Part1 Introduction

statistical inference统计推断

Course Objectives

3

1. Recognize the importance of data collection, identify limitations in data collection methods.
2. Use software to summarize data numerically and visually
3. Be able to perform basic data analysis
4. Have a conceptual understanding of statistical inference
5. Apply estimation and testing methods to analyze variables to understand the biological phenomena and make decisions
6. Interpret results correctly
7. Understand the basic statistical concepts of clinical trials

清华大学统计学研究中心

大部分都是计算 不会要求有那么多（？？）证明题

什么是医学统计

Example1**辛普森悖论**：数据整体比例和分组时方向相反（数据会骗人/呈现方式不同，结论不同）
?**混杂因素**confounding variable
应用：流行病学

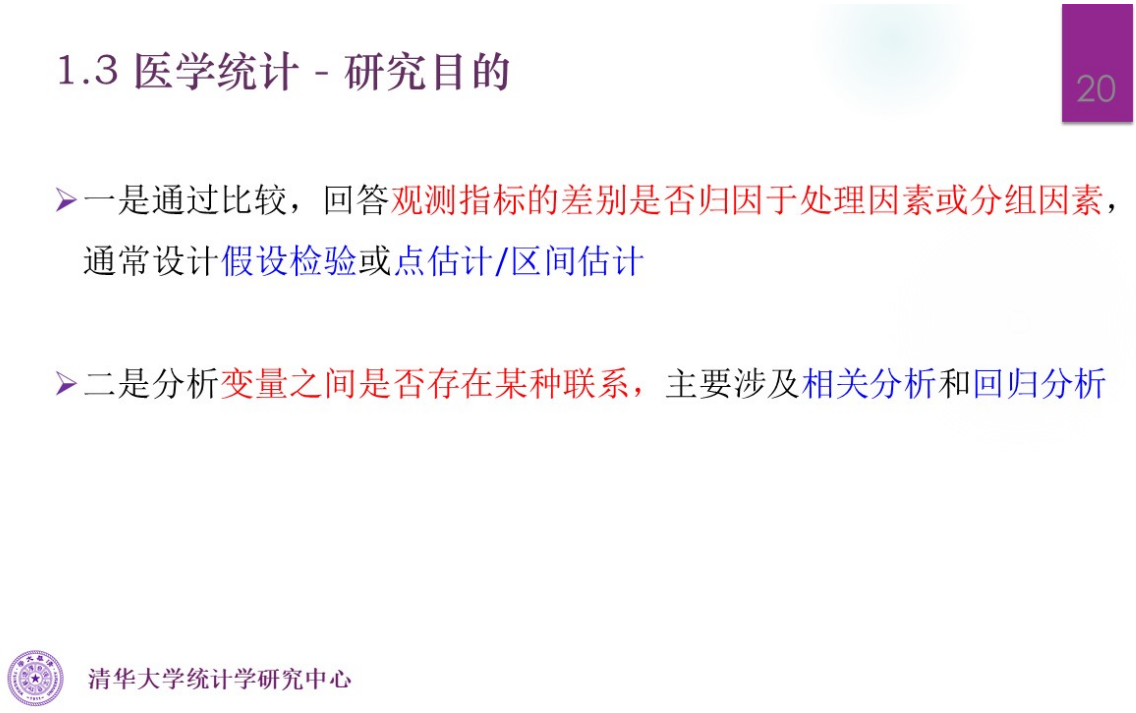
Example2
回答的可靠性

statistics is a science dealing with *random phenomena*

研究目的

- 变量与处理因素、分组因素ry归因（假设检验/点估计、区间估计）

- 变量之间的联系（相关分析/回归分析）



数据/资料采集

随机变量 *大写字母* random variable
代表随机现象的结果

资料data

- numerical data数值型资料aka quantitative data量化型资料
  - 连续型（设备精度的问题不代表数值本身取不到）
  - 离散型
- categorical data分类型资料aka qualitative data
  - 无序型（不能比大小）
  - 有序型（eg ABCDEF/还可以、恶化、死了）
  - 定距型?对连续型进行分段切分（收入档次）

  数据类型、分析对象决定统计方法
- 计数资料
- 生存资料

数据收集方法：①系统性地收集数据②需要预先计划

- 观测法 流行病学
- 面试
- 问卷
- 文件

- 公共设备?云计算中心

2.4 Common problems in data collection

- Lack of adequate time
- Expense
- Inadequately trained and experienced staff
- Public data service
- Invasion of privacy
- Bias
- Etc.

结果不可靠

- primary data（节省资源：时间、人力……）
- secondary data
  收集到的数据应能反映想研究的问题

数据/资料的描述性统计与呈现

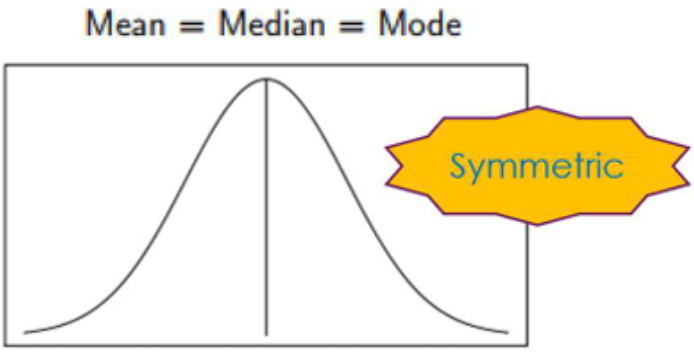
收集数据
↓
Step1数据分析（整体看） **探索性数据分析EDA**

- descriptive summary
- frequency table
- graphs

Descriptive summary

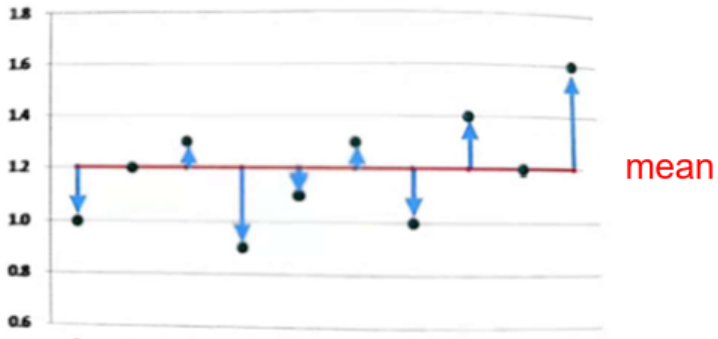
- central tendency
  - **mean/median/mode**

- y轴代表**频数/频率**；x轴代表**数值**注意左右(尾巴方向)对应



  - 在有偏移的情况下，mean不适合代表中心趋势（*mean往偏态方向移动*）；median相对稳定，比较合适
  - **outlier离群值** the mean is extremely sensitive to unusual value (i.e., outlier);the median is NOT as sensitive
- Dispersion 分散/变异程度（离散性）
  - **variance方差/standard deviation标准差** *单位一致* **/range极差/interquartile range四分位数间距** ←不太受outlier影响(?)
  - first quartile = 25th percentile = lower quartile(third upper 75th;second middle median)
  - **离均差平方和 SS**=方差×数据量n

Sum of squares of deviations from mean (SS) 离均差平方和



- Var
- Sta

  - 四分位数 **间距**：两个数之间的距离
    IQR=Q3-Q1 25-median-75 Q1-Q2-Q3
- Relative position相对位置
  - **percentile百分位数**

    For example, 15% of data values are less than or equal to the 15th percentile.

    记为P_r

- **Standard Z-score**
  Z=（X-μ）/σ （和正态分布有点像
- (categorical variables) rate and proportion *对分类型*
  - ☆☆☆Rate ≠ Proportion
  - rate隐含指发生率

Part2 Plot your data

标题
数据的可视化
如何使用图表

Frequency table频数表

分类型数据

Pie chart

分类型数据 代表频数的相对比例 3-10类会用（占比太小的类合并称为其它类并具体列出来）
信息过于不丰富、角度比大小不直观

```
data(Arthritis)
counts<- table(Arthritis$Improved)
counts
head(Arthritis)
```

Bar chart柱状图/条形图

y 频数 x类别
类别之间有gap，大小适中

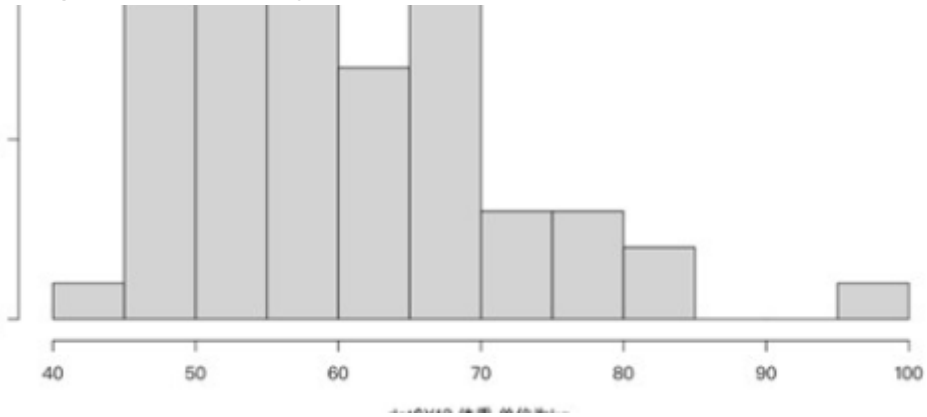
直观、易比较长短、可以添加额外信息，有助于比较

Histogram直方图

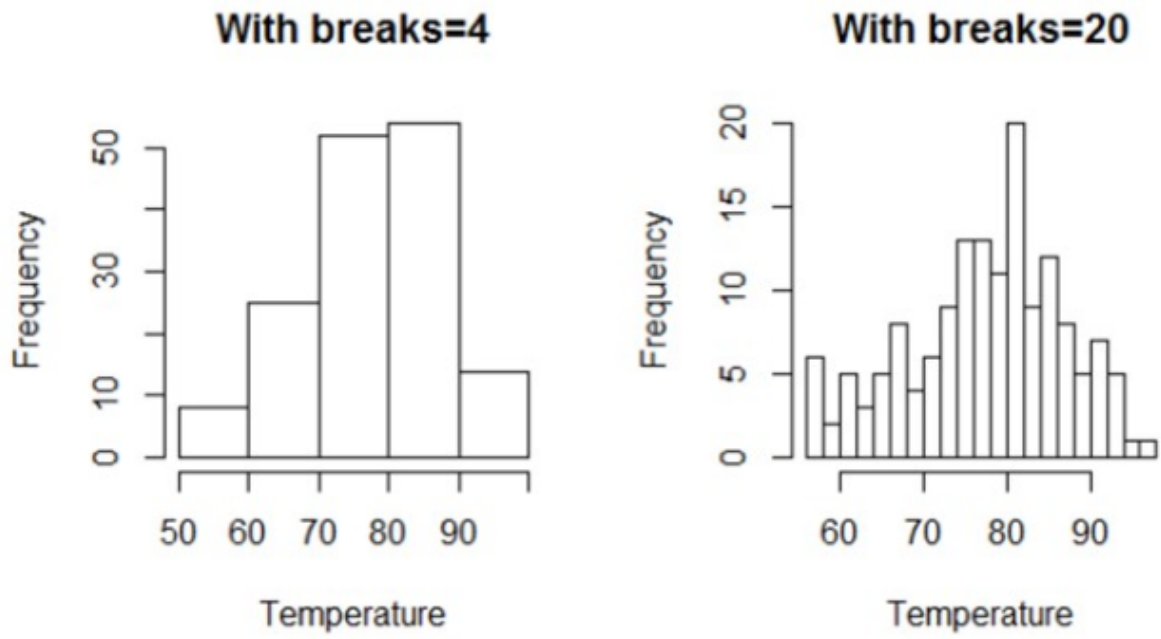
y频数
展示 **连续数据分布** 最常用的工具

```
hist（x,breaks,freq)
```

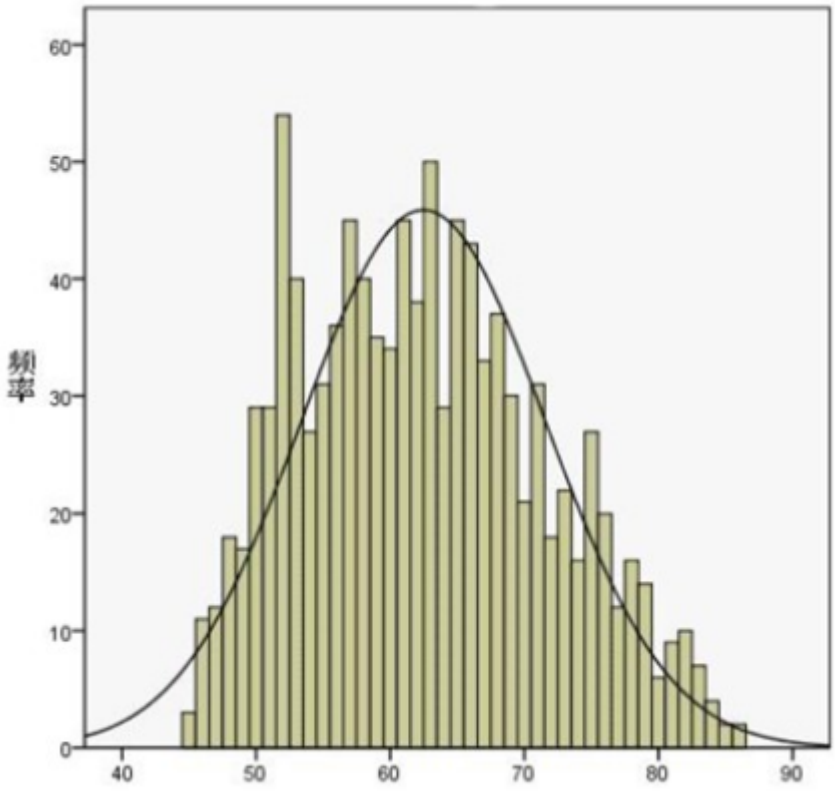
没有gap（频数为0 √ 是gap ×）

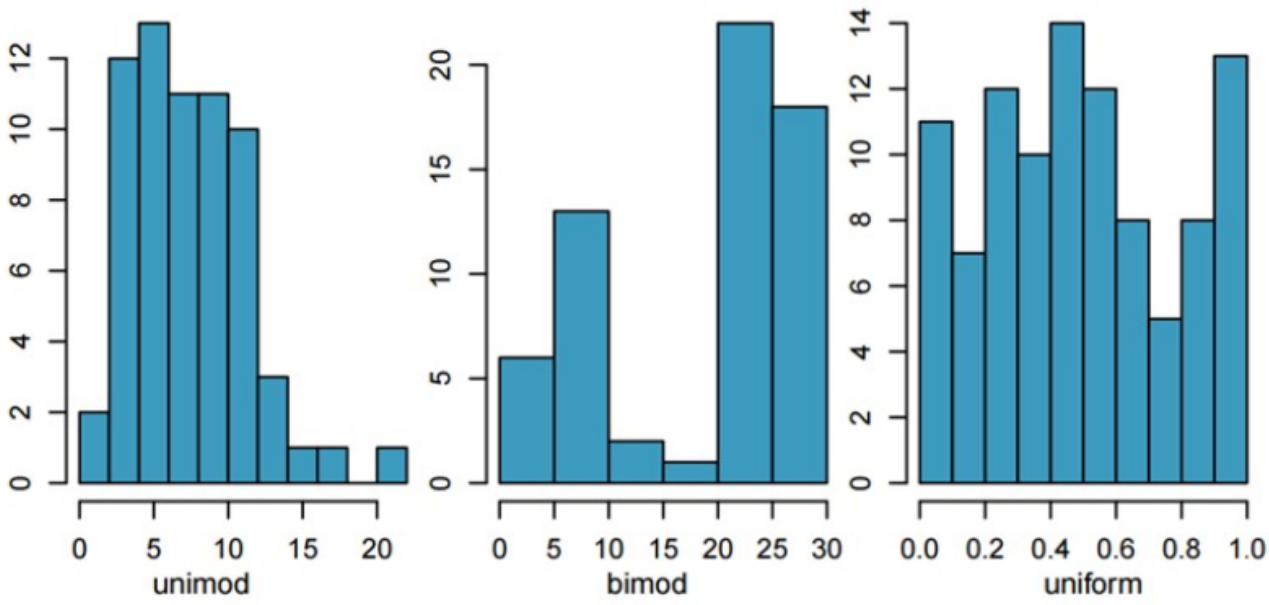


离群值↑



条数=break+1
preferable when sample size is large but individual observations are not of interest
可以看形状：对称/左偏/右偏 单峰umimod/双峰bimod/平均uniform；可以看中心位置；可以看数据密度（任意区间内）；可以看出离群值
光滑曲线：**估计**分布情况





biomod可能表示有隐含因素让结果呈现完全不同的情况（eg有人完全没学，有人学了）

Stem-and-leaf Plots茎叶图

直方图适用于大量数据，茎叶图适用于小数据&位数相近
匹配

Age Interval	Observations
20-29	5 6 9
30-39	2 5 6 8
40-49	4 9
50-59	1

也有点像直方图

☆Boxplot箱线图

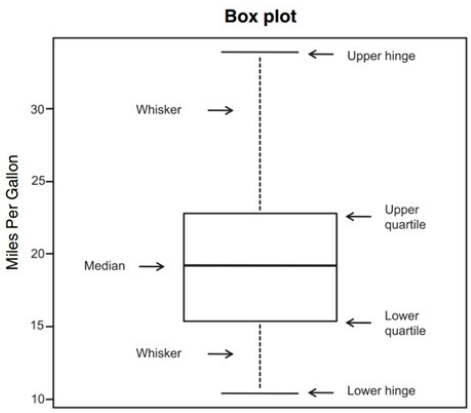
数值型数据

Boxplot (箱线图)

21

- 从四分位数的角度出发描述数值型数据的分布
- 最大值，上四分位数Q_3,中位住Q_2 ， 下四分位数Q_1，最小值
- 通过每一段数据占据的长度，来推断数据的集中或离散趋势；长度越短，数据在该区间越密集，反之越稀疏。

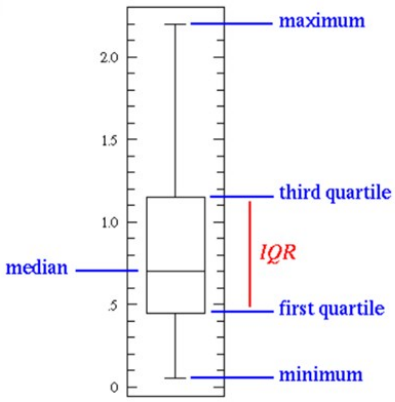
```
boxplot(mtcars$mpg,
main="Box plot",
ylab="Miles per Gallon")
```

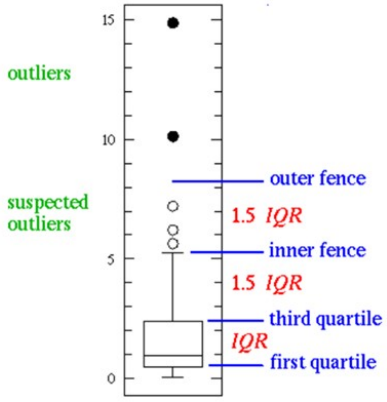


一个图里可以有多个箱子，比较很方便
可能有几条会重合；线之间不一定有值
总有50%的数据在箱子之外
max&min; upper hinge&lower hinge 3xIQR

Boxplot (箱线图)

没有异常值

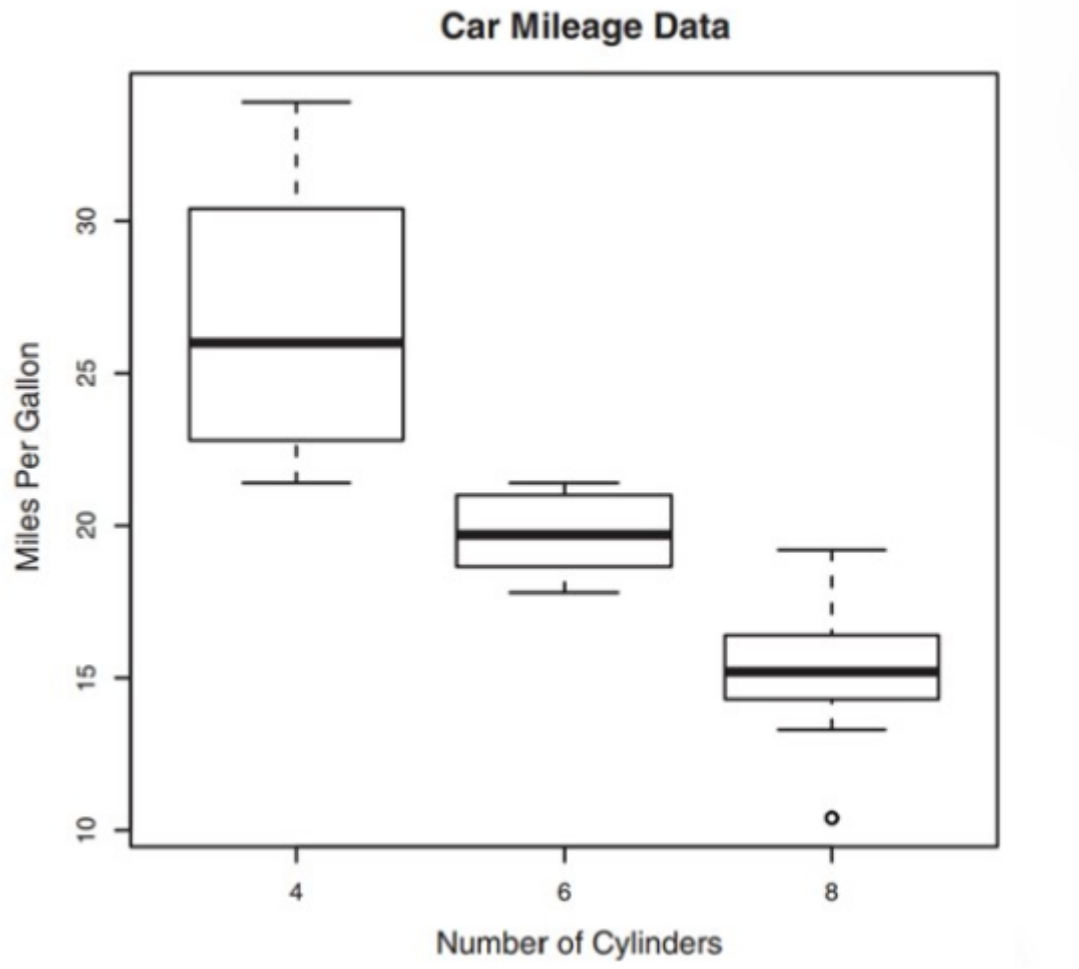




异常值(outliers): 在上四分位数之上 3× IQR 或更高，或在下四分位数以下3× IQR 或更低。

可疑异常值(suspected outliers):在上四分位数之上 1.5 × IQR 或更高，或在下四分位数以下1.5 × IQR 或更低。

inner fence-outer fence 空心圈 ○
outer fence 离群值 实心圈 ●
当然外面也可能什么都没有



3. Based on the survey above, answer the following questions:
   (a) Except "ID", how many variables in the dataset?
   13
   (b) List all numerical variables.
   Age, yesterday's bed time, credits, height, and weight.
   (c) Which are discrete variables? Which are continuous variables?
   Discrete variables include
   (d) List all categorical variables.
   (d) Is there any ordinal variable? List them.