

第五章 参数估计

统计推断是根据样本信息来对产生样本的总体性质和特征进行推测和估计的统计方法,参数估计是统计推断的一大内容,包括参数的点估计和区间估计两类。本章将解释如何用样本信息估计总体的均值或总体比例,以及怎样做这些参数的区间估计。关于参数点估计的评价标准也在本章有简单的描述。最后,我们将以如何确定能够满足区间估计精度要求的样本容量的讨论作为本章的结束。

第一节 参数估计的一般问题

一、什么是参数估计

在抽样分布一章主要是讨论在已知总体分布以及有关参数的情况下,通过抽样得到统计量的分布。但是在现实生活中,总体参数常常是未知的,而是要根据样本的数据及其分布来推断总体参数,这就是统计推断的目的。从推断的逻辑来讲,概率论和抽样分布所应用的是演绎的方法,从一般到具体,而统计推断则是一种归纳的方法,从具体到一般。统计推断主要有估计和检验两种方式。本章主要是讨论估计问题。

参数从狭义讲是指决定某一理论分布的函数中一个或若干个数值,它决定了随机变量的分布状况。如正态分布有两个参数 μ 和 σ ,它决定了正态分布重心的位置及离散情况;又如泊松分布有一个参数 λ 决定了分布的形状。二项分布的参数有试验的次数 n 和每次试验中某事件出现的概率 p 等。从广义上讲参数是反映总体特征和决定有关模型的数值,如总体的均值、总体的总值、总

体的比例、总体的方差、两个变量中回归模型的回归系数、相关模型中的相关系数等等。统计是研究客观现象的数量表现及其规律性的,因此统计工作的目的就是要取得各种参数。除了全面调查以外就要通过抽样来加以估计。因此本章所要讨论的参数也是指广义的参数。例如:要估计全国的粮食产量;某一地区居民户的平均收入;某公司的经理要对经营产品的未来销售量进行估计;某广告公司想了解该公司广告的收视率及其效果;财务人员需要估计账户上呆账的总值等等。取得这些总体参数对决策者进行决策具有重要作用。

二、估计量和估计值

总体是客观现象的全体,总体的范围一旦确定以后,参数通常是一个不变的常数。在对总体参数进行估计时要利用样本的统计量,这些统计量的具体取值随着抽到不同样本单元而变化,因而是随机变量。这就要应用上一章抽样分布的知识。这些统计量及其数值有两个术语:估计量(estimator)和估计值(estimate)。估计量指的是用来估计总体参数的统计量的名称,也即样本函数的名称。它说明如何对样本数据进行加工计算。例如,通常用样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 来估计总体均值,样本的比例 P 来估计总体比例 p ,样本均值和样本比例就是估计量。估计值是某一估计量用来估计参数时计算结果的具体数值。抽取一个样本并计算这个样本估计量所得到的具体值就得到了估计值。因此估计量是对所有可能出现的估计值的抽象概括(通常采用大写英文字母表示),而估计值则是估计量的具体体现(一般用对应的小写英文字母表示)。把估计量与估计值加以区别是重要的,因为二者之间既有联系又有区别。一个好的估计量有可能得到差的估计值,而一个差的估计量也有可能得到好的估计值。比如要估计某批钢材的拉力强度,假设总体的平均拉力强度是 5 000 公斤,从中随机抽取 3 件作为样本,分别测试后的结果为 4 900, 4 999 和 5 011 公斤,从样本测试结果可

以采用两个估计量,一个是样本均值 \bar{X} ,一个是样本中位数,从这一样本计算出 $\bar{x} = 4970$,中位数 4999,这就是估计值,从这一具体估计值看,中位数 4999 公斤最接近总体均值,似乎最好。但再抽一个样本就可能不是这样的结果了,因此应从总体上看,也即从大量试验结果来看。可以证明估计量中样本均值要优于样本中位数这一估计量。由于在估计时,总体真值是未知的,因此要选择一个好的估计量。

三、估计的类型

估计可以分为点估计和区间估计。

点估计是指从抽到的具体样本,经过调查或观察把所得到的数据算出的单个估计值用作相应总体参数的值。设某个未知总体参数为 θ ,估计量为 $\hat{\theta}$,根据某个随机样本计算得出估计量 $\hat{\theta}$ 的具体估计值就是总体参数的点估计。例如要估计企业某批产品的次品率,通常以样本中次品的比例作为估计量,假设抽取了 100 个产品,发现有 9 个是次品,样本的次品率为 9%,作为总体的次品率,这就是点估计。由于抽样的随机性,估计量 $\hat{\theta}$ 是一个随机变量,而点估计就以随机变量中的某一个值来作出估计,显然会有抽样误差产生,这种误差是由样本的代表性引起的,如果误差比较小,那么这个点估计值还是一个好的估计值,如果误差很大,那么这个点估计就失去意义。当选择的是一个好的估计量,那么大部分的估计值应该是接近总体参数,由于这一估计方式简单直观,所以也得到广泛应用。但点估计无法指出抽样误差的大小,这是它的局限性。

区间估计是在点估计的基础上给出一个估计的范围,使总体参数包括在这个范围之内,而且可以推断总体参数有多大的把握被涵盖在这个范围之内,因此区间估计是包含总体参数的一个值域,同时在结论中指出未知参数所在值域的上下限和该结论的可靠性。例如根据样本资料估计出某地居民户的平均年收入在 7 000~8 000 元之间,这一结论的可靠性是 95%,这就是一个区间估计。

四、估计量的评价标准

要对总体参数进行估计,需要选择一个估计量,因为总体参数有时有不止一个估计量,而是可以有若干个估计量。比如前面介绍估计量时对于钢材的平均拉力强度,可以用样本均值来估计,也可以用样本的中位数来估计。究竟应该选择哪一个估计量呢?当然要选择一个最接近总体均值的估计量,然而在估计时并不知道总体均值,又如何能找到最接近总体均值的估计量呢?这就要比较估计量的抽样分布,因为抽样分布比较全面地概括了统计量的所有可能结果。统计学家们在评价估计量方面提出了一些标准,这里介绍主要的几个:

(一) 无偏性

无偏性就是指估计量抽样分布的数学期望要等于总体的参数。用数学的语言表述:设总体参数为 θ ,要选择的估计量为 $\hat{\theta}$,如果 $E(\hat{\theta}) = \theta$,称 $\hat{\theta}$ 为 θ 的无偏估计量。从图 5.1 直观地显示 $\hat{\theta}_1$ 为无偏估计量, $\hat{\theta}_2$ 是有偏估计量。

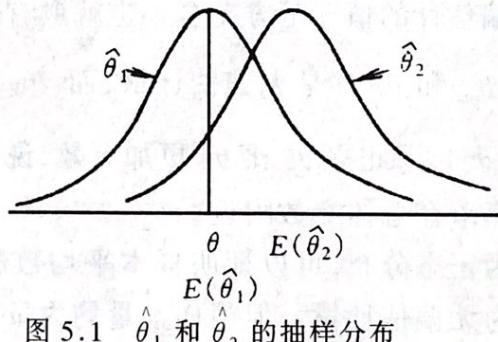


图 5.1 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 的抽样分布

如果估计量 $\hat{\theta}$ 不是总体参数 θ 的无偏估计量的话,则 $E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta \neq 0$ 。称 $E(\hat{\theta}) - \theta$ 为偏差。 θ 的无偏估计量 $\hat{\theta}$ 就是使偏差 $E(\hat{\theta} - \theta)$ 等于零的估计量。我们知道,在随机抽样中,有时会抽中偏小的单位,有时则会抽中偏大的单位。在无偏估计的情况下,这种误差没有系统性的方向,随着样本量的增大,(在数学

期望的意义下)这种有大有小的误差会相互抵消,因此我们称无偏估计量没有系统性误差。有偏的估计量则不然,它在抽样中表现出来的误差不会随着样本量的增大而消失,而是具有一定的方向,我们称有偏估计量具有系统性误差。

对于总体参数 θ 的任一估计值 $\hat{\theta}$, 我们称 $\hat{\theta} - \theta$ 为抽样误差。注意抽样误差与偏差的关系是: 偏差是大量抽样的平均抽样误差, 而抽样误差是某一次具体抽样的误差。

统计学证明了样本的均值是总体均值的无偏估计量, 即 $E(\bar{X}) = \mu$, 其中: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 为样本均值, μ 为总体均值。也证明了样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 在重复抽样的条件下是总体方差 σ^2 的无偏估计量。

(二) 有效性

一个无偏估计量并不意味着这一估计量一定非常接近待估计的参数, 它还必须是与总体参数的离散程度比较小。离散程度通常是以方差来衡量的, 因此也就是要求估计量抽样分布的方差比较小。在无偏估计的情况下方差愈小也就愈有效。图 5.2 显示了两个估计量 $\hat{\theta}_a$ 和 $\hat{\theta}_b$ 都是无偏估计量, 而 $\hat{\theta}_a$ 的方差小于 $\hat{\theta}_b$, 即 $D(\hat{\theta}_a) < D(\hat{\theta}_b)$, 因此称 $\hat{\theta}_a$ 比 $\hat{\theta}_b$ 更加有效, 说明该估计量有更大比例的估计值落在总体参数附近。

当总体为正态分布, 可以证明样本平均数和样本中位数这两个估计量均为无偏估计量。但当样本量均为 n 时, 样本均值 \bar{X} 的方差为 $\frac{\sigma^2}{n}$, 而中位数的方差为 $\frac{\pi\sigma^2}{2n} \doteq \frac{1.57\sigma^2}{n}$, 说明样本均值 \bar{X} 比中位数更加有效。在抽样中有时会遇到无偏估计量有较大的方差, 而另一个有偏估计量有较小的方差。这种情况下选择的标准仍以离总体真值愈小愈好, 衡量的标准是均方误, 即 $E(\hat{\theta} - \theta)^2$, 它可以分解为方差和偏差的平方两部分, 即 $E(\hat{\theta} - \theta)^2 = E[(\hat{\theta} -$

$E(\hat{\theta})) + (E(\hat{\theta}) - \theta)^2 = D(\hat{\theta}) + b^2$, 其中 b 代表偏差。

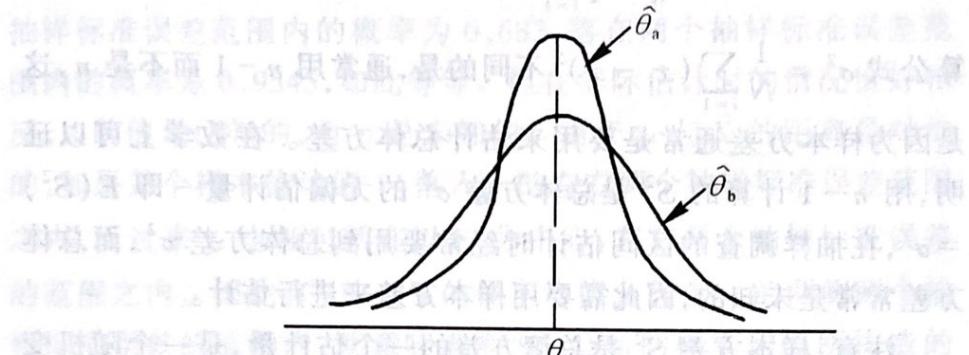


图 5.2 $\hat{\theta}_a$ 与 $\hat{\theta}_b$ 的抽样分布(图示有效性)

(三) 一致性

又称相合性,是指随着样本容量的增大,估计值就愈来愈接近于总体参数值。如果一个估计量是一致估计量,那么样本愈大就愈精确,就可以通过增加样本容量来提高估计精度和增加可靠性,如果不是一致估计量,抽取大样本就会浪费时间和费用。以后将会看到一些常用的估计量都是一致的估计量。

五、样本的数字特征与总体参数的点估计

在抽样调查中通常以样本的估计值对总体的特征进行估计,最常用的样本估计值是样本均值、样本方差和样本标准差。

(一) 样本均值

设随机变量 X , 从中抽取样本容量为 n 的样本, 得到其观测为 $X_i (i = 1, 2, \dots, n)$, 则样本均值定义为 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 。在等概率抽选的情况下, 通常用样本均值来估计总体均值, 因它是总体均值 μ 的无偏估计量, 即 $E(\bar{X}) = \mu$, 而且它有一些比较好的数学性质, 特别地, 样本均值 \bar{X} 的方差 $D(\bar{X}) = \frac{\sigma^2}{n}$, 其中 σ^2 为总体方差。

(二) 样本方差

样本方差是反映样本数据离散程度的指标, 通常用来估计总体的方差。设随机变量 X , 从中抽取样本容量为 n 的样本, 则样

本方差定义为 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, 这里与有限总体方差计算公式 $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$ 不同的是, 通常用 $n-1$ 而不是 n , 这是因为样本方差通常是要用来估计总体方差。在数学上可以证明, 用 $n-1$ 计算的 S^2 是总体方差 σ^2 的无偏估计量 – 即 $E(S^2) = \sigma^2$, 在抽样调查的区间估计时经常要用到总体方差 σ^2 , 而总体方差常常是未知的, 因此需要用样本方差来进行估计。

注意: 样本方差 S^2 是总方差的一个估计量, 是一个随机变量; 而样本均值的方差 $D(\bar{X})$ 是一个常数(也许是未知的)。这是两个不同的概念, 读者应当注意。

(三) 样本标准差

样本标准差是样本方差的算术平方根, 其定义为 $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$ 。其作用与样本方差相似。需要注意的是, 样本标准差 S 不是总体标准差 σ 的无偏估计量。

第二节 总体均值的区间估计

一、区间估计的基本原理

估计总体均值是实际工作中最常见的一种参数估计。例如, 欲估计一批灯泡的平均使用寿命; 某一居民区的户均年收入; 某种作物的平均单产; 某出租汽车公司每辆车的平均行驶里程等等。区间估计是根据样本的估计值计算出一个数值范围, 推断总体参数位于这一数值范围内的概率, 这一数值范围称作置信区间, 其边界则称为置信限, 位于这一数值范围内的概率称作置信水平。区间估计的原理要利用第四章的抽样分布, 例如要对总体均值 μ 进行区间估计, 就需要知道样本均值 \bar{X} 与总体均值之间的关系, 当样本均值 \bar{X} 的抽样分布为正态分布, 其数学期望为 μ , 抽样标准

误差为 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ 时, 可以知道样本均值 \bar{X} 落在 μ 的两侧各为一个

抽样标准误差范围内的概率为 0.683, 落在两个抽样标准误差范围内的概率是 0.9545, 如此等等。但在实际估计时的情况恰好相反, \bar{x} 的值是已知的, 而 μ 是未知的。由于 μ 与 \bar{X} 的距离是对称的, 如果某个样本估计值 \bar{x}_i 落入 μ 的左右两个抽样标准误差的范围之内, 反过来 μ 也被包括在以 \bar{x}_i 为中心, 左右两个抽样标准误差的范围之内。因此有 95.45% 的样本均值会落在 μ 的两侧两个抽样标准误差的范围内, 就意味着有 95.45% 的样本均值所构造的两个抽样标准误差区间会包括 μ 。实际上抽取的只是一个容量为 n 的样本, 该样本的均值 \bar{X} , 落入 μ 两侧两个抽样标准误差的概率为 95.45%, 同样总体均值 μ 被包括在这一样本构造的区间之内的概率也是 95.45%。图 5.3 表示区间估计的原理。由于对总体的信息掌握不同, 抽样的方式不同, 其区间上、下限的确定也有所差别, 现分别不同情况加以讨论。

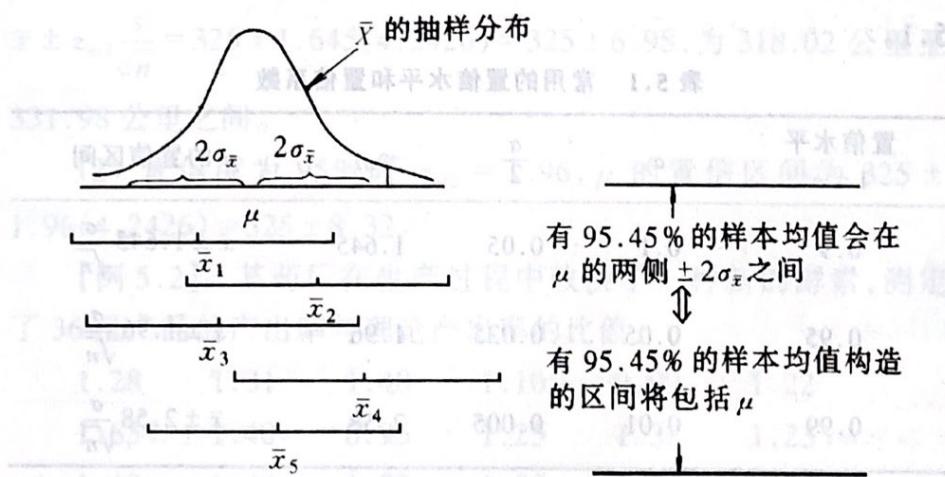


图 5.3 区间估计示意图

二、正态总体方差 σ^2 已知或非正态总体方差未知大样本

在上述情况下样本均值 \bar{X} 的抽样分布均为正态分布, 其数学

期望为总体均值 μ , 抽样分布的方差为 $\frac{\sigma^2}{n}$, 即

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

由正态分布的性质可推出

$$P\left(\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

也即 $P\left(|\mu - \bar{X}| < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$, 因此 μ 的置信区间为:

$$\left(\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \quad (5.1)$$

其中 $\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ 称作 μ 的置信下限, $\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ 为 μ 的置信上限,

$z_{\alpha/2}$ 是正态分布条件与置信水平相联系的系数, α 为选择的风险, 也就是总体均值 μ 不包括在置信区间的概率, $1 - \alpha$ 称作置信水平或置信度。要根据估计的对象来确定, 常用的置信水平如表 5.1。

表 5.1 常用的置信水平和置信系数

置信水平 $1 - \alpha$	α	$\frac{\alpha}{2}$	$z_{\alpha/2}$	μ 的置信区间
0.9	0.1	0.05	1.645	$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}$
0.95	0.05	0.025	1.96	$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
0.99	0.01	0.005	2.58	$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$

这里还需要补充两点:

- 如果是总体方差未知, 而采用大样本, 则可用样本方差 S^2 代替总体方差 σ^2 , 从而 μ 的置信区间为:

$$\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}} \quad (5.2)$$

2. 如果是采取不重复抽样,而且抽样比 n/N 比较大,则抽样分布的方差应乘以有限总体不重复抽样的修正系数 $\frac{N-n}{N-1}$,从而 μ 的置信区间为:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad \text{或} \quad \bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (5.3)$$

【例 5.1】 某汽车租赁公司欲估计全年每个租赁汽车的顾客每次租赁平均行驶的里程。由于全年汽车租赁量很大,随机抽取了 200 个顾客,根据记录计算平均行驶里程 $\bar{x} = 325$ 公里,标准差 $s = 60$ 公里。试估计全年所有租赁汽车每次平均行驶里程的置信区间。置信水平分别为(1)0.90,(2)0.95。

解:由于样本量 $n = 200$ 为大样本,故 \bar{X} 的抽样分布为正态分布, \bar{X} 的标准差的估计值为 $\frac{s}{\sqrt{n}} = \frac{60}{\sqrt{200}} = 4.2426$

(1) 置信度为 90% 时, $z_{\alpha/2} = 1.645$,由公式(5.2),置信区间为 $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 325 \pm 1.645(4.2426) = 325 \pm 6.98$,为 318.02 公里至 331.98 公里之间。

(2) 置信度为 95% 时 $z_{\alpha/2} = 1.96$, μ 的置信区间为 $325 \pm 1.96(4.2426) = 325 \pm 8.32$ 。

【例 5.2】 某药厂在生产过程中改换了一种新的酵素,测定了 36 批产品的产出率与理论产出率的比值:

1.28	1.31	1.48	1.10	0.99	1.22
1.65	1.40	0.95	1.25	1.32	1.23
1.43	1.24	1.73	1.35	1.31	0.92
1.10	1.05	1.39	1.16	1.19	1.41
0.98	0.82	1.22	0.91	1.26	1.32
1.71	1.29	1.17	1.74	1.51	1.25

要求:(1) 计算这一比值 95% 的置信区间;

(2) 得出上述结论时作了什么假设;

(3) 能否以 95% 的置信水平说明新酵素的产出率提高了。

解：(1) 计算得到 $\bar{x} = 1.268$, $s = 0.228$, 置信度为 95% 时 $z_{\alpha/2} = 1.96$, 故置信区间为 $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 1.268 \pm 1.96 \left(\frac{0.228}{\sqrt{6}} \right)$ 得 $1.194 < \mu < 1.342$ 。

(2) 假设 36 批的样本是随机的。

(3) 说明新的酵素的产出率提高了, 因为置信下限已超过 1。

三、正态总体, 方差未知, 小样本

总体为正态分布时, 小样本的样本均值仍服从正态分布, 即 $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, 而当方差 σ^2 未知, 用 S^2 代替时, 由上一章的抽样分布可知 $\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ 就服从 t 分布, 自由度为 $n - 1$, 也即 $P(t_{1-\alpha/2}(n-1) < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < t_{\alpha/2}(n-1)) = 1 - \alpha$, 因此 μ 的置信区间为:

$$\bar{X} \pm t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}} \quad (5.4)$$

t 分布的置信区间要宽于正态分布, 随着样本量 n 的增大就逐渐趋向于正态分布。不同样本量的情况下, 95% 置信度, t 分布的置信区间如表 5.2

表 5.2 不同样本量 t 分布 95% 置信度的区间

样本量(n)	自由度	μ 的 95% 的置信区间
5	4	$\bar{X} \pm 2.78 \left(\frac{S}{\sqrt{n}} \right)$
10	9	$\bar{X} \pm 2.26 \left(\frac{S}{\sqrt{n}} \right)$

续表

样本量(n)	自由度	μ 的 95% 的置信区间
20	19	$\bar{X} \pm 2.09 \left(\frac{S}{\sqrt{n}} \right)$
30	29	$\bar{X} \pm 2.05 \left(\frac{S}{\sqrt{n}} \right)$
∞	∞	$\bar{X} \pm 1.96 \left(\frac{S}{\sqrt{n}} \right)$

其他的样本量可查阅书后附录 t 分布表。

【例 5.3】 为研究独生子女的每月零花钱, 从某小学随机抽取了 20 个独生子女的家庭, 得到 $\bar{x} = 107$, $s = 40$, 试以 95% 的置信度估计该校独生子女小学生家庭平均每月零花钱的置信区间。

解: 因为 t 分布适用于正态总体, 因此研究这一问题应首先假设独生子女家庭的子女零花钱应服从正态分布, 在小样本, 总体方差未知用 S^2 代替时, $\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$, 由公式(5.4)其置信区间为:

$$\bar{x} \pm t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} = 107 \pm 2.09 \left(\frac{40}{\sqrt{20}} \right) = (88.30 \sim 125.7 \text{ 元})$$

【例 5.4】 某灯泡厂要估计一种新型灯泡的平均使用寿命, 在生产线上随机抽取 9 个灯泡进行测试, 取得了下列使用时间数据(小时):

5 100	5 100	5 400
5 260	5 400	5 100
5 320	5 180	4 940

假设灯泡的使用寿命服从正态分布, 以 95% 的置信度, 说明这种新型灯泡平均使用寿命的置信区间。

解: 经计算 $\bar{x} = 5 200$, $s = 156.5$, 因为 $1 - \alpha = 0.95$, $t_{\alpha/2}(8) = 2.306$, 这种新型灯泡平均使用寿命的置信区间为 $\bar{x} \pm$

$$t_{a/2} (n-1) \frac{s}{\sqrt{n}} = 5200 \pm 2.306 \left(\frac{156.5}{3} \right), \text{ 即 } 5079 \sim 5320 \text{ 小时之间。}$$

若总体为非正态分布,在小样本的情况下样本均值就不近似正态分布,因此既不能使用正态分布,又不能使用 t 分布,如果需要估计总体均值的置信区间,在总体方差已知的情况下可用切比雪夫不等式

$$P \left(| \bar{X} - \mu | < K \frac{\sigma}{\sqrt{n}} \right) \geq 1 - \frac{1}{K^2} \quad (5.5)$$

【例 5.5】 在例 5.4 的数据中假定总体不服从正态分布,计算的样本标准差作为总体标准差,试估计 95% 置信度的置信区间。

解:因为 $1 - \frac{1}{K^2} = 0.95$, 所以 $K^2 = 20$ $K = 4.472$,

$$\bar{x} \pm K \frac{\sigma}{\sqrt{n}} = 5200 \pm 4.472 \left(\frac{156.5}{3} \right)$$

即 4966.7—5433.3,显然其置信区间比 t 分布更宽。

总体均值区间估计的置信区间归纳如表 5.3。

表 5.3 不同情况下总体均值的区间估计

总体分布	样本量	σ 已知	σ 未知
正态分布	大样本 ($n \geq 30$)	$\bar{x} \pm z_{a/2} \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm z_{a/2} \frac{s}{\sqrt{n}}$
	小样本 ($n < 30$)	$\bar{x} \pm z_{a/2} \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{a/2} \frac{s}{\sqrt{n}}$
非正态分布	大样本 ($n \geq 30$)	$\bar{x} \pm z_{a/2} \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm z_{a/2} \frac{s}{\sqrt{n}}$
	小样本 ($n < 30$)	$\bar{x} \pm K \frac{\sigma}{\sqrt{n}}$	置信度为 $1 - \frac{1}{K^2}$

第三节 总体比例的区间估计

总体比例也是常见的参数之一,它是指总体单元中具有某种特征的单元数所占的比例。实践中研究这样的问题是很多的,例如需要了解一批产品中次品的比例或一等品的比例;一个学校的学生中女生所占的比例;一片林木中受某种病害的比例;新栽树木中成活的比例;企业中亏损企业所占的比例等等。它实际上是前面所说总体均值的一种特例,即每个单元的变量值是一种属性变量,设为 X ,当单元具有某种特征时用 1 表示,不具有某种特征用 0 表示。若总体单元数为 N ,则总体的比例 $p = \frac{1}{N} \sum_{i=1}^N x_i$,若从中随机抽取 n 个单元作样本,则样本比例 $P = \frac{1}{n} \sum_{i=1}^n X_i$,可以证明样本比例是总体比例的一个无偏估计量,可以在此基础上进行区间估计。由于样本大小不同与抽样方式不同,在进行区间估计时可分几种不同情况。

一、大样本重复抽样的估计方法

由上一章样本比例的抽样分布可知,当样本容量 n 增大时,样本比例趋近正态分布,其方差为 $\frac{pq}{n}$ ($q = 1 - p$),见公式(4.2.a),因此有:

$$P\left(p - z_{\alpha/2} \sqrt{\frac{pq}{n}} < P < p + z_{\alpha/2} \sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

$$P\left(|p - P| < z_{\alpha/2} \sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

和样本均值的置信区间一样,在抽样取得样本比例 P 后,置信水平为 $1 - \alpha$,总体比例 p 的置信区间为

$$P \pm Z_{\alpha/2} \sqrt{\frac{pq}{n}} \quad (5.6.a)$$

但是这里需要说明,在(5.6.a)确定置信区间时需要用到总体比例 p ,而这总体比例 p 恰恰是待估的参数,因此是未知的。由于在大样本时总体比例与样本比例相差不大,可以用样本比例来代替总体比例,因此区间估计的界限就改为

$$P \pm z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}} \quad (5.6.b)$$

【例 5.6】 某地区调查下岗女工中女性的比例,随机抽选了 36 个下岗工人,其中 20 人为女性。要求估计:(1)下岗工人中女性比例的点估计;(2)以 95% 的置信系数估计该地区下岗工人中女性比例的置信区间;(3)能否认为下岗工人中女性超过男性?

解:(1)以样本的比例作为总体比例的点估计, $\hat{p} = P = \frac{20}{36} = 0.5555 = 0.56$ 。

$$(2) S_p = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{(0.56)(0.44)}{36}} = \sqrt{\frac{0.2464}{36}} = 0.083$$

置信系数 $1 - \alpha = 0.95$ $z_{\alpha/2} = 1.96$, 置信区间为

$$P \pm z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}} = 0.56 \pm 1.96(0.083) = (0.40—0.72)$$

(3) 由于该区间包括了 0.5, 说明 p 有可能低于 0.5 因此尚不能肯定下岗工人中的女性比例超过男性。

二、大样本不重复抽样

当采用不重复抽样,大样本时 P 也近似正态分布,在抽样比 $\frac{n}{N}$ 不能忽略不计时, P 的抽样方差应乘以有限总体不重复抽样的修正系数即 $S_p = \sqrt{\frac{P(1-P)}{n} \cdot \frac{N-n}{N-1}}$, 置信系数为 $1 - \alpha$ 时, p 的置信区间为

$$P \pm z_{\alpha/2} \sqrt{\frac{P(1-P)}{n} \left(\frac{N-n}{N-1} \right)} \quad (5.7)$$

【例 5.7】 某企业欲实行一项改革,在职工中征求意见,整个企业有 1 000 名职工,随机抽取了 200 人,其中有 120 人表示同意,80 人表示反对。试求:(1)同意改革的职工占总职工人数比例的点估计。(2)以 95% 的置信系数确定同意人数比例的置信区间,能否认为同意的人数超过半数?

解:(1)样本比例估计值 $P = \frac{120}{200} = 0.6$ 可以作为同意人数的点估计。

(2) $n = 200$ 为大样本, P 的抽样分布趋近于正态分布, 抽样比 $\frac{200}{1000} = 0.2 > 0.05$, 计算方差应考虑有限总体不重复抽样的修正系数。

$$s_p = \sqrt{\frac{P(1-P)}{n} \left(\frac{N-n}{N-1} \right)} = \sqrt{\frac{(0.6)(0.4)}{200} \left(\frac{1000-200}{1000-1} \right)} = 0.031$$

当置信系数为 95% 时, $Z_{\alpha/2} = 1.96$, p 的置信区间为

$$P \pm z_{\alpha/2} \sqrt{\frac{P(1-P)}{n} \left(\frac{N-n}{N-1} \right)} = 0.6 \pm 1.96(0.031) = (0.54—0.66)$$

其中下限 0.54 已超过 0.5, 可以认为同意的人数比例已超过 50%。

当总体很大, 抽样比 $\frac{n}{N} < 5\%$ 时, 虽为不重复抽样, 其修正系数也可忽略不计。

三、比例置信区间的一些特殊情况

1. 稀有事件的小比例估计问题。若总体中具有某种特征的单元数很少, 因而 P 很小, 即使当 n 很大时 $nP \leq 5$ 。这时 P 就不宜用正态分布近似计算。由概率论的知识可知, 这时 n 个样本单元中具有某种特征的单元数 X 服从泊松分布, 可由泊松分布来求置信区间。泊松分布的置信区间通常已编制成表, 可以直接查阅。