统计模拟和实证分析介绍

秦永松

(广西师范大学 数学与统计学院, 广西 桂林 541006)

我们在此简单介绍基于模型(model-based)的统计方法的**统计模拟**和**实证分析**，统计模拟和实证分析不仅仅局限于基于模型的统计方法，但基于模型的统计方法是数理统计的最大分支。在假定已有数据满足模型的前提下，对模型参数（未知量）进行估计或者对模型涉及的统计假设进行检验，这就是数理统计的理论（简称统计理论）或者方法。统计模拟就是依据设定的统计模型，通过计算机模拟相应的数据，并依据统计理论的结果，计算模拟数据下的结果并由此检验统计方法的性能或者优劣，其道理完全等同于物理或者化学实验，不同的是物理和化学实验是在实验室通过实物材料进行操作（比如甲醇制作芳烃的实验），统计模拟是利用计算机软件(常用的有R 和Matlab)进行操作，所以统计模拟又可以称为**统计实验**；实证分析是在已有实际数据的情况下，利用统计理论结果，代入公式，给出实际数据下的结论。

1 基于正态逼近的方法

统计理论

定理1 (基于正态逼近方法) . 设总体$X \in R$ 的期望(设为μ)和方差均存在，$X_1, X_2, \cdots, X_n$ 为其简单随机样本，则μ 的（一个）矩估计为

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

且由中心极限定理知, 当$n \to \infty$ 时，有

$$\sqrt{n}(\hat{\mu} - \mu)/s \xrightarrow{d} N(0, 1),$$

其中, s^2 为样本方差，即$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. 由此得到$\mu$ 的置信水平为$1 - \alpha$ $(0 < \alpha < 1)$ 的渐近置信区间为

$$[\bar{X} - su_{1-\alpha/2}/\sqrt{n}, \bar{X} + su_{1-\alpha/2}/\sqrt{n}],$$

其中，$u_{1-\alpha/2}$ 为$N(0, 1)$ 的$1 - \frac{\alpha}{2}$ 分位数。

统计模拟

下面用统计实验证实上述理论。

1. 点估计$\hat{\mu}$ 的优劣

在模型条件“总体$X \in R$ 的期望(设为μ)和方差均存在，$X_1, X_2, \cdots, X_n$ 为其简单随机样本”下，独立重复生成$m(= 1, 000)$ 个随机样本：$X_{k1}, X_{k2}, \cdots, X_{kn}, k = 1, 2, \cdots, m$。有以下几个衡量估计优劣的标准：

a. Mean: $M = \frac{1}{m}\sum_{k=1}^{m} \bar{X}_k$，其中, $\bar{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ki}, k = 1, 2, \cdots, m$;

b. Empirical Bias: $M - \mu_0$, 其中，μ_0 为总体均值（取样前设定）；

c. Empirical 标准差(E-SD)：$\bar{X}_k, k=1,2,\cdots,m$ 的标准差，即E-SD= $\left\{\frac{1}{m-1}\sum_{k=1}^m(\bar{X}_k-M)^2\right\}^{1/2}$；

d. Empirical root mean square error (RMSE)：E-RMSE= $\left\{\frac{1}{m-1}\sum_{k=1}^m(\bar{X}_k-\mu_0)^2\right\}^{1/2}$。

R 程序举例：取$\mu_0=10$, 总体设为正态分布$N(\mu_0,2)$, 样本容量$n=100$, 重复次数$m=1000$:

$$
\begin{aligned}
&n=100\\
&nsim=1000\\
&m=1\\
&y=rep(0,nsim)\\
&for(m\ in\ 1:nsim)\{\\
&x=rnorm(n=100,mean=10,sd=sqrt(2))\\
&y[m]=sum(x)/n\\
&m=m+1\\
&\}\\
&Mean=sum(y)/nsim\\
&Bias=Mean-10\\
&SD=sqrt(sum(y-Mean)^2/(nsim-1))\\
&RMSE=sqrt(sum(y-10)^2/(nsim-1))\\
&Mean\\
&Bias\\
&SD\\
&RMSE
\end{aligned}
$$

Mean 9.999951，Bias -4.92973e-05，SD 6.23836e-15，RMSE 0.001559698

为了使模拟结果更有说服力，可以加大模拟量，比如，样本容量可以变化，总体均值可以变化，总体分布可以变化，等等。

2. 区间估计的优劣

覆盖概率(CP) 和区间长度（AL）两个指标：

a. 覆盖概率: CP=m 次模拟中真值μ_0 落在区间$[\bar{X}-su_{1-\alpha/2}/\sqrt{n},\bar{X}+su_{1-\alpha/2}/\sqrt{n}]$ 内的频率；

b. 区间（平均）长度：AL=m 个区间的平均长度= $2\left(\sum_{k=1}^m s_k u_{1-\alpha/2}/n\right)/m$，其中$s_k$ 为第k 次模拟的样本标准差。

R 程序举例：取$\mu_0=10$, 总体设为正态分布$N(\mu_0,2)$, $\alpha=0.05$, 样本容量$n=100$, 重复次数$m=1000$:

$$
\begin{aligned}
&n=100\\
&nsim=1000\\
&mu0=10\\
&m=1\\
&CP=0
\end{aligned}
$$

$AL = 0$

$for(m\, in\, 1 : nsim)\{$

$x = rnorm(n = 100, mean = mu0, sd = sqrt(2))$

$bar = sum(x)/n$

$t0 = sum((x - bar)^2)/(n - 1)$

$temp = sqrt(t0) * 1.96/sqrt(n)$

$if(bar - temp <= mu0\ \&\ mu0 <= bar + temp)CP = CP + 1$

$AL = AL + 2 * temp$

$m = m + 1$

$\}$

$CP = CP/nsim$

$AL = AL/nsim$

CP

AL

模拟结果如下：

CP 0.955，AL 0.5523

3. Q-Q 图

检验$\sqrt{n}(\hat{\mu} - \mu)/s \xrightarrow{d} N(0, 1)$结论的正确性，可用Q-Q图表示。

R 程序举例：取$\mu_0 = 10$, 总体设为正态分布$N(\mu_0, 2)$, 样本容量$n = 100$, 重复次数$m = 1000$:

$n = 100$

$nsim = 1000$

$mu0 = 10$

$m = 1$

$T = rep(0, nsim)$

$for(m\, in\, 1 : nsim)\{$

$x = rnorm(n = 100, mean = mu0, sd = sqrt(2))$

$bar = sum(x)/n$

$t0 = sum((x - bar)^2)/(n - 1)$

$T[m] = sqrt(n) * (bar - mu0)/sqrt(t0)$

$m = m + 1$

$\}$

$qqnorm(T); qqline(T, col = 2)$

2 基于经验似然方法（单总体情形）

统计理论

定理2 (基于经验似然方法) . 设总体$X \in R$ 的期望(设为μ)和方差均存在，$X_1, X_2, \cdots, X_n$ 为其简单随机样本，则μ 的经验似然比统计量为

$$\ell(\mu) = 2\sum_{i=1}^{n} \log\{1 + m \cdot (X_i - \mu)\},$$

其中，$m = m(\mu)$ 满足

$$\frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \mu}{1 + m(X_i - \mu)} = 0$$

且

$$\ell(\mu) \xrightarrow{d} \chi_1^2,$$

其中, χ_1^2 表示自由度为1 的服从卡方分布的随机变量。由此得到μ 的置信水平为$1 - \alpha$ $(0 < \alpha < 1)$ 的渐近置信区间为

$$\{\mu : \ell(\mu) \le \chi_1^2(1 - \alpha)\},$$

其中，$\chi_1^2(1 - \alpha)$ 为χ_1^2 的$1 - \alpha$ 分位数。

统计模拟

下面用统计实验证实上述理论。

在模型条件“总体$X \in R$ 的期望(设为μ)和方差均存在，$X_1, X_2, \cdots, X_n$ 为其简单随机样本”下，独立重复生成$m(= 1,000)$ 个随机样本：$X_{k1}, X_{k2}, \cdots, X_{kn}, k = 1, 2, \cdots, m$。

1. 区间估计的优劣

覆盖概率(CP) 和区间长度（AL）两个指标。

R 程序举例：取$\mu_0 = 10$, 总体设为正态分布$N(\mu_0, 2)$, $\alpha = 0.05$, 样本容量$n = 100$, 重复次数$m = 1000$。

注：程序中的$< -$ 表示赋值，等同于=。

```
lambda <- function(x, mu)
{
L = -1/max(x - mu)
R = -1/min(x - mu)
dif <- 1
tol <- 1e-08
while(dif > tol){
M <- (L + R)/2
glam <- sum((x - mu)/(1 + M * (x - mu)))
if(glam > 0)L <- M
```

$if(glam < 0)R < -M$

$dif < -abs(glam)$

$\}$

$return(M)$

$\}$

$nsim < -1000$

$a < -0.95$

$cut < -qchisq(a, 1)$

$mu0 = 10$

$f < -0$

$L < -0$

$U < -0$

$m = 1$

$for(m\ in\ 1 : nsim)\{$

$x < -rnorm(n = 100, mean = mu0, sd = sqrt(2))$

$tol < -1e - 08$

$t1 < -mean(x)$

$t2 < -max(x)$

$dif < -t2 - t1$

$while(dif > tol)\{$

$tau < -(t1 + t2)/2$

$M < -lambda(x, tau)$

$elratio < -2 * sum(log(1 + M * (x - tau)))$

$if(elratio > cut)t2 < -tau$

$if(elratio <= cut)t1 < -tau$

$dif < -t2 - t1$

$\}$

$UB < -(t1 + t2)/2$

$t1 < -mean(x)$

$t2 < -min(x)$

$dif < -t1 - t2$

$while(dif > tol)\{$

$tau < -(t1 + t2)/2$

$M < -lambda(x, tau)$

$elratio < -2 * sum(log(1 + M * (x - tau)))$

$if(elratio > cut)t2 < -tau$

$$
\begin{aligned}
&if(elratio <= cut)t1 < -tau\\
&dif < -t1 - t2\\
&\}\\
&LB < -(t1 + t2)/2\\
&el < -2 * sum(log(1 + lambda(x, mu0) * (x - mu0)))\\
&if(el <= cut)f = f + 1\\
&L = LB + L\\
&U = UB + U\\
&m = m + 1\\
&\}\\
&fL = L/nsim\\
&fU = U/nsim\\
&CP = f/nsim\\
&AL = fU - fL\\
&fL\\
&fU\\
&CP\\
&AL
\end{aligned}
\tag{1}
$$

模拟结果如下：

置信区间[9.72797, 10.28292]，CP 0.946，AL 0.554955

2. Q-Q 图

检验$\ell(\mu) \xrightarrow{d} \chi_1^2$结论的正确性，可用Q-Q图表示。

R 程序举例：取$\mu_0 = 10$, 总体设为正态分布$N(\mu_0, 2)$, $\alpha = 0.05$, 样本容量$n = 100$, 重复次数$m = 1000$。

$$
\begin{aligned}
&lambda < -function(x, mu)\\
&\{\\
&L < --1/max(x - mu)\\
&R < --1/min(x - mu)\\
&dif < -1\\
&tol < -1e - 08\\
&while(dif > tol)\{\\
&M < -(L + R)/2\\
&glam < -sum((x - mu)/(1 + M * (x - mu)))\\
&if(glam > 0)L < -M\\
&if(glam < 0)R < -M
\end{aligned}
$$

$dif < -abs(glam)$

$\}$

$return(M)$

$\}$

$n = 100$

$mu0 = 10$

$nsim < -1000$

$T = rep(0, nsim)$

$m = 1$

$for(m\ in\ 1 : nsim)\{$

$x = rnorm(n = 100, mean = mu0, sd = sqrt(2))$

$T[m] = 2 * sum(log(1 + lambda(x, mu0) * (x - mu0)))$

$m = m + 1$

$\}$

$qqplot(qchisq(ppoints(1000), df = 1), T,$

$main = expression(Q - Q\ plotfor\ \ \chi^2[nu == 1]))$

$qqline(T, distribution = function(p)qchisq(p, df = 1),$

$probs = c(0.1, 0.6), col = 2)$

3 基于经验似然方法（线性模型）

统计理论

定理3 (线性模型中的经验似然方法) . 设有线性模型

$$Y_i = X_i'\beta + \epsilon_i, 1 \le i \le n,$$

其中，$\beta \in R^k$, $\{\epsilon_i, 1 \le i \le n\}$ 为均值为0 且3 阶矩有限的独立随机变量序列, 则β 的经验似然比统计量为

$$\ell(\beta) = 2\sum_{i=1}^{n} \log\{1 + m'(\beta)X_i(Y_i - X_i'\beta)\},$$

其中，$m = m(\beta)$ 满足

$$\frac{1}{n}\sum_{i=1}^{n} \frac{X_i(Y_i - X_i'\beta)}{m'(\beta)X_i(Y_i - X_i'\beta)} = 0$$

则在一定的正则条件下，当$n \to \infty$ 时，有

$$\ell(\beta) \xrightarrow{d} \chi_k^2,$$

其中, χ_k^2 表示自由度为k 的服从卡方分布的随机变量。由此得到β 的置信水平为$1 - \alpha$ $(0 < \alpha < 1)$

的渐近置信区间为

$$\{\beta : \ell(\beta) \leq \chi_k^2(1-\alpha)\},$$

其中，$\chi_k^2(1-\alpha)$ 为χ_k^2 的$1-\alpha$ 分位数。

统计模拟

下面用统计实验证实上述理论。

1. 区间估计的优劣

覆盖概率(CP)

R 程序举例：取$\epsilon \sim U(-0.5, 0.5)$, $\alpha = 0.05$, 样本容量$n = 200$, 重复次数$m = 1000$。

注：程序中的$<-$ 表示赋值，等同于=。

```
lambda <- function(u){
M <- c(0,0)
dif <- 1
tol <- 1e-8
k <- 0
while(dif > tol & k <= 15){
D1 <- c(0,0)
DD <- D1%*%t(D1)
for(i in 1:n){
aa <- as.numeric(1+t(M)%*%u[i,])
D1 <- D1+u[i,]/aa
DD <- DD-(u[i,]%*%t(u[i,]))/aa^2
}
D2 <- solve(DD,D1,tol=1e-8)
dif <- max(abs(D2))
rule <- 1
while(rule > 0){
rule <- 0
if(min(1+t(M-D2)%*%t(u)) <= 0)rule <- rule+1
if(rule > 0)D2 <- D2/2
}
M <- M-D2
k <- k+1}
if(k >= 15)M <- c(0,0)
return(M)
}
```

$nsim = 1000$

$a < -0.95$

$k < -2$

$n = 200$

$cut < -qchisq(a, k)$

$beta = c(0, 1)$

$f < -0$

$m = 1$

$for(m\; in\; 1 : nsim)\{$

$x0 < -rnorm(n)$

$x1 < -c(x0, x0^2)$

$x = matrix(x1, nrow = n, ncol = 2)$

$e = runif(n, -0.5, 0.5)$

$e = t(t(e))$

$y = x\% * \%beta + e$

$el = 0$

$z = matrix(rep(0, 2 * n), nrow = n, ncol = 2)$

$i = 1$

$for(i\; in\; 1 : n)\{$

$z[i,] = x[i,] * e[i]$

$i = i + 1$

$\}$

$lam = lambda(z)$

$for(i\; in\; 1 : n)\{$

$el = el + 2 * log(1 + t(lam)\% * \%x[i,] * e[i])$

$i = i + 1$

$\}$

$if(el <= cut)f = f + 1$

$m = m + 1$

$\}$

$CP = f/nsim$

CP

模拟结果如下：

CP 0.937

2. Q-Q 图

检验$\ell(\beta) \xrightarrow{d} \chi_k^2$结论的正确性，可用Q-Q图表示。

R 程序举例：取$\epsilon \sim U(-0.5, 0.5)$, 样本容量$n = 200$, 重复次数$m = 1000$。

$lambda <- function(u)\{$

$M <- c(0, 0)$

$dif <- 1$

$tol <- 1e - 8$

$k <- 0$

$while(dif > tol \; \& \; k <= 15)\{$

$D1 <- c(0, 0)$

$DD <- D1\% * \%t(D1)$

$for(i \; in \; 1 : n)\{$

$aa <- as.numeric(1 + t(M)\% * \%u[i,])$

$D1 <- D1 + u[i,]/aa$

$DD <- DD - (u[i,]\% * \%t(u[i,]))/aa^2$

$\}$

$D2 <- solve(DD, D1, tol = 1e - 8)$

$dif <- max(abs(D2))$

$rule <- 1$

$while(rule > 0)\{$

$rule <- 0$

$if(min(1 + t(M - D2)\% * \%t(u)) <= 0)rule <- rule + 1$

$if(rule > 0)D2 <- D2/2$

$\}$

$M <- M - D2$

$k <- k + 1\}$

$if(k >= 15)M <- c(0, 0)$

$return(M)$

$\}$

$nsim = 1000$

$k <- 2$

$n = 200$

$beta = c(0, 1)$

$T = rep(0, nsim)$

$m = 1$

$for(m \; in \; 1 : nsim)\{$

$x0 <- rnorm(n)$

$x1 <- c(x0, x0^2)$

$x = matrix(x1, nrow = n, ncol = 2)$

$e = runif(n, -0.5, 0.5)$

$e = t(t(e))$

$y = x\% * \%beta + e$

$el = 0$

$z = matrix(rep(0, 2 * n), nrow = n, ncol = 2)$

$i = 1$

$for(i\ in\ 1 : n)\{$

$z[i,] = x[i,] * e[i]$

$i = i + 1$

$\}$

$lam = lambda(z)$

$for(i\ in\ 1 : n)\{$

$el = el + 2 * log(1 + t(lam)\% * \%x[i,] * e[i])$

$i = i + 1$

$\}$

$T[m] = el$

$m = m + 1$

$\}$

$qqplot(qchisq(ppoints(1000), df = 2), T,$

$main = expression("Q - Qplotfor"\ \ chi^2[nu == 2]))$

$qqline(T, distribution = function(p)qchisq(p, df = 2),$

$probs = c(0.1, 0.6), col = 2)$