Bootstrap Intervel

統計模擬期末報告

統計碩一

110354029 陳槐廷

摘要:

拔靴法在資料來源未知的情況下，可用拔靴法去做估計，此篇用到 normal bootstrap，percentile bootstrap,t bootstrap 以上方法來對指數分配的最大概似估計量做信賴區間，並以覆蓋機率，平均信賴區間長度，信賴區間長度標準差，作為評判標準，有母數的效果較無母數佳，t bootstrap 有較好的效果。

研究動機:

拔靴法是一種統計分析的方法，可用來估計樣本統計量，常應用在估計和財務、風管等領域。拔靴法是 Efron 於 1979 年提出是一種重複抽樣的方法，早期是用來做變異數的估計，以求得因觀測值不足而無法得知資料的特性，將一組觀測資料$\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \ldots, \mathrm{x}_\mathrm{n})$當作母體，並做重複抽樣，藉以估計統計量的分配。

以拔靴法為基準來建置信賴區間，當樣本來自已知分配則為有母數拔靴法，若來自未知分配的話則為無母數拔靴法，並以不同的方法來取得信賴區間，並比較這幾個方法的優劣。

拔靴法介紹:

無母數拔靴法:

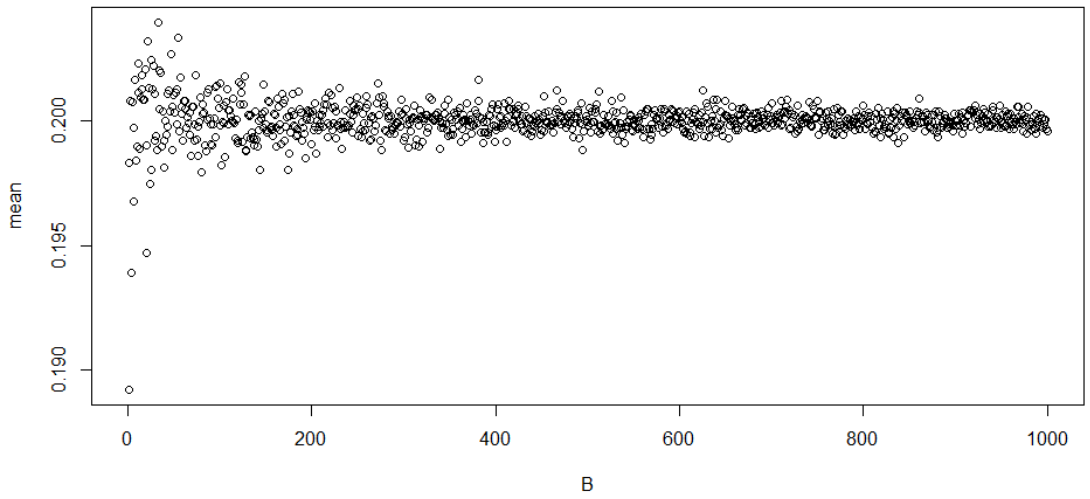
當觀測值$x = (x_1, x_2, x_3, \ldots, x_n)$來自未知母體時，由 x 中抽取 n 個樣本，且$(x_1, x_2, x_3, \ldots, x_n)$抽取的機率皆為 1/n，抽出的樣本為 bootstrap 樣本，因為要選取 B 組，所以重複以上 B 次，每一組的 bootstrap 樣本都可以得出一個估計值，並可以用估計值來得出信賴區間。

有母數拔靴法:

當觀測值$x = (x_1, x_2, x_3, \ldots, x_n)$來自已知母體時，因為母體已知，所以可以利用觀測值來估計母體的參數，並利用觀測值產生參數的模型來進行隨機抽取 n 個值，重複 B 次，且每一組產生的值都可以得到一個估計值也可以得出信賴區間。

B 值的選取:

我以 exp(5)的統計模型來看，建立 B 的 function，利用迴圈跑 1~1000 次作為 B 的組數 1~1000 組，每組內皆有隨機的 exp(5) 1000 個值，把生成 1~1000 組的 B 個別平均作為估計值，並用 plot 畫出 ，在 B 選擇較少組來估計時的結果會比較不準確，所以這裡選擇 B=1000 為組數。



信賴區間模擬

指數分配:

樣本來自$x = (x_1, x_2, x_3, \ldots, x_n)$服從指數分配，求得指數分配的概似函數，取對數並對$\boldsymbol{\beta}$進行微分令此為 0，將得出的$\hat{\beta}$帶入對概似函數二次微分小於 0，最後得出$\hat{\beta} = \bar{x}$。

生成隨機 30 個 exp(5)的值，並建置最大概似估計量的信賴區間和近似的信賴區間。因為$\sum_{i=1}^{30} x_i \sim \text{gamma}(30,5)$經由變數變換可知，$2 * 30 * \bar{x} \sim x^2(60)$，可得出最大概似函數的信賴區間為:

```
beta的95%信賴區間為[ 3.117792 , 6.415356 ]
```

因為$(\beta - \hat{\beta}) / \sqrt{\hat{\beta}/_{30}}$近似於 N(0,1)分配，可得最大概似估計量的近似信賴區間為:

```
beta的95%信賴區間為[ 2.779511 , 5.877317 ]
```

這裡會提出三個方法來建置最大概似估計量的信賴區間分別為 normal bootstrap、bootstrap percentile、bootstrap t，分別比較這幾個方法在無母數拔靴法和有母數拔靴法下的信賴區間並進行比較。

Bootstrap normal:

估計最大概似估計量的信賴區間，這裡採用$\hat{\beta}\sim N(\beta+\text{bias}(\hat{\beta}),\sigma)$近似的常態分配可得近似的信賴區間$[\hat{\beta}-\text{bias}(\hat{\beta})-z_{\alpha/2}*\hat{\sigma},\hat{\beta}-\text{bias}(\hat{\beta})+z_{\alpha/2}*\hat{\sigma}]$

無母數拔靴法

建立一個 beta 為空向量，並建立 B=1000 次的迴圈，因為為無母數拔靴法無法知道分配，一開始所得到$\hat{\beta}=\bar{x}$將每次產生的 30 個值取平均，並把誤差和標準差算出，將估計值帶入近似信賴區間中。

```
beta <- c()
for (i in 1:1000){
  z <- sample(x,30,T)
  beta[i] <- mean(z)
}
bias1 <- mean(beta)-beta_mle
sigma1 <- (sum((beta-mean(beta))^2)/999)^(1/2)
```

```
normal bootstrap的95%信賴區間為[ 2.764245 , 5.910663 ]
```

有母數拔靴法:

建立 beta_parametric 為空向量，因為為有母數拔靴法已知分配下，建立 B=1000 次迴圈，因為把觀測值當作母體，所以 exp 分配的參數為觀測值期望值的倒數，並從指數分配隨機抽取 30 個取期望值為最大概似估計量。

```r
beta_parametric <- c()
for (i in 1:1000){
  y <- rexp(30,1/mean(x))
  beta_parametric[i] <- mean(y)
}
bias2 <- mean(beta_parametric)-beta_mle
sigma2 <- (sum((beta_parametric-mean(beta_parametric))^2)/999)^(1/2)
hist(beta_parametric)
cat("normal bootstrap的95%信賴區間為[",beta_mle-bias2-1.96*sigma2,",",beta_mle-bias2+1.96*sigma2,"]")
```

```
normal bootstrap的95%信賴區間為[ 2.772776 , 5.786772 ]
```

Percentile bootstrap :

將由 B=1000 組所得到的 1000 個最大概似估計量的值由小到大進行排序，且他的下界為 100*α百分位，上界為 100*$(1-\alpha)$百分位。

無母數拔靴法:

令 beta 為空向量，並進行 B=1000 次迴圈，因為母體分配未知，所以將 $\hat{\beta} = \bar{x}$將每次產生的 30 個值取平均放進 beta 中，並將 beta 內的值用 sort 從小排到大，取出第 25 個和第 975 個值作為下界和上界。

```r
beta <- c()
for (i in 1:1000){
  z <- sample(x,30,T)
  beta[i] <- mean(z)
}
cat("bootstrap percentile的95%信賴區間為[",sort(beta)[25],",",sort(beta)[975],"]")
```

```
bootstrap percentile的95%信賴區間為[ 3.015049 , 6.137707 ]
```

有母數拔靴法:

令 beta_parametric 為空向量，進行 B=1000 次迴圈，並將觀測值取期望值的倒數作為指數分配的參數，並隨機抽出 30 個值取期望值放入

beta_parametric 中用 sort 從小排到大，取出第 25 個和第 975 個值作為下界和上界。

```
beta_parametric <- c()
for (i in 1:1000){
  y <- rexp(30,1/mean(x))
  beta_parametric[i] <- mean(y)
}
cat("bootstrap percentile_parametric的95%信賴區間為[",sort(beta_parametric)[25],",",sort(beta_parametric)[975],"]")
```

```
bootstrap percentile_parametric的95%信賴區間為[ 2.917322 , 6.005396 ]
```

Bootstrap-t :

生成 B 組的獨立樣本，並算出$\hat{\theta}(b)$為各組的估計量和$\hat{\sigma}(b)$，並求出各組的$\hat{t}(b) = \frac{\hat{\beta}(b)-\theta}{\hat{\sigma}(b)}$，得出 1000 個 t 值並以由小到大進行排序，求出 100*α百分位和 100*$(1-\alpha)$百分位。

Bootstrap-t 的 95%信賴區間為$[\hat{\beta} - \hat{t}^{97.5} * \hat{\sigma}(\hat{\beta}), \hat{\beta} - \hat{t}^{2.5} * \hat{\sigma}(\hat{\beta})]$

無母數拔靴法:

令 beta、sigma、t 為空向量，從 30 個觀測值以取出放回的方式抽 30 個，並取期望值作為最大概似估計量，而他的標準差則是以 MLE 不變性帶$(1/\hat{\beta}\text{^}2)^{1/2}$，並計算 1000 個 t 從小到大排列，並將 100*α百分位和 100*$(1-\alpha)$百分位帶入信賴區間中。

```r
beta <- c()
sigma <- c()
t <- c()
for (i in 1:1000){
  z <- sample(x,30,T)
  beta[i] <- mean(z)
  sigma[i] <- (beta[i]^2)^(1/2)
  t[i] <- (beta[i]-beta_mle)/sigma[i]
}
```

```
bootstrap t_nonparametric的95%信賴區間為[ 3.161974 , 5.040808 ]
```

有母數拔靴法:

令 beta_parametric、sigma3、t1 為空向量，並將觀測值取期望值的倒數作為指數分配的參數，beta_parametric 為每組的最大概似估計量，標準差則是以 MLE 不變性帶$(1/\hat{\beta}^2)^{1/2}$，並得出 1000 個 t 從校排到大，將所得出來的值帶入信賴區間中。

```r
beta_parametric <- c()
sigma3 <- c()
t1 <- c()
for (i in 1:1000){
  y <- rexp(30,1/mean(x))
  beta_parametric[i] <- mean(y)
  sigma3[i] <- (mean(y)^2)^(1/2)
  t1[i] <- (beta_parametric[i]-beta_mle)/sigma3[i]
}
```

```
bootstrap t_nonparametric的95%信賴區間為[ 3.163708 , 6.362281 ]
```

統整以上結果:

	下界	上界	長度
MLE	3.799868	7.818836	4.018968
MLE 近似	3.387581	7.163091	3.77551

無母數:

方法	下界	上界	長度
normal bootstrap	2.764245	5.910663	3.146418
bootstrap percentile	3.015049	6.137707	3.122658
bootstrap t	3.161974	5.040808	1.878834

有母數:

方法	下界	上界	長度
normal bootstrap	2.772776	5.786772	3.013996
bootstrap percentile	2.917322	6.005396	3.088074
bootstrap t	3.152391	6.362281	3.20989

方法比較:

利用覆蓋機率、信賴區間的平均長度和長度的標準差來判斷，覆蓋機率為模擬 1000 次真實參數有落在信賴區間的次數，平均長度為有覆蓋到真實參數的信賴區間的長度除上有覆蓋到的次數。

	覆蓋率	**平均長度**	**長度標準差**
MLE	0.949	3.765016	0.5992178
MLE 近似	0.927	3.669919	0.5808076

無母數:

方法	**覆蓋率**	**平均長度**	**長度標準差**
normal bootstrap	0.605	2.304894	0.04478616
bootstrap percentile	0.884	2.275087	0.05894485
bootstrap t	1	2.387534	0.08142516

有母數:

方法	覆蓋率	平均長度	長度標準差
normal bootstrap	1	2.764641	0.06546492
bootstrap percentile	1	2.745991	0.08368389
bootstrap t	1	2.936051	0.1061469

結論:

分別對有母數和無母數拔靴法去建置 normal bootstrap、bootstrap percentile、bootstrap t 的信賴區間，由 B 的選取圖來看分的組數越多結果越好，但越多電腦計算時間長，越沒效率。在有母數拔靴法我認為是分配的選取導致真實參數百分之百落在信賴區間中，但有母數還是優於無母數的，以平均長度來看 bootstrap t 的平均長度比其他兩種方式長，覆蓋率也比較高。所以用 bootstrap t 來建置指數分配的信賴區間是不錯的。

文獻探討:

Efron, B., Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman & Hall, New York.

周心怡(2004). 拔靴法(Bootstrap)之探討及其應用