第四章：区间估计

苏勤亮

中山大学计算机学院

超算中心503M

suqliang@mail.sysu.edu.cn

主要内容

- 置信区间
- 正态总体参数的置信区间
- 非正态总体参数的置信区间

- 区间估计

假设$(X_1,\cdots,X_n)$是总体X的一个样本，区间估计是给出两个统计量$\underline{\theta}=\underline{\theta}(X_1,\cdots,X_n), \bar{\theta}=\bar{\theta}(X_1,\cdots,X_n)$，使区间$(\underline{\theta},\bar{\theta})$以一定的概率覆盖真实的参数值$\theta$

- 区间估计与点估计的区别
  - 点估计是根据观察到的样本给出未知参数θ的一个估计值$\hat{\theta}$，但点估计并不能提供任何关于估计值$\hat{\theta}$ 可靠性的信息
  - 区间估计则要由样本给出参数θ的一个估计范围，并指出该区间包含θ可能性的大小

置信度、置信系数

定义：设总体$X_1, X_2, \cdots, X_n$是取自总体$F_\theta(x)$的一个样本，假如$\hat{\theta}_L(\boldsymbol{X})$与$\hat{\theta}_U(\boldsymbol{X})$是在参数空间$\Theta$上取值的两个统计量，且$\hat{\theta}_L(\boldsymbol{X}) < \hat{\theta}_U(\boldsymbol{X})$，则称随机区间$[\hat{\theta}_L, \hat{\theta}_U]$为参数$\theta$的一个区间估计。该区间覆盖参数$\theta$的概率$P_\theta\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right)$称为置信度。该置信度在参数空间$\Theta$上的下确界$\inf\limits_{\theta\in\Theta} P_\theta\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right)$称为该区间估计的置信系数

1) $[\hat{\theta}_L, \hat{\theta}_U]$为参数$\theta$的一个区间估计
2) $P_\theta\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right)$称为该区间估计的置信度
3) $\inf\limits_{\theta\in\Theta} P_\theta\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right)$称为该区间估计的置信系数

- 显然，区间估计的评价准则
  - 置信度（置信系数）越大越好
  - 区间估计$[\hat{\theta}_L, \hat{\theta}_U]$的平均长度越短越好

例：设$x_1, x_2, \cdots, x_n$是来自正态总体$N(\mu, \sigma^2)$的一个样本。用样本均值$\bar{x}$和样本方差s^2可以给出正态均值μ的区间估计

$$\left[\bar{x} - \frac{ks}{\sqrt{n}}, \bar{x} + \frac{ks}{\sqrt{n}}\right]$$

可计算出概率

$$P\left(\bar{x} - \frac{ks}{\sqrt{n}} \le \mu \le \bar{x} + \frac{ks}{\sqrt{n}}\right) = P\left(\left|\frac{\sqrt{n}(\bar{x} - \mu)}{s}\right| \le k\right)$$

$$=?$$

随机变量$\frac{\sqrt{n}(\bar{x}-\mu)}{s}$服从什么分布？

由前面章节t分布的知识可知，

$$\frac{\sqrt{n}(\bar{x}-\mu)}{s} \sim t_{n-1}(z)$$

由于分布 $t_{n-1}(z)$不依赖于未知参数μ和σ，所以可以用t分布来计算置信度

$$P\left(\left|\frac{\sqrt{n}(\bar{x}-\mu)}{s}\right| \le k\right) = P(|t| \le k)$$

其中$t \sim t_{n-1}(z)$. 因此，

$$P\left(\bar{x}-\frac{s}{\sqrt{n}} \le \mu \le \bar{x}+\frac{s}{\sqrt{n}}\right) = P(|t| \le 1) = 0.654$$

$$P\left(\bar{x}-\frac{2s}{\sqrt{n}} \le \mu \le \bar{x}+\frac{2s}{\sqrt{n}}\right) = P(|t| \le 2) = 0.933$$

$$P\left(\bar{x}-\frac{3s}{\sqrt{n}} \le \mu \le \bar{x}+\frac{3s}{\sqrt{n}}\right) = P(|t| \le 3) = 0.990$$

容易看出，在样本数n固定的情况下，置信系数与估计区间的长度成正比

- 区间增长，置信系数会增大，但由于区间太长，估计的精度下降
- 区间缩短，估计精度增加，但置信系数会减小

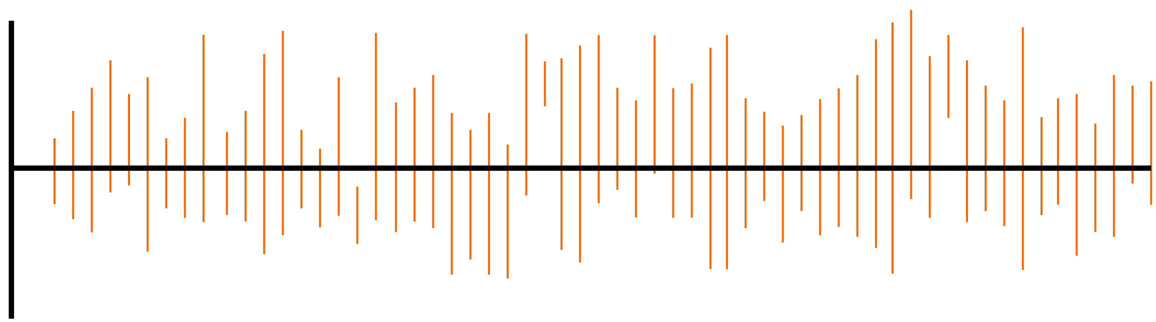
置信区间

定义：设θ是总体的一个参数，其参数空间为Θ，设$x_1, x_2, \cdots, x_n$是来自该总体的一个样本，对给定α $(0 < \alpha < 1)$，确定两个统计量$\hat{\theta}_L = \hat{\theta}_L(x_1, x_2, \cdots, x_n)$与$\hat{\theta}_U = \hat{\theta}_U(x_1, x_2, \cdots, x_n)$，若有

$$P\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right) \ge 1 - \alpha, \qquad \forall\, \theta \in \Theta$$

则称随机区间$[\hat{\theta}_L, \hat{\theta}_U]$是$\theta$的置信水平为$1-\alpha$的置信区间，或简称$[\hat{\theta}_L, \hat{\theta}_U]$是$\theta$的$1-\alpha$置信区间，$\hat{\theta}_L$与$\hat{\theta}_U$分别称为$1-\alpha$置信区间的置信下限与置信上限

- 置信水平与置信区间的含义：设法构造一个随机区间$[\hat{\theta}_L, \hat{\theta}_U]$，它能覆盖未知参数$\theta$的概率至少为$1-\alpha$
  - 区间是一个随机变量，会随着观察值的变化而改变
  - 若反复抽样多次, 每次的抽样值都会导致一个区间$(\hat{\theta}_L, \hat{\theta}_U)$



  - 这样的区间或者包含θ的真值, 或者不包含θ的真值

    当$\alpha = 0.01$，即置信水平为99%时, 它表示100个计算得到的区间$(\underline{\theta}, \bar{\theta})$中，约有99个包含真实的$\theta$值

定义：如果对于给定的α $(0 < \alpha < 1)$恒有

$$P\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right) = 1 - \alpha, \qquad \forall\, \theta \in \Theta$$

则称随机区间$\left[\hat{\theta}_L, \hat{\theta}_U\right]$是$\theta$的$1-\alpha$的同等置信区间

- 在一些问题中，我们可能只关心某未知参数的下限或上限
  - 例如，对于某种合金钢的强度来讲，人们总是希望其强度越大越好，这时强度的下限是一个很重要的指标
  - 对某些药物的毒性来讲，人们总希望其毒性越小越好，这时毒性的上限便成了一个重要的指标

定义：设θ是总体的某个未知参数，对给定的$\alpha\ (0<\alpha<1)$，由来自该总体的样本$x_1, x_2, \cdots x_n$确定的统计量$\hat{\theta}_L=\hat{\theta}_L(x_1, x_2, \cdots, x_n)$满足

$$P\left(\theta \geq \hat{\theta}_L\right) \geq 1-\alpha, \qquad \forall\, \theta \in \Theta$$

则称$\hat{\theta}_L$为θ的置信水平是$1-\alpha$的单侧置信下限，简称$1-\alpha$单侧置信下限。若等号对一切$\theta \in \Theta$成立，则称$\hat{\theta}_L$为θ的$1-\alpha$单侧同等置信下限

类似的，可以定义单侧置信上限、单侧同等置信上限

枢轴量法

- 构造未知参数θ的置信区间的一种常用方法是枢轴量法

1) 从θ的一个点估计$\hat{\theta}$出发，构造$\hat{\theta}$与θ的一个函数$G(\hat{\theta},\theta)$，使得G的分布已知且与θ无关

2) 选取两个适当的常数c与d，使得

$$P\left(c \leq G(\hat{\theta},\theta) \leq d\right) \geq 1-\alpha$$

3) 将不等式$c \leq G(\hat{\theta},\theta) \leq d$进行等价变形，得到形如$\hat{\theta}_L \leq \theta \leq \hat{\theta}_U$的不等式，那么这时

$$P\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\right) = P\left(c \leq G(\hat{\theta},\theta) \leq d\right) \geq 1-\alpha$$

这时，可以说$\left[\hat{\theta}_L,\hat{\theta}_U\right]$是$\theta$的$1-\alpha$置信区间

- 枢轴量法的难点在于如何构造一个函数（统计量）$G(\hat{\theta},\theta)$，使其概率分布与参数θ无关

- 构造好枢轴量$G(\hat{\theta},\theta)$后，接下来需要确定常数c和d的值
  - 当枢轴量G的分布对称时（如标准正态分布），可取d，使得

$$P(-d \le G \le d) = P(|G| \le d) = 1-\alpha$$

  - 当枢轴量G的分布非对称时（如χ^2分布），可这样取c和d，使得

$$P(G \le c) = \alpha/2 \qquad P(G \ge d) = \alpha/2$$

这样获得的置信区间称为等尾置信区间

主要内容

- 置信区间
- 正态总体参数的置信区间
- 非正态总体参数的置信区间

（一）单个正态总体$N(\mu,\sigma^2)$的情形

均值μ的置信区间——σ^2已知

问题：设$X_1, X_2, \cdots, X_n$来自$N(\mu, \sigma^2)$，$\bar{X}$为样本均值，试确定μ可能取值范围（假设σ^2已知）

- 由给定条件可知，$\bar{X}$服从分布

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \Rightarrow \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

因此，可以得到如下不等式

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

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

因此，μ的置信水平为$1-\alpha$的置信区间为$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$，简写为$\left(\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$

- 从以上分析过程，我们归纳出计算置信区间的一般步骤如下

1) 寻找参数θ的一个良好的点估计$T(X_1, X_2, \ldots, X_n)$

2) 寻找一个参数θ和估计量T的函数$G(T,\theta)$, 使其分布完全已知且不包含任何未知参数

3) 对于给定的置信水平$1-\alpha$，根据$G(T,\theta)$的分布，确定常数c, d，使得

$$P(c < G(T,\theta) < d) = 1-\alpha$$

4) 对$c < G(T,\theta) < d$做等价变形, 得到如下形式

$$\underline{\theta} < \theta < \overline{\theta}$$

$(\underline{\theta}, \ \overline{\theta})$就是$\theta$的置信水平为$1-\alpha$的置信区间

均值μ的置信区间——σ^2未知

- 由于σ^2未知，不能基于概率分布$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$来推导置信区间
- 由第二章可知，当σ^2用样本方差$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$代替时，有

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

- 由该分布可以知道如下不等式成立

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

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

- 因此， 在σ^2未知情况下，μ置信水平为$1-\alpha$的置信区间为

$$\left(\bar{X} - \frac{S}{\sqrt{n}}t_{\alpha/2}(n-1), \bar{X} + \frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)\right)$$

例：设某种植物的高度X(cm)服从正态分布$N(\mu,\sigma^2)$, 随机选取36棵, 其平均高度为15cm. 就以下两种情形求μ的95％双侧置信区间

1) $\sigma^2 = 16$;　　2) σ^2未知，$S^2 = 16$

解：　1) 由题目信息可知

$$n = 36, \bar{X} = 15, \sigma = 4,\ \ z_{\alpha/2} = z_{0.025} = 1.96$$

在σ^2已知的条件下，　$1-\alpha$置信区间为

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2},\ \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$$

将上述数值带入，可以得到μ的95%置信区间为

$$(13.693, 16.307)$$

2) 由题目信息可知

$$n = 36, \bar{X} = 15, S^2 = 16$$

在σ^2未知的条件下，$1-\alpha$置信区间为

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1), \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right)$$

通过查表可知$t_{0.025}(35) = 2.0301$

将上述数值带入，可以得到μ的95%置信区间为

$$(13.647, 16.353)$$

- 比较1) 和2) 两种情形下μ的置信区间

  - σ^2已知的置信区间：(13.693,16.307)

  - σ^2未知的置信区间：(13.647,16.353)

第一种情形区间更短（精度更高），但第二种情形往往更实用，因为多数时候，σ^2是未知的

样本数量确定

- 一般而言，样本数量n越大，未知参数估计精度越高。但大样本的获得通常费用较高、时间较长，不容易实现

在满足一定的精度要求下，最少需要多少样本？

- 当标准差σ已知时，均值μ的$1-\alpha$置信区间为$\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$. 若要该置信区间长度不超过$2d$，则有

$$2z_{\alpha/2}\sigma/\sqrt{n} \le 2d$$

可以得到

$$n \ge \left(\frac{z_{\alpha/2}\sigma}{d}\right)^2$$

例：设一个物体的重量 μ 未知，为估计其重量，可以用天平去称，假设称重服从正态分布。如果已知称量的误差的标准差为0.1克，为使 μ 的95%的置信区间的长度不超过0.2，那么至少应称多少次？

解：已知$\sigma = 0.1$，当要求置信区间的长度不超过0.2时，由公式

$n \geq \left(\frac{z_{\alpha/2}\sigma}{d}\right)^2$可直接得出

$$n \geq \left(1.96\times\frac{0.01}{0.1}\right)^2 = 3.84$$

故，至少需要称$n = 4$次

当要求置信区间的长度不超过0.1时，可计算得到

$$n \geq \left(1.96\times\frac{0.01}{0.05}\right)^2 = 15.37$$

故，至少需要称$n = 16$次

- 当标准差σ未知时，均值μ的$1-\alpha$置信区间为$\bar{x} \pm t_{\alpha/2}(n-1)s/\sqrt{n}$. 若要该置信区间长度不超过$2d$，则有

$$2t_{\alpha/2}(n-1)s/\sqrt{n} \leq 2d$$

假设s有通过n_0个样本估计得到，记为s_0，那么用s_0代替s，可获得

$$n \geq \left(\frac{t_{\alpha/2}(n_0-1)s_0}{d}\right)^2$$

例：为了对垫圈总体的平均厚度做出估计，我们所取的风险是允许在100次估计中有5次误差超过0.02cm，近期从另一批产品中抽得一个容量为10的样本，得到标准差的估计为$s_0 = 0.0359$，问现在该取多少样品为宜？

解：这里的“风险”就是指样本均值落在置信区间外的概率α，如今$\alpha = 0.05$。“估计误差超过0.02”表明$d = 0.02$。用10个样本估计得到的样本标准差$s_0 = 0.0359$，$n_0 = 10$，因此$t_{\alpha/2}(9) = 2.262$. 将这些值代入 $n \geq \left(\frac{t_{\alpha/2}(n_0-1)s_0}{d}\right)^2$可得

$$n \geq \left(\frac{0.0359 \times 2.262}{0.02}\right)^2 = 16.49$$

因此，至少需要取17个样品

方差σ^2的置信区间

- 由第二章学习可知，

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

- 根据上述分布，可知如下不等式成立

$$P\left\{\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right\} = 1-\alpha$$

即：$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right\} = 1-\alpha$

- 因此，σ^2的$1-\alpha$置信区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$$

例: 一个园艺科学家正在培养一个新品种的苹果, 这种苹果除了口感好和颜色鲜艳以外, 另一个重要特征是单个重量差异不大。为了评估新苹果, 她随机挑选了25个测试重量(单位：克), 其样本方差为$S^2=4.25$. 试求σ^2的置信度为95％和的99％的置信区间（假设苹果重量服从正态分布）

解：置信度为95%时，查表可得

$$\chi^2_{0.975}(24)=12.4, \qquad \chi^2_{0.025}(24)=39.4$$

带入置信区间$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$，计算得$\sigma^2$的95%置信区间为$(2.59, 8.23)$

置信度为99%时, $\chi^2_{0.995}(24)=9.89$，$\chi^2_{0.005}(24)=45.6$，可计算得到其置信区间为$(2.24,10.31)$

（二）两个正态总体$N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$的情形

第二章结论回顾

定理：设样本$(X_1,\cdots,X_{n_1})$和$(Y_1,\cdots,Y_{n_2})$分别来自总体$N(\mu_1,\sigma_1^2)$和$N(\mu_2,\sigma_2^2)$并且它们相互独立，其样本方差分别为S_1^2和S_2^2, 则：

1) $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1,n_2-1)$

2) $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1)$

3) 当${\sigma_1}^2={\sigma_2}^2=\sigma^2$时，$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2)$

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

$\mu_1 - \mu_2$的置信区间

问题：$X_1, \cdots, X_{n_1}$来自$N(\mu_1, \sigma_1^2)$，$Y_1, \cdots, Y_{n_2}$来自$N(\mu_2, \sigma_2^2)$，$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$，$\bar{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$，$S_1^2$和$S_2^2$分别为第一、二个总体的样本方差，试确定$\mu_1 - \mu_2$的置信区间

- σ_1^2，σ_2^2 已知的情况

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

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1)$$

  可以求得$\mu_1 - \mu_2$的置信区间为

$$\left((\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right)$$

- $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 但$\sigma^2$未知的情况

由第二章知识可知

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2)$$

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

可计算得到置信区间为

$$\left((\bar{X}-\bar{Y}) \pm t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{1/n_1+1/n_2}\right)$$

$\frac{\sigma_1^2}{\sigma_2^2}$的置信区间

- 由 $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$可知

$$P\left\{F_{1-\alpha/2}(n_1-1, n_2-1) < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{\alpha/2}(n_1-1, n_2-1)\right\} = 1-\alpha$$

即：$P\left\{\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right\} = 1-\alpha$

可以计算得到置信区间为：

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1, n_2-1)}, \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1, n_2-1)}\right)$$

例：两台机床生产同一个型号的滚珠，从甲机床生产的滚珠中抽取8个，从乙机床生产的滚珠中抽取9个，测得这些滚珠的直径(毫米)如下：

甲机床 15.0 14.8 15.2 15.4 14.9 15.1 15.2 14.8

乙机床 15.2 15.0 14.8 15.1 14.6 14.8 15.1 14.5 15.0

设两机床生产的滚珠直径分别为X, Y, 且$X \sim N(\mu_1, \sigma_1^2)$， $Y \sim N(\mu_2, \sigma_2^2)$

1) $\sigma_1 = 0.18, \sigma_2 = 0.24$, 求$\mu_1 - \mu_2$的置信度为0.90的置信区间

2) 若$\sigma_1 = \sigma_2 = \sigma$未知，求$\mu_1 - \mu_2$的置信度为0.90的置信区间

3) 若μ_1, μ_2未知，求$\frac{\sigma_1^2}{\sigma_2^2}$的置信度为0.90的置信区间

解：$n_1 = 8, \bar{x} = 15.05$， $s_1^2 = 0.0457, n_2 = 9$， $\bar{y} = 14.9$， $s_2^2 = 0.0575$

1） 当$\sigma_1 = 0.18, \sigma_2 = 0.24$时, $\mu_1 - \mu_2$的置信度为0. 90的置信区间为

$$\left(\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right)$$

查表得$z_{0.05} = 1.645$, 从而所求区间为$(-0.018, 0.318)$

2） 当$\sigma_1 = \sigma_2 = \sigma$未知时，$\mu_1 - \mu_2$的置信度为0. 90的置信区间为

$$\left(\bar{X} - \bar{Y} \pm t_{\alpha/2}(n_1 + n_2 - 2)S_W\sqrt{1/n_1 + 1/n_2}\right)$$

代入$t_{0.05}(15) = 1.7531, S_W = 0.228, \sqrt{1/n_1 + 1/n_2} = 0.486$，可得置信区间为$(-0.044, 0.344)$

3) 当μ_1，μ_2未知时，$\frac{\sigma_1^2}{\sigma_2^2}$的置信度为0. 90的置信区间为

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)},\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right)$$

将$F_{0.05}(7,8)=3.50, F_{0.95}(7,8)=\frac{1}{F_{0.05}(8,7)}=\frac{1}{3.73}=0.268$代入，计算得到置信度为0. 9的置信区间为(0. 227, 2. 965)

主要内容

- 置信区间
- 正态总体参数的置信区间
- 非正态总体参数的置信区间

小样本方法——指数分布参数的置信区间

问题：设$x_1, x_2, \cdots, x_n$是从指数分布$\exp(1/\theta)$中抽取的一个样本，其密度函数为

$$p_\theta(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \qquad x \geq 0$$

其中，$\theta > 0$为总体均值，即：$E[x] = \theta$，现要求θ的$1-\alpha$置信区间$(0 < \alpha < 1)$

- $T_n = x_1 + x_2 + \cdots + x_n$是$\theta$的充分统计量，$\frac{T_n}{n}$是$\theta$的无偏估计。由于指数分布是伽玛分布的特例，即：$x_i \sim Ga\left(1, \frac{1}{\theta}\right)$. 利用伽玛分布的性质可知

$$T_n \sim Ga(n, 1/\theta)$$

$$Ga(x; \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$\frac{2T_n}{\theta} \sim Ga\left(n, \frac{1}{2}\right) = \chi^2(2n)$$

可见，$2T_n/\theta$的分布不依赖于θ，可取其为枢轴量

- 对给定的置信水平$1-\alpha$，利用χ^2分布的$\frac{\alpha}{2}$和$1-\frac{\alpha}{2}$分位数可得

$$P\left(\chi^2_{\alpha/2}(2n) \le \frac{2T_n}{\theta} \le \chi^2_{1-\alpha/2}(2n)\right) = 1-\alpha$$

- 利用不等式等价变形可得

$$P\left(2T_n/\chi^2_{1-\alpha/2}(2n) \le \theta \le 2T_n/\chi^2_{\alpha/2}(2n)\right) = 1-\alpha$$

这样就获得θ的$1-\alpha$同等置信区间

$$\left[2T_n/\chi^2_{1-\alpha/2}(2n), 2T_n/\chi^2_{\alpha/2}(2n)\right]$$

补充：Gamma分布性质

如果$X \sim \Gamma(a_1, b)$，$Y \sim \Gamma(a_2, b)$，那么当X和Y相互独立时，$X + Y \sim \Gamma(a_1 + a_2, b)$

证：令$Z = X + Y$，那么由$f(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z - x) dx$可知

$$
\begin{aligned}
f(z) &= \int_0^z \frac{x^{a_1-1} e^{-x}}{\Gamma(a_1)} \frac{(z-x)^{a_2-1} e^{-(z-x)}}{\Gamma(a_2)} \mathrm{d}x \\
&= e^{-z} \int_0^z \frac{x^{a_1-1} (z-x)^{a_2-1}}{\Gamma(a_1)\Gamma(a_2)} \mathrm{d}x \\
&= e^{-z} z^{a_1+a_2-1} \int_0^1 \frac{t^{a_1-1} (1-t)^{a_2-1}}{\Gamma(a_1)\Gamma(a_2)} \mathrm{d}t \\
&= \frac{e^{-z} z^{a_1+a_2-1}}{\Gamma(a_1 + a_2)}
\end{aligned}
$$

Beta分布

例如，某产品的寿命服从指数分布$\exp\left(\frac{1}{\theta}\right)$，如今从中随机抽取9个样品进行寿命试验，获得如下9个寿命数据（单位：小时）：152，457，505，531，607，645，707，822，903。可计算得到

$$T_n = 5329$$

若取$\alpha = 0.1$，可从χ^2分布α分位数查表得

$$\chi^2_{0.05}(18) = 9.39 \qquad \chi^2_{0.95}(18) = 28.87$$

于是带入$\left[2T_n/\chi^2_{1-\alpha/2}(2n), 2T_n/\chi^2_{\alpha/2}(2n)\right]$可得平均寿命$\theta$的0. 9同等置信区间为$[369.17, 1135.04]$

小样本方法——均匀分布参数的置信区间

问题：设$\boldsymbol{X} = (X_1, \cdots, X_n)$ 为自均匀分布总体$U(0,\theta)$中抽取的简单随机样本，求θ的置信系数为$1-\alpha$的置信区间

- 设$T(\boldsymbol{X}) = X_{(n)} = max\{X_1, \cdots, X_n\}$，由前面学习已知$T(X)$为充分统计量，$\frac{n+1}{n}T$是$\theta$的无偏估计（也是UMVUE）
- 取枢轴变量为$\frac{T}{\theta}$. 由第二章学习已经知道T的密度函数为

$$f(t) = n\frac{t^{n-1}}{\theta^n} I_{[0<t<\theta]}$$

因而，$\widetilde{\mathrm{T}} = T/\theta$的密度函数为

$$f(t) = nt^{n-1} I_{[0<t<1]}$$

- 由于T/θ的概率密度与参数无关，可取其为枢轴量。现在确定c_1和c_2（$0 < c_1 < c_2 \le 1$），使得

$$P_\theta\left(c_1 < \frac{T}{\theta} < c_2\right) = \int_{c_1}^{c_2} nt^{n-1}dt = c_2^n - c_1^n = 1 - \alpha$$

- 由于$f(t)$是严格增函数，故$c_2 = 1$的置信区间精度最高。这时，可计算得到

$$c_1 = \sqrt[n]{\alpha}$$

- 因此，θ的置信系数为$1 - \alpha$的置信区间可从$\sqrt[n]{\alpha} < \frac{T}{\theta} < 1$变换得到，即：

$$T < \theta < T/\sqrt[n]{\alpha}$$

大样本方法——基于中心极限定理的置信区间

问题：设$\boldsymbol{X}=(X_1,\cdots,X_n)$为来自伯努利分布$b(1,p)$中抽取的简单随机样本，求$p$的置信系数为$1-\alpha$的置信区间

- 令$S_n=\sum_{i=1}^n X_i$，可知$S_n \sim b(n,p)$． S_n的均值和方差分别为np和$np(1-p)$． 当n很大时，由中心极限定理可知

$$\frac{S_n-np}{\sqrt{np(1-p)}}=\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \xrightarrow{\mathcal{L}} N(0,1)$$

其中$\hat{p}=S_n/n$

- 因此，当n充分大时，$T=\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$的渐进分布是$N(0,1)$，与未知参数$p$无关。因此，可以取$T$为枢轴变量

- 因此，当n充分大时，有

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

- 由不等式 $\left|\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right| < z_{\alpha/2}$可知（记$\gamma = z_{\alpha/2}$），

$$(\hat{p}-p)^2 < \frac{\gamma^2 p(1-p)}{n}$$

- 通过展开二次项并解不等式，可知$c_1 < p < c_2$，其中

$$c_1, c_2 = \frac{n}{n+\gamma^2}\left[\hat{p} + \frac{\gamma^2}{2n} \pm \gamma \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\gamma^2}{4n^2}}\right]$$

- 当$n \to +\infty$时，由大数定律知$\hat{p} \to p$，因而 $\frac{\sqrt{p(1-p)}}{\sqrt{\hat{p}(1-\hat{p})}} \to 1$。因此，

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \frac{\sqrt{p(1-p)}}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow{\mathcal{L}} N(0,1)$$

- 可以取枢轴量为$T = \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$，其极限分布与参数$p$无关。因此，可以知道

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

- 因此，可以得到p的置信系数为$1-\alpha$的置信区间为

$$\left[\hat{p} - z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}\right]$$

问题：设$\boldsymbol{X}=(X_1,\cdots,X_n)$为抽自Poisson总体$P(\lambda)$的简单随机样本，求$\lambda$的置信系数为$1-\alpha$的置信区间

- 令$S_n=\sum_{i=1}^{n}X_i$，当$n\to+\infty$时，由中心极限定理可知

$$\frac{S_n-n\lambda}{\sqrt{n\lambda}}=\frac{\sqrt{n}(\hat{\lambda}-\lambda)}{\sqrt{\lambda}}\xrightarrow{\mathcal{L}}N(0,1)$$

其中$\hat{\lambda}=S_n/n$

- 将随机变量 $T=\frac{\sqrt{n}(\hat{\lambda}-\lambda)}{\sqrt{\lambda}}$作为枢轴变量，其极限分布与未知参数无关

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

- 由不等式 $\left|\frac{\sqrt{n}(\hat{\lambda}-\lambda)}{\sqrt{\lambda}}\right| < z_{\alpha/2}$ 可知

$$(\hat{\lambda}-\lambda)^2 = \frac{\lambda z_{\alpha/2}^2}{n}$$

- 通过展开二次项并解不等式，可知$d_1 < \lambda < d_2$，其中

$$d_1, d_2 = \hat{\lambda} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{z_{\alpha/2}^2}{4n^2} + \frac{\hat{\lambda}}{n}}$$

- 由于泊松分布的期望是λ，因此，当$n \to +\infty$时，由大数定律知 $\hat{\lambda} \to \lambda$，因而 $\sqrt{\lambda}/\sqrt{\hat{\lambda}} \to 1$。因此，

$$\frac{\sqrt{n}(\hat{\lambda}-\lambda)}{\sqrt{\hat{\lambda}}} = \frac{\sqrt{n}(\hat{\lambda}-\lambda)}{\sqrt{\lambda}} \cdot \frac{\sqrt{\lambda}}{\sqrt{\hat{\lambda}}} \xrightarrow{\mathcal{L}} N(0,1)$$

- 令 $T = \frac{\sqrt{n}(\hat{\lambda}-\lambda)}{\sqrt{\hat{\lambda}}}$为枢轴变量，由$\frac{\sqrt{n}(\hat{\lambda}-\lambda)}{\sqrt{\hat{\lambda}}} \xrightarrow{\mathcal{L}} N(0,1)$可知

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

- 因此，可以得到p的置信系数为$1-\alpha$的置信区间为

$$\left[\hat{\lambda} - z_{\alpha/2}\sqrt{\hat{\lambda}/n}, \hat{\lambda} + z_{\alpha/2}\sqrt{\hat{\lambda}/n}\right]$$

大样本方法——基于最大似然估计的置信区间

- 由第三章内容可知，最大似然估计$\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \cdots, x_n)$在$n \to +\infty$时服从渐进正态分布

$$N\left(\theta, \sigma_n^2(\theta)\right)$$

其中$\sigma_n^2(\theta) = \frac{1}{nI(\theta)}$，$I(\theta) = E\left[\left(\frac{\partial \log f(x,\theta)}{\partial \theta}\right)^2\right]$

- 因此，可以知道

$$\frac{\hat{\theta}_n - \theta}{\sigma_n(\theta)} \xrightarrow{\mathcal{L}} N(0,1)$$

- 因为最大似然估计是相合估计，即在$n \to +\infty$时，$\hat{\theta}_n \to \theta$。

因此，可以认为 $\frac{\hat{\theta}_n - \theta}{\sigma_n(\hat{\theta}_n)} \xrightarrow{\mathcal{L}} N(0,1)$

- 令$T = \frac{\hat{\theta}_n - \theta}{\sigma_n(\hat{\theta}_n)}$为枢轴变量，它的极限分布与$\theta$无关，且

$$P\left(\left|\frac{\hat{\theta}_n - \theta}{\sigma_n(\hat{\theta}_n)}\right| < z_{\alpha/2}\right) \approx 1 - \alpha$$

- 解括号中的不等式可得θ的置信系数为$1 - \alpha$的置信区间为

$$\left[\hat{\theta}_n - z_{\alpha/2}\sigma_n(\hat{\theta}_n), \hat{\theta}_n + z_{\alpha/2}\sigma_n(\hat{\theta}_n)\right]$$

例：设$x_1, x_2, \cdots, x_n$是来自指数分布$p(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} (x > 0)$的一个样本，求其大样本下的$1-\alpha$置信区间

解：该总体的Fisher信息量为

$$I(\theta) = \theta^{-2}$$

参数θ的最大似然估计$\hat{\theta}_n = \bar{x}$，它的渐进分布为

$$\hat{\theta}_n \sim N\left(\theta, \frac{\theta^2}{n}\right)$$

当$n \to +\infty$时，$\hat{\theta}_n \to \theta$，$I(\hat{\theta}_n) \to I(\theta)$。因此，可近似认为

$$\hat{\theta}_n \sim N\left(\theta, \frac{\hat{\theta}_n^2}{n}\right)$$

令$T=\frac{\hat{\theta}_n-\theta}{\hat{\theta}_n/\sqrt{n}}$为枢轴变量，可知

$$P\left(\left|\frac{\hat{\theta}_n-\theta}{\hat{\theta}_n/\sqrt{n}}\right|<z_{\alpha/2}\right)\approx 1-\alpha$$

即：θ的置信系数为$1-\alpha$的置信区间为

$$\left[\bar{x}-\bar{x}z_{\alpha/2}/\sqrt{n},\bar{x}+\bar{x}z_{\alpha/2}/\sqrt{n}\right]$$

- 与前面小样本条件下获得的精确置信区间比较

$$\left[2T_n/\chi^2_{1-\alpha/2}(2n),2T_n/\chi^2_{\alpha/2}(2n)\right]$$

其中$T_n=x_1+\cdots+x_n=n\bar{x}$

- 显然，这两种方法获得的置信区间不一样

- 在大样本条件下，这两种方法获得的区间会比较接近

例如，某产品的寿命X服从指数分布$\exp(1/\theta)$，θ为其平均寿命。若从中抽取60个样品作寿命试验，试验到全部失效为止，所得60个寿命数据之和$T_n = 45079$（小时）。故其平均寿命的估计值为$\hat{\theta}_n = T_n/n = 751.32$小时，现求其0.9的置信区间

1）按大样本方法，θ的0.9置信区间为

$$\bar{x}\left(1 \pm \frac{z_{0.05}}{\sqrt{n}}\right) = (591.74, 910.90)$$

2）按小样本方法，θ的0.9置信区间为

$$\left[2T_n/\chi^2_{1-\alpha/2}(2n), 2T_n/\chi^2_{\alpha/2}(2n)\right] = (615.12, 942.09)$$

可以看出，两者较为接近

作业

- 3.1节：1，2
- 3.2节：1，2，8，10，11，12
- 3.3节：1，2，6