

## الگوریتم بوت استرپ بیزی

نصراله ایران‌پناه<sup>۱</sup>

### چکیده

فرض کنید  $X = (X_1, \dots, X_n)$  یک نمونه تصادفی مستقل با توزیعهای یکسان (iid) از توزیع نامعلوم  $F$  با پارامتر  $\theta$  باشد. برآوردکننده  $\hat{\theta} = \hat{\theta}(X)$  را برای  $\theta$  و  $R(X, F)$  به صورت تابعی از  $X$  و  $F$  مانند  $R(X, F) = \hat{\theta} - \theta$  در نظر می‌گیریم. افران روش بوت‌استرپ را با استفاده از طرح نمونه‌گیری مجدد از مشاهدات برای برآورد توزیع نمونه‌گیری  $R(X, F)$  و مشخصه‌های آن مانند میانگین و واریانس ارایه کرد.

در استنباط بیزی فرض می‌شود  $\theta$  یک متغیر تصادفی است که دارای یک توزیع پیشین است و هدف پیدا کردن توزیع پسین  $R(X, F)$  است. رویین روش بوت‌استرپ بیزی ناپارامتری را برای تقریب توزیع پسین  $R(X, F)$  ارایه کرده است [۴].

در این مقاله، دوروش بوت‌استرپ افران و بوت‌استرپ بیزی رویین ارایه و مقایسه می‌شود. الگوریتم دوروش در یک مثال تشریح و برنامه رایانه‌ای با استفاده از نرم‌افزار آماری S-PLUS ارایه می‌شود.

### ۱ الگوریتم بوت‌استرپ افران (BE)

$F$  با پارامتر  $\theta$  باشد. برآوردکننده  $\hat{\theta} = \hat{\theta}(X)$  را برای  $\theta$  و  $R(X, F)$  را به صورت تابعی از  $X$  و  $F$  مانند  $R(X, F) = \hat{\theta} - \theta$  در نظر می‌گیریم. هدف، تقریب توزیع نمونه‌گیری  $R(X, F)$  و

افران در سال ۱۹۷۹ روشی را به نام بوت‌استرپ برای تقریب توزیع نمونه‌گیری و اندازه‌های دقت (مانند اریبی و واریانس) برآوردکننده‌ها ارایه کرد [۲]. فرض کنید  $X = (X_1, \dots, X_n)$  یک نمونه تصادفی (iid) از توزیع نامعلوم

<sup>۱</sup> نصراله ایران‌پناه، گروه آمار، دانشکده علوم، دانشگاه اصفهان

۱۹۸۱ یک روش بوت استرپ بیزی ناپارامتری را در حالتی که توزیع پیشین نامعلوم است، ارائه کرد. الگوریتم BBR در سه مرحله زیر خلاصه می شود:

مرحله ۱ - فرض کنید  $U_1, \dots, U_{n-1}$  متغیرهای تصادفی (iid) از  $U(0, 1)$  باشند. آماره های مرتب  $0 = U_{(0)} < U_{(1)} < \dots < U_{(n-1)} < U_{(n)} = 1$  فواصل  $g_i = U_{(i)} - U_{(i-1)}$ ;  $(i = 1, \dots, n)$  را در نظر می گیریم. اگر  $D_n$  تابع توزیع تصادفی با وزنهای  $g_i$  برای هر  $X_i$  به صورت زیر باشد:

$$D_n(X) = \sum_{i=1}^n g_i I_{X_i \leq x}$$

آنگاه با محاسبه  $\theta = \theta(D_n)$  می توان از توزیع شرطی  $R(X, D_n) = \hat{\theta} - \bar{\theta}$  (به شرط  $X$ ) به عنوان تقریب توزیع پسین  $R(X, F)$  استفاده کرد. توزیع  $R(X, D_n)$  جنبه ی نظری دارد و می توان آن را از روش مونت کارلو تقریب زد.

مرحله ۲ - در روش مونت کارلو، مرحله ۱ را به طور مستقل  $B$  بار تکرار کرده،  $D_{nb}$  و از روی آن  $\bar{\theta}^b$  و سپس  $R_b = R(X, D_{nb}) = \hat{\theta} - \bar{\theta}^b$  را  $(b = 1, \dots, B)$  محاسبه می کنیم.

مرحله ۳ - تابع توزیع تجربی  $R_1, \dots, R_B$  (که جرم احتمال  $1/B$  را به هر  $R_b$  نسبت می دهد) توزیع  $R(X, D_n)$  را برای  $B$  های بزرگ تقریب می زند.

مشخصه های آن، مانند میانگین و واریانس است. الگوریتم BE در سه مرحله زیر خلاصه می شود:

مرحله ۱ - از تابع توزیع تجربی  $F_n$  یک نمونه تصادفی (iid) به صورت  $X^* = (X_1^*, \dots, X_n^*)$  تولید می کنیم. به عبارت ساده تر نمونه بوت استرپ  $X^*$  را به روش نمونه گیری تصادفی ساده با جایگذاری از  $X$  به دست می آوریم. با محاسبه آماره بوت استرپ  $\hat{\theta}^* = \hat{\theta}(X^*)$  می توان از توزیع  $R(X^*, F_n) = \hat{\theta}^* - \hat{\theta}$  به عنوان تقریب توزیع نمونه گیری  $R(X, F)$  استفاده کرد. توزیع  $R(X^*, F_n)$  جنبه ی نظری دارد و می توان آن را از روش مونت کارلو تقریب زد.

مرحله ۲ - در روش مونت کارلو، نمونه گیری از  $F_n$  را به طور مستقل  $B$  بار تکرار کرده، نمونه بوت استرپ  $X^{*b}$  و از روی آن آماره بوت استرپ  $\hat{\theta}^{*b} = \hat{\theta}(X^{*b})$  و سپس  $R_b = R(X^{*b}, F_n) = \hat{\theta}^{*b} - \hat{\theta}$  را  $(b = 1, \dots, B)$  محاسبه می کنیم.

مرحله ۳ - تابع توزیع تجربی  $R_1, \dots, R_B$  (که جرم احتمال  $1/B$  را به هر  $R_b$  نسبت می دهد)، توزیع  $R(X^*, F_n)$  را برای  $B$  های بزرگ تقریب می زند.

## ۲ الگوریتم بوت استرپ بیزی روبین (BBR)

در استنباط بیزی فرض می شود  $\theta$  یک متغیر تصادفی است که دارای یک توزیع پیشین است و  $X = (X_1, \dots, X_n)$  یک نمونه تصادفی (iid) از  $F$  (توزیع شرطی  $X$  به شرط  $\theta$ ) است. استنباط بیزی بر اساس توزیع پسین  $R(X, F)$  یعنی توزیع شرطی  $R(X, F)$  به شرط  $X$  است. اگر  $F$  متعلق به یک خانواده پارامتری با پارامتر  $\theta$  باشد، توزیع پسین  $R(X, F)$  در صورتی قابل محاسبه است که توزیع پیشین معلوم باشد. روبین در سال

## ۳ مقایسه BE و BBR

دو روش BE و BBR را برای پارامتر  $\theta$ ، میانگین توزیع نامعلوم  $F$  یعنی  $\theta = \mu = \int x dF(x)$  مقایسه می کنیم. نمونه تصادفی مشاهده شده  $x_1, \dots, x_n$  و برآورد  $\bar{x} = \sum x_i/n$  را برای  $\theta$

در نظر می‌گیریم.

با به دست آوردن تمام تکرارهای مرحله ۱ (که در عمل کاری غیر ممکن است) و سپس محاسبه  $\bar{y}_i$  توزیع BBR برای  $n$  به دست می‌آید که به عنوان تقریب توزیع پسین  $\pi$  می‌تواند استفاده شود. از روش مونت کارلو برای تقریب توزیع بوت‌استرپ بیزی  $\bar{\mu}$  استفاده می‌کنیم. اگر مرحله ۱ الگوریتم BBR را به طور مستقل  $B$  بار تکرار کنیم، در نتیجه  $\bar{\mu}^1, \dots, \bar{\mu}^B$  به دست می‌آید (مرحله ۲). تابع توزیع  $\bar{\mu}^1, \dots, \bar{\mu}^B$  می‌تواند به عنوان تقریب توزیع بوت‌استرپ بیزی  $\bar{\mu}$  و در نتیجه تقریب توزیع پسین  $\pi$  استفاده شود (مرحله ۳).

الگوریتم دوروش BE و BBR فقط در تخصیص احتمال  $f_i$  و  $g_i$  به هر  $x_i$  باهم اختلاف دارند. در روش BE توزیع نمونه‌گیری برآوردکننده  $\theta$  شبیه‌سازی و تقریب زده می‌شود، در صورتی که در روش BBR توزیع پسین پارامتر  $\theta$  شبیه‌سازی و تقریب زده می‌شود.

به سادگی می‌توان نشان داد که دو بردار  $f = (f_1, \dots, f_n)$  و  $g = (g_1, \dots, g_n)$  دارای توزیع  $n$  جمله‌ای  $(\frac{1}{n}, \dots, \frac{1}{n})$  و دیریکله  $(1, \dots, 1)$  هستند. و همچنین،

$$E(f_i) = E(g_i) = \frac{1}{n};$$

$$var(f_i) = \frac{n-1}{n^2}; \quad var(g_i) = \frac{n-1}{n^2(n+1)};$$

$$cov(f_i, f_j) = cov(g_i, g_j) = \frac{-1}{n^2}.$$

#### ۴ یک مثال دو بعدی

در این بخش دو الگوریتم BE و BBR را در حالتی که  $x$  یک متغیر دو بعدی  $(y, z)$  و پارامتر  $\theta$  ضریب همبستگی جامعه  $(\rho)$  است، همراه با یک مثال کاربردی (۱)، (صفحه ۴۳) ارائه می‌کنیم.

از بین دانشجویان رشته الکترونیک دانشگاه صنعتی شریف، برای مطالعه نمرات درک مطلب ( $y$ ) و دستور زبان ( $z$ ) درس زبان انگلیسی، نمونه‌ای تصادفی به حجم  $n = 15$

در روش BE، نمونه بوت‌استرپ  $x_1^*, \dots, x_n^*$  را به روش نمونه‌گیری تصادفی ساده با جایگذاری از  $x_1, \dots, x_n$  به دست می‌آوریم (مرحله ۱). فرض کنید برای هر  $i = 1, 2, \dots, n$ ،  $n_i$  فراوانی و  $f_i = n_i/n$  فراوانی نسبی  $x_i$  باشد که در نمونه بوت‌استرپ ظاهر شده است و می‌تواند مقادیر  $0, 1, \dots, n$  را اختیار کند. واضح است که  $n_i$  دارای توزیع  $b(1, n)$  است. میانگین نمونه بوت‌استرپ  $\hat{\theta}^* = \bar{x}^* = \sum x_i^*/n = \sum f_i x_i$  را محاسبه می‌کنیم. با به دست آوردن تمام نمونه‌های ممکن بوت‌استرپ که تعداد آنها  $m = n^n$  است و محاسبه میانگین نمونه بوت‌استرپ روی آنها،  $\bar{x}^{*1}, \dots, \bar{x}^{*m}$  به دست می‌آید. توزیع نمونه‌گیری  $\bar{x}$  می‌تواند با استفاده از توزیع تجربی  $\bar{x}^{*1}, \dots, \bar{x}^{*m}$  تقریب زده شود. واضح است که به دلیل بزرگ بودن  $m$ ، انجام این روش حتی با رایانه نیز کاری مشکل است. در نتیجه از روش مونت کارلو برای تقریب بوت‌استرپ  $\bar{x}^*$  استفاده می‌کنیم. نمونه‌گیری بوت‌استرپ در مرحله ۱ را به طور مستقل  $B$  بار تکرار کرده و  $\bar{x}^{*1}, \dots, \bar{x}^{*B}$  را محاسبه می‌کنیم (مرحله ۲). تابع توزیع تجربی  $\bar{x}^{*1}, \dots, \bar{x}^{*B}$  می‌تواند به عنوان تقریب توزیع بوت‌استرپ  $\bar{x}^*$  و در نتیجه تقریب توزیع نمونه‌گیری  $\bar{x}$  برای  $B$  های بزرگ ( $200 \leq B \leq 1000$ ) استفاده شود (مرحله ۳).

نکته مهم در روش BE این است که در تقریب توزیع نمونه‌گیری  $\bar{x}$ ، فرض شده که تابع توزیع نمونه همان تابع توزیع جامعه است. به عبارت دیگر، نمونه‌های بوت‌استرپ به طور مستقل از تابع توزیع نمونه (یعنی  $F_n$ ) تولید شده‌اند.

در روش BBR بردار  $g = (g_1, \dots, g_n)$  را مشابه مرحله ۱ الگوریتم BBR به دست می‌آوریم (مرحله ۱). هر احتمال  $g_i$  به صورت یک احتمال پسین برای  $x_i$  است. واضح است که  $g_i$  دارای توزیع  $beta(1, n-1)$  است. با فرض  $\theta = \mu$  هر تکرار مرحله ۱، یک میانگین به صورت  $\mu = \sum g_i x_i$  تولید می‌کند.

انتخاب و در شکل ۱ نشان داده شده است. ضریب همبستگی نمونه  $r = 0/662$  است.

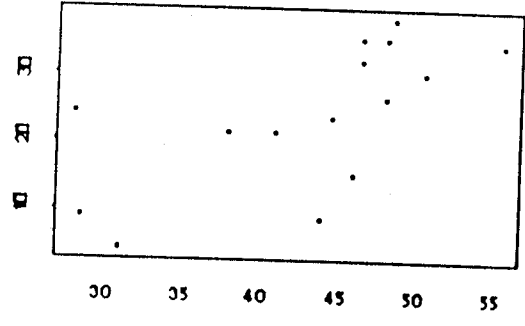
همبستگی  $r$  را به صورت زیر به دست می آوریم:

$$bias^*(r) = \frac{1}{1000} \sum_{b=1}^{1000} r^{*b} - r$$

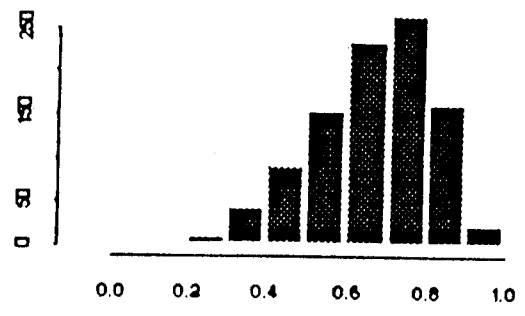
$$= 0/662 - 0/662 = 0/001$$

$$SE^*(r) = \frac{1}{999} \sum_{b=1}^{1000} (r^{*b} - \frac{1}{1000} \sum_{b=1}^{1000} r^{*b})^2 \frac{1}{999}$$

$$= 0/152$$



شکل ۱



شکل ۲

• الگوریتم BE را در این مثال برای تقریب توزیع نمونه گیری  $r$  و مشخصه های آن در سه مرحله زیر ارایه می کنیم:

مرحله ۱ - نمونه تصادفی مشاهده شده را به صورت زوجی  $x_i = (y_i, z_i); i = 1, \dots, 15$  در نظر می گیریم. از نمونه زوجی  $(x_1, \dots, x_{15})$  به روش نمونه گیری تصادفی ساده با جایگذاری، نمونه زوجی بوت استرپ  $(x_1^*, \dots, x_{15}^*)$  را به دست می آوریم.

مرحله ۲ - ضریب همبستگی نمونه زوجی بوت استرپ  $r^*$  را محاسبه می کنیم.

مرحله ۳ - مراحل ۱ و ۲ را ۱۰۰۰ بار تکرار کرده و ۱۰۰۰ ضریب همبستگی بوت استرپ  $(r^{*1}, \dots, r^{*1000})$  را به دست می آوریم. شکل ۲ بافت نگار بوت استرپ ضریب همبستگی حاصل از ۱۰۰۰ تکرار مرحله ۳ را نشان می دهد که می تواند به عنوان تقریب توزیع نمونه گیری  $r$  استفاده شود. از مشخصه های توزیع بوت استرپ استفاده کرده، برآورد بوت استرپ ارببی و خطای معیار ضریب

• همچنین الگوریتم BBR را در این مثال برای تقریب توزیع پسین  $\rho$  و مشخصه های آن در سه مرحله زیر ارایه می کنیم:

مرحله ۱ - بردار تصادفی مشاهده شده  $g = (g_1, \dots, g_n)$  را مشابه با مرحله ۱ الگوریتم BBR به دست می آوریم.

مرحله ۲ - ضریب همبستگی پسین  $\bar{\rho}$  را با استفاده از بردار مرحله ۱، به صورت زیر محاسبه می کنیم:

$$\bar{\rho} = \frac{\sum g_i y_i z_i - (\sum g_i y_i)(\sum g_i x_i)}{\{[\sum g_i y_i^2 - (\sum g_i y_i)^2][\sum g_i z_i^2 - (\sum g_i z_i)^2]\}^{\frac{1}{2}}}$$

مرحله ۳ - مراحل ۱ و ۲ را ۱۰۰۰ بار تکرار کرده و ۱۰۰۰ ضریب همبستگی پسین  $(\rho^1, \dots, \rho^{1000})$  را به دست می آوریم. شکل ۳ بافت نگار توزیع پسین BBR ضریب همبستگی حاصل از ۱۰۰۰

قابل اجرا است. الگوریتم BBR با استفاده از توابع gcor و grubin قابل اجراست که تابع gcor مقدار ضریب همبستگی پسین  $\bar{\rho}$  را محاسبه می کند. در این توابع  $x$ ، نمونه تصادفی زوجی،  $n$  حجم نمونه و  $b$  تعداد تکرارهای بوت استرپ است. خروجی هر دو تابع bootcor و grubin تکرارهای بوت استرپ  $r^*$  و  $\bar{\rho}^*$  است.

تکرار مرحله ۳ را نشان می دهد که می تواند به عنوان تقریب توزیع نمونه گیری پسین  $\rho$  استفاده شود. از مشخصه های توزیع پسین BBR استفاده کرده، برآورد بوت استرپ اریبی و خطای معیار ضریب همبستگی پسین  $\rho$  را به صورت زیر به دست می آوریم:

$$bias^*(\rho) = \frac{1}{1000} \sum_{b=1}^{1000} \bar{\rho}^b - r$$

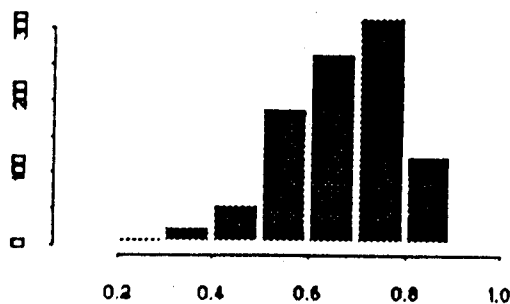
$$= 0/663 - 0/663 = 0$$

$$SE^*(\rho) = \frac{1}{999} \sum_{b=1}^{1000} (\bar{\rho}^b - \frac{1}{1000} \sum_{b=1}^{1000} \bar{\rho}^b)^2$$

$$= 0/131$$

```
bootcor<-function(n,b,x){
  istar<-matrix(sample(1:n,size=n*b,replace=T)
                ,nrow=b)
  cc<-c(1:b)
  for(i in 1:b){
    xstar<-matrix(c(x[istar[i,1]],x[istar[i,],2])
                 ,ncol=2)
    cc[i]<-cor(xstar[,1],xstar[,2])
  }
  return(cc)}
```

```
gcor<-function(g,x){
  y<-sum(g*x[,1])
  z<-sum(g*x[,2])
  yz<-sum(g*x[,1]*x[,2])
  y2<-sum(g*x[,1]*x[,1])
  z2<-sum(g*x[,2]*x[,2])
  rg<-(yz-(y*z))/(sqrt((y2-(y*y))*(z2-(z*z))))
  return(rg)}
```



شکل ۳

```
grubin<-function(n,b,x){
  rg<-c(1:b)
  g<-c(1:b)
  for(i in 1:b){
    u<-c(0,sort(runif(n-1),1))
    for(j in 1:n){
      g[j]<-(u[j+1]-u[j])
    }
    rg[i]<-gcor(g,x)
  }
  return(rg)}
```

## ۵ برنامه های رایانه ای

نرم افزار S-PLUS یکی از نرم افزارهای جدید و توانای آماری است که امکان برنامه نویسی را نیز دارد. مثال بخش ۴ با استفاده از برنامه های نوشته شده توسط این نرم افزار حل شده است. الگوریتم BE برای پارامتر  $\rho$  با استفاده از تابع bootcor

## مراجع

- [۱] ایران‌پناه، نصراله و پاشا، عین‌اله، آشنایی با الگوریتم بوت استرپ، مجله اندیشه آماری، سال دوم، شماره ۱، فروردین ۱۳۷۶.
- [2] Efron, B. (1979), *Bootstrap methods; another look at the Jackknife*, Ann. Statist., 7, 1-26.
- [3] Lo, A., Y. (1987), *A Large sample study of the Bayesian Bootstrap*, Ann. Statist., 15, 360-375.
- [4] Rubin, D. B. (1981), *The Bayesian Bootstrap*, Ann. Statist., 9, 130-134.
- [5] Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag.

- هر پدیده‌ای که اتفاق افتد، مشاهداتی حاصل می‌شود که اگر آنها را ثبت نکنیم، هیچگونه‌ارزشی ندارند.

«آمار»

- داده‌های جمع‌آوری شده را اگر خلاصه و گویا نکنیم، ارزش آنها ناچیز می‌ماند.

«آمار توصیفی»

- مطالعات انجام شده روی داده‌ها را اگر مورد تجزیه و تحلیل قرار ندهیم و تفسیر نکنیم، ارزش واقعی آنها را از دست می‌دهیم.

«آمار استنباطی»

- اگر نتایج به دست آمده از تجزیه و تحلیل‌های آماری را در برنامه‌ریزیها به کار نبریم، همه‌ی تلاشهای انجام گرفته بی‌ارزش، و اتلاف وقت بوده‌است.

«آمار کاربردی»