

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



数理统计

目录

第一章	统计量与抽样分布
第二章	点估计
第三章	区间估计
第四章	假设检验
第五章	分布的检验
第六章	Bayes统计与统计判决理论
第七章	非参数统计
第八章	抽样调查

目 录

3.1 置信区间

3.2 正态总体参数的置信区间

3.3 大样本置信区间

区间估计

定义 3.1.1 (区间估计(interval estimation)) 设 $X_1, \dots, X_n$ 为来自参数分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$ 的样本, θ 为一维未知参数. 如果 $\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})$ 为两个统计量, 且 $\hat{\theta}_L(\mathbf{X}) \leq \hat{\theta}_U(\mathbf{X})$, 则称随机区间 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 为 θ 的一个区间估计.

置信系数

定义 3.1.2 (置信系数) 设随机区间 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 为 θ 的一个区间估计, 则称

$$\inf_{\theta \in \Theta} P_{\theta} \{ \hat{\theta}_L(\mathbf{X}) \leq \theta \leq \hat{\theta}_U(\mathbf{X}) \} \quad (3.1.1)$$

为该区间估计的置信系数.

注 3.1.1 如果置信度不依赖于未知参数, 则置信系数就是置信度.

注 3.1.2 区间估计有时要用开区间或半开半闭区间, 但从置信度的角度看, 这几种区间估计没有本质的区别.

注 3.1.3 在计算某区间估计的置信度时, 我们应知道 $\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})$ 的联合分布. 如不知道其联合分布, 则很难求得其置信系数. 这就是构造置信区间的技巧之所在.

置信系数

例 3.1.1 设 $X_1, \dots, X_n$ 为来自 $N(\mu, \sigma^2)$ 的 iid 样本, 其中 μ, σ^2 未知. 考虑 μ 的区间估计.

解 前面我们讲过的点估计中, 多数是用样本均值估计总体均值. 对于本问题, 我们知道, 样本均值是 μ 的一个很好的点估计 (矩估计、MLE, UMVUE 等). 关于 μ 的区间估计, 一个自然的想法是基于样本均值, 即我们可以用

$$(\bar{X} - kS_n/\sqrt{n}, \bar{X} + kS_n/\sqrt{n})$$

来作为 μ 的区间估计, 其中 k 是一常数, S_n 为样本标准差. 则它的置信度为

$$\alpha_k = P\{\bar{X} - kS_n/\sqrt{n} \leq \mu \leq \bar{X} + kS_n/\sqrt{n}\} = P\left\{\left|\frac{\sqrt{n}(\bar{X} - \mu)}{S_n}\right| \leq k\right\} = P\{|T_n| \leq k\},$$

其中

$$T_n = \frac{\sqrt{n}(\bar{X} - \mu)}{S_n} \sim t(n-1).$$

□

评价区间估计的好坏

一个未知参数的区间估计有很多，如何评价其好坏呢？常用的标准有如下两个：

（1）置信度（或置信系数）越大越好，因为人们对给出的区间估计可覆盖未知参数的概率越大越放心；

（2）随机区间的平均长度越短越好，因为平均长度越短表示估计的精度越高。

置信区间

定义 3.1.3 (置信区间(Confident Interval)) 设 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 是参数 θ 的一个区间估计, 如果对给定的 $\alpha \in (0, 1)$, 有

$$P_{\theta}\{\hat{\theta}_L(\mathbf{X}) \leq \theta \leq \hat{\theta}_U(\mathbf{X})\} \geq 1 - \alpha, \forall \theta \in \Theta, \quad (3.1.2)$$

则称 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 为 θ 的置信水平为 $1 - \alpha$ 的置信区间, $\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})$ 分别称为置信水平为 $1 - \alpha$ 的置信区间的置信下限与置信上限或双侧置信上(下)限.

注 3.1.4 虽然上面定义的置信区间的置信度不小于 $1 - \alpha$, 但在实际中我们均是求取满足

$$P_{\theta}\{\hat{\theta}_L(\mathbf{X}) \leq \theta \leq \hat{\theta}_U(\mathbf{X})\} = 1 - \alpha, \forall \theta \in \Theta \quad (3.1.3)$$

的置信区间, 有的书上称这种置信区间为同等置信区间, 但在不致于引起混淆的时候, 我们也简称它为置信区间.

置信区间

定义 3.1.4 (单侧置信限) 设 $\hat{\theta}_L(\mathbf{X})$ 与 $\hat{\theta}_U(\mathbf{X})$ 是两个统计量, 如对于给定的 $\alpha \in (0, 1)$, 有

$$P_{\theta}\{\hat{\theta}_L(\mathbf{X}) \leq \theta\} \geq 1 - \alpha, \forall \theta \in \Theta, \quad (3.1.3)$$

$$P_{\theta}\{\hat{\theta}_U(\mathbf{X}) \geq \theta\} \geq 1 - \alpha, \forall \theta \in \Theta, \quad (3.1.4)$$

则分别称 $\hat{\theta}_L(\mathbf{X})$ 与 $\hat{\theta}_U(\mathbf{X})$ 为 θ 的置信水平为 $1 - \alpha$ 的单侧置信下限与上限.

单侧置信限

例 7.4.3 从某批灯泡中随机取 5 只作寿命试验, 其寿命 (以小时计) 如下: 1050, 1100, 1120, 1250, 1280. 设寿命服从正态分布, 试求其置信度为 95% 的置信下限.

解 在本例中, 总体 $N(a, \sigma)$ 的方差未知, 故应用 t - 分布. 我们知道

$$\frac{(\bar{\xi} - a)\sqrt{n-1}}{S} \sim t(n-1),$$

这里我们要求

$$P\left(\frac{(\bar{\xi} - a)\sqrt{n-1}}{S} \leq c\right) = 1 - \alpha.$$

按以前定义知 $c = t_{2\alpha}(n-1)$, 即

$$P\left(\frac{(\bar{\xi} - a)\sqrt{n-1}}{S} \leq t_{2\alpha}(n-1)\right) = 1 - \alpha.$$

于是得 a 的单侧置信度为 $1-\alpha$ 的置信区间为 $\left[\bar{\xi} - t_{2\alpha}(n-1) \frac{S}{\sqrt{n-1}}, +\infty\right)$.

单侧置信限

对于给定的数据, 具体计算如下:

$$\bar{x} = \frac{1}{5}(1050 + 1100 + 1120 + 1250 + 1280) = 1160,$$

$$S^2 = \frac{1}{5} \left((1050 - 1160)^2 + (1100 - 1160)^2 + (1120 - 1160)^2 \right. \\ \left. + (1250 - 1160)^2 + (1280 - 1160)^2 \right) = \frac{39800}{5} = 7960.$$

又 $1 - \alpha = 0.95$, $\alpha = 0.05$, $t_{0.1}(5 - 1) = 2.1318$, 故寿命均值的置信度为 0.95 的单侧置信区间为 $\left[1160 - 2.1318 \frac{\sqrt{7960}}{2}, +\infty \right) = [1065, +\infty)$.
1065 就是所求的置信下限.

单侧置信限与双侧置信限的关系

至于单侧置信限与双侧置信限的关系如下：

定理 3.1.1 设 $\hat{\theta}_L(\mathbf{X})$ 与 $\hat{\theta}_U(\mathbf{X})$ 分别是参数 θ 的置信水平为 $1 - \alpha_1$ 和 $1 - \alpha_2$ 的单侧(同等)置信下限与上限, 且对于样本 $X_1, \dots, X_n$, 均有 $\hat{\theta}_L(\mathbf{X}) \leq \hat{\theta}_U(\mathbf{X})$. 则 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 是 θ 的置信水平为 $1 - (\alpha_1 + \alpha_2)$ 的(同等)置信区间.

证明 因为 $\{\hat{\theta}_L(\mathbf{X}) \leq \theta \leq \hat{\theta}_U(\mathbf{X})\}$, $\{\theta \leq \hat{\theta}_L(\mathbf{X})\}$ 与 $\{\theta \geq \hat{\theta}_U(\mathbf{X})\}$ 是三个互不相容且其并为全事件的三个事件, 又因为

$$P_{\theta}\{\theta < \hat{\theta}_L(\mathbf{X})\} = 1 - P_{\theta}\{\theta \geq \hat{\theta}_L(\mathbf{X})\} = 1 - \alpha_1,$$

$$P_{\theta}\{\hat{\theta}_U(\mathbf{X}) < \theta\} = 1 - P_{\theta}\{\theta \leq \hat{\theta}_U(\mathbf{X})\} = 1 - \alpha_2,$$

所以

$$P_{\theta}\{\hat{\theta}_L(\mathbf{X}) \leq \theta \leq \hat{\theta}_U(\mathbf{X})\} = 1 - \alpha_1 - \alpha_2.$$

□

置信域

定义 3.1.5 (置信域(Confident Region)) 设 $X_1, \dots, X_n$ 是来自参数分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta \subset R^k\}$ 的样本, $\theta = (\theta_1, \dots, \theta_k)'$. 如果统计量 $S(\mathbf{X})$ 满足

- 对任一样本观测值 $\mathbf{x}$, $S(\mathbf{x})$ 是 Θ 的一个子集,
- 对给定的 $\alpha \in (0, 1)$, $P_\theta\{\theta \in S(\mathbf{X})\} \geq 1 - \alpha, \forall \theta \in \Theta$,

则称 $S(\mathbf{X})$ 是 θ 的置信水平为 $1 - \alpha$ 的置信域, 而概率 $P_\theta\{\theta \in S(\mathbf{X})\}$ 在 Θ 上的下确界就称为置信系数.

枢轴量法

为了引入用枢轴量法求取参数的置信区间，我们先看一个简单的例子.

例 3.2.1 设 $X_1, \dots, X_n$ 为来自正态总体 $N(\mu, \sigma^2)$ 的iid样本, 其中 σ^2 已知, μ 未知. 求 μ 的 $1-\alpha$ 的置信区间.

解 由上一章内容知, $\bar{X}$ 是 μ 的一个很好的点估计, 且知道

$$\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1). \quad (3.2.3)$$

对于本例, 由(3.2.3)式可知

$$P \left\{ -u_{\alpha/2} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq u_{\alpha/2} \right\} = \Phi(u_{\alpha/2}) - \Phi(-u_{\alpha/2}) = 1 - \alpha,$$

即

$$P \left\{ \bar{X} - \frac{\sigma}{\sqrt{n}}u_{\alpha/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}}u_{\alpha/2} \right\} = 1 - \alpha,$$

这说明

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}}u_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}}u_{\alpha/2} \right]$$

即是 μ 的置信水平为 $1 - \alpha$ 的置信区间.

□

枢轴量法

从上个例子可以看出，我们在求取某参数的置信区间时，一般是先给一个好的点估计，之后再通过这个点估计的分布求取感兴趣参数的置信区间。这就是枢轴量的基本想法。下面，我们把用枢轴量法求取置信区间的步骤总结如下：

- ① 找一个与待估参数 $g(\theta)$ 无关的统计量 T 。一般是它的一个很好的点估计(如例3.2.1中的 $\bar{X}$)。
- ② 设法找出 T 和 $g(\theta)$ 的某函数 $S(T, g(\theta))$ ，使得 $S(T, g(\theta))$ 的分布 $F(x)$ 与 θ 无关(如在例3.2.1中的 $\sqrt{n}(\bar{X} - \mu)/\sigma$)， S 就称为枢轴量。
- ③ 适当地选取两个常数 c, d ，使对给定的 $\alpha \in (0, 1)$ ，有

$$P_{\theta}\{c \leq S(T, g(\theta)) \leq d\} = 1 - \alpha, \quad (3.2.4)$$

即 $F(d) - F(c) = 1 - \alpha$ (一般取 $d = F_{\alpha/2}$, $c = F_{1-\alpha/2}$)。

- ④ 如果能把(3.2.4)中的不等式 $c \leq S(T, g(\theta)) \leq d$ 等价地改写成 $\hat{\theta}_L(\mathbf{X}) \leq g(\theta) \leq \hat{\theta}_U(\mathbf{X})$ ，其中 $\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})$ 只与 c, d 和 T 有关，而与 θ 无关。则 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 就是 $g(\theta)$ 的置信水平为 $1 - \alpha$ 的置信区间。

枢轴量法

例 3.2.2 设 $X_1, \dots, X_n$ 为来自正态总体 $N(\mu, \sigma^2)$ 的 iid 样本, μ, σ^2 均未知, 试求 μ 的置信水平为 $1 - \alpha$ 的置信区间.

解 依枢轴量法的四个步骤, 我们如下求取:

- ❶ 由于 $\bar{X}$ 是 μ 的一个好的点估计, 故我们在第一步中取 $T = \bar{X}$.
- ❷ 由于 σ 未知, 故我们不能取 $\sqrt{n}(\bar{X} - \mu)/\sigma$ 作为枢轴量(虽然它的分布与参数 μ 无关, 但是它与另一未知的讨厌参数 σ 有关, 故第四步行不通). 此时, 既然 σ 未知, 就用样本标准差估计之, 于是, 我们可以取 $S = \sqrt{n}(\bar{X} - \mu)/S_n$.
- ❸ 因为 $\sqrt{n}(\bar{X} - \mu)/S_n \sim t(n-1)$, 故取 $c = t_{1-\alpha/2}(n-1) = -t_{\alpha/2}(n-1), d = t_{\alpha/2}(n-1)$.
- ❹ 因为

$$\begin{aligned} -t_{\alpha/2}(n-1) &\leq \sqrt{n}(\bar{X} - \mu)/S_n \leq t_{\alpha/2}(n-1) \\ \iff \bar{X} - \frac{S_n}{\sqrt{n}}t_{\alpha/2}(n-1) &\leq \mu \leq \bar{X} + \frac{S_n}{\sqrt{n}}t_{\alpha/2}(n-1), \end{aligned}$$

故

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

即为 μ 的置信水平为 $1 - \alpha$ 的置信区间.

□

枢轴量的广泛存在性

设 $x_1, x_2, \dots, x_n$ 是来自某分布函数 $F(x; \theta)$ 的一个样本，若此分布函数 $F(x; \theta)$ 既是 $\mathbf{x}$ 的连续函数，又是 θ 的单调函数，则可构造枢轴量，获得参数 θ 的置信区间。

目 录

3.1 置信区间

3.2 正态总体参数的置信区间

3.3 大样本置信区间

正态总体参数的置信区间

正态分布 $N(\mu, \sigma^2)$ 用途很广, 寻求它的两个参数 μ 与 σ^2 的置信区间是实际中经常会遇到的问题. 下面我们分几种情况来讨论这个问题:

一, 正态均值 μ 的置信区间

1, σ 已知时 μ 的置信区间

2, σ 未知时 μ 的置信区间

二, 正态方差 σ^2 的置信区间

三, 两正态总体的置信区间

正态总体均值的置信区间

设总体 $\xi \sim N(a, \sigma)$, $\xi_1, \dots, \xi_n$ 为其子样, 试求 a 及 σ^2 的区间估计.

1. 参数 a 的区间估计

(1) 在方差 $\sigma^2 = \sigma_0^2$ 为已知时, 求 a 的区间估计.

因为 $\bar{\xi} \sim N(a, \frac{\sigma_0}{\sqrt{n}})$, 故 $U = \frac{\bar{\xi} - a}{\sigma_0/\sqrt{n}} \sim N(0, 1)$. 给定置信水平 $1 - \alpha$, 查 $N(0, 1)$ 分布表得临界值 u_α , 使

$$P\left(\left|\frac{\bar{\xi} - a}{\sigma_0/\sqrt{n}}\right| \leq u_\alpha\right) = 1 - \alpha,$$

即

$$P\left\{\bar{\xi} - u_\alpha \frac{\sigma_0}{\sqrt{n}} \leq a \leq \bar{\xi} + u_\alpha \frac{\sigma_0}{\sqrt{n}}\right\} = 1 - \alpha \quad (3.0.2)$$

就是说, 随机区间 $\left[\bar{\xi} - u_\alpha \frac{\sigma_0}{\sqrt{n}}, \bar{\xi} + u_\alpha \frac{\sigma_0}{\sqrt{n}}\right]$ 是数学期望 a 的置信水平为 $1 - \alpha$ 的区间估计.

正态总体均值的置信区间

对于不同的置信水平 $1 - \alpha$, a 的区间估计随之不同, 如下表 7.4.1 所示:

表 7.4.1

α	u_α	$\left[\bar{\xi} - u_\alpha \frac{\sigma_0}{\sqrt{n}}, \bar{\xi} + u_\alpha \frac{\sigma_0}{\sqrt{n}} \right]$
0.01	2.58	$\left[\bar{\xi} - 2.58 \frac{\sigma_0}{\sqrt{n}}, \bar{\xi} + 2.58 \frac{\sigma_0}{\sqrt{n}} \right]$
0.05	1.96	$\left[\bar{\xi} - 1.96 \frac{\sigma_0}{\sqrt{n}}, \bar{\xi} + 1.96 \frac{\sigma_0}{\sqrt{n}} \right]$
0.10	1.65	$\left[\bar{\xi} - 1.65 \frac{\sigma_0}{\sqrt{n}}, \bar{\xi} + 1.65 \frac{\sigma_0}{\sqrt{n}} \right]$

正态总体均值的置信区间

(2) 在方差 σ^2 未知时, 求 a 的区间估计.

当方差未知时, 可用子样方差 S^2 估计 σ^2 , 已知

$$t = \sqrt{n-1} \frac{\bar{\xi} - a}{S} \sim t_{(n-1)},$$

给定置信水平 $1 - \alpha$ ($0 < \alpha < 1$), 在 t -分布表中查自由度 $n - 1$ 对应的临界值 $t_{(n-1)}(\alpha)$, 使得

$$P\left(\left|\sqrt{n-1} \frac{\bar{\xi} - a}{S}\right| > t_{(n-1)}(\alpha)\right) = \alpha,$$

即

$$P\left\{\bar{\xi} - t_{(n-1)}(\alpha) \frac{S}{\sqrt{n-1}} \leq a \leq \bar{\xi} + t_{(n-1)}(\alpha) \frac{S}{\sqrt{n-1}}\right\} = 1 - \alpha \quad (3.0.3)$$

就是说, $\left[\bar{\xi} - t_{(n-1)}(\alpha) \frac{S}{\sqrt{n-1}}, \bar{\xi} + t_{(n-1)}(\alpha) \frac{S}{\sqrt{n-1}}\right]$ 是 a 的置信度为 $1 - \alpha$ 的区间估计.

正态总体均值的置信区间

例 7.4.2 为了确定某种溶液的甲醛浓度, 取样得 4 个独立测定值的平均值 $\bar{x} = 8.34\%$, 标准差 $s = 0.03$, 并设被测总体服从正态分布, 求总体均值 μ 的 95% 的置信区间.

解因为 $1 - \alpha = 0.95$, $n = 4$, $n - 1 = 3$, 所以

$$t_{(n-1)}(\alpha) = t_3(0.05) = 3.1824, \quad \frac{s}{\sqrt{n-1}} = \frac{0.03}{\sqrt{3}},$$

于是 μ 的置信度是 0.95 的置信区间是

$$\left[8.34 - 3.1824 \times \frac{3}{\sqrt{3}}, 8.34 + 3.1824 \times \frac{3}{\sqrt{3}} \right].$$

正态总体方差的置信区间

2. 参数 σ^2 的区间估计

若总体 $\xi \sim N(a, \sigma)$, 则 $\frac{nS^2}{\sigma^2} \sim \chi^2_{(n-1)}$. 给定置信水平 $1-\alpha$, $0 < \alpha < 1$, 在 χ^2 -分布表中查临界值 $\chi^2_{(n-1)}(\frac{\alpha}{2})$ 及 $\chi^2_{(n-1)}(1 - \frac{\alpha}{2})$, 使得

$$P\left(\frac{nS^2}{\sigma^2} \geq \chi^2_{(n-1)}(\frac{\alpha}{2})\right) = \frac{\alpha}{2}, \quad P\left(\frac{nS^2}{\sigma^2} \geq \chi^2_{(n-1)}(1 - \frac{\alpha}{2})\right) = 1 - \frac{\alpha}{2},$$

即

$$P\left\{\chi^2_{(n-1)}(1 - \frac{\alpha}{2}) \leq \frac{nS^2}{\sigma^2} \leq \chi^2_{(n-1)}(\frac{\alpha}{2})\right\} = 1 - \alpha,$$

或

$$P\left\{\frac{nS^2}{\chi^2_{(n-1)}(\frac{\alpha}{2})} \leq \sigma^2 \leq \frac{nS^2}{\chi^2_{(n-1)}(1 - \frac{\alpha}{2})}\right\} = 1 - \alpha. \quad (3.0.4)$$

就是说, $\left[\frac{nS^2}{\chi^2_{(n-1)}(\frac{\alpha}{2})}, \frac{nS^2}{\chi^2_{(n-1)}(1 - \frac{\alpha}{2})}\right]$ 为方差 σ^2 的置信度为 $1 - \alpha$ 的区间估计.

单正态总体的置信区间

由于正态总体的实用性，故我们在结束本节之前，把关于单个正态总体参数的置信区间总结如下：

表3.1 单样本正态总体参数的置信区间

参数情况	μ 的置信区间	σ^2 的置信区间
σ^2 已知	$\left[\bar{X} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$	
σ^2 未知	$\left[\bar{X} - t_{\alpha/2}(n-1) \frac{S_n}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \frac{S_n}{\sqrt{n}} \right]$	
μ 已知		$\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{\alpha/2}^2(n)}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{1-\alpha/2}^2(n)} \right]$
μ 未知		$\left[\frac{(n-1)S_n^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)S_n^2}{\chi_{1-\alpha/2}^2(n-1)} \right]$

两正态总体的置信区间

设 $X_1, \dots, X_m$ 和 $Y_1, \dots, Y_n$ 为分别来自正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的样本, 且全样本是独立的, 其中 $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 为参数.

§3.3.1 Behrens-Fisher问题

我们感兴趣的是求 $\delta = \mu_2 - \mu_1$ 的置信区间. 这个问题是由Behrens于1929年提出的, 由于Fisher也讨论过这个问题, 故人们习惯称这个问题为Behrens-Fisher问题. 求取 δ 的置信区间的问题之所以成为一个很有名的问题, 是由于当我们对 σ_1^2, σ_2^2 一无所知时, 不知如何求取, 虽然, 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知时, 此问题已得到了很好的解决.

设样本均值为 $\bar{X} = \sum_{i=1}^m X_i/m, \bar{Y} = \sum_{i=1}^n Y_i/n$, 样本方差为 $S_{1m}^2 = \sum_{i=1}^m (X_i - \bar{X})^2/(m-1), S_{2n}^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/(n-1)$. 下面分几种情况讨论 δ 的置信区间.

两正态总体的置信区间

(1) 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知时

因为统计量

$$T = \sqrt{\frac{mn(m+n-2)}{m+n}} \frac{\bar{Y} - \bar{X} - \delta}{\sqrt{(m-1)S_{1m}^2 + (n-1)S_{2n}^2}} \sim t(m+n-2), \quad (3.3.1)$$

故可以取它作为枢轴量. 于是, 由此可求得 δ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\bar{Y} - \bar{X} \pm t_{\alpha/2}(m+n-2) \sqrt{\frac{m+n}{mn(m+n-2)}} \sqrt{(m-1)S_{1m}^2 + (n-1)S_{2n}^2} \right].$$

两正态总体的置信区间

(2) 当 $\theta = \sigma_2^2/\sigma_1^2$ 已知时

因为统计量

$$T = \sqrt{\frac{mn(m+n-2)}{m\theta + n}} \frac{\bar{Y} - \bar{X} - \delta}{\sqrt{(m-1)S_{1m}^2 + (n-1)S_{2n}^2/\theta}} \sim t(m+n-2), \quad (3.3.2)$$

于是, δ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\bar{Y} - \bar{X} \pm t_{\alpha/2}(m+n-2) \sqrt{\frac{m\theta + n}{mn(m+n-2)}} \sqrt{(m-1)S_{1m}^2 + (n-1)S_{2n}^2/\theta} \right].$$

两正态总体的置信区间

(3) 当 $m = n$ 时

因为 $Z_i = Y_i - X_i \sim N(\delta, \sigma_1^2 + \sigma_2^2)$, $i = 1, \dots, n$, 且相互独立, 于是知 $\bar{Z} \sim N(\delta, (\sigma_1^2 + \sigma_2^2)/n)$, $\sum_{i=1}^n (Z_i - \bar{Z})^2 / (\sigma_1^2 + \sigma_2^2) \sim \chi^2(n-1)$. 故有

$$\frac{\sqrt{n(n-1)}(\bar{Z} - \delta)}{\sqrt{\sum_{i=1}^n (Z_i - \bar{Z})^2}} \sim t(n-1), \quad (3.3.3)$$

由此可得 δ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\bar{Z} \pm t_{\alpha/2}(n-1) \frac{\sqrt{\sum_{i=1}^n (Z_i - \bar{Z})^2}}{\sqrt{n(n-1)}} \right],$$

两正态总体的置信区间

(4) 当 m, n 都充分大时

因为 $S_{1m}^2 \xrightarrow{P} \sigma_1^2$, $S_{2n}^2 \xrightarrow{P} \sigma_2^2$, 且 $\bar{Y} - \bar{X} \sim N(0, \sigma_1^2/m + \sigma_2^2/n)$, 故由Slutsky定理知

$$\tilde{T} = \frac{\bar{Y} - \bar{X} - \delta}{\sqrt{S_{1m}^2/m + S_{2n}^2/n}} \xrightarrow{\mathcal{L}} N(0, 1), \quad (3.3.4)$$

于是, δ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\bar{Y} - \bar{X} \pm u_{\alpha/2} \sqrt{S_{1m}^2/m + S_{2n}^2/n} \right],$$

两正态总体的置信区间

(5) 一般情况

当 m, n 不是很大时, (3.3.4)式左边的随机变量既不是 t 分布, 也不是正态分布. Welch于1938年通过研究认为 $\tilde{T}$ 的分布近似地与自由度为 r 的 t 分布相同, 其自由度的选取可如下考虑: 定义

$$T = \frac{\bar{Y} - \bar{X} - \delta}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}, \quad S_{mn}^{*2} = \frac{S_{1m}^2}{m} + \frac{S_{2n}^2}{n}, \quad Z = \frac{S_{mn}^{*2}}{\sigma_1^2/m + \sigma_2^2/n},$$

显然 $T \sim N(0, 1)$, Z 接近 χ^2 分布, 并且(3.3.4)式的 $\tilde{T}$ 可以改写成 $\tilde{T} = T/\sqrt{Z}$, 于是, 可以认为 $\tilde{T}$ 近似地服从一个 t 分布. 如果 $\tilde{T}$ 的自由度为 r , 则可以近似地认为 rZ 近似地服从 $\chi^2(r)$ 分布, 即 $\text{Var}Z \doteq 2/r$. 又由 Z 的定义可求

$$\text{Var}Z = \frac{\frac{2\sigma_1^4}{m^2(m-1)} + \frac{2\sigma_2^4}{n^2(n-1)}}{(\sigma_1^2/m + \sigma_2^2/n)^2},$$

两正态总体的置信区间

故可得

$$r \doteq \frac{(\sigma_1^2/m + \sigma_2^2/n)^2}{\frac{\sigma_1^4}{m^2(m-1)} + \frac{\sigma_2^4}{n^2(n-1)}},$$

但由于 σ_i^2 是未知的，故我们可用样本方差估计它，这样就得到了 r 的近似值为

$$r \doteq \frac{S_{mn}^{*4}}{\frac{S_{1m}^4}{m^2(m-1)} + \frac{S_{2n}^4}{n^2(n-1)}}. \quad (3.3.5)$$

由此可得 δ 的置信水平近似为 $1 - \alpha$ 的置信区间为

$$[\bar{Y} - \bar{X} \pm t_{\alpha/2}(r)S_{mn}^*].$$

两正态总体的置信区间

例 3.3.1 某厂利用两条自动化流水线装蕃茄酱罐头. 现分别从两条流水线上随机地各抽取一组样本容量为6和7的样本, 分别称重后得(单位: 克)

$$\bar{X} = 10.6, \bar{Y} = 10.1, S_{1m}^2 = 0.0125, S_{2n}^2 = 0.01.$$

假设两条流水线上所装蕃茄酱的重量都服从正态分布, 其均值分别为 μ_1, μ_2 , 而方差不等. 求 $\mu_1 - \mu_2$ 的置信水平为0.90的置信区间.

解 对于此问题, 我们只能利用(5)的一般情形近似计算之. 因为

$$S_{mn}^{*2} = \frac{0.0125}{6} + \frac{0.01}{7} = 0.003512, S_{mn}^* = 0.05926,$$

$$r = 0.003512^2 / \left(\frac{0.0125^2}{5 \times 6^2} + \frac{0.01^2}{6 \times 7^2} \right) = 10.209 \approx 10.$$

又因为 $t_{\alpha/2}(10) = 1.812$, 于是 $\mu_1 - \mu_2$ 的置信水平为0.9的置信区间为 $0.5 \pm 0.05926 \times 1.812 = [0.393, 0.607]$. □

两正态总体的置信区间

§3.3.2 方差比 σ_1^2/σ_2^2 的置信区间

因为 $(m-1)S_{1m}^2/\sigma_1^2 \sim \chi^2(m-1)$, $(n-1)S_{2n}^2/\sigma_2^2 \sim \chi^2(n-1)$, 且二者是独立的, 于是

$$F = \frac{S_{1m}^2/\sigma_1^2}{S_{2n}^2/\sigma_2^2} \sim F(m-1, n-1),$$

可以作为枢轴量, 并且

$$P \left\{ F_{1-\alpha/2}(m-1, n-1) \leq \frac{S_{1m}^2/\sigma_1^2}{S_{2n}^2/\sigma_2^2} \leq F_{\alpha/2}(m-1, n-1) \right\}$$

这样, 就得到了 σ_1^2/σ_2^2 的置信区间为

$$\left[\frac{S_{1m}^2/S_{2n}^2}{F_{\alpha/2}(m-1, n-1)}, \frac{S_{1m}^2/S_{2n}^2}{F_{1-\alpha/2}(m-1, n-1)} \right].$$

两正态总体的置信区间

我们把两样本正态总体参数的置信区间总结如下：

表3.2 两样本正态总体参数的置信区间

讨厌参数	待估参数	置信区间
$\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知	$\mu_2 - \mu_1$	$\left[\bar{Y} - \bar{X} \pm t_{\alpha/2}(m+n-2) \sqrt{\frac{m+n}{mn(m+n-2)}(Q_1^2 + Q_2^2)} \right]$
$\sigma_2^2/\sigma_1^2 = \theta$ 已知	$\mu_2 - \mu_1$	$\left[\bar{Y} - \bar{X} \pm t_{\alpha/2}(m+n-2) \sqrt{\frac{m\theta+n}{mn(m+n-2)}(Q_1^2 + Q_2^2/\theta)} \right]$
$m = n$ 时	$\mu_2 - \mu_1$	$\left[\bar{Z} \pm t_{\alpha/2}(n-1) \frac{\sqrt{\sum_{i=1}^n (Z_i - \bar{Z})^2}}{\sqrt{n(n-1)}} \right]$
m, n 充分大	$\mu_2 - \mu_1$	$[\bar{Y} - \bar{X} \pm u_{\alpha/2} S_{mn}^*]$
一般情况	$\mu_2 - \mu_1$	$[\bar{Y} - \bar{X} \pm t_{\alpha/2}(r) S_{mn}^*]$
μ_1, μ_2 未知	σ_1^2/σ_2^2	$\left[\frac{S_{1m}^2/S_{2n}^2}{F_{\alpha/2}(m-1, n-1)}, \frac{S_{1m}^2/S_{2n}^2}{F_{1-\alpha/2}(m-1, n-1)} \right]$

其中 $Q_1^2 = \sum_{i=1}^m (X_i - \bar{X})^2$, $Q_2^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$, $Z_i = Y_i - X_i$, $S_{mn}^{*2} = S_{1m}^2/m + S_{2n}^2/n$.

其他分布类型的区间估计

前面讨论的区间估计是在总体 $\xi \sim N(a, \sigma)$ 的假设下讨论的, 当总体 ξ 不服从正态分布时, 若其数学期望 $E(\xi)$ 及方差 $D(\xi)$ 都有限, 由中心极限定理知, $\eta_n = \frac{\bar{\xi} - a}{\sigma/\sqrt{n}}$ 的极限分布为正态 $N(0, 1)$ 分布, 故当 n 充分大 ($n > 50$), σ_0^2 已知时, 子样均值 $\bar{\xi}$ 渐近服从 $N(a, \frac{\sigma_0}{\sqrt{n}})$, 此时可用 (2) 式近似给出 a 的区间估计; 如果方差 σ^2 未知, 用子样方差 S^2 估计 σ^2 , 则可用 (3) 式近似地给出 a 的区间估计.

当然并不是所有的非正态分布都用 (2) 式和 (3) 式作近似计算, 对一些特殊分布, 可以直接导出子样均值 $\bar{\xi}$ 的精确分布, 从而求出参数的区间估计. 看下例:

其他分布类型的区间估计

例 7.4.4 设总体 ξ 服从指数分布, 密度函数为

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \lambda > 0 \\ 0 & \text{其它} \end{cases}$$

求参数 λ 的置信水平为 $1 - \alpha$ 的单侧上置信限及 λ 的置信水平为 $1 - \alpha$ 的置信区间.

其他分布类型的区间估计

解 我们知道 ξ 的特征函数为

$$\varphi(t) = \left(1 - j \frac{t}{\lambda}\right)^{-1},$$

于是, 子样均值 $\bar{\xi}$ 的特征函数

$$\varphi_{\bar{\xi}}(t) = \left(1 - j \frac{t}{n\lambda}\right)^{-n}.$$

记 $\eta = 2n\lambda\bar{\xi}$, 则 η 的特征函数

$$\varphi_{\eta}(t) = (1 - 2jt)^{-n}.$$

可见 $\eta \sim \chi^2(2n)$.

给定置信水平 $1 - \alpha$, 查得临界值 $\chi_{2n}^2(\alpha)$, 使得

$$P(2n\lambda\bar{\xi} \leq \chi_{2n}^2(\alpha)) = 1 - \alpha,$$

其他分布类型的区间估计

亦即

$$P\left(0 < \lambda \leq \frac{1}{2n\bar{\xi}} \chi_{2n}^2(\alpha)\right) = 1 - \alpha = P\left(-\infty < \lambda \leq \frac{1}{2n\bar{\xi}} \chi_{2n}^2(\alpha)\right) \quad (\lambda > 0)$$

因此, $\frac{1}{2n\bar{\xi}} \chi_{2n}^2(\alpha)$ 是参数 λ 的置信度为 $1 - \alpha$ 的单侧上置信限, 而注意到题设 $\lambda > 0$, 因此 $\left(0, \frac{1}{2n\bar{\xi}} \chi_{2n}^2(\alpha)\right]$ 是参数 λ 的置信度为 $1 - \alpha$ 的单侧置信区间.

再取

$$P\left(2n\lambda\bar{\xi} \geq \chi_{2n}^2\left(\frac{\alpha}{2}\right)\right) = \frac{\alpha}{2}, \quad P\left(2n\lambda\bar{\xi} \geq \chi_{2n}^2\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \frac{\alpha}{2}$$

其他分布类型的区间估计

亦即

$$P\left(\chi_{2n}^2\left(1 - \frac{\alpha}{2}\right) \leq 2n\lambda\bar{\xi} \leq \chi_{2n}^2\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha,$$

即

$$P\left(\frac{1}{2n\bar{\xi}} \chi_{2n}^2\left(1 - \frac{\alpha}{2}\right) \leq \lambda \leq \frac{1}{2n\bar{\xi}} \chi_{2n}^2\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha.$$

因此, $\left[\frac{1}{2n\bar{\xi}} \chi_{2n}^2\left(1 - \frac{\alpha}{2}\right), \frac{1}{2n\bar{\xi}} \chi_{2n}^2\left(\frac{\alpha}{2}\right)\right]$ 是参数 λ 的置信度为 $1 - \alpha$ 的区间估计.

目 录

3.1 置信区间

3.2 正态总体参数的置信区间

3.3 大样本置信区间

精确置信区间与近似置信区间

前面所讲的枢轴量法是构造精确置信区间的方法, 其特点是: 对给定的置信水平 $1 - \alpha$, 按照这些方法可以获得置信系数恰好为 $1 - \alpha$ 的置信区间. 这类方法常在小样本场合使用, 当然也可用于大样本场合.

还有一类构造置信区间的方法, 它们仅能在大样本场合使用, 所得的置信区间的置信系数不能精确地达到预先设定的置信水平 $1 - \alpha$, 只能近似于给定的置信水平 $1 - \alpha$, 这类方法统称为大样本方法, 所得的置信区间称为近似置信区间, 或称大样本置信区间.

本节主要介绍以下两种求大样本置信区间的方法:

基于 **MLE** 的近似置信区间

基于中心极限定理的近似置信区间

基于MLE的近似置信区间

定理 2.7.9 设参数空间 Θ 是开区间, 总体PDF $f(x; \theta)$ 满足

- (1) 在参数真值 θ_0 的邻域内, $\partial^i \ln f / \partial \theta^i$, $i = 1, 2, 3$ 存在;
- (2) 在参数真值 θ_0 的邻域内, $|\partial^3 \ln f / \partial \theta^3| \leq H(x)$, 且 $EH(X) < \infty$;
- (3) 在参数真值 θ_0 处,

$$E_{\theta_0} \left[\frac{f'(X, \theta_0)}{f(X, \theta_0)} \right] = 0, \quad E_{\theta_0} \left[\frac{f''(X, \theta_0)}{f(X, \theta_0)} \right] = 0, \quad I(\theta_0) = E_{\theta_0} \left[\frac{f'(X, \theta_0)}{f(X, \theta_0)} \right]^2 > 0.$$

记 $\hat{\theta}_n$ 为当 $n \rightarrow \infty$ 时似然方程的相合解, 则

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, I^{-1}(\theta_0)). \quad (2.7.9)$$

基于MLE的近似置信区间

由上述定理我们可以得到

$$\frac{\hat{\theta}_n - \theta}{\sigma_n(\theta)} \rightarrow N(0, 1), n \rightarrow \infty$$

在一般场合若用 $MLE\hat{\theta}_n$ 代替 $\sigma_n(\theta)$ 中的 θ , 上式仍然成立.

此时给定显著性水平 $1 - \alpha$, 我们有

$$P(-u_{1-\alpha/2} < \frac{\hat{\theta}_n - \theta}{\sigma_n(\theta)} < u_{1-\alpha/2}) = 1 - \alpha$$

从而可得 θ 的近似 $1 - \alpha$ 的等尾置信区间

$$\hat{\theta}_n \pm u_{1-\alpha/2}\sigma_n(\hat{\theta}_n)$$

基于MLE的近似置信区间

例题: 设 $x_1, x_2, \dots, x_n$ 是来自指数分布 $p(x; \theta) = \frac{1}{\theta}e^{-x/\theta} (x > 0)$ 的一个样本, 求参数 θ 的置信水平为 $1 - \alpha$ 的置信区间.

基于MLE的近似置信区间

譬如, 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 个样品做寿命试验, 试验到全部失效为止, 所得的 60 个寿命数据之和为 $T_n = 45079$. 求其 0.9 的置信区间.

基于中心极限定理的近似置信区间

一，中心极限定理

若 $\xi_1, \xi_2, \dots, \xi_n, \dots$ 是一串相互独立相同分布的随机变量序列, 且

$$E\xi_k = \mu, \quad D\xi_k = \sigma^2 \quad (5.3.10)$$

我们来讨论标准化随机变量和

$$\zeta_n = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^n (\xi_k - \mu) \quad (5.3.11)$$

的极限分布.

林德伯格 (Lindeberg) 与莱维建立了下列中心极限定理.

定理 5.3.2 (林德伯格-莱维) 对于标准化和 (5.3.11), 若 $0 < \sigma^2 < \infty$, 则

$$\lim_{n \rightarrow \infty} P\{\zeta_n < x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \quad (5.3.12)$$

基于中心极限定理的近似置信区间

由上述中心极限定理, 我们可以构造出总体均值 μ 的近似 $1 - \alpha$ 的等尾置信区间:

$$\bar{x} \pm u_{1-\alpha} \sigma \sqrt{n}$$

若总体中 σ 尚未知, 可用 σ^2 的相合估计代替.

基于中心极限定理的近似置信区间

例 3.2.7 设 $X_1, \dots, X_n$ 为来自 Bernoulli 总体 $b(1, p)$ 的 iid 样本, 试求 p 的置信水平为 $1 - \alpha$ 的置信区间.

基于中心极限定理的近似置信区间

解 对于此问题, 由于我们知道 $\sum_{i=1}^n X_i$ 是充分统计量, 故枢轴量应与它有关. 又因为 $\sum_{i=1}^n X_i \sim B(n, p)$, 故我们不可能由此构造一个满足枢轴量法条件的②与④. 但由中心极限定理可知

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{npq}} \xrightarrow{\mathcal{L}} N(0, 1), \text{ 当 } n \rightarrow \infty \text{ 时,}$$

即当 n 充分大时, 我们有

$$P \left\{ \frac{T_n - np}{\sqrt{npq}} < x \right\} \approx \Phi(x),$$

且与 p 无关. 也就是说, 当 n 充分大时,

$$P \left\{ -u_{\alpha/2} \leq \frac{T_n - np}{\sqrt{npq}} \leq u_{\alpha/2} \right\} \approx 1 - \alpha. \quad (*)$$

又因为

$$-u_{\alpha/2} \leq \frac{T_n - np}{\sqrt{npq}} \leq u_{\alpha/2} \iff n(\bar{X} - p)^2 \leq p(1-p)u_{\alpha/2}^2,$$

基于中心极限定理的近似置信区间

而上式右端的关于 p 的二次三项式: $n(\bar{X} - p)^2 - p(1 - p)u_{\alpha/2}^2$ 的判别式

$$\Delta = (-2n\bar{X} - u_{\alpha/2}^2)^2 - 4(n + u_{\alpha/2}^2)n\bar{X}^2 = 4nu_{\alpha/2}^2\bar{X}(1 - \bar{X}) + u_{\alpha/2}^4 > 0,$$

且其平方项系数大于零, 故可以求得 $n(\bar{X} - p)^2 - p(1 - p)u_{\alpha/2}^2 = 0$ 的两个根为

$$p = \frac{n}{n + u_{\alpha/2}^2} \left[\bar{X} + \frac{u_{\alpha/2}^2}{2n} \pm u_{\alpha/2} \sqrt{\frac{\bar{X}(1 - \bar{X})}{n} + \frac{u_{\alpha/2}^2}{4n^2}} \right],$$

分别以 $\hat{p}_L < \hat{p}_U$ 记之. 则当 $p \in (\hat{p}_L, \hat{p}_U)$ 时, 上述二次三项式小于零. 于是由(*)式知,

$$P\{\hat{p}_L \leq p \leq \hat{p}_U\} \approx 1 - \alpha,$$

即说明 $[\hat{p}_L, \hat{p}_U]$ 是 p 的置信水平近似 $1 - \alpha$ 的置信区间. □

我们注意到, 上述的近似置信水平 $1 - \alpha$ 的置信区间, 当 n 充分大时, 它非常接近水平 $1 - \alpha$. 在实际中, 当 $n \geq 30$, 就可以认为是充分大了.

基于中心极限定理的近似置信区间

譬如, 在某电视节目的收视率调查中, 调查了 400 人, 其中 100 人收看了该电视节目, 试求该节目收视率 p 的置信水平为 0.95 的置信区间.

我们在这里

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他人回复问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

