# Introductory Statistics Using R

MAOYING WU

Department of Bioinformatics & Biostatistics

Shanghai Jiao Tong University

# 参数估计与假设检验

- 参数估计（parameter estimation）
  - 参数（parameter） vs 统计量（statistic）
  - 如何从样本统计量获取总体的参数？

- 假设检验（hypothesis testing）
  - 提出假设（$H_0$）与对立假设（$H_1$）
  - 参数方法（parametric）还是非参数方法（nonparametric）？
  - 选择假设检验方法
  - 结论

# 参数 VS 统计量

- 参数是对于总体而言的，而统计量是针对样本的；
- 参数是固定不变的（fixed），而统计量是可变的（variable）
- 参数绝大多数情况下是未知的，但是可以通过样本进行推断，这就是所谓的inferential statistics

# 判断估计量优劣的标准

- 无偏性（unbiased）
- 有效性
- 一致性（consistency）
- 充分性

# 参数估计的基本方式

- 点估计（point estimation）
- 区间（interval）估计

# 总体参数的点估计

给定总体 $X \sim N(\mu, \sigma^2)$，现有一个样本$(x_1, x_2, \ldots, x_n)$

- 总体均数: $\hat{\mu} = \overline{x}$
- 总体标准偏差: $\hat{\sigma} = S$

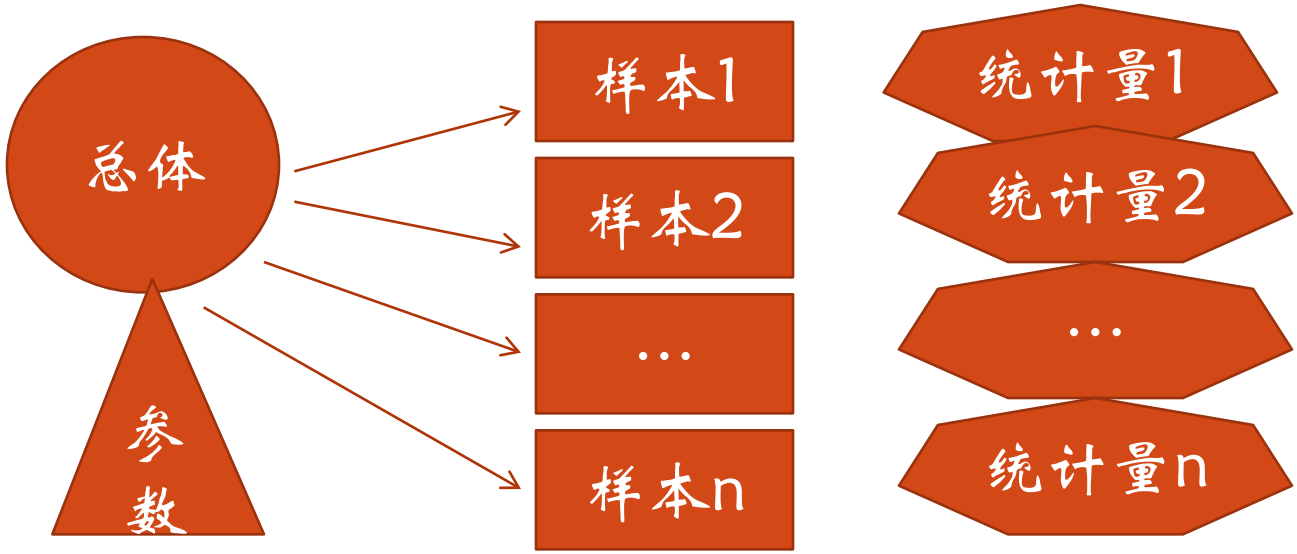
# 抽样误差和标准误差

- 抽样误差（sampling error）

  由于个体差异而导致的抽样样本统计量与总体参数之间的差别。

- 标准误差（standard error）

  多次抽样形成的多个统计量之间的差别



# 标准误差

若总体 $x \sim N(u,\sigma^2)$ 或总体分布不明但样本含量很大时，样本平均数服从或近似服从正态分布，

即：

$$\bar{x} \sim N(u,\frac{\sigma^2}{n})$$

$\overline{X}$ 的离散程度反映了抽样误差的大小

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

# 标准误差的计算

总体标准偏差一般是未知的，应用中常以样本标准偏差代替

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

# 区间估计

- 点估计值仅仅是未知参数的一个近似值，它没有反映出这个近似值的误差范围，使用起来把握不大. 区间估计正好弥补了点估计的这个缺陷。
- 简单的说，区间估计就是用一个区间去估计未知参数，把未知参数估计在某两个界限之间

- 置信区间

  按照预先给定的概率（1- α ）确定的包含未知总体参数的可能范围。它是以上下置信限（L1 , L2）为界

# 置信水平

- 置信水平或置信度

  指在区间估计中，预先选定（规定）的概率，用 1-α表示，常取95%或99%。

- 显著性水平

  在使用置信区间作估计时，被估计的参数不在该区间内的概率，用α表示，一般α取值要求较小。

# 置信区间的计算

- 明确问题：求什么参数的置信区间μ
- 设定置信水平1-α
- 确定未知参数的点估计 $\hat{\mu} = \overline{X}$
- 确定待确定参数和估计量的函数，并且其分布已知

$$U = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \qquad P\{|\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}| \leq u_{\alpha/2}\} = 1 - \alpha$$

# 置信区间的计算（续）

$$\hat{\mu} = \overline{X} \pm \frac{\sigma}{\sqrt{n}} u_{\alpha/2}$$

# Ⅰ类错误(Type I error)和Ⅱ类错误

- **Ⅰ类错误**

  $H_0$实际成立，假设检验的结果却拒绝

  常用α表示

- **Ⅱ类错误**

  $H_0$实际不成立，假设检验的结果却接受

  常用β表示

- power