第四章　统计抽样与抽样分布

本章为推断性统计学的基础章节，将系统介绍统计抽样的基本概念以及整个推断性统计学中所涉及的几种与正态分布有关的概率分布。

4.1　关于抽样的基本概念

4.1.1 为什么要抽样

为了收集必要的资料，对所研究的对象（总体）的全部元素逐一进行观测，往往不很现实。一种情形是研究的总体元素非常多，搜集数据费时，费用大，不及时而使所得的数据无意义（如在质量检验中，全部检查使废品数量又增加了许多）。另一种情形是检查具有破坏性，如对炮弹、灯管、砖的检查等，因此必须进行抽样。

4.1.2 简单随机抽样

不同的抽样方式，样本与总体的关系不一样，构成不同的抽样技术，本书全部都是指简单随机抽样。

首先介绍一下有关样本随机性的知识。把总体看成随机变量 X，对其进行 n 次观测，得到一个容量为 n 的样本：

$$x_1^{(1)},\quad x_2^{(1)},\ldots\ldots,x_n^{(1)}$$

如另作n次观测，则会得到由不同的观测结果$x_1^{(2)},x_2^{(2)},\ldots\ldots,x_n^{(2)}$所组成第二个样本。如继续下去，会得到很多不同的样本，从容量为N的总体中抽取容量为n的样本，则有 C_N^n 个。

尽管我们实际中只抽取一个样本，但是在观测之前，样本的出现具有随机性。因此，样本的每一个观测值，例如第一个观测值，在观测之前就是一个随机变量，记作X_1，观测得到它的取值记作x_1，第二个元素，第三个元素依次类推。所以一个容量为n的样本，在观测之前，就是一个n维向量，即（x_1, x_2,……, x_n）。

简单随机抽样是指这 n 个随机变量组成样本时，要具备以下两个条件：

① 这 n 个随机变量与总体 X 具有相同的概率分布；

② 它们之间相互独立。

4.1.3 样本统计量与抽样分布

前面采取的简单随机抽样，样本具有随机性，样本的随机数 $\bar{x}$ ，s^2等也会随着样本不同而不同，故它们是样本的函数。记为g（x_1， x_2,……, x_n）称为样本统计量。

统计量的概率分布称为抽样分布（Sample distribution）

4.2 几种与正态分布有关的概率分布

通常我们把总体看作是一个随机变量X，有它自身的分布，（大多数均视为正态分布），其分布中有参数，这些参数往往与总体特征数有关，正态分布有两个参数，μ，σ^2，其中μ就是X的期望，σ^2就是X的方差。所以我们常把总体的特征数叫做总体参数。这些总体特征数不宜直接求出，由于样本是总体的一部分，故可根据样本统计量的信息推断总体参数。为了介绍总统参数的推断，这里先来介绍几个与正态分布有关的概率分布。

4.2.1 正态分布

1. 若随机变量 X 的概率密度函数为：

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\qquad -\infty<x<+\infty \tag{4-1}$$

记为 $X\sim N(\mu,\sigma^2)$

$$P(X\le x)=\int_{-\infty}^{x}f(t)\mathrm{d}t \tag{4-2}$$

当 $\mu=0,\sigma^2=1$时，$\varphi(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$

记为 $\mu\sim N(0,1)$

令 $U=\frac{X-\mu}{\sigma}$，$P(U\le u)=\int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\mathrm{d}t=\phi(u)$

标准正态分布概率密度函数如图 4-1 所示：

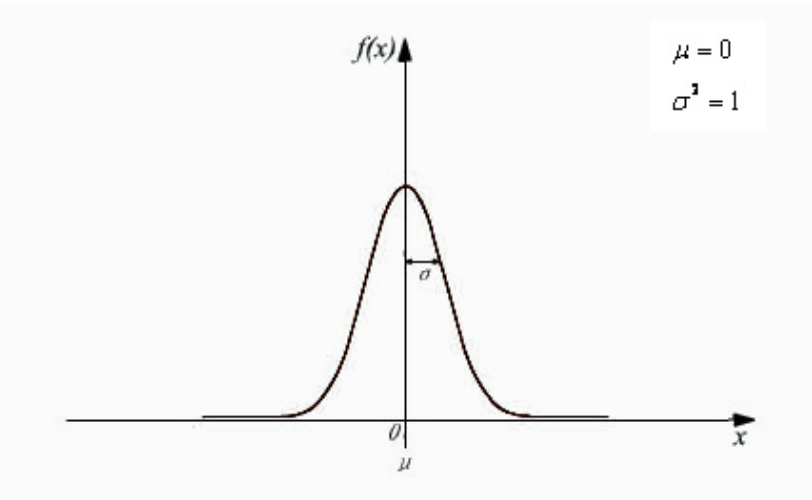

图 4-1 标准正态分布概率密度函数

2. 查表

当 *u* 大于零时，可查正态分布表，但如果 *u<0* 时，则可由下式 *φ（-u）=1-φ（u）*求出。

若求当$x_1 \le X \le x_2$时的概率，可由下面的推导得到：

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(t)dt = \int_{x_1}^{x_2} \frac{1}{o\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{dx}$$

$$\xrightarrow{U_1=\frac{x_1-\mu}{\sigma}, U_2=\frac{x_2-\mu}{\sigma}} \int_{u_1}^{u_2} \frac{1}{\sqrt{2\pi}} e^{\frac{u^2}{2}} \mathrm{d}u = \phi(U_2) - \phi(U_1)$$

$$= \phi(\frac{x_2-\mu}{\sigma}) - \phi(\frac{x_1-\mu}{\sigma})$$

见图 4-2 和图 4-3

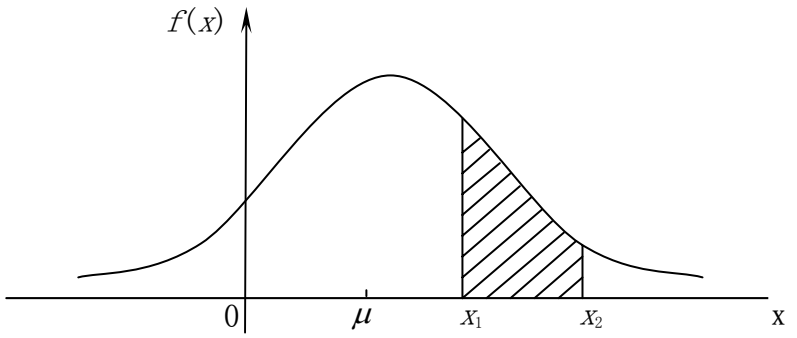

图 4-2 正态函数概率密度计算示意图

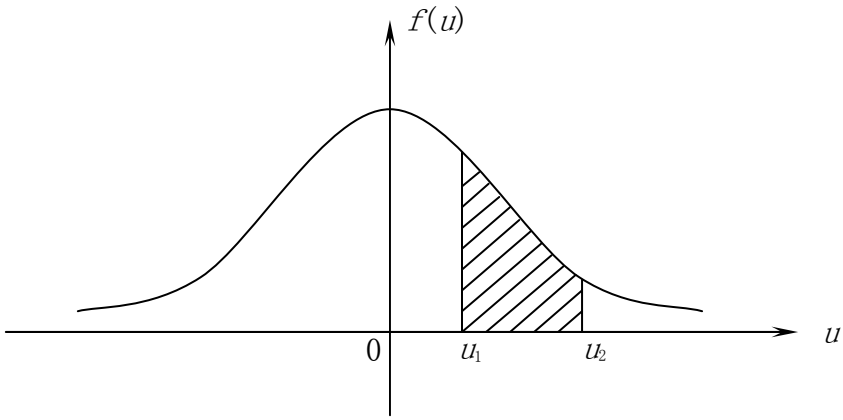

图 4-3 标准正态函数概率密度计算示意图

3. 正态分布的线性性质：

如果 $X_i(i=1,2,\cdots,n)$ 服从正态分布，$X_i \sim N(\mu_i,\sigma_i^2)$，且相互独立。对于常数 a_i，有下式成立：

(1) $\sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n}\mu_i,\sum_{i=1}^{n}\sigma_i^2)$

(2) $aX_i \sim N(a\mu_i,a^2\sigma_i^2)$

可以看出，正态随机变量的线性组合仍然为正态随机变量。

4.2.2 χ^2 分布

1. 定义：$x_1,x_2,\cdots,x_n$ 是相互独立且服从 N（0，1）分布的随机变量，则称随机变量 $\chi^2=\sum_{i=1}^{n}x_i^2$ 所服从的分布是自由度为 n 的 χ^2 分布，且记 $\chi^2 \sim \chi^2(n)$，其概率密度函数为：

$$f(x;n)=\begin{cases} A_n e^{-\frac{x}{2}}x^{\frac{n}{2}-1}, & x>0 \\ 0, & ,x\le 0 \end{cases} \qquad (4\text{-}3)$$

其中，A_n 是仅与 n 有关的常数。f（x）的图形随 n 的不同而不同，如图 4-4 所示。

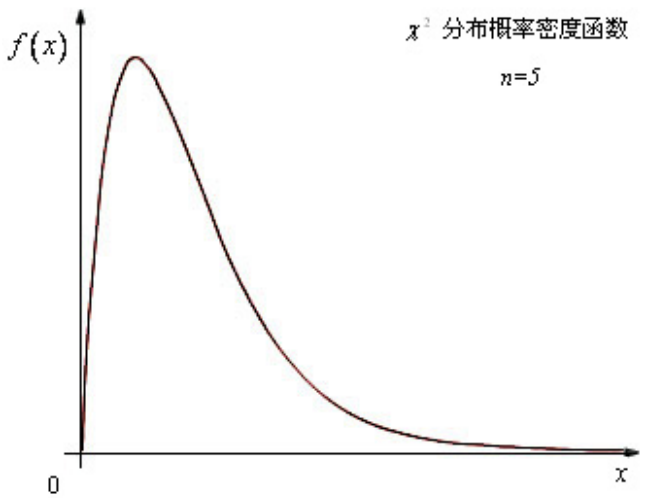

图 4-4 χ^2 分布概率密度函数

2. χ^2 分布的随机变量的期望与方差为：

$$E(\chi^2)=n \tag{4-4a}$$

$$D(\chi^2)=2n \tag{4-4b}$$

3. 查表：对于给定的 α，0<α<1，可在 χ^2 分布表中查得，即

$$P\{\chi^2(n)>\chi_\alpha^2\}=\int_{x_\alpha^2}^{+\infty} f(x,n)\mathrm{d}x=\alpha \tag{4-5}$$

例如：$\chi^2_{0.1}(10)=15.987$，即指

$P\{\chi^2(10)>15.978\}=\int_{15.978}^{+\infty} f(x;n)\mathrm{d}x=0.1$，见图 4-5。

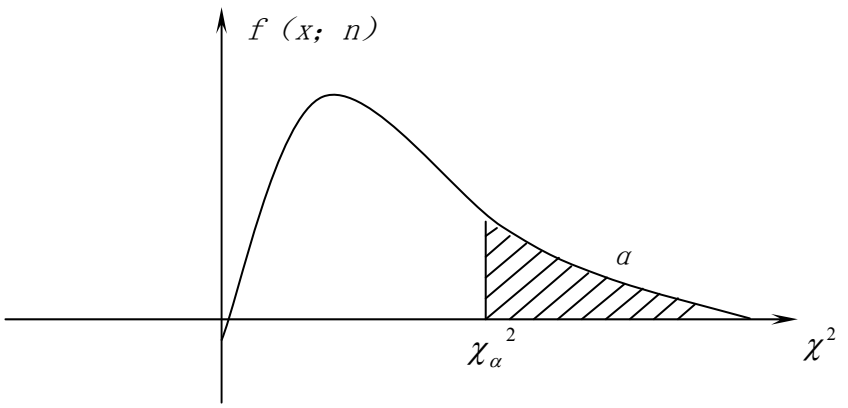

图 4-5 χ^2 分布函数概率密度计算示意图

4. χ^2 分布的性质：

① 如果 $X \sim N(0,1)$，则 $X^2 \sim \chi^2(1)$

② 设 $\chi_1^2 \sim \chi^2(n_1), \chi_2^2 \sim \chi^2(n_2)$ 且相互独立，则 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$

③ 若 $\chi_3^2 = \chi_1^2 + \chi_2^2$，已知 χ_1^2 ,χ_2^2 相互独立 $\chi_1^2 \sim \chi^2(n_1), \chi_3^2 \sim \chi^2(n)$，则

$\chi_2^2 \sim \chi^2(n - n_1)$

④ 总体 $X \sim N(\mu, \sigma^2)$，$x_1, x_2, \cdots, x_n$ 是 X 的一个样本，$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

样本的平均数，$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ 为样本的方差。则

a. $\bar{x}$与s^2 相互独立

b. $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

4.2.3 F 分布

1. 定义：设相互独立的随机变量V和W分别服从自由度为 n_1, n_2 的 χ^2 分布，

即$V \sim \chi^2(n_1), W \sim \chi^2(n_2)$，则随机变量$F = \frac{V/n_1}{W/n_2}$所服从的分布为F分布。$n_1$，$n_2$分别是它的第一自由度和第二自由度，且通常记为$F \sim F(n_1, n_2)$。其概率密度函数如下，如图 4-6 所示：

$$f(x;n_1,n_2)=\begin{cases} B(n_1,n_2)\cdot\dfrac{x^{\frac{n_1}{2}-1}}{(1+\frac{n_1 x}{n_2})^{\frac{n_1+n_2}{2}}} & x\geq 0 \\ 0 & x<0 \end{cases} \qquad (4\text{-}6)$$

其中$B(n_1,n_2)=\dfrac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}\cdot(\frac{n_1}{n_2})^{\frac{n_1}{2}}$。

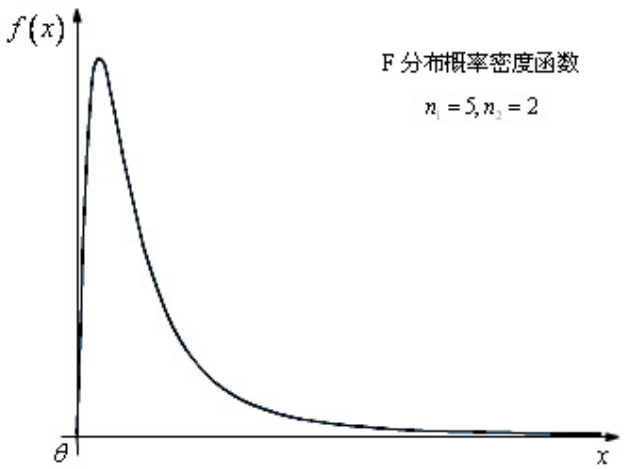

图 4-6 F 分布概率密度函数

2. F 分布的期望与方差：

$$E(F)=\frac{n_2}{n_2-2} \qquad (n_2>2) \qquad (4\text{-}7a)$$

$$D(F)=\frac{n_2^2}{(n_2-2)^2}\cdot\frac{2(n_1+n_2-2)}{n_1(n_2-4)} \qquad (n_2>4) \qquad (4\text{-}7b)$$

3. 查表：$P(F > F_\alpha) = \int_{F_\alpha}^{\infty} f(x)\mathrm{d}x = \alpha \quad (0 < \alpha < 1)$。

4. 性质：

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}$$

F 分布表给出了 F 分布的上侧 100α 百分位数，表中没有列出的某些值可利用上面提到的性质求出。

4.2.4 t 分布 （Students 分布）

1. 定义：设随机变量 U 服从标准正态分布，随机变量 W 服从自由度为 n 的 χ^2 分布，且 U 与 W 相互独立，则称随机变量 $T = \frac{U}{\sqrt{W/n}}$ 所服从的分布为自由度为 n 的 t 分布，且记 T~t（n）。

2. t 分布的概率密度为：

$$f(t) = C_n\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad \left(-\infty < t < +\infty\right) \tag{4-8}$$

$$C_n = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}。$$

$f(t)$ 的图形随自由度 n 不同而不同，与正态分布的形状相似，如图 4-7 显示了 $n=1$， $n=10$ 及 $n=\infty$ 时的 t 分布图。

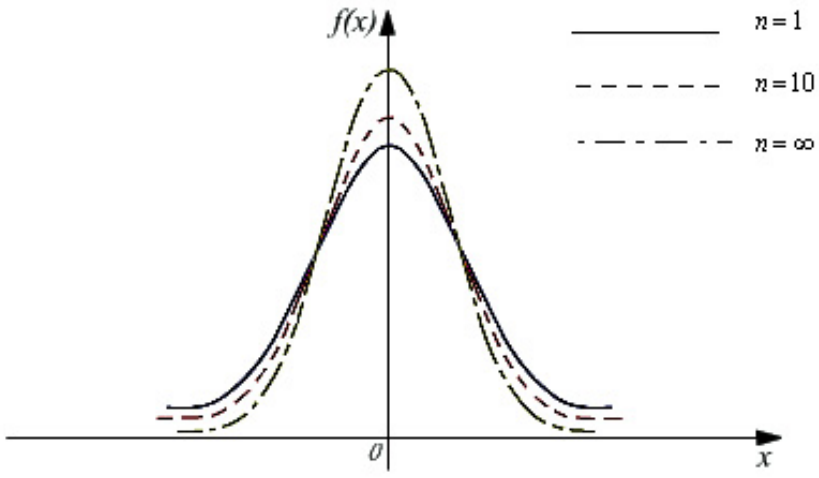

图 4-7 自由度分别为 $n=1$， $n=10$ 及 $n=\infty$ 时
t 分布概率密度函数

3. 查表： $P\{|t|>t_{\alpha/2}(n)\}=\alpha$ 或 $P\{t>t_{\alpha}(n)\}=\alpha$ 。

4. 性质：

当n很大时， $\lim\limits_{n\to\infty} f(t)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ 此时，$t_{\alpha/2}\approx u_{\alpha/2}$，t分布近似标准正态分布。

5. t 分布的期望与方差：

对于 t~t（n），

E（t）=0，D（t）=n/n-2 （4-9）

其中：n>2。

4.3 样本平均数的抽样分布

设总体 $x\sim N(\mu,\sigma^2)$， $x_1,x_2,\cdots,x_n$ 是总体 $\boldsymbol{X}$ 的随机样本，样本平均数

为 $\overline{X}=\sum x_i/n$，则容易推出 $\overline{X}$ 抽样分布的均值和方差为：

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

证明：

$$E(\overline{X})=E(\frac{x_1+x_2+\cdots+x_n}{n})=\frac{1}{n}E(x_1+x_2+\cdots+x_n)$$

$$=\frac{1}{n}\left[E(x_1)+E(x_2)+\cdots+E(x_n)\right]=\frac{1}{n}\cdot n\mu=\mu$$

$$D(\overline{X})=D(\frac{x_1+x_2+\cdots+x_n}{n})=\frac{1}{n^2}D(x_1+x_2+\cdots+x_n)$$

$$=\frac{1}{n^2}\cdot n\sigma^2=\frac{\sigma^2}{n}$$

当 ***X*** 不服从正态分布时，根据中心极限定理，$\overline{X}$ 随 *n* 的增加而近似正态分布，即对于足够大的 *n*，有

$$P\{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\le u\}\to\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}e^{-\frac{x^2}{2}}\mathrm{d}x$$

上述的关于均值和方差的公式以及中心极限定理都是对无限总体而言的。

但对于有限总体若采取有放回抽样，则与无限总体等价。若有限总体容量为 N 而采取无放回抽样，且 n/N≤0.1，仍可视为无限总体，而当 n/N>0.1 时则

$$E(\overline{X})=\mu\qquad D(\overline{X})=\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}\qquad\sigma(\overline{X})=\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

称式 $\sqrt{\frac{N-n}{N-1}}$ 为有限总体的修正系数。

4.4 中心极限定理

确定 $\overline{x}$ 抽样分布特征的最后一步是确定 $\overline{x}$ 概率分布的形式。我们考虑

两种情形：一种是总体分布未知，另一种为已知总体分布为正态分布。总体分布未知时，我们依赖于统计学中最重要的定理之一 —— 中心极限定理。中心极限定理在抽样分布中的应用如下：

中心极限定理

从总体中抽取样本容量为 n 的简单随机样本，当样本容量很大时，样本均值 x 的抽样分布可用正态概率分布近似。

图 4-8 说明对于三个不同总体中心极限定理的作用。在每种情形下，显然总体是非正态的。然而，我们注意到随着样本容量的增加，$\overline{x}$ 抽样分布开始发生变化。例如，当样本客量为 2 时，我们看到$\overline{x}$ 抽样分布开始呈现与总体分布不同的外形；样本容量为 5 时，三个抽样分布都开始呈现—种钟形外形；最后，当样本容量为 30 时，三个抽样分布近似于一种正态。因而，当样本容量足够大时，$\overline{x}$ 抽样分布可用正态概率分布近似。但是，样本容量应该达到多大时，我们才可以假定能够使用中心极限定理呢？统计研究人员通过研究各种总体不同样本容量下$\overline{x}$ 的抽样分布，来研究该问题。当总体分布是对称坡形形状时，样本容量为 5 到 10 时即可适用中心极限定理。然而，如果总体分布严重偏态或明显非正态，则需要更大的样本容量。通常在统计实践中，假定对多数应用，当样本容量大于等于 30 时，$\overline{x}$ 的抽样分布可用正态概率分布近似。实际上，样本容量为 30 或更多时，即假定满足中心极限定理大样本条件。这一结果非常重要，我们再次重申一下，当样本容量很大的时候，$\overline{x}$ 的抽样分布可用正态概率分布来近似。大样本的条件可假定为简单随机样本样本容量为 30 或更多。当总体分布未知时，中心极限定理是确定$\overline{x}$ 抽样分布形式的关键。然而，我们也可能遇到这样一些假定或认为总体是正态概率分布的抽样情形。在这种情形下，下面的结果定义了$\overline{x}$ 抽样分布的形式：

当总体为正态概率分布时，对任何样本容量，$\overline{x}$ 的抽样分布均为正态分布。

总之，若我们用一个大的简单随机样本（n>=30）时，中心极限定理使我们可以用正态概率分布近似$\overline{x}$ 的抽样分布。在简单随机样本是小样本（n＜30 时，仅当我们假定总体为正态概率分布时，$\overline{x}$ 的抽样分布才为正态的。

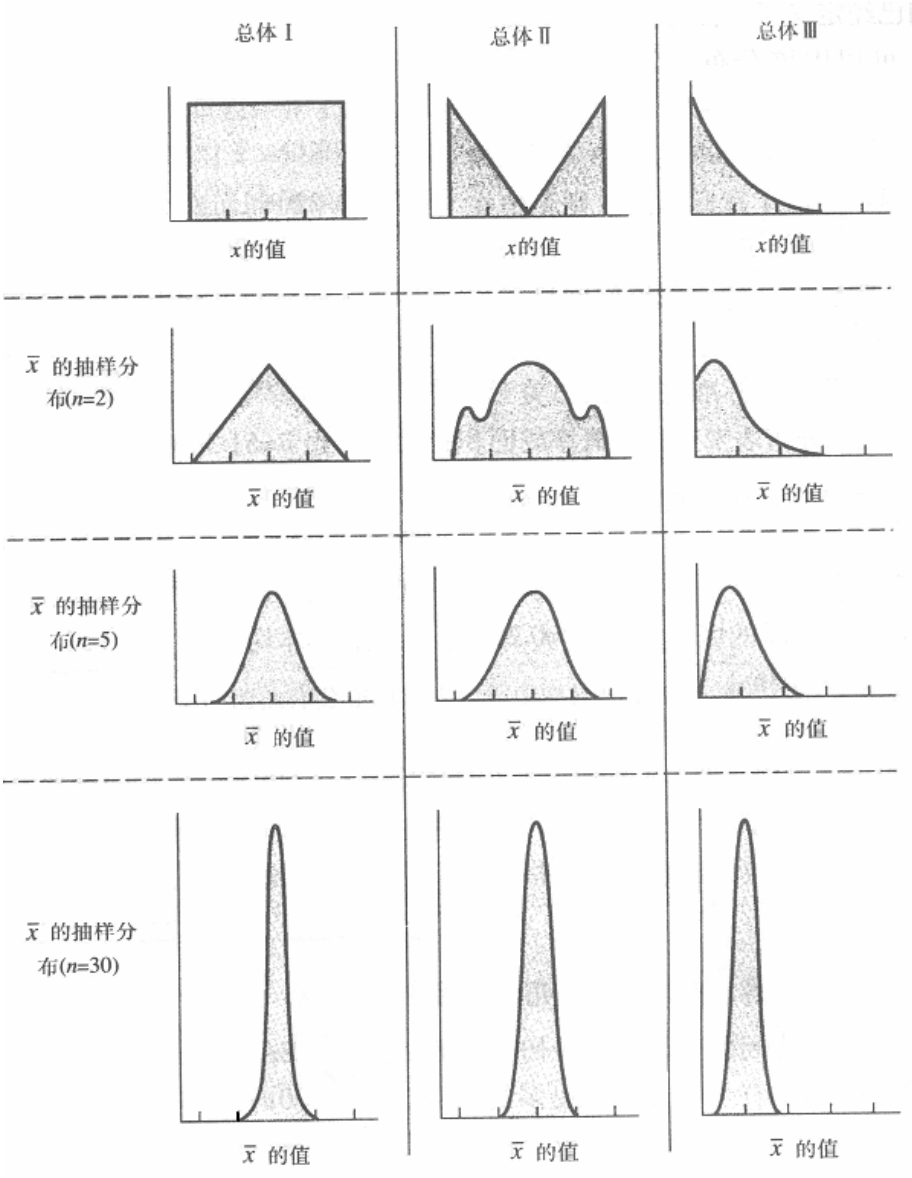

图 4-8 中心极限定理在三个不同总体中的作用

习题

1、订阅者阅读《青年报》的平均时间是 49 分钟，假定标准差是 16 分钟且时间呈正态分布。

a. 一名订阅者至少花 1 小时读报的概率是多少？

b. 一名订阅者读报的时间不超过 30 分钟的概率是多少？

c. 10%花费最多时间读报的人的时间范围是多少？

2、高速公路巡警保留着一份有关从事故报告到官员抵达事故现场所需时间的记录资料。一个由 10 条记录组成的简单随机样本数据（以分钟为单位）如下：

126	34	48	50	68	23	36	81	25	103

a. 从事故报告到官员抵达事故现场所需时间总体平均值的点估计为多少？

b. 从事故报告到官员抵达事故现场所需时间总体标准差的点估计为多少？

3、从均值为 200，标准差为 50 的总体，抽取 n＝100 的简单随机样本，样本均值 $\bar{x}$ 用于估计总体均值。

a. $\bar{x}$ 的数学期望是多少？

b. $\bar{x}$ 的标准差是多少？

c. $\bar{x}$ 的抽样分布是什么？

d. $\bar{x}$ 的抽样分布说明什么？

4、一项家庭旅行调查表明，旅行时四口之家日平均花费为 215.60 元，假定四口之家日花销的总体均值为 215.60 元，总体标准差为 50 元。选择 40 个家庭组成一个简单随机样本进行进一步研究。

a. 说明样本均值 $\bar{x}$ 的抽样分布，其中 $\bar{x}$ 是四口之家日花销的均值。

b. 这 40 个家庭组成的简单随机样本的样本平均值在总体均值左右 20 元以内的概率是多少？

c. 这 40 个家庭组成的简单随机样本的样本平均值在总体均值左右 10 元以内的概率是多少？

5、一图书馆每天平均登记 μ＝320 本书，标准差 σ＝75 本，考虑营业的 30 天为一个样本，$\bar{x}$ 为每天登记书的数量的样本均值。

a. 说明 $\bar{x}$ 的抽样分布。

b. $\bar{x}$ 的标准差为多少？

c. 这 30 天里样本均值在 300 到 400 本的概率为多少？

d. 这 30 天里样本均值大于等于 325 本的概率为多少？