文章编号:1004-4337(2013)06-0681-02 **中图分类号**:R311 **文献标识码**:A · 统计分析 ·

非正态总体下的小样本区间估计问题

杭国明 祝国强

(复旦大学数学科学学院 上海 200433)

摘 要: 在非正态总体的条件下,给出的样本是小样本时,根据总体的不同情况,可以用确切概率计算法、Fisher正态近似法、切比雪夫不等式法等方法来确定总体未知参数的置信区间。

关键词: 小样本; 置信区间; 总体率; 样本率

doi:10.3969/j.issn.1004-4337.2013.06.017

参数估计问题是数理统计中的三大问题之一,其要解决的问题是对总体中的未知参数进行估计。在现实生活中,总体的分布可能是未知的;也可能根据以往的经验得知某总体的分布类型,但总体中的某些参数往往是未知的;还可能随着时间的推移,总体中的某些参数已发生了变化。例如,根据经验知,成年人的身高值服从正态分布 ,不同地区的成年人的身高值是不同的,同一地区的成年人在不同时期的身高值也是不一样的。

在许多情况下,我们需要对总体中的未知参数作一个估计。对总体未知参数的估计,通常有点估计和区间估计两种方法,本文只讨论区间估计问题。当总体分布已知时,常用的区间估计方法是从总体中抽取一个样本,构造一个合适的函数,根据该函数的分布,再得到未知参数的置信区间。而当总体分布未知时,区间估计可以借助切比雪夫不等式来解决。

通常,在区间估计问题中,我们根据总体分布类型的不同,把总体分成两类:一类是正态总体,另一类是非正态总体。不管是正态总体或是非正态总体,只要样本是大样本,那么根据中心极限定理,都可以把它归结为正态总体问题去处理。当遇到的总体为非正态总体或者总体的分布未知,而给出的样本又是小样本(样本容量)时,就要寻求其他的方法来求总体中未知参数的置信区间。

1 确切概率计算法

1.1 总体率 P 的置信区间

设某总体的容量为 N,其中具有某种特征的个体数为 M,则称 $P=\frac{M}{N}$ 为具有某种特征的个体的总体率,如治愈率、成功率等。

假设总体率 P 未知,如何求出总体率 P 的 $1-\alpha$ 置信区间呢?

我们先从总体中随机抽取一个样本容量为 n 的样本,假设其中具有某种特征的个体数为 m,则 $m\sim B(n,P)$。

设 $A=$ {具有某种特征的事件},令 $P(P\leqslant p_1)\leqslant\frac{\alpha}{2}$,$P(P\geqslant p_2)\leqslant\frac{\alpha}{2}$ 。对于事先给定的 α 值,只有 P 较小时,事件 A 发生的次数才较少,则此时事件 A 发生的次数大于等于 m 是小概率事件,于是满足方程 $\sum_{k=m}^{n}C_n^k p^k(1-p)^{n-k}=\frac{\alpha}{2}$ 的值就是 p_1,即 p_1 为总体率 P 的 $1-\alpha$ 置信区间下限;同理,只有 P 较大时,事件 A 发生的次数才较大,则此时事件 A 发生的次数小

7 Vernon RB, Lane TF, Angello JC, et al. Adhesion, shape, proliferation, and gene expression of mouse Leydig cells are influenced by extracellular matrix in vitro. Biol Reprod, 1991, 44(1): 157~170.

8 Nagle DL, McGrail SH, Vitale J, et al. The mahogany protein is a receptor involved in suppression of obesity. Nature, 1999, 398(6723): 148~152.

9 Gunn TM, Miller KA, He L, et al. The mouse mahogany locus encodes a transmembrane form of human attractin. Nature, 1999, 398(6723): 152~156.

10 Gunn TM, Barsh GS. Mahogany/attractin: en route from phenotype to function. Trends Cardiovasc Med, 2000, 10(2): 76~81.

11 Nakadate K, Sakakibara S, Ueda S. Attractin/mahogany protein expression in the rodent central nervous system. J Comp Neurol, 2008, 508(1): 94~111.

12 Lu X, Gunn TM, Shieh K, et al. Distribution of Mahogany/Attractin mRNA in the rat central nervous system. FEBS Lett, 1999, 462(1-2): 101~107.

13 隋聪, 熊承良, 沈士亮. 敲除Attractin或Mahoganoid基因对雄性小鼠生殖内分泌激素的影响. 中国计划生育学杂志, 2007, 15(12): 729~731.

收稿日期:2013-02-11

通讯作者:祝国强

* 第二军医大学 基础医学部 数理教研室

于等于 m 也是小概率事件，于是满足方程 $\sum_{k=0}^{m}C_n^k p^k(1-p)^{n-k}=\frac{\alpha}{2}$ 的值就是 p_2，即 p_2 为总体率 P 的 $1-\alpha$ 置信区间上限。这样，就可以得到总体率 P 的 $1-\alpha$ 置信区间为 (p_1,p_2)。

1.2　Poisson 分布参数 的置信区间

假设 $m\sim P(\lambda)$，其中 m 为样本的观察值。与情况 1 类似，满足方程 $\sum_{k=m}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}=\frac{\alpha}{2}$ 的值 λ_1 为总体率 P 的 $1-\alpha$ 置信区间下限，满足方程的 $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}=\frac{\alpha}{2}$ 的值 λ_2 为总体率 P 的 $1-\alpha$ 置信区间上限。这样，即得到总体率 P 的 $1-\alpha$ 置信区间为 (p_1,p_2)。

2　费歇(Fisher)正态近似法

设具有某种特征的总体率为 P，从总体中随机抽取一个样本容量为 n 的样本，其中具有某种特征的个体数为 m，则称 $p=\frac{m}{n}$ 为样本率。

作变换 $\varphi=2\arcsin\sqrt{p}$，则 φ(单位：弧度)近似服从正态分布 $N(\mu_\varphi,\sigma_\varphi^2)$，其中均数 $\mu_\varphi=\varphi=2\arcsin\sqrt{p}$，标准差 $\sigma_\varphi\approx\frac{1}{\sqrt{n}}$。由 $\left(\left|\frac{\varphi-\phi}{\sigma_\varphi}\right|\leqslant Z_{\frac{\alpha}{2}}\right)=1-\alpha$，可以得到 φ 的 $1-\alpha$ 置信区间为 (φ_1,φ_2)，其中 $\varphi_1=\varphi-Z_{\frac{\alpha}{2}}\sigma_\varphi$，(其中 $Z_{\frac{\alpha}{2}}$ 为标准正态分布的双侧临界值)，再算出 $p_1=\sin^2\frac{\varphi_1}{2}$，$p_2=\sin^2\frac{\varphi_2}{2}$，就可以得到总体率 P 的 $1-\alpha$ 置信区间为 (p_1,p_2)。

3　切比雪夫不等式法

当随机变量 X 的分布未知时，若 X 的数学期望 $E(X)$ 和方差 $V(X)$ 存在，且 $V(X)$ 已知，则可利用切比雪夫不等式 $P(|X-E(X)|<\varepsilon)\geqslant 1-\frac{V(X)}{\varepsilon^2}$(式中的 ε 为任意的正数)求出总体均数 $E(X)$ 的 $1-\alpha$ 置信区间。

假设从总体中抽的样本为 $X_1,X_2\cdots,X_n$，则 $E(\bar{X})=E(X)$，$V(\bar{X})=\frac{V(X)}{n}$，由切比雪夫不等式可知，成立

$$P(|\bar{X}-E(\bar{X})|<\varepsilon)\geqslant 1-\frac{V(\bar{X})}{\varepsilon^2}=1-\frac{V(X)}{n\varepsilon^2}$$

取 $\varepsilon=\sqrt{\frac{V(X)}{n\alpha}}$，则

$$P\left(\bar{X}-\sqrt{\frac{V(X)}{n\alpha}}<E(X)<\bar{X}+\sqrt{\frac{V(X)}{n\alpha}}\right)\geqslant 1-\alpha$$

所以，总体均数 $E(X)$ 的 $1-\alpha$ 置信区间为 $\left(\bar{X}-\sqrt{\frac{V(X)}{n\alpha}},\bar{X}+\sqrt{\frac{V(X)}{n\alpha}}\right)$。

若方差 $V(X)$ 未知，则可以用样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 代替 $V(X)$，从而得到总体均数 $E(X)$ 的 $1-\alpha$ 置信区间为 $\left(\bar{X}-\frac{S}{\sqrt{n\alpha}},\bar{X}+\frac{S}{\sqrt{n\alpha}}\right)$，其中 $S=\sqrt{S^2}$ 为样本标准差。

参　考　文　献

1　祝国强. 医药数理统计方法 . 第 2 版. 北京：高等教育出版社，2009，112～114.

2　周怀捂. 医药应用概率统计. 上海：百家出版社，1989，152～153.

Small Sample Interval Estimation of Non-normal Population

Hang Guoming，et al

(*School of Mathematical Sciences*，*Fudan University*，*Shanghai* 200433)

Abstract　When the population is non-normal and the sample is a small one, we can use such methods as exact probability calculation, Fisher normal approximation and Chebyshev inequality to determine the confidence interval of unknown parameters according to the difference conditions of the population.

Key words　small sample; confidence interval; population rate; sample rate