第 5 章　　数理统计的基本概念

数理统计是运用概率论的知识，研究如何有效地对带有随机性影响的数据进行收集、整理、分析和推断的学科，由于随机性现象广泛存在于工、农业生产、工程技术、自然科学和社会科学等领域中，因此数理统计有着最广泛的应用。

5.1　基本概念

1.总体和样本

在数理统计中，我们将研究对象的全体称为**总体**或**母体**，而把组成总体的每个元素称为**个体**。例如研究一批灯泡的平均寿命时，该批灯泡的全体构成了研究的总体，其中每个灯泡就是个体。

在实际问题中，研究对象往往是很具体的事物或现象，而我们所关心的不是每一个个体的种种具体的特征，而是其中某项或某几项数量指标，记为 X。在上例中，X即指该批灯泡的寿命。对不同的个体，X的取值一般是不同的。例如在试验中观察若干个个体就会得到X的一种数值但在试验或观察之前，无法确定会得到一组什么样的数值，所以 X是一个随机变量或随机向量，而 X的分布也就完全描述了我们所关心的指标，即总体的分布。为方便起见，以后我们将 X的可能取值的全体组成的集合称为总体，或直接称 X为总体，X的分布也就是总体的分布。

总体分布一般是全部或部分未知的，为了研究总体 X的分布规律，我们需要对总体进行若干次观察。由观察得到总体指标 X的一组数值$(x_1, x_2, \cdots, x_n)$，其中 x_i为第 i次观察结果，并称$(x_1, x_2, \cdots, x_n)$为总体 X的一组**容量为 n 的样本观察值**，样本观察值是对总体分布进行分析、推断的基础。这种从总体中随机地抽出若干个个体进行观察或实验，称为随机抽样观察，从总体中抽出的若干个个体称为**样本**，一般记为$(X_1, X_2, \cdots, X_n)$，而一次具体的观察结果$(x_1, x_2, \cdots, x_n)$是完全确定的一组数值，但它又随着每次抽样观察而改变。因此，容量为 n的样本（$X_1, X_2, \ldots, X_n$）是 n维随机向量，而具体的观察值（$x_1, x_2, \ldots, x_n$）是随机变量（$X_1, X_2, \ldots, X_n$）的一个样本观察值。样本（$X_1, X_2, \ldots, X_n$）所有可能取值的全体称为**样本空间**，记为$\aleph$，而样本观察值（$x_1, x_2, \ldots, x_n$）是$\aleph$中的一个样本点。

随机抽样的目的是为了对总体 X的分布进行各种分析推断，所以要求抽取的样本能很好地反映总体的特性，为此我们要求随机抽取的样本（$X_1, X_2, \ldots, X_n$）满足：

（1）具有代表性。即样本的每个分量 X_i与 X有相同的分布；

（2）具有独立性。即 $X_1, X_2, \ldots, X_n$是相互独立的随机变量，也就是说，n次观察值之间是互相独立的；

满足上述两条的样本称为简单随机样本，今后如无特别说明，所说的样本均指简单

随机样本。

例 1　对一批 N 件产品情况进行检查，从中有放回的抽取 n 件。分别以 1，0 表示某件产品为合格品和次品，以 $\theta(0 \le \theta \le 1)$ 表示产品的合格路率，则总体指标 X 服从参数为 θ 的（0-1）分布，即 $P(X=x)=\theta^{x}(1-\theta)^{1-x}, x=0,1$。这样抽取得到的观察结果为 $X_1, X_2, \ldots, X_n$ 为一个简单随机样本，也就是说 $X_1, X_2, \ldots, X_n$ 是相互独立且均服从参数为 θ 的（0-1）分布，故样本（$X_1, X_2, \ldots, X_n$）的联合分布律为

$$P(X_1=x_1, X_2=x_2, ..., X_n=x_n)=\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i},$$

$x_n=0$，1，i=1，2，…n。

每组观察值（$x_1, x_2, \ldots, x_n$）为由 0，1 组成的一个 n 维向量，其样本空间为

$\aleph$={（$x_1, x_2, \ldots, x_n$）| x_i=0，1，i=1，2，…n}。

共有 2^n 个样本点。

一般地，若总体 X 的概率密度或联合分布律为 $f(x)$，则样本($X_1, X_2, \ldots, X_n$)的联合密度或联合分布律为

$$L(x_1, x_2, ..., x_n)=\prod_{i=1}^{n} f(x_i)$$

并称 L（$x_1, x_2, \ldots, x_n$）为样本（$X_1, X_2, \ldots, X_n$）的似然函数。

对于个体为有限的总体来说，采用有放回随机抽样就能得到简单随机样本。但有放回抽样使用起来很不方便。又由于当总体的个体为无限时，有放回抽样与不放回抽样没有什么区别，因此，在实际问题中，当总体中个体数 N 很大，而样本容量 n 相应较小时，可把总体看作是无限的，从而可将不放回抽样当作有放回抽样来处理。

2.统计量和样本矩

样本是我们进行分析和推断的起点，但实际上我们往往并不直接利用样本进行推断，而需要对样本进行一番“加工”和“提炼”，将分散于样本中的信息集中起来。为此我们引进统计量的概念。

设 $X_1, X_2, \ldots, X_n$ 为来自总体 X 的一个样本，g（$X_1, X_2, \ldots, X_n$）为一个 n 元连续函数，若 g（$X_1, X_2, \ldots, X_n$）中不含任何未知参数，则称 g（$X_1, X_2, \ldots, X_n$）为一个统计量。显然统计量也是一个随机变量。以后，针对不同的问题我们总是构造相应的统计量以实现对总体的统计推断。

例如，设总体 X 服从正态分布 N（μ，σ^2）其中 μ，σ^2 未知。$X_1, X_2, \ldots, X_n$ 是从

正态总体 X 中抽取的一个样本，则 $\frac{1}{n}\sum_{i=1}^{n}X_i$，$\sum_{i=1}^{n}X_i^2$，

均是样本的统计量，而 $\frac{1}{n}\sum_{i=1}^{n}x_i-\mu$，$\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i^2$，都不是统计量.

下面介绍一类常用的统计量——样本矩。

设（$X_1, X_2, \ldots, X_n$）为一个简单随机样本，则称

$$A_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r, r = 1, 2, \cdots$$

为 ***r*阶样本原点矩**，特别地，称 A_1 为**样本均值**，并记为 $\overline{X}$，即 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

称 $$B_r=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^r(r=2,3,\cdots)$$

为 ***r*阶样本中心矩**。其中的 B_2 称为 2 阶**样本中心矩**。但为了今后的需要，我们定义样本方差如下：

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

若总体 X 的期望 $\mu=E(X)$ 和方差 $\sigma^2=D(X)$ 存在，则

$$E(\overline{X})=E(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\mu$$

$$D(\overline{X})=D(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{\sigma^2}{n}$$

$$E(S^2)=\frac{1}{n-1}E[\sum_{i=1}^{n}(X_i-\overline{X})^2]=\frac{1}{n-1}E[\sum_{i=1}^{n}X_i^2-n\overline{X}^2]$$

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

$$=\frac{1}{n-1}\sum_{i=1}^{n}\{D(X_i)+[E(X_i)]^2\}-\frac{1}{n-1}\{D(\overline{X})+[E(\overline{X})]^2\}$$

$$=\frac{1}{n-1}\sum_{i=1}^{n}(\sigma^2+\mu^2)-\frac{n}{n-1}\left(\frac{\sigma^2}{n}+\mu^2\right)=\sigma^2$$

5.2 抽样分布

统计量是我们对总体的分布规律或数字特征进行推断的基础。在使用统计量进行推断时必须要知道它的分布。在数理统计中，统计量的分布称为抽样分布，因而确定统计量的分布是数理统计的基本问题之一。下面我们介绍三类重要的分布。

1. χ^2 分布

定义1 设 $X_1, X_2, \cdots, X_n$ 相互独立且均服从标准正态分布,即 $X_i \sim N(0,1), i=1,2,\cdots,n,$ 则随机变量 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2 = \sum_{i=1}^{n} X_i^2$ 服从自由度为 n 的 χ^2 分布，记为 $\chi^2 \sim \chi^2(n)$。

这里自由度 n 是指独立变量的个数。

1).χ^2 分布具有可加性，即若 $Y_1 \sim \chi^2(n_1)$， $Y_2 \sim \chi^2(n_2)$，且 Y_1，Y_2 相互独立，则

$$Y_1+Y_2 \sim \chi^2(n_1+n_2)$$

2).当 $X_i \sim N(0,1)$，$i=1$，…，n，则 $X_i^2 \sim \chi^2(1)$

3).利用求随机变量函数的分布的方法即可求得 χ^2 分布的密度函数为

$$f(y)=\begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} & ,y>0 \\ 0 & ,y<0 \end{cases},$$

其中 $\Gamma(\frac{n}{2})$ 为 Γ 函数，其定义为 $\Gamma(\alpha)=\int_0^{+\infty} x^{\alpha-1}e^{-x}dx$

下图给出 $n=1$，4，10，20 时的 χ^2 分布的密度函数的曲线。

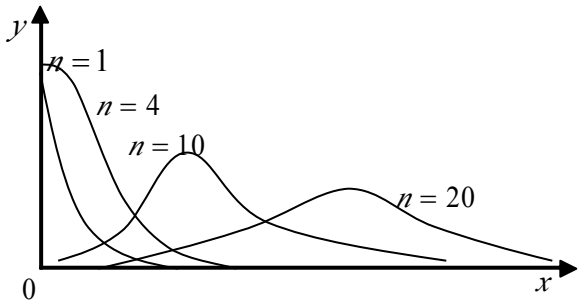

图5-1 χ^2 分布密度函数曲线

4).设 $X \sim \chi^2(n)$ 根据定义，容易验证 $E(X)=n, D(X)=2n$

5).下面介绍分布的上 α 分位点的概念，在后面将会经常用到。

定义2 设随机变量 X 的密度函数为 $f(x)$，对给定的 $\alpha(0<\alpha<1)$, 称满足条件

$$P\{X \geq x_\alpha\} = \int_{x_\alpha}^{+\infty} f(x)dx$$

的实数 x_α 为 X 的上 α 分位点.

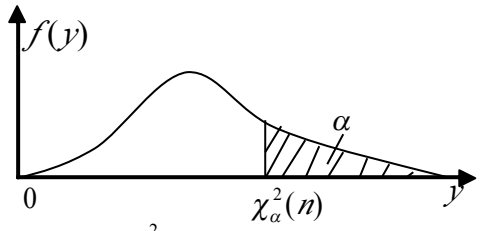

图5-2　χ^2 −分布的上 α 分位点

例如，随机变量 $\chi^2 \sim \chi^2(n)$，则称 $P\{\chi^2 > \chi_\alpha^2(n)\} = \alpha$ 的点 $\chi_\alpha^2(n)$ 为 $\chi^2(n)$ 分布的上 α 分位点，见图 6-2。x^2 分布的上 α 分位点已制成表格。如 $\alpha = 0.01, n = 10$，则查表可得 $x_{0.01}^2(n) = 23.209$，又如 $\alpha = 0.005, n = 6$，则 $x_{0.005}^2(6) = 18.548$。

若随机变量 $X \sim N(0,1)$，则它的上 α 分位点常用 Z_α 来表示。由 $P\{X > Z_\alpha\} = \alpha$ 可知，$Z_{0.005} = 1.645, Z_{0.025} = 1.96$，见图 5-3。通过查标准正态分布表即可得到。

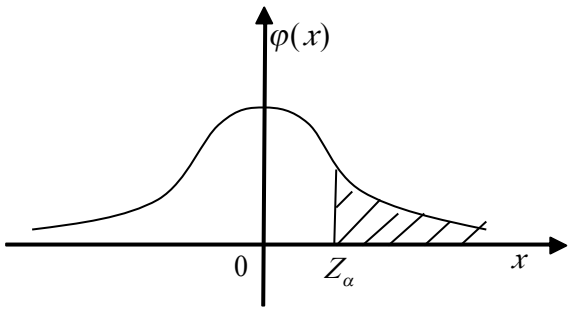

图5-3 标准正态分布的上 α 分位

这是因为 $P\{X \leq Z_\alpha\} = 1 - \alpha$，故 $P\{X \leq 1.96\} = 0.975 = 1 - 0.025$。

例 设 $X_1, X_2, \cdots, X_n$ 为来自正态总体 $N(\mu, \sigma^2)$ 的一个样本，其中 μ 为已知常数，求统计量 $\dfrac{\Gamma}{\sigma^2} = \sum_{k=1}^{n} \dfrac{(X_k - \mu)^2}{\sigma^2}$ 的分布。

解 记 $Y_k = \dfrac{X_k - \mu}{\sigma}, k = 1,2,\cdots,n$，则 $Y_1, Y_2, \cdots, Y_n$ 相互独立且都服从 $N(0,1)$ 分布，于是

$$\frac{\Gamma}{\sigma^2} = \sum_{k=1}^{n} \left(\frac{X_k - \mu}{\sigma} \right)^2$$

故 Γ 服从 χ^2（n）分布

2.t 分布

定义 3　设 $X \sim N(0,1), Y \sim \chi^2(n)$，且 X 与 Y 相互独立，则称随机变量

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

服从自由度为 n 的 t **分布**，记为 $T \sim t(n)$。

通过计算可得 t 分布的密度函数为 $f(y) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}(1+\frac{y^2}{n})^{-\frac{n+1}{2}}, -\infty \prec y \prec +\infty$

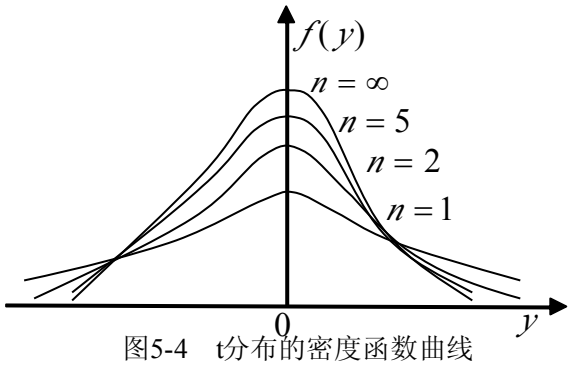

图5-4　t分布的密度函数曲线

图 5-4 给出了 n=1，5，10 时 t 分布的密度函数。以 t_α（n）记为 t 分布的上α 分位点，见图 6-5。由　　$P\{T > t_\alpha(n)\} = \alpha$ ，

查 t 分布表可得 t_α（n）的值。由于 t 分布有对称性，因此 $t_{1-\alpha}(n) = -t_\alpha(n)$

注意到

$$\lim_{n\to\infty}(1+\frac{y^2}{n})^{-\frac{n+1}{2}} = e^{-\frac{y^2}{2}}$$

即 n 很大时，t 分布接近标准正态分布。因此，在应用中，当 n>45 时有 t_α（n）≈z $_\alpha$。

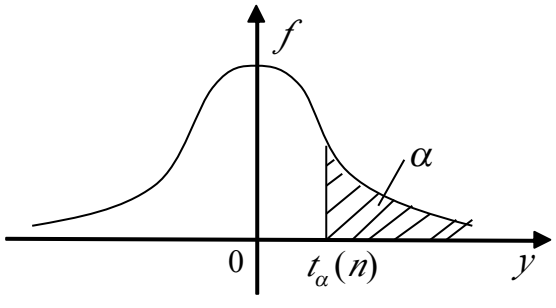

图6-5 t 分布的上α 分位点

3. F 分布

定义 4 设 X,Y 相互独立,分别服从自由度为 n,m 的 x^2 分布,则随机变量

$$F=\frac{\frac{X}{n}}{\frac{Y}{m}}=\frac{X}{Y}\cdot\frac{m}{n}$$

服从自由度为 (n,m) 的 F 分布，记为 $F(n,m)$。显然 $\frac{1}{F}\sim F(m,n).$

通过计算，可求得 $F(n,m)$ 的概率密度函数

$$f(y)=\left\{\frac{\Gamma\left[\frac{n_1+n_2}{2}\right]\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}y^{\frac{n_1}{2}-1}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)\left[1+\frac{n_2}{n_2}y\right]^{\frac{n_1+n_2}{2}}}\right\},y>0$$

比较 t 分布与 F 分布的定义，易知 $t^2(n)=F(1,n)$ 。图 5-6 给出了一些 F 分布的密度函数的图象

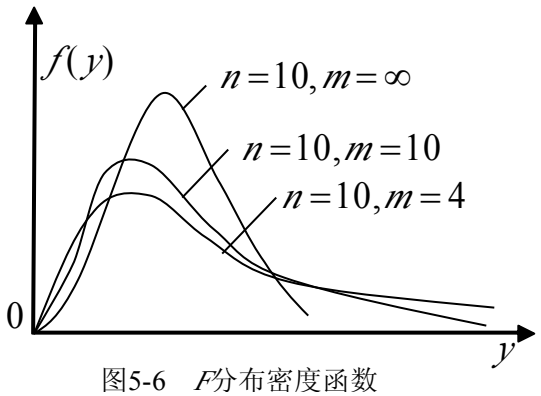

图5-6　F分布密度函数

关于 F 分布的上 α 分位点，我们称满足

$$P\{F>F_\alpha(n,m)\}=\int_{F_\alpha(n,m)}^{+\infty}f(y)dy=\alpha$$

的点 $F_\alpha(n,m)$ 为 $F(n,m)$ 分布的上 α 分位点，见图 6-7。F 分布的上 α 分位点有如下性质：

$$F_{1-\alpha}(n,m)=\frac{1}{F_\alpha(m,n)}.$$

事实上，设 $F\sim F(n,m)$，则

$$\frac{1}{F}\sim F(n,m),$$

且

$$\alpha=P\{F\geq F_\alpha(n,m)\}=P\left\{\frac{1}{F}\leq\frac{1}{F_\alpha(n,m)}\right\}$$

$$=1-P\left\{\frac{1}{F}\geq\frac{1}{F_\alpha(n,m)}\right\}=1-P\left\{\frac{1}{F}\geq\frac{1}{F_\alpha(n,m)}\right\},$$

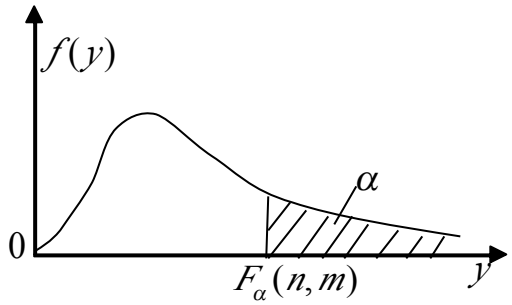

图6-7　F分布的上 α 分布点

于是

$$P\left\{\frac{1}{F}\geq\frac{1}{F_\alpha(n,m)}\right\}=1-\alpha,$$

由α分位点的定义，显然　$F_{1-\alpha}(m,n)=\dfrac{1}{F_\alpha(n,m)}$成立。

理论上，若总体的分布已知，统计量的分布总是确定的。但对一般的总体分布，统计量的分布计算往往很复杂，甚至不能求出。这里我们考虑正态总体分布的抽样分布。一方面是因为其抽样分布较容易求出，另一方面是正态分布可以作为很多统计问题中总体分布的近似。

定理1　设 $X_1,X_2....,X_n$ 是从正态总体 $N(\mu,\sigma^2)$ 中抽取的一个简单随机样本，$\overline{X}$ 与 S^2 分别为样本均值和样本方差，则

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

(2) $\dfrac{(n-1)S^2}{\sigma^2}\sim x^2(n-1)$;

(3) $\overline{X}$与S^2相互独立。(证明略)

推论 1 $$T=\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1).$$

证 由定理知

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\sim N(0,1),\ \frac{(n-1)S^2}{\sigma^2}\sim x^2(n-1)$$

且二者相互独立,由定义 6.3 可知

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\Bigg/\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}=\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1)$$

即$T\sim(n-1)$

设$X_1,X_2....,X_n$与$Y_1,Y_2....,Y_m$分别为来自正态总体$N(\mu_1,{\sigma_1}^2)$和$N(\mu_1,{\sigma_2}^2)$的简单随机样本，且两样本之间相互独立，若

$${S_1}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2,{S_2}^2=\frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2$$

则

(1) $F=\frac{S_1^2}{S_2^2}\cdot\frac{{\sigma_2}^2}{{\sigma_1}^2}\sim F(n-1,m-1);$

(2) 若进一步假设${\sigma_1}^2={\sigma_2}^2$，有

$$T=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim t(n+m-2)$$

其中 $${S_w}^2=\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{i=1}^{m}(Y_i-\overline{Y})^2}{n+m-2}$$

以上结论在后面将经常用到，必须记住。另外，对其它总体，虽然很难求到其精确的抽样分布，但我们可以利用中心极限定理等理论得到当 n 较大时的近似分布，这就是

统计问题中的大样本问题，在此我们不加讨论。

例 从正态总体 $N(\mu,\sigma^2)$ 中抽取容量为 16 的样本，试求：

(1) 已知 $\sigma^2=25$；(2) σ^2 为知，但已知样本方差 $S^2=20.8$ 的情况下，样本均值 $\bar{x}$ 与总体均值 μ 之差的绝对值小于 2 的概率。

解 (1) 由于统计量

$$\mu = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1),$$

因此在 σ^2 已知时，

$$P\{|\bar{x}-\mu|<2\}=P\left\{\frac{|\bar{x}-\mu|}{\frac{\sigma}{\sqrt{n}}}<\frac{2}{\frac{\sigma}{\sqrt{n}}}\right\}=P\left\{\frac{|\bar{x}-\mu|}{\frac{5}{4}}<\frac{2\times 4}{5}\right\}$$

$$=P\{|u|<1.6\}=\phi(1.6)-\phi(-1.6)=2\phi(1.6)-1=2\times 0.9452-1=0.8904;$$

(2) 由于 σ^2 未知，但 S^2=20.8，这时统计量

$$t=\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1),$$

因此

$$P\{|\bar{x}-\mu|<2\}=P\left\{\frac{|\bar{x}-\mu|}{S/\sqrt{n}}<\frac{2}{S/\sqrt{n}}\right\}=P\left\{\frac{|\bar{x}-\mu|}{S/\sqrt{n}}<\frac{2}{4.56/\sqrt{16}}\right\}$$

$$=P\{|t|<1.754\}=1-P\{|t|\geq 1.754\}$$

查 t 分布表得 $t_{0.05}(16\text{-}1)=1.753$，$P(t\geq 1.753)=0.05$。由此可得

$$P\{|\bar{x}-\mu|<2\}\approx 1-2\times 0.05=0.90$$

例 设总体 X 服分布 $N(72,100)$，为使样本均值大于 70 的概率不小于 90%，则样本容量应取多少?

解 设所需样本容量为 n，由于

$$\frac{\bar{x}-\mu}{\sigma}\sqrt{n}\sim N(0,1),$$

则

$$P\{\overline{X}>70\}=P\left\{\frac{\overline{X}-72}{10}\sqrt{n}>\frac{70-72}{10}\sqrt{n}\right\}=1-P\left\{\frac{\overline{X}-72}{10}\sqrt{n}<-0.2\sqrt{n}\right\}$$

$$=1-\phi\left(-0.2\sqrt{n}\right)=\phi\left(0.2\sqrt{n}\right)\geq 0.9$$

查标准正态分布得

$$0.2\sqrt{n}\geq 1.29,$$

即 $n\geq 41.6025$，故样本容量至少为 42，才能使样本均值不大于 90%。