中国人民大学 陈传波 chris@ruc.edu.cn

第四讲 估计量及其性质

估计量是随机变量，可用分布理论来描述

核心问题：如何加工观察值来更好逼近未知参数值？

建立模型——明确待估参数——收集数据——寻找最优估计量——评估估计量——估计量的抽样分布——推断估计量与待估参数之间的关系——未知分布下的渐近推断

一、问题

天安门城楼到底有多高，其“真实”的高度究竟是多少？恐怕没有人能够给出完全精确的答案，即使给出来也未必令人信服。尽管实践是检验真理的唯一标准，可是如果我们去测量，每次测量的结果都不同，又应该相信哪一次的结果呢？如何来处理这很多次的测量结果，以便最好地逼近真实高度呢？

上述问题可以转化为如下数学模型：

$$y = \beta + u$$

其中 β 为天安门城楼的高度，β 是一个客观存在，是有唯一精确值的未知数。y 为测量结果，u 称为测量误差，y 和 u 都可看作随机变量，尽管一次测量完成后，我们知道 y 的值，但测量之前却不可能知道。u 是我们的理论构造，在真实世界中并不存在。显然，只有 y 是能够观察到的，而真实高度与误差却无法观察到，核心问题是：**如何用观察到的 y 来求得未知的 β**？

我们可以把总体也可视为一种数据生成机制（DGP，data generating process）。天门安城楼高度的每一个测量结果均由 $y = \beta + u$ 这一机制生成。在前一讲，我们是已知总体分布（包括分布类型的所有参数取值），根据分布来推断随机变量落在特定区间的概率。在这一讲，总体信息未知或部分未知，首要任务是先获得总体的分布，β 是总体的一个未知参数，由于 β 未知，所以目前这个 DGP 仍然是一个黑箱，计量分析的任务就是要通过收集样本来确定其中的未知常参数，以打开这个黑箱。

二、抽样

最容易理解的抽样是从有限总体中抽取一个样本，比如从 100 个混有红球和黑球的暗箱中摸出 8 个球来。有限样本的抽样又分为放回和不放回两种，如果放回，则同一个球可能被抽中多次。

从无限总体中抽取一个样本可被视为某个数据生成过程（DGP）的一次实现。比如天门安城楼高度的测量结果 y 是一个无限总体，某一次的测量结果可视为按照公式 $y=\beta+u$ 所确定的数据产生机制生成的一个数据 y_i。相应地 n 次测量的结果可被视为一个 n 维随机向量（样本）。

如果给定样本容量 n，即每次抽取 n 个观察值得到一个样本。

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

如果重新再去测量 n 次，将得到另外一个样本容量为 n 的新样本。显然，两个不同的样本中 y 的取值不同，因此，我们可以将其视为随机向量。

简单随机抽样是指每个样本被抽取的可能性等同，也就是事前不知道会抽中哪个样本，每个样本都有同样的可能性被抽中。

独立指的是各个试验或观察得到的样本间是相互独立的。独立性要求每一次取样的结果不影响另一次取样的结果。

注意独立和随机是两回事，随机样本并不一定相互独立，而相互独立的两个样本并不一定随机。

样本联合概率密度：已知随机变量 y,从该总体中随机的取一个容量为 n 的样本，其联合概率密度可记为 $f_J(\boldsymbol{y})$。

独立同分布是从服从同一分布的总体中随机独立地抽取样本，其联合概率密度满足公式：

$$f_J(\boldsymbol{y})=f_J(y_1,y_2,\cdots y_n)=\prod_{i=1}^{n}f(y_i)$$

每一个被抽中的样本都满足模型

$$y = \beta + u$$

也可以记为

$$y_i = \beta + u_i$$

还可以用向量的形式表达为

$$\boldsymbol{y} = \boldsymbol{\iota}\beta + \boldsymbol{u}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \beta + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

三、估计方法

取得样本后，我们有了多个数据，如何处理这些数据呢？同样的问题曾困惑着 18 世纪和 19 世纪初的许多天文学家和数学家。那个时代的人热衷于测量天体（比如慧星）的轨道长度，他们在很多地方建立天文台，反复测量，得到大量的数据。“每次测量都有误差，次数越多，误差累积越多，但把次数减少并不是解决问题的办法，用什么办法来恰当地使用大量的数据呢”？勒让德（Legendre,1752-1833）提出了“最小二乘法”，解决了如何从数据中得出准确结论的问题。而著名的数学家高斯（1777-1855）也声称他发明了最小二乘法，并用此方法准确地预测了谷神星在天空中出现的时间和位置。

最小二乘法的核心思想是：使样本点与参数的距离最小，这种距离通常以平方和来表示，因此称为最小二乘估计。

$$\min_{\arg b} \sum_{i=1}^{n} (y_i - b)^2$$

根据这个式子，我们就可以计算出

$$b = \bar{y}$$

b 称为 β 的最小二乘估计量（OLS）。上式右边可被看作函数 $g(y_1,y_2,\ldots,y_n)$。实际上，估计量是一个处理随机样本的法则，这个法则是抽样之前就已制定好的，不管实际上得到的是什么数据，这个法则都不变。

既然估计量是随机变量的函数，它也是一个随机变量，其随机性由样本决定，随着样本而变，代入不同的样本值，同一个估计量会得到不同的估计值。

上机 1：估计量与估计值

```
sysuse auto, clear
sample 10
sum price
```

反复执行上面的三行命令，每一次我们都得到不同的均值。

同样，反复执行下述三行命令，每一次我们也得到不同的估计值

```
drawnorm u,n(8) clear
g    y=10+u
reg     y
sum    y
```

最小二乘估计（OLS）是一种获得样本加工法则的思路，运用这种思路推导，可以得到一个样本的函数（即估计量）。类似地，矩估计方法和极大似然方法是另外两种思路，运用这两种思路也可以得到一种加工样本数据的法则（函数），该函数可能与 OLS 相同，也可能不同。

四、估计量

（一）线性无偏估计量

对同一个样本，可以定义无穷多的估计量，这些估计量仅依赖于总体的性质和定义估计量的函数。我们不能控制总体特征，因为它是由客观分布规律所决定的，而客观分布规律又是由自然规律或社会力量来决定。但是我们可以选择定义估计量的函数，即可选择加工处理样本数据的方法。问题是我们该选择什么样的函数来处理观察到的样本呢？什么样的函数更好处理，更容易计算逼近我们想要的那个参数呢？

潜在的函数既可以是线性的也可以是非线性的，但线性的往往比较容易处理。因为**线性估计量**是因变量 y 的线性函数（线性组合），相当于给每个样本点某个权重，然后加权求和。

其次，既然估计量是随机变量，它也就具有期望和方差等数字特征，而估计量的期望既取决于样本特征，也取决于我们所选择的函数形式（数据处理法则）。**无偏估计量是一类特殊估计量，无偏估计量的期望等于总体参数真值。**注意估计量的**无偏性评价的是估计法则的特性，而不是特定样本**。一个估计量的无偏性和可能偏误的大小既依赖于 Y 的分布，也依赖于函数 g（），通常 Y 的分布是我们不能选择的，但函数 g（）的选择操纵在我们手中，如果我们想要得到一个无偏估计量，我们就要对 g（）做相应的选择，放弃那些导致有偏的加工处理数据方法，比如将权重全部取为零，也是一种估计，但这样估计得到的结果必然为 0，它离天安门的真实高度相关甚远。

无偏性反映的是有限样本的性质，它可以理解为穷尽所有可能的抽样，然后利用每个样本按照 g（）计算出估计值，各估计值依概率（样本出现的概率）加权求和，得到的期望应等于总体参数真值。

线性无偏估计量是同时满足线性和无偏性的估计量。

在测量天安门高度的例子中，估计量 b 是线性的吗？是无偏估计量吗？是线性无偏估计量吗？如果不是，需要满足什么条件才是一个线性无偏估计量呢？

首先 b 是一个线性估计量， b 是 n 个观察值 y_i 的加权，每个观察值的权重为 1/n。

其次，要使 b 成为一个无偏估计量，则 $Eb=\beta$

$$E\hat{b} = E[(\iota'\iota)^{-1}\iota'y] = E[(\iota'\iota)^{-1}\iota'\iota b + (\iota'\iota)^{-1}\iota'u] = b + (\iota'\iota)^{-1}\iota'Eu = b$$

即当 $Eu=0$ 时，b 为线性无偏估计量。

如果假设 $Eu=0$ 不成立，则 b 是有偏的，在什么情况下，误差为零的假设不成立呢？比如测量时用的工具并不准确，总是偏大。再比某测量员总是倾向于高估测量结果等。

上机 2：回归

```
clear
scalar   b=int(10*uniform())+1      //生成一个随机的b，事先不知
drawnorm   u,n(10)                  //Eu=0
g   y=b+u
reg   y                  //猜测b为多少，看看实际结果与猜测之间的差异
```

```
scalar  list  b        //真实的b

clear
scalar  b=int(10*uniform())+1
drawnorm  u,n(10)  m(10)        //Eu>0,存在系统误差
g  y=b+u
reg  y                  //猜测b为多少，看看实际结果与猜测之间的差异
scalar  list  b         //真实的b
```

（二）有效估计量

除了上述线性无偏估计量外，考虑另一个线性无偏估计量 $\tilde{\beta}=Y_1$，因为

$$E(\tilde{\beta})=EY_1=\beta+Eu_1=\beta$$

显然也是线性无偏估计量，我们又如何在这两个估计量中间选择更好的一个呢？办法是进一步比较估计量的方差，并选择方差最小的那一个。如果两个无偏估计量 b_1 和 b_2，总有 $Var(b_1)<Var(b_2)$，则称 b_1 比 b_2 相对有效。如果不限于考虑无偏估计量，那么比较方差大小就毫无意义。比如，无论取到什么样本，我们都设定 一个等于 0 的估计量，其方差最小，但毫无意义。同时满足线性、无偏、最小方差的估计量称为**最小方差线性无偏估计量（BLUE）**。

$$Var(u)=\sigma^2 \Rightarrow Var(y)=\sigma^2$$

$$Var[b]=Var[\bar{y}]=Var(\frac{y_1+y_2+\cdots+y_n}{n}) \xrightarrow{iid} \sum_{i=1}^{n} Var(\frac{y_i}{n})=\frac{\sigma^2}{n}$$

若把 n 个样本视为 n 维随机向量 ***y*** 和 **u**，则其方差将变为

$$Var(\boldsymbol{u})=E(\boldsymbol{uu}')=\sigma^2 I=\begin{bmatrix}\sigma^2 & 0 & & 0\\ 0 & \sigma^2 & & 0\\ & & \cdot & \\ 0 & 0 & & \sigma^2\end{bmatrix}$$

该假定称为同方差假定，在同方差假定下，主对角线元素均相同，如果样本之间独立（至少不相关），则非主角线元素均为0。这一假定并不必然成立，后面我们会进一步讨论不成立时的情形。

$$Var(b) = Var[(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{y}] = Var[\boldsymbol{\beta} + (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{u}] = (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'Var(\boldsymbol{u})\boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} = \frac{\sigma^2}{n}$$

证明：在零均值和同方差假设下，OLS 估计量 b 为 BLUE 估计量

$$Var(b) = Var[(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{y}] = Var[\boldsymbol{\beta} + (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{u}] = (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'Var(\boldsymbol{u})\boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} = \frac{\sigma^2}{n}$$

$$\tilde{\beta} = Ay$$

$$E\tilde{\beta} = E(Ay) = E(A\boldsymbol{\iota}\beta + A\boldsymbol{u}) \xrightarrow{E\boldsymbol{u}=0} = A\boldsymbol{\iota}\beta = \beta \rightarrow A\boldsymbol{\iota} = 1$$

$$Var(\tilde{\beta}) - Var(b) = \sigma^2[AA' - (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}] = \sigma^2[AA' - A\boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'A'] = \sigma^2 AMA' \geq 0$$

五、误差方差的估计

尽管我们已得到

$$Var(b) = \frac{\sigma^2}{n}$$

但是，由于σ^2未知，仍然无法求出具体的值。同β一样，σ^2是模型中的另一个未知常参数，同样可以用样本观察值 y 来进行估计。这样的估计量也有无穷多个，我们需要找到具有优良性质的估计量，如无偏估计量或一致估计量。

为此，我们定义残差

$$e_i = y_i - \hat{y} = y_i - b = y_i - \bar{y}$$

考虑残差平方和

$$\sum e_i^2 = \sum (y_i - \bar{y})^2 = \boldsymbol{y}'\boldsymbol{M}\boldsymbol{y} = \boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}$$

两边取期望

$$E[\sum e_i^2] = E(\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}) = (n-1)\sigma^2$$

上式的推导过程中，要用到独立同分布假设，由于独立,当 i≠j 时，$E(u_iu_j)=0$，i=j 时 $E(u_iu_j)=\sigma^2$，因此只有 M 的主对角线元素成为系数，但这些系数乘以同一个常数σ^2并求和，可先求和再乘，求和正好等于 n-1，于是上式成立

无偏估计量为

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-1} = \frac{e'e}{n-1}$$

由于$Var(b)=\frac{\sigma^2}{n}$，估计量 b 的方差的无偏估计为

$$\hat{Var}(b) = \frac{\hat{\sigma}^2}{n} = \frac{\boldsymbol{e'e}}{n(n-1)}$$

其平方根称为标准误 se

$$se(b) = \sqrt{\frac{\hat{\sigma}^2}{n}} = \sqrt{\frac{\boldsymbol{e'e}}{n(n-1)}}$$

注意比较下面的五个概念：

总体方差：

$$Var(y) = \sigma^2$$

均方差(mean squared error,MSE）定义为：

$$MSE[b] = E[b-\beta]^2 = Var(b) + [Eb-\beta]^2$$

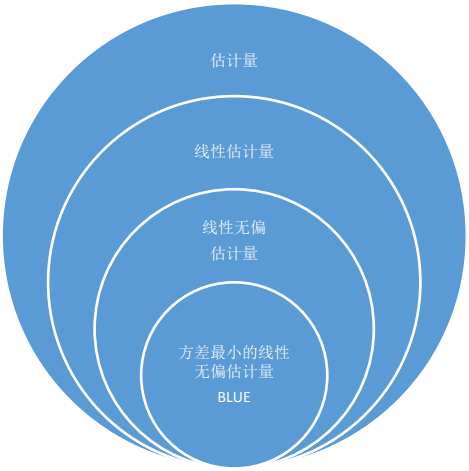
样本方差：

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-1} = \frac{e'e}{n-1}$$

估计量方差：既然估计量 b 是随机变量，它也有方差，其方差为σ^2/n

估计量方差的估计：是对估计量 b 的方差的一个估计。

$$\hat{Var}(b) = \frac{\boldsymbol{e}'\boldsymbol{e}}{n(n-1)}$$



六、估计量的分布

（一）抽样分布

既然估计量是一个随机变量，它就有相应的分布函数，称之为抽样分布。b 服从什么分布呢？

$$u \rightarrow N(0, \sigma^2)$$

由于均值相当于随机向量的一个函数（线性组合）。组合之后，均值仍然为随机的，而且成为一个随机变量。由于正态随机变量的线性组合仍然服从正态分布。因此 b 也服从正态分布，正态分布由均值和方差确定，故

$$b \sim N(\beta, \frac{\sigma^2}{n})$$

上机 3：估计量的抽样分布

有限总体抽样

```
clear
set obs 10000
```

```
g   y=10+rnormal()
save   temp,replace
capt   prog   drop   sd
prog   sd
u   temp,clear
sample   8,count
reg   y
end
***将上述抽样试验进行 100 次，得到 100 个均值和标准差
simulate   _b, reps (100):sd
sum                    //比较两者的均值和标准差。
tw (kdensity _b) (function y=normalden(x,10,1/sqrt(8)),range(9 11))
```

无限总体抽样

下面的例题，首先生成一个均值为 0，标准差为 1 的随机误差项，然后生成 Y，再抽取 8 个样本，计算其均值。重复上述程序 1000 次，得到 1000 个估计值，做这些估计值的直方图，可以发现，它服从正态分布。

```
capt prog drop sd
prog   sd
drawnorm u,n(8) clear   //8 个期望为 10 的正态随机样本
g   y=10+u
reg   y
end
***将上述抽样试验进行 1000 次，得到 1000 个均值和标准差
simulate   _b, reps (1000):sd
sum                    //比较两者的均值和标准差
```

```
mean    y     _b              //mean 计算标准误，su 计算总体标准差
tw (kdensity _b) (function y=normalden(x,10,1/sqrt(8)),range(5 15))
```

（二）误差方差的分布

$\hat{\sigma}^2$ 是一个估计量，自然是一个随机变量，那么这个随机变量服从什么分布呢？

由 $\boldsymbol{u} \rightarrow \boldsymbol{N(0, \sigma^2 I)}$，可得 $\frac{\boldsymbol{u}}{\sigma} \rightarrow \boldsymbol{N(0, I_n)}$，于是

$$\frac{(n-1)\hat{\sigma}^2}{\sigma^2} = \frac{\boldsymbol{e'e}}{\sigma^2} \xrightarrow{\boldsymbol{e=Mu,}} = \frac{\boldsymbol{u'}}{\sigma} M \frac{\boldsymbol{u}}{\sigma} \xrightarrow{rank(M)=n-1,\ \frac{\boldsymbol{u}}{\sigma} \rightarrow \boldsymbol{N(0, I_n)}} \rightarrow \chi^2_{n-1}$$

最后一步根据第三章关于标准正态分布随机变量的二次型服从卡方分布的定理得到。

上机 4：误差方差分布

```
clear
capt   prog   drop   sd
prog   sd
drawnorm   u,n(8)   clear   //8 个期望为 10 的正态随机样本
g   y=10+u
reg   y
scalar   s=7*(e(rmse))^2
end
```

***将上述抽样试验进行1000次，得到1000个均值和标准差

```
simulate   s, reps (1000):sd
g   chi=rchi2(7)
tw   (kdensity    _s) (kdensity   chi)
```

（三）t 估计量

在上面的分布中，β 和 σ 是未知的常参数，因而仍然无法确定估计量 b 的具体分布。怎么办呢？能否在 σ 未知的情况下得到某个具体的分布？

办法是构造 t 值，t 值是一个含有未知常参数 β 的估计量（因为 b 和 S 都是样本的函数），而且 t 值的分布函数仅有样本容量 n 唯一确定。

$$b \to N(\beta, \frac{\sigma^2}{n}) \Rightarrow \frac{b-\beta}{\sqrt{\sigma^2/n}} \to N(0,1)$$

$$\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \to \chi^2_{n-1}$$

根据 t 分布的定义：标准正态分布与卡方分布除以自由度后两者之比

$$t = \frac{\frac{b-\beta}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)\hat{\sigma}^2}{\sigma^2}/(n-1)}} = \frac{b-\beta}{\sqrt{\frac{\hat{\sigma}^2}{n}}} = \frac{b-\beta}{se(b)} \to t(n-1)$$

注意到 t 值实际上也是样本的一个函数，然而当总体服从正态分布时，t 值成为一个仅与样本容量有关的统计量。注意到上式中仅有一个未知常参数 β，我们把这种统计量称为枢柚量。

七、区间估计

区间估计的含义是：总体参数 β（真值）被由样本和置信水平构造的区间覆盖住的概率。根据一个样本的观察值给出总体参数的估计范围，并给出总体参数落在这一区间的概率

t 分布仅有一个参数，即样本容量 n，当 n 的大小被确定，分布即被决定。随机变量 t 落在（-∞，$-t_{0.025}$）和（$+t_{0.025}$，+∞）内的概率为 0.05,t 落在（$-t_{0.025}$,$+t_{0.025}$）的概率为 0.95。

$$0.95 = P(|t| < t_{0.025}) = P(-t_{0.025} < t < t_{0.025})]$$
$$= P(|\frac{b-\beta}{se}| < t_{0.025}) = P(b - t_{0.025}se < \beta < b + t_{0.025}se)$$

而 t 由 n，b，se 及 β 四个变量所决定。给定样本，随样本变化，b 和 se 会随之变化，而 β 为未知参数，但 β 落在区间 $(b-t_{0.025}se, b+t_{0.025}se)$ 的概率为 0.95。

大致意思是如果随机抽取样本容量相同（均为 n）的样本很多很多次，每次都计算出相应的 se,b，代入上式计算出许许多多的区间，则所有区间中约有 95%将包含总体参数 β，有 5%不包含 β。真值约有 95%次穿过区间，但约有 5%次在区间两个端点之外。

对某一次抽样来说，可信区间一旦形成，它要么包含总体参数，要么不包含总体参数，二者必居其一，无概率可言，因此所谓 95％的可信度是针对可信区间的构建方法而言的。

区间估计与点估计不同，它寻求一个区间，该区间以一定的概率保证真正的总体参数值包含在其中，当然，对于一个特定的样本，它可能包含参数真值，也可能不包含。

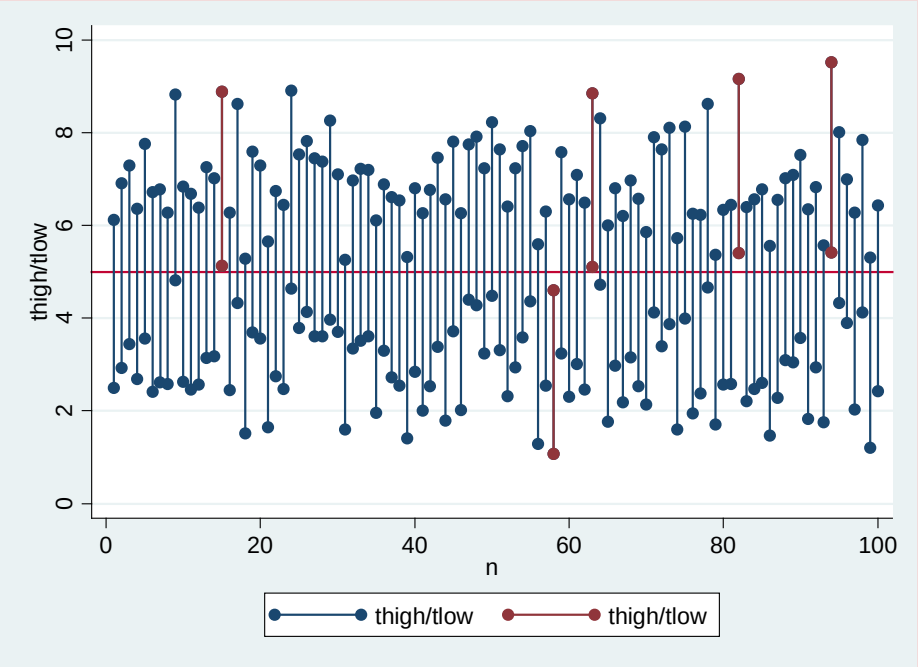
上机 5：区间估计

```
capt  prog  drop  bb
prog  bb
drawnorm  u,n(10)  d  clear
/*生成一个标准差 o=10 的正态随机变量样本，样本容量为 100*/
g  y=10+u
quietly  reg  y
end
***将上述抽样试验进行 100 次，得到 100 个样本均值 mean 和标准误
simulate  _b  _se, reps (100): bb
g  n=_n
*在总体方差未知的前提下，用样本标准差 sd 替代，需要借助 t 统计量
gen  tlow=_b-invttail(9,0.025)*_se
gen  thigh=_b+invttail(9,0.025)*_se
*考察总体均值是否在子样本的 95%置信区间内，如不在则标记为 1，否则为零
gen  tsign=(tlow<10 & thigh>10)
*统计没有包括总体均值的子样本 95%置信区间个数
```

tab tsign

***图示**

tw rcapsym thigh tlow n, yline(10) || rcapsym thigh tlow n if thigh<10 | tlow>10



在通常的研究中，我们只进行一次抽样，只构造出一个区间，并推测这一个区间有 95%的可能属于包含总体参数的区间簇，有 5%的可能属于不包含总体参数的区间簇。

八、假设检验

真正的总体参数 β 是一个常数，但具体等于多少，却是未知的。我们假设总体参数等于一个值 β_0=8，这个值是我们假设出来的，它也是一个常数。由于不知道 β 的取值，我们用猜测出来的 β_0 替代 β，于是有原假设（H_0：$\beta=\beta_0$），假设值 β_0 可能正好等于原总体的参数值 β，也可能不等。想一想，你能一次性地准确猜测出真正的总体值吗？另外，注意到在原假设与对立假设中，并不涉及到估计量。

利用估计量 b（随机变量）和假设值 β_0 构造一个 T 估计量（随机变量），这个 T 估计量小于临界值 t_α 的概率为

$$\gamma = P(T < t_\alpha) = P(\frac{b-\beta_0}{se} < t_\alpha) = P(\frac{b-\beta+\beta-\beta_0}{se} < t_\alpha) = P(\frac{b-\beta}{se} < t_\alpha - \frac{\beta-\beta_0}{se})$$

注意：上式中真正服从 t 分布的不是$\frac{b-\beta_0}{se}$而是$\frac{b-\beta}{se}$。

令$\lambda=\frac{\beta-\beta_0}{se}$,则有

$\gamma=t_{n-1}(t_\alpha-\lambda)$

如果原假设恰好成立，也即当原假设为真（$\beta=\beta_0$）时，有 $\lambda=0$，于是

$\gamma=t_{n-1}(t_\alpha)$

γ 是随机变量$\frac{b-\beta_0}{se}$落在t_α左边的概率，由于临界值t_α意味着其左边的面积为$1-\alpha$，故

$\gamma\geq 1-\alpha$

当α取值较小时（通常为 0.1、0.05 或 0.01）,意味着随机变量$\frac{b-\beta_0}{se}$出现在t_α右边的概率就很小。

当我们抽取一个特定的样本，计算后得到一个估计值 b^*（注意区别 β，β_0，b，b^*），这个估计值 b^*是估计量 b（为随机变量）的一个实现，是可以计算出具体取值的，如果$\frac{b^*-\beta_0}{se}$出现在t_α右边，意味着在一次取样中，不太可能出现的小概率事件出现了，于是我们倾向于认为原假设不对，拒绝（H_0：$\beta=\beta_0$），也就是认为$\beta\neq\beta_0$。

即使我们的假设是正确的，即 β 确实等于 β_0，但因为我们只抽得了一个样本，并利用这个样本计算出 T 值，这个 T 值有α的可能出现在t_α的右边。但我们却认这是一个小概率事件而拒绝原假设，认为 $\beta\neq\beta_0$，这一拒绝是错误的选

择，错误缘于抽样的偏误，使我们可能恰好在一次抽样中得到一个过大的 T 值，从而否定正确的原假设，这种错误叫做弃真错误，但是在原假设为真的前提下，发生这种错误的可能性只有 5%。

$$当\beta=\beta_0时, P(\frac{b-\beta}{se}>t_\alpha)=P(\frac{b-\beta_0}{se}>t_\alpha)=\alpha$$

在 STATA 统计软件中，默认的 $\beta_0=0$,根据特定样本计算出来的 T 值为

$$T^*=\frac{b^*-0}{se^*}=\frac{b^*}{se^*}$$

其中的“*”号表示根据某一个被抽取的样本计算得到的估计值。

以这个 T^*值为临界点，服从 $t_{(n-1)}$ 分布的随机变量 T 落入两端的概率称为 P 值，即

$$Pvalue=P(|\frac{b-\beta}{se}|>|T^*|)=P(|\frac{b-\beta}{se}|>|\frac{b^*-0}{se^*}|)$$

上机 6：假设检验

情形 1：总体均值已知，为 $\beta=8$。但我们假装不知道，却做出了对总体均值正确的原假设，认为它等于 $\beta_0=10$，则抽样进行假设检验如下

```
drawnorm y,n(100) m(8) sds(10) d clear
*生成一个均值为8,标准差为10的正态随机变量，作为研究总体
quietly sum y
di "从样本计算t统计值为："(r(mean)-8)/(sqrt(100)*r(sd))
di "根据t统计量临界值为："as error invttail(99,0.025)
```

对这次实验，拒绝还是接受？

由于我们通常只取一次样，所以有可能碰巧得到的样本正好是导致我们拒绝真的原假设的样本。这时我们就会犯错误。然而，弃真错误的可能性比较小。在100次这样的抽样研究中，大概有5次左右。

将上述试验进行100次，统计一下有多少次拒绝，多少次接受？

```
capt    prog    drop    bb
prog  bb
drawnorm  u,n(10)  d  clear
g    y=10+u
quietly    reg    y
end
```

*****将上述抽样试验进行 100 次，得到 100 个样本均值 mean 和标准误**

```
Simulate    _b    _se, reps (100) : bb
gen    t=abs((_b-10)/_se)>invttail(9,0.025)
tab    t    //其中的 1 表示在 100 次中拒绝原假设的次数。
```

情形 2：总体均值已知为 10。但我们假装不知道，并做出了对总体均值错误的原假设，如认为它等于 5，则抽样进行假设检验如下

```
capt    prog    drop  bb
prog    bb
drawnorm    u,n(10)  d    clear
g    y=10+u
quietly    reg    y
end
```

*****将上述抽样试验进行 100 次，得到 100 个样本均值 mean 和标准误**

```
simulate    _b    _se, reps (100): bb
gen    t=abs((_b-5)/_se)>invttail(9,0.025)
tab    t
```

这时，我们 100 次地拒绝了原假设，认为原总体的均值不可能为 5。

显著性：你和朋友来进行横跨西伯利亚的越野车比赛，一个月后，你以一秒之差击败他，显然你不能吹嘘自己比他快。你可能受助于某些东西，或者只

是随机因素使然，别无其他。那一秒不够显著，没有办法据此得出什么结论。“自行车骑手 A 比 B 优秀，因为他平常吃菠菜，而 B 吃豆腐，所在 A 在 3000 里的比赛中比 B 快了 1 秒”。