概率论与数理统计

第六章: 数理统计的基本概率与抽样分布

李立颖 (lily@sustech.edu.cn)

南方科技大学数学系

2023 秋季

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

数理统计分为三大部分:

- 抽样分布
- 参数估计
- 假设检验

数理统计学是一门以数据为基础的学科.
数理统计学的任务就是如何获得样本和利用样本, 从而对事物的某些未知方面进行分析、推断并作出一定的决策.

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应用.

数理统计概论

概率论是在 (总体) X **分布已知** 的情况下, 研究 X 的性质及统计规律性.
数理统计是在 (总体) X **分布未知** (或部分未知) 的情况下, 对总体 X 的分布作出的推断和预测.

数理统计的研究方法

通过从总体抽取部分个体 (样本), 通过对样本的研究, 对总体作出推断或预测. 一种**由部分推测整体** 的方法.

什么叫数理统计

数理统计是以概率论为理论基础, 关于实验数据的收集、整理、分析、推断的一门数学学科.

实际背景

- 某工厂生产了一大批产品, 从中随机抽检了 n 件产品, 发现有 m 件次品. 如何估计整批产品的次品率 p?
- 要求某种元件的平均使用寿命不得低于 10000 小时. 现从这批元件中随机抽取 25 件, 测得其寿命的平均值为 9500 小时. 试问这批元件是否达到了要求?

统计推断的基本内容

点估计、区间估计、参数假设检验、非参数假设检验、方差分析、回归分析.

什么是实验数据

科学实验, 或对某事件、现象进行观察获得的数据称为 **实验数据**.
特点: 数据受随机因素的影响.

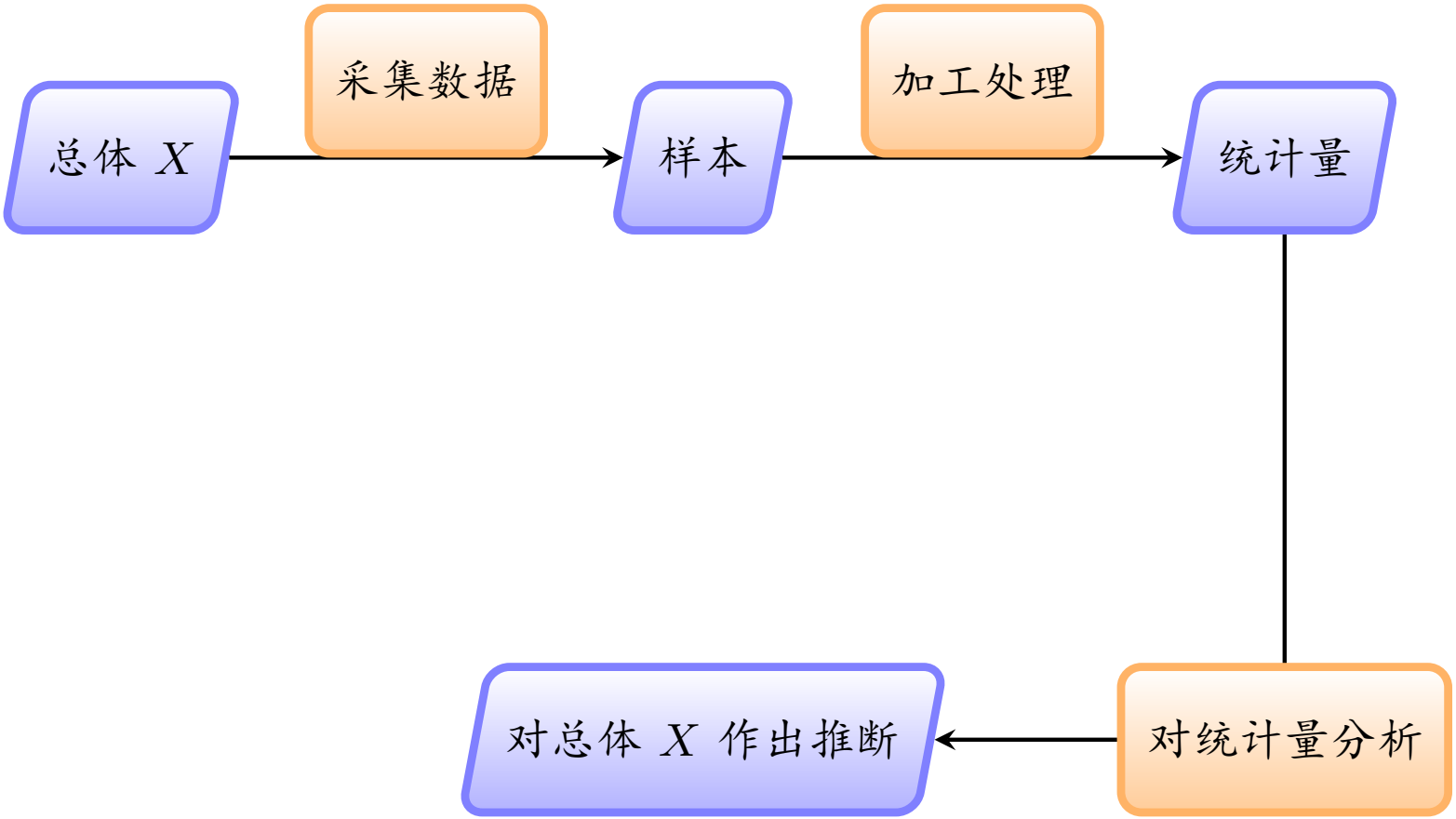
处理实验数据的过程?

收集、整理、分析、推断. (我们将围绕这四个过程来研究)

如何收集和整理数据?

本章将研究 ‘‘收集’’ 和 ‘‘整理’’ 数据的科学含义.

数理统计研究方法流程图



数理统计的核心问题是由样本推断总体, 即**统计推断问题**.

总体与个体

定义

总体 研究对象的全体称为 **总体 (母体)**.

个体 组成总体的每个研究对象称为个体.

概率模型

由于个体的出现是随机的, 所以总体是一个随机变量, 我们用 X 表示. 总体分为有限总体和无限总体.

例

我们研究某批灯泡的寿命, 则**该批灯泡寿命的全体就是总体.**

数量指标

在研究中, 往往关心每个个体的一个 (或) 几个**数量指标**和该数量指标在总体中的分布情况. 这时, 每个个体具有的**数量指标的全体**就是总体.

总体 研究对象的数量指标 $X \sim F(x)$

个体 随机变量 X 的值.

例子

- 考察某班级学生的英语课程学习成绩 X. 因为每个学生的成绩都在全班平均成绩 μ 的附近波动, 所以总体可以视为 $X \sim \mathcal{N}(\mu, \sigma^2)$.
- 考察某工厂生产的某批灯泡的寿命 X. 因为每个灯泡的寿命都在该批灯泡平均寿命 μ 的附近波动, 所以总体可视为 $X \sim \mathcal{N}(\mu, \sigma^2)$.
- 考察某工厂生产和零件是否合格, 记

$$X = \begin{cases} 0, & \text{零件合格}, \\ 1, & \text{零件不合格}. \end{cases}$$

则总体可视为 $X \sim \mathrm{Bin}(1, p)$, p 为零件的次品率.

收集数据

抽样

从研究对象中任取 n 个 "个体", 观察它们的数量指标 $X_1, X_2, \ldots, X_n$, 这一过程称为 **抽样**, $X_1, \ldots, X_n$ 称为**容量** 为 n 的 **样本**.

抽样的特点:

在相同条件下对总体 X 进行 n 次重复、独立观察:

独立性 要求各次抽样的结果互不影响

代表性 每次取出的样品与总体有相同的分布.

样本的二重性

- 观察前: $X_1, \ldots, X_n$ 是相互独立, 与总体同分布的随机变量.
- 观察后: 样本观察值 $X_1, \ldots, X_n$ 是 n 个独立的观察数据.

简单随机抽样

定义

简单随机抽样要求抽取的样本满足下面两点:

1. 代表性: $X_1, \ldots, X_n$ 中每一个与总体 X 有相同的分布.
2. 独立性: $X_1, \ldots, X_n$ 是相互独立的随机变量.

简单随机抽样即为随机地独立地抽取, 如有放回抽样. 无放回抽样当总体很大, 样本容量很小时, 认为是近似的简单随机抽样.

简单随机样本

由简单随机抽样抽得的样本 $X_1, \ldots, X_n$ 称为**简单随机样本 (样本)**.

概率论解释

显然, 样本就是来自总体 X 的 n 个相互独立的且与总体同分布的随机变量 $X_1, \ldots, X_n$. 可看成 n 维随机向量 $(X_1, \ldots, X_n)$. 若总体的分布函数为 $F(x)$, 则简单随机抽样的联合分布函数为

$$F(x_1) \cdot \cdots \cdot F(x_n).$$

非简单随机抽样的例子: 购物 app 数据跟踪, 时间序列观测, Alpha Go 决策树, ...

连续型总体

例子

某厂生产了**一大批**灯泡. 现从中随机抽取 5 只进行检测, 测得其寿命 (小时) 分别为

$$983, 969, 1037, 1302, 852$$

- 总体为灯泡的寿命: $X \sim \mathcal{N}(\mu, \sigma^2)$
- 样本容量为 5, 样本为 X_1, X_2, X_3, X_4, X_5
- 样本观察值为 983, 969, 1037, 1302, 852 (样本二重性)

例子

对长度为 μ 的工件进行了 6 次测量, 测量值为

$$29.1, 30.2, 29.3, 29.1, 30.3, 29.5.$$

- 总体为工件长度 $X \sim \mathcal{N}(\mu, \sigma^2)$.
- 样本容量为 6, 样本为 $X_1, X_2, \ldots, X_6$.
- 样本观察值为 29.1, 30.2, 29.3, 29.1, 30.3, 29.5.

离散型总体

例子

考察某工厂生产的零件是否合格, 从该厂生产的一批产品中随机抽检了 100 个, 若合格则记为 0, 若不合格则记为 1, 100 个产品的检查结果为 $x_1, x_2, \ldots, x_{100}$.

- 总体: $X \sim \mathrm{Bin}(1, p)$ **零件合格或不合格**
- 总体频率函数为 $\mathsf{P}(X = 1) = p$, $\mathsf{P}(X = 0) = 1 - p$, 其中 p 为零件的次品率.

样本二重性

- 样本 $X_1, X_2, \ldots, X_{100}$: 独立同分布 0-1 分布随机变量
- 样本观察值 $x_1, x_2, \ldots, x_{100}$.

对样本的一些认识

设 $X_1, X_2, \ldots, X_n$ 是来自于总体 $X \sim F(x)$ 的样本.

- $X_1, X_2, \ldots, X_n$ 是一堆 ‘‘杂乱无章" 的数据.
- $X_1, X_2, \ldots, X_n$ 包含了有关总体的 ‘‘信息".
- $X_1, X_2, \ldots, X_n$ 是对总体进行推断的依据
- 在观察前 $X_1, \ldots, X_n$ 是一组独立同分布的随机变量,
  在观察后 $x_1, \ldots, x_n$ 是一组具体的数据.

简单随机样本的分布

问: 样本 $X_1, \ldots, X_n$ 作为多维随机变量, 服从什么分布?

若总体分布函数为 $F(x)$, 则样本联合**分布函数**为

$$F^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i).$$

若总体密度函数为 $f(x)$, 则样本联合**密度函数**为

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i).$$

例子: n 维正态分布

设 $X_1, \ldots, X_n$ 为来自于总体 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的样本, 则样本的分布密度函数为

$$f^*(x_1, \cdots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

例子

设 $X_1, X_2, \ldots, X_n$ 为来自总体 $X \sim \mathrm{Bin}(1, p)$, $p \in (0, 1)$ 的样本, 求样本分布.

解

总体的频率函数为

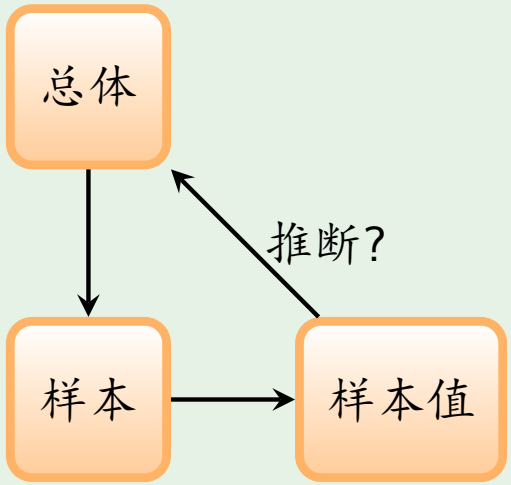
$$\mathsf{P}(X = 1) = p, \quad \mathsf{P}(X = 0) = 1 - p.$$

样本 $X_1, X_2, \ldots, X_n$ 是独立同服从 0-1 分布的随机变量, 因此样本分布为

$$\mathsf{P}(X_1 = x_1, \cdots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}.$$

总体、样本和样本值的关系

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**.

例子

生产厂家声称他们生产的灯泡平均寿命不低于 6000 小时, 如何验证厂家说法的真伪? 由于灯泡寿命试验是破坏性试验, 不可能把整批灯泡逐一检测, 只能抽取一部分灯泡进行检验, 通过这部分灯泡的寿命数据来推断整批灯泡的平均寿命. **以部分数据信息来推断整体未知参数**, 就是整理统计研究问题的基本方式.

为什么用样本的特性去估计总体的特性?

在工农业生产和科学研究领域里, 将研究对象全体进行鉴定是不可能的.

- 在许多情况下, 总体包含的个体数很多;
- 有时从总体中抽取个体是破坏性的试验, 在这种情况下, 不允许逐个抽取, 并且抽取的数量不可能太多, 而样本是总体的一部分, 它的特性在某种程度上能反映总体的特性, 所以需要用样本的特性去估计总体的相应特性.

实例

了解某校大学生 “做过/正在做家教” 的比例

总体是该校大学生全体. 这是一个**有限总体**, 每个大学生有许多指标, 比如性别、年龄、身高、体重、高考成绩等. 现在我们只关心的是学生 “是否做过家教” 这一指标

了解某城市的空气质量情况及 PM2.5 值

总体是城市上空一定范围内的空气, 这是一个**无限总体**. 描述空气质量有许多指标, 而我们公关心 PM2.5 值.

药厂研究某种药物在人体中的吸收情况

总体是全体国民, 这是一个**有限总体**, 但数量非常巨大. 我们可以把它看成无限总体.

抽样与概率论关系

- 总体的某个指标 X, 对于不同的个体来说有不同的取值, 这些取值构成一个分布, 因此 X 可以看成一个**随机变量**.
- 有时候直接将 X 称为总体. 假设 X 的分布函数为 $F(x)$, 也称总体 X 具有分布 $F(x)$.

例

了解某校学生 "做过家教" 的情况, 对每个学生来说, 以 $\{X=1\}$ 表示 "做过家教", 以 $\{X=0\}$ 表示 "未做过家教". 则总体为

$$X = \mathrm{Bin}(1, p),$$

其中 p 是全体学生中做过家教所占的比例, 是一个未知量. 但我们可以写出 X 的分布

$$\mathsf{P}(X=x) = p^x(1-p)^{1-x}, \quad x=0,1.$$

例题

四个同学参加了《概统》课程考试, 成绩分别为 88, 75, 70, 63. 以 X 表示这四人的成绩.

1. 写出总体 X 的分布律, 数学期望和方差
2. 从总体中抽取容量为 2 的样本, 列出全部的样本值.

解

X	88	75	70	63
p	1/4	1/4	1/4	1/4

$$\mathbb{E}X = (88+75+70+63)/4 = 74,$$

$$\mathrm{Var}(X) = 83.5 = \frac{(88-74)^2+(75-74)^2+(70-74)^2+(63-74)^2}{4}.$$

样本 (X_1, X_2) 的取值 (x_1, x_2) 有 16 个样本值, $\mathsf{P}(X_1 = x_1,\ X_2 = x_2) = \dfrac{1}{16}$. 全体取值为

$$\{88, 75, 70, 63\} \times \{88, 75, 70, 63\}.$$

例 1

为了解某区八年级学生的身高，有关部门从八年级中抽 200 名学生测量他们的身高，然后根据这一部分学生的身高去估计此区所有八年级学生的平均身高. 说出总体、个体、样本和样本容量.

例 2

要了解一片水稻田里所有单株水稻的产量情况，从中抽取 500 株水稻单株产量去估计这片田里所有水稻的单株产量. 说出总体、个体、样本和样本容量.

例 3

为了考察某商店一年中每天的营业额，从中抽查了 30 天的每天营业额. 说出总体、个体、样本和样本容量.

例 4

为了了解参加某运动会的 2000 名运动员的年龄情况，从中抽取了 100 名运动员的年龄. 说出总体、个体、样本和样本容量.

例 5

为了解初三年级 400 名学生的身高情况, 从中抽取 40 名学生进行测量, 这 40 名学生的身高是

A. 总体的一个样本;

B. 个体;

C. 总体;

D. 样本容量.

A.

例 6

为了解我省中考数学考试的情况, 抽取 2000 名考生的数学试卷进行分析, 2000 叫做:

A. 个体;

B. 样本;

C. 样本容量;

D. 总体.

C.

例 7

为了考察某班学生的身高情况, 从中抽取 20 名学生进行身高测算, 下列说法正确的是

A. 这个班级的学生是总体;

B. 抽测的 20 名学生是样本;

C. 抽测的 20 名学生的身高的全体就是总体;

D. 样本容量是 20.

D.

为了解 1000 台新型电风扇的寿命, 从中抽取 10 作连续运转实验, 在这个问题中, 下列说法正确的是

A. 1000 台风扇是总体;

B. 每台风扇是个体;

C. 抽取的 10 台风扇是样本容量;

D. 抽取的 10 台风扇的使用寿命是样本.

D.

小结

总体、个体和样本容量

一般地，我们把要考察的对象的全体叫做**总体**，其中**每一个考察对象**叫做个体，从总体中被抽取的考察对象的集体叫做总体的**一个样本**，样本中**个体的数目**叫做样本容量.

- 总体和样本是相对而言的.
- 样本的特性反映了总体的相应特性.

统计推断的基础: 收集数据

从总体 $X \sim F(x)$ 抽取样本 $X_1, X_2, \ldots, X_n$.
怎样从“杂乱无章”的数据中提炼出有用的信息?

数据的加工整理: 统计量

设 $X_1, X_2, \ldots, X_n$ 为来自总体 $X \sim F(x)$ 的样本, $g(x_1, \ldots, x_n)$ 为 n 元函数. 若随机变量 $g(X_1, X_2, \ldots, X_n)$ 不含任何未知参数, 则称 $g(x_1, X_2, \ldots, X_n)$ 为**统计量**.

统计量的二重性

- 试验前: $g(X_1, X_2, \ldots, X_n)$ 是随机变量.
- 试验后: $g(x_1, x_2, \ldots, x_n)$ 是具体数值.

例子

某班级《高等数学》课程考试成绩单列出 n 个学生的成绩分别为 $X_1, X_2, \ldots, X_n$. 下面哪一个量能更好评价全班整体学习情况?

$$\frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \max_{1\le i\le n} X_i, \qquad \min_{1\le i\le n} X_i.$$

常见统计量

样本均值 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$

(它们与均值和方差有什么不同?)

样本标准差 $S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

样本 k 阶矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k,\ k = 1, 2, \ldots$

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k,\ k = 1, 2, \ldots$

常见统计量 II

将样本 $X_1, X_2, \ldots, X_n$ 从小到大排列为 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.

顺序统计量 $(X_{(1)}, \ldots, X_{(n)})$.

极小值 $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$.

极大值 $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$.

Khinchin (辛钦) 大数定律

设 $\{X_n\}$ 独立同分布, $\mathbb{E}X_1 = \mu$ 存在, 则 $\{X_n\}$ 服从弱大数律, 即 $\forall \varepsilon > 0$, 有

$$\lim_{n\to\infty} \mathsf{P}\Big(\Big|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\Big| < \varepsilon\Big) = 1.$$

样本矩的特性

设 $X_1, X_2, \ldots, X_n$ 为来自总全 $X \sim F(x)$ 的样本, 总体 k 阶矩 $\mu_k = \mathbb{E}X^k$, $k = 1, 2, \ldots$ 都存在.

- $X_1, \ldots, X_n$ 独立且与总体同分布 $\Rightarrow X_1^k, \ldots, X_n^k$ 独立且与 X^k 同分布
- $\mathbb{E}A_k = \dfrac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i^k = \dfrac{1}{n}\sum_{i=1}^{n} \mathbb{E}X^k = \mu_k.$
- $A_k = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{\mathsf{P}} \mu_k$ 推出 $g(A_1, \ldots, A_k) \xrightarrow{\mathsf{P}} g(\mu_1, \mu_2, \ldots, \mu_k)$, $n \to \infty$.