# Seminar-数理统计

## 绪论与第一周作业


**姜青娥**

克莱登大学


2023 年 4 月 26 日

上一次修改于: April 26, 2023

# Contents

# Chapter 1

# 绪论

## 1.1 基本概念

**Example 1.** *一批产品共 10000 件，其中有正品有废品，为估计废品率，从中抽取 100 件进行检查.*

*10000 件产品为总体，每件样本为个体，100 件为样本，100 叫样本大小，或样本容量. 这个行为叫抽样.*

若总体中的个体的数目为有限个，则称为有限个体. 无限个体的例子比如产品的寿命问题，时间的取值是无限的.

我们一般关心的是个体上的数量指标，比如寿命，尺寸等. 个体上的数量指标带有随机性，因此可以把该数量指标看成随机变量 (random variable, r.v.)

数量指标在总体上的分布情况就是随机变量的分布. 数量指标就是可以理解为用数字代替的特征. 比如

$$X=\begin{cases}1 & \text{废品}\\ 0 & \text{正品}\end{cases} \tag{1.1}$$

中，特定个体上的数量指标是 $r.v.X$ 的观察值.

**Definition 2.** 统计问题研究的对象的全体称为总体. 在数理统计这门课中，总体可以用一个随机变量及其概率分布来描述.

有时候总体可以用 $r.v.X$ 、分布函数 $F$ 来表示，若 $F$ 也有密度 $f$, 也可以用密度函数来表示.

当从一个总体中抽取样本大小为 $n$ 的样本 $X_1,\cdots,X_n$ 时，它们一定是独立同分布的，记作 **i.i.d.** .

当个体上的数量指标不止一项时，则用随机向量来表示. 比如研究某地区的气温和降雨量，则可以用 $T$ 表示温度，$R$ 表示降雨量, 总体用二维随机变量 $(T,R)$ 或者联合分布 $F(t,r)$ 来表示. 假如 $F$ 有密度 $f$ , 也可以用联合密度 $f(t,r)$ 来表示.

**Remark 3.** 通过上面的表达，我们发觉，分布可能不存在密度.

**Definition 4.** 样本 $\mathbf{X}=(X_1,\cdots,X_n)$ 可能取值的总体，构成样本空间 $\mathscr{X}$ .

**Example 5.** 打靶试验，每次三发，考察中靶的环数. 如样本 $\mathbf{X}=(5,1,9)$ 表示反三次打靶分别中 *5* 环，*1* 环和 *9* 环. 此时的样本空间为

$$\mathscr{X}=\{(x_1,x_2,x_3):x_i=0,\cdots,10,i=1,2,3\} \tag{1.2}$$

**Definition 6.** 设 $X_1,\cdots,X_n$ 为从总体 $F$ 中抽取的容量为 $n$ 的样本，若

- $X_1,\cdots,X_n$ 相互独立
- $X_1,\cdots,X_n$ 同分布

则称 $X_1,\cdots,X_n$ 为简单随机样本.

有放回的抽样获得的样本就是简单随机样本.

**Remark 7.** 统计模型就是样本分布. 统计上把出现在样本分布中的未知的常数成为参数，多个参数组成参数向量. 如正态分布中的 $(a,\sigma)$.

这些参数需要通过样本去估计. 参数取值的范围称为参数空间. 比如正态分布中的 $\Theta=\{(a,\sigma):a>0,\sigma>0\}$

**Remark 8.** 样本分布包含未知参数，由于参数的取值不同，因此样本的分布就不止一个，当参数取不同的值得到的不同的分布一起构成一个分布族. 记为 $\mathscr{F}=\{f(x,\lambda):\lambda>0\}$

因此更确切地说，统计模型就是样本分布族.

## 1.2 统计量

由样本算出来的量就叫做统计量. 统计量是一个函数，但是在数理统计中的统计量是指具体的函数，不能泛指，不能含有未知参数.

**Definition 9.** 设 $\mathbf{X} \sim P_\theta(\theta \in \Theta)$ 是一个统计模型，则定义在样本空间上的任何函数 $T(x)(x \in \mathscr{X})$ 都称为统计量.

1. 样本均值

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.3}$$

2. 样本方差

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{1.4}$$

   $n-1$ 称为自由度.

3. 次序统计量设 $X_1, \cdots, X_n$ 是从总体中抽取的样本，将其按照大小排列为 $X_{(1)}, \cdots, X_{(n)}$. 按照排列组成的向量 $(X_{(1)}, \cdots, X_{(n)})$，叫做样本的次序统计量.

4. 样本变异系数设 $X_1, \cdots, X_n$ 为从总体 $F$ 中抽取的样本，则称

$$\hat{\nu} = \frac{S_n}{\overline{X}} \tag{1.5}$$

5. 样本的 $k$ 阶原点矩

$$a_{n,k}=\frac{2}{n}\sum_{i=1}^{n}X_i^k, k\in 1,\cdots,n \tag{1.6}$$

6. 样本的 $k$ 阶中心矩

$$m_{n,k}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k, k\in 1,\cdots,n \tag{1.7}$$

7.

**Definition 10.** 设 $X_1,\cdots,X_n$ 为自总体 $F(x)$ 中抽取的 $i.i.d.$ 样本，将其按大小排列为 $X_{(1)}\leq\cdots\leq X_{(n)}$，对任意实数 $x$，称下列函数

$$F_n(x)\begin{cases}0 & x\leq X_{(1)}\\ \frac{k}{n} & X_{(k)}<X\leq X_{(k+1)}, k=1,\cdots,n-1\\ 1 & X_{(n)}<X\end{cases} \tag{1.8}$$

为经验分布函数.

经验分布函数是单调、非降，左连续函数. 并且它仅依赖于样本 $X_1,\cdots,X_n$ 的函数，因此它是统计量.

1. 由中心极限定理，当 $n\to\infty$ 时有

$$\frac{\sqrt{n}(F_n(x)-F(x))}{\sqrt{F(x)(1-F(x))}}\xrightarrow{\mathscr{L}}N(0,1) \tag{1.9}$$

其中 $\mathscr{L}$ 表示依分布收敛

2. 由 Bernoulli 或 (辛钦) 大数定律，则当 $n \to \infty$ 时有

$$F_n(x) \xrightarrow{P} F(x) \tag{1.10}$$

3. 由强大数定律，则有

$$P\left(\lim_{n\to\infty} F_n(x) = F(x)\right) = 1 \tag{1.11}$$

4. 更进一步，有格里汶科定理 (Glivenko-Cantelli Theorem)

**Definition 11.** 设 $F(x)$ 为 $r.v.X$ 的分布函数，$X_1, \cdots, X_n$ 为取自总体 $F(x)$ 的简单随机样本，$F_n(x)$ 为其经验函数，记 $D_n = \sup\limits_{-\infty<x<\infty} |F_n(x) - F(x)|$，则有

$$P\left(\lim_{n\to\infty} D_n = 0\right) = 1 \tag{1.12}$$

## 1.3 第一周课后作业参考答案

1. 试举出一个有限总体的例子，并指出其概率分布. (2 分)

**【解】** 检验一批产品（假设有 $n$ 件）的质量，每一件产品的检验结果为合格或不合格，记录检验结果中合格品的数量 $X$，则 $X$ 的可能取值为 $0,1,2,\ldots,n$，这是一个有限总体的问题。如果记任意一件产品为合格品的概率为 $p$，则这一批产品当中合格品数量 $X$ 的概率分布为

$$P\left(X=k\right)=\binom{n}{k}p^{k}\left(1-p\right)^{n-k}\quad,\quad k=0,1,2,\ldots,n$$

**Remark 12.** 一般来说，若随机变量 $X$ 服从参数为 $n$ 和 $p$ 的二项分布，我们记作 $X\sim b(n,p)$ 或 $X\sim B(n,p)$ 。$n$ 次试验中正好得到 $k$ 次成功的概率由分布函数或概率质量函数给出：

$$f(k,n,p)=\Pr(X=k)=\binom{n}{k}p^{k}(1-p)^{n-k}$$

对于 $k=0,1,2,\cdots,n$，其中

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

是二项式系数（这就是二项分布的名称的由来），又记为 $C(n,k),{}_nC_k$，或 ${}^nC_k$ . 该公式可以用以下方法理解：我们希望有 $k$ 次成功（概率为 $p^k$ ）和 $n-k$ 次失败（概率为 $(1-p)^{n-k}$ ）。然而，$k$ 次成功可以在 $n$ 次试验的任何地方出现，而把 $k$ 次成功分布在 $n$ 次试验中共有 $C(n,k)$ 个不同的方法.

2. 试举出一个无限总体的例子，并指出其概率分布. (2 分)

**【解】** 检验一批产品的寿命 $X$（单位：小时），则其可能的取值为 $[0,\ +\infty)$，这是一个无限总体的问题。进一步假设产品的寿命 $X$ 服从指数分布，则其概率密度函数为

$$f_X(x;\ \lambda) = \lambda \mathrm{e}^{-\lambda x} \cdot \mathrm{I}_{[0,\ +\infty)}(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

**Remark 13.** 示性函数

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & \text{其它} \end{cases} \tag{1.13}$$

3. 一个总体有 $N$ 个元素，其指标分别为 $a_1 > a_2 > \cdots > a_N$，指定自然数 $M < N$，$n < N$，并设 $m = \frac{nM}{N}$ 为整数. 在 $(a_1,\ a_2,\ \ldots,\ a_M)$ 中不放回地随机抽出 $m$ 个，在 $(a_{M+1},\ a_{M+2},\ \ldots,\ a_N)$ 中不放回地随机抽出 $n-m$ 个. 写出所得样本的分布. (2 分)

**Remark 14.** 从包含有 $n$ 个不同的元素的总体中取出 $r$ 个来进行排列，既要考虑到取出的元素也要顾及取出顺序. 这种排列分为两类

- 有放回地选取. 从 $n$ 个不同的元素中取出 $r$ 个元素进行排列，这种排列称为有重复的排列，其总数共有 $n^r$ 种.

- 无放回地选取. 从 $n$ 个不同的元素中取出 $r$ 个元素进行排列，其总数为

$$A_n^r = n(n-1)(n-2)\cdots(n-r+1) \tag{1.14}$$

这种排列叫选排列，当 $r=n$ 时，称为全排列.

- $n$ 个不同的元素的全排列数为

$$P_n = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1 = n! \tag{1.15}$$

**【解】** 假设所抽取的样本为 $(X_1, X_2, \ldots, X_n)$，则前面的 $m$ 个个体 $(X_1, X_2, \ldots, X_m)$ 是在 $(a_1, a_2, \ldots, a_M)$ 中不放回地抽取，共有

$$\mathrm{A}_M^m = \binom{M}{m} = \frac{M!}{m!}$$

种等可能的结果. 后面 $n-m$ 个个体 $(X_{m+1}, X_{m+2}, \ldots, X_n)$ 是在 $(a_{M+1}, a_{M+2}, \ldots, a_N)$ 中不放回地抽取, 共有

$$\mathrm{A}_{N-M}^{n-m} = \binom{N-M}{N-M} = \frac{(N-M)!}{(n-m)!}$$

种等可能的结果. 故 $(X_1, X_2, \ldots, X_n)$ 共有

$$\mathrm{A}_M^m \cdot \mathrm{A}_{N-M}^{n-m} = \frac{M!}{m!}\cdot\frac{(N-M)!}{(n-m)!}$$

种等可能的结果. 于是

$$\begin{aligned} P(X_1 = x_1, \dots, X_m = x_m, X_{m+1} = x_{m+1}, \dots, X_n = x_n) &= \frac{1}{\mathrm{A}_M^m \cdot \mathrm{A}_{N-M}^{n-m}} \\ &= \frac{m!(n-m)!}{M!(N-M)!} \end{aligned}$$

4. 一物体的重量 $a$ 未知, 有两架天平可用, 其随机误差分别服从正态分布 $N\left(0, \sigma_1^2\right)$ 和 $N\left(0, \sigma_2^2\right)$, 其中 $\sigma_1^2$ 和 $\sigma_2^2$ 都未知. 先把物体在第一架天平上称两次得 $X_1, X_2$, 再在第二架天平上称两次得 $X_3, X_4$, 然后视 $|X_1 - X_2| \le |X_3 - X_4|$ 与否而在第一架或第二架天平上再称 $n-4$ 次得 $X_5, \dots, X_n$. 写出 $(X_1, X_2, \dots, X_n)$ 的密度. (2 分)

**【解】** 由题意可知 $X_1, X_2 \sim N\left(a, \sigma_1^2\right)$, 该总体的概率密度函数为

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \mathrm{e}^{-\frac{(x-a)^2}{2\sigma_1^2}}$$

同样, $X_3, X_4 \sim N\left(a, \sigma_2^2\right)$, 该总体的概率密度函数为

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_2} \mathrm{e}^{-\frac{(x-a)^2}{2\sigma_2^2}}$$

当 $|X_1-X_2|\le|X_3-X_4|$ 时，$X_i\sim N\left(a,\ \sigma_1^2\right)$，$i=5,\ 6,\ \ldots,\ n$，因此

$$f_{X_i}(x_i)=\frac{1}{\sqrt{2\pi}\sigma_1}\mathrm{e}^{-\frac{(x_i-a)^2}{2\sigma_1^2}},\ i=5,\ 6,\ \ldots,\ n$$

当 $|X_1-X_2|>|X_3-X_4|$ 时，$X_i\sim N\left(a,\ \sigma_2^2\right)$，$i=5,\ 6,\ \ldots,\ n$，因此

$$f_{X_i}(x_i)=\frac{1}{\sqrt{2\pi}\sigma_2}\mathrm{e}^{-\frac{(x_i-a)^2}{2\sigma_2^2}},\ i=5,\ 6,\ \ldots,\ n$$

再根据简单随机样本的定义，我们有 $(X_1,\ X_2,\ X_3,\ X_4,\ X_5,\ X_6,\ \ldots,\ X_n)$ 相互独立，于是其联合概率密度函数为

$$f_{X_1,X_2,\ldots,X_n}\left(x_1,x_2,\ldots,x_n\right)=f_{X_1}\left(x_1\right)f_{X_2}\left(x_2\right)f_{X_3}\left(x_3\right)f_{X_4}\left(x_4\right)f_{X_5}\left(x_5\right)\cdots f_{X_n}\left(x_n\right)$$

$$
= \begin{cases} \dfrac{1}{\left(\sqrt{2\pi}\right)^n \sigma_1^{n-2}\sigma_2^2} \exp\left[-\dfrac{\sum\limits_{i=1}^{2}(x_i-a)^2+\sum\limits_{j=5}^{n}(x_j-a)^2}{2\sigma_1^2}-\dfrac{\sum\limits_{k=3}^{4}(x_k-a)^2}{2\sigma_2^2}\right], \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad |X_1-X_2|\le|X_3-X_4| \\ \dfrac{1}{\left(\sqrt{2\pi}\right)^n \sigma_1^{2}\sigma_2^{n-2}} \exp\left[-\dfrac{\sum\limits_{i=1}^{2}(x_i-a)^2}{2\sigma_1^2}-\dfrac{\sum\limits_{j=3}^{n}(x_j-a)^2}{2\sigma_2^2}\right], \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad |X_1-X_2|>|X_3-X_4| \end{cases}
$$

5. 设总体 $X$ 服从两点分布 $b(1,\ p)$ （即 $P(X=1)=p$，$P(X=0)=1-p$），其中 $p$ 是未知参数，$\boldsymbol{X}=(X_1,\ X_2,\ X_3,\ X_4,\ X_5)$ 为从此总体中抽取的简单样本，

(a) 写出样本空间 $\mathscr{X}$ 和 $\boldsymbol{X}$ 的概率分布. (2 分)

**【解】** 样本空间为

$$
\mathscr{X}=\{(X_1,\ X_2,\ X_3,\ X_4,\ X_5)\ :\ X_i=0 \text{ 或 } 1,\ i=1,\ 2,\ 3,\ 4,\ 5\}
$$

$\boldsymbol{X}$ 的概率分布为

$$P\left(X_1=x_1,\ X_2=x_2,\ X_3=x_3,\ X_4=x_4,\ X_5=x_5\right)=p^{\sum\limits_{i=1}^{5}x_i}(1-p)^{5-\sum\limits_{i=1}^{5}x_i}$$

其中 $x_i=0$ 或 1，$i=1,\ 2,\ 3,\ 4,\ 5$. 或者 $\sum\limits_{i=1}^{5}X_i\sim b(5,\ p)$.

(b) 指出 $X_1+X_2$，$\min\limits_{1\leq i\leq 5}X_i$，$X_5+2p$，$X_5-E\left(X_1\right)$，$\dfrac{\left(X_5-X_1\right)^2}{D\left(X_1\right)}$ 哪些是统计量，哪些不是统计量，并说明理由. (2 分)

【解】 因为 $X_1+X_2$ 和 $\min\limits_{1\leq i\leq 5}X_i$ 是样本 $\boldsymbol{X}=\left(X_1,\ X_2,\ X_3,\ X_4,\ X_5\right)$ 的函数，且不含未知参数，因此它们是统计量. $X_5+2p$，$X_5-E\left(X_1\right)$ 和 $\dfrac{\left(X_5-X_1\right)^2}{D\left(X_1\right)}$ 虽然也是样本 $\boldsymbol{X}=\left(X_1,\ X_2,\ X_3,\ X_4,\ X_5\right)$ 的函数，但其中含有未知参数 $p$，所以它们不是统计量.

6. 设 $a\neq 0$ 和 $b$ 皆为常数，令 $y_i=ax_i+b$，$i=1,\ 2,\ \ldots,\ n$.

(a) 证明 $y_1$，$y_2$，$\ldots$，$y_n$ 的样本均值 $\overline{y}$ 与 $x_1$，$x_2$，$\ldots$，$x_n$ 的样本均值 $\overline{x}$ 之间的关系为 $\overline{y}=a\overline{x}+b$. (2 分)

**【证明】** 证明过程如下：

$$\begin{aligned}\overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\ &= a\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{1}{n}\sum_{i=1}^{n} b \\ &= a\overline{x} + b\end{aligned}$$

(b) 证明 $y_1$，$y_2$，$\ldots$，$y_n$ 的样本方差 $S_y^2$ 与 $x_1$，$x_2$，$\ldots$，$x_n$ 的样本方差 $S_x^2$ 之间的关系为 $S_y^2 = a^2 S_x^2$. (2 分)

【证明】 证明过程如下：

$$
\begin{aligned}
S_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[(ax_i+b)-(a\overline{x}+b)]^2 \\
&= \frac{a^2}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \\
&= a^2 S_x^2
\end{aligned}
$$

(c) 根据上述结果，利用适当的变换，求下列数据的样本均值和样本方差: (2 分)

$$480,\quad 550,\quad 500,\quad 590,\quad 510,\quad 560,\quad 490,\quad 600,\quad 580.$$

【解】 做变换 $y = 10x + 550$，则当 $x$ 取值

$$-7,\quad 0,\quad -5,\quad 4,\quad -4,\quad 1,\quad -6,\quad 5,\quad 3$$

时，$y$ 取值即为

$$480,\quad 550,\quad 500,\quad 590,\quad 510,\quad 560,\quad 490,\quad 600,\quad 580.$$

容易求得

$$\overline{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = -1, \quad s_x^2 = \frac{1}{9-1}\sum_{i=1}^{9}(x_i - \overline{x})^2 = 21$$

根据上述公式，我们就有

$$\overline{y} = 10\overline{x} + 550 = 10 \times (-1) + 550 = 540, \quad s_y^2 = 10^2 s_x^2 = 2100$$