数理统计疑问解析

姜炳麟 贾玉心

华中理工大学出版社

数理统计疑问解析

姜炳麟　贾玉心

华中理工大学出版社

内 容 提 要

本书以问答的形式深入浅出地阐述了数理统计中的基本概念、原理和方法。全书共包括五章：基本概念与抽样分布、参数估计、假设检验、方差分析与回归分析。本书所论述的问题多为初学者在学习数理统计过程中易产生的疑问。对帮助读者深入理解数理统计中的有关概念、原理和方法是有益的。

本书可供高等工科院校各专业本科生和研究生参阅，也可供自学数理统计的各类人员学习时参考。

数理统计疑问解析

姜炳麟 贾玉心

责任编辑 李立鹏

*

华中理工大学出版社出版发行

（武昌喻家山·430074）

新华书店湖北发行所经销

武汉大学出版社印刷总厂印刷

*

开本：850×1168 1/32 印张：10.375 字数：260 000

1993年6月第1版 1993年6月第1次印刷

印数：1-1 000

ISBN7-5609-0752-0/O·102

定价：2.50元

（鄂）新登字第10号

前　言

数理统计的任务是用收集到的试验数据来分析和研究随机现象的规律性，并对所研究的对象作出合理的统计推断。随着科学技术的发展，数理统计方法越来越广泛地被人们所重视。目前，许多大专院校都开设了数理统计课。

对工科院校本科生、研究生学习数理统计课程的要求应是正确理解基本概念和原理，熟练掌握统计的基本方法。学生在学习过程中感到困难的往往不是某些数学形式的推导，而是在于不善于正确把握和深刻理解统计思想。我们认为作为高等院校的数理统计课不应只从纯技术性这一点出发，仅教给学生一些现成的统计方法如何使用，而更重要的是应培养学生树立正确的统计思想，从而才能使学生能灵活、准确运用统计的概念、原理和方法，去观察、研究随机现象的规律性。

当学生在认真对待和逐渐深入了解数理统计基本概念、原理和方法，从而体会统计思想时，自然会遇到种种疑问。我们希望通过本书中问题的解析能对初学者深入理解数理统计中有关概念、原理等方面有所帮助。

本书所讨论的问题涉及一般工科院校数理统计课内容的各个方面。它包括数理统计基本概念、抽样分布、点估计、区间估计、假设检验、方差分析及回归分析等问题。我们力求做到以下几点：(1)所选问题既有普遍性，又有典型性；(2)通过大量例子及直观方法使问题解析简明易懂；(3)为了使问题的解答准确而有说服力，也配有必要的数学上的分析和论证；(4)当需要涉及到工科院校一般数理统计教材之外内容时，不作过多的阐述。

由于编者水平有限，不可避免会有谬误。恳请读者批评指正。

作者

目　录

第一章　基本概念与抽样分布

内容摘要

基本概念

总体与个体：所研究对象的全体称为**总体**. 当总体是数量指标时，常记为 X，它是一个随机变量. 总体中的元素称为**个体**.

样本与样本值：对总体进行 n 次观测得到 n 个观测值 $x_1, x_2, \cdots, x_n$，称向量 $(x_1, x_2, \cdots, x_n)$ 为来自此总体的**样本值**. 由于抽样是随机的，故 $(x_1, x_2, \cdots, x_n)$ 是**随机向量** $(X_1, X_2, \cdots, X_n)$ 的一个可能取值，称 $(X_1, X_2, \cdots, X_n)$ 为来自此总体的**样本**.

简单随机样本：如果样本的分量，即观测 $X_1, X_2, \cdots, X_n$ 是独立的且与总体同分布，则称此样本为简单随机样本，简称为样本.

如果总体的分布函数为 $F(x)$，则简单随机样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布函数为

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

统计量：样本 $(X_1, X_2, \cdots, X_n)$ 的不含任何未知参数的函数 $T = T(X_1, X_2, \cdots, X_n)$ 称为**统计量**. 常见统计量有

样本 r 阶原点矩 $A_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$，特别称 $A_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 为**样本均值**.

样本 r 阶中心矩 $B_r = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^r$，特别称 $B_2^* = S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$ 为**样本方差**.

顺序统计量：设$(x_1,x_2,\cdots,x_n)$是样本$(X_1,X_2,\cdots,X_n)$的一个样本值，将样本值分量按大小顺序排列：

$$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$$

定义统计量$X_{(i)}$，其取值为$x_{(i)}$，称它为样本的第i个**顺序统计量**. 称$(X_{(1)},X_{(2)},\cdots,X_{(n)})$为样本的一组顺序统计量. 特别称$X_{(1)}$为最小项统计量，$X_{(n)}$为最大项统计量. 称$R=X_{(n)}-X_{(1)}$为极差.

经验分布函数：设$(X_{(1)},X_{(2)},\cdots,X_{(n)})$为样本$(X_1,X_2,\cdots,X_n)$的一组顺序统计量，则称函数

$$F_n^*(x)=\begin{cases}0, & x\leqslant X_{(1)};\\ k/n, & X_{(k)}<x\leqslant X_{(k+1)},k=1,2,\cdots,n-1;\\ 1, & x>X_{(n)}.\end{cases}$$

为样本$(X_1,X_2,\cdots,X_n)$的**经验分布函数**.

若$F(x)$为总体的分布函数，$F_n^*(x)$为来自此总体样本的经验分布函数，则有

$$P\{\lim_{n\to\infty}\sup_{-\infty<x<+\infty}|F_n^*(x)-F(x)|=0\}=1.$$

常用的三个分布

(1)χ^2分布

设随机变量$X_1,X_2,\cdots,X_n$相互独立且同分布于正态分布$N(0,1)$，则称随机变量$\chi^2=\sum_{i=1}^{n}X_i^2$服从自由度为$n$的$\chi^2$分布，记为$\chi^2\sim\chi^2(n)$. 其分布密度为

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x>0;\\ 0, & x\leqslant 0.\end{cases}$$

χ^2分布有以下性质

(i)$E(\chi^2)=n,D(\chi^2)=2n$;

(ii)若$\chi_1^2\sim\chi^2(n_1),\chi_2^2\sim\chi^2(n_2)$，且独立，则$\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2)$.

(iii)若 $\chi_n^2 \sim \chi^2(n)$，$n=1,2,\cdots$，则对任意 x 有

$$\lim_{n\to\infty} P\left\{\frac{\chi_n^2-n}{\sqrt{2n}}<x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt.$$

(2)t 分布

设随机变量 $X\sim N(0,1)$，$Y\sim\chi^2(n)$且独立，则称随机变量 $T=X/\sqrt{Y/n}$服从自由度为 n 的 t 分布，记为 $T\sim t(n)$. 其分布密度为

$$f(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},\quad -\infty<x<+\infty.$$

t 分布有以下性质：

(i) $E(T)=0(n>1)$，$D(T)=\dfrac{n}{n-2}(n>2)$

(ii) 若 $T_n\sim t(n)$，$n=1,2,\cdots$，则对任意 x 有

$$\lim_{n\to\infty} P\{T_n<x\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt.$$

(3)F 分布

设随机变量 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$且独立，则称随机变量 $F=\dfrac{X/n_1}{Y/n_2}$服从自由度为(n_1,n_2)的 F 分布，记为 $F\sim F(n_1,n_2)$，其分布密度为

$$f(x)=\begin{cases}\dfrac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)\left(\frac{n_1}{n_2}x\right)^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x>0;\\ 0, & x\leqslant 0.\end{cases}$$

F 分布有以下性质：

(i) $E(F)=\dfrac{n}{n-2}\quad(n>2)$；

(ii) 若 $t\sim t(n)$，则 $t^2\sim F(1,n)$；

(iii) 若 $F\sim F(n_1,n_2)$则$\dfrac{1}{F}\sim F(n_2,n_1)$.

正态总体下的抽样分布

主要结果有

(1)设总体 $X\sim N(\mu,\sigma^2)$,$(X_1,X_2,\cdots,X_n)$为来自此总体的样本,则有

(i) $\overline{X}\sim N\left(\mu,\dfrac{\sigma^2}{n}\right)$;

(ii) $\dfrac{nS_n^2}{\sigma^2}\sim\chi^2(n-1)$;

(iii) $\overline{X}$ 与 S_n^2 独立.

(2)设总体 $X\sim N(\mu,\sigma^2)$,$(X_1,X_2,\cdots,X_n)$为来自此总体的样本,则有

$$\frac{\overline{X}-\mu}{S_n}\sqrt{n-1}\sim t(n-1).$$

(3)设有两个总体 $X\sim N(\mu_1,\sigma_1^2)$,$Y\sim N(\mu_2,\sigma_2^2)$,又设$(X_1,X_2,\cdots,X_{n_1})$和$(Y_1,Y_2,\cdots,Y_{n_2})$为分别来自 X 和 Y 的两个独立样本,则有

$$\frac{\dfrac{n_1S_{n_1}^2}{\sigma_1^2}\Big/n_1-1}{\dfrac{n_2S_{n_2}^2}{\sigma_2^2}\Big/n_2-1}\sim F(n_1-1,n_2-1).$$

当 $\sigma_1^2=\sigma_2^2$ 时,又有

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{n_1S_{n_1}^2+n_2S_{n_2}^2}}\sqrt{\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}\sim t(n_1+n_2-2),$$

其中 $S_{n_1}^2=\dfrac{1}{n_1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2$,$S_{n_2}^2=\dfrac{1}{n_2}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2$.

§1.1 基本概念

问1 什么是简单随机样本?怎样抽样可得到简单随机样本?

答　设$(X_1, X_2, \cdots, X_n)$是来自总体 X 的样本，如果它满足以下两个条件，则称它为简单随机样本：

(1)　$X_1, X_2, \cdots, X_n$ 与 X 同分布；

(2)　$X_1, X_2, \cdots, X_n$ 彼此独立.

由于简单随机样本的分量具有独立同分布的特性，使利用样本或统计量对总体统计特性进行统计分析变得简单化. 因为此时，样本分布与总体分布有着简单而直接联系：

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

其中 F_n 和 F 分别是样本和总体的分布函数.

例如，当总体 $X \sim N(\mu, \sigma^2)$时，便可立即写出样本$(X_1, X_2, \cdots, X_n)$的分布密度

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

便可毫不费气力地在点估计问题中求出 μ 和 σ^2 的最大似然估计，也可以在假设检验问题中求得检验统计量 $T = \dfrac{\overline{X}-\mu_0}{S_n}\sqrt{n-1}$ 及 $\chi^2 = \dfrac{(n-1)S_n^2}{\sigma_0^2}$ 等等的分布，以及使其他统计分析问题变得简单和可行.

对总体进行随机地独立重复观测便可以获得简单随机样本. 这里抽样的随机性是指对总体的每一个个体有相同的机会被抽取，因为这样可使样本对总体具有代表性. 例如，为了考查某车间在某一天内所生产的滚珠的直径是否符合规格，需要进行抽样分析. 现从车间中的某个熟练工人所生产的滚珠中抽取 n 个测其直径，便可得一个容量为 n 的样本$(X_1, X_2, \cdots, X_n)$. 但是它不是此车间所生产滚珠直径这个总体的简单随机样本. 因为它不能反映整个车间所生产滚珠直径的统计特性. 如设整个车间生产滚珠直径 $X \sim N(\mu, \sigma^2)$，而此熟练工人生产滚珠的直径为 $X_1 \sim N(\mu_1, \sigma_1^2)$. 由于工人技术熟练，产品质量稳定，一般应有 $\sigma_1^2 < \sigma^2$，可见 X 与 X_1 的

分布不同，即此时的样本对 X 而言不具有代表性. 再如，从含有 10 个次品的 100 个产品中，抽取容量为 10 的样本，当我们抽取样本时保证 100 个产品中的每一个抽取的机会相同，这样得到的样本中所含次品的比例便可约为 1/10. 其统计特性便可真实反映总体的统计特性.

抽样的独立性是指每次抽样结果发生的可能性程度不受其他次抽样结果的影响. 对于从一个有限总体(即由有限个个体组成的总体)，如 N 个产品中，抽取样本时，随机有放回地抽取便可以保证样本分量的独立性，而当总体所含个体数目 N 比较大而所抽取样本的容量 n 相对比较小时(实践中通常要求 $N:n\geqslant 10:1$)，尽管抽取是无放回的，所得样本的分量也可以近似看成是相互独立的. 例如，一批产品 30 个，其中有 15 个一级品，现从中随机地有放回地取出 20 个，则其中恰好有 10 个一级品的概率为

$$p_1 = C_{20}^{10}\left(\frac{1}{2}\right)^{20} \approx 0.176$$

若是无放回地取出 20 个，则其中恰有 10 个一级品的概率却是

$$p_2 = \frac{C_{20}^{10}A_{15}^{10}A_{15}^{10}}{A_{30}^{20}} \approx 0.3.$$

可见，当样本容量与总体中个体数目相差不大时，得到的结果有明显的不同. 如果在此例中，抽出样本容量为 2，在有放回抽取情况下，其中恰好含有一个一级品的概率为

$$p_1' = C_2^1\left(\frac{1}{2}\right)^2 = 0.5,$$

而无放回地抽取时，2 个产品中恰好含有一个一级品的概率为

$$p_2' = C_2^1\frac{A_{15}^1A_{15}^1}{A_{30}^2} \approx 0.517.$$

可见，所得结果的差别比前者要小. 由概率论的知识可知，对于固定的样本容量，当总体所含个体数目趋于无穷时，有放回抽取和无放回抽出所得到样本的分布将趋于一致.

问 2 总体的概念应当如何理解？

答 所谓总体是指所有研究对象的全体构成的集合. 总体中的每个元素称为个体.

例 1 如果研究的对象是某个工厂在一天中所生产的所有1000 个灯泡,那么这批灯泡便组成了一个总体. 而每个灯泡便是此总体的一个个体.

然而,在实际应用中我们关心的往往是所研究对象的某个数量指标,如例 1 中灯泡的寿命. 即便有时使我们感兴趣的研究对象不是某个数量指标,也可以适当定义一个"数量指标"来表示它.

例 2 在考查一万件产品时,我们关心的是产品中的正品和次品情况. 此时,就可以用两个不同的数,如 0 和 1 分别表示正品和次品,数字 0 和 1 便是代表每件产品的"数量指标".

于是,可将研究对象的数量指标的全体称为总体(为了区别,不妨将前者称为实际总体,后者称为指标总体). 如,在例 1 中1000 个灯泡的寿命所构成的集合便是一个指标总体. 在例 2 中,代表 1 万件产品的正品和次品的 0 和 1 组成的集合也是一个指标总体.

然而,对于指标总体我们往往不能得到总体中的每一个个体. 如,在例 1 中,若将所有 1000 支灯泡的寿命测得,那么这一天生产的所有灯泡也就全部报废了. 在例 2 中将一万件产品逐一进行检查也是不现实的. 实际上,总体内的个体的分布通常有一定规律,因此对于个体具有分布规律的指标总体中的每一个个体可以看成是一个具有相应概率分布随机变量的一个可能取值. 此随机变量的分布可以完整描述总体内个体的分布规律. 于是了解了此随机变量的分布也就了解了相应的总体. 在以后的统计推断与分析中将这样的随机变量称为总体. 常记为 X. 其分布称为总体分布. 有时也将其分布称为总体. 例如,在例 1 中,总体便是这样一个随机变量,它表示任取一个灯泡的寿命,而工厂一天中生产的 1000 只灯泡的寿命便是它的可能取值. 如果知道了它的分布,就可知道这些灯泡寿命的分布情况及其统计特性. 在例 2 中,总体可定义为

$$X=\begin{cases}1,\text{任取一产品为次品;}\\0,\text{任取一产品为正品,}\end{cases}$$

其分布为二点分布 $B(1,p)$，只要知道了概率 p，对于一万个产品中的正品与次品情况就有所了解了。

问3 什么是统计量？为什么要引进统计量？为什么要求统计量中不含任何未知参数？统计量的分布是否也必不含未知参数？

答 (一)所谓统计量是指不含任何未知参数的样本$(X_1,X_2,\cdots,X_n)$的函数 $T=T(X_1,X_2,\cdots,X_n)$.

例1 设总体 $X\sim U\left[\frac{1}{2}-\theta,\frac{1}{2}+\theta\right]$，其中 $\theta>0$ 为未知参数，又设$(X_1,X_2,\cdots,X_n)$为来自此总体的样本. 试判断样本函数 $S_i^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_k)^2$是否为统计量，其中

(1)$\mu_1=\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$(样本均值)；

(2)$\mu_2=\begin{cases}X\left(\frac{n+1}{2}\right), & \text{当 } n \text{ 为奇数时(样本中位数)；}\\ \frac{1}{2}\left[X\left(\frac{n}{2}\right)+X\left(\frac{n}{2}+1\right)\right], & \text{当 } n \text{ 为偶数时.}\end{cases}$

(3)$\mu_3=E(X)$(总体期望)；

(4)$\mu_4=\frac{b_3}{\sigma^3}$(总体偏度)，其中 $b_3=E[X-E(X)]^3$，$\sigma=\sqrt{D(X)}$.

解 μ_1,μ_2 显然是统计量.

μ_3 也是统计量. 因为 $\mu_3=E(X)=\frac{\left(\theta+\frac{1}{2}\right)+\left(\frac{1}{2}-\theta\right)}{2}=\frac{1}{2}$，于是，$S_3^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\frac{1}{2}\right)^2$ 不含未知参数.

μ_4 不是统计量，因为

$$D(X)=\frac{\left(\frac{1}{2}+\theta-\frac{1}{2}+\theta\right)^2}{12}=\frac{\theta^2}{3},$$

$$\sigma^3=(\sqrt{D(X)})^3=\frac{\theta^3}{3\sqrt{3}},$$

$$b_3=E\left(X-\frac{1}{2}\right)^3=\int_{\theta-\frac{1}{2}}^{\theta+\frac{1}{2}}\left(x-\frac{1}{2}\right)^3\frac{1}{2\theta}\mathrm{d}x$$

$$=\frac{1}{8\theta}[\theta^4-(\theta-1)^4],$$

所以

$$\mu_4=\frac{b_3}{\sigma_3}=\frac{3\sqrt{3}}{8\theta^4}[\theta^4-(\theta-1)^4],$$

从而 $S_4^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_4)^2=\frac{1}{n}\sum_{i=1}^{n}\left\{X_i-\frac{3\sqrt{3}}{8\theta^4}[\theta^4-(\theta-1)^4]\right\}^2$ 含有未知参数 θ，故它不是统计量.

（二）引进统计量的目的是为了将杂乱无章的样本值整理成便于对所研究问题进行统计推断、分析的形式. 将样本中所含的有关所研究问题的信息集中起来，从而更有效地揭示出问题的实质，进而得到解决问题的办法. 如，为了估计总体的期望值 μ，可将样本中关于总体取值平均值的信息集中起来，这一信息便可集中体现在样本分量 $X_1,X_2,\cdots,X_n$ 的算术平均值 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 上. 因为若总体期望比较大时，取自总体的观测值的平均值自然也应有偏大倾向；反之也将有偏小倾向. 这样就比较清楚地提出了估计 μ 的办法，而若直接考虑样本就显得没有头绪. 再如，对于正态总体 $N(\mu,\sigma^2)$ 中的假设 $H_0:\mu=\mu_0$ 检验问题中，统计量 $T=\frac{\overline{X}-\mu_0}{S_n}\sqrt{n-1}$ 便集中了关于假设是否成立的信息，它揭示了问题的实质. 不难看出，它的大小反映了 H_0 的成立与否. 从而也就找到了解决问题的办法.

此外，虽然样本中含有关于所研究问题的全部信息，除了上述它没有将有用信息集中起来以外，直接用它——一个 n 维的随机向量进行统计推断和分析总是没有使用适当统计量——一个一维

的随机变量简单.

当然，选择的统计量应较好地集中样本中所含的关于所研究问题的信息，而不过多丢失有用信息.

(三)在(二)中已经指出，统计量的使用目的在于对所研究的问题进行统计推断和分析. 如，用统计量对未知参数进行估计时，若统计量本身仍含有未知参数，那么就无法根据所测得的样本值求得未知参数的估计值. 利用统计量估计未知参数将失去意义. 再如，在假设检验中，若检验统计量中含有未知参数，那么由样本值就无法求出相应的检验统计量的值，也就无法与相应的临界值进行比较，从而使得通过统计量表示的拒绝域失去意义. 总之，从统计量的意义上看，要求它不含未知参数是自然的.

(四)统计量本身虽然不含未知参数，但是它的分布却可能含有未知参数. 如，对正态总体 $N(\mu,\sigma^2)$，其中 μ 和 σ^2 为未知参数，则统计量 $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$，可见其分布中却含有未知参数 μ 和 σ^2. 然而，含有未知参数的样本函数其分布却不一定含有未知参数. 如，在上例中，含有未知参数 μ 和 σ^2 的样本函数 $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ 却服从不含任何未知参数的标准正态分布 $N(0,1)$. 再如样本函数 $\frac{nS_n^2}{\sigma^2}\sim\chi^2(n-1)$.

问 4 什么叫大样本和小样本？它们之间的区别是否是以样本容量大小来区分的？

答 在样本容量固定条件下，进行的统计推断、分析问题称为**小样本问题**，而在样本容量趋于无穷的条件下，进行的统计推断、分析问题称为**大样本问题**.

然而，众多统计推断与分析问题与统计量或样本的函数的分布相关联. 能否得到有关统计量或样本的函数的分布常成为解决问题的关键. 所以大、小样本的区分常与这一分布能否得到相联系.

对于固定的样本容量，如果能得到有关统计量或样本函数的精确分布，相应统计推断、分析问题通常便属于小样本问题. 此时，在样本容量有限情况下，能够较精确、令人满意地讨论各种统计推断与分析问题.

如，设总体 $X\sim N(\mu,\sigma^2)$，其中 μ、σ^2 未知，要求未知参数 μ 在一定置信水平下的置信区间，需要考虑样本函数 $T=\dfrac{\overline{X}-\mu}{S_n}\sqrt{n-1}$ 的分布. 而在样本容量 n 固定时，不难证明 $T\sim t(n-1)$. 此问题便属于小样本问题.

但是，在一般情况下要确定一个统计量或样本函数的精确分布不是一件容易的事. 仅在一些特殊情况下的特殊统计量(如正态总体下的某些常见统计量)才能求得它们的精确分布. 如果统计量或样本函数的精确分布求不出或者其表达式过于复杂而难于应用时，如能求出在样本容量趋于无穷时的极限分布，利用此极限分布作为其近似分布进行统计推断、分析，此类问题便属于大样本问题.

如，若 X 是具有二阶矩的非正态分布的总体，$(X_1,X_2,\cdots,X_n)$ 是来自此总体样本. 此时，当样本容量 n 固定时，统计量 $\overline{X}$ 的精确分布往往不易求得. 然而，可以证明，当 n 足够大时，$\overline{X}$ 近似服从 $N\left(\mu,\dfrac{\sigma^2}{n}\right)$ 分布. 于是此时依据 $\overline{X}$ 所做的统计推断、分析问题便属于大样本问题.

大样本与小样本决不可以以样本容量的大和小来区分. 样本容量的大小受多种因素的影响. 有时虽属小样本问题，但要求的样本容量却可能比较大；反之对某些大样本问题，有可能要求其样本容量却不大.

例 设总体 $X\sim N(\mu,\sigma^2)$，其中 μ 未知，σ^2 已知. 若用样本均值 $\overline{X}$ 代替总体期望 μ 的误差不少于 10%的概率不大于 5%，试问样本容量 n 应取多大？

解 设来自此总体的样本为 $(X_1,X_2,\cdots,X_n)$. 由于样本均值

$\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$，所以有

$$P\left\{|\overline{X}-\mu|<\frac{\sigma}{\sqrt{n}}x\right\}=\Phi(x)-\Phi(-x)$$
$$=2\Phi(x)-1.$$

按题意，令 $2\Phi(x)-1\geqslant 0.95$，解得 $x\geqslant 1.96$.

再令

$$0.1=\frac{\sigma}{\sqrt{n}}x\geqslant\frac{1.96\sigma}{\sqrt{n}},$$

解得 $n\geqslant 384.16\sigma^2$.

可见，对此小样本问题，当方差 σ^2 比较大时，所要求的样本容量可以是很大的.

问 5 经验分布函数是否就是分布函数？

答 所谓经验分布函数是指当顺序统计量为

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

时的函数

$$F_n^*(x)=\begin{cases}0, & x\leqslant X_{(1)};\\ \dfrac{k}{n}, & X_{(k)}<x\leqslant X_{(k+1)},k=1,2,\cdots,n-1;\\ 1, & x>X_{(n)}.\end{cases}$$

可见，它既是实数 x 的函数，又是顺序统计量 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 的函数，它具有“两重性”，解释如下.

(1)当样本 $(X_1,X_2,\cdots,X_n)$ 取定一样本值 $(x_1,x_2,\cdots,x_n)$ 后，经验分布函数便是一个分布函数. 事实上，此时

$$F_n^*(x)=\begin{cases}0, & x\leqslant x_{(1)};\\ \dfrac{k}{n}, & x_{(k)}<x\leqslant X_{(k+1)},k=1,2,\cdots,n-1;\\ 1, & x>x_{(n)}.\end{cases}$$

不难看出，它具有三条性质：(i)单调不减性，即当 $x_1<x_2$ 时，$F_n^*(x_1)\leqslant F_n^*(x_2)$；(ii)有界性，即 $0\leqslant F_n^*(x)\leqslant 1$ 且 $F_n^*(-\infty)=0$，