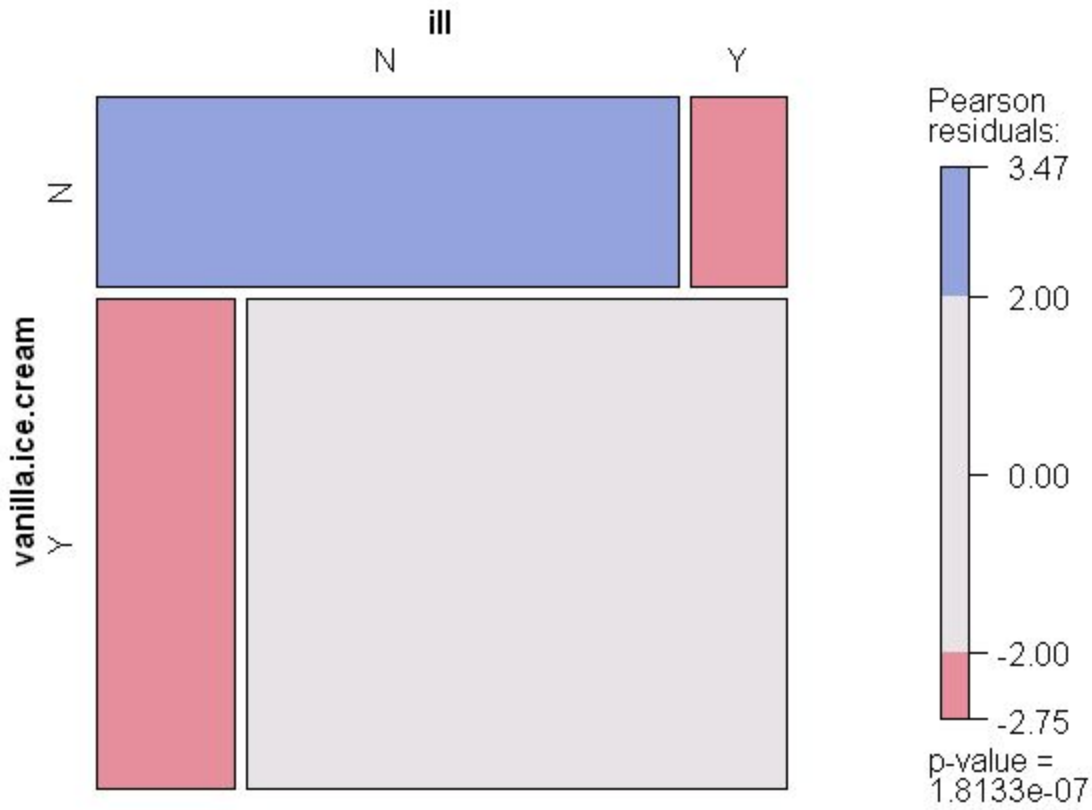
Графически


ill
N
Y
rolls
N
Y
Pearson
residuals:
0.448
0.000
-0.442
p-value =
0.42192

Графически


ill
N
Y
brown.bread
N
Y
Pearson
residuals:
0.354
0.000
-0.446
p-value =
0.47688

Графически


ill
N
Y
N
Y
vanilla.ice.cream
Pearson
residuals:
3.47
2.00
0.00
-2.00
-2.75
p-value =
1.8133e-07

Биостатистика – инструмент эпидемиологии

- Анализ показал, что причиной пищевого отравления было ванильное мороженное
- Стало возможным найти причину заболевания, даже не располагая результатами посева или иными методами
- Именно поэтому биостатистика стала основным инструментом эпидемиологии

Три основные задачи биостатистики:

- Суммирование и описание данных
- Обнаружение общих закономерностей на основании полученных данных
- Обнаружение взаимосвязей, оценка различий между группами и влияния случайностей на результат

Описательная статистика

- Графические методы

- Численные методы
  - показатели центральной тенденции
  - показатели разброса

Доказательная статистика

- Тестирование статистических гипотез
- Мультивариантная статистика
- Data Mining («*обнаружение знаний в базах данных*», *интеллектуальный анализ данных*) - *обнаружение в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний*

Этапы научно-практического исследования:

1. Формулирование целей и задач
2. Организация исследования
3. Сбор информации
4. Обработка информации
5. Анализ результатов исследования
6. Распространение и внедрение результатов исследования в практику

Формулирование цели и задач

- Рабочая гипотеза (ожидаемые результаты)
- Размер выборки

Организация исследования

Выбор **объекта** наблюдения:

объект наблюдения – статистическая совокупность, состоящая из отдельных предметов или явлений – единиц наблюдений, являющихся носителями признаков и их значений.

Организация исследования

Типы признаков:

- Качественные, категориальные:
  - номинальные
  - дихотомические
  - порядковые, ординальные, ранжируемые
- Количественные, интервальные
  - дискретные
  - непрерывные

Сбор информации

Регистрационный документ (анкета, бланк, карта и т.п):

- Включает обязательные вопросы (номер, дата, название организации, и т.п.)
- Предполагает унифицированность заполнения, однозначность формулировок вопросов
- Удобен для чтения и заполнения, а также для шифровки и обработки данных (альтернативные ответы или подсказы)

Фрагмент анкеты

1. ФИО ребенка _____
2. Возраст ребенка _____ лет
3. Пол: 3.1. муж. 3.2. жен
4. Рост _____ см
5. Количество детей в семье:
6. Образование матери:
   - 6.1. неполное среднее
   - 6.2. среднее
   - 6.3. специальное среднее
   - 6.4. высшее
   - 6.5. ученая степень
7. В чем заключается ваше общение с ребенком:
   - 7.1. проверка уроков
   - 7.2. совместные прогулки
   - 7.3. …
   - 7.6. другое (указать что)

Обработка данных

- Создание и подготовка базы данных

Анализ результатов исследования

1. Описание результатов исследования
2. Сравнение различных статистических совокупностей
3. Дифференциация, оценка взаимодействия и интеграция факторов
4. Анализ динамики явлений (динамические или временные ряды)

Результаты исследования

- Результаты должны быть воспроизводимыми
- Каждый шаг анализа должен быть задокументирован
- Профессиональные статистические системы всегда базируются на программном языке, а не на интерфейсе «укажи и кликни»
  - С этим интерфейсом полное документирование и воспроизведение невозможно.

Контроль качества статистического анализа

- На этапе планирования
  - План статистического анализа
    - Краткое содержание протокола
    - Описание всех измеряемых переменных
    - Описание всех производных переменных и как они рассчитываются
    - Для каждой зависимой переменной – метод анализа
    - Дополнительные методы анализа и модели с обоснованием их выбора

- На этапе анализа
  - Подготовительный этап
  - Адекватность выбранной методологии
  - Правильность используемых программ
  - Полнота анализа и его понятность
  - Проверка качества

Подготовка анализа

- Анализ статистического плана на наличие стандартных и нестандартных подходов
- Подготовленность персонала к выполнению статистического анализа
- Анализ протокола исследования
- Анализ статистического плана
- Анализ адекватности базы данных

Адекватность базы данных

- Использование валидизированных методов импорта и экспорта данных
- Проводится анализ протоколов ведения базы данных
- Если гарантий точности базы нет анализируется случайная выборка 5% случаев (или 100) – сравниваются записи в базе данных с ИРК

Адекватность выбранной методологии

- Адекватность методов целям и задачам исследования
- Адекватность выбранной методологии
- Адекватность выбранных моделей
  - Может требовать консультации с другим статистиком
- Важно соответствие статистическому плану исследования

Правильность используемых программ

- Использование валидизированных макро или программ, включенных в стандартные статистические пакеты (SAS/S-plus).
- Все макро написанные кем-то еще валидизируются и хранятся вместе с данными.
- Все программы, которые используются для расчетов, построения таблиц, рисунков и списков хранятся вместе с данными.

Пакеты прикладных программ

- **SPSS (SPSS Inc.,USA)**
- **SAS**
- **STATA**
- **STATISTICA (StatSoft, USA; StatSoft-Russia)**
- **BIOSTATISTICA (S.A. Glantz, McGraw Hill, перевод на русский язык – «Практика», 1998)**
- **EpiInfo 2000 (Centers for Disease Control and Prevention, USA)**

Спасибо за внимание!