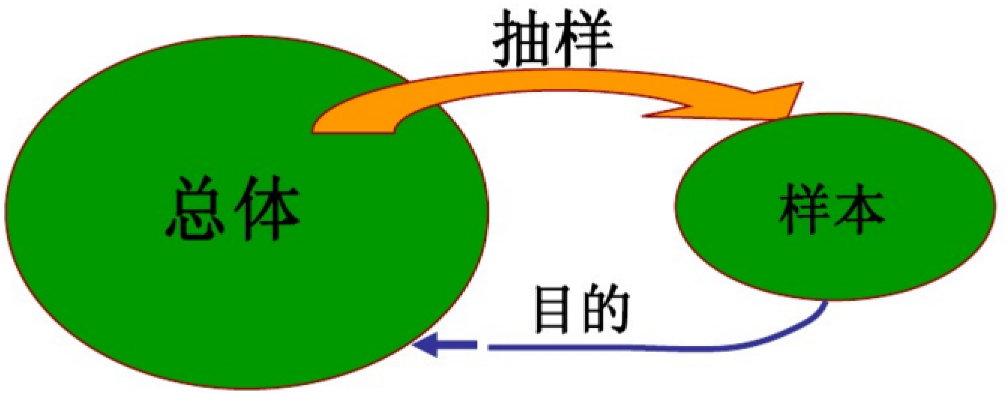
第5章 参数估计

李镇清

中国科学院大学

统计学基本问题：研究总体与样本的关系

参数
(parameter)：总体特征数，也称参量。



用样本信息来推断总体特征

统计量
(staistic)：样本特征数，也称统计数，由样本算出的量，或者说统计量就是样本的函数。统计量只依赖于样本，而不能与任何未知的量有关。

抽样分布：样本统计量的分布

样本统计量推断总体参数，以抽样分布为基础
统计推断:假设检验，参数估计

参数的区间估计与点估计

一、参数区间估计与点估计的原理

二、总体平均数的区间估计与点估计

三、两个总体平均数差数的区间估计与点估计

参数估计（parameter estimation）

参数估计（parameter estimation），统计推断的一种。根据从总体中抽取的随机样本来估计总体分布中未知参数的过程。从估计形式看，区分为点估计与区间估计；从构造估计量的方法讲，有矩法估计、最小二乘估计、似然估计、贝叶斯估计等。

要处理两个问题：

（1）求出未知参数的估计量；

（2）在一定置信度（可靠程度）下指出所求的估计量的精度。信度一般用概率表示，如置信程度为95%；精度用估计量与被估参数（或待估参数）之间的接近程度或误差来度量。

用估计量估计总体参数

估计量是估计总体参数的统计量，两种估计，一种是点估计(point estimation)，即用估计量的实现值来近似相应的总体参数。

另一种是区间估计(interval estimation)；它是包括估计量在内（有时是以估计量为中心）的一个区间；该区间被认为很可能包含总体参数。

点估计给出一个数字，用起来很方便；而区间估计给出一个区间，说起来留有余地；不像点估计那么绝对。

点估计

最常用的估计量：样本均值、样本标准差(s)和(Bernoulli试验的)成功比例(x/n)；

用它们来分别估计总体均值(μ)、总体标准差(σ)和成功概率(或总体中的比例)p.

点估计

用什么样的估计量来估计参数呢？

- 衡量一个估计量好坏的标准。每个标准一般都仅反映估计量的某个方面。
- 这样就出现了按照这些标准定义的各种名目的估计量（如无偏估计量等）。
- 另一些估计量则是由它们的计算方式来命名的（如最大似然估计和矩估计等）。

好估计量的标准

（1）无偏性

无偏性是指估计量抽样分布的数学期望等于总体参数的真值。无偏性的含义是，估计量是一随机变量，对于样本的每一次实现，由估计量算出的估计值有时可能偏高，有时可能偏低，但这些估计值平均起来等于总体参数的真值。在平均意义下，无偏性表示没有系统误差。

（2）有效性

有效性是指估计量与总体参数的离散程度。如果两个估计量都是无偏的，那么离散程度较小的估计量相对而言是较为有效的。离散程度是用方差度量的，因此在无偏估计量中，方差愈小愈有效。

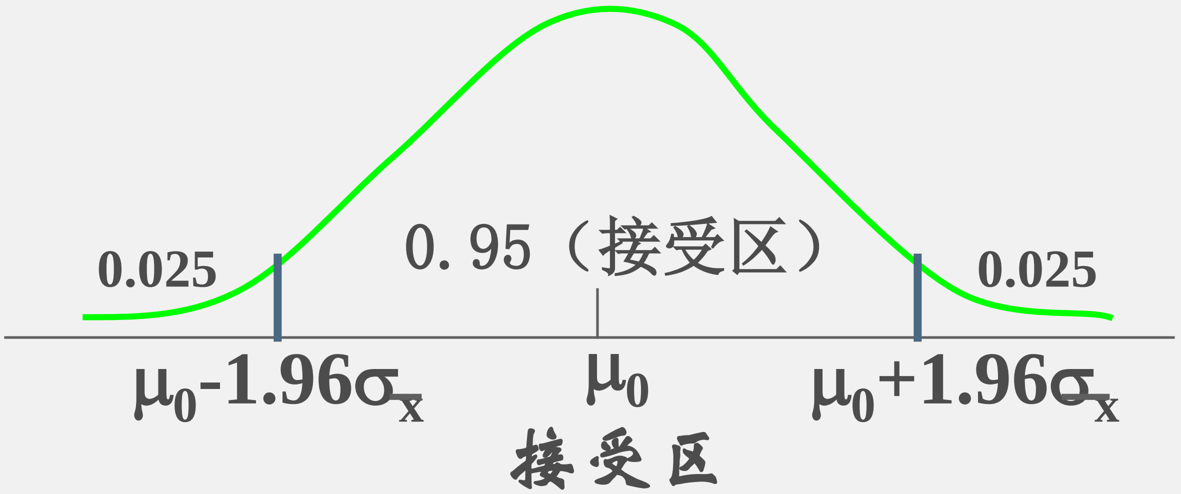
（3）一致性

一致性，又称相合性，是指随着样本容量的增大，估计量愈来愈接近总体参数的真值

一、参数区间估计与点估计的原理

参数的区间估计与点估计是建立在一定理论基础上的一种方法。

由中心极限定理和大数定律，只要抽样为大样本，不论其总体是否为正态分布，其样本平均数都近似服从正态分布$N(\mu, \sigma^2/n)$。



临界值 $\mu \pm u_{\alpha/2}\sigma_{\bar{x}}$

$$P(\mu - 1.96\sigma_{\bar{x}} < \bar{x} < \mu + 1.96\sigma_{\bar{x}}) = 0.95$$

$$P(\bar{x} > \mu + 1.96\sigma_{\bar{x}}) + P(\bar{x} > \mu - 1.96\sigma_{\bar{x}}) = 0.05$$

$$P(\mu - 2.58\sigma_{\bar{x}} < \bar{x} < \mu + 2.58\sigma_{\bar{x}}) = 0.99$$

$$P(\bar{x} > \mu + 2.58\sigma_{\bar{x}}) + P(\bar{x} > \mu - 2.58\sigma_{\bar{x}}) = 0.01$$

$$P(\mu - 1.96\sigma_{\bar{x}} < \bar{x} < \mu + 1.96\sigma_{\bar{x}}) = 0.95$$

$$P(-1.96\sigma_{\bar{x}} < \bar{x} - \mu < 1.96\sigma_{\bar{x}}) = 0.95$$

$$P(-\bar{x} - 1.96\sigma_{\bar{x}} < -\mu < -\bar{x} + 1.96\sigma_{\bar{x}}) = 0.95$$

$$P(\bar{x} + 1.96\sigma_{\bar{x}} > \mu > \bar{x} - 1.96\sigma_{\bar{x}}) = 0.95$$

$$P(\bar{x} - 1.96\sigma_{\bar{x}} < \mu < \bar{x} + 1.96\sigma_{\bar{x}}) = 0.95$$

$$P(\mu - 2.58\sigma_{\bar{x}} < \bar{x} < \mu + 2.58\sigma_{\bar{x}}) = 0.99$$

$$P(\bar{x} - 2.58\sigma_{\bar{x}} < \mu < \bar{x} + 2.58\sigma_{\bar{x}}) = 0.99$$

$$P(\bar{x} - 1.96\sigma_{\bar{x}} < \mu < \bar{x} + 1.96\sigma_{\bar{x}}) = 0.95$$

$$P(\bar{x} - 2.58\sigma_{\bar{x}} < \mu < \bar{x} + 2.58\sigma_{\bar{x}}) = 0.99$$

$$P(\bar{x} \quad u_{\alpha/2}\sigma_{\bar{x}} < \mu < \bar{x} + u_{\alpha/2}\sigma_{\bar{x}}) = 1 - \alpha$$

u_α： 正态分布下置信度P=1- α时的u临界值

α： 置信水平

$$P(\bar{x} - u_{\alpha/2}\sigma_{\bar{x}} < \mu < \bar{x} + u_{\alpha/2}\sigma_{\bar{x}}) = 1 - \alpha$$

知道 $\bar{x}$，但不知道μ

1- α置信区间、置信距 $(\bar{x} - u_{\alpha/2}\sigma_{\bar{x}},\ \bar{x} + u_{\alpha/2}\sigma_{\bar{x}})$

$$(L_1 = \bar{x} - u_{\alpha/2}\sigma_{\bar{x}}, L_2 = \bar{x} + u_{\alpha/2}\sigma_{\bar{x}})$$

$$\left(L_1 = \bar{x} - u_{\alpha/2}\sigma_{\bar{x}}, L_2 = \bar{x} + u_{\alpha/2}\sigma_{\bar{x}}\right)$$

用样本平均数 $\bar{x}$ 对总体平均数μ的置信度为P=1-α的区间估计。

$$L = \bar{x} \pm u_{\alpha/2}\sigma_{\bar{x}}$$

用样本平均数 $\bar{x}$ 对总体平均数μ的置信度为P=1-α的点估计。

二、总体平均数μ的区间估计和点估计

当为大样本时，不论总体方差σ^2为已知或未知，可以利用样本平均数x和总体方差σ^2作出置信度为P＝1-α的中体平均数的区间估计为：

$$(L_1 = \bar{x} - u_{\alpha/2}\sigma_{\bar{x}}, L_2 = \bar{x} + u_{\alpha/2}\sigma_{\bar{x}})$$

其置信区间的下限L₁和上限L₂为

$$L_1 = \bar{x} - u_{\alpha/2}\sigma_{\bar{x}}$$

$$L_2 = \bar{x} + u_{\alpha/2}\sigma_{\bar{x}}$$

总体平均数的点估计L为

$$L = \bar{x} \pm u_{\alpha/2}\sigma_{\bar{x}}$$

当样本为小样本且总体方差σ^2未知时，σ^2需由样本方差s^2来估计，于是置信度为$P=1-\alpha$的总体平均数μ的置信区间可估计为

$$\left(\bar{x}-t_{\alpha/2}s_{\bar{x}},\ \bar{x}+t_{\alpha/2}s_{\bar{x}}\right)$$

其置信区间的下限L_1和上限L_2为：

$$(L_1 = \bar{x} - t_{\alpha/2} s_{\bar{x}}, L_2 = \bar{x} + t_{\alpha/2} s_{\bar{x}})$$

总体平均数的点估计L为：

$$L = \bar{x} \pm t_{\alpha/2} s_{\bar{x}}$$

T_a为正态分布下置信度P＝1－ α时的t临界值

例4.14测得某批25个小麦样本的平均蛋白质含量＝14.5%，已知σ ＝2.50%，试进行95%置信度下的蛋白质含量的区间估计和点估计。

分析：本例σ 为已知, 置信度P＝1- α =0. 95，$u_{0.05}$=1. 96。

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.50}{\sqrt{25}} = 0.50$$

$$L_1 = \bar{x} - u_{\alpha/2}\sigma_{\bar{x}} = 14.5 \quad 1.96 \times 0.50 = 13.52(\%)$$

$$L_2 = \bar{x} + u_{\alpha/2}\sigma_{\bar{x}} = 14.5 + 1.96 \times 0.50 = 15.48(\%)$$

蛋白质含量的点估计为：

$$L = \bar{x} \pm u_{\alpha/2}\sigma_{\bar{x}} = 14.5 \pm 1.96 \times 0.50 = 14.5 \pm 0.98$$

说明小麦蛋白质含量有95％的把握落在13.52％～15.48％的区间里。

例题　从某渔场收对虾的总体中，随机取20尾对虾，测的平均体长x＝120mm，标准差是＝15mm，试估计置信度为99％的对虾总体平均数

本例中，由于总体方差σ ^{2}未知，需用s^2估计σ^2，当df＝20－1＝19时，$t_{0.01/2}$＝2.861。具体计算如下

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{15}{\sqrt{20}} = 3.354$$

于是对虾体长的区间估计为

$$L_1 = \bar{x} - t_{\alpha/2} s_{\bar{x}} = 120 - 2.861 \times 3.354 = 110.4(mm)$$

$$L_2 = \bar{x} + t_{\alpha/2} s_{\bar{x}} = 120 + 2.861 \times 3.354 = 129.6(mm)$$

对虾体长的点估计为：

$$L = \bar{x} \pm t_{\alpha/2} s_{\bar{x}}$$

$$= 120 \pm 2.861 \times 3.354 = 120 \pm 9.6(mm)$$

说明对虾体长有99％把握落在110.4mm～129.6mm区间里

三、两个总体平均数差数

$\mu_1-\mu_2$的区间估计与点估计

当两个总体方差σ_1^2和σ_2^2为已知，或总体方差σ_1^2和σ_2^2未知但为大样本时，在置信度为P＝1- α下，两个总体平均数差数μ_1-μ_2的区间估计为：

$$\left[\left(\bar{x}_1 \quad \bar{x}_2\right) - u_{\alpha/2}\sigma_{\bar{x}_1 \quad \bar{x}_2}, \left(\bar{x}_1 \quad \bar{x}_2\right) + u_{\alpha/2}\sigma_{\bar{x}_1 \quad \bar{x}_2}\right]$$

其置信区间的下限L $_1$和上限L $_2$为：

$$\left[L_1=\left(\overline{x_1}-\overline{x_2}\right)-u_{\alpha/2}\sigma_{\overline{x_1}-\overline{x_2}},L_2=\left(\overline{x_1}-\overline{x_2}\right)+u_{\alpha/2}\sigma_{\overline{x_1}-\overline{x_2}}\right]$$

两个总体平均数差数$\mu_1-\mu_2$的点估计L为

$$L=\left(\overline{x_1}-\overline{x_2}\right)\pm u_{\alpha/2}\sigma_{\overline{x_1}-\overline{x_2}}$$

当两个样本为小样本，总体方差σ_1^2和σ_2^2未知，当两总体方差相等，即$\sigma_1^2 = \sigma_2^2 = \sigma^2$时，可由两样本方差$s_1^2$和$s_2^2$估计总体方差$\sigma_1^2$和$\sigma_2^2$，在置信度为$P = 1-\alpha$下，两总体平均数差数$\mu_1-\mu_2$的区间估计为：

$$\left[\left(\overline{x_1} - \overline{x_2}\right) - t_{\alpha/2} s_{\overline{x_1} - \overline{x_2}}, \left(\overline{x_1} - \overline{x_2}\right) + t_{\alpha/2} s_{\overline{x_1} - \overline{x_2}}\right]$$

其置信区间的下限L_1和上限L_2为：

$$\left[L_1=(\overline{x_1}-\overline{x_2})-t_{\alpha/2}s_{\overline{x_1}-\overline{x_2}},\ L_2=(\overline{x_1}-\overline{x_2})+t_{\alpha/2}s_{\overline{x_1}-\overline{x_2}}\right]$$

两个总体平均数差数$\mu_1-\mu_2$的点估计L为：

$$L=(\overline{x_1}-\overline{x_2})\pm t_{\alpha/2}s_{\overline{x_1}-\overline{x_2}}$$

当两个样本为小样本，总体方差σ_1^2和σ_2^2未知，且两总体方差不相等，即$\sigma_1^2 \neq \sigma_2^2$时，可由两样本方差$s_1^2$和$s_2^2$对总体方差$\sigma_1^2$和$\sigma_2^2$的估计而算出的t值，已不是自由度df＝n1+n2-2的t分布，而是近似的服从自由度df '的t分布，在置信度为P＝1-α下，两总体平均数差数$\mu_1-\mu_2$的区间估计为：

$$\left[\left(\overline{x_1} \quad \overline{x_2}\right) \quad t_{\alpha/2(df')} s_{\overline{x_1} \quad \overline{x_2}}, \left(\overline{x_1} \quad \overline{x_2}\right) + t_{\alpha/2(df')} s_{\overline{x_1} \quad \overline{x_2}}\right]$$

其置信区间的下限 L 1和上限L2为:

$$\left[L_1 = \left(\overline{x_1} - \overline{x_2}\right) \quad t_{\alpha/2(df')} s_{\overline{x_1}\ \overline{x_2}}, L_2 = \left(\overline{x_1} - \overline{x_2}\right) + t_{\alpha/2(df')} s_{\overline{x_1} - \overline{x_2}} \right]$$

两个总体平均数差数μ_1-μ_2的点估计 L 为：

$$L = \left(\overline{x_1} - \overline{x_2}\right) \pm t_{\alpha, df'} s_{\overline{x_1} - \overline{x_2}}$$

上面三式中，$t_{\alpha/2,\ df'}$为置信度为P=1- α时自由度为df '的t临界值。

当两样本为成对资料时，在置信度为P＝1- α时，两总体平均数差数μ_1-μ_2的置信区间可估计为：

$$\left(\bar{d} - t_{\alpha/2} s_{\bar{d}}, \bar{d} + t_{\alpha/2} s_{\bar{d}}\right)$$

其置信区间的下限L 1和上限L2为：

$$\left(L_1 = \bar{d} - t_{\alpha/2} s_{\bar{d}}, L_2 = \bar{d} + t_{\alpha/2} s_{\bar{d}}\right)$$

两个总体平均数差数μ_1-μ_2的点估计 L 为：

$$L = \bar{d} \pm t_{\alpha/2} s_{\bar{d}}$$

例题 用高蛋白和低蛋白两种饲料饲养一月龄大白鼠，在三个月时，测定两组大白鼠的增重重量（g），两组的数据分别为：

高蛋白组：134，146，106，119，124，161，107，83，113，129，97，123

低蛋白组：70，118，101，85，107，132，94

试进行置信度为95％时两种蛋白饲料饲养的大白鼠增重的差数区间估计和点估计。

已算得 $\overline{x_1}=120.17g, \overline{x_2}=101.00g$

$s_{\overline{x_1}-\overline{x_2}}=10.005$ $df=17, t_{0.05}=2.110$

其置信度为95%时两种蛋白饲料饲养的大白鼠增重的差数区间估计为：

$$L=\left(\overline{x_1}\quad\overline{x_2}\right)\quad t_{\alpha/2}s_{\overline{x_1}\quad\overline{x_2}}$$

$$=(120.17\quad 101.00)\quad 2.110\times 10.005=\quad 1.94(g)$$

$$L=\left(\overline{x_1}-\overline{x_2}\right)+t_{\alpha/2}s_{\overline{x_1}\quad\overline{x_2}}$$

$$=(120.17-101.00)+2.110\times 10.005=40.284(g)$$

两种蛋白质饲料饲养的大白鼠增重的差数点估计为：

$$L=\left(\overline{x_1}-\overline{x_2}\right)\pm t_{\alpha/2}s_{\overline{x_1}-\overline{x_2}}$$
$$=(120.17-101.00)\pm 2.110\times 10.005$$
$$=19.17\pm 21.11(g)$$

说明两种蛋白饲料饲养下大白鼠增重的差数有95％的把握落在-1.94g～40.284g的区间里。

例题 试对表4－1资料进行置信度为99％的区间估计和点估计。

已算得

$$\bar{d} = 812.5 IU \cdot g^{-1},$$

$$s_{\bar{d}} = 193.13 IU \cdot g^{-1}$$

$$df = 7, t_{0.01} = 3.499$$

于是，两种饲料饲养下动物肝脏中维生素A含量差数的区间估计为：

$$
\begin{aligned}
L &= \bar{d} - t_{\alpha/} s_{\bar{d}} \\
&= 812.5 - 3.499 \times 193.13 \\
&= 136.74 (IU \quad g^{\ 1})
\end{aligned}
$$

$$
\begin{aligned}
L &= \bar{d} + t_{\alpha/2} s_{\bar{d}} \\
&= 812.5 + 3.499 \times 193.13 \\
&= 1488.26 (IU \quad g^{\ 1})
\end{aligned}
$$

两种饲料饲养下动物肝脏中维生素A含量差数的点估计为：

$$L = \overline{d} + t_{\alpha/2} s_{\overline{d}}$$
$$= 812.5 \pm 3.499 \times 193.13$$
$$= 812.5 \pm 675.76 (IU \quad g^{\ 1})$$

说明两种饲料饲养下动物肝脏中维生素A含量差数的有99％的把握落在136.74IU·g-1～148.26 IU·g-1的区间里。