4.4 区间估计

4.4.1 基本概念

如前所述,点估计是用一个点(即一个数)去估计未知参数.顾名思义,区间估计就是用一个区间去估计未知参数,即把未知参数值估计在某两界限之间.例如,估计一个人的年龄在30到35岁之间;估计所需费用在1000—1200元之间等等.区间估计是一种很常用的估计形式,其好处是把可能的误差用醒目的形式标出来了.你估计费用需1000元,我相信多少会有误差.误差多少?单从你提出的1000这个数字还给不出什么信息,你若估计费用在800~1200元之间,则人们会相信你在作出这估计时,已把可能出现的误差考虑到了,多少给人们以更大的信任感.

现今最流行的一种区间估计理论是原籍波兰的美国统计学家J.奈曼在本世纪30年代建立起来的.他的理论的基本概念很简单.为书写简单计,我们暂设总体分布只包含一个未知参数θ,且要估计的就是θ本身.如果总体分布包含若干个未知参数$\theta_1,\cdots,\theta_k$,而要估计的是$g(\theta_1,\cdots,\theta_k)$,基本概念并无不同.这将在后面的例子中看到.

设$X_1,\cdots,X_n$是从该总体中抽出的样本.所谓θ的区间估计,就是满足条件$\hat{\theta}_1(X_1,\cdots,X_n)\leqslant\hat{\theta}_2(X_1,\cdots,X_n)$的两个统计量$\hat{\theta}_1$,$\hat{\theta}_2$为端点的区间$[\hat{\theta}_1,\hat{\theta}_2]$.一旦有了样本$X_1,\cdots,X_n$,就把$\theta$估计在区间$[\hat{\theta}_1(X_1,\cdots,X_n),\hat{\theta}_2(X_1,\cdots,X_n)]$之内,不难理解,这里有两个要求:

1. θ要以很大的可能性落在区间$[\hat{\theta}_1,\hat{\theta}_2]$内,也就是说,概率

$$P_\theta(\hat{\theta}_1(X_1,\cdots,X_n)\leqslant\theta\leqslant\hat{\theta}_2(X_i,\cdots,X_n)) \qquad (4.1)$$

要尽可能大.

2. 估计的精密度要尽可能高.比方说,要求区间的长度$\hat{\theta}_2-\hat{\theta}_1$尽可能小,或某种能体现这个要求的其他准则.

例如,估一个人的年龄在某一区间内,例如[30,35]内.我们要求这估计尽量可靠,即该人的年龄有很大把握确在这区间内,同时,也要求区间不能太长:比如,估计一人的年龄在10—90岁之间,当然可靠了,但精度太差,用处不大.

但这两个要求是相互矛盾的.区间估计理论和方法的基本问题,莫不在于在已有的样本资源的限制下,怎样找出更好的估计方法,以尽量提高此二者——可靠性和精度,但终归有一定的限度.奈曼所提出并为现今所广泛接受的原则是:先保证可靠度,在这个前提下尽量使精度提高.为此他引进了如下的定义.

定义 4.1 给定一个很小的数$\alpha>0$.如果对参数θ的任何值,概率(4.1)都等于$1-\alpha$,则称区间估计$[\hat{\theta}_1,\hat{\theta}_2]$的置信系数为

$1-\alpha$.

区间估计也常称为“置信区间”.字面上的意思是:对该区间能包含未知参数 θ 可置信到何种程度.

有时,我们无法证明概率(4.1)对一切 θ 都恰好等于 $1-\alpha$,但知道它不会小于 $1-\alpha$,则我们称 $1-\alpha$ 是$[\hat{\theta}_1,\hat{\theta}_2]$的“置信水平”.按此,置信水平不是一个唯一的数.因为,若概率(4.1)总不小于0.8,那它也总不小于0.7,0.6,…等.就是说,若 β 为置信水平,则小于 β 的数也是置信水平,置信系数是置信水平中的最大者.在实用上,人们并不总是把这两个术语严加区别,这要看各人的习惯.

定义4.1中的 α,一般以取为0.05的最多,还有0.01,0.10,以至0.001等,也视情况需要而使用.这几个数字本身并无特殊意义,主要是这样标准化了以后对造表方便.

区间估计理论的主要问题,按奈曼的上述原则,就是在保证给定的置信系数之下,去寻找有优良精度的区间估计.而这个“优良”,也可以有种种准则.这方面现已有了一些结果,但在本课程范围之内,我们无法去涉及这些较深的理论问题,我们所能做的,就是从直观出发如何去构造看来是合理的区间估计.这就是下面两段要讨论的问题.

4.4.2 枢轴变量法

从一个简单例子入手.设 $X_1,\cdots,X_n$ 为抽自正态总体 $N(\mu,\sigma^2)$的样本,σ^2 已知,要求 μ 的区间估计.

先找一个 μ 的良好的点估计.在此可选择样本均值$\overline{X}$.由总体为正态易知

$$\sqrt{n}(\overline{X}-\mu)/\sigma \sim N(0,1) \tag{4.2}$$

以 Φ 记$N(0,1)$的分布函数.对 $0<\beta<1$(一般是 β 很小),用方程

$$\Phi(u_\beta)=1-\beta \tag{4.3}$$

定义记号 u_β. u_β 称为分布$N(0,1)$的“上 β 分位点”.其意义是:

$N(0,1)$分布中大于 u_β 的那部分的概率,就是 β. 图 4.2 中画出的是 $N(0,1)$的密度函数 $\varphi(x)=(\sqrt{2\pi})^{-1}e^{-x^2/2}$的图形,涂黑部分标出的面积为 β.

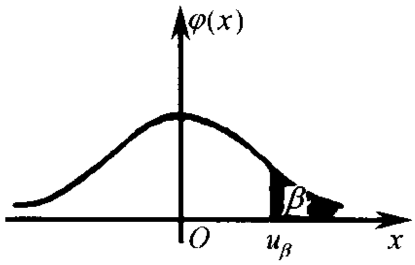

图 4.2

上 β 分位点的概念可推广到任何分布 F:满足条件 $F(v_\beta)=1-\beta$ 的点 v_β,就是分布函数 F 的上 β 分位点. 在数理统计学的应用中,除正态分布外,"统计三大分布"的上分位点很常用. 以后,我们分别用 $\chi_n^2(\beta)$, $t_n(\beta)$和 $F_{n,m}(\beta)$记自由度 n 的卡方分布,自由度 n 的 t 分布,以及自由度为(n,m)的 F 分布的上 β 分位点,这些都有表可查.

另外,读者还须注意:在有的著作中使用"下分位点",分布函数 F 的下 β 分位点是指满足条件 $F(w_\beta)=\beta$ 的点 w_β. 上、下分位点之间的换算不难:分布 F 的 β 下分位点,就是其 $1-\beta$ 上分位点. 当分布 F 的密度函数 f 关于原点对称(即 $f(-x)=f(x)$)时,F 的上、下 β 分位点只相差一个符号,本书以后只使用上分位点.

现在回到 μ 的区间估计问题. 由(4.2)及 μ_β 的定义,并注意到 $\Phi(-t)=1-\Phi(t)$,有

$$\begin{aligned}P\left(-u_{\alpha/2}\leqslant\sqrt{n}(\overline{X}-\mu)/\sigma\leqslant u_{\alpha/2}\right)&=\Phi(u_{\alpha/2})-\Phi(-u_{\alpha/2})\\&=(1-\alpha/2)-\alpha/2=1-\alpha\end{aligned}$$

此式可改写为

$$P\left(\overline{X}-\sigma u_{\alpha/2}/\sqrt{n}\leqslant\mu\leqslant\overline{X}+\sigma u_{\alpha/2}/\sqrt{n}\right)=1-\alpha$$

此式指出

$$[\hat{\theta}_1,\hat{\theta}_2]=\left[\overline{X}-\sigma u_{\alpha/2}/\sqrt{n},\overline{X}+\sigma u_{\alpha/2}/\sqrt{n}\right]\qquad(4.4)$$

可作为 μ 的区间估计,置信系数为 $1-\alpha$.

由这个例子悟出一种找区间估计的一般方法,可总结为以下几条:

1° 找一个与要估计的参数 $g(\theta)$有关的统计量 T,一般是其

一良好的点估计(此例 T 为$\overline{X}$);

2° 设法找出 T 和$g(\theta)$的某一函数 $S(T,g(\theta))$,其分布 F 要与θ 无关(在此例中,$S(T,g(\theta))$为$\sqrt{n}(\overline{X}-\mu)/\sigma$,分布 F 就是Φ). S 称为"枢轴变量";

3° 对任何常数 $a<b$,不等式 $a\leqslant S(T,g(\theta))\leqslant b$ 要能改写为等价的形式 $A\leqslant g(\theta)\leqslant B$,$A$,$B$ 只与 T,a,b 有关而与θ 无关;

4° 取分布 F 的上 $\alpha/2$ 分位点 $w_{\alpha/2}$ 和上$(1-\alpha/2)$分位点 $w_{1-\alpha/2}$. 有 $F(w_{\alpha/2})-F(w_{1-\alpha/2})=1-\alpha$. 因此

$$P(w_{1-\alpha/2}\leqslant S(T,g(\theta))\leqslant w_{\alpha/2})=1-\alpha$$

根据第 3 条,不等式 $w_{1-\alpha/2}\leqslant S(T,g(\theta))\leqslant w_{\alpha/2}$可改写为 $A\leqslant g(\theta)\leqslant B$ 的形式,A,B 与 T 有关因而与样本有关. $[A,B]$就是 $g(\theta)$的一个置信系数 $1-\alpha$ 的区间估计.

现在举一些例子来说明这个方法,这些例子包含了许多常用的重要区间估计.

例 4.1 从正态总体 $N(\mu,\sigma^2)$中抽样本 $X_1,\cdots,X_n$,μ 和σ^2 都未知,求 μ 的区间估计.

μ 的点估计仍取为样本均值$\overline{X}$. 作为枢轴变量,再取$\sqrt{n}(\overline{X}-\mu)/\sigma$ 已不行. 因为虽然这变量的分布 $N(0,1)$与参数无关,但因 σ 未知,条件 3°已不满足. 现把 σ 改为样本标准差 S,则枢轴变量一切条件都满足了,因为(见第二章(4.34))变量$\sqrt{n}(\overline{X}-\mu)/S$ 服从自由度为$n-1$ 的 t 分布,与参数无关. 由此出发用 4°,并注意 t 分布密度关于0 对称因而 $t_{n-1}(1-\alpha/2)=-t_{n-1}(\alpha/2)$,得 μ 的区间估计

$$\left[\overline{X}-St_{n-1}(\alpha/2)/\sqrt{n},\overline{X}+St_{n-1}(\alpha/2)/\sqrt{n}\right] \qquad (4.5)$$

置信系数为 $1-\alpha$. 它称为"一样本 t 区间估计".

例如,为估计一物件的重量 μ,把它在天平上重复秤了 5 次,得结果为(单位为克)

$$5.52, 5.48, 5.64, 5.51, 5.43$$

假定此天平无系统误差且随机误差服从正态分布. 则总体分布为

$N(\mu,\sigma^2)$，μ 即未知的重量，方差 σ^2 也未知. 算出

$$\overline{X} = (5.52 + \cdots 5.43)/5 = 5.516$$

$$S = \sqrt{\frac{1}{5-1}[(5.52-5.516)^2 + \cdots + (5.43-5.516)^2]}$$

$$= \frac{1}{2}\sqrt{0.02412} = 0.078$$

查表，知 $t_4(0.025) = 2.776$. 以这些数值代入(4.5)，得 μ 的置信系数 0.95 的区间估计为[5.419,5.613].

[5.419,5.613]是一个具体的区间，μ 是一个虽然未知，但其值确定的数. [5.419,5.613]这区间或者包含 μ，或者不包含，二者只居其一. 说这区间的置信系数为 0.95，其确切意义应当是：它是根据所有的数据，用一个其置信系数为 0.95 的方法作出的. 可见置信系数一词是针对方法：用这方法作出的区间估计，平均 100 次中 95 次确包含所要估计的值. 一旦算出具体区间，就不能再说它有 95% 的机会包含要估计的值了. 这一点意义上的理解必须分清，正如说一个人长于挑西瓜：他挑的瓜，平均 100 个中有 95 个好的. 某天他给你挑一个，结果或好或坏，必居其一，不是 95% 的好. 但是，考虑到他挑瓜的技术，我对他挑的比较放心，这就是置信系数.

区间估计(4.5)叫做一样本 t 区间估计. “一样本”是指这里只有一个总体，因而只有一组样本，以别于下例.

例 4.2 设有两个正态总体，其分布分别为 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$. 注意方差相同. 设 μ_1,μ_2,σ^2 都未知. 现从这两个总体分别抽出样本 $X_1,\cdots,X_n$ 和 $Y_1,\cdots,Y_m$. 要求 $\mu_1-\mu_2$ 的区间估计.

记 $\overline{X}$ 和 $\overline{Y}$ 分别为 X_i 和 Y_j 的样本均值，而

$$S = \left[\sum_{i=1}^{n}(X_i-\overline{X})^2 + \sum_{j=1}^{m}(Y_j-\overline{Y})^2\right]^{1/2} \Big/ \sqrt{n+m-2}$$

据第二章(4.36)式，知

$$T = \sqrt{\frac{mn}{m+n}} \cdot ((\overline{X}-\overline{Y}) - (\mu_1-\mu_2))/S \sim t_{n+m-2}$$

的分布不依赖于参数 μ_1,μ_2,σ^2. 它适合于作为枢轴变量的条件，按 4°，定出 $\mu_1-\mu_2$ 的区间估计为

$$\left[(\overline{X}-\overline{Y})-St_{n+m-2}(\alpha/2)\sqrt{\frac{n+m}{nm}},\right.$$

$$\left.(\overline{X}-\overline{Y})+St_{n+m-2}(\alpha/2)\sqrt{\frac{n+m}{nm}}\right] \tag{4.6}$$

置信系数为 $1-\alpha$. 这个区间称为“两样本 t 区间估计”，是应用上常用的区间估计之一.

如考虑上例，设有另一物件，其重量 μ_2 也未知. 在这同一架天平上秤 4 次，得结果为

$$5.45, 5.40, 5.34, 5.51$$

把上例中的 μ 记为 μ_1. 因是同一架天平，方差不变. 要对两物件重量之差 $\mu_1-\mu_2$ 作区间估计. 可用(4.6). 算出

$$\overline{Y}=(5.45+\cdots+5.51)/4=5.425$$

$$\sum_{j=1}^{n}(Y_j-\overline{Y})^2=(5.45-5.425)^2+\cdots+(5.51-5.425)^2$$

$$=0.01570$$

结合前例数据，算出

$$\overline{X}-\overline{Y}=-0.091, S=\sqrt{0.02412+0.01570}/\sqrt{5+4-2}$$

$$=0.075$$

又 $\sqrt{(n+m)/nm}=\sqrt{9/20}=0.671$. 取 $\alpha=0.05$，查 t 分布表得 $t_7(0.025)=2.365$. 把这些都代入(4.6)，算出 $\mu_1-\mu_2$ 的区间估计为 $[-0.210, 0.028]$，置信系数 0.95.

在实际问题中，两总体方差相等的假定往往只是近似成立. 当方差之比接近 1 时，用(4.6)产生的误差不大(这里的“误差”一词是指实际的置信系数与名义的置信系数 $1-\alpha$ 有出入). 如果差别较大，则必须假定两正态总体分别有方差 σ_1^2 和 σ_2^2，σ_1^2 和 σ_2^2 都未知. 在这样的假定下求 $\mu_1-\mu_2$ 的区间估计问题，是数理统计学上一个著名的问题，叫贝伦斯-费歇尔问题. 因为这两位学者分别在 1929 和 1930 年研究过这个问题，他们以及后来的研究者提出过

一些解法,但还没有一个被公认为是最满意的.

例 4.3 再考虑例 4.1,但现在要求作 σ^2 的区间估计.

据第二章(4.33),有 $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. 于是 $(n-1)S^2/\sigma^2$ 适合枢轴变量的条件. 按 4°,得 σ^2 的区间估计为

$$[(n-1)S^2/\chi^2_{n-1}(\alpha/2),(n-1)S^2/\chi^2_{n-1}(1-\alpha/2)] \tag{4.7}$$

置信系数为 $1-\alpha$. 类似地,若另有一正态总体 $N(\mu,\sigma_2^2)$ 及从中抽出的样本 $Y_1,\cdots,Y_m$,要作方差比 σ_1^2/σ_2^2 的区间估计. 记 S_1^2 和 S_2^2 分别为 $X_1,\cdots,X_n$ 和 $Y_1,\cdots,Y_m$ 的样本方差,按第二章(4.35),有

$$(S_2^2/\sigma_2^2)/(S_1^2/\sigma_1^2) \sim F_{m-1,n-1}$$

即 $\lambda \cdot S_2^2/S_1^2 \sim F_{m-1,n-1}$,其中 $\lambda=\sigma_1^2/\sigma_2^2$. 于是得到枢轴变量. 按 4°,得出比值 λ 的置信系数 $1-\alpha$ 的区间估计为

$$[(S_1^2/S_2^2)F_{m-1,n-1}(1-\alpha/2),(S_1^2/S_2^2)F_{m-1,n-1}(\alpha/2)] \tag{4.8}$$

例 4.4 设 $X_1,\cdots,X_n$ 为抽自指数分布总体的样本,要求其参数 λ 的区间估计.

在第二章 2.4.3 小节中曾证明 $2n\lambda\overline{X} \sim \chi^2_{2n}$. 故 $2n\lambda\overline{X}$ 可作为枢轴变量. 由 4°,得 λ 的区间估计为

$$[\chi^2_{2n}(1-\alpha/2)/(2n\overline{X}),\chi^2_{2n}(\alpha/2)/(2n\overline{X})] \tag{4.9}$$

置信系数为 $1-\alpha$. 若要求总体均值 $1/\lambda$ 的区间估计,则为

$$[2n\overline{X}/\chi^2_{2n}(\alpha/2),2n\overline{X}/\chi^2_{2n}(1-\alpha/2)] \tag{4.10}$$

从这些例子可以看出"枢轴变量法"这名称的由来. 拿本例来说,变量 $2n\lambda\overline{X}$ 起了一个"轴心"的作用,把一个变量(即 $2n\lambda\overline{X}$)介于某两个界限之间的不等式轻轻一转,就成为未知参数 λ 介于某两个界限之间的不等式.

对离散型变量来说,枢轴变量法不易使用. 不仅由于满足条件 1°—4°的枢轴变量 $S(T,g(\theta))$ 大多不存在,即使存在了,由于其分布 F 为离散,对指定的 β,一般也不一定存在确切的上 β 分位点. 对离散型总体的参数去找具有所指定的置信系数的区间估计方法,超出本书范围之外. 在下一段中,对二项和波哇松分布参数

这两个重要情况，将给出一种基于极限分布的方法.

在实用中，除了指定的置信系数外，往往还对于区间估计的长度，或其他某种反映其精度的量，有一定的要求. 在有些情况下这个问题比较好处理. 例如，$N(\mu,\sigma^2)$当σ^2已知时，μ的区间估计(4.4)的长为$2\sigma u_{\alpha/2}/\sqrt{n}$. 为要使这个长度不超过指定的$L>0$，只须取$n$为不小于$(2\sigma u_{\alpha/2}/L)^2$的最小整数即可.

对例4.3正态分布方差或方差比的估计，由于方差本身的意义，在实际问题中，考虑估计值与它相差多少倍，往往比考虑估计值与其差的绝对值更好. 这就要求，例如，区间(4.7)的右端不超过左端的L倍($L>1$)，即

$$\chi_{n-1}^2(\alpha/2)/\chi_{n-1}^2(1-\alpha/2) \leqslant L$$

在给定了L之后，可以查χ^2分布表，找一个最小的n使上式成立即可. 对方差比的情况，以及指数分布参数λ(或$1/\lambda$)的情况，也完全类似地处理.

对t区间估计，则情况不同. 拿一样本t区间估计(4.5)来说，其长$2St_{n-1}(\alpha/2)/\sqrt{n}$与$S$有关，而$S$与样本有关，故无法决定这样一个$n$，它能保证在任何情况下都有$2St_{n-1}(\alpha/2)/\sqrt{n}\leqslant L$. 1945年，美国统计学家斯泰因提出了一个“两阶段抽样”的方法来解决这个问题：先抽出样本$X_1,\cdots,X_n$，算出样本标准差S如前. 根据S的大小决定追加抽样的数目：S愈大，追加抽样次数愈多. 具体公式如下：先引进记号$[\alpha]$=不超过α的最大整数，例如$[3.12]=3$，$[2]=2$等，追加抽样次数m的公式为

$$m=\begin{cases}0, & \text{若 } n\geqslant\left[4t_{n-1}^2(\alpha/2)S^2/L^2\right]+1\\ n-1-\left[4t_{n-1}^2(\alpha/2)S^2/L^2\right], & \text{其他情况}\end{cases}$$

记原有样本和追加样本全体的样本均值为$\widetilde{X}$，则可以证明，长为L的区间估计$[\widetilde{X}-L/2,\widetilde{X}+L/2]$有置信系数$1-\alpha$.

4.4.3 大样本法

大样本法就是利用极限分布，主要是中心极限定理，以建立枢

轴变量,它近似满足枢轴变量的条件 2°.最好通过例子来说明.

例 4.5 某事件 A 在每次试验中发生的概率为 p.作 n 次独立试验,以 Y_n 记 A 发生的次数,要求 p 的区间估计.

设 n 相当大,则按定理 4.3,近似地有 $(Y_n-np)/\sqrt{np(1-p)}\sim N(0,1)$.于是 $(Y_n-np)/\sqrt{np(1-p)}\sim N(0,1)$ 可取为枢轴变量.由

$$P\left(-u_{\sigma/2}\leqslant(Y_n-np)\Big/\sqrt{np(1-p)}\leqslant u_{\sigma/2}\right)\approx 1-\alpha \tag{4.11}$$

可改写为

$$P(A\leqslant p\leqslant B)\approx 1-\alpha \tag{4.12}$$

其中 A,B 是二次方程

$$(Y_n-np)^2/(np(1-p))=u_{\alpha/2}^2$$

的两个根,即

$$A,B=\frac{n}{n+u_{\alpha/2}^2}\left(\hat{p}+\frac{u_{\alpha/2}^2}{2n}\pm u_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{u_{\alpha/2}^2}{4n^2}}\right) \tag{4.13}$$

A 取负号,B 取正号,$\hat{p}=Y_n/n$.

因为(4.11)和(4.12)只是近似的,故区间估计 $[A,B]$ 的置信系数,也只是近似地等于 $1-\alpha$.当 n 较大,例如 $n\geqslant30$ 时,相去不远,实际上,n 太小时,找 p 的区间估计意义不大.因为这种区间都失之过长,实际意义不大.这可由下面的分析看出:由于 $0\leqslant\hat{p}\leqslant1$,$\hat{p}(1-\hat{p})$ 的最大值可为 $1/4$.这时,区间 $[A,B]$ 之长,在把 $\hat{p}(1-\hat{p})$ 改为 $1/4$ 后,为 $u_{\alpha/2}\Big/\sqrt{n+u_{\alpha/2}^2}$.取 $\alpha=0.05$,有 $u_{\alpha/2}=1.96$.若要求这区间之长不超过 0.3(这是一个很低的要求),必须 $1.96/\sqrt{n+(1.96)^2}\leqslant0.3$.算出 n 至少应为 39.可以看出:在试验次数 n 低于 40 时,求 p 的区间估计没有多大实用意义.

例 4.6 设 $X_1,\cdots,X_n$ 为抽自有波哇松分布 $P(\lambda)$ 的总体的样本,求 λ 的区间估计.

记 $Y_n = X_1 + \cdots + X_n$. 设 n 相当大,注意到波哇松分布的均值方差都是 λ,由第三章定理 4.2,知$(Y_n - n\lambda)\big/\sqrt{n\lambda}$ 近似地有分布 $N(0,1)$. 仿前例的做法,即得到 λ 的区间估计$[A,B]$,A,B 为二次方程

$$(Y_n - n\lambda)^2 = n\lambda u_{\alpha/2}^2$$

的两根,即

$$A,B = \overline{X} + u_{\alpha/2}^2/(2n) \pm u_{\alpha/2}\sqrt{u_{\alpha/2}^2/(4n^2) + \overline{X}/n} \tag{4.14}$$

A 取负号,B 取正号,$\overline{X} = Y_n/n$.

例 4.7 设某总体有均值 θ,方差 σ^2. θ 和 σ^2 都未知,从这总体中抽出样本 $X_1,\cdots,X_n$,要作 θ 的区间估计.

因为对总体分布没有作任何假定,要作出满足条件 1°—4°的枢轴变量是不可能的. 但是,若 n 相当大,则据中心极限定理(第三章定理 4.2),有$\sqrt{n}(\overline{X} - \theta)/\sigma \sim N(0,1)$. 但此处 σ 未知,仍不能以$\sqrt{n}(\overline{X} - \theta)/\sigma$ 作为枢轴变量. 因为 n 相当大,样本均方差 S 是σ 的一个相合估计,故可近似地用 S 代σ,得

$$\sqrt{n}(\overline{X} - \theta)/S \sim N(0,1)$$

由此就不难得出 θ 的区间估计

$$\left[\overline{X} - S\mu_{\alpha/2}\big/\sqrt{n}, \overline{X} + S\mu_{\alpha/2}\big/\sqrt{n}\right]$$

它的置信系数,当 n 相当大时,近似地为 $1-\alpha$. 近似的程度如何不仅取决于 n 的大小,还要看总体的分布如何.

例 4.8 考虑在例 4.2 中提出的贝伦斯-费歇尔问题:$X_1,\cdots,X_n$ 是从正态总体$N(\mu,\sigma_1^2)$中抽出的样本,$Y_1,\cdots,Y_m$ 是从正态总体 $N(\mu_2,\sigma_2^2)$中抽出的样本,要求 $\mu_1 - \mu_2$ 的区间估计.

在本例中有

$$\left[(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)\right]\Big/\sqrt{\sigma_1^2/n + \sigma_2^2/m} \sim N(0,1) \tag{4.15}$$

这里没有近似:分布是严格成立的. 但是,由于 σ_1,σ_2 未知,(4.15)

并不构成枢轴变量.如果 n,m 都相当大,则 σ_1^2 和 σ_2^2 分别可用 X 样本的样本方差 S_1^2 和 Y 样本的样本方差 S_2^2 近似地代替之,得

$$[(\overline{X}-\overline{Y})-(\mu_1-\mu_2)]\Big/\sqrt{S_1^2/n+S_2^2/m}\sim N(0,1) \tag{4.16}$$

与(4.15)不同,(4.16)只是近似而非严格.(4.16)可作为枢轴变量,而得出 $\mu_1-\mu_2$ 的区间估计.当然,其置信系数只是近似的.

例 4.5—4.8 所导出的区间估计,叫"大样本区间估计".一般如果一个统计方法是基于有关变量的当样本大小 n 很大时的极限分布,则称这一统计方法为"大样本方法".反之,若依据的是有关变量的确切分布,则称为"小样本方法".如例 4.1—4.4 导出的区间估计就是小样本区间估计.这不在于 n 多大多小:在例 4.1—4.4 中,即使样本大小 $n=10^{10}$,仍是小样本方法.对例 4.5 而言,因使用的是极限分布,即使 $n=40$,仍算是大样本方法,不言而喻,大样本方法只有在样本大小较大时才宜于使用.

4.4.4 置信界

在实际问题中,有时我们只对参数 θ 的一端的界限感兴趣.例如,θ 是在一种物质中某种杂质的百分率,则我们可能只关心其上界,即要求找到这样一个统计量 $\overline{\theta}$,使 $\{\theta\leqslant\overline{\theta}\}$ 的概率很大. $\overline{\theta}$ 就称为 θ 的置信上界(或上限).又如,θ 是某种材料的强度,则我们可能只关心其下界,即要求找到这样一个统计量 $\underline{\theta}$,使 $\{\theta\geqslant\underline{\theta}\}$ 的概率很大. $\underline{\theta}$ 就称为 θ 的置信下界(或下限).下面给出正式的定义,为行文简单,就以一个参数 θ 的情况为例.

定义 4.2 设 $X_1,\cdots,X_n$ 是从某一总体中抽出的样本,总体分布包含未知参数 θ,$\overline{\theta}=\overline{\theta}(X_1,\cdots,X_n)$ 和 $\underline{\theta}=\underline{\theta}(X_1,\cdots,X_n)$ 都是统计量(它们与 θ 无关),则

1. 若对 θ 的一切可取的值有

$$P_\theta(\overline{\theta}(X_1,\cdots,X_n)\geqslant\theta)=1-\alpha \tag{4.17}$$

则称 $\overline{\theta}$ 为 θ 的一个置信系数为 $1-\alpha$ 的置信上界;

2. 若对 θ 的一切可取的值有

$$P_\theta(\underline{\theta}(X_1,\cdots,X_n)\leqslant\theta)=1-\alpha \tag{4.18}$$

则称$\underline{\theta}$为 θ 的一个置信系数为 $1-\alpha$ 的置信下界.

把(4.17)与(4.18)与区间估计的置信系数定义去比较,看出:置信上、下界无非是一种特殊的置信区间,其一端为∞或$-\infty$.因此,前面用于求区间估计的方法,都很容易平行地移至此处.例如,找 $N(\mu,\sigma^2)$ 的均值 μ 的置信下界,假定 σ^2 已知,以 $\sqrt{n}(\overline{X}-\mu)/\sigma$ 为枢轴变量,其分布为 $N(0,1)$.有

$$P(\sqrt{n}(\overline{X}-\mu)/\sigma\leqslant u_\alpha)=1-\alpha$$

此式可改写为

$$P(\mu\geqslant\overline{X}-u_\alpha\sigma/\sqrt{n})=1-\alpha \tag{4.19}$$

把(4.19)与(4.18)比较,即知$\overline{X}-u_\alpha\sigma/\sqrt{n}$为 μ 的一个置信下界,置信系数为 $1-\alpha$.将这个方法用于以前讨论过的诸例,得出一些置信上、下界的结果,例如(记号均见有关各例):

1. **例 4.1** μ 的置信上、下界分别为(正号为上界)

$$\overline{X}\pm St_{n-1}(\alpha)/\sqrt{n}$$

2. **例 4.2** $\mu_1-\mu_2$ 的置信上、下界分别为(正号为上界)

$$(\overline{X}-\overline{Y})\pm St_{n+m-2}(\alpha)\sqrt{\frac{m+n}{mn}}$$

3. **例 4.3** σ^2 的置信上界为$(n-1)S^2/\chi_{n-1}^2(1-\alpha)$,下界为$(n-1)S^2/\chi_{n-1}^2(\alpha)$.

以上置信系数都是 $1-\alpha$,其余各例都与此类似,我们注意到一点:在置信区间中的 $\alpha/2$ 在这里都被 α 取代.这是由于区间估计是双侧的.共为 α 的概率由两边均分,各占 $\alpha/2$.而置信界则是单侧的.

4.4.5 贝叶斯法

用贝叶斯法处理统计问题的基本思想,已在 4.2 节 4.2.4 中阐述过了.用它来处理区间估计问题,概念上和做法上都很简单:

沿用4.2节4.2.4中的记号.在有了先验分布密度 $h(\theta)$和样本 $X_1,\cdots,X_n$ 后,算出后验密度 $h(\theta|X_1,\cdots,X_n)$.再找两个数$\hat{\theta}_1,\hat{\theta}_2$都与 $X_1,\cdots,X_n$ 有关,使

$$\int_{\hat{\theta}_1}^{\hat{\theta}_2} h(\theta|X_1,\cdots,X_n)\mathrm{d}\theta = 1-\alpha \tag{4.20}$$

区间$[\hat{\theta}_1,\hat{\theta}_2]$的意思是:在所得后验分布之下,$\theta$ 落在这区间内的概率为 $1-\alpha$.因此,$[\hat{\theta}_1,\hat{\theta}_2]$可作为 θ 的一区间估计,其后验信度为 $1-\alpha$."后验"是指"有了样本以后"的意思.因此,所谓"后验信度为 $1-\alpha$",可以解释为:在已有了样本以后,我对区间$[\hat{\theta}_1,\hat{\theta}_2]$能包含未知参数 θ 的相信程度为 $1-\alpha$.这与奈曼理论中的置信系数的含义相似,但理论观念上有别.因为这里整个架构根本不同.

如果要找贝叶斯上下界,则只须把(4.20)分别改为

$$\int_{-\infty}^{\hat{\theta}} h(\theta|X_1,\cdots,X_n)\mathrm{d}\theta = 1-\alpha(\text{上界}) \tag{4.21}$$

和

$$\int_{\hat{\theta}}^{\infty} h(\theta|X_1,\cdots,X_n)\mathrm{d}\theta = 1-\alpha(\text{下界}) \tag{4.22}$$

对(4.20)而言还有一个问题:满足条件(4.20)的$\hat{\theta}_1,\hat{\theta}_2$很多,如何决定一对?一般是以使$\hat{\theta}_1-\hat{\theta}_2$ 最小为原则*(也可以是使$\hat{\theta}_2/\hat{\theta}_1$最小,这要看参数的性质与实际问题中的要求如何而定).下面将通过例子解释这一点.

例 4.9 考虑例2.14.在该例中所规定的先验分布之下,找 θ

---

* 另一种可取的方法是找$\hat{\theta}_1,\hat{\theta}_2$,使

$$\int_{-\infty}^{\hat{\theta}_1} h(\theta|X_1,\cdots,X_n)\mathrm{d}\theta = \alpha/2,\quad \int_{\hat{\theta}_2}^{\infty} h(\theta|X_1,\cdots,X_n)\mathrm{d}\theta = \alpha/2$$

的区间估计.

在该例中已找出 θ 的后验分布为 $N(t,\eta^2)$, t, η^2 分别由(2.17),(2.18)决定,这个密度函数在 t 点处达到最大值,然后在两边对称地下降.由此易见,如要找 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 满足(4.20)式,它只有在 $\hat{\theta}_1,\hat{\theta}_2=t\pm c$ 时才能使 $\hat{\theta}_2-\hat{\theta}_1$ 最小.由正态分布即知,c 必须取为 $\eta\mu_{\alpha/2}$.于是得出贝叶斯区间估计

$$[t-\eta\mu_{\alpha/2},t+\eta\mu_{\alpha/2}]$$

其后验信度为 $1-\alpha$.

例 4.10 考虑例 2.13.在此已求出当取 $R(0,1)$ 为先验分布时,p 的后验密度为

$$h(p|X_1,\cdots,X_n) = p^X(1-p)^{n-X}/\beta(X+1,n-X+1),0\leqslant p\leqslant 1 \quad (4.23)$$

要找 $\hat{p}_1,\hat{p}_2$,使

$$\int_{\hat{p}_1}^{\hat{p}_2} p^X(1-p)^{n-X}\mathrm{d}p/\beta(X+1,n-X+1)=1-\alpha$$

并使 $\hat{p}_2-\hat{p}_1$ 最小,问题就麻烦些.(4.23)的图形大致如图 4.3.它在点 $p=X/n$ 处达到最大,然后往两边下降.故只有图中 c,d 那种对子,才能使 $d-c$ 最小.方法是:先在 X/n 左边取定一个值 c.由方程

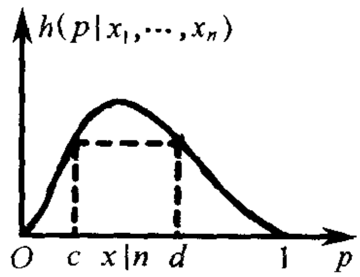

图 4.3

$$c^X(1-c)^{n-X}=p^X(1-p)^{n-X}$$

以 p 为未知量,解出 $p=d$.从图 4.3 看出,d 必大于 X/n.计算积分

$$\int_c^d p^X(1-c)^{n-X}\mathrm{d}p/\beta(X+1,n-X+1)=A$$

若 $A>1-\alpha$,表示 c 取得太小.若 $A<1-\alpha$,则表示 c 取得太大.经过几次调整后即可找到足够接近的近似值.

与奈曼的理论相比,我们看出,这里求区间估计的过程容易多

了.固然,在寻找适合(4.20)的$\hat{\theta}_1$和$\hat{\theta}_2$时,往往计算很繁,但并无原则困难,用计算机也很容易实现.但用奈曼的方法,则涉及到麻烦的分布问题.如例4.1—4.4这几个例,就基于有关的统计量服从t分布,卡方和F分布等.这不是常有的情况,而只是少见的几个特例(幸好这几个特例在实用中用得很多).往往由于分布问题无法解决,而只好求助于大样本理论.实用上往往样本不很大,使我们对由此而产生的误差(即实际的置信系数与名义的置信系数的距离)不甚了然.贝叶斯方法不存在这些问题.当然,贝叶斯方法有其自身的问题,即先验分布如何定,这一点我们在前面已提过了.