【例 5.8】 有一大批产品次品率比较低，从中抽取 200 个作检验，发现有 5 个次品，试以 95% 的置信水平确定次品率的置信区间。

解：$P=\frac{5}{200}=0.025$，比例比较小，属于稀少事件，$n=200$ 为大样本，$nP=5$。样本中的次品数服从参数为 μ 的泊松分布，μ 的置信区间可查泊松分布 μ 的置信区间表。该表的一侧为具有某种特征单位数，用 C 表示，因此可以找到 $C=5$ 处，表的顶端为置信系数，找到 $1-\alpha=0.95$，在交错的位置上为 1.62—11.67，因为这是 200 个产品中的次品数，把它转换成比例应分别除以 200，得 μ 的置信区间为 $\left(\frac{1.62}{200},\frac{11.67}{200}\right)$，即在 0.008 1 和 0.058 75 之间。

如果 $P>0.5$ 而 Q 较小，当 $nQ\leqslant 5$ 时也应用泊松分布计算。

2. 比例 P 虽然随着样本量 n 的增加而近似正态分布，但究竟多大才能使 P 近似正态分布呢？这与 p 的取值大小有关，当 p 接近于 0.5 时，用较少的样本即可趋近正态分布，但当 p 接近于 0 和 1 时就要很大的样本量才能趋向正态分布，统计学家科克伦(W.G.Cochran)提出了一个标准可供参考(见表 5.4)：

表 5.4 比例近似正态分布要求的样本量

P	近似正态分布要求样本量
0.5	30
0.4—0.6	50
0.3—0.7	80
0.2—0.8	200
0.1—0.9	600

第四节　两个均值或两个比例之差的区间估计

一、两个总体均值之差的区间估计

在日常生活中人们也会遇到两总体之差的区间估计。例如城市居民户每周看电视的小时数是否多于郊区，究竟多多少；甲厂的产品质量是否高于乙厂，究竟高多少等，这就需要在各自的总体中抽取样本进行比较，并由此推断总体。

(一) 两个总体为正态分布或大样本

估计两个总体均值之差的估计量显然为两个样本均值之差。设两个总体均值分别为 μ_1 和 μ_2，方差分别为 σ_1^2 和 σ_2^2。为了估计两总体均值之差，分别独立地抽取两个样本，其样本量为 n_1 和 n_2，得到样本均值为 $\bar{X}_1$ 和 $\bar{X}_2$，则可以证明 $\bar{X}_1 - \bar{X}_2$ 为总体均值之差 $\mu_1 - \mu_2$ 的无偏估计，即

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

由 $\bar{X}_1$ 和 $\bar{X}_2$ 分别为正态分布，其差也服从正态分布，而且其方差为各自的方差 $\frac{\sigma_1^2}{n_1}$ 和 $\frac{\sigma_2^2}{n_2}$ 之和，即

$$D(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

因为是分别独立抽取的，其协方差为 0，从而有

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

当置信度为 $1-\alpha$ 时，$\mu_1 - \mu_2$ 之差的置信区间为

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{5.8}$$

在大样本情况下，(5.8)式中的 σ_1^2 和 σ_2^2 可以用样本方差 S_1^2 和 S_2^2 替代。

【例 5.9】 某罐头食品厂，有两条装罐头的生产线，有一条生产线已老化，装的数量偏多，而差异较大，现欲测定两条生产线上所装数量的差别。于是在每条生产线上各抽 100 个罐头，一条生产线上罐头的平均重量为 103.02 克，方差为 0.64，另一条生产线的平均重量为 102.83 克，方差为 0.36。试构造两条生产线重量之差 90% 的置信区间。

解：因为 $1-\alpha=0.9$ $z_{\alpha/2}=z_{0.05}=1.64$，两条生产线的总体方差未知，由于是大样本，可以用样本方差估计总体方差，故

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}=\sqrt{\frac{0.64}{100}+\frac{0.36}{100}}=\sqrt{\frac{1}{100}}=\frac{1}{10}=0.1$$

因此 $\mu_1-\mu_2$ 的置信区间为

$$\bar{x}_1-\bar{x}_2\pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}=103.02-102.83\pm 1.64(0.1)$$

置信下限为 $103.02-102.83-1.64(0.1)=0.026$

置信上限为 $103.02-102.83+1.64(0.1)=0.354$

因此第一条生产线比第二条生产线多装 0.026—0.354 克。

（二）两个正态总体，方差未知，小样本

当比较两个正态总体的均值而方差未知时，在大样本的情况下可以用样本方差来代替总体方差，但在小样本的情况下就不能简单进行代替。假设两个总体的方差是相等的，这比较简单，可以将两组样本数据合并起来估计方差，即

$$\hat{\sigma}^2=S_{合}^2=\frac{\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)^2+\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_2)^2}{n_1+n_2-2}$$

$$=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

因为 $$\bar{X}_1-\bar{X}_2\sim N\left(\mu_1-\mu_2,\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}\right)$$

所以 $$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim N(0,1)$$

当用 $S_{合}^2$ 来代替 σ^2 时

$$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_{合}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

即 t 分布自由度为 n_1+n_2-2，$\mu_1-\mu_2$ 置信度为 $1-\alpha$ 的置信区间为：

$$(\bar{X}_1-\bar{X}_2)\pm t_{\alpha/2}(n_1+n_2-2)S_{合}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \qquad (5.9)$$

【例 5.10】 某汽车零件要更换一种新的替代产品，为比较两种零件的行驶里程，分别抽取了 8 个样本作记录，其数据如下（千公里）

原有零件 39.6,34.2,47,40.9,50.6,27.5,43.5,36.3

替代零件 35.7,52,46.8,58.5,45.7,52.4,41.3,43.8

假设行驶里程服从正态分布，试以 0.90 的置信水平，估计两种零件平均行驶里程的差别。

解：从样本数据可以得到：

原零件 $\bar{x}_1=39.95$ $s_1^2=54.02$

替代零件 $\bar{x}_2=47.03$ $s_2^2=51.22$

由于总体方差未知，并假设其相等，需要合并估计：

$$s_{合}=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}=\sqrt{\frac{7(54.02)+7(51.22)}{14}}=7.25$$

要求置信度 $1-\alpha=0.9$，$t_{\alpha/2}(n_1+n_2-2)=t_{\alpha/2}(14)=1.761$

$\mu_1-\mu_2$ 的置信区间为

$$\bar{x}_1-\bar{x}_2\pm 1.761(7.25)\sqrt{\frac{1}{8}+\frac{1}{8}}=-7.08\pm 6.39$$

下限 $=-13.47$，上限 $=-0.69$，即原零件要比替代零件少行驶 690—13470 公里。

(三)成对观测的两个正态总体均值之差的估计

有时需要通过成对的观测来估计两个总体均值之差,例如社会学家研究每对夫妇中丈夫平均工资与妻子平均工资之差,企业要对工人培训前后平均产量的差别进行估计。要估计这种成对比较差别的置信区间,需要先计算出每一对的差别 $D_i = X_{1i} - X_{2i}$,于是 $\mu_1 - \mu_2$ 之差可用 d 表示,从总体中抽取 n 对样本,计算 $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ 作为 d 的估计量,并计算方差 $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$,由于两个总体服从正态分布,故 $\bar{D}$ 也服从正态分布,$\frac{\bar{D} - d}{\frac{S_D}{\sqrt{n}}} \sim t(n-1)$,$d$ 的区间估计为

$$\bar{D} \pm t_{\alpha/2}(n-1) \cdot \left(\frac{S_D}{\sqrt{n}}\right) \tag{5.10}$$

【例 5.11】 为估计服用某种药物对人体某种指标值的变化,抽取了 10 个人服用这种药物前后的指标,测得数据如下:

样本	1	2	3	4	5	6	7	8	9	10
服用前	41	60.3	23.9	36.2	52.7	22.5	67.5	50.3	50.9	24.6
服用后	46.9	64.5	33.3	36	43.5	56.8	60.7	57.3	65.4	41.9

假设该指标变量值服从正态分布,试以 0.90 的置信水平估计服用该药后指标值之差的置信区间。

解:以服用后的指标值减去服用前的指标值得到 d_i,并计算 $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = 7.64$,$s_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2} = 12.57$,由于指标值服从正态分布,$\bar{D} \sim N\left(d, \frac{\sigma_D^2}{n}\right)$,当用样本方差 S_D^2 代替总体方差时

$$\frac{\bar{D} - d}{\frac{S_D}{\sqrt{n}}} \sim t(n-1) \text{ 分布}$$

由公式(5.10),d 的置信区间为

$$\bar{d} \pm t_{\alpha/2}(n-1)\frac{s_D}{\sqrt{n}} = 7.64 \pm 1.833(3.97)$$

即服用该药后指标值增加 0.36—14.92。

二、两个比例之差的区间估计

当研究的目的是估计两个总体的比例差 $p_1 - p_2$ 时,由前面单个总体比例的差可知,通常需要大样本才能使样本比例服从正态分布,样本比例之差 $P_1 - P_2$ 是总体比例之差的无偏估计。在两个样本均为大样本,且有 $p_1 < 0.5, n_1 p_1 \geqslant 5, p_2 < 0.5, n_2 p_2 \geqslant 5$ 时,$P_1 - P_2$ 也近似正态分布。$p_1 - p_2$ 的置信区间为:

$$P_1 - P_2 \pm z_{\alpha/2}\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} \quad (5.11)$$

【例 5.12】 在某个节目的电视收视率调查中,农村调查了 400 人,有 32% 的人收看了该节目,城市抽了 500 人,有 35% 的人收看了该节目,试以 90% 的置信度估计城市与农村收视率差别的置信区间。

解:$P_2 - P_1 = 0.35 - 0.32 = 0.03$,当置信度为 90% 时,$1 - \alpha = 0.9$,$z_{\alpha/2} = z_{0.05} = 1.645$,因此置信区间为

$$P_2 - P_1 \pm 1.645\sqrt{\frac{(0.35)(0.65)}{500} + \frac{(0.32)(0.68)}{400}} = 0.03 \pm 0.052$$

即 $-0.022 - 0.082$。

从置信区间的范围看其差别从负到正,因此尚不能以 90% 的置信度断定城市收视率高于农村,因为抽样中有随机的原因。

第五节 样本容量的确定

一、参数估计中置信度、置信区间与样本量的关系

在参数估计中人们总是愿意提高估计的置信程度,但是在一

定样本量和抽样方式下，欲提高置信度就会扩大置信区间。而过宽的置信区间在实际的估计问题中是没有意义的。例如估计对某项意见的支持率在40%～80%之间，估计某一总体中人的平均年龄在30～60岁之间，显然是没有多大意义的。反过来如果要缩小置信区间，就会降低置信度，太低的置信度，比如说50%的置信度也同样是没有意义的。二者的关系如图5.4所示。

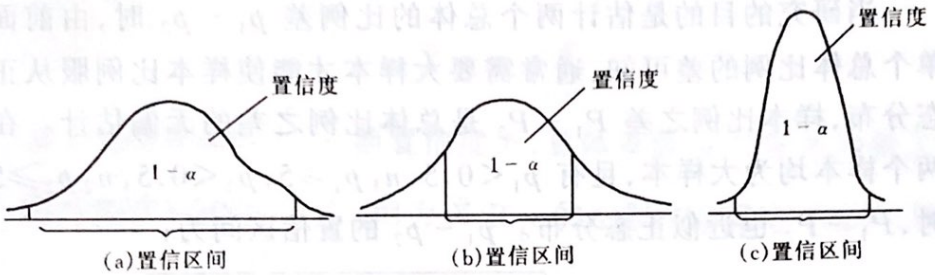

图5.4　置信区间与置信度的关系

在样本量和抽样方式相同时其抽样分布是一致的，如图5.4中的(a)和(b)。在图(a)中置信度$1-\alpha$比较高，置信区间也就比较宽。在图(b)中置信区间小了，置信度$1-\alpha$也随之缩小。如果既要缩小置信区间，同时又不降低置信度，就要增加样本容量，因为增加样本量可以使统计量抽样分布的离散程度缩小，从而使更多的样本落在总体真值周围，如图5.4中的(c)。但是样本量的增大又会带来工作量的增大，费用增加，调查时间延长以及又可能会使调查误差增大等等问题。因此样本量也不是愈大愈好。适当地确定样本量是抽样估计的一个重要问题。这要取决于：

1．对某一项估计要求什么样的精度，即希望得到的估计值与总体真值之间的离差在什么样的范围以内。换句话说，想构造多宽的置信区间。

2．对于规定的置信区间来说想要多大的置信度，也即总体真值在这区间内的结论有多大的可靠性。

3．对该项估计所能承担的费用。

二、估计总体均值时样本容量的确定

(一) 重复抽样或抽样比$\frac{n}{N}$比较小可以忽略不计时

在总体均值的区间估计时,在大样本的情况下,样本均值的抽样分布服从正态分布。因此总体均值 μ 的置信区间由公式 5.1 可知为

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

这一公式表示在 $1-\alpha$ 的置信度下,总体参数 μ 与样本均值 $\bar{X}$ 的绝对离差不超过 $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$,因为 $\bar{X}$ 在 μ 的左右两边都是可能的,因此 $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ 实际上为置信区间长度的 1/2,通常称作允许误差,可用 Δ 表示。其中 $z_{\alpha/2}$ 是相应于置信度 $1-\alpha$ 在正态分布情况下的置信系数,可以查正态分布下的面积表得到。最常用的有:

置信度$(1-\alpha)$	$z_{\alpha/2}$
0.9	1.645
0.95	1.96
0.99	2.58

σ 是总体变量的标准差,也是确定样本容量所必要的。由公式 $\Delta = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ 可以推导出:

$$n = \left(\frac{z_{\alpha/2}\sigma}{\Delta}\right)^2 \tag{5.12}$$

从这一公式反映出样本量取决于置信度 $1-\alpha$,即相应的 $z_{\alpha/2}$,还取决于总体方差 σ^2 和估计时的允许误差 Δ。这三者与样本量的关系如下:

1. 置信程度与样本量成正比,当 σ 和 Δ 保持不变时,置信程

度要求愈高，样本量也要愈大。

2. 总体方差与样本量成正比。总体的差异愈大，要求的样本量也要大。

3. 允许误差与样本量成反比，允许误差放大，也就是置信区间放宽，样本量可以减少。

【例 5.13】 某超级市场欲估计每个顾客平均每次购物的金额，根据过去的经验，标准差大约为 160 元，现要求以 95% 的置信度估计每个顾客的购物金额，并要求允许误差不超过 20 元，应抽多少顾客作样本？

解：因为 $1-\alpha=0.95 \quad z_{\alpha/2}=z_{0.025}=1.96$

已知 $\sigma=160, \Delta=20$ 代入公式得

$$n=\frac{(z_{\alpha/2})^2\sigma^2}{\Delta^2}=\frac{(1.96)^2(160)^2}{20^2}=246$$

即应抽 246 位顾客作调查。

（二）有限总体不重复抽样下的样本容量

在实际抽样问题中，总体往往是有限的，而且采用不重复抽样，这时 μ 的置信区间由公式(5.3)可知：

$$\bar{X}\pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}}$$

因此：

$$\Delta=z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

由此推导出

$$n=\frac{Nz_{\alpha/2}^2\sigma^2}{(N-1)\Delta^2+z_{\alpha/2}^2\sigma^2} \tag{5.13}$$

【例 5.14】 某养鸡专业户饲养了 1 000 只小鸡，一个月后欲估计这批小鸡的总重量，要求估计的误差不超过 2 公斤(2 000 克)，而每只鸡重量的方差约为 49，若置信水平为 95% 应抽多少只鸡作样本？

解：要求估计的是总重量，允许误差为 2 000 克，换算成每只

鸡的允许误差为$\frac{2\ 000}{1\ 000}=2$克，即$\Delta=2$，又知$\sigma^2=49$，$z_{\alpha/2}=1.96$，所以

$$n=\frac{1\ 000(1.96)^2(49)}{999(4)+(1.96)^2 49}=\frac{188\ 238.4}{4\ 184.238\ 4}=45$$

应随机抽取45只鸡作样本。

不重复抽样估计样本量的公式不易记忆，可以用重复抽样样本量的公式作进一步推算。设重复抽样的样本量为n_0，则不重复抽样的样本量为：

$$n=\frac{n_0}{1+\frac{n_0}{N}} \tag{5.14}$$

【例5.15】 以例14的数据用公式(5.14)计算样本量。

解：先用重复抽样的公式(5.13)计算样本量

$$n_0=\frac{z_{\alpha/2}^2\sigma^2}{\Delta^2}=\frac{1.96^2(49)}{2^2}=47$$

再用公式(5.14)

$$n=\frac{n_0}{1+\frac{n_0}{N}}=\frac{47}{1+\frac{47}{1000}}=45$$

其计算结果是一致的。从上可以看出不重复抽样的效率要高于重复抽样，当抽样比例比较大时就更为明显。

三、总体比例样本量的确定

1. 在重复抽样或抽样比$\frac{n}{N}<0.05$时，由

$$\Delta=z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

可以推导出

$$n=\frac{z_{\alpha/2}^2 P(1-P)}{\Delta^2} \tag{5.15}$$

应注意，这里的允许误差 Δ 是绝对的误差，因为估计的是比例问题，容易与相对误差相混淆。如估计总体比例为40%时，$\Delta=2\%$，意味着估计时应在38%和42%之间。相对误差是指$\frac{\Delta}{P}$，也是用百分比表示，如估计的比例 $P=40\%$，允许误差 $\Delta=2\%$，则相对的允许误差为 $r=\frac{\Delta}{P}=\frac{2}{40}=5\%$。假如预先规定是相对误差 r，则样本容量公式就改为：

$$n=\frac{z_{\alpha/2}^2(1-P)}{r^2P} \tag{5.16}$$

【例5.16】 若某一城市欲调查出生率，根据以往经验，该城市的出生率约为13‰，现要求相对误差不超过10%，置信度为95%，采用简单随机抽样应抽取多少人作样本？

解：这个问题可以有两种解法，第一种解法是把相对误差换算成绝对误差，然后用公式(5.15)计算，第二种解法是直接用公式(5.16)计算：

解法一：因为相对误差为10%，$P=0.013$ 因此 $\Delta=0.013\times10\%=0.001\ 3$，因此用公式(5.15)

$$n=\frac{z_{\alpha/2}P(1-P)}{\Delta^2}=\frac{1.96^2(0.013)(0.987)}{(0.001\ 3)^2}$$

$$=\frac{0.049\ 291\ 569}{1.69\times10^{-6}}=29\ 166\text{ 人}$$

解法二：直接用公式(5.16)

$$n=\frac{z_{\alpha/2}^2(1-P)}{r^2P}=\frac{1.96^2(0.987)}{(0.1)^2(0.013)}=\frac{3.791\ 659\ 2}{0.000\ 13}=29\ 166\text{ 人}$$

两种方法计算的结果是一致的。

2. 有限总体不重复抽样。用同样的方法可求得样本量的公式为

$$n=\frac{Nz_{\alpha/2}^2P(1-P)}{(N-1)\Delta^2+z_{\alpha/2}^2P(1-P)} \tag{5.17}$$

或是分两步计算：