

الگوریتم بوت استرپ بیزی

نصراله ایران‌پناه^۱

چکیده

فرض کنید $X = (X_1, \dots, X_n)$ یک نمونه تصادفی مستقل با توزیعهای یکسان (iid) از توزیع نامعلوم F با پارامتر θ باشد. برآوردکننده‌ی $\hat{\theta} = \hat{\theta}(X)$ را برای θ 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 θ 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، دو روش بوت‌استرپ افران و بوت‌استرپ بیزی رویین ارایه و مقایسه می‌شود. الگوریتم دو روش در یک مثال تشریح و برنامه رایانه‌ای با استفاده از نرم‌افزار آماری S-PLUS ارایه می‌شود.

۱ الگوریتم بوت‌استرپ افران (BE)

F با پارامتر θ باشد. برآوردکننده‌ی $\hat{\theta} = \hat{\theta}(X)$ را برای θ 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

^۱ نصراله ایران‌پناه، گروه آمار، دانشکده علوم، دانشگاه اصفهان

مشخصه‌های آن، مانند میانگین و واریانس است. الگوریتم BE در سه مرحله زیر خلاصه می‌شود:

مرحله ۱ - از تابع توزیع تجربی F_n یک نمونه تصادفی (iid) به صورت $X^* = (X_1^*, \dots, X_n^*)$ تولید می‌کنیم. به عبارت ساده‌تر نمونه بوت‌استرپ X^* را به روش نمونه‌گیری تصادفی ساده با جایگذاری از X به دست می‌آوریم. با محاسبه آماره بوت‌استرپ $\hat{\theta}^* = \hat{\theta}(X^*)$ می‌توان از توزیع $R(X^*, F_n) = \hat{\theta}^* - \hat{\theta}$ به عنوان تقریب توزیع نمونه‌گیری $R(X, F)$ استفاده کرد. توزیع $R(X^*, F_n)$ جنبه‌ی نظری دارد و می‌توان آن را از روش مونت کارلو تقریب زد.

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

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

۲ الگوریتم بوت‌استرپ بیزی روبین (BBR)

در استنباط بیزی فرض می‌شود θ 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 θ) 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 θ باشد، توزیع پسین $R(X, F)$ در صورتی قابل محاسبه است که توزیع پیشین معلوم باشد. روبین در سال

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

مرحله ۱ - فرض کنید $U_1, \dots, U_{n-1}$ متغیرهای تصادفی (iid) از $U(0, 1)$ باشند. آماره‌های 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 می‌گیریم. 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 محاسبه‌ی $\bar{\theta} = \theta(D_n)$ می‌توان از توزیع شرطی $R(X, D_n) = \bar{\theta} - \hat{\theta}$ (به شرط X) به عنوان تقریب توزیع پسین $R(X, F)$ استفاده کرد. توزیع $R(X, D_n)$ جنبه‌ی نظری دارد و می‌توان 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 $R_b = R(X, D_{nb}) = \bar{\theta}^b - \bar{\theta}$ را محاسبه می‌کنیم.

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

۳ مقایسه BE و BBR

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

در نظر می گیریم.

در روش BE، نمونه بوت استرپ $x_1^*, \dots, x_n^*$ را به روش نمونه گیری تصادفی ساده با جایگذاری از $x_1, \dots, x_n$ به دست می آوریم (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 بوت استرپ ظاهر شده است و می تواند مقادیر $0, 1, \dots, n$ را اختیار کند. واضح است که n_i دارای توزیع $b(1, n)$ است. میانگین نمونه بوت استرپ $\bar{x}^* = \sum x_i^*/n = \sum f_i x_i$ را محاسبه می کنیم. با به دست آوردن تمام نمونه های ممکن بوت استرپ که تعداد آنها $m = n^n$ است و محاسبه میانگین نمونه بوت استرپ روی آنها، $\bar{x}^{*1}, \dots, \bar{x}^{*m}$ به دست می آید. توزیع نمونه گیری $\bar{x}$ می تواند 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 بوت استرپ $\bar{x}^*$ استفاده می کنیم. نمونه گیری بوت استرپ در مرحله ۱ را به طور مستقل B بار تکرار کرده و $\bar{x}^{*1}, \dots, \bar{x}^{*B}$ را محاسبه می کنیم (مرحله ۲). تابع توزیع تجربی $\bar{x}^{*1}, \dots, \bar{x}^{*B}$ می تواند به عنوان تقریب توزیع بوت استرپ $\bar{x}^*$ و در نتیجه تقریب توزیع نمونه گیری $\bar{x}$ برای B های بزرگ ($1000 \leq B \leq 2000$) استفاده شود (مرحله ۳).

نکته مهم در روش BE این است که در تقریب توزیع نمونه گیری $\bar{x}$ ، فرض شده که تابع توزیع نمونه همان تابع توزیع جامعه است. به عبارت دیگر، نمونه های بوت استرپ به طور مستقل از تابع توزیع نمونه (یعنی F_n) تولید شده اند.

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

با به دست آوردن تمام تکرارهای مرحله ۱ (که در عمل کاری غیر ممکن است) و سپس محاسبه $\bar{\mu}$ ، توزیع BBR برای μ به دست می آید که به عنوان تقریب توزیع پسین μ می تواند استفاده شود. از روش مونت کارلو برای تقریب توزیع بوت استرپ بیزی $\bar{\mu}$ استفاده می کنیم. اگر مرحله ۱ الگوریتم BBR را به طور مستقل B بار تکرار کنیم، در نتیجه $\bar{\mu}^1, \dots, \bar{\mu}^B$ به دست می آید (مرحله ۲). تابع توزیع $\bar{\mu}^1, \dots, \bar{\mu}^B$ می تواند به عنوان تقریب توزیع بوت استرپ بیزی $\bar{\mu}$ و در نتیجه تقریب توزیع پسین μ 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 نمونه گیری برآوردکننده θ شبیه سازی و تقریب زده می شود، در صورتی که در روش BBR توزیع پسین پارامتر θ شبیه سازی و تقریب زده می شود.

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

$$\begin{aligned} E(f_i) &= E(g_i) = \frac{1}{n}; \\ \text{var}(f_i) &= \frac{n-1}{n^2}; \quad \text{var}(g_i) = \frac{n-1}{n^2(n+1)}; \\ \text{cov}(f_i, f_j) &= \text{cov}(g_i, g_j) = \frac{-1}{n^2}. \end{aligned}$$

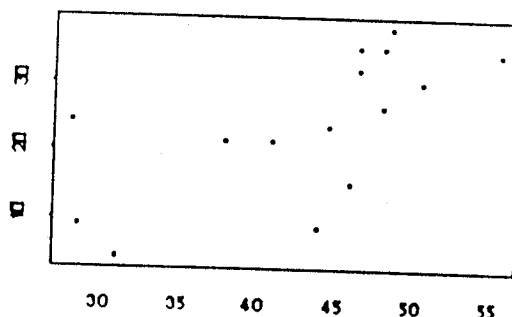
۴ یک مثال دو بعدی

در این بخش دو الگوریتم BE و BBR را در حالتی که x یک متغیر دو بعدی $x = (y, z)$ و پارامتر θ ضریب همبستگی جامعه (ρ) است، همراه با یک مثال کاربردی ([۱]، صفحه ۴۳) ارائه می کنیم.

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

انتخاب و در شکل ۱ نشان داده شده است. ضریب همبستگی

نمونه $r = 0/663$ است.



شکل ۱

• الگوریتم BE را در این مثال برای تقریب توزیع نمونه‌گیری r و مشخصه‌های آن در سه مرحله زیرارایه می‌کنیم:

مرحله ۱ - نمونه تصادفی مشاهده شده را به

صورت زوجی $x_i = (y_i, z_i)$; $i = 1, \dots, 15$ در نظر می‌گیریم. از نمونه زوجی $(x_1, \dots, x_{15})$ به روش نمونه‌گیری تصادفی ساده با جایگذاری، نمونه زوجی بوت‌استرپ $(x_1^*, \dots, x_{15}^*)$ را به دست می‌آوریم.

مرحله ۲ - ضریب همبستگی نمونه زوجی بوت‌استرپ r^* را محاسبه می‌کنیم.

مرحله ۳ - مراحل ۱ و ۲ را ۱۰۰۰ بار تکرار کرده

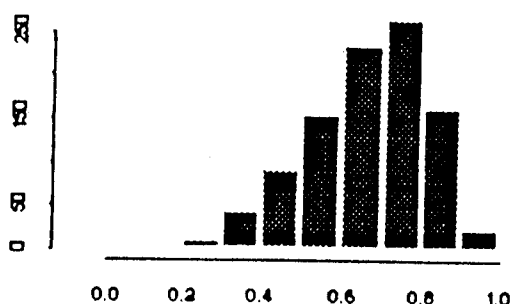
و ۱۰۰۰ ضریب همبستگی بوت‌استرپ $(r^{*1}, \dots, r^{*1000})$ را به دست می‌آوریم. شکل ۲

بافت‌نگار بوت‌استرپ ضریب همبستگی حاصل از ۱۰۰۰ تکرار مرحله ۳ را نشان می‌دهد که می‌تواند به عنوان تقریب توزیع نمونه‌گیری r استفاده شود. از مشخصه‌های توزیع بوت‌استرپ استفاده کرده، برآورد بوت‌استرپ ارببی و خطای معیار ضریب

همبستگی r را به صورت زیر به دست می‌آوریم:

$$\begin{aligned} bias^*(r) &= \frac{1}{1000} \sum_{b=1}^{1000} r^{*b} - r \\ &= 0/662 - 0/663 = 0/001 \end{aligned}$$

$$\begin{aligned} SE^*(r) &= \frac{1}{999} \sum_{b=1}^{1000} (r^{*b} - \frac{1}{1000} \sum_{b=1}^{1000} r^{*b})^2 \Bigg)^{\frac{1}{2}} \\ &= 0/152 \end{aligned}$$



شکل ۲

• همچنین الگوریتم BBR را در این مثال برای تقریب توزیع پسین ρ و مشخصه‌های آن در سه مرحله زیرارایه می‌کنیم:

مرحله ۱ - بردار تصادفی مشاهده شده

$g = (g_1, \dots, g_n)$ را مشابه با مرحله ۱ الگوریتم BBR به دست می‌آوریم.

مرحله ۲ - ضریب همبستگی پسین $\bar{\rho}$ را با استفاده از

بردار مرحله ۱، به صورت زیر محاسبه می‌کنیم:

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

مرحله ۳ - مراحل ۱ و ۲ را ۱۰۰۰ بار تکرار کرده و

۱۰۰۰ ضریب همبستگی پسین $(\rho^1, \dots, \rho^{1000})$ را به دست می‌آوریم. شکل ۳ بافت‌نگار 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 می‌کند. در این توابع؛ x ، نمونه تصادفی زوجی، n حجم نمونه و b تعداد تکرارهای بوت‌استرپ است. خروجی هر دو تابع bootcor و grubin تکرارهای بوت‌استرپ r^* و $\bar{\rho}^*$ است.

تکرار مرحله ۳ را نشان می‌دهد که می‌تواند به عنوان تقریب توزیع نمونه‌گیری پسین ρ استفاده شود. از مشخصه‌های توزیع پسین BBR استفاده کرده، برآورد بوت‌استرپ اریبی و خطای معیار ضریب همبستگی پسین ρ را به صورت زیر به دست می‌آوریم:

$$\begin{aligned} bias^*(\rho) &= \frac{1}{1000} \sum_{b=1}^{1000} \bar{\rho}^b - r \\ &= 0/663 - 0/663 = 0 \end{aligned}$$

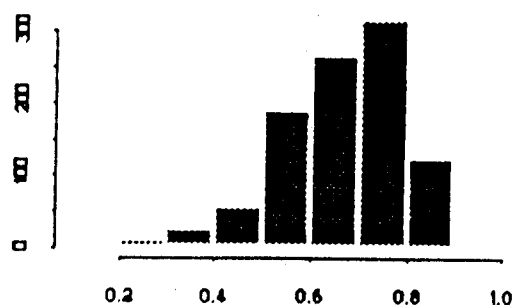
$$\begin{aligned} SE^*(\rho) &= \frac{1}{999} \sum_{b=1}^{1000} (\bar{\rho}^b - \frac{1}{1000} \sum_{b=1}^{1000} \bar{\rho}^b)^2 \}^{\frac{1}{2}} \\ &= 0/131 \end{aligned}$$

```
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 برنامه‌های رایانه‌ای

نرم‌افزار S-PLUS یکی از نرم‌افزارهای جدید و توانای آماری است که امکان برنامه‌نویسی را نیز دارد. مثال بخش ۴ با استفاده از برنامه‌های نوشته شده توسط این نرم‌افزار حل شده است. الگوریتم BE برای پارامتر ρ 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»