概率论与数理统计

第二十九讲
数理统计的基本概念

从宿舍到教室需要花多少时间?

相信大家心里对此都有一个大概的“数”.

- **问题**　你是怎么得到这个“数”的?

这就是一个典型的统计思维过程

数据 → 归纳 → 结果

数理统计就是一个归纳推断过程

第29讲　数理统计的基本概念

数理统计是以概率论为基础，关于实验数据的收集、整理、分析与推断的一门科学与艺术

- **问题** 什么是实验数据?

  科学试验,或对某事物、现象进行观察获得的数据称为试验数据

- **特点** 数据受随机因素的影响

  --可以通过某种概率分布来描述

第29讲　数理统计的基本概念

- **问题**　实验数据的处理过程?

数据 收集，整理，分析，推断

《数理统计》围绕这四个过程来进行研究

本讲主要介绍数据“收集”和“整理”环节中的一些相关概念

第29讲　数理统计的基本概念

- **总体**　研究对象的全体称为总体
- **个体**　总体中的每一个具体对象称为个体

例　分析某班级学生的英语考试成绩

总体 -- 该班级所有学生的英语考试成绩

个体 -- 每一个学生的英语考试成绩

例 分析某工厂生产的灯泡的使用寿命

总体 -- 该厂生产的所有灯泡的使用寿命

个体 -- 每一个灯泡的使用寿命

总体　研究对象的数量指标 X

$$X \sim F(x)$$

个体　总体 X 的可能取值

例　分析某工厂生产的灯泡的使用寿命

总体　-- 该厂生产的所有灯泡的使用寿命 X

$$X \sim N(\mu,\ \sigma^2)$$

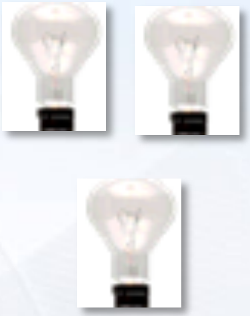
个体　-- 每一个灯泡的使用寿命，即 X 的一个可能取值

第29讲　数理统计的基本概念

- **问题** 如果对总体完全了解的情况下，能否对个体进行预测?

第29讲　数理统计的基本概念

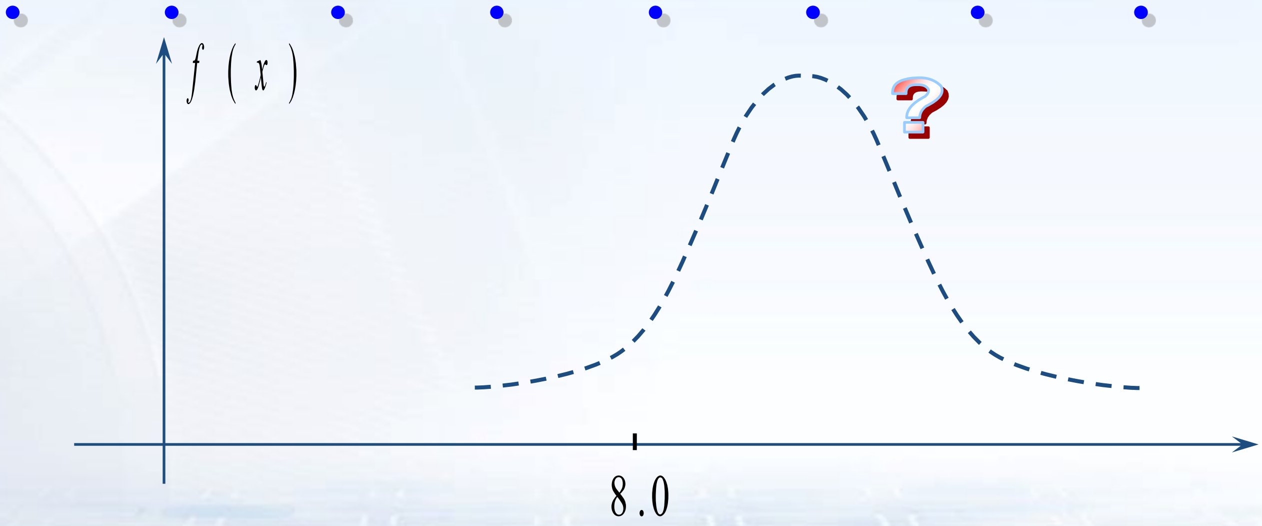
问题　如果知道部分个体的值，能否预测总体？



灯泡寿命　$X \sim N(\mu, \sigma^2)$　　　$\mu = ?\ \sigma^2 = ?$

例 设某台机床加工的零件的长度 $X \sim N(\mu,\ 0.1)$ 实测了其中 8 个零件，得到它们的测量值为

$$8.3,\quad 7.7,\quad 8.6,\quad 8.0,\quad 8.6,\quad 7.7,\quad 8.6,\quad 8.0$$



定义1 从总体 X 中抽取的部分个体，得到的数量指标 $X_1, X_2, \ldots, X_n$，若满足下列条件：

(1) $X_1, X_2, \ldots, X_n$ 与 X 同分布；

(2) $X_1, X_2, \ldots, X_n$ 相互独立.

则称 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的一个简单随机样本，简称样本.

对样本 $X_1, X_2, \ldots, X_n$ 进行观测后，得到的观测值：$x_1, x_2, \ldots, x_n$ 称为**样本观测值**.

注：

观测前：$X_1, X_2, \ldots, X_n$ 是随机变量；

观测后：$x_1, x_2, \ldots, x_n$ 是具体的数据.

样本的联合分布

设总体 $X \sim F(x)$，则样本 $X_1, X_2, \ldots, X_n$ 的联合分布函数为：

$$F(x_1, \cdots, x_n) = P\{X_1 \le x_1, \cdots, X_n \le x_n\} = \prod_{i=1}^{n} F(x_i)$$

若总体 X 的密度函数为 $f(x)$，则样本 $X_1, X_2, \ldots, X_n$ 的联合密度函数为：

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

样本的联合分布

若总体 X 的分布律为：

$$P\{X=a_k\}=p_k,\quad k=1,2,\cdots$$

则样本 X_1，X_2，…，X_n 的联合分布律为：

$$P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$$

$$=P\{X_1=x_1\}P\{X_2=x_2\}\cdots P\{X_n=x_n\}$$

$$=\prod_{i=1}^{n}P\{X=x_i\}$$

例 设 $X_1, X_2, \ldots, X_n$ 是来自总体 $N(\mu, \sigma^2)$ 的样本，则样本的联合密度函数为：

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

$$=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

例 设 $X_1, X_2, \ldots, X_n$ 是来自总体 $B(1, p)$ $(0 < p < 1)$ 的样本,则样本的联合分布律为：

$$P\{X_1 = x_1, \mathrm{L} , X_n = x_n\} = \prod_{i=1}^{n} P\{X = x_i\}$$

$$= \prod_{i=1}^{n} p^{x_i} p^{(1-x_i)}$$

$$= p^{\sum_{i=1}^{n} x_i} p^{\sum_{i=1}^{n} (1-x_i)}$$

总体， 样本， 样本值的关系

对样本的一些认识

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

1. $X_1, X_2, \ldots, X_n$ 是一堆“杂乱无章”的数据；
2. $X_1, X_2, \ldots, X_n$ 包含总体的相关“信息”；
3. $X_1, X_2, \ldots, X_n$ 是对总体进行推断的依据；
4. 观测前，$X_1, X_2, \ldots, X_n$ 是 $i.i.d.$ 随机变量，
   观测后，$x_1, x_2, \ldots, x_n$ 是具体的数据.

- **统计推断的基础：收集数据**

从总体 $X \sim F(x)$中抽取样本：

$$\boxed{X_1,\ X_2,\ \ldots,\ X_n}$$

“杂乱无章”的数据　　**包含了有用的“信息”**

- **问题**

如何提炼出有用的“信息”？

第29讲　数理统计的基本概念

例 设某班级英语考试后，全班同学的成绩分别为：$X_1, X_2, \ldots, X_n$

- **问题** 你除了希望知道自己的成绩外，还关心哪个成绩?
- **问题** 如何评价该班级的英语整体学习情况?

$$\max\{X_1, X_2, \cdots, X_n\} \qquad \frac{1}{n}\sum_{i=1}^{n} X_i$$

--对样本进行“整理”后得到的数据

数据的整理：统计量

定义2 设 $X_1, X_2, \ldots, X_n$ 是来自总体$X \sim F(x)$的样本，$g(x_1, x_2, \ldots, x_n)$ 是n元实值连续函数，若函数$g(x_1, x_2, \ldots, x_n)$ 不含未知参数，则称随机变量 $g(X_1, X_2, \ldots, X_n)$为**统计量**.

例 设X_1，X_2，…，X_n是来自总体$N(\mu,\sigma^2)$的样本，其中μ，σ^2均未知．则下列随机变量中哪些是统计量?

$$\frac{1}{n}\sum_{i=1}^{n}X_i,\quad \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2,\quad \min\{X_1,X_2,\cdots,X_n\}$$

常用统计量

设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本，则称

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

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

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

(3) $A_k = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k$ 　为样本的 k 阶原点矩

(4) $B_k = \dfrac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k$ 　为样本的 k 阶中心矩

(5) 将样本 X_1，X_2，…，X_n 按由小到大的顺序排成

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

则称统计量 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为顺序统计量

称 $X_{(1)} = \min\{X_1, X_2, \cdots, X_n\}$ 为样本极小值

称 $X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$ 为样本极大值

称 $R_n = X_{(n)} - X_{(1)}$ 为样本极差

样本均值与样本方差的数字特征

命题1　设 $X_1, X_2, \ldots, X_n$ 是来总体 X 的样本，且总体的均值与方差存在，记为

$$E(X)=\mu,\quad D(X)=\sigma^2$$

则有

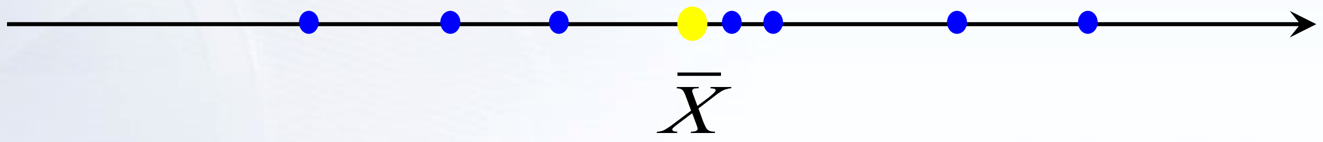
(1)　$E(\bar{X})=\mu,\quad D(\bar{X})=\frac{1}{n}\sigma^2$

(2)　$E(S^2)=\sigma^2$

样本均值与样本方差的含义

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$ —— 是观测数据X_1，X_2，…，X_n的平均值

是观测数据 X_1，X_2，…，X_n 的“中心”

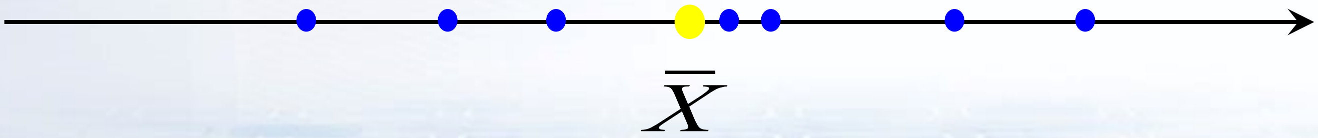


样本均值与样本方差的含义

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

—— 反映了观测数据 $X_1, X_2, \ldots, X_n$ 与观测数据中心点的偏离程度.

反映了观测数据 $X_1, X_2, \ldots, X_n$ 的离散程度.



问题

下结果说明了什么?

$$E(\bar{X}) = \mu,\quad D(\bar{X}) = \frac{\sigma^2}{n},\quad E(S^2) = \sigma^2$$