TONGJI UNIVERSITY

生物统计学

biostatistics

张敬 教授

zhangjing@tongji.edu.cn

统计学基础知识

- 变异（Variation)

  遗传因素

  环境因素

  发育噪音

  个体变异—表示为条件相同的个体，各项特征仍存 在着差异。

  变异是由众多的、偶然的、次要的因素造成。

  随机测量变异—表现为同一个体多次观测结果不完全相同。

同质与变异

homogeneity and variation

- 同质——指被研究的指标的影响因素相同，包括事物的性质、影响条件或背景等相同或非常相近。
- 变异——指同质的个体之间的差异 。

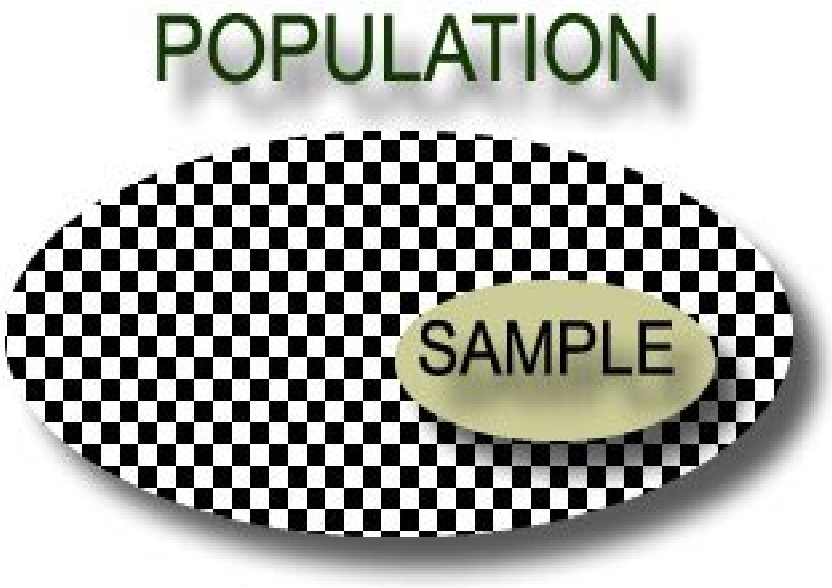
例 子

- 调查**2003**年上海市**7**岁男童的身高和体重

同质：**2003**年、上海市、**7**岁男童

变异：身高和体重各不相同

总体与样本
population and sample



总体——根据研究目的确定的同质研究对象的全体（集合）。分为有限总体与无限总体。

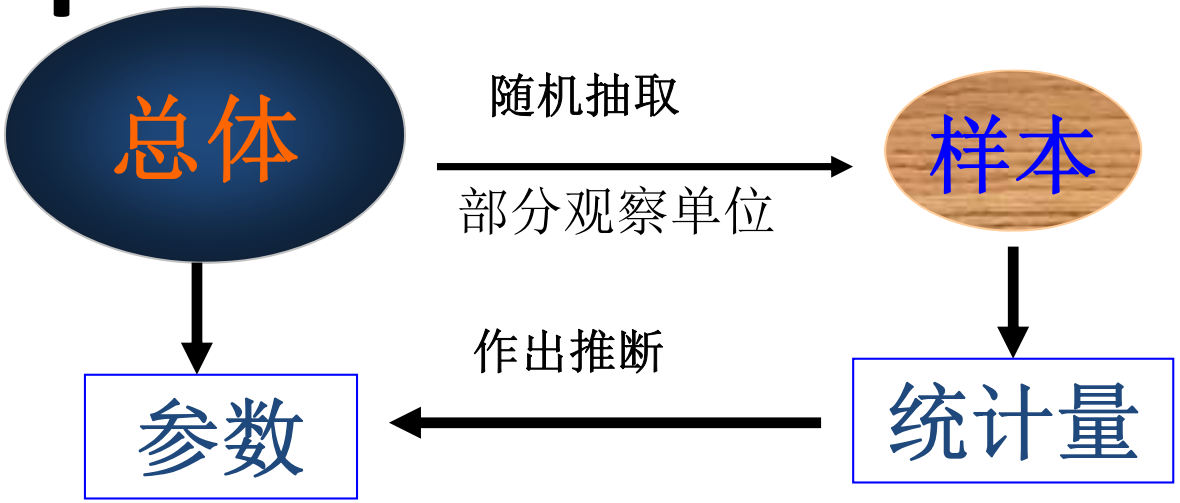
样本——从总体中随机抽取的部分观察单位。

参数（**parameter**）—根据总体的分布特征而计算的总体数值。用希腊字母表示。如：总体均数（μ）、总体率（π）、总体标准差（σ）等。

统计量（**statistic**）—从总体中随机抽取样本而计算的相应指标。用拉丁字母表示。如：样本均数（$\bar{x}$）、样本率（***p***）、样本标准差（**S**）等。

参数与统计量

parameter and statistic



参数——总体的统计指标，如总体均数、标准差，采用希腊字母分别记为 μ、σ。固定的常数。

统计量——样本的统计指标，如样本均数、标准差，采用拉丁字母分别记为 $\bar{X}$、S。参数附近波动的随机变量。

- 抽样（sampling)

  抽样（ **sample** ）—从总体中抽取部分个体的过程称抽样。

样本应具有：

代表性（**representative)**

随机性 **(randomization)**

可靠性 **(reliability)**

可比型 **(comparable)**

四种随机抽样方法

单纯随机抽样—将观察单位逐一编号，然后用随机数字表、抽签或电脑等方法随机抽取 部分观察单位组成样本。为最基本的抽样方法。

系统抽样—按一定顺序机械地每隔若干个观察单位抽取一个观察单位以组成样本。又称间隔抽样、机械抽样、等距抽样。

整群抽样—从总体中随机抽取若干个“群体”以组成样本。这个群体可以是班级、街道社区等。

分层抽样—先按影响观察值变异较大的某种特征，将总体分为若干类型或组别（统计上叫“层”），再从每一层内随机抽取一定数量的观察单位，以组成样本。也即分类抽样。

变量（variable）—观察单位的某项特征。

变量值（value of variable）—对变量的测定和观察结果。

数值变量（numerical variable）—其变量值是定量的，表现为数值的大小，一般有度量衡单位。又叫计量资料。

分类变量（categorical variable）—其变量值是定性的，表现为互不相容的类别或属性。有两种情况：

无序分类—各类间无程度上的差异，亦称计数资料。

有序分类—各类之间有程度上的差别，有“半定量”特性。亦称等级资料。

变量

编号（ID）	性别（X）	体重（kg）（Y）	疗效（Z）
张1	1	66	0
李2	1	78	1
王3	0	57	2
…	…	…	…

变量值

资料的类型

- 计量资料（度量数据，measurement data)

  单位：身高（cm）体重（kg）浓度（mg/L）

- 计数资料（计数数据，enumeration data,count data)

- 等级资料

  化验结果：- + ++ +++ ±

变量间的转化

例：一组20~40岁成年人的血压

计量资料

等级资料

<8	低血压
8~	正常血压
12~	轻度高血压
15~	中度高血压
17~	重度高血压

计数资料

以12kPa为界分为正常与异常两组，统计每组例数

三种资料的相互转换

血红蛋白

每个人的血红蛋白——计量资料 12g%

正常、异常——计数资料

分等级——等级资料

<6(g%)	重度贫血
6~ (g%)	中度贫血
9~ (g%)	轻度贫血
12.5~ (g%)	血红蛋白正常
>16 (g%)	血红蛋白增高

- 误差（error)

系统误差（system error)

方向性，可克服。

随机测量误差（random measurement error)

无方向性，可控制。

抽样误差（sampling error）—由抽样而引起的样本指标与总体指标的差异。

不可避免，样本含量控制。

四种随机抽样方法抽样误差由大到小顺序为：

整群抽样(cluster sampling) >

单纯随机抽样(simple random sampling) >

系统抽样(systematic sampling) >

分层抽样(stratified sampling)

- 概率***P*(A)**

  概率（probability）—事件发生可能性大小的度量。

  ***P*=f/N　f—发生数或频数 N---观察总数**

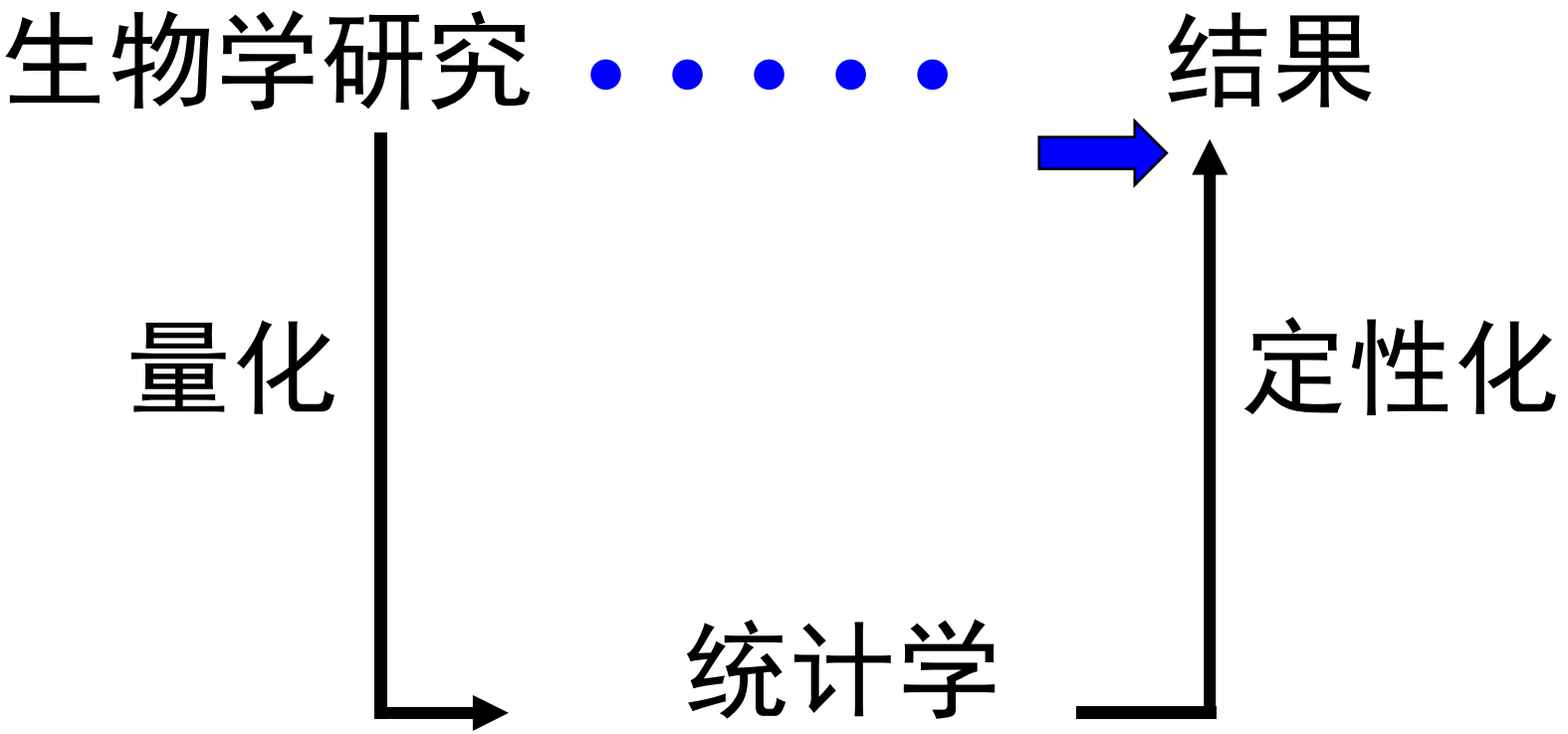
  ***P*=1**　必然事件

  ***P*=0**　不可能事件

  ***P*≤0.05**　小概率事件　差别有统计意义

  ***P* ≤0.01**　小概率事件　差别有高度统计意义

统计学与生物学研究



统计工作流程

收集资料

↓

资料整理、制表

↓

初步统计分析
（统计描述）

- 计量资料
  - 集中趋势指标
  - 离散趋势指标
- 计数资料：率等

↓

进一步统计分析
（统计推断）

- 计量资料
  - 1.正常值范围、标准误、可信区间
  - 2.差别的显著性检验—u、t 、F和非参数统计等
  - 3.相关与回归分析
- 计数资料
  - 1.率的标准误、总体率的可信区间
  - 2.差别的显著性检验—u、χ^2检验

↓

统计结果的表达，
回答和解决实际问题

资料收集

- 原始资料（raw data) 准确的、完整的、充满信息的
- 质量控制
  - 统一性
  - 确切性
  - 可重复性

精度（precision)

偏度 (bias)

整理资料

科学加工 → 系统化、合理化 → 统计分析

注意：

资料的逻辑检查

(1)从专业的角度对资料的合理性进行检查

（2从专业的角度对资料的一致性进行检查

检查报表纵向、横向的合计和总的合计

统计归纳

组段（d）	人数	组段（d）	人数
0～	6	35～	1
5～	21	40～	1
10～	14	45～	5
15～	13	50～	1
20～	6	55～	3
25～	2	60～	3
30～	3	合　计	79

制频数分布表（frequency distribution table)

- 表1-1 每10名新生儿中体重超过3kg的人数的频数（率）表

组值	频数计算	频数	频率
0		0	0.000
1		0	0.000
2		0	0.000
3	—	1	0.008
4	T	2	0.017
5	正正T	12	0.100
6	正正正	19	0.158
7	正正正正正正正	39	0.325
8	正正正正正正	34	0.283
9	正正	10	0.083
10	F	3	0.025
总计		120	0.999

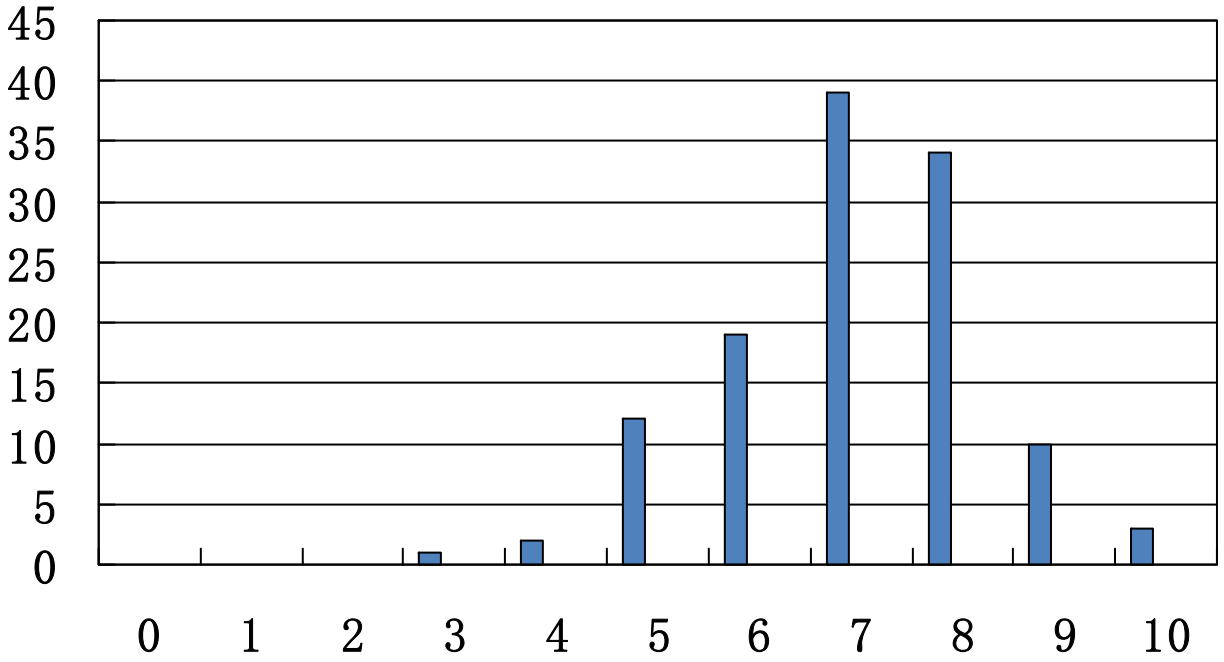


图1-1 频数图

分析资料

- 统计描述
- 统计推断
- 阐明事物规律性

Thank You

同济大学生命科学与技术学院