概率论与数理统计

第16讲

数理统计的基本概念

一. 数理统计介绍

前五章讲述了概率论的基本内容，现在将讲述数理统计的内容。在概率论中，所研究的随机变量，其概率分布都假设已知，在这一前提下研究随机变量的性质、特点及其规律。

但是在许多实际问题中，描述随机现象的随机变量的概率分布可能完全不知道；或者由于现象的某些事实知道其分布形式，但是不知道其分布函数中所含的参数。

一. 数理统计介绍

例1.某种型号的电视机的寿命X服从什么分布完全未知。

（1）希望知道它的分布函数$F(x)$；

（2）或者，希望知道该分布的一些特征值，如期望和方差等。

随机变量X的概率分布完全未知

一. 数理统计介绍

例2.某商业部门收购一批工业产品，每件产品或为合格品，或为不合格品。令

$$X=\begin{cases}1, & \text{产品为合格品}\\ 0, & \text{产品为不合格品}\end{cases}$$

易知$X\sim b(1,p)$，（即0—1分布），其中p为合格率，但p未知。

→ 我们希望知道p是多少。

随机变量的概率分布形式已知，但是参数p的值未知。

例3.某工厂生产大批电子元件，已知元件的寿命X服从指数分布，若按国家规定，这种元件的平均寿命应大于5000小时。

（1）元件的平均寿命是多少?

（2）该工厂生产的元件是否符合国家的规定?

易知X的密度形式为 $f(x)=\begin{cases}\frac{1}{\theta}e^{-\frac{1}{\theta}x}, & x>0\\ 0, & x\le 0\end{cases}$

随机变量的概率分布形式已知，参数θ的值未知。

问题（1）实际上是问θ是多少?

问题（2）实际上是需要推断$\theta>5000$是否成立?

一. 数理统计介绍

类似的问题在生产和经济活动中是经常会遇到的，数理统计就是在解决这些实际问题中逐渐形成的一个独立的学科。

从理论上讲，只要我们对随机现象进行大量的观察或试验，它的统计规律性就会呈现出来。

然而，在实际中常常是办不到的。如：元件寿命试验。

由于元件寿命试验是破坏性试验，不可能把每一个元件进行测试，只能抽取一部分元件进行测试，通过抽取这部分元件的寿命数据来推断整批元件的寿命情况。

一. 数理统计介绍

数理统计方法具有“局部推断整体”的特点。

局部既然是整体的一部分，它必然能反映出整体的某些信息；但是局部又不是整体，它决不能准确地反映出整体的全部信息。

一个好的统计方法，是使由局部推断出的有关整体的信息尽可能地准确。

由于是对随机现象进行观察或试验，因此，观察或试验数据是带有随机性的。为此需要我们从中尽可能地排除随机性的干扰，以作出合理的推断。

一. 数理统计介绍

数理统计的任务就是研究怎样以有效的方式收集、整理和分析带有随机性的数据，在此基础上，对所研究的问题作出推断，直至对可能作出的决策提供依据和建议。

数理统计研究的内容

第一： 怎样对随机现象进行有限次的观察或试验，即如何收集数据？

——试验设计与抽样方法

第二： 如何对这有限次的观察或试验所得到的带有随机性的数据进行合理的分析，作出科学的推断？

——统计推断

一. 数理统计介绍

概率论

已知随机变量的分布，可求得：某个随机事件出现的概率，随机变量落在某个区间的概率，随机变量的数字特征(均值，方差，协方差，相关系数)等等。

数理统计

已知随机变量的取值（数据），去求随机变量的分布，或一些数字特征（均值，方差，协方差，相关系数）等。

一. 数理统计介绍

数理统计与概率论的关系

概率论是数理统计的基础，数理统计是概率论的重要应用。

它们是两个并列的数学分支学科，并无从属关系。

- 数理统计
  - 抽样分布
  - 参数估计
  - 假设检验

二. 基本概念

总体： 在一个统计问题中，所研究问题涉及到的研究对象全体组成的集合。

个体： 组成总体的每个元素。

例如：研究某批灯泡的质量

总体：该批灯泡

例如：研究北京理工大学学生的学习情况

总体：北京理工大学全体学生

二. 基本概念

总体中所包含个体的个数称为总体的容量，容量为有限的总体称为有限总体，容量为无限的总体称为无限总体。

但是，实际问题中，我们关心的并不是总体或这些个体本身，而是关心与个体性能相联系的某一项(或某几项)数量指标以及这些数量指标在总体中的分布情况。表征个体特征的量称为数量指标。

一般来说，用X表示数量指标(可以是一维也可以是多维)，称数量指标X的全体组成的集合称为总体，而其中每个数量指标称为个体。

例4 研究100万只灯泡的使用寿命，这100万只灯泡的使用寿命组成总体，而其中每个灯泡的寿命是个体。

例5 研究北京理工大学全体在读学生的身高和体重，北京理工大学全体在读学生的身高和体重组成总体，而其中每个学生的身高和体重是个体。

二. 基本概念

对于一个总体，用X表示其数量指标，由于每个个体的出现是随机的，那么每个个体上数量指标的出现也带有随机性。

因此X是一个随机变量。此时，把总体与一个随机变量(如灯泡寿命)对应起来。

因此，对总体的研究转化为对表示总体的随机变量X的统计规律性的研究。随机变量X的分布函数和数字特征也称为总体的分布函数和数字特征。今后将不区分总体与相应的随机变量，统称为总体X。

例6. 灯泡使用寿命这一总体是指数分布总体，指总体中的观察值是指数分布随机变量的值。

例7. 检查自生产线上生产出来的零件是次品还是正品，定义随机变量X如下

$$X=\begin{cases}1, & \text{次品}\\ 0, & \text{正品}\end{cases}$$

X是一个参数为p的0-1分布的随机变量，其中$p=P\{X=1\}$为次品率。我们将它说成0-1分布总体，指总体中的观察值是0-1分布随机变量的值。

注：当总体分布为指数分布时，称为指数分布总体；当总体分布为正态分布时，称为正态分布总体或正态总体；当总体分布为二项分布时，称为二项分布总体等等。

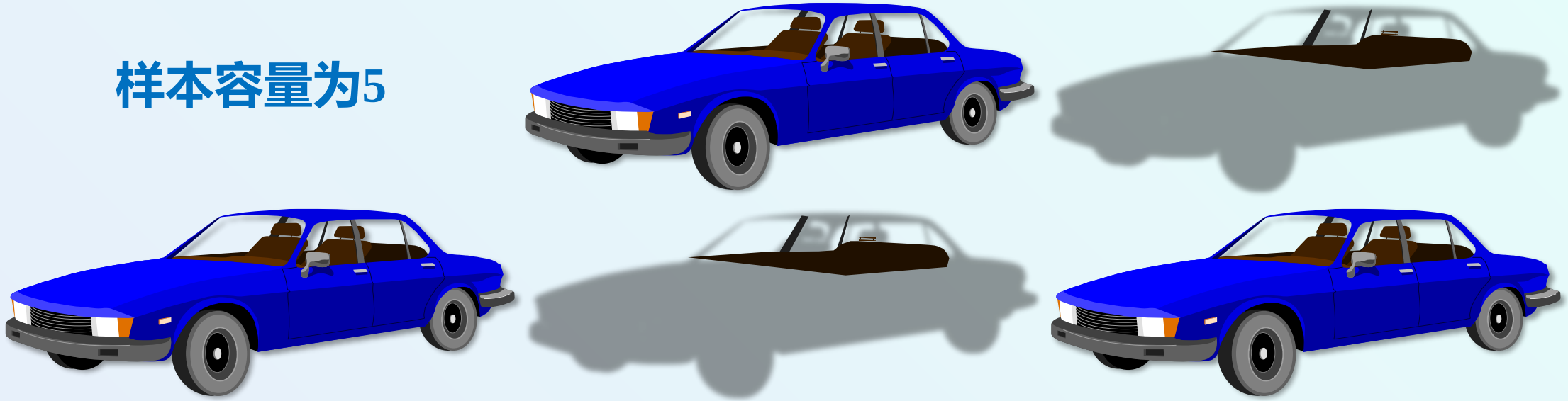
二. 基本概念

为推断总体分布及其各种特征，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。

从总体中抽取一个个体，就是对总体X进行一次观察并记录其结果。我们对总体X进行n 次独立重复观察，将n 次观察结果依次记为$X_1, X_2,\ldots, X_n$。由于某n 次抽样与另外n 次抽样所得的X_i (i=1,2,…,n)一般取不同的数值。因此重复抽样中每一个X_i 应该看作是一个随机变量。

例如：从国产轿车中抽5辆进行耗油量试验

样本容量为5



随机抽5辆：用$X_1,\ldots,X_5$表示待抽的第1到第5辆汽车的耗油量

此时，$(X_1,X_2,\ldots,X_5)$是5维随机变量

但是，一旦取定一组样本，得到的是5个具体的数$(x_1, x_2, \ldots, x_5)$

二. 基本概念

容量为n的样本可以看作n维随机变量$(X_1, X_2, \ldots, X_n)$。

但是，一旦取定一组样本，得到的是n个具体的数$(x_1, x_2, \ldots, x_n)$，称为样本的观察值，简称样本值。

由于抽样的目的是为了对总体进行统计推断，为了使抽取的样本能很好地反映总体的信息，必须考虑抽样方法。

最常用的一种抽样方法满足如下要求：

代表性

总体中的每一个个体X_i 都有同等机会被抽取，即要求每个个体X_i 与总体X有相同的分布，因此每一个被抽取的个体都具有代表性。

独立性

每一次抽样不影响下一次抽样的情况，即要求$X_1, X_2, \ldots, X_n$相互独立。

上述抽样方法称为简单随机抽样，利用简单随机抽样方法得到的样本，称为简单随机样本。

定义1： 设有一个总体X，若X_1，X_2，…，X_n相互独立且每个X_i (i=1,2,…,n)与总体X有相同的分布，则称X_1，X_2，…，X_n为从总体X中抽取的容量为n的简单随机样本，简称样本。

若总体X的分布函数为$F(x)$，也称$X_1, X_2, \ldots, X_n$为从总体分布$F(x)$中抽取的容量为n的样本，记为$X_1, X_2, \ldots, X_n$ *i.i.d.*，且$X_i \sim F(x)$，i=1,2,…,n。

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。

二. 基本概念

样本作为随机变量有概率分布，称这个概率分布为样本分布，样本分布可由总体分布完全确定。样本分布是样本受随机性影响最完整的描述。

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

$$F(x_1, x_2, \ldots, x_n) = F(x_1)\, F(x_2) \ldots F(x_n)$$

若总体X为离散型随机变量，其分布律为$f(x)$，即 $f(x) = P\{X=x\}$

则样本$X_1, X_2, \ldots, X_n$ 的联合分布律为

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} \\ &= P\{X_1 = x_1\}P\{X_2 = x_2\}\ldots P\{X_n = x_n\} \\ &= P\{X = x_1\}P\{X = x_2\}\ldots P\{X = x_n\} = f(x_1)f(x_2)\ldots f(x_n) \end{aligned}$$

例8.设总体X服从参数为p的两点分布，$X_1, X_2, \ldots, X_n$为来自该总体的样本，求$X_1, X_2, \ldots, X_n$的联合分布律。

解：由于X的分布律为 $P\{X = x\} = p^x q^{1-x}, x = 0, 1$

所以$X_1, X_2, \ldots, X_n$的联合分布律为

$$f(x_1, x_2, \ldots, x_n) = P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$$

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

$$= p^{x_1}q^{1-x_1}\, p^{x_2}q^{1-x_2} \ldots p^{x_n}q^{1-x_n}$$

$$= p^{\sum_{i=1}^{n} x_i}\, q^{n-\sum_{i=1}^{n} x_i}$$

$x_i = 0, 1,\ i = 1, 2, \ldots, n$

若总体X的概率密度为$f(x)$，则样本$X_1, X_2, \ldots, X_n$的联合概率密度函数为

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

例9. 设总体$X \sim N(\mu, \sigma^2)$，$\mu \in R$，$\sigma > 0$，$X_1, X_2, \ldots, X_n$为从总体X中抽取的样本，求样本$X_1, X_2, \ldots, X_n$的概率密度函数。

解：总体X的概率密度函数为 $f(x;\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

因此，样本$X_1, X_2, \ldots, X_n$ 的概率密度函数是

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

数理统计的重要任务之一就是通过样本推断总体的特性。但是由于样本中所含总体信息较为分散，有时显得杂乱无章，一般不宜直接用于统计推断。

因此，为将这些分散在样本中的有关总体的信息集中起来以反映总体的各种特征，常常要把样本中的信息进行整理、加工、提炼，集中样本中的有用信息，针对不同的研究问题构造样本的适当函数，再利用样本的函数进行统计推断。

例10. 从某厂生产的自行车头盔中抽取10件进行检测，结果是前3件为不合格品，后面7件为合格品，依此估计不合格品率p。

分析：总体为$X\sim b(1,p)$，$0<p<1$，样本为$X_1, X_2, \ldots, X_{10}$

样本提供的信息为

（1）10次试验中1出现的次数，即：不合格品的个数；

（2）不合格品在哪次试验中出现。 不重要

如何利用这些数据信息来估计未知参数。

三. 统计量

定义 设$X_1, X_2, \ldots, X_n$为来自总体X的样本，$T(X_1, X_2, \ldots, X_n)$是样本$X_1, X_2, \ldots, X_n$的函数，且不含任何未知参数，则称$T(X_1, X_2, \ldots, X_n)$是统计量。

若$x_1, x_2, \ldots, x_n$是相应于样本$(X_1, X_2, \ldots, X_n)$的样本值，则称$T(x_1, x_2, \ldots, x_n)$是$T(X_1, X_2, \ldots, X_n)$的观察值。

说 明

1. 统计量是不含未知参数的样本的函数。
2. 统计量既然依赖于样本，而后者又是随机变量，即统计量是随机变量的函数，故统计量是随机变量。

例11.设$X_1, X_2, \ldots, X_n$为来自正态总体$N(\mu, \sigma^2)$的一个样本，其中μ未知，σ^2已知，下列随机变量中哪些是统计量。

(1) $\frac{1}{n}\sum_{i=1}^{n} X_i$　　(2) $\frac{1}{n}\sum_{i=1}^{n} X_i - \mu$　　(3) $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$

(4) $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$　　(5) $X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$　　(6) $X_1 + X_2$

答：（1）（3）（5）（6）是统计量，（2）（4）不是

设$X_1, X_2, \ldots, X_n$为来自总体X的一个样本，$x_1, x_2, \ldots, x_n$是这一样本的观察值，定义如下常用统计量。

1. 样本均值 $\frac{1}{n}\sum_{i=1}^{n} X_i$ 反映了总体均值EX的信息。

2. 样本k阶（原点）矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \cdots$

反映了总体k阶(原点)矩EX^k的信息。

3.样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 反映了总体方差DX的信息。

4.样本标准差 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$ 反映了总体标准差 $\sqrt{DX}$ 的信息。

5.样本k阶中心矩 $B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k, k=2,3,\cdots$

反映了总体标准差 $E[(X-E(X)]^k$ 的信息。

注意：B_2和S^2的区别。

它们的观察值分别为 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$　$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^{\,k},k=1,2\cdots$

$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$　$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$　$b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k,k=2,3\cdots$

仍分别称为样本均值，样本k阶（原点）矩，样本方差，样本标准差，样本k阶中心矩。

6. 设$X_1, X_2, \ldots, X_n$ 是来自总体X的样本，将$X_1, X_2, \ldots, X_n$ 按照从小到大排列为

$$X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$$

称$(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$为$X_1, X_2, \ldots, X_n$ 的顺序统计量。

$X_{(k)}$，$1 \le k \le n$ 为第k个顺序统计量；

$X_{(1)} = \min\limits_{1 \le i \le n} X_i = \min(X_1, X_2, \cdots, X_n)$ 为最小顺序统计量；

$X_{(n)} = \max\limits_{1 \le i \le n} X_i = \max(X_1, X_2, \cdots, X_n)$ 为最大顺序统计量；

$R = X_{(n)} - X_{(1)}$ 为样本极差，反映了样本观察值的最大波动程度；

$$\tilde{X}=\begin{cases} X_{\left(\frac{n+1}{2}\right)}, & n=2m+1, \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right), & n=2m. \end{cases}$$

称为样本中位数，它是把样本分为两部分，而样本中位数恰好是分界线。

作业：1,2,3

第16讲

谢谢观看