数理统计概述

- 统计学是与“生活”及“工作”有密切关系的一门学科.如果能够掌握统计学的知识，那么你的生活将会变得更加方便.
  - 预测食堂今天可以卖出多少个馒头
  - 预测本学期《概率统计》考试能否通过
  - 比较服用药剂X和不服用药剂X两种情况下的治愈率
  - ……

概率论与数理统计的关系

- 概率论是数理统计的理论基础；
  数理统计是概率论的重要应用.
- 概率论：在(总体/R.V.) X 分布已知的情况下，研究 X 的性质及统计规律性.
- 数理统计：在(总体/R.V.) X 分布未知 (或部分未知) 的情况下，对总体 X 的分布作出推断和预测.

数理统计学

数理统计学是一门应用性很强的学科。

它使用概率论和数学的方法，研究怎样收集（通过试验或观察）带有随机误差的数据，并在设定的模型（称为统计模型）之下，对这种数据进行分析（称为统计分析），以便对研究对象的性质、特点作出推断（称为统计推断）。

数理统计研究的内容

1. 有效地收集数据
   - 收集数据的方法有: 全面观察(或普查)、抽样调查和安排试验等方式

   ① 抽样方法：简单随机抽样、分层抽样、等距抽样、整群抽样、分阶段抽样……

   ② 试验设计与分析

   - 一个重要问题： 数据必须具有随机性!
2. 有效的使用数据
   - 建立一个统计模型来描述数据
   - 制定评判不同统计推断方法优劣的准则
   - 参数估计、假设检验、方差分析、回归分析

数理统计的分类

- 描述（descriptive）统计学 § 6：研究如何取得反映客观现象的数据，并通过图表形式或数学方法对所搜集的数据进行加工处理和显示，进而通过综合概括和分析对数据的分布状态、数字特征和随机变量之间关系进行估计和描述的方法。
  - 描述统计分为集中趋势分析、离中趋势分析和相关分析三大部分
- 推断（inferential）统计学：研究如何根据样本数据去推断总体数量特征的方法，它是在对样本数据进行描述的基础上，对统计总体的未知数量特征做出以统计形式表述的推断。
  - 统计推断的基本问题分为参数估计 § 7和假设检验 § 8两大类方法
    - 参数估计：用样本统计量去估计总体的未知参数。
    - 假设检验：先对总体参数提出一个假设值，然后利用样本信息判断这一假设是否成立。

数理统计举例

【例】某工厂生产搭配的电子元件，假定元件的寿命服从指数分布。实际应用中，感兴趣的问题如：

（1）元件的平均寿命如何？　　　　**参数估计**

（2）使用单位要求平均寿命能达到某个指定的数 l，例如5000小时，问这批元件可否被接受？

- 统计模型 —— 元件的寿命服从指数分布（参数 λ 未知）
- 如何抽样？（随机/等概率）
- 用样本的算术平均值来估计未知的平均寿命，误差可能有多大？产生指定大小的误差的概率有多大？为了降低**误差概率**，抽出的元件个数n至少应达到多少？

数理统计举例

【例】某工厂生产搭配的电子元件，假定元件的寿命服从指数分布。实际应用中，感兴趣的问题如：

（1）元件的平均寿命如何?

（2）使用单位要求平均寿命能达到某个指定的数 l，例如5000小时，问这批元件可否被接受? **假设检验**

- 怎么确定 l？ l 定得越高，检验越严格
- 有一定概率出现两种错误：

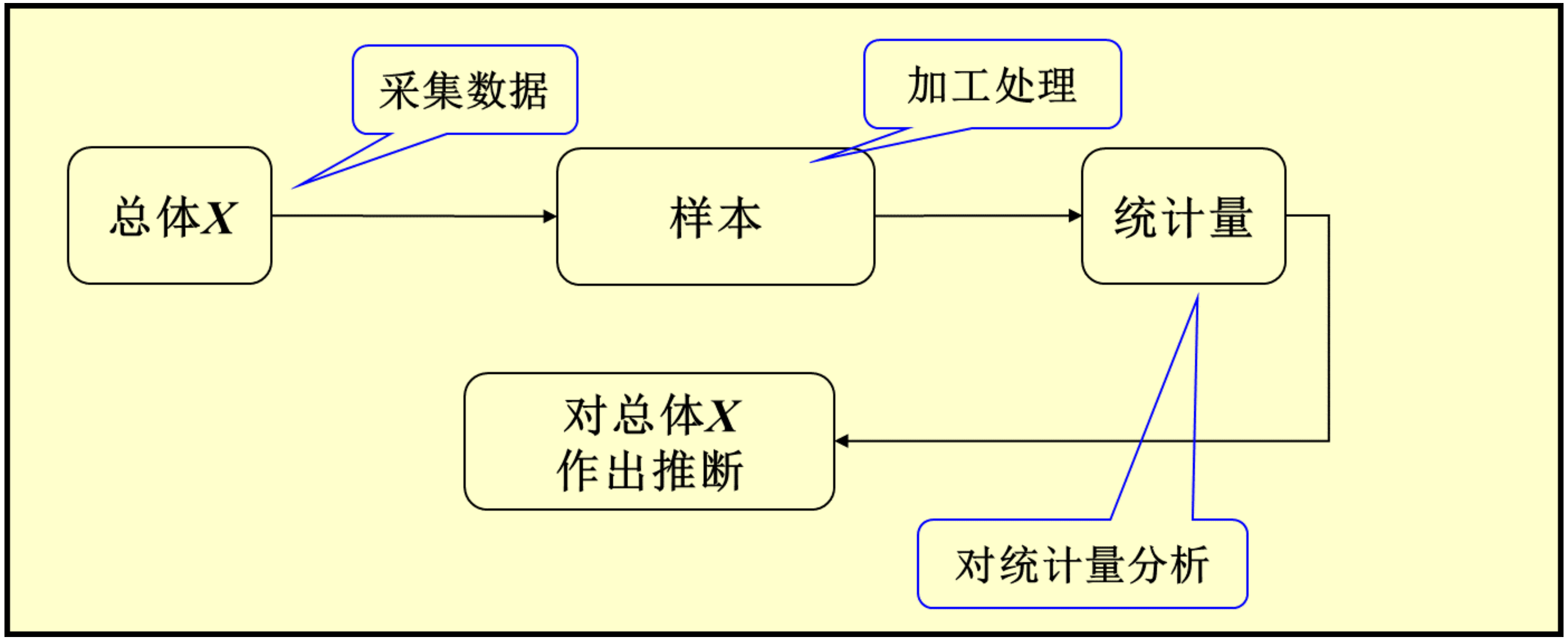
  ① 元件平均寿命达到要求却被**拒绝**；

  ② 元件平均寿命达不到要求却被**接受**

数理统计方法的归纳性质

- 统计方法的本质是归纳式的，而数学则是演绎式的。
- 统计方法的归纳性质，源于它在作结论时是根据所观察到的大量的“个别”情况，“归纳”起来所得。而不是从一些假设、命题或已知事实出发按一定的逻辑推理得出来的(这后者称为演绎推理)
- 归纳推理是要冒风险的. 事实上归纳推理的不确定性的出现，是一种逻辑的必然. 人们不可能做出十分肯定的结论, 因为归纳推理所依据的数据具有随机性.
- 统计学的作用之一就是提供归纳推理和计算不确定性程度的方法.

数理统计研究方法的流程



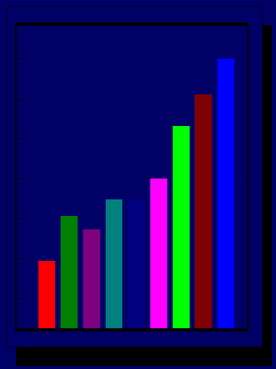
第一节　随机样本

- 总体和样本
- 小结

由于大量随机现象必然呈现出规律性，只要对随机现象进行足够多次观察，被研究的规律性一定能清楚地呈现出来.

逐一观察每个被研究对象，要么根本不可能，要么缺乏可操作性

- **某国14岁男孩的平均身高**
- **大西洋中鳕鱼的平均重量**
- **某工厂生产的一批灯泡的平均寿命**
- **观看某一电视节目的观众人数**
- ……

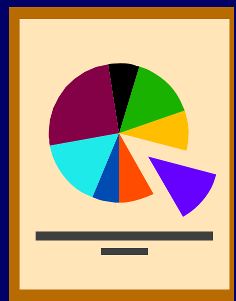


客观上， 往往只允许我们对研究对象进行次数不多的观察试验.

数理统计的任务就是研究有效地收集、整理、分析所获得的有限的资料，对所研究的问题，尽可能地作出精确而可靠的结论.

在数理统计中，不是对所研究的对象全体 (称为总体)进行观察，而是抽取其中的部分(称为样本)进行观察获得数据（抽样），并通过这些数据对总体进行推断.

数理统计方法具有“部分推断整体”的特征 .



一、总体和样本

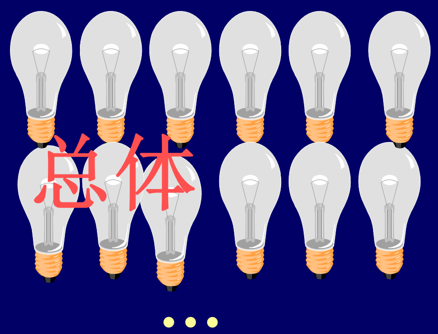
1.总体

一个统计问题总有它明确的研究对象.

研究对象的全体称为总体,

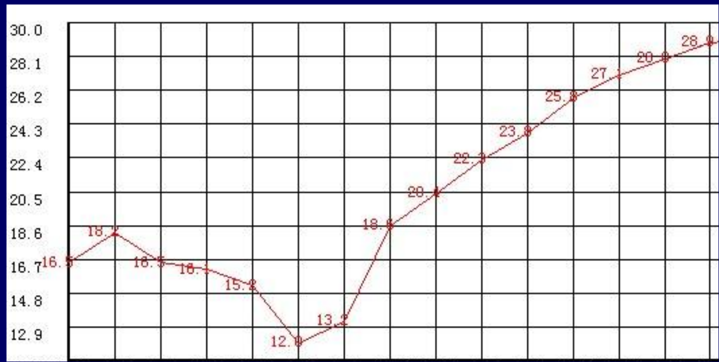
总体中每个成员称为个体,

总体中所包含的个体的个数称为总体的容量.



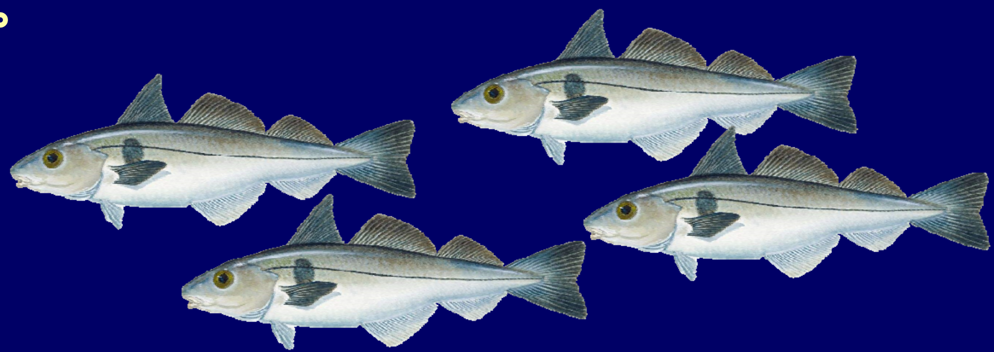
研究某批灯泡的质量

总体
- 有限总体
- 无限总体



观察并记录某一地点每天的最高气温

有些有限总体，容量很大，也可以把它当成无限总体来处理。



考察大西洋中鳕鱼的重量

我们关心的是总体中的个体的某项指标(如人的身高、灯泡的寿命,汽车的耗油量…）.

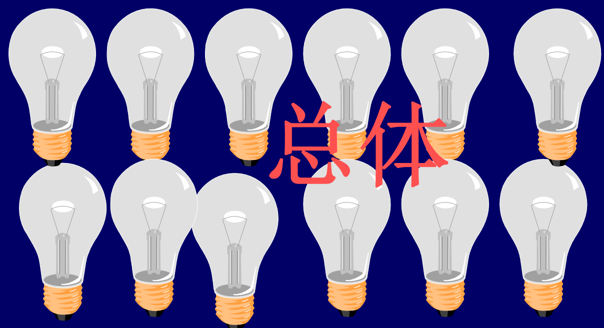
- 该数量指标的所有可能取值就是总体；该数量指标的每个观察值就是个体。

由于每个个体的出现是随机的，所以相应的数量指标的出现也带有随机性．从而可以把这种数量指标看作一个随机变量**X**，因此随机变量**X**的分布就是该数量指标在总体中的分布.

总体就可以用一个随机变量及其分布来描述.

因此在理论上可以把总体与概率分布等同起来.

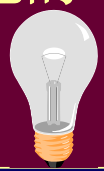
例如:研究某批灯泡的寿命时，关心的数量指标就是寿命，那么，此总体就可以用随机变量**X**表示，或用其分布函数$F(x)$表示.



寿命 X 可用一概率（指数）分布来刻划

寿命总体是指数分布总体

某批
灯泡的寿命



鉴于此，常用随机变量的记号或用其分布函数表示总体. 如说总体**X**或总体$F(x)$.

类似地，在研究某地区中学生的营养状况时，若关心的数量指标是身高和体重，我们用X 和Y 分别表示身高和体重，那么此总体就可用二维随机变量(X,Y)或其联合分布函数 $F(x,y)$来表示.

统计中，总体这个概念的要旨是：总体就是一个概率分布.

2. 样本

总体分布一般是未知，或只知道是包含未知参数的分布，为推断总体分布及各种特征，按一定规则从总体中抽取若干个体进行观察试验，以获得有关总体的信息，这一抽取过程称为“抽样”，所抽取的部分个体称为样本. 样本中所包含的个体数目称为样本容量.

从国产轿车中抽5辆
进行耗油量试验
样本容量为5

抽到哪5辆是随机的

对总体X在相同的条件下，进行n次重复、独立观察，其结果依次记为X_1，X_2,…，X_n.

这样得到的随机变量$X_1, X_2, \cdots X_n$是来自总体X的一个简单随机样本，与总体随机变量具有相同的分布. n称为这个样本的容量.

一旦取定一组样本X_1，… ,X_n ,得到n个具体的数$(x_1,x_2,\ldots,x_n)$，称为样本的一次观察值，简称样本值 .

最常用的一种抽样叫作“简单随机抽样”，其特点：

1. 代表性：$X_1,X_2,\ldots,X_n$中每一个与所考察的总体有相同的分布.

2. 独立性：$X_1,X_2,\ldots,X_n$是相互独立的随机变量.

【定义】

设X是具有分布函数F的随机变量，若$X_1, X_2, \cdots, X_n$是具有同一分布函数F的、相互独立的随机变量，则称$X_1, X_2, \cdots, X_n$为从分布函数F(或总体F、或总体X）得到的容量为n的简单随机样本，简称样本，它们的观察值$x_1, x_2, \cdots, x_n$称为样本值，又称为X的n个独立的观察值.

由简单随机抽样得到的样本称为简单随机样本，它可以用与总体独立同分布的n个相互独立的随机变量$X_1, X_2, \ldots, X_n$表示.

样本具有双重性

观察前：$X_1, X_2, \ldots, X_n$是相互独立，与总体同分布的随机变量.

观察后：样本值$x_1, x_2, \ldots, x_n$为n个具体的观察数据.

若总体的分布函数为$F(x)$、概率密度函数为$f(x)$,则其简单随机样本的联合分布函数为

$$F^*(x, x_2, \cdots, x_n) = F(x_1)\,F(x_2)\,\ldots\,F(x_n)$$

其简单随机样本的联合概率密度函数为

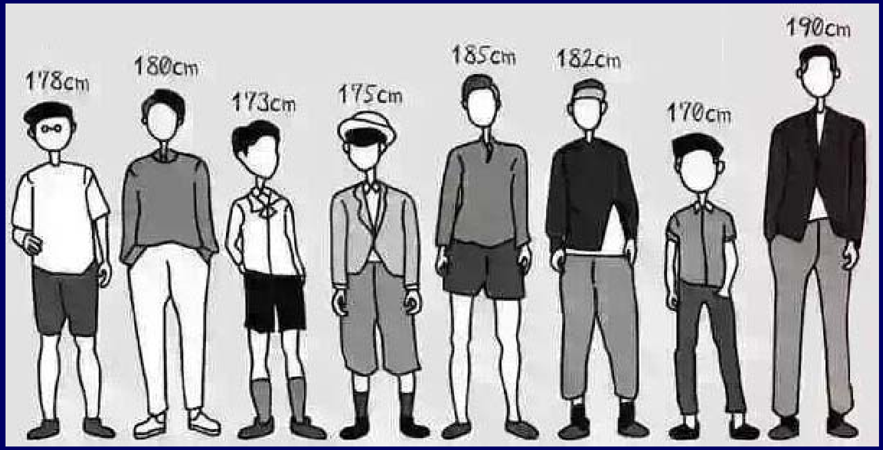
$$f^*(x, x_2, \cdots, x_n) = f(x_1)\,f(x_2)\,\ldots\,f(x_n)$$

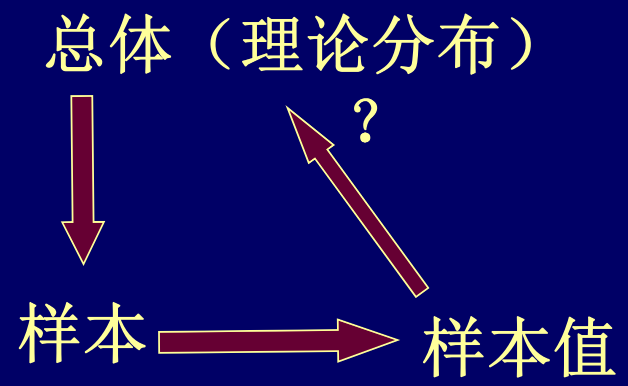
简单随机样本是应用中最常见的情形，今后，当说到“$X_1, X_2, \ldots, X_n$是取自某总体的样本”时，若不特别说明，就指简单随机样本.

- 对于有限总体，采用放回抽样，就能得到简单随机样本。
  - 当个体的总数N比样本容量n大得多时，实际中可将不放回抽样近似地当做放回抽样来处理
- 对于无限总体，总是用不放回抽样。

3. 总体、样本、样本值的关系

事实上我们抽样后得到的资料都是具体的、确定的值. 如我们从某班大学生中抽取10人测量身高, 得到10个数，它们是样本取到的值而不是样本. 我们只能观察到随机变量取的值而见不到随机变量.





统计是从手中已有的资料--样本值，去推断总体的情况---总体分布$F(x)$的性质.

样本是联系二者的桥梁

总体分布决定了样本取值的概率规律，也就是样本取到样本值的规律，因而可以由样本值去推断总体.

二、小结

研究对象的全体称为总体

总体中每个成员称为个体

设X是具有分布函数F的随机变量，若$X_1, X_2,$ $\cdots, X_n$是具有同一分布函数F的、相互独立的随机变量，则称$X_1, X_2, \cdots, X_n$为从分布函数F(或总体F、或总体X）得到的容量为n的简单随机样本.

简称样本.