

آشنایی با الگوریتم بوت استرپ

نصراله ایران پناه*

عین‌اله پاشا[†]

چکیده

روشهای باز نمونه‌گیری^۱، دستیابی به هدف بالا را بدون ضعفهای دیدگاه سنتی میسر می‌سازند. از جمله روشهای باز نمونه‌گیری روش بوت استرپ^۲ است که به وسیله افرون^۳ [13] در سال ۱۹۷۹ ارائه شده است.

افرون در سال ۱۹۷۹ روش بوت استرپ را برای برآورد اریبی، واریانس، و توزیع نمونه‌ای آماره‌ها ارائه کرد. در این مقاله روش بوت استرپ که مبتنی بر ایده باز نمونه‌گیری از مشاهدات است، به همراه چند مثال کاربردی ارائه می‌شود.

۱.۱ آماره و توزیع نمونه‌ای آماره

موضوع اساسی تحلیل آماری استخراج حداکثر اطلاعات از داده‌ها و استنتاج ویژگیهای جامعه است. تحلیل آماری به طور کلی بر اساس آماره است، که تابعی از داده‌ها است و مطابق بعضی اصول (از جمله، اصل حداکثر درستنمایی، اصل جانیشینی^۴ و غیره) انتخاب می‌شود. قبل از انتخاب داده‌ها، یک آماره یک متغیر تصادفی و دارای یک توزیع احتمال معینی است که توزیع نمونه‌ای آماره نامیده می‌شود. اکثر روشهای آگاهی برای تحلیل نیاز به بعضی اطلاعات از توزیع نمونه‌ای دارند.

از طرف دیگر، در یک مسأله برآورد به علت اینکه هر برآورد کننده ممکن است یک خطای برآورد داشته باشد، داشتن یک اندازه دقت برای برآورد کننده بسیار مهم است. واریانس، اریبی و میانگین توان دوم خطا از جمله اندازه‌های دقت برآورد کننده‌ها هستند. این اندازه‌های دقت مشخصاتی از توزیع نمونه‌ای برآورد کننده‌ها هستند. اندازه‌های دقت همچنین می‌توانند برای انتخاب بهترین برآورد کننده در میان یک رده از برآورد کننده‌های مناسب استفاده شوند.

توزیع نمونه‌ای و مشخصه‌های آن معمولاً به جامعه تحت بررسی بستگی دارند و بنابراین نامعلوم‌اند. در اکثر مسائل استنباط آماری از داده‌های مشاهده شده برای برآورد یا تقریب توزیع نمونه‌ای و مشخصه‌های آن استفاده می‌کنیم. روشهای باز نمونه‌گیری از جمله بوت استرپ و جک‌نایف^۵ روشهایی

۱ پیشگفتار

نظریه آمار همواره تلاش می‌کند که به سه پرسش اساسی زیر پاسخ دهد:

- (۱) اطلاعات را چگونه جمع‌آوری کنیم؟
- (۲) اطلاعات جمع‌آوری شده را چگونه خلاصه و تحلیل کنیم؟
- (۳) دقت خلاصه اطلاعات چقدر است؟

پرسش ۳ و پاسخ به آن بخش مهم و بزرگی از استنباط آماری را تشکیل می‌دهد. مسأله زیر را در نظر بگیرید:

فرض کنید $X_1, \dots, X_n$ نمونه تصادفی مستقل و هم‌توزیع (iid) از توزیع نامعلوم F و $T_n = T(X_1, \dots, X_n)$ برآورد کننده پارامتر نامعلوم θ است. هدف برآورد اندازه‌های دقت آماره T_n از جمله اریبی و واریانس یا برآورد توزیع نمونه‌ای آماره T_n است. اریبی و واریانس برای مقایسه آماره‌ها و توزیع نمونه‌ای برای ساختن فواصل اطمینان مفید هستند. دیدگاه سنتی برای برآورد اندازه‌های دقت به ویژه واریانس عموماً متکی به مشتق‌گیریهای پیچیده است که محاسبه آن گاهی غیر ممکن است. برآورد توزیع نمونه‌ای نیز عموماً مبتنی به قضایای حدی است و بنابراین برای حجم نمونه بزرگ اعتبار دارد.

(*) گروه آمار دانشگاه اصفهان † مؤسسه ریاضیات دکتر مصاحب دانشگاه تربیت معلم

1) resampling 2) bootstrap 3) Efron 4) substitution 5) Jackknife

در نظریه سنتی سعی می‌کنیم مسأله را با در نظر گرفتن تقریبها یا بسطهای مجانبی $\text{var}(T_n)$ ساده‌تر کنیم. اغلب تحت تعدادی شرایط نظم می‌توان تعیین کرد که

$$\lim_{n \rightarrow \infty} [n \text{var}(T_n)] = \sigma_F^2, \quad (3)$$

که در آن $\sigma_F^2 = \sigma^2(F)$ یک تابع ساده از F و یا $\sigma_F^2 = \sigma^2(\gamma)$ یک تابع از بردار γ شامل پارامترهای نامعلوم است. معمولاً $\text{var}(T_n)$ را با مقایسه تجربی تقریب یعنی $\sigma^2(\hat{F})/n$ یا $\sigma^2(\hat{\gamma})/n$ و $\hat{F}$ و $\hat{\gamma}$ برآورد کننده‌های F و γ هستند برآورد می‌کنیم.

مثال ۱- (ادامه). در حالت $T_n = \bar{X}_n^2$ ، آنگاه (۳) برای (۲) به‌ازای $\sigma_F^2 = 4\mu^2\alpha_2$ برقرار است. یعنی (۲) می‌تواند به صورت ساده‌تر زیر تقریب و برآورد شود:

$$\text{var}(\bar{X}_n^2) \simeq \frac{4\mu^2\alpha_2}{n}, \quad \widehat{\text{var}}(\bar{X}_n^2) = \frac{4\bar{X}_n^2\hat{\alpha}_2}{n}.$$

مثال ۲- میانگین نمونه پیراسته. میانگین نمونه α -پیراسته $\bar{X}_n^{(\alpha)}$ را در برآورد نیرومند مرکزیک توزیع F متقارن در نظر می‌گیریم ($0 < \alpha < 1/2$).

$$\bar{X}_n^{(\alpha)} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)},$$

که در آن $[t]$ جزء صحیح عدد حقیقی t و $X_{(i)}$ آماره مرتب نام است. یک شکل دقیق و صریح برای $\text{var}(\bar{X}_n^{(\alpha)})$ وجود ندارد، اما می‌توان نشان داد (لهمن^۶ [24]) که (۳) برقرار است با

$$\sigma_F^2 = \frac{2}{(1 - 2\alpha)^2} \left[\int_0^{F^{-1}(1-\alpha)} x^2 dF(x) + \alpha(F^{-1}(1-\alpha))^2 \right],$$

که در آن

$$F^{-1}(t) = \inf\{x : F(x) \geq t\}. \quad (4)$$

بنابراین، می‌توان $\text{var}(\bar{X}_n^{(\alpha)})$ را بصورت σ_F^2 برآورد کرد، $\hat{F}$ برآورد کننده F است. برای مثال، توزیع تجربی F_n یک برآورد کننده F است که به صورت زیر تعریف می‌شود:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\} =$$

$$\#\{X_i \leq x; i = 1, \dots, n\}/n,$$

که در آن $I\{A\}$ تابع نشانگر مجموعه A است. در واقع F_n جرم $\frac{1}{n}$ را به هر X_i نسبت می‌دهد.

در زیر به بعضی از ضعفها و معایب دیدگاه سنتی را بیان می‌کنیم:

هستند که برای برآورد توزیع نمونه‌ای آماره و مشخصه‌های آن به‌کار می‌روند.

۲.۱ دیدگاه سنتی

قبل از معرفی بوت‌استرپ به‌طور مختصر به دیدگاه سنتی در برآورد اندازه‌های دقت می‌پردازیم. برای مثال واریانس را به صورت یک اندازه دقت در نظر می‌گیریم. فرض کنید $X_1, \dots, X_n$ متغیرهای تصادفی iid با مقادیر حقیقی از یک توزیع نامعلوم F باشند. آماره $T_n = T(X_1, \dots, X_n)$ را در نظر می‌گیریم، واریانس T_n به صورت زیر است:

$$\begin{aligned} \text{var}(T_n) &= E[T_n - E(T_n)]^2 \\ &= \int [T_n(x) - \int T_n(y) d \prod_{i=1}^n F(y_i)]^2 d \prod_{i=1}^n F(x_i) \quad (1) \end{aligned}$$

که در آن $x = (x_1, \dots, x_n)$ و $y = (y_1, \dots, y_n)$ معمولاً $\text{var}(T_n)$ به صورت تابعی از پارامترهای نامعلوم است که با جانشینی برآورد این پارامترها، برآورد می‌شود.

مثال ۱- توابعی از میانگین نمونه. وقتی T_n میانگین نمونه است، آنگاه

$$\begin{aligned} T_n &= \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \\ \text{var}(\bar{X}_n) &= \frac{1}{n} \text{var}(X_1). \end{aligned}$$

بنابراین، می‌توان $\text{var}(\bar{X}_n)$ را با $\text{var}(X_1)$ به صورت یک پارامتر نامعلوم برآورد کرد. اگر F متعلق به خانواده پارامتری نباشد، آنگاه $\text{var}(X_1)$ معمولاً با استفاده از واریانس نمونه S^2 برآورد می‌شود،

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

وقتی T_n تابع ساده‌ای از $\bar{X}_n$ است، $\text{var}(T_n)$ بدون هیچ فرمول پیچیده‌ای به دست می‌آید. برای مثال، اگر $T_n = \bar{X}_n^2$ ، آنگاه

$$\text{var}(\bar{X}_n^2) = \frac{4\mu^2\alpha_2}{n} + \frac{4\mu\alpha_2}{n^2} + \frac{\alpha_2}{n^3}, \quad (2)$$

که در آن $\mu = E(X_1)$ میانگین جامعه و $\alpha_k = E(X_1 - \mu)^k$ گشتاور مرکزی k ام X_1 است. می‌توان $\text{var}(\bar{X}_n^2)$ را با جانشینی μ و α با $\bar{X}_n$ و $\hat{\alpha}_k$ برآورد کرد،

$$\hat{\alpha}_k = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^k, \quad k = 2, 3, 4.$$

اما همواره T_n مانند میانگین نمونه ساده نیست و بنابراین تعیین یک فرمول دقیق و صریح برای $\text{var}(T_n)$ در (۱)، خیلی مشکل و یا غیرممکن است.

بوت استرپ برای T_n است. وقتی $\text{var}^*(T_n^*)$ تابع صریحی از $X_1, \dots, X_n$ نیست، ممکن است که در کاربردهای عملی مورد استفاده قرار نگیرد. اگر $\text{var}^*(T_n^*)$ یک تابع صریحی از $X_1, \dots, X_n$ باشد، آنگاه واقعاً یک برآوردکننده جانشینی $\text{var}(T_n)$ است.

مثال ۱- (ادامه). وقتی $T_n = \bar{X}_n$ و $\hat{F} = F_n$ توزیع تجربی است، با جانشینی F_n به جای F در عبارت

$$\text{var}(\bar{X}_n) = \frac{1}{n} \text{var}(X_1) = \frac{1}{n} \int [x - \int y dF(y)]^2 dF(x),$$

برآورد کننده بوت استرپ واریانس $\bar{X}_n$ را به صورت زیر به دست می آوریم:

$$\text{var}^*(\bar{X}_n^*) = \frac{1}{n^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

وقتی $T_n = \bar{X}_n^*$ ، با استفاده از (۲)، $\mu = \int x dF(x)$ و $\alpha_k = \int (x - \mu)^k dF(x)$ به دست می آوریم:

$$\text{var}^*(\bar{X}_n^*) = \frac{4 \bar{X}_n^* \tilde{\alpha}_2}{n} + \frac{4 \bar{X}_n^* \tilde{\alpha}_2}{n} + \frac{\tilde{\alpha}_2}{n^2},$$

که در آن $\tilde{\alpha}_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^k$.

مثال ۴- میانه نمونه. فرض کنید $T_n = F_n^{-1}(\frac{1}{2})$ میانه نمونه باشد، که در آن F_n^{-1} با جانشینی F_n به جای F در (۴) تعریف می شود. اگر فرض کنیم $n = 2m - 1$ ، آنگاه $T_n = X_{(m)}$. اگر $\hat{F} = F_n$ ، برآورد کننده واریانس بوت استرپ $X_{(m)}$ را به صورت زیر به دست می آوریم. فرض کنید که $X_1^*, \dots, X_n^*$ یک نمونه تصادفی iid از F_n است و N_i^* تعداد X_i هایی است که در نمونه بوت استرپ انتخاب می شود،

$$N_i^* = \#\{X_j^* = X_i | X_1, \dots, X_n\},$$

$$i = 1, \dots, n, j = 1, \dots, n.$$

واضح است که بردار $N^* = (N_1^*, \dots, N_n^*)$ دارای توزیع n جمله ای است که احتمال $\frac{1}{n}$ را به هر جمله نسبت می دهد و بنابراین امید ریاضی انتخاب شدن هر X_i در نمونه بوت استرپ یک است. حال آماره مرتب $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ و مطابق آن $N_{(1)}^* \leq N_{(2)}^* \leq \dots \leq N_{(n)}^*$ را در نظر می گیریم. احتمالات بوت استرپ شرطی به شرط $X_1, \dots, X_n$ را به ازای $k = 1, \dots, n$ در نظر می گیریم،

(۱) چون از فرمول تقریبی σ_F^2 در (۳) استفاده می کنیم، در اکثر مواقع نیاز به حجم نمونه n بزرگ برای داشتن واریانس دقیق (یا اندازه دقت دیگر) برآورد کننده ها داریم.

(۲) فرمول نظری و یا تقریب آن بر اساس مدل فرض شده است. وقتی مدل کمی نادرست است، برآورد کننده دقت به دست آمده ممکن است که دیگر هیچ اعتباری نداشته باشد.

(۳) با به کار بردن دیدگاه سنتی در بعضی از مسائل مختلف، مشتق یک فرمول نظری را به صورت σ_F^2 در (۳) برای هر مسأله در نظر می گیریم. این مشتقگیریها ممکن است مشکل یا پردردسر باشند. بعلاوه، برای مشتقگیری فرمول نظری نیاز به اطلاع کافی از ریاضیات و آمار نظری داریم.

۲ آشنایی با بوت استرپ

افرون [13] در سال ۱۹۷۹ روش بوت استرپ را برای برآورد اندازه دقت و توزیع نمونه ای آماره ها ارائه کرد. این روش براساس ایده باز نمونه گیری از داده ها است، که در چند سال اخیر با استفاده از کامپیوتر گسترش فراوانی یافته است. مباحث تحت بررسی بوت استرپ همان مباحث سنتی آمار هستند، تنها ابزار مورد استفاده تغییر یافته است. کامپیوترها این مباحث را قابل انعطاف، سریع، آسان و با حداقل فرضیات ریاضی ممکن می سازند.

۱.۲ برآورد واریانس

فرض کنید $X_1, \dots, X_n$ متغیرهای تصادفی iid از F هستند و F با $\hat{F}$ برآورد می شود. با جانشینی $\hat{F}$ به جای F در (۱)، برآورد کننده بوت استرپ واریانس T_n را به دست می آوریم،

$$\begin{aligned} \text{var}^*[T(X_1^*, \dots, X_n^*) | X_1, \dots, X_n] &= E^*[T_n^* - E^*(T_n^*)]^2 \\ &= \int [T_n(x) - \int T_n(y) d \prod_{i=1}^n \hat{F}(y_i)]^2 d \prod_{i=1}^n \hat{F}(x_i), \end{aligned} \quad (5)$$

که در آن $\{X_1^*, \dots, X_n^*\}$ یک نمونه تصادفی iid از $\hat{F}$ است و نمونه بوت استرپ نامیده می شود و $\text{var}^*[\cdot | X_1, \dots, X_n]$ واریانس شرطی به شرط $\{X_1^*, \dots, X_n^*\}$ است. به بیان ساده تر، نمونه بوت استرپ $\{X_1^*, \dots, X_n^*\}$ با نمونه گیری تصادفی ساده با جایگذاری از نمونه اولیه $\{X_1, \dots, X_n\}$ به دست می آید. به طور کلی، نمونه گیری iid از $\hat{F}$ یک اندازه احتمال، امید ریاضی و واریانس شرطی به شرط $X_1, \dots, X_n$ تعریف می کند که به صورت E^*, P^*, var^* نشان می دهیم. برای سادگی، از این به بعد قسمت شرط را حذف می کنیم. همچنین $T_n^* = T(X_1^*, \dots, X_n^*)$ را آماره بوت استرپ می نامیم. معادله (۵) شکل نظری برآورد کننده واریانس

واضح است که با استفاده از قانون قوی اعداد بزرگ، وقتی که $B \rightarrow \infty$ آنگاه،

$$\widehat{\text{var}}^*(T_n^*) \xrightarrow{a.s.} \text{var}^*(T_n^*).$$

هم $\text{var}^*(T_n^*)$ و هم تقریب مونته کارلو آن $\widehat{\text{var}}^*(T_n^*)$ برآورد کننده های واریانس بوت استرپ نامیده می شوند. اغلب در کاربردهای عملی مفید است، در حالی که در مطالعات نظری معمولاً روی $\text{var}^*(T_n^*)$ دقت می کنیم.

بنابراین، روش بوت استرپ ترکیبی از دو روش است: اصل جانشینی و تقریب مونته کارلو. وقتی $\text{var}^*(T_n^*)$ در (۵) تابع صریحی از $X_1, \dots, X_n$ است، بوت استرپ دقیقاً با دیدگاه جانشینی سنتی منطبق می شود، در غیر این صورت بوت استرپ از (۷) برای تقریب استفاده می کند. در حالیکه، دیدگاه سنتی ابتدا به طور تحلیلی $\text{var}(T_n)$ را تقریب می کند و سپس پارامترهای نامعلوم در فرمول تقریبی $\text{var}(T_n)$ را برآورد می کند. این مفهوم بوت استرپ کمک می کند که آن را به مسائل پیچیده تر تعمیم دهیم.

۳.۲ برآورد توزیع نمونه ای

فرض کنید $X_1, \dots, X_n$ متغیرهای تصادفی iid از F هستند، هدف برآورد توزیع نمونه ای آماری $T_n = T(X_1, \dots, X_n)$ است. چون معمولاً از T_n برای ساختن فواصل اطمینان برای یک پارامتر نامعلوم θ (θ به F وابسته است) استفاده می کنیم، به جای T_n یک ریشه آن $R_F = R(X_1, \dots, X_n, F)$ را در نظر می گیریم. ریشه R_F ممکن است به صورت $T_n - \theta$ یا آماره استودنت شده $(T_n - \theta)/S_n$ باشد، که در آن S_n برآورد کننده انحراف معیار T_n است. پس نیاز داریم که توزیع نمونه ای R_F را به دست آوریم.

$$H_F(x) = P\{R(X_1, \dots, X_n, F) \leq x\}, \quad (A)$$

R_F و H_F در واقع به n بستگی دارند که برای سادگی نمادها حذف شده اند. در دیدگاه سنتی، ابتدا به دنبال یک فرمول نظری ساده دقیق یا تقریبی برای $H_F(x)$ هستیم و سپس پارامترهای نامعلوم در فرمول نظری را با برآوردهای آن جانشین می کنیم. برای مثال، وقتی $R_F = T_n - \theta$ ، در بعضی از حالتها $H_F(x)$ با $\Phi(\sqrt{n}x/\sigma)$ تقریب می شود، که در آن $\Phi(\cdot)$ توزیع نرمال استاندارد و σ یک پارامتر نامعلوم وابسته به F است. اگر $\hat{\sigma}$ یک برآورد کننده σ باشد، آنگاه $H_F(x)$ با $\Phi(\sqrt{n}x/\hat{\sigma})$ برآورد می شود. برای $R_F = (T_n - \theta)/S_n$ ، در بعضی از حالتها $H_F(x)$ با $\Phi(x)$ تقریب می شود. این دیدگاه، به هر حال ضعیف و معایبی دارد که در بخش ۲-۱ توضیح داده شد.

مشابه دستورالعملی که برای برآورد کننده بوت استرپ واریانس در بخش ۱-۲ و تقریب آن در بخش ۲-۲ دادیم، اینک برآورد کننده بوت استرپ $H_F(x)$ را ارائه می کنیم. ابتدا $\hat{F}$ را به جای F در (A) قرار می دهیم،

$$\begin{aligned} p_k &= P^*\{X_{(m)}^* = X_{(k)}\} \\ &= P^*\{X_{(m)}^* > X_{(k-1)}\} - P^*\{X_{(m)}^* > X_{(k)}\} \quad (۶) \\ &= P^*\left\{\sum_{i=1}^{k-1} N_{(i)}^* \leq m-1\right\} - P^*\left\{\sum_{i=1}^k N_{(i)}^* \leq m-1\right\} \\ &= P\left\{\text{binomial}\left(n, \frac{k-1}{n}\right) \leq m-1\right\} \\ &\quad - P\left\{\text{binomial}\left(n, \frac{k}{n}\right) \leq m-1\right\} \\ &= \sum_{j=0}^{m-1} \binom{n}{j} \left(\frac{k-1}{n}\right)^j \left(\frac{n-k+1}{n}\right)^{n-j} \\ &\quad - \sum_{j=0}^{m-1} \binom{n}{j} \left(\frac{k}{n}\right)^j \left(\frac{n-k}{n}\right)^{n-j} \\ &= \sum_{j=0}^{m-1} \binom{n}{j} \frac{(k-1)^j (n-k+1)^{n-j} - k^j (n-k)^{n-j}}{n^n}. \end{aligned}$$

فرض کنید $F_{X_{(m)}}$ توزیع $X_{(m)}$ باشد. چون

$$\text{var}(X_{(m)}) = \int [x - \int y dF_{X_{(m)}}(y)]^2 dF_{X_{(m)}}(x)$$

با جایگذاری توزیع بوت استرپ $X_{(m)}^*$ (که از رابطه (۶) به دست می آید) به جای $F_{X_{(m)}}$ نتیجه زیر را به دست می آوریم،

$$\text{var}^*(X_{(m)}^*) = \sum_{k=1}^n p_k (X_{(k)} - \sum_{j=1}^n p_j X_{(j)})^2.$$

به هر حال، همان طور که قبلاً عنوان شد، رابطه های (۱) یا (۵) معمولاً پیچیده هستند و $\text{var}^*(T_n^*)$ تابع صریحی از $X_1, \dots, X_n$ نیست. در نتیجه از روش مونته کارلو برای تقریب استفاده می کنیم.

۲.۲ روش مونته کارلو

وقتی طرف راست (۱) ساده نیست، حتی اگر F معلوم باشد، نمی توان $\text{var}(T_n)$ را به طور دقیق محاسبه نمود. در آمار روشی قدیمی به نام مونته کارلو وجود دارد که وقتی F معلوم است، با استفاده از آن می توان $\text{var}(T_n)$ را به طور عددی تقریب کرد. یعنی اینکه، به طور تکراری مجموعه داده های جدیدی از F به دست می آوریم و آنگاه آماره T_n را روی هر مجموعه داده محاسبه می کنیم، واریانس نمونه مقادیر T_n یک تقریب عددی $\text{var}(T_n)$ است. وقتی $\hat{F}$ معلوم است، این ایده می تواند برای تقریب $\text{var}^*(T_n^*)$ استفاده شود. یعنی، با فرض معلوم بودن $X_1, \dots, X_n, B$ نمونه مستقل بوت استرپ $\{X_{1b}^*, \dots, X_{nb}^*\}$ را از $\hat{F}$ به دست می آوریم. سپس آماره های بوت استرپ $T_{n,b}^* = T(X_{1b}^*, \dots, X_{nb}^*)$ را محاسبه می کنیم، $\text{var}^*(T_n^*)$ با استفاده از تقریب مونته کارلو به صورت زیر به دست می آید:

$$\widehat{\text{var}}^*(T_n^*) = \frac{1}{B-1} \sum_{b=1}^B (T_{n,b}^* - \frac{1}{B} \sum_{j=1}^B T_{n,j}^*)^2. \quad (۷)$$

مثال ۵- در حالت $n = 2$ ، نمونه (X_1, X_2) را با فرض $X_1 > X_2$ در نظر می‌گیریم. F_n جرم $\frac{1}{2}$ را به X_1 و X_2 نسبت می‌دهد. نمونه بوت استرپ (X_1^*, X_2^*) با احتمال $\frac{1}{2}$ مقادیر (X_1, X_2) ، (X_1, X_1) و (X_2, X_2) را می‌گیرد و چون ترتیب اهمیتی ندارد (X_1^*, X_2^*) مقادیر (X_1, X_1) ، (X_2, X_2) و (X_1, X_2) را با احتمالهای $\frac{1}{2}$ ، $\frac{1}{2}$ و $\frac{1}{2}$ می‌گیرد. بوت استرپ میانگین بخش ۱-۳ را در نظر می‌گیریم که در آن

$$\bar{X} = (X_1 + X_2)/2, S = (X_1 - X_2)/2, R = \sqrt{2}(\bar{X} - \mu)/S.$$

هدف تقریب توزیع $\gamma(F) = P(R \leq x)$ با استفاده از بوت استرپ است. چون $\bar{X}^*$ مقادیر $X_1, \bar{X}, X_2$ را با احتمالهای $\frac{1}{2}$ ، $\frac{1}{2}$ و $\frac{1}{2}$ می‌گیرد، پس $R^* = \sqrt{2}(\bar{X}^* - \bar{X})/S$ مقادیر $0, \sqrt{2}$ و $-\sqrt{2}$ را با احتمالهای $\frac{1}{2}$ ، $\frac{1}{2}$ و $\frac{1}{2}$ می‌گیرد، بنابراین

$$\gamma^* = \gamma(F_n) = P^*(R^* \leq x) = \begin{cases} 0 & x < -\sqrt{2}, \\ \frac{1}{2} & -\sqrt{2} \leq x < 0, \\ \frac{2}{2} & 0 \leq x < \sqrt{2}, \\ 1 & \sqrt{2} \leq x. \end{cases}$$

در حالت کلی تعداد نمونه‌های بوت استرپ مجزا برابر $m = \binom{n-1}{n}$ است که آن را به صورت $\{(X_{1j}^*, \dots, X_{nj}^*), j = 1, \dots, m\}$ نشان می‌دهیم. حال اگر در نمونه بوت استرپ $(X_1^*, \dots, X_n^*)$ ، n_1 بار X_1 ، n_2 بار X_2 ، ... و n_n بار X_n ظاهر شده باشند، احتمال به دست آوردن نمونه بوت استرپ از توزیع n جمله‌ای زیر تبعیت می‌کند:

P (مشاهده نمونه بوت استرپ)

$$= \binom{n}{n_1, \dots, n_n} \prod_{i=1}^n \left(\frac{1}{n}\right)^{n_i} = \frac{n!}{n_1! \dots n_n!} \left(\frac{1}{n}\right)^n,$$

$$\sum_{i=1}^n n_i = n, n_i = 0, 1, \dots, n.$$

اگر در بین m نمونه بوت استرپ مجزا، احتمال به دست آوردن k امین نمونه بوت استرپ را با p_k ، $(k = 1, \dots, m)$ نشان دهیم (به صورت احتمال بالا محاسبه می‌شود)، آنگاه آماره بوت استرپ k ام T^{*k} مقدار $T(X_{1k}^*, \dots, X_{nk}^*)$ را با احتمال p_k می‌گیرد و در نتیجه γ^* قابل محاسبه است. مثلاً γ^* را به صورت اندازه دقت واریانس (۵) در نظر می‌گیریم، در این صورت

$$\begin{aligned} \gamma^* &= \text{var}^*(T^*) = E^*[T^* - E^*(T^*)]^2 \\ &= \sum_{k=1}^m p_k [T^{*k} - \sum_{j=1}^m p_j T^{*j}]^2 \end{aligned}$$

$$H^*(x) = H_F(x) =$$

$$P^*\{R(X_1^*, \dots, X_n^*, \hat{F}) \leq x | X_1, \dots, X_n\}, \quad (9)$$

که در آن $X_1^*, \dots, X_n^*$ از iid $\hat{F}$ هستند. اگر $H^*(x)$ تابع صریحی از $X_1, \dots, X_n$ باشد، آنگاه برآورد کننده بوت استرپ $H_F(x)$ خواهد بود. در غیر این صورت، می‌توانیم از تقریب مونت کارلو استفاده کنیم،

$$\hat{H}^*(x) = \frac{1}{B} \sum_{b=1}^B I\{R(X_{1b}^*, \dots, X_{nb}^*, \hat{F}) \leq x\} \quad (10)$$

که در آن $\{X_{1b}^*, \dots, X_{nb}^*\}$ ، $b = 1, \dots, B$ نمونه‌های بوت استرپ مستقل از $\hat{F}$ هستند.

۴.۲ کاربردهای دیگر

گاهی اندازه دقت یک آماره T_n را به صورت مشخصه توزیع نمونه‌ای T_n در نظر می‌گیریم. برآورد کننده‌های بوت استرپ اندازه دقت می‌توانند با استفاده از مشخصه H^* یا $\hat{H}^*$ به دست آورده شوند. برای مثال، برآورد کننده بوت استرپ واریانس $\text{var}^*(T_n^*)$ در (۵) یا $\widehat{\text{var}}^*(T_n^*)$ در (۷) واریانس توزیع بوت استرپ H^* در (۹) یا $\hat{H}^*$ در (۱۰) است با $R_n = T_n$ همچنین برآورد کننده بوت استرپ اریبی، میانگین توان دوم خطا و برد میان چارکی توزیع نمونه‌ای T_n ، اریبی، میانگین توان دوم خطا و برد میان چارکی H^* یا $\hat{H}^*$ است. فواصل اطمینان بوت استرپ برای یک پارامتر نامعلوم θ می‌تواند با استفاده از صدکهای H^* یا $\hat{H}^*$ به دست آید. برای مثال، با فرض $R_n = T_n$ ، فاصله اطمینان صدکی بوت استرپ $(1 - 2\alpha)$ درصد برای θ صدکهای α ام و $(1 - \alpha)$ ام H^* یا $\hat{H}^*$ است. برای توضیح بیشتر در مورد فواصل اطمینان بوت استرپ می‌توان به مراجع [10]، [11] و [14] مراجعه کرد. همچنین فواصل اطمینان بوت استرپ بهتری در مراجع [12] و [9] ارائه شده‌اند. روش بوت استرپ در زمینه‌های دیگری از جمله آزمون فرض، پیش‌بینی و انتخاب مدل مطرح است. برای دیدن یک سری مثالهای کاربردی روش بوت استرپ همراه با بعضی الگوریتمهای مفید می‌توان به کتاب مقدمه‌ای بر بوت استرپ افرون و تی‌بشیرانی^۷ [15] مراجعه کرد.

۵.۲ علت استفاده از تقریب مونت کارلو

برآورد کننده‌های بوت استرپ (۵) و (۹) چون تابعی از توزیع تجربی F_n اند، مقادیری معلوم‌اند. چرا از روش مونت کارلو برای تقریب آنها استفاده می‌کنیم؟ علت این است که برای n های بزرگ محاسبه آنها عملاً غیر ممکن است. مشکل بودن محاسبات را برای حالت $n = 2$ در مثال زیر نشان می‌دهیم.

از نمونه اولیه $X_1, \dots, X_n$ به روش نمونه‌گیری تصادفی ساده با جایگذاری نمونه بوت‌استرپ $X_1^*, \dots, X_n^*$ را به دست می‌آوریم. $\bar{X}_n^*$ و S_n^{*2} ریشه بوت‌استرپ R_n^* را به صورت زیر تعریف می‌کنیم:

$$\begin{aligned}\bar{X}_n^* &= \frac{1}{n} \sum_{i=1}^n X_i^*, \\ S_n^{*2} &= \frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2, \\ R_n^* &= \sqrt{n}(\bar{X}_n^* - \bar{X}_n)/S_n^*.\end{aligned}$$

انتظار می‌رود که رفتار ریشه R_n^* از R_n تقلید کند. بنابراین، توزیع R_n^* (که می‌تواند تنها با استفاده از نمونه مشاهده شده محاسبه شود) می‌تواند برای تقریب توزیع نمونه‌ای نامعلوم R_n استفاده شود. یا حتی به طور دقیقتر، توزیع بوت‌استرپ $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ می‌تواند برای تقریب توزیع نمونه‌ای $\sqrt{n}(\bar{X}_n - \mu)$ استفاده شود. بیکل و فریدمن [2] نشان دادند که وقتی که $n \rightarrow \infty$ ، آنگاه:

الف-۱) توزیع شرطی $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ به طور ضعیف همگرا به توزیع نرمال $N(0, \sigma^2)$ است.

الف-۲) در احتمال شرطی $S_n^* \rightarrow \sigma$ ، یعنی اینکه، برای هر $\varepsilon > 0$ ،

$$P^*\{|S_n^* - \sigma| > \varepsilon | X_1, \dots, X_n\} \xrightarrow{a.s.} 0.$$

بیکل و فریدمن به طور مشابه وقتی X_i ها متغیرهای تصادفی بردار-مقدار هستند سازگاری بوت‌استرپ را نشان دادند.

ب) بوت‌استرپ میانه. با توجه به فرضهای قسمت (الف)، فرض کنید $\bar{X}_n$ میانه نمونه واقعی $X_1, \dots, X_n$ و $\bar{X}_n^*$ میانه نمونه بوت‌استرپ $X_1^*, \dots, X_n^*$ باشد. می‌توان نشان داد (لهم [24]) که اگر F دارای یک میانه یکتا m و f دارای یک مشتق مثبت و پیوسته در همسایگی m باشد، آنگاه

$$R_n = \sqrt{n}(\bar{X}_n - m) \xrightarrow{L} N\left(0, \frac{1}{4f^2(m)}\right).$$

بیکل و فریدمن نشان دادند که $R_n^* = \sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ نیز به طور ضعیف به همان توزیع نرمال میل می‌کند. اگر فرض کنیم $R_n = \sqrt{n}(\bar{X}_n - m)/\sqrt{4f^2(m)}$ و بخواهیم در برآورد بوت‌استرپ ریشه یعنی R_n^* به جای واریانس $1/4f^2(m)$ برآورد بوت‌استرپ واریانس میانه را قرار دهیم، آنگاه بوت‌استرپ بی‌اعتبار خواهد بود. زیرا بدون فرض بیشتر نمی‌دانیم که برآورد بوت‌استرپ واریانس میانه به سمت واریانس حدی $1/4f^2(m)$ میل می‌کند.

واضح است که محاسبه دقیق γ^* برای $n \geq 5$ به دلیل بزرگ بودن m عملاً غیر ممکن است. حتی انجام محاسبات با استفاده از کامپیوتر نیز با صرف وقت بسیار همراه است. برای مثال، برای حجم نمونه کوچک $n = 10$ باید $m = 92378$ آماره T^{*k} و احتمال p_k را محاسبه نمود. در نتیجه به جای m نمونه بوت‌استرپ B نمونه بوت‌استرپ را تولید، T^{*k} را محاسبه و با استفاده از تقریب مونت کارلو $\hat{\gamma}^*$ را به دست می‌آوریم.

۳ استنباطهای مجانبی

بیشتر مطالعات نظری در زمینه بوت‌استرپ، رفتار مجانبی برآورد کننده‌های بوت‌استرپ است. اگر چه در کاربردهای واقعی حجم نمونه نمی‌تواند به طور نامحدودی افزایش یابد، با این حال نظریه مجانبی ملاک مناسبی برای استفاده از روش بوت‌استرپ خواهد بود.

۱.۳ مفهوم سازگاری بوت‌استرپ

هریک از اندازه‌های دقت مانند واریانس (۱)، همچنین توزیعهای نمونه‌ای مانند (۸) توابعی از n و F هستند که می‌توان آنها را به صورت تابع $\gamma_n(F)$ نشان داد. با تغییر کوچکی مشابه (۳) اغلب داریم:

$$\lim_{n \rightarrow \infty} \gamma_n(F) = \gamma(F).$$

حال اگر $\gamma_n(F_n)$ یا γ_n^* برآورد کننده بوت‌استرپ $\gamma_n(F)$ باشد، برآورد کننده بوت‌استرپ را سازگار می‌گوئیم اگر وقتی که $n \rightarrow \infty$ ،

$$\gamma_n(F_n) \rightarrow \gamma(F).$$

بیکل و فریدمن^۸ [2] سازگاری بوت‌استرپ را برای میانگین نمونه، تابعهای فون میزس^۹، فرایندهای تجربی و فرایندهای چندک از قبیل میانه، فواصل صدکی و برآورد کننده L (ترکیب خطی از آماره‌های مرتب) مانند میانگین α -پیراسته بررسی کردند. در زیر تنها برای دو آماره میانگین و میانه نمونه توضیح مختصر داده و شرایط سازگاری را بیان می‌کنیم.

الف) بوت‌استرپ میانگین. فرض کنید $X_1, \dots, X_n$ متغیرهای تصادفی حقیقی-مقدار iid از توزیع نامعلوم F بامیانگین μ و واریانس محدود σ^2 باشند. اگر $T_n = \bar{X}_n$ میانگین نمونه و $S_n^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ آنگاه با استفاده از قضیه حد مرکزی،

$$R_n = \sqrt{n}(\bar{X}_n - \mu)/S_n \xrightarrow{L} N(0, 1).$$

از توزیع تجربی F_n که جرم $\frac{1}{n}$ را به هر X_i نسبت می‌دهد، به صورت iid نمونه بوت‌استرپ $X_1^*, \dots, X_n^*$ را به دست می‌آوریم. به عبارت دیگر،

۲.۳ دقت بوت استرپ

سینگ^{۱۱} [30] با فرض $E(X^2) < \infty$ سازگاری قوی به طور یکنواخت برآوردکننده بوت استرپ توزیع ریشه $\bar{X}_n$ را نشان داد. یعنی وقتی که $n \rightarrow \infty$

$$\sup_x |P^* \{ \sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x \} - P \{ \sqrt{n}(\bar{X}_n - \mu) \leq x \}| \xrightarrow{a.s.} 0$$

همچنین با فرض $E|X|^2 < \infty$ نشان داد که برآورد کننده بوت استرپ توزیع استودنت شده $\bar{X}_n$ دارای دقت مرتبه دوم است. یعنی وقتی که $n \rightarrow \infty$

$$\sup_x |P^* \{ \sqrt{n}(\bar{X}_n^* - \bar{X}_n)/S_n \leq x \} - P \{ \sqrt{n}(\bar{X}_n - \mu)/\sigma \leq x \}| = o(\frac{1}{\sqrt{n}}) \quad a.s.$$

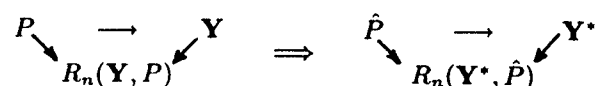
سازگاری بالا نشان می دهد که تقریب بوت استرپ توزیع، دقیقتر از توزیع نرمال حدی است. یکی از موارد استفاده تقریب توزیع در به دست آوردن فواصل اطمینان است. سینگ همچنین نرخ دقیق همگرایی را برای تقریب بوت استرپ توزیع نمونه ای چندکها تعیین کرد و نشان داد که بادقت بالا می توان توزیع $\sqrt{n}[F_n^{-1}(\alpha) - F^{-1}(\alpha)]$ را با توزیع بوت استرپ $\sqrt{n}[F_n^{*-1}(\alpha) - F_n^{-1}(\alpha)]$ تقریب کرد.

۴ الگوریتم بوت استرپ

در این بخش ابتدا روش بوت استرپ را در یک حالت کلی خلاصه می کنیم و سپس الگوریتمی برای محاسبات کاربردی و برنامه های کامپیوتری ارائه می کنیم.

۱.۴ خلاصه روش بوت استرپ

فرض کنید نمونه $Y = (Y_1, \dots, Y_n)$ را داریم (لازم نیست iid باشند) و P یک مدل آماری است که تحت آن نمونه ها به دست آمده اند. معمولاً، P می تواند با استفاده از توزیع توأم Y یا بعضی پارامترها که به طور یکتایی این توزیع توأم را تعیین می کنند، مشخص شود. فرض کنید $R_n(Y, P)$ یک ریشه است و می خواهیم توزیع آن را برآورد کنیم. ابتدا مدل P را با استفاده از نمونه Y برآورد می کنیم. فرض کنید Y^* یک نمونه بوت استرپ باشد که از مدل برآورد شده $\hat{P}$ تولید شده است. توزیع شرطی $R_n(Y^*, \hat{P})$ به شرط Y برآوردکننده بوت استرپ توزیع $R_n(Y, P)$ است. افرون و تییشیرانی [4] این فرایند را به صورت نمودار زیر خلاصه کردند:



هنر بوت استرپ استفاده از تقلید رفتار نمونه گیری سه تایی $(\hat{P}, Y^*, R_n(Y^*, \hat{P}))$ از $(P, Y, R_n(Y, P))$ است، به طوری که ارتباط میان $\hat{P}, Y^*$ و $R_n(Y^*, \hat{P})$ مشابه ارتباط میان P, Y و $R_n(Y, P)$ است. اگر $\hat{P} = P$ ، آنگاه توزیع $R_n(Y^*, \hat{P})$ دقیقاً مشابه توزیع $R_n(Y, P)$ است. حتی اگر $\hat{P} \neq P$ ، توزیعهای $R_n(Y^*, \hat{P})$ و $R_n(Y, P)$ ممکن است شبیه هم باشند.

با اینکه بوت استرپ براساس اصل جانشینی و تقلید کردن رفتار نمونه گیری است، در عمل معمولاً با بازنمونه گیری انجام می شود. یعنی وقتی که توزیع شرطی $R_n(Y^*, \hat{P})$ تابع صریحی از Y نیست، روش مونت کارلو برای محاسبه کردن برآورد کننده های بوت استرپ مورد نیاز است. بوت استرپ می تواند در تمام حالت هایی که مدل P می تواند معلوم فرض شود و یا با استفاده از $\hat{P}$ برآورد شود، کاربرد داشته باشد. این مسأله را در دو مثال زیر نشان می دهیم.

مثال ۶- مسأله یک نمونه ای. فرض کنید $X_1, \dots, X_n$ یک نمونه تصادفی iid از توزیع F باشند. توزیع توأم $X_1, \dots, X_n$ با استفاده از F تعیین می شود. بنابراین، $P = F$. اگر F متعلق به خانواده پارامتری باشد، آنگاه $P = F_\theta$ ، به طوری که θ یک بردار از پارامترهای نامعلوم است. در حالت پارامتری، ابتدا θ با استفاده از برآوردکننده $\hat{\theta}$ برآورد می شود و سپس P را با استفاده $\hat{P} = F_{\hat{\theta}}$ برآورد می کنیم. اکنون نمونه بوت استرپ $X_1^*, \dots, X_n^*$ از $\hat{P}$ تولید می شود. این روش بوت استرپ اغلب بوت استرپ پارامتری نامیده می شود. در حالت ناپارامتری، P با استفاده از توزیع تجربی $\hat{P} = F_n$ برآورد می شود. سپس نمونه بوت استرپ $X_1^*, \dots, X_n^*$ را از F_n تولید می کنیم. این روش بوت استرپ اغلب بوت استرپ ناپارامتری نامیده می شود. بوت استرپ ناپارامتری می تواند برای هر دو مدل پارامتری و ناپارامتری استفاده شود.

بوت استرپ پارامتری بستگی به فرض معلوم بودن مدل پارامتری دارد. در صورتی که بوت استرپ ناپارامتری آزاد از فرض معلوم بودن مدل است. وقتی که مدل پارامتری دقیق است، بوت استرپ پارامتری کاراتر از بوت استرپ ناپارامتری است. به طور کلی، روش بوت استرپ متکی به این است که ما چگونه به خوبی مدل را تشخیص دهیم یا برآورد کنیم. حتی در حالت بوت استرپ ناپارامتری، وقتی که می دانیم F هموار است، می توانیم برآورد کننده هموار F را به جای F_n قرار دهیم. در این صورت برآورد کننده بوت استرپ ناپارامتری بهتری به دست می آید.

مثال ۷- مدل رگرسیون خطی. فرض می کنیم $Y_1, \dots, Y_n$ یک نمونه تصادفی مستقل باشد، به طوری که $Y_i = (y_i, x_i')$ ، $i = 1, \dots, n$ ، بردار p متغیره x' و ترانهاده x است. با توجه به اینکه x_i تصادفی باشد یا نه، در این مسأله با دو مدل مختلف روبرو هستیم.

[4]، بسط اجورث^{۱۱} و تقریب بوت استرپ رگرسیون کاکس [19] و بوت

استرپ رگرسیون لجستیک [26] از دیگر بحثهای بوت استرپ است.

(ج) نمونه‌گیری. بوت استرپ در نمونه‌گیری طبقه بندی در [2]، بسطهای اجورث و بوت استرپ در نمونه‌گیری طبقه بندی در [8] و بوت استرپ در نمونه گیری سیستماتیک در [25] بحث شده است.

(د) چند متغیره. کاربرد بوت استرپ در تحلیل عاملی در [6]، تحلیل مبتن در [9] و تحلیل خوشه‌ای در [22] و [28] ارائه شده است.

(و) استنباط بیزی. ایده باز نمونه‌گیری داده‌ها می‌تواند در محاسبه احتمال پسین توزیع‌ها در تحلیل بیزی کاربرد داشته باشد. ایده بوت استرپ بیزی به وسیله روبین [29] ارائه گردید. گسترش این ایده در زمینه استنباط بیزتجربی براساس نمونه بوت استرپ در [25] و استنباط بیزناپارامتری در [27] ارائه شده است.

(ه) مشاهدات وابسته و سریهای زمانی. روش بوت استرپ در مشاهدات مستقل محققاً در مشاهدات وابسته کاربرد ندارد و با توجه به اینکه ساختمان وابستگی در مشاهدات وابسته معلوم (مانند سریهای زمانی) و یا نامعلوم باشد، تفاوت دارد. در زمینه بوت استرپ مشاهدات وابسته و به خصوص سریهای زمانی در چند سال اخیر تحقیقات و مقالات بسیاری ارائه گردیده است که برای آشنایی با این روشها همراه با مثالهای کاربردی و برنامه‌های کامپیوتری با نرم افزار S-PLUS می‌توان به [1] مراجعه نمود.

۳.۴ الگوریتم بوت استرپ

فرض کنید $X_1, \dots, X_n$ نمونه تصادفی iid از توزیع نامعلوم F باشد و $T_n = T(X_1, \dots, X_n)$ یک برآوردکننده پارامتر نامعلوم θ باشد. هدف برآورد اندازه‌های دقت آماره T_n و توزیع نمونه‌ای آماره T_n و یا ریشه‌های $R_n(T_n, F)$ به صورت $\gamma_n(F)$ است. در زیر مراحل سه گانه روش بوت استرپ را برای محاسبه برآوردهای مورد نظر ارائه می‌کنیم.

مرحله ۱- نمونه بوت استرپ. با نمونه گیری iid از توزیع تجربی F_n نمونه بوت استرپ $X_1^*, \dots, X_n^*$ را تولید می‌کنیم. به عبارت ساده‌تر، نمونه بوت استرپ $X_1^*, \dots, X_n^*$ را با نمونه‌گیری تصادفی ساده با جایگذاری از نمونه اولیه $X_1, \dots, X_n$ به دست می‌آوریم.

مرحله ۲- آماره و سازگاری بوت استرپ. با تأثیر آماره T_n بر نمونه بوت استرپ $X_1^*, \dots, X_n^*$ آماره بوت استرپ $T_n^* = T(X_1^*, \dots, X_n^*)$ و در صورت لزوم ریشه بوت استرپ $R_n^* = R(T_n^*, F_n)$ را محاسبه می‌کنیم. اکنون برآوردکننده‌های بوت استرپ اندازه دقت و توزیع نمونه‌ای را به صورت γ_n^* به دست می‌آوریم. باید توجه کنیم که بوت استرپ وقتی معتبر

حالت الف) اگر x_i تصادفی نباشد، $y_i = x_i' \beta + \varepsilon_i$ به طوری که β یک بردار p متغیره نامعلوم از پارامترها و $\varepsilon_1, \dots, \varepsilon_n$ iid از توزیع نامعلوم F_ε با میانگین صفر هستند. در این حالت، P می‌تواند به صورت (β, F_ε) مشخص شود. فرض کنید $\hat{\beta}$ برآورد کننده β باشد (برای مثال برآورد کننده حداقل مربعات). آنگاه F_ε می‌تواند با توزیع تجربی $\hat{F}_\varepsilon$ برآورد شود. $\hat{F}_\varepsilon$ جرم $\frac{1}{n}$ را با به هر $\varepsilon_i = y_i - \hat{\beta}' x_i', i = 1, \dots, n$ ، $\hat{\varepsilon}_i = \varepsilon_i - \frac{1}{n} \sum_{j=1}^n \varepsilon_j$ نسبت می‌دهد، به طوری که $\hat{\varepsilon}_i = y_i - \hat{\beta}' x_i'$ ، $i = 1, \dots, n$ ، اکنون P با $(\hat{\beta}, \hat{F}_\varepsilon) = \hat{P}$ برآورد می‌شود. برای تولید نمونه بوت استرپ $Y_1^*, \dots, Y_n^*$ ابتدا داده‌های $\varepsilon_1^*, \dots, \varepsilon_n^*$ را به صورت iid از $\hat{F}_\varepsilon$ تولید می‌کنیم و سپس $y_i^* = x_i' \hat{\beta} + \varepsilon_i^*$ ، $i = 1, \dots, n$ ، آنگاه $Y_i^* = (y_i^*, x_i')$ ، $i = 1, \dots, n$ این روش به نام بوت استرپ کردن باقیمانده‌ها معروف است.

حالت ب) x_i تصادفی است، $(y_i, x_i'), i = 1, \dots, n$ iid از توزیع نامعلوم $(p+1)$ متغیره F هستند $E(y_i | x_i) = x_i' \beta$ در این حالت، $P = F$ می‌تواند با توزیع تجربی F_n برآورد شود. F_n جرم $\frac{1}{n}$ را به هر جفت $(y_i, x_i'), i = 1, \dots, n$ نسبت می‌دهد. نمونه بوت استرپ $Y_1^* = (y_1^*, x_1'), i = 1, \dots, n$ را به صورت iid از F_n تولید می‌کنیم. این روش به نام بوت استرپ زوجها معروف است.

۲.۴ بوت استرپ در ساختمان داده‌های پیچیده‌تر

هر چند بوت استرپ در ابتدا برای نمونه‌های iid ارائه شد، معمولاً تعمیم آن به نمونه‌های غیر iid ساده است. در بخش قبل نشان دادیم که چگونه بوت استرپ می‌تواند برای نمونه Y غیر iid استفاده شود. طبیعتاً، بعضی از خاصیت‌هایی که برآوردکننده‌های بوت استرپ در حالت iid دارند ممکن است در حالت غیر iid مضر باشند. کاربرد کورکورانه بوت استرپ ممکن است منجر به نتایج غلط شود. در زیر قسمتی از حوزه وسیع فعالیت بوت استرپ را در برخورد با شاخه‌های مختلف آمار ارائه می‌کنیم که می‌تواند مرجع مناسبی برای خواننده باشد.

الف) مدل‌های خطی. فریدمن [16] بوت استرپ مدل‌های رگرسیونی را در دو حالت بردار x ثابت و تصادفی بررسی کرده است. او همچنین در [17] بوت استرپ برآوردهای حداقل مربعات دو مرحله‌ای در مدل‌های خطی شامل مدل رگرسیونی، مدل‌های یویا و مدل‌های اقتصادسنجی را ارائه کرده است. برای دیدن برخی نتایج تجربی در زمینه بوت استرپ معادلات رگرسیونی می‌توان به [18] مراجعه کرد.

ب) مدل غیر خطی. بوت استرپ رگرسیون ناپارامتری [21]، فواصل اطمینان بوت استرپ در رگرسیون ناپارامتری [20] و نرخ همگرایی بوت استرپ در رگرسیون ناپارامتری [5]، همچنین فواصل اطمینان بوت استرپ مدل کاکس

الگوریتم بوت استرپ

مرحله ۱- نمونه بوت استرپ $X_1^*, \dots, X_n^*$ را به روش نمونه‌گیری تصادفی ساده با جایگذاری از نمونه مشاهده شده $x_1, \dots, x_n$ به دست می‌آوریم.

مرحله ۲- آماره بوت استرپ $T^* = T(X_1^*, \dots, X_n^*)$ را محاسبه می‌کنیم.

مرحله ۳- مراحل ۱ و ۲ را B بار تکرار کرده و B آماره بوت استرپ $\{T_b^*; b = 1, \dots, B\}$ را محاسبه می‌کنیم. برآورد کننده بوت استرپ واریانس و توزیع نمونه‌ای T_n به صورت زیر ارائه می‌شود:

$$\frac{1}{B-1} \sum_{b=1}^B (T_b^* - \frac{1}{B} \sum_{j=1}^B T_j^*)^2,$$

$$\# \{T_b^*; b = 1, \dots, B\} / B.$$

۵ مثالهای کاربردی

در این بخش برای آشنایی بیشتر با روش و الگوریتم بوت استرپ تعدادی مثال کاربردی ارائه می‌کنیم. محاسبات و نمودارها با استفاده از برنامه‌های نوشته شده در نرم‌افزار S-PLUS انجام شده است.

۱.۵ حالت یک متغیره، میانگین و میانه

در یک تحقیق اثر نوعی دارو را بر روی هفت موش آزمایش کرده‌اند که نتایج زیر به دست آمده است:

$$(x_1, \dots, x_7) = (94, 197, 16, 38, 99, 141, 23).$$

میانگین، خطای معیار و میانه نمونه برابر است با:

$$\bar{x} = 86.86, \quad \hat{\sigma} = 61.81, \quad \tilde{x} = 94$$

هدف برآورد انحراف معیار آماره‌های $\bar{X}$ و $\bar{X}$ و در صورت امکان محاسبه فاصله اطمینان برای میانگین μ و میانه m جامعه با استفاده از الگوریتم بوت استرپ است.

مرحله ۱- از نمونه مشاهده شده $(x_1, \dots, x_7)$ به روش نمونه‌گیری تصادفی ساده با جایگذاری نمونه بوت استرپ $(x_1^*, \dots, x_7^*)$ را به دست می‌آوریم. برای مثال ممکن است نمونه زیر را به دست آوریم:

$$(x_1^*, \dots, x_7^*) = (x_5, x_7, x_5, x_2, x_7, x_2, x_1) \\ = (99, 23, 99, 38, 23, 16, 94).$$

مرحله ۲- میانگین و میانه نمونه بوت استرپ $(x_1^*, \dots, x_7^*)$ مرحله ۱ را به دست می‌آوریم،

$$\bar{x}^* = 56, \quad \tilde{x}^* = 38.$$

است که برآورد کننده γ_n^* سازگار باشد، یعنی با فرض اینکه $\gamma_n(F) \rightarrow \gamma(F)$ باید $\gamma_n^* = \gamma_n(F_n) \rightarrow \gamma(F)$.

مرحله ۳- برآورد کننده بوت استرپ. برآورد کننده بوت استرپ γ_n^* محاسبه شده در مرحله ۲ در بیشتر حالتها به خصوص در مورد توزیع نمونه‌ای عبارت صریحی نیست و بیشتر برای محاسبات نظری و استنباط‌های مجانبی مفید است. در نتیجه از روش مونت‌کارلو برای تقریب برآورد کننده‌های بوت استرپ γ_n^* مرحله ۲ استفاده می‌کنیم. به این صورت که، مراحل ۱ و ۲ را B بار تکرار کرده B (در عمل بزرگ انتخاب می‌کنیم) نمونه بوت استرپ $\{(X_{1b}^*, \dots, X_{nb}^*); b = 1, \dots, B\}$ را تولید و B آماره بوت استرپ $\{T_{n,b}^* = T_n(X_{1b}^*, \dots, X_{nb}^*); b = 1, \dots, B\}$ را به دست می‌آوریم. اکنون γ_n^* را با استفاده از روش مونت‌کارلو به صورت $\hat{\gamma}_n^*$ تقریب می‌کنیم. در روش بوت استرپ باید به نکات زیر توجه کرد.

نکته ۱- نمونه‌گیری iid از F_n در واقع یک اندازه احتمال، امید ریاضی و واریانس شرطی به شرط $X_1, \dots, X_n$ به صورت $E^*(\cdot), P^*(\cdot)$ و $\text{var}^*(\cdot)$ تعریف می‌کند. قسمت شرط را برای سادگی نمادها حذف کرده‌ایم و در واقع به صورت $E^*(\cdot | X_1, \dots, X_n), P^*(\cdot | X_1, \dots, X_n)$ و $\text{var}^*(\cdot | X_1, \dots, X_n)$ هستند.

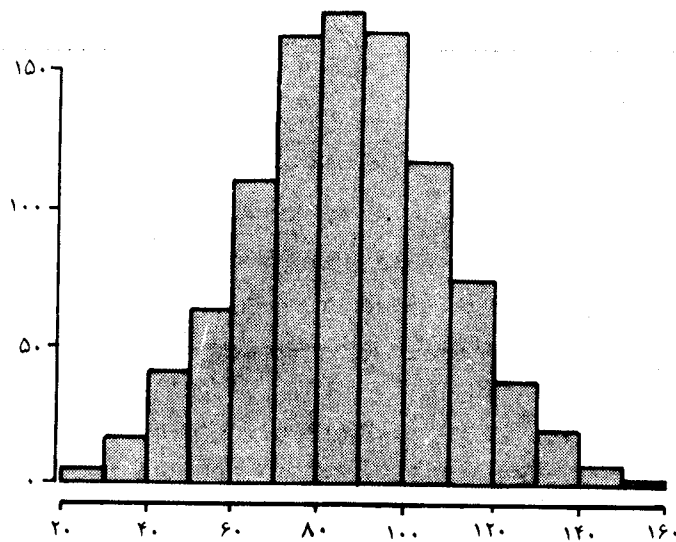
نکته ۲- مقدار B را در عمل می‌توان بسیار بزرگ انتخاب کرد. مقدار B برای برآورد اندازه‌های دقت بین ۵۰ تا ۲۰۰ و برای برآورد توزیع نمونه‌ای بین ۲۰۰ تا ۱۰۰۰ پیشنهاد شده است. با استفاده از قانون قوی اعداد بزرگ وقتی که $B \rightarrow \infty$ ، آنگاه $\gamma_n^* \xrightarrow{a.s.} \gamma_n^*$.

نکته ۳- متغیر تصادفی X را یک متغیره فرض کرده‌ایم که می‌تواند p متغیره هم فرض شود. در نتیجه آماره T_n هم می‌تواند p متغیره باشد.

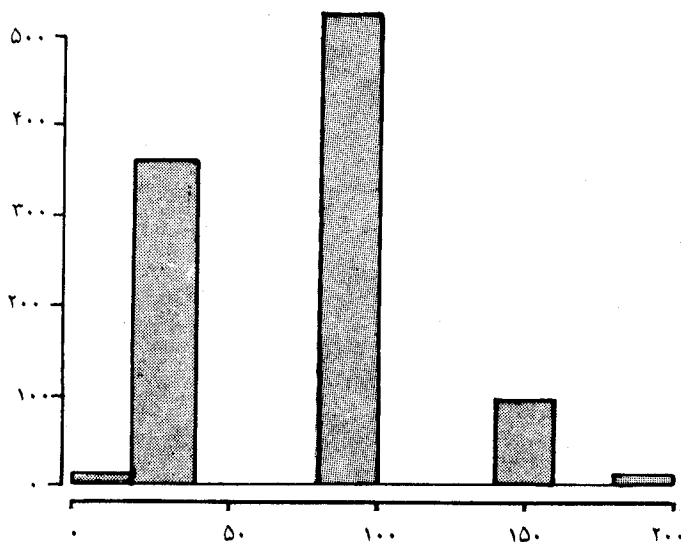
نکته ۴- مراحل سه‌گانه روش بوت استرپ برای ساختمان داده‌های پیچیده‌تر با اندکی تغییر قابل اجرا است. روش بوت استرپ در مشاهدات وابسته، تنها مرحله اول یعنی تولید نمونه بوت استرپ تغییر خواهد کرد. و دو مرحله بعدی مشابه است.

نکته ۵- برآوردهای بوت استرپ توزیع نمونه‌ای آماره T_n برای ساختن فواصل اطمینان پارامتر θ مفید هستند. در نتیجه می‌توان از یافتن آماره بوت استرپ $\{T_{n,b}^*; b = 1, \dots, B\}$ برای محاسبه فاصله اطمینان صدکی استفاده کرد.

اکنون مراحل سه‌گانه روش بوت استرپ را به صورت الگوریتم زیر ارائه می‌کنیم که برای برنامه‌سازی با استفاده از کامپیوتر نیز مفید است.



شکل ۱- بافتنگار بوت‌استرپ میانگین.



شکل ۲- بافتنگار بوت‌استرپ میانه.

فاصله اطمینان ۰/۹۵ میانگین جامعه با استفاده از توزیع t -استودنت بدون استفاده از بوت‌استرپ به صورت زیر است:

$$\mu \in [\bar{x} \pm t_{n-1, 0.025} \frac{s}{\sqrt{n}}] = [86.86 \pm 2.447 (\frac{66.77}{\sqrt{V}})] \\ = [25.11, 148.61].$$

همچنین فاصله اطمینان ۰/۹۵ میانگین جامعه را با استفاده از فاصله اطمینان صدکی بوت‌استرپ محاسبه می‌کنیم. یعنی صدک‌های ۰/۰۲۵ و ۰/۹۷۵ بافتنگار شکل ۱ را به دست می‌آوریم، به عبارت دیگر اگر $B = 1000$ آماره بوت‌استرپ $\bar{x}_1^*, \dots, \bar{x}_B^*$ را مرتب کنیم، فاصله اطمینان صدکی ۰/۹۵ میانگین جامعه به صورت زیر است:

مرحله ۳- مراحل ۱ و ۲ را B با تکرار کرده، در نتیجه B آماره بوت‌استرپ $\bar{x}_b^*$ و $\bar{x}_b^*$ را به دست می‌آوریم. اکنون انحراف معیار میانگین و میانه را با استفاده از روش مونت‌کارلو به صورت زیر تقریب می‌کنیم:

$$\widehat{sd}^*(\bar{x}^*) = \left\{ \frac{1}{B-1} \sum_{b=1}^B (\bar{x}_b^* - \frac{1}{B} \sum_{j=1}^B \bar{x}_j^*)^2 \right\}^{1/2},$$

$$\widehat{sd}^*(\tilde{x}^*) = \left\{ \frac{1}{B-1} \sum_{b=1}^B (\tilde{x}_b^* - \frac{1}{B} \sum_{j=1}^B \tilde{x}_j^*)^2 \right\}^{1/2}$$

جدول ۱ برآورد بوت‌استرپ انحراف معیار میانگین و میانه را به ازای مقادیر مختلف B نشان می‌دهد. مشاهده می‌شود که میانگین به دلیل داشتن انحراف معیار کوچکتر از دقت بالاتری نسبت به میانه برخوردار است.

B	۵۰	۱۰۰	۲۵۰	۵۰۰	۱۰۰۰	∞
$\widehat{sd}^*(\bar{x}^*)$	۲۱/۱۹	۲۵/۶۹	۲۰/۴۳	۲۲/۷۸	۲۲/۵۷	۲۳/۳۶
$\widehat{sd}^*(\tilde{x}^*)$	۳۵/۵۶	۳۸/۲۴	۳۵/۶۲	۳۸/۲۹	۳۸/۴۴	۳۸/۵۱

جدول ۱- برآورد بوت‌استرپ انحراف معیار میانگین و میانه.

با استفاده از ادامه مثال ۱ بخش ۱-۲ عبارت صریحی برای برآورد بوت‌استرپ انحراف معیار میانگین می‌توان به دست آورد. در این صورت ستون آخر جدول ۱ به دست می‌آید، وقتی که $B \rightarrow \infty$.

$$\widehat{sd}^*(\bar{X}^*) \xrightarrow{a.s.} sd^*(\bar{x}^*) = \frac{\hat{\sigma}}{\sqrt{n}} = 23.36.$$

همچنین با استفاده از مثال ۴ بخش ۱-۲ و رابطه (۶) احتمالهای p_k ($k = 1, \dots, 7$) را به صورت زیر به دست می‌آوریم:

$$(p_1, \dots, p_7) = (0.0102, 0.0981, 0.2386, 0.3062, 0.2386, 0.0981, 0.0102).$$

در نتیجه برآورد بوت‌استرپ انحراف معیار میانه به صورت زیر محاسبه می‌شود:

$$sd^*(\tilde{x}^*) = \left\{ \sum_{k=1}^7 p_k (x_{(k)})^2 - \left(\sum_{j=1}^7 p_j x_{(j)} \right)^2 \right\}^{1/2} = 38.51.$$

در این صورت ستون آخر جدول ۱ به دست می‌آید، وقتی که $B \rightarrow \infty$.

$$\widehat{sd}^*(\tilde{X}^*) \xrightarrow{a.s.} sd^*(\tilde{x}^*) = 38.51.$$

شکلهای ۱ و ۲ بافتنگار بوت‌استرپ میانگین و میانه مرحله ۳ را به ازای $B = 1000$ تکرار نشان می‌دهد. واضح است که شکل ۱ به توزیع نرمال بسیار نزدیک است، این نزدیکی در بوت‌استرپ میانگین بخش ۱-۳ نیز نشان داده شده است. در بوت‌استرپ کردن میانه بخش ۱-۳ شرط نزدیکی بافتنگار بوت‌استرپ به توزیع نرمال را پیوستگی مشتق f در همسایگی m ذکر کردیم که واضح است در نمونه گسسته $(x_1, \dots, x_7)$ این شرط برقرار نیست و در نتیجه شکل ۲ گسسته است.

هدف برآورد اریبی و انحراف معیار آماره r و همچنین محاسبه فاصله اطمینان برای ضریب همبستگی ρ جامعه با استفاده از الگوریتم بوت استرپ است. باید متوجه باشیم که متغیر X دو بعدی به صورت زوج نمرات درک مطلب و دستور زبان (C, G) است.

مرحله ۱- نمونه تصادفی مشاهده شده را به صورت زوجی $x_i = (c_i, g_i), i = 1, \dots, 15$ در نظر می‌گیریم. از نمونه زوجی $(x_1, \dots, x_{15})$ به روش نمونه‌گیری تصادفی ساده با جایگذاری، نمونه زوجی بوت استرپ $(x_1^*, \dots, x_{15}^*)$ را به دست می‌آوریم.

مرحله ۲- ضریب همبستگی نمونه زوجی بوت استرپ $(x_1^*, \dots, x_{15}^*)$ را به صورت $r^* = \text{Corr}^*(c^*, g^*)$ محاسبه می‌کنیم.

مرحله ۳- مراحل ۱ و ۲ را B بار تکرار کرده و B ضریب همبستگی بوت استرپ $r_1^*, \dots, r_B^*$ را به دست می‌آوریم. اریبی و انحراف معیار ضریب همبستگی نمونه را با استفاده از روش مونت کارلو به صورت زیر تقریب می‌کنیم:

$$\widehat{\text{bias}}^*(r^*) = \frac{1}{B} \sum_{b=1}^B r_b^* - r,$$

$$\widehat{\text{sd}}^*(r^*) = \left\{ \frac{1}{B-1} \sum_{b=1}^B (r_b^* - \frac{1}{B} \sum_{j=1}^B r_j^*)^2 \right\}^{1/2}$$

جدول ۲ برآورد بوت استرپ اریبی و انحراف معیار ضریب همبستگی را به ازای مقادیر مختلف B نشان می‌دهد.

B	۵۰	۱۰۰	۲۵۰	۵۰۰	۱۰۰۰	۱۰۰۰ R.S.
$\widehat{\text{bias}}^*(r^*)$	-۰٫۰۰۰۱	-۰٫۰۰۱۴	۰٫۰۰۱۵	۰٫۰۰۰۲	-۰٫۰۰۰۱	-۰٫۰۰۱۲
$\widehat{\text{sd}}^*(r^*)$	۰٫۱۴۷	۰٫۱۴۶	۰٫۱۴۳	۰٫۱۴۹	۰٫۱۵۲	۰٫۲۰۲

جدول ۲- برآورد بوت استرپ اریبی و انحراف معیار ضریب همبستگی.

در این مثال چون نمرات جامعه $N = 98$ دانشجو را در اختیار داریم با استفاده از روش و نمونه‌گیری مونت کارلو، ۱۰۰۰ بار نمونه تصادفی زوجی به حجم $n = 15$ را تولید و در نتیجه ۱۰۰۰ ضریب همبستگی نمونه $r_1, \dots, r_{1000}$ را محاسبه می‌کنیم. در نتیجه با استفاده از روش مونت کارلو اریبی و خطای معیار ضریب همبستگی را به صورت زیر برآورد می‌کنیم (ستون آخر جدول ۲ مقادیر زیر هستند):

$$\widehat{\text{bias}}(r) = \frac{1}{1000} \sum_{i=1}^{1000} r_i - \rho = -۰٫۰۰۱۲,$$

$$\widehat{SE} = \left\{ \frac{1}{999} \sum_{i=1}^{1000} (r_i - \frac{1}{1000} \sum_{j=1}^{1000} r_j)^2 \right\}^{1/2} = ۰٫۲۰۲.$$

$$\mu \in [\bar{x}_{(15)}^*, \bar{x}_{(15)}^*] = [۴۱٫۷۵, ۱۳۰٫۸۶].$$

فاصله اطمینان صدکی بوت استرپ کوتاه‌تر از فاصله اطمینان μ استودنت است. در نتیجه بادقت بیشتری μ را شامل می‌شود. فاصله اطمینان بوت استرپ برای میانه جامعه را با استفاده از محاسبات مثال ۴ به صورت زیر ارائه می‌کنیم. چون

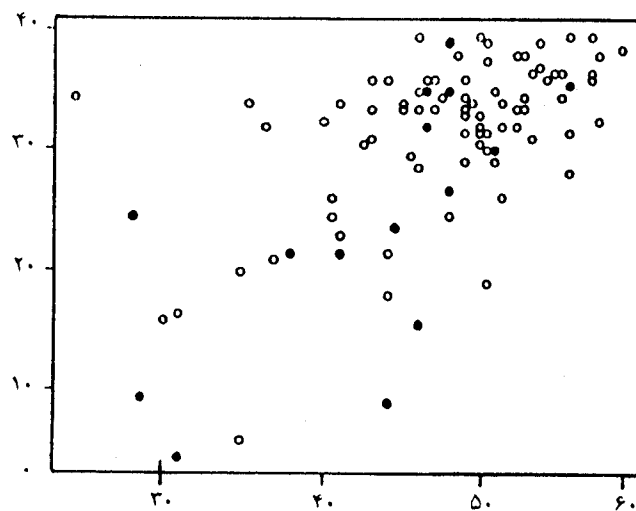
$$P^*\{x_{(1)} < \bar{X}^* < x_{(7)}\} = 1 - p_1 - p_7 \cong ۰٫۹۸,$$

$$P^*\{x_{(7)} < \bar{X}^* < x_{(9)}\} = p_7 - p_7 - p_9 \cong ۰٫۷۸$$

در نتیجه فاصله اطمینان بوت استرپ ۰٫۹۸ و ۰٫۷۸ برای میانه جامعه به صورت [۲۳، ۹۹] و [۳۸، ۹۹] است.

۲.۵ حالت دو متغیره، ضریب همبستگی

در زمستان سال ۱۳۷۳ تحقیقی با همکاری آقای دکتر مشیری استاد درس زبان انگلیسی در دانشکده الکترونیک دانشگاه صنعتی شریف انجام گردید که در این مثال و مثال بعد از اطلاعات این تحقیق استفاده شده است. از بین $N = 98$ دانشجوی رشته الکترونیک دانشگاه صنعتی شریف، برای مطالعه ارتباط نمرات درک مطلب (c) و دستور زبان (g) درس زبان انگلیسی، نمونه‌ای تصادفی به حجم $n = 15$ انتخاب کرده‌ایم. نمرات جامعه با $(+)$ و $(-)$ و نمونه با $(\circ)$ در شکل ۳ نشان داده شده است، محور افقی نمرات درک مطلب و محور عمودی نمرات دستور زبان دانشجویان است.



شکل ۳- نمرات درس زبان انگلیسی دانشجویان.

ضریب همبستگی جامعه و نمونه نمرات به صورت زیر است:

$$\rho = \text{Corr}(c, g) = ۰٫۵۹۷, \quad r = \widehat{\text{Corr}}(c, g) = ۰٫۶۶۳.$$

۳.۵ حالت چند متغیره، بزرگترین مقدار ویژه ماتریس کواریانس

از بین دانشجویان دانشگاه صنعتی شریف $n = 98$ دانشجو به تصادف انتخاب و نمرات میان ترم و پایان ترم درس زبان انگلیسی شامل سه قسمت درک مطلب، لغت و دستور زبان را ثبت کرده‌ایم.

مجموعه داده‌ها یک ماتریس 98×6 است که هر سطر آن برداری به صورت $x_i = (x_{i1}, \dots, x_{i6})$, $(i = 1, 2, \dots, 98)$ به ترتیب شامل نمرات درک مطلب، لغت و دستور زبان میان ترم و درک مطلب، لغت و دستور زبان پایان ترم است. بردار میانگین و ماتریس کواریانس نمونه را که برآورد جانشینی بردار میانگین μ و ماتریس کواریانس Σ است به صورت زیر محاسبه کرده‌ایم:

$$\bar{x} = (14,78, 16,37, 14,27, 16,35, 13,35, 15,81).$$

$$S = \begin{pmatrix} 7,42 & & & & & \\ 1,61 & 7,05 & & & & \\ 5,42 & 5,27 & 25,62 & & & \\ 3,94 & 2,21 & 6,21 & 8,12 & & \\ 4,38 & 5,55 & 9,55 & 5,77 & 13,99 & \\ 4,92 & 4,25 & 13,61 & 5,75 & 8,29 & 13,48 \end{pmatrix}$$

همچنین مقادیر ویژه ماتریس S به صورت زیر است:

$$(\hat{\lambda}_1, \dots, \hat{\lambda}_6) = (47,49, 10,36, 6,75, 3,86, 3,76, 3,46).$$

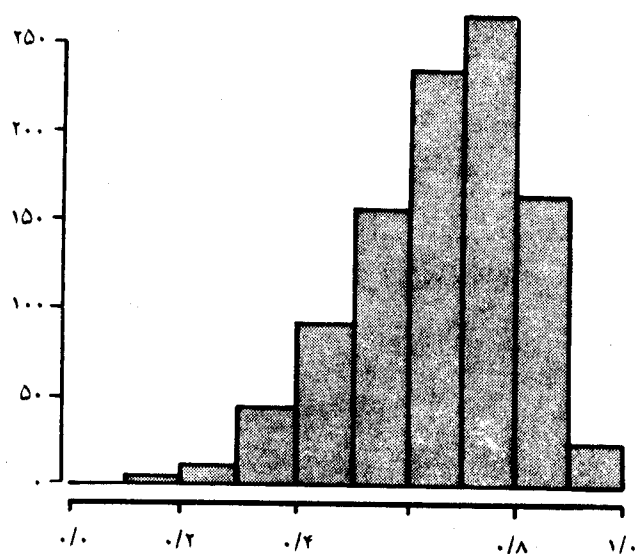
پارامتر $\theta = \lambda_1 / \sum_{i=1}^6 \lambda_i = 0,6275$ را با برآورد $\hat{\theta} = \hat{\lambda}_1 / \sum_{i=1}^6 \hat{\lambda}_i$ در نظر می‌گیریم. هدف برآورد اریبی و انحراف معیار آماره $\hat{\theta}$ و همچنین محاسبه فاصله اطمینان برای θ جامعه با استفاده از الگوریتم بوت‌استرپ است.

مرحله ۱- از نمونه تصادفی شش متغیره مشاهده شده $(x_1, \dots, x_{98})$ به روش نمونه‌گیری تصادفی ساده با جایگذاری نمونه بوت‌استرپ شش متغیره $(x_1^*, \dots, x_{98}^*)$ را به دست می‌آوریم.

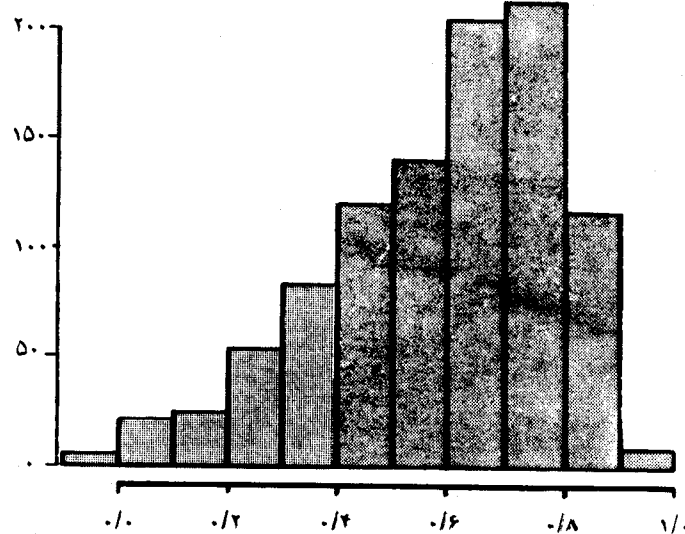
مرحله ۲- ماتریس کواریانس نمونه بوت‌استرپ $(x_1^*, \dots, x_{98}^*)$ را به دست آورده و مقادیر ویژه $\lambda_1^*, \dots, \lambda_6^*$ ماتریس S^* را محاسبه می‌کنیم و آنگاه $\theta^* = \lambda_1^* / \sum_{i=1}^6 \lambda_i^*$ را به دست می‌آوریم.

مرحله ۳- مراحل ۱ و ۲ را $B = 1000$ بار تکرار کرده $\theta_1^*, \dots, \theta_B^*$ را به دست می‌آوریم. اریبی و انحراف معیار برآورد کننده $\hat{\theta}$ را با استفاده از روش مونت‌کارلو به صورت زیر تقریب می‌کنیم:

شکل ۴ بافتگر بوت‌استرپ ضریب همبستگی حاصل از ۱۰۰۰ تکرار $\theta_1^*, \dots, \theta_{1000}^*$ را نشان می‌دهد. شکل ۵ بافتگر شبیه‌سازی ضریب همبستگی حاصل از ۱۰۰۰ تکرار $\theta_1^*, \dots, \theta_{1000}^*$ نمونه تصادفی است. واضح است که شکل ۴ تقریب بسیار خوبی از شکل ۵ است.



شکل ۴- بافتگر بوت‌استرپ ضریب همبستگی.

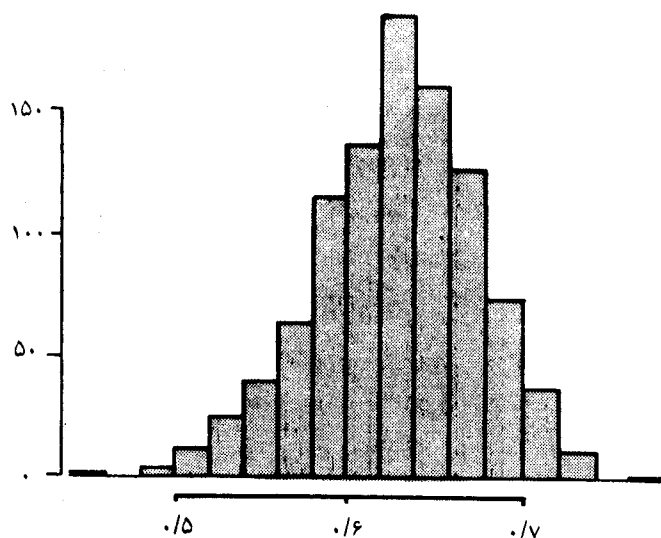


شکل ۵- بافتگر شبیه‌سازی ضریب همبستگی.

فاصله اطمینان $0,95$ ضریب همبستگی جامعه را با استفاده از فاصله اطمینان صدکی بوت‌استرپ محاسبه می‌کنیم. یعنی صدکهای $0,025$ و $0,975$ بافتگر شکل ۴ را به دست می‌آوریم، به عبارت دیگر اگر 1000 آماره بوت‌استرپ $\theta_1^*, \dots, \theta_{1000}^*$ را مرتب کنیم، فاصله اطمینان صدکی بوت‌استرپ $0,95$ میانگین جامعه به صورت زیر است:

$$[r_{(25)}^*, r_{(975)}^*] = [0,3222, 0,899].$$

دیده می‌شود که ضریب همبستگی جامعه $\rho = 0,597$ به فاصله اطمینان بالا خلق دارد.

شکل ۶- بافتنگار بوت استرپ θ^*

۰٫۹۵ تقریبی برای θ به صورت زیر محاسبه می‌شود:

$$\theta \in [\hat{\theta}(1 \pm z_{0.975} \sqrt{\frac{2}{n}})] = [0.628(1 \pm 1.96 \sqrt{\frac{2}{98}})] \\ = [0.452, 0.804].$$

فاصله‌های اطمینان بوت استرپ کوتاه‌تر از تقریب نرمال هستند، در نتیجه با دقت بیشتری θ را در بر می‌گیرند.

$$\widehat{bias}^*(\theta^*) = \frac{1}{1000} \sum_{b=1}^{1000} \theta_b^* - \hat{\theta} \\ = 0.6279 - 0.6275 = 0.0004 \approx 0,$$

$$\widehat{sd}^*(\theta^*) = \left\{ \frac{1}{999} \sum_{b=1}^{1000} (\theta_b^* - \frac{1}{1000} \sum_{j=1}^{1000} \theta_j^*)^2 \right\}^{1/2} = 0.046.$$

شکل ۶ بافتنگار بوت استرپ θ^* حاصل از ۱۰۰۰ تکرار $\theta_1^*, \dots, \theta_{1000}^*$ مرحله ۳ را نشان می‌دهد، واضح است که شکل ۶ به توزیع نرمال بسیار نزدیک است. در نتیجه فاصله اطمینان ۰٫۹۵ برای θ را با توجه به توزیع نرمال θ^* می‌توان به صورت زیر محاسبه کرد:

$$\theta \in [\hat{\theta} \pm z_{0.975} \widehat{sd}^*(\theta^*)] = [0.628 \pm 1.96(0.046)] \\ = [0.538, 0.718].$$

همچنین فاصله اطمینان ۰٫۹۵ برای θ را با استفاده از فاصله اطمینان صدکی بوت استرپ یعنی با استفاده از ۱۰۰۰ تکرار $\theta_1^*, \dots, \theta_{1000}^*$ مرحله ۳ به صورت زیر محاسبه می‌کنیم:

$$\theta \in [\theta_{(25)}^*, \theta_{(975)}^*] = [0.527, 0.709].$$

در آنالیز چند متغیره با فرض اینکه $X \sim N_6(\mu, \Sigma)$ است، فاصله اطمینان

مراجع

- [۱] ایران پناه، نصراله (۱۳۷۵). الگوریتم بوت استرپ در سریهای زمانی و مشاهدات وابسته، رساله کارشناسی ارشد، گروه آمار، موسسه ریاضیات دکتر مصاحب، دانشگاه تربیت معلم.
- [2] Bickel, P.J and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap, *Ann. Statist.*, 9, 1196-1217.
- [3] Bickel, P. J. and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling, *Ann. Statist.*, 12, 470-482.
- [4] Burr, D.(1994). A comparison of certain bootstrap confidence intervals in the Cox model, *J. Amer. Statist. Assoc.*, 89, 1290-1302.
- [5] Cao-Abad, R.(1991). Rate of convergence for the wild bootstrap in nonparametric regression, *Ann. Statist.*, 19, 2226-2231.
- [6] Chatterjee, S.(1984). Variance estimation in factor analysis, an application of bootstrap, *British J. Math. statist. Psycho.*, 37, 252-262.
- [7] Chen, J. and Sitter, R.R. (1993). Edgeworth expansions and the bootstrap for stratified sampling without replacement from a finite population, *Canadian J. statist.*, 21, 347-357.
- [8] Chen, Y. and Tu, D.(1987). Estimating the error rate in discriminate analysis: By the delta, jackknife and bootstrap methods, *chinese J. Appl. Prob. Statist.*, 3, 203-210.
- [9] Diccio, T.J. and Efron, B.(1992). More accurate confidence intervals in exponential families, *Biometrika*, 79, 231-245.

- [10] Efraon, B. (1981). Nonparametric standard errors and confidence intervals (with discussions), *Canadian J. statist.*, 9, 139-172.
- [11] Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems, *Biometrika*, 72, 45-58.
- [12] Efron, B. (1987). Better bootstrap confidence intervals (with discussions) *J. Amer. Statist. Assoc.*, 82, 171-200.
- [13] Efron, B. (1979). Bootstrap methods : Another look at the jackknife, *Ann. Statist.*, 7, 1-26.
- [14] Efron, B. and Tibshirani, R.J. (1986). Bootstrap methods for standard errors, Confidence intervals, and other measures of statistical accuracy, *Statist. Science*, 1, 54-77.
- [15] Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.
- [16] Freedman, D.A. (1981). Bootstrapping regression models, *Ann. statist.*, 9, 1218-1228.
- [17] Freedman, D.A. (1984). On bootstrapping two-stage least squares estimates in stationary linear models, *Ann. Statist.*, 12, 827-842.
- [18] Freedman, D.A. and Peters, S.C. (1984). Bootstrapping a regression equation: Some empirical results, *J. Amer. Statist. Assoc.*, 79, 97-106.
- [19] Gu, M. (1992). On the Edgeworth expansion and bootstrap approximation for the Cox regression model under random censorship, *Canadian J. Statist.*, 20, 399-414.
- [20] Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression, *Ann. Statist.*, 20, 695-711.
- [21] Hardle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands, *J. Amer. Statist. Assoc.*, 83, 102-110.
- [22] Jhun, M. (1990). Bootstrapping K-means clustering, *J. Japanese Soc. Compu. Statist.*, 3, 1-14.
- [23] Kuk, A. Y. C. (1989). Double bootstrap estimation of variance under systematic sampling with probability proportional to size, *J. Statist. Compu. Simul.*, 31, 73-82.
- [24] Lehmann, E.L. (1983). *Theory of Point Estimation*, Wiley, New York.
- [25] Laird, N.M. and Louis, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussions), *J. Amer. Statist. Assoc.*, 82, 739-757.
- [26] Lee, K.W. (1990). Bootstrapping logistic regression models with random regressors, *Comm. Statist. A*, 19, 2527-2539.
- [27] Newton, M. A. and Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussions), *J.R. Statist. Soc. B*, 56, 3-48.
- [28] Peck, R., Fisher, L. and Van Ness, J. (1989). Bootstrap confidence intervals for the numbers of clusters in cluster analysis, *J. Amer. Statist. Assoc.*, 84, 184-191.
- [29] Rubin, D.B. (1981). The Bayesian bootstrap, *Ann. Statist.*, 9.
- [30] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.*, 9, 1187-1195.