参数估计

参数估计是统计推断的主要内容，一般在总体分布已知但参数未知的情况下，利用样本构造合理的统计量对参数进行估计，可分为点估计和区间估计

点估计

设总体的分布为 $F(x;\theta)$ ，其中 $\theta=(\theta_1,\theta_2,\dots,\theta_k)$ 为参数，则根据样本 $X_1,X_2,\dots,X_n$ 构造一个**统计量** $\hat\theta(X_1,X_2,\dots,X_n)$ 作为参数 θ 的估计。称

$$\hat\theta(X_1,X_2,\dots,X_n)$$

为 θ 的**估计量**，如果将样本的值带入 $\hat\theta$ 则得到

$$\hat\theta(x_1,x_2,\dots,x_n)$$

为 θ 的估计值。这样的估计方法称为点估计

点估计主要可分为矩估计和极大似然估计

矩估计

矩估计的基本思想就是：用样本矩作为总体矩的估计

如果参数 $\theta=(\theta_1,\theta_2,\dots,\theta_n)$ 可以表示为总体矩的函数 $\theta_i=h_i(\mu_1,\mu_2,\dots,\mu_k)$ ，则用样本矩 $A_1,A_2,\dots,A_k$ 代替总体矩，得到的估计量即为矩估计量，即

$$\hat\theta_i=h_i(A_1,A_2,\dots,A_k)$$

矩估计的基本原理是，根据大数定律，样本矩按概率收敛到总体矩，若 h 为已知连续函数，则有

$$h(A_1,A_2,\dots,A_k)\xrightarrow{P}h(\mu_1,\mu_2,\dots,\mu_k)$$

矩估计十分简单，因为并没有用到总体分布的信息，但是也正因为没有充分利用总体分布的信息，精度不高

极大似然估计

极大似然估计的基本思想是，要为参数选一个合理的估计值，就要使参数在取该值时，样本发生的概率最大

对于离散型总体 X，样本 X_i 取值 x_i 的概率为 $p(X=x_i;\theta)$ ，则所有样本取观测值的概率为

$$L(x_1,x_2,\ldots,x_n;\theta)=L(\theta)=\prod_{i=1}^{n}p(X=x_i;\theta)$$

上述函数称为极大似然函数，极大似然估计就是选取使上述概率达到最大的参数值 $\hat{\theta}$ 作为参数 θ 的估计，即

$$\hat{\theta}=\arg\max L(\theta)$$

对于连续型总体 X，若其密度函数为 $p(x;\theta)$ ，则样本的概率密度函数为 $\prod_{i=1}^{n}p(x_i;\theta)$ ，样本取值在 $(x_1,x_2,\ldots,x_n)$ 邻域的概率可近似为 $\prod_{i=1}^{n}p(x_i;\theta)dx_i$ ，则为方便起见，选取极大似然函数为

$$L(x_1,x_2,\ldots,x_n;\theta)=L(\theta)=\prod_{i=1}^{n}p(x_i;\theta)$$

极大似然函数一般为连乘形式，而由于 $L(\theta)>0$ ，且 $\ln L(\theta)$ 为 $L(\theta)$ 的单调函数，有相同的最大值点，故可以用 $\ln L(\theta)$ （称为对数似然函数）

写出极大似然函数后求其偏导，令偏导数为 0，解方程（组）即可得到极大似然估计量/值

有时也会有解方程法失效的情况，此时需要结合极大似然估计的定义来分析，即选择能使极大似然函数取值尽可能大的参数值

极大似然估计的不变性：若 $\hat{\theta}$ 是 θ 的极大似然估计，$\phi(\theta)$ 有单值反函数，则 $\phi(\hat{\theta})$ 是 $\phi(\theta)$ 的极大似然估计

估计量的评价标准

无偏性

设 $\hat{\theta}(X_1,X_2,\ldots,X_n)$ 为参数 θ 的一个估计量，若对任意 $\theta\in\Theta$ 满足

$$E[\hat{\theta}(X_1,X_2,\ldots,X_n)]=\theta$$

则称 $\hat{\theta}$ 是 θ 的**无偏估计量**

样本原点矩是总体原点矩的无偏估计，即

$$E[A_k]=E\left[\frac{1}{n}\sum_{i=1}^{n}X_i^k\right]=\frac{1}{n}\sum_{i=1}^{n}E[X_i^k]=\mu_k$$

样本方差也是总体方差的无偏估计

均方误差准则

用 $E[(\hat{\theta}-\theta)^2]$ 来估计 $\hat{\theta}$ 与 θ 偏离的程度，记为

$$M(\hat{\theta},\theta)=E[(\hat{\theta}-\theta)^2]$$

称其为**均方误差**

显然，均方误差越小越好

对于均方误差的计算，有

$$M(\hat{\theta},\theta)=D[\hat{\theta}]+(E[\hat{\theta}]-\theta)^2$$

显然在 $\hat{\theta}$ 无偏时，均方误差就是 $\hat{\theta}$ 的方差

一致性

一致性的含义是当样本量越来越大时，估计量也应当越来越接近真实参数

设 $\hat{\theta}(X_1,X_2,...,X_n)$ 为参数 θ 的一个估计量，若对任意 $\theta\in\Theta$ 满足

$$\hat{\theta}\xrightarrow{P}\theta$$

则称 $\hat{\theta}$ 是 θ 的**一致估计量**

区间估计

基本概念

除了构造一个点来估计未知参数以外，也可以构造一个区间 $(\hat{\theta_1},\hat{\theta_2})$ 来估计参数 θ 的范围。

区间包含参数 θ 的概率称为区间的置信度

如果设 θ 是总体 X 的未知参数，$X_1,X_2,...,X_n$ 是来自总体的样本，若对事先给定的常数 $\alpha(0<\alpha<1)$ 存在两个统计量

$$\hat{\theta_1}(X_1,X_2,...,X_n)$$
$$\hat{\theta_2}(X_1,X_2,...,X_n)$$

满足

$$P(\hat{\theta_1}<\theta<\hat{\theta_2})=1-\alpha$$

则称区间 $(\hat{\theta_1},\hat{\theta_2})$ 为 θ 的置信度为 $1-\alpha$ 的**置信区间**

需要注意的是，置信度代表在取多个区间时，大约有 $100(1-\alpha)\%$ 的区间包含 θ ，而对于某个确定的区间来说 θ 只有在其中和不在其中两种可能

有时候仅关注未知参数的上限/下限，此时称为单侧置信区间

如果设 θ 是总体 X 的未知参数，$X_1, X_2, \ldots, X_n$ 是来自总体的样本，若对事先给定的常数 $\alpha(0<\alpha<1)$ 存在统计量 $\hat{\theta_1}$ 使得

$$P(\hat{\theta_1} < \theta) = 1-\alpha$$

则称 $(\hat{\theta_1}, \infty)$ 为 θ 的置信度为 $1-\alpha$ 的单侧置信区间，上限同理

枢轴变量法

对于区间估计，有一种一般的方法

- 寻找一个样本函数 $U(X_1, X_2, \ldots, X_n)$ 包含待估计的参数 θ ，但是不含其他未知参数。并且 U 的分布要已知。这个函数称为枢轴变量
- 由于 U 的分布已知，故可以根据给定的置信度 $1-\alpha$ ，找到两个常数 a, b 满足
$$P(a < U < b) = 1-\alpha$$
- 变形不等式，解出 $\hat{\theta_1} < \theta < \hat{\theta_2}$ ，即为所求置信区间

正态总体的置信区间

正态总体均值的置信区间

给定置信度 $1-\alpha$ ，设 $X_1, X_2, \ldots, X_n$ 为来自总体 $X \sim N(\mu, \sigma^2)$ 的一组样本

若 σ^2 已知，则可以取枢轴变量

$$U = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

解不等式得到置信区间为

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

若 σ^2 未知，则取枢轴变量

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

解不等式得到置信区间为

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

正态总体方差的置信区间

给定置信度 $1-\alpha$ ，设 $X_1,X_2,\ldots,X_n$ 为来自总体 $X\sim N(\mu,\sigma^2)$ 的一组样本

若 μ 未知，取枢轴变量

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

解不等式即可得到置信区间为

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

两个正态总体均值差的置信区间

给定置信度 $1-\alpha$ ，设 $X_1,X_2,\ldots,X_{n_1}$ 为来自总体 $X\sim N(\mu_1,\sigma_1^2)$ 的一组样本，$Y_1,Y_2,\ldots,Y_{n_2}$ 是来自总体 $Y\sim N(\mu_2,\sigma_2^2)$ 的一组样本，且两样本相互独立

若 σ_1^2,σ_2^2 已知，则取枢轴变量

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

解不等式得到置信区间为

$$\left(\overline{X}-\overline{Y}-u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\overline{X}-\overline{Y}+u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

可以类比单个正态总体的情况，基本原理为

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

若 $\sigma_1^2=\sigma_2^2=\sigma^2$ 但 σ^2 未知，则取枢轴变量

$$T=\sqrt{\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{(n_1-1)S_1^2+(n_2-1)S_2^2}}\sim t(n_1+n_2-2)$$

为了方便表示，设

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

则置信区间为

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

同样可以类比单个正态总体的情况

两个正态总体方差比的置信区间

给定置信度 $1-\alpha$ ，设 $X_1,X_2,\dots,X_{n_1}$ 为来自总体 $X\sim N(\mu_1,\sigma_1^2)$ 的一组样本，$Y_1,Y_2,\dots,Y_{n_2}$ 是来自总体 $Y\sim N(\mu_2,\sigma_2^2)$ 的一组样本，且两样本相互独立

若 μ_1,μ_2 未知，则取枢轴变量

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

则置信区间为

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

非正态总体的区间估计

总体分布非正态时，通常采用大样本法，求得枢轴变量的近似分布

设 $X_1,X_2,\dots,X_n$ 为来自均值为 μ ，方差为 σ^2 的总体的一组样本，给定置信度 $1-\alpha$ ，求均值 μ 的置信区间，则根据中心极限定理，有

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\approx N(0,1)$$

由于 σ^2 未知，使用样本标准差 S 代替，枢轴变量为

$$U=\frac{\overline{X}-\mu}{S/\sqrt{n}}\approx N(0,1)$$

解不等式，置信区间为

$$\left(\overline{X}-u_{\alpha/2}\frac{S}{\sqrt{n}},\overline{X}+u_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$