第六讲：参数估计

6.2 区间估计

对于一个未知量, 人们在测量和计算时, 常不以得到近似值为满足, 还需要估计误差, 及要求知道近似值的精确程度 (亦即所求真值所在的范围). 类似的, **对于未知的参数 θ, 除了求出它的点估计 $\hat{\theta}$ 外, 我们还希望估计出一个范围, 并希望知道这个范围包含参数 θ 真值得可信程度.** 这样的范围通常以区间形式给出, 同时还给出此区间包含真值的可信程度. 这种形式的估计称为**区间估计**.

↑Example

比如你估计月花费支出是 500, 我们相信多少会有误差, 但是误差有多大? 单从你提出的 500 这个数字还给不出什么信息, 若你给出估计支出是 400-600 之间, 则人们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. 因此区间估计也是常用的一种估计方式.

↓Example

现在最流行的一种区间估计理论是 J. Neyman 在上世纪 30 年代建立起来的. 他的理论的基本概念很简单, 为表达方便, 我们暂时假定总体分布只包含一个未知参数 θ, 且要估计的就是 θ 本身. 如果总体分布中包含若干位置参数 $\theta_1, \cdots, \theta_k$, 而要估计的是 $g(\theta_1, \cdots, \theta_k)$, 则基本概念和方法并无不同. 这在后面的例子里可以看出.

6.2.1 置信区间

Neyman 建立起来的区间估计也叫**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**.

假设 $X_1, \cdots, X_n$ 是从该总体中抽取的样本, 所谓 (一维未知)θ 的区间估计, 就是要

- **寻求统计量 $\underline{\theta}(X_1, \cdots, X_n) < \bar{\theta}(X_1, \cdots, X_n)$ 所构成的区间 $[\underline{\theta}, \bar{\theta}]$.**
- **该区间满足一定的要求**

不难理解, 这里有两个要求

- **θ 以很大概率被包含在区间 $[\underline{\theta}, \bar{\theta}]$ 内, 也就是说**

$$P_\theta(\underline{\theta} \leq \theta \leq \bar{\theta}) = 1 - \alpha$$

  尽可能大, 即要求估计尽量可靠.

- **估计的精度要尽可能高，比如要求区间 $[\underline{\theta}, \bar{\theta}]$ 要尽可能的短, 或者某种能体现这个要求的其他准则。**

比如估计一个人的年龄, 如 [30,35], 我们自然希望这个人的年龄有很大把握在这个区间之内, 并且希望这个区间不能太长. 如果估计是 [10,90], 当然可靠了, 但是精度太差, 用处不大.

但这两个要求是相互矛盾的，因此区间估计的原则是在已有的样本资源限制下，找出更好的估计方法以尽量提高可靠性和精度。Neyman 提出了广泛接受的准则：**先保证可靠性，在此前提下尽可能提高精度**。为此，引入如下定义:

Definition

设总体分布 $F(x,\theta)$ 含有一个或多个未知的参数 θ，$\theta \in \Theta$，对给定的值 $\alpha,(0<\alpha<1)$，若由样本 $X_1,\cdots,X_n$ 确定的两个统计量 $\bar{\theta}=\bar{\theta}(X_1,\cdots,X_n)$ 和 $\underline{\theta}=\underline{\theta}(X_1,\cdots,X_n)$，满足

$$P_\theta(\underline{\theta} \le \theta \le \bar{\theta}) = 1-\alpha \qquad \forall\ \theta \in \Theta$$

称 $1-\alpha$ 为置信系数或置信水平，而称 $[\underline{\theta},\bar{\theta}]$ 为 θ 的置信水平为 $1-\alpha$ 的置信区间。

置信区间就是在给定的置信水平之下，去寻找有优良精度的区间。

一般，我们首先寻求参数 θ 的一个估计 (多数是基于其充分统计量构造的)，然后基于此估计量构造参数 θ 的置信区间，介绍如下:

1. 枢轴变量法 设待估参数为 $g(\theta)$，

1. 找一个与待估参数 $g(\theta)$ 有关的统计量 T，一般是其一个良好的点估计 (多数是通过极大似然估计构造)；

2. 设法找出 T 与 $g(\theta)$ 的某一函数 $S(T, g(\theta))$ 的分布，其分布 F 要与参数 θ 无关 (S 即为枢轴变量);

3. 对任何常数 $a < b$，不等式 $a \le S(T, g(\theta)) \le b$ 要能表示成等价的形式 $A \le g(\theta) \le B$，其中 A, B 只与 T, a, b 有关而与参数无关；

4. 取分布 F 的上 $\alpha/2$ 分位数 $\omega_{\alpha/2}$ 和上 $(1-\alpha/2)$ 分位数 $\omega_{1-\alpha/2}$，有 $F(\omega_{\alpha/2}) - F(\omega_{1-\alpha/2}) = 1 - \alpha$. 因此

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

由 3 我们就可以得到所求的置信区间.

设 $X_1, \cdots, X_n$ 为从正态总体 $N(\mu, \sigma^2)$ 中抽取得样本，求参数μ, σ^2 的 $1-\alpha$ 置信区间。

↑Example

↓Example

解： 由于 μ, σ^2 的估计 $\bar{X}, S^2$ 满足

$$T_1 = \sqrt{n}(\bar{X} - \mu)/S \sim t_{n-1}$$
$$T_2 = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$$

所以 T_1, T_2 就是我们所要寻求的枢轴变量，从而易得参数 μ, σ^2 的 $1-\alpha$ 置信区间分别为

$$\left[\bar{X} - \frac{1}{\sqrt{n}} S t_{n-1}(\alpha/2), \bar{X} + \frac{1}{\sqrt{n}} S t_{n-1}(\alpha/2)\right],$$
$$\left[\frac{(n-1)S^2}{\chi^2_{n-1}(\alpha/2)}, \frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha/2)}\right].$$

设 $X_1,\cdots,X_n$ 为从正态总体 $N(\mu_1,\sigma_1^2)$ 中抽取得样本，$Y_1,\cdots,Y_m$ 为从正态总体 $N(\mu_2,\sigma_2^2)$ 中抽取得样本，两组样本相互独立。求参数$\mu_1-\mu_2, \sigma_1^2/\sigma_2^2$ 的 $1-\alpha$ 置信区间。

解： 方法完全类似于前面的例子，由于 $\mu_1,\mu_2,\sigma_1^2,\sigma_2^2$为 的估计分别 $\bar{X},\bar{Y},S_X^2,S_Y^2$，且注意到 $\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m})$, $(n-1)S_X^2/\sigma_1^2\sim\chi^2_{n-1}$ 以及 $(m-1)S_Y^2/\sigma_2^2\sim\chi^2_{m-1}$, 结合两组样本的独立性可知

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

从而可得 σ_X^2/σ_Y^2 的置信区间. 对 $\mu_1-\mu_2$ 的置信区间, 当 σ_X^2,σ_Y^2 已知或者相等但未知情形，容易得到其置信区间；当两者不全已知且不相等时，不存在 $\mu_X-\mu_Y$的精确置信区间 (Behrens-Fisher problem).

2. 大样本法

大样本法就是利用极限分布，以建立枢轴变量。通过以下例子说明:

↑Example

某事件 A 在每次实验中发生的概率都是 p，作 n 次独立的实验，以 Y_n 记 A 发生的次数。求 p 的 $1-\alpha$ 置信区间。

↓Example

解： 设 n 比较大，令 $q=1-p$, 则由中心极限定理知，近似有 $(Y_n-np)/\sqrt{npq}\sim N(0,1)$，从而 $(Y_n-np)/\sqrt{npq}$ 可以作为枢轴变量。由

$$P(-u_{\alpha/2}\leq (Y_n-np)/\sqrt{npq}\leq u_{\alpha/2})\approx 1-\alpha \qquad (*)$$

可以等价表示成

$$P(A\leq p\leq B)\approx 1-\alpha$$

其中 A, B 为方程

$$(Y_n - np)/\sqrt{npq} = u_{\alpha/2}$$

的解, 即

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

A 取负号，B 取正号，$\hat{p} = Y_n/n$。

由于 (*) 式只是近似成立，故区间估计也只是近似成立，当 n 较大时才相去不远。详细的说明参见课本 p203。我们还可以先假定方差是 “已知” 的，最后再将其估计，得到如下 Wald 置信区间:

$$\hat{p} \pm u_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}.$$

6.2.2 置信界

在实际中，有时我们只对参数 θ 的一端的界限感兴趣。比如果汁的最低含量, 有害物质的最高含量等等.

设总体分布 $F(x,\theta)$ 含有一个未知的参数 θ, $\theta\in\Theta$, 对给定的值 $\alpha,(0<\alpha<a)$, 若由样本 $X_1,\cdots,X_n$ 确定的两个统计量 $\bar{\theta}=\bar{\theta}(X_1,\cdots,X_n)$ 和 $\underline{\theta}=\underline{\theta}(X_1,\cdots,X_n)$,

1．若

$$P_\theta(\theta\leq\bar{\theta})\geqq 1-\alpha \qquad \forall\ \theta\in\Theta$$

则称 $\bar{\theta}$ 为 θ 的一个置信系数为 $1-\alpha$ 的**置信上界**.

2．若

$$P_\theta(\theta\geq\underline{\theta})\geqq 1-\alpha \qquad \forall\ \theta\in\Theta$$

则称 $\underline{\theta}$ 为 θ 的一个置信水平为 $1-\alpha$ 的**置信下界**.

Definition

而 $(-\infty,\bar{\theta}]$ 和 $[\underline{\theta},+\infty)$ 都称为是单边的置信区间。寻求置信上、下界的方法和寻求置信区间的方法完全类似。

6.2.3 确定样本大小

在以区间长度为精度准则下，置信区间越窄就越好，为什么呢？作为一个一般的原则，我们已经知道更多的测量可以得到更精确的推断。有时候，对精度是有要求的，甚至于是在测量之前就提出此要求，因此相应的样本大小就要事先确定下来。我们以如下的例子说明如何确定样本大小，一般的方法类似。

假设某种成分的含量服从正态分布 $N(\mu,\sigma^2)$，σ^2 已知。要求平均含量 μ 的 $(1-\alpha)$ 置信区间的长度不能长于 ω。试确定测量样本大小。

↑Example

↓Example

解: 由于 σ^2 已知，我们已经知道可以根据 $\bar{X} \sim N(\mu, \sigma^2/n)$ 来构造μ 的 95% 置信区间。因此易知区间长度为 $2u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. 从而由

$$2u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \omega$$

得到

$$n \geq \left(\frac{2u_{\alpha/2}\sigma}{\omega}\right)^2.$$

比如当 $\sigma = 0.1, \omega = 0.05, \alpha = 0.05$, 可以得到 $n \geq \left(\frac{2\times 1.96\times 0.1}{0.05}\right)^2 = 61.4656$. 即为达到要求至少需要测量 62 次。