آشنایی با الگوریتم بوت‌استرپ

نصراله ایران پناه*
عین‌اله پاشا†

چکیده

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

۱ پیشگفتار

نظریهٔ 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 خلاصهٔ اطلاعات چقدر است؟

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

فرض کنید $X_1, \ldots, X_n$ نمونهٔ تصادفی مستقل و همتوزیع (iid) از توزیع نامعلوم F و $T_n = T(X_1, \ldots, X_n)$ برآورد کنندهٔ پارامتر نامعلوم θ است. هدف برآورد اندازه‌های دقت آمارهٔ T_n از جمله اریبی و واریانس یا برآورد توزیع نمونه‌ای آمارهٔ 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 که‌محاسبهٔ آن گاهی غیر ممکن است. برآورد توزیع نمونه‌ای نیز عموماً مبتنی به قضایای حدی است و بنابراین برای حجم نمونهٔ بزرگ اعتبار دارد.

روشهای باز نمونه‌گیری[1]، دستیابی به هدف بالا را بدون ضعفهای دیدگاه سنتی میسر می‌سازند. از جمله روشهای بازنمونه‌گیری روش بوت‌استرپ[2] است که بوسیلهٔ افرون[3] [13] در سال ۱۹۷۹ ارائه شده‌است.

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 جانشینی[4] و غیره) انتخاب می‌شود. قبل از انتخاب داده‌ها، یک آماره یک متغیر تصادفی و دارای یک توزیع احتمال معینی است که‌توزیع نمونه‌ای آماره نامیده می‌شود. اکثر روشهای آماری برای تحلیل نیاز به بعضی اطلاعات از توزیع نمونه‌ای دارند.

از طرف دیگر، در یک مسألهٔ برآورد به علت اینکه هر برآورد کننده ممکن است یک خطای برآورد داشته باشد، داشتن یک اندازهٔ 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 برآوردکننده در میان یک رده از برآورد کننده‌های مناسب استفاده شوند.

توزیع نمونه‌ای و مشخصه‌های آن معمولاً به جامعهٔ 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 بوت‌استرپ و جک‌نایف[5] روشهایی

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

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

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 اندازهٔ دقت در نظر می‌گیریم. فرض کنید $X_1,\ldots,X_n$ متغیرهای تصادفی iid با مقادیر حقیقی از یک توزیع نامعلوم F باشند. آمارهٔ $T_n = T(X_1,\ldots,X_n)$ را در نظر می‌گیریم، واریانس T_n به صورت زیر است:

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

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

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

$$T_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$
$$\operatorname{var}(\bar{X}_n) = \frac{1}{n}\operatorname{var}(X_1).$$

بنابراین، می‌توان $\operatorname{var}(\bar{X}_n)$ را با $\operatorname{var}(X_1)$ به صورت یک پارامتر نامعلوم برآورد کرد. اگر F متعلق به خانوادهٔ پارامتری نباشد، آنگاه $\operatorname{var}(X_1)$ معمولاً با استفاده از واریانس نمونهٔ 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، $\operatorname{var}(T_n)$ بدون هیچ فرمول پیچیده‌ای به دست می‌آید. برای مثال، اگر $T_n = \bar{X}_n^2$، آنگاه

$$\operatorname{var}(\bar{X}_n^2) = \frac{4\mu^2\alpha_2}{n} + \frac{4\mu\alpha_3}{n^2} + \frac{\alpha_4}{n^3}, \qquad (2)$$

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

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

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

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

$$\lim_{n\to\infty}[n\operatorname{var}(T_n)] = \sigma_F^2, \qquad (3)$$

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 $\operatorname{var}(T_n)$ را با مقایسهٔ تجربی تقریب یعنی $\sigma^2(\hat{F})/n$ یا $\sigma^2(\hat{\gamma})/n$ ($\hat{F}$ و $\hat{\gamma}$ برآوردکننده‌های F و γ هستند) برآورد می‌کنیم.

مثال ۱ـ (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:

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

مثال ۲ـ میانگین نمونهٔ پیراسته. میانگین نمونهٔ αـپیراستهٔ $\bar{X}_n^{(\alpha)}$ را در برآورد نیرومند مرکز یک 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 $X_{(i)}$ آمارهٔ مرتب iام است. یک شکل دقیق و صریح برای $\operatorname{var}(\bar{X}_n^{(\alpha)})$ وجود ندارد، اما می‌توان نشان داد(لهمن[6] [24]) که (۳) برقرار است با

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

که در آن

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

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

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\} = \#\{X_i \leq x; i = 1,\ldots,n\}/n,$$

که در آن $I\{A\}$ تابع‌نشانگر مجموعهٔ 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:

6) Lehmann

(۱) چون از فرمول تقریبی σ_F^2 در (۳) استفاده می‌کنیم، در اکثرمواقع نیاز به حجم نمونهٔ n بزرگ برای داشتن واریانس دقیق (یا اندازهٔ دقت دیگر) برآورد کننده‌ها داریم.

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

(۳) با بکار 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 می‌گیریم. 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 بوت‌استرپ

افرون [13] در سال ۱۹۷۹ روش بوت‌استرپ را برای برآورد اندازهٔ دقت و توزیع نمونه‌ای آماره‌ها ارائه کرد. این روش براساس ایدهٔ باز نمونه‌گیری از داده‌ها است، که در چند سال اخیر با استفاده از کامپیوتر گسترش فراوانی یافته است. مباحث تحت بررسی بوت‌استرپ 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 می‌سازند.

۱.۲ برآورد واریانس

فرض کنید $X_1,\ldots,X_n$، متغیرهای تصادفی iid از F هستند و F با $\hat{F}$ برآورد می‌شود. با جانشینی $\hat{F}$ به‌جای F در (۱)، برآورد کنندهٔ بوت استرپ واریانس T_n را به‌دست می‌آوریم،

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

که در آن $\{X_1^*,\ldots,X_n^*\}$ یک نمونهٔ تصادفی iid از $\hat{F}$ است و نمونهٔ بوت‌استرپ نامیده می‌شود و $\mathrm{var}^*[.|X_1,\ldots,X_n]$ واریانس شرطی به شرط $X_1,\ldots,X_n$ است. به بیان ساده‌تر، نمونهٔ بوت‌استرپ $\{X_1^*,\ldots,X_n^*\}$ با نمونه‌گیری تصادفی سادهٔ با جایگذاری از نمونهٔ اولیه $\{X_1,\ldots,X_n\}$ به دست می‌آید. به طور کلی، نمونه‌گیری iid از $\hat{F}$ یک اندازهٔ احتمال، امیدریاضی و واریانس شرطی به شرط $X_1,\ldots,X_n$ تعریف می‌کند که به صورت E^*, P^* و var^* نشان می‌دهیم. برای سادگی، از این به بعد قسمت شرط را حذف می‌کنیم. همچنین $T_n^*=T(X_1^*,\ldots,X_n^*)$ را آمارهٔ بوت‌استرپ می‌نامیم. معادلهٔ (۵) شکل نظری برآورد کنندهٔ واریانس بوت‌استرپ برای T_n است. وقتی $\mathrm{var}^*(T_n^*)$ تابع صریحی از $X_1,\ldots,X_n$ نیست، ممکن است که درکاربردهای عملی مورد استفاده قرار نگیرد. اگر $\mathrm{var}^*(T_n^*)$ یک تابع صریحی از $X_1,\ldots,X_n$ باشد، آنگاه واقعاً یک برآوردکنندهٔ جانشینی $\mathrm{var}(T_n)$ است.

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

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

برآورد کنندهٔ بوت‌استرپ واریانس $\bar{X}_n$ را به صورت زیر به‌دست می‌آوریم:

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

وقتی $T_n=\bar{X}_n^2$، با استفاده از (۲)، $\mu=\int x\,dF(x)$ و $\alpha_k=\int(x-\mu)^k dF(x)$ به دست می‌آوریم:

$$\mathrm{var}^*(\bar{X}_n^{*2})=\frac{4\bar{X}_n^2\tilde{\alpha}_2}{n}+\frac{4\bar{X}_n\tilde{\alpha}_3}{n}+\frac{\tilde{\alpha}_4}{n^3},$$

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

مثال ۴ـ میانهٔ نمونه. فرض کنید $T_n=F_n^{-1}(\frac{1}{2})$ میانهٔ نمونه باشد، که در آن F_n^{-1} با جانشینی F_n به‌جای F در (۴) تعریف می‌شود. اگر فرض کنیم $n=2m-1$، آنگاه $T_n=X_{(m)}$. اگر $\hat{F}=F_n$، برآورد کنندهٔ واریانس بوت‌استرپ $X_{(m)}$ را به صورت زیر به‌دست می‌آوریم. فرض کنید که $X_1^*,\ldots,X_n^*$ یک نمونهٔ تصادفی iid از F_n است و N_i^* تعداد X_iهایی است که در نمونهٔ بوت‌استرپ انتخاب می‌شود،

$$N_i^*=\#\{X_j^*=X_i|X_1,\ldots,X_n\},\qquad i=1,\ldots,n,\ j=1,\ldots,n.$$

واضح است که بردار $N^*=(N_1^*,\ldots,N_n^*)$ دارای توزیع n جمله‌ای است که احتمال $\frac{1}{n}$ را به هر جمله نسبت می‌دهد و بنابراین امید ریاضی انتخاب شدن هر X_i در نمونهٔ بوت‌استرپ یک است. حال آمارهٔ مرتب $X_{(1)}\le X_{(2)}\le\ldots\le X_{(n)}$ و مطابق آن $N_{(1)}^*\le N_{(2)}^*\le\ldots\le N_{(n)}^*$ را در نظر می‌گیریم. احتمالات بوت‌استرپ شرطی به شرط $X_1,\ldots,X_n$ را به ازای $k=1,\ldots,n$ در نظر می‌گیریم،

$$
\begin{aligned}
p_k &= P^*\{X^*_{(m)} = X_{(k)}\}\\
&= P^*\{X^*_{(m)} > X_{(k-1)}\} - P^*\{X^*_{(m)} > X_{(k)}\} \qquad (6)\\
&= P^*\{\sum_{i=1}^{k-1} N^*_{(i)} \leq m-1\} - P^*\{\sum_{i=1}^{k} N^*_{(i)} \leq m-1\}\\
&= P\{binomial(n, \frac{k-1}{n}) \leq m-1\}\\
&- P\{binomial(n, \frac{k}{n}) \leq m-1\}\\
&= \sum_{j=0}^{m-1} \binom{n}{j} \left(\frac{k-1}{n}\right)^j \left(\frac{n-k+1}{n}\right)^{n-j}\\
&- \sum_{j=0}^{m-1} \binom{n}{j} \left(\frac{k}{n}\right)^j \left(\frac{n-k}{n}\right)^{n-j}\\
&= \sum_{j=0}^{m-1} \binom{n}{j} \frac{(k-1)^j(n-k+1)^{n-j} - k^j(n-k)^{n-j}}{n^n}.
\end{aligned}
$$

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 بوت‌استرپ $X^*_{(m)}$ (که از رابطهٔ (۶) به‌دست می‌آید) به‌جای $F_{X_{(m)}}$ نتیجهٔ زیر را به‌دست می‌آوریم،

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

به هر حال، همان طور که قبلاً عنوان شد، رابطه‌های (۱) یا (۵) معمولاً پیچیده هستند و $\text{var}^*(T^*_n)$ تابع صریحی از $X_1, \ldots, X_n$ نیست. در نتیجه از روش مونت کارلو برای تقریب استفاده می‌کنیم.

۲.۲ روش مونت کارلو

وقتی طرف راست (۱) ساده نیست، حتی اگر F معلوم باشد، نمی‌توان $\text{var}(T_n)$ را به طور دقیق محاسبه نمود. در آمار روشی قدیمی به‌نام مونت کارلو وجود دارد که وقتی F معلوم‌است، با استفاده از آن می‌توان $\text{var}(T_n)$ را به‌طور عددی تقریب کرد. یعنی اینکه، به‌طور تکراری مجموعه داده‌های جدیدی از F به‌دست می‌آوریم و آنگاه آمارهٔ T_n را روی هر مجموعهٔ داده محاسبه می‌کنیم، واریانس نمونهٔ مقادیر T_n یک تقریب عددی $\text{var}(T_n)$ است. وقتی $\hat{F}$ معلوم است، این ایده می‌تواند برای تقریب $\text{var}^*(T^*_n)$ استفاده شود. یعنی، با فرض معلوم بودن $X_1, \ldots, X_n$، B نمونهٔ مستقل بوت‌استرپ $\{X^*_{1b}, \ldots, X^*_{nb}\}, b = 1, \ldots, B$ را از $\hat{F}$ به‌دست می‌آوریم. سپس آماره‌های بوت‌استرپ $T^*_{n,b} = T(X^*_{1b}, \ldots, X^*_{nb})$ را محاسبه می‌کنیم، $\text{var}^*(T^*_n)$ با استفاده از تقریب مونت‌کارلو به صورت زیر به‌دست می‌آید:

$$\widehat{\text{var}}^*(T^*_n) = \frac{1}{B-1} \sum_{b=1}^{B} (T^*_{n,b} - \frac{1}{B} \sum_{j=1}^{B} T^*_{n,j})^2. \qquad (7)$$

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

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

هم $\text{var}^*(T^*_n)$ و هم تقریب مونت‌کارلو آن $\widehat{\text{var}}^*(T^*_n)$، برآورد کننده‌های واریانس بوت‌استرپ نامیده می‌شوند. $\widehat{\text{var}}^*(T^*_n)$ اغلب در کاربردهای عملی مفید است، در حالی که در مطالعات نظری معمولاً روی $\text{var}^*(T^*_n)$ دقت می‌کنیم.

بنابراین، روش بوت‌استرپ ترکیبی از دو روش است: اصل جانشینی و تقریب مونت‌کارلو. وقتی $\text{var}^*(T^*_n)$ در (۵) تابع صریحی از $X_1, \ldots, X_n$ است، بوت‌استرپ دقیقاً با دیدگاه جانشینی سنتی منطبق می‌شود، در غیر این صورت بوت استرپ از (۷) برای تقریب استفاده می‌کند. درحالیکه، دیدگاه سنتی ابتدا به طور تحلیلی $\text{var}(T_n)$ را تقریب می‌کند و سپس پارامترهای نامعلوم در فرمول تقریبی $\text{var}(T_n)$ را برآورد می‌کند. این مفهوم بوت‌استرپ کمک می‌کند که آن را به مسائل پیچیده‌تر تعمیم دهیم.

۳.۲ برآورد توزیع نمونه‌ای

فرض کنید $X_1, \ldots, X_n$، متغیرهای تصادفی iid از F هستند، هدف برآورد توزیع نمونه‌ای آمارهٔ $T_n = T(X_1, \ldots, X_n)$ است. چون معمولاً از T_n برای ساختن فواصل اطمینان برای یک پارامتر نامعلوم θ (θ به F وابسته است) استفاده می‌کنیم، به‌جای T_n یک ریشهٔ آن $R_F = R(X_1, \ldots, X_n, F)$ را در نظر می‌گیریم. ریشهٔ R_F ممکن است به صورت $T_n - \theta$ و یا آمارهٔ استودنت شده $(T_n - \theta)/S_n$ باشد، که در آن S_n برآورد کنندهٔ انحراف معیار T_n است. پس نیاز داریم که توزیع نمونه‌ای R_F را به‌دست آوریم،

$$H_F(x) = P\{R(X_1, \ldots, X_n, F) \leq x\}, \qquad (8)$$

R_F و H_F در واقع به n بستگی دارند که برای سادگی نمادها حذف شده‌اند.

در دیدگاه سنتی، ابتدا به‌دنبال یک فرمول نظری سادهٔ دقیق یا تقریبی برای $H_F(x)$ هستیم و سپس پارامترهای نامعلوم در فرمول نظری را با برآوردهای‌آن جانشین می‌کنیم. برای مثال، وقتی $R_F = T_n - \theta$، در بعضی از حالتها $H_F(x)$ با $\Phi(\sqrt{n}x/\sigma)$ تقریب می‌شود، که در آن $\Phi(.)$ توزیع نرمال استاندارد و σ یک پارامتر نامعلوم وابسته به F است. اگر $\hat{\sigma}$ یک برآورد کنندهٔ σ باشد، آنگاه $H_F(x)$ با $\Phi(\sqrt{n}x/\hat{\sigma})$ برآورد می‌شود. برای $R_F = (T_n - \theta)/S_n$، در بعضی از حالتها $H_F(x)$ با $\Phi(x)$ تقریب می‌شود. این دیدگاه، به هر حال ضعفها و معایبی دارد که در بخش ۱ـ۲ توضیح داده شد.

مشابه دستورالعملی که برای برآوردکنندهٔ بوت‌استرپ واریانس در بخش ۱ـ۲ و تقریب آن در بخش ۲ـ۲ دادیم، اینک برآوردکنندهٔ بوت‌استرپ $H_F(x)$ را ارائه می‌کنیم. ابتدا $\hat{F}$ را به‌جای F در (۸) قرار می‌دهیم،

$$H^*(x) = H_{\hat F}(x) = P^*\{R(X_1^*,\dots,X_n^*,\hat F) \le x | X_1,\dots,X_n\}, \quad (۹)$$

که در آن $X_1^*,\dots,X_n^*$، iid از $\hat F$ هستند. اگر $H^*(x)$ تابع صریحی از $X_1,\dots,X_n$ باشد، آنگاه برآوردکنندهٔ بوت‌استرپ $H_F(x)$ خواهد بود. در غیر این صورت، می‌توانیم از تقریب مونت‌کارلو استفاده کنیم،

$$\hat H^*(x) = \frac{1}{B}\sum_{b=1}^{B} I\{R(X_{1b}^*,\dots,X_{nb}^*,\hat F)\le x\} \quad (۱۰)$$

که درآن $\{X_{1b}^*,\dots,X_{nb}^*\}, b=1,\dots,B$ نمونه‌های بوت‌استرپ مستقل از $\hat F$ هستند.

۴.۲ کاربردهای دیگر

گاهی اندازهٔ دقت یک آمارهٔ T_n را به صورت مشخصهٔ توزیع نمونه‌ای T_n در نظر می‌گیریم. برآوردکننده‌های بوت‌استرپ اندازهٔ دقت می‌توانند با استفاده از مشخصهٔ H^* یا $\hat H^*$ به‌دست آورده شوند. برای مثال، برآوردکنندهٔ بوت‌استرپ واریانس $\mathrm{var}^*(T_n^*)$ در (۵) [یا $\widehat{\mathrm{var}}^*(T_n^*)$ در (۷)] واریانس توزیع بوت‌استرپ H^* در(۹) [یا $\hat H^*$ در (۱۰)] است با $R_n = T_n$ همچنین برآوردکننده بوت‌استرپ 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 بوت‌استرپ برای یک پارامتر نامعلوم θ می‌تواند با استفاده‌از صدکهای H^* یا $\hat H^*$ به‌دست آید. برای مثال، بافرض $R_n = T_n$، فاصلهٔاطمینان صدکی بوت‌استرپ $(1-2\alpha)$ درصد برای θ، صدکهای αام و $(1-\alpha)$ام H^* یا $\hat H^*$ است. برای توضیح بیشتر در مورد فواصل اطمینان بوت‌استرپ می‌توان به مراجع [10]، [11] و [14] مراجعه کرد. همچنین فواصل اطمینان بوت‌استرپ بهتری در مراجع [12] و [9] ارائه شده‌اند.روش بوت‌استرپ در زمینه‌های دیگری از جمله آزمون فرض، پیش‌بینی و انتخاب مدل مطرح است. برای دیدن یک سری مثالهای کاربردی روش بوت‌استرپ همراه با بعضی الگوریتمهای مفید می‌توان به کتاب مقدمه‌ای بر بوت‌استرپ افرون و تیبشیرانی[7] [15] مراجعه کرد.

۵.۲ علت‌استفاده از تقریب مونت‌کارلو

برآوردکننده‌های بوت‌استرپ (۵) و (۹) چون تابعکی از توزیع تجربی F_nاند، مقادیری معلوم‌اند. چرا از روش مونت‌کارلو برای تقریب آنهااستفاده‌می‌کنیم؟ علت این است که برای nهای بزرگ محاسبهٔ آنها عملاً غیر ممکن است. مشکل بودن محاسبات را برای حالت $n=2$ در مثال زیر نشان می‌دهیم.

مثال ۵ـ درحالت $n=2$، نمونهٔ (X_1,X_2) را بافرض $X_1 > X_2$ در نظر می‌گیریم. F_n جرم $\frac{1}{2}$ را به X_1 و X_2 نسبت می‌دهد. نمونهٔ بوت‌استرپ (X_1^*,X_2^*) با احتمال $\frac{1}{4}$ مقادیر $(X_1,X_1),(X_1,X_2),(X_2,X_1)$ و (X_2,X_2) را می‌گیرد و چون ترتیب اهمیتی ندارد (X_1^*,X_2^*) مقادیر (X_1,X_1)، (X_1,X_2) و (X_2,X_2) را با احتمالهای $\frac{1}{4}$، $\frac{1}{2}$ و $\frac{1}{4}$ می‌گیرد. بوت‌استرپ میانگین بخش ۳ـ۱ 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 $\gamma(F) = P(R\le x)$ با استفاده از بوت‌استرپ است. چون $\bar X^*$ مقادیر X_1، $\bar X$ و X_2 را با احتمالهای $\frac{1}{4}$، $\frac{1}{2}$ و $\frac{1}{4}$ می‌گیرد، پس $R^* = \sqrt{2}(\bar X^*-\bar X)/S$ مقادیر $\sqrt{2}$، ۰ و $-\sqrt{2}$ را با احتمالهای $\frac{1}{4}$، $\frac{1}{2}$ و $\frac{1}{4}$ می‌گیرد، بنابراین

$$\gamma^* = \gamma(F_n) = P^*(R^*\le x) = \begin{cases} 0 & x < -\sqrt{2}, \\ \frac{1}{4} & -\sqrt{2} \le x < 0, \\ \frac{3}{4} & 0 \le x < \sqrt{2}, \\ 1 & \sqrt{2} \le x. \end{cases}$$

در حالت کلی تعداد نمونه‌های بوت‌استرپ مجزا برابر $m=\binom{2n-1}{n}$ است که آن را به صورت $\{(X_{1j}^*,\dots,X_{nj}^*), j=1,\dots,m\}$ نشان می‌دهیم. حال اگر در نمونهٔ بوت‌استرپ $(X_1^*,\dots,X_n^*)$، n_1 بار X_1، n_2 بار X_2، ... و n_n بار X_n ظاهر شده باشند، احتمال به‌دست آوردن نمونهٔ بوت‌استرپ از توزیع n جمله‌ای زیر تبعیت می‌کند:

$$P(\text{مشاهدهٔ نمونهٔ بوت‌استرپ}) = \binom{n}{n_1,\dots,n_n}\prod_{i=1}^{n}(\frac{1}{n})^{n_i} = \frac{n!}{n_1!\dots n_n!}(\frac{1}{n})^n,$$

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

اگر در بین m نمونهٔ بوت‌استرپ مجزا، احتمال به‌دست آوردن k امین نمونهٔ بوت‌استرپ را باp_k، $(k=1,\dots,m)$ نشان دهیم (p_k به صورت احتمال بالا محاسبه‌می‌شود)، آنگاه آمارهٔ بوت‌استرپ kام 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 اندازهٔ دقت واریانس (۵) در نظر می‌گیریم، در این صورت،

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

7) Tibshirani

واضح است که محاسبهٔ 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 نمونهٔ کوچک $n = 10$ باید $m = 92378$ آمارهٔ T^{*b} و احتمال p_k را محاسبه نمود. در نتیجه به‌جای m نمونهٔ بوت‌استرپ B نمونهٔ بوت‌استرپ را تولید، T^{*b} را محاسبه و با استفاده از تقریب مونت‌کارلو $\hat{\gamma}^*$ را به‌دست می‌آوریم.

۳ استنباطهای مجانبی

بیشتر مطالعات نظری در زمینهٔ بوت‌استرپ، رفتار مجانبی برآورد کننده‌های بوت‌استرپ است. اگر چه در کاربردهای واقعی حجم‌نمونه نمی‌تواند به طور نامحدودی افزایش یابد، با این حال نظریهٔ 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 تابعک $\gamma_n(F)$ نشان داد. با تغییر کوچکی مشابه (۳) اغلب داریم:

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

حال اگر $\gamma_n(F_n)$ یا γ_n^* برآورد کنندهٔ بوت‌استرپ $\gamma_n(F)$ باشد، برآورد کنندهٔ بوت‌استرپ را سازگار می‌گوییم اگر وقتی که $n \to \infty$،

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

بیکل[8] و فریدمن[9] [2] سازگاری بوت استرپ را برای میانگین نمونه، تابعکهای فون میزس[10]، فرایندهای تجربی و فرایندهای چندک از قبیل میانه، فواصل صدکی و برآوردکنندهٔ L(ترکیب خطی از آماره‌های مرتب) مانند میانگین αـپیراسته بررسی کردند. در زیر تنها برای دو آمارهٔ میانگین و میانهٔ نمونه توضیح مختصر داده و شرایط سازگاری را بیان می‌کنیم.

(الف) بوت‌استرپ میانگین. فرض کنید $X_1, \ldots, X_n$ متغیرهای تصادفی حقیقی‌ـمقدار 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 بااستفاده از قضیهٔ 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 نمونهٔ بوت‌استرپ $X_1^*, \ldots, X_n^*$ را به‌دست می‌آوریم. به عبارت دیگر، از نمونهٔ اولیهٔ $X_1, \ldots, X_n$ به‌روش نمونه‌گیری تصادفی سادهٔ با جایگذاری نمونهٔ بوت‌استرپ $X_1^*, \ldots, X_n^*$ را به‌دست می‌آوریم. $\bar{X}_n^*$، S_n^{*2} و ریشهٔ بوت‌استرپ R_n^* را به صورت زیر تعریف می‌کنیم:

$$\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*,$$

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

$$R_n^* = \sqrt{n}(\bar{X}_n^* - \bar{X}_n)/S_n^*.$$

انتظار می‌رود که رفتار ریشهٔ R_n^* از R_n تقلید کند. بنابراین، توزیع R_n^* (که می‌تواند تنها با استفاده از نمونهٔ مشاهده شده محاسبه شود) می‌تواند برای تقریب توزیع نمونه‌ای نامعلوم R_n استفاده شود. یا حتی به‌طور 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 $n \to \infty$، آنگاه:

الفـ۱) توزیع شرطی $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ به طور ضعیف همگرا به توزیع نرمال $N(0, \sigma^2)$ است.

الفـ۲) در احتمال شرطی $S_n^* \to \sigma$. یعنی اینکه، برای هر $\varepsilon > 0$،

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

بیکل و فریدمن به‌طور مشابه وقتی X_iها متغیرهای تصادفی بردارـمقدار هستند سازگاری بوت‌استرپ را نشان دادند.

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

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

بیکل و فریدمن نشان دادند که $R_n^* = \sqrt{n}(\tilde{X}_n^* - \tilde{X}_n)$ نیز به طور ضعیف به همان توزیع نرمال میل می‌کند. اگر فرض کنیم $R_n = \sqrt{n}(\tilde{X}_n - m)/4f^2(m)$ و بخواهیم در برآورد بوت استرپ ریشه یعنی R_n^* به‌جای 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 به‌سمت واریانس حدی $1/4f^2(m)$ میل می‌کند.

8) Bickel 9) Freedman 10) Von mises

۲.۳ دقت بوت‌استرپ

سینگ[11] [30] با فرض $E(X^{2}) < \infty$ سازگاری قوی به‌طور یکنواخت برآوردکنندهٔ بوت‌استرپ توزیع ریشهٔ $\bar{X}_n$ را نشان داد.یعنی وقتی‌که $n \to \infty$،

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

همچنین با فرض $E|X|^{3} < \infty$، نشان داد که برآورد کنندهٔ بوت‌استرپ توزیع استودنت شدهٔ $\bar{X}_n$ دارای دقت مرتبهٔ دوم است. یعنی وقتی که $n \to \infty$،

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

سازگاری بالا نشان می‌دهد که تقریب بوت‌استرپ توزیع، دقیقتر از توزیع نرمال حدی است. یکی از موارد استفادهٔ تقریب توزیع در به‌دست آوردن فواصل اطمینان است. سینگ همچنین نرخ دقیق همگرایی را برای تقریب بوت‌استرپ توزیع نمونه‌ای چندکها تعیین کرد و نشان داد که بادقت بالا می‌توان توزیع $\sqrt{n}[F_n^{-1}(\alpha) - F^{-1}(\alpha)]$ را با توزیع بوت‌استرپ $\sqrt{n}[F_n^{*-1}(\alpha) - F_n^{-1}(\alpha)]$ تقریب کرد.

۴ الگوریتم بوت‌استرپ

در این بخش ابتدا روش بوت استرپ را در یک حالت کلی خلاصه می‌کنیم و سپس الگوریتمی برای محاسبات کاربردی و برنامه‌های کامپیوتری ارائه می‌کنیم.

۱.۴ خلاصهٔ روش بوت‌استرپ

فرض کنید نمونهٔ $\mathbf{Y} = (Y_1, \dots, Y_n)$ را داریم (لازم نیست iid باشند) و P یک مدل آماری است که تحت آن نمونه‌ها به‌دست آمده‌اند. معمولاً، P می‌تواند با استفاده از توزیع توأم $\mathbf{Y}$ یا بعضی پارامترها که به‌طور یکتایی این توزیع توأم را تعیین می‌کنند، مشخص شود. فرض کنید $R_n(\mathbf{Y}, P)$ یک ریشه است و می‌خواهیم توزیع آن را برآورد کنیم. ابتدا مدل P را با استفاده از نمونهٔ $\mathbf{Y}$ برآورد می‌کنیم. فرض کنید $\mathbf{Y}^*$ یک نمونهٔ بوت‌استرپ باشد که از مدل برآورد شدهٔ $\hat{P}$ تولید شده است. توزیع شرطی $R_n(\mathbf{Y}^*, \hat{P})$ به شرط $\mathbf{Y}$ برآوردکنندهٔ بوت‌استرپ توزیع $R_n(\mathbf{Y}, P)$ است. افرون و تیبشیرانی [4] این فرایند را به صورت نمودار زیر خلاصه کردند:

$$\begin{tikzcd} P \arrow[rr] \arrow[dr] & & \mathbf{Y} \arrow[dl] & \Longrightarrow & \hat{P} \arrow[rr] \arrow[dr] & & \mathbf{Y}^* \arrow[dl] \\ & R_n(\mathbf{Y}, P) & & & & R_n(\mathbf{Y}^*, \hat{P}) & \end{tikzcd}$$

هنر بوت‌استرپ استفاده از تقلید رفتار نمونه‌گیری سه‌تایی $(P, \mathbf{Y}, R_n(\mathbf{Y}, P))$ از $(\hat{P}, \mathbf{Y}^*, R_n(\mathbf{Y}^*, \hat{P}))$ است، به‌طوری که ارتباط میان $\hat{P}, \mathbf{Y}^*$ و $R_n(\mathbf{Y}^*, \hat{P})$ مشابه ارتباط میان $P, \mathbf{Y}$ و $R_n(\mathbf{Y}, P)$ است. اگر $\hat{P} = P$، آنگاه توزیع $R_n(\mathbf{Y}^*, \hat{P})$ دقیقاً مشابه توزیع $R_n(\mathbf{Y}, P)$ است. حتی اگر $\hat{P} \neq P$، توزیعهای $R_n(\mathbf{Y}^*, \hat{P})$ و $R_n(\mathbf{Y}, P)$ ممکن است شبیه هم باشند.

بااینکه بوت‌استرپ براساس اصل جانشینی و تقلید کردن رفتار نمونه‌گیری است، در عمل معمولاً با بازنمونه‌گیری انجام می‌شود. یعنی وقتی که توزیع شرطی $R_n(\mathbf{Y}^*, \hat{P})$ تابع صریحی از $\mathbf{Y}$ نیست، روش مونت‌کارلو برای محاسبه کردن برآورد کننده‌های بوت‌استرپ مورد نیاز است. بوت استرپ می‌تواند در تمام حالتهایی که مدل P می‌تواند معلوم فرض شود و یا با استفاده از $\hat{P}$ برآورد شود، کاربرد داشته باشد. این مسأله را در دو مثال زیر نشان می‌دهیم.

مثال ۶ـ مسألهٔ یک نمونه‌ای. فرض کنید $X_1, \dots, X_n$ یک نمونهٔ تصادفی iid از توزیع F باشند. توزیع توأم $X_1, \dots, X_n$ با استفاده از F تعیین می‌شود. بنابراین، $P = F$. اگر F متعلق به خانوادهٔ پارامتری باشد، آنگاه $P = F_\theta$ ، به‌طوری که θ یک بردار از پارامترهای نامعلوم است. در حالت پارامتری، ابتدا θ با استفاده از برآوردکنندهٔ $\hat{\theta}$ برآورد می‌شود و سپس P را با استفاده $\hat{P} = F_{\hat{\theta}}$ برآورد می‌کنیم. اکنون نمونهٔ بوت‌استرپ $X_1^*, \dots, X_n^*$ از $F_{\hat{\theta}}$ تولید می‌شود. این روش بوت استرپ اغلب بوت‌استرپ پارامتری نامیده می‌شود. درحالت ناپارامتری، P با استفاده از توزیع تجربی $\hat{P} = F_n$ برآورد می‌شود. سپس نمونهٔ بوت‌استرپ $X_1^*, \dots, X_n^*$ را از F_n تولید می‌کنیم. این روش بوت‌استرپ اغلب بوت‌استرپ ناپارامتری نامیده می‌شود. بوت‌استرپ ناپارامتری می‌تواند برای هر دومدل پارامتری و ناپارامتری استفاده شود.

بوت‌استرپ پارامتری بستگی به فرض معلوم بودن مدل پارامتری دارد، در صورتی که بوت‌استرپ ناپارامتری آزاد از فرض معلوم بودن مدل است. وقتی که مدل پارامتری دقیق است، بوت‌استرپ پارامتری کاراتر از بوت‌استرپ ناپارامتری است. به‌طور کلی، روش بوت‌استرپ متکی به این است که ما چگونه به خوبی مدل را تشخیص دهیم یا برآورد کنیم. حتی در حالت بوت‌استرپ ناپارامتری، وقتی که می‌دانیم F هموار است، می‌توانیم برآورد کنندهٔ هموار F را به جای F_n قرار دهیم. در این صورت برآورد کنندهٔ بوت‌استرپ ناپارامتری بهتری به‌دست می‌آید.

مثال ۷ـ مدل رگرسیون خطی. فرض می‌کنیم $Y_1, \dots, Y_n$ یک نمونهٔ تصادفی مستقل باشد، به‌طوری‌که $Y_i = (y_i, x_i'), i = 1, \dots, n$ ، x_i بردار p متغیره و x' ترانهادهٔ x است. با توجه به اینکه x_i تصادفی باشد یا نه، در این مسأله با دو مدل مختلف روبرو هستیم.

11) Singh

حالت الف) اگر x_i تصادفی نباشد، $y_i = x'_i\beta + \varepsilon_i$، به طوری که β یک بردار p متغیرهٔ نامعلوم از پارامترها و $\varepsilon_1, \ldots, \varepsilon_n$، iid از توزیع نامعلوم F_ε با میانگین صفر هستند. در این حالت، P می‌تواندبه صورت (β, F_ε) مشخص شود. فرض کنید $\hat{\beta}$ برآورد کنندهٔ β باشد(برای مثال برآورد کنندهٔ 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 $\hat{\varepsilon}_i - \frac{1}{n}\sum_{j=1}^{n}\hat{\varepsilon}_j, i = 1, \ldots, n$ نسبت می‌دهد، به طوری که $\hat{\varepsilon}_i = y_i - x'_i\hat{\beta}$، iامین باقیمانده است. اکنون P با $\hat{P} = (\hat{\beta}, \hat{F}_\varepsilon)$ برآورد می‌شود. برای تولید نمونهٔ بوت‌استرپ $Y_1^*, \ldots, Y_n^*$، ابتدا داده‌های $\varepsilon_1^*, \ldots, \varepsilon_n^*$ را به‌صورت iid از $\hat{F}_\varepsilon$ تولید می‌کنیم‌و سپس $y_i^* = x'_i\hat{\beta} + \varepsilon_i^*$، آنگاه $Y_i^* = (y_i^*, x'_i), i = 1, \ldots, n$. این روش به‌نام بوت استرپ کردن باقیمانده‌ها معروف است.

حالت ب) x_i تصادفی است، $iid, (y_i, x'_i), i = 1, \ldots, n$ از توزیع نامعلوم $(p+1)$ متغیرهٔ F هستند و $E(y_i|x_i) = x'_i\beta$. در این حالت، $P = F$ می‌تواند با توزیع تجربی F_n برآورد شود. F_n جرم $\frac{1}{n}$ را به هر جفت $(y_i, x'_i), i = 1, \ldots, n$ نسبت می‌دهد. نمونهٔ بوت‌استرپ $Y_i^* = (y_i^*, x_i^*), i = 1, \ldots, n$ را به صورت iid از F_n تولید می‌کنیم. این روش به نام بوت‌استرپ زوجها معروف است.

۲.۴ بوت‌استرپ در ساختمان داده‌های پیچیده‌تر

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

(الف) مدلهای خطی. فریدمن [16] بوت‌استرپ مدل‌های رگرسیونی را در دو حالت بردار x ثابت و تصادفی بررسی کرده است. او همچنین در [17] بوت‌استرپ برآوردهای حداقل مربعات دو مرحله‌ای در مدلهای خطی شامل مدل رگرسیونی، مدلهای پویا و مدلهای اقتصاد سنجی را ارائه کرده است. برای دیدن برخی نتایج تجربی در زمینهٔ بوت‌استرپ معادلات رگرسیونی می‌توان‌به [18] مراجعه کرد.

(ب) مدل غیر خطی. بوت‌استرپ رگرسیون ناپارامتری [21]، فواصل اطمینان بوت‌استرپ در رگرسیون ناپارامتری [20] و نرخ همگرایی بوت‌استرپ در رگرسیون ناپارامتری [5]، همچنین فواصل اطمینان بوت‌استرپ مدل کاکس [4]، بسط اجورث[12] و تقریب بوت‌استرپ رگرسیون کاکس [19] و بوت استرپ رگرسیون لجستیک [26] از دیگر بحثهای بوت‌استرپ است.

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

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

(و) استنباط بیزی. ایدهٔ باز نمونه‌گیری داده‌ها می‌تواند در محاسبهٔ احتمال پسین توزیع‌ها در تحلیل بیزی کاربرد داشته باشد. ایدهٔ بوت استرپ بیزی به وسیلهٔ روبین [29] ارائه گردید. گسترش این ایده در زمینهٔ استنباط بیزتجربی براساس نمونهٔ بوت‌استرپ 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 زمینهٔ بوت‌استرپ مشاهدات وابسته و به‌خصوص 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 نرم‌افزار S-PLUS می‌توان به [1] مراجعه نمود.

۳.۴ الگوریتم بوت‌استرپ

فرض کنید $X_1, \ldots, X_n$ نمونهٔ تصادفی iid از توزیع نامعلوم F باشد و $T_n = T(X_1, \ldots, X_n)$ یک برآوردکنندهٔ پارامتر نامعلوم θ باشد. هدف برآورد اندازه‌های دقت آمارهٔ T_n وتوزیع نمونه‌ای آمارهٔ T_n و یا ریشه‌های $R_n(T_n, F)$ به صورت $\gamma_n(F)$ است. در زیر مراحل سه گانهٔ روش بوت استرپ را برای محاسبهٔ برآوردهای مورد نظر ارائه می‌کنیم.

مرحلهٔ۱ـ نمونهٔ بوت‌استرپ. با نمونه گیری iid از توزیع تجربی F_n نمونهٔ بوت‌استرپ $X_1^*, \ldots, X_n^*$ را تولید می‌کنیم. به عبارت ساده‌تر، نمونهٔ بوت‌استرپ $X_1^*, \ldots, X_n^*$ را با نمونه‌گیری تصادفی سادهٔ با جایگذاری‌از نمونهٔاولیه $X_1, \ldots, X_n$ به‌دست می‌آوریم.

مرحلهٔ۲ـ آماره و سازگاری بوت‌استرپ. با تأثیر آمارهٔ T_n بر نمونهٔ بوت‌استرپ $X_1^*, \ldots, X_n^*$ آمارهٔ بوت‌استرپ $T_n^* = T(X_1^*, \ldots, X_n^*)$ و در صورت لزوم ریشهٔ بوت‌استرپ $R_n^* = R(T_n^*, F_n)$ را محاسبه می‌کنیم.اکنون برآورد کننده‌های بوت‌استرپ اندازهٔ دقت و توزیع نمونه‌ای را به صورت γ_n^* به‌دست می‌آوریم. باید توجه کنیم که بوت استرپ وقتی معتبر

12) edgeworth expansion

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

مرحلهٔ۳ـ برآوردکنندهٔ بوت‌استرپ. برآوردکنندهٔ بوت‌استرپ γ_n^* محاسبه شده در مرحلهٔ ۲ در بیشتر حالتها به‌خصوص در مورد توزیع نمونه‌ای عبارت صریحی نیست و بیشتر برای محاسبات نظری و استنباط‌های مجانبی مفید است. در نتیجه از روش مونت‌کارلو برای تقریب برآوردکننده‌های بوت‌استرپ γ_n^* مرحلهٔ ۲ استفاده می‌کنیم. به این صورت که، مراحل ۱و۲ را B بارتکرار کرده (B را در عمل بزرگ انتخاب می‌کنیم)، B نمونهٔ بوت استرپ $\{(X_{1b}^*, \ldots X_{nb}^*); b = 1, \ldots, B\}$ را تولید و B آمارهٔ بوت‌استرپ $\{T_{n,b}^* = T_n(X_{1b}^*, \ldots, X_{nb}^*); b = 1, \ldots, B\}$ را به‌دست می‌آوریم. اکنون γ_n^* را بااستفاده از روش مونت‌کارلو به صورت $\hat{\gamma}_n^*$ تقریب می‌کنیم. در روش بوت‌استرپ باید به نکات زیر توجه کرد.

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

نکتهٔ ۲ـ 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 $\gamma_n^* \xrightarrow{\text{a.s.}} \hat{\gamma}_n^*$.

نکتهٔ ۳ـ متغیر تصادفی X را یک متغیره فرض کرده‌ایم که می‌تواند p متغیره هم فرض شود. در نتیجه آمارهٔ T_n هم می‌تواند p متغیره باشد.

نکتهٔ ۴ـ مراحل سه‌گانهٔ روش بوت‌استرپ برای ساختمان داده‌های پیچیده‌تر با اندکی تغییر قابل اجرا است. روش بوت‌استرپ در مشاهدات وابسته، تنها مرحلهٔ اول یعنی تولید نمونهٔ بوت‌استرپ تغییر خواهد کرد. و دو مرحلهٔ بعدی مشابه است.

نکتهٔ ۵ـ برآوردهای بوت‌استرپ توزیع نمونه‌ای آمارهٔ T_n برای ساختن فواصل اطمینان پارامتر θ مفید هستند. در نتیجه می‌توان از بافتنگار آمارهٔ بوت‌استرپ $\{T_{n,b}^*; b = 1, \ldots, B\}$ برای محاسبهٔ فاصلهٔ اطمینان صدکی استفاده کرد.

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

الگوریتم بوت‌استرپ

مرحلهٔ ۱ـ نمونهٔ بوت‌استرپ $X_1^*, \ldots, X_n^*$ را به روش نمونه‌گیری تصادفی سادهٔ با جایگذاری از نمونهٔ مشاهده شدهٔ $x_1, \ldots, x_n$ به‌دست می‌آوریم.

مرحلهٔ ۲ـ آمارهٔ بوت‌استرپ $T^* = T(X_1^*, \ldots, X_n^*)$ را محاسبه می‌کنیم.

مرحلهٔ ۳ـ مراحل ۱و۲ را B بار تکرار کرده و B آمارهٔ بوت‌استرپ $\{T_b^*; b = 1, \ldots, B\}$ را محاسبه می‌کنیم. برآوردکنندهٔ بوت استرپ واریانس و توزیع نمونه‌ای T_n به صورت زیر ارائه می‌شود:

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

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

۵ مثالهای کاربردی

در این بخش برای آشنایی بیشتر با روش و الگوریتم بوت‌استرپ 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 برروی هفت موش آزمایش کرده‌اند که نتایج زیر به‌دست آمده‌است:

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

میانگین، خطای معیار و میانهٔ 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 $\tilde{X}$ و در صورت امکان محاسبهٔ فاصلهٔ اطمینان برای میانگین μ و میانهٔ m جامعه با استفاده از الگوریتم بوت‌استرپ است.

مرحلهٔ۱ـ از نمونهٔ مشاهده شده $(x_1, \ldots, x_7)$ به روش نمونه‌گیری تصادفی سادهٔ باجایگذاری نمونهٔ بوت‌استرپ $(x_1^*, \ldots, x_7^*)$ را به‌دست می‌آوریم. برای مثال ممکن است نمونهٔ زیر را به‌دست آوریم:

$$(x_1^*, \ldots, x_7^*) = (x_5, x_7, x_5, x_4, x_7, x_3, x_1)$$
$$= (99, 23, 99, 38, 23, 16, 94).$$

مرحلهٔ ۲ـ میانگین و میانهٔ نمونهٔ بوت استرپ $(x_1^*, \ldots, x_7^*)$ مرحلهٔ ۱ را به دست می‌آوریم،

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

مرحلهٔ ۳ـ مراحل ۱و۲ را B با تکرار کرده، در نتیجه B آمارهٔ بوت‌استرپ $\bar{x}_b^*$ و $\tilde{x}_b^*$، $(b = 1, \ldots, B)$ را به‌دست می‌آوریم. اکنون انحراف معیار میانگین و میانه را با استفاده از روش مونت‌کارلو به صورت زیر تقریب می‌کنیم:

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

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

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 ـ برآورد بوت‌استرپ انحراف معیارمیانگین و میانه.

با استفاده از ادامهٔ مثال ۱ بخش ۲ـ۱ عبارت صریحی برای برآورد بوت‌استرپ انحراف معیار میانگین می‌توان به دست آورد. در این صورت ستون آخر جدول ۱ به دست می‌آید، وقتی که $B \to \infty$،

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

همچنین با استفاده از مثال ۴ بخش ۲ـ۱ و رابطهٔ (۶) احتمالهای $p_k(k = 1, \ldots, 7)$ را به صورت زیر به‌دست می‌آوریم:

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

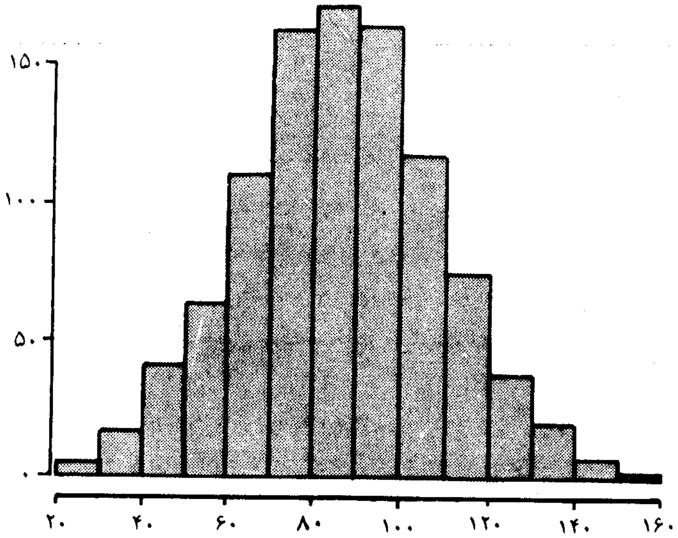
در نتیجه برآورد بوت‌استرپ انحراف‌معیار میانه به صورت زیر محاسبه می‌شود:

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

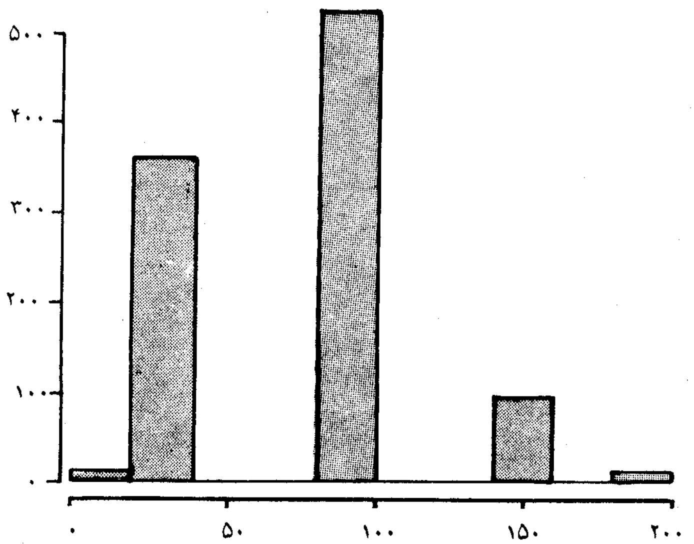
در این صورت ستون آخر جدول ۱ به‌دست می‌آید، وقتی که $B \to \infty$،

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

شکلهای ۱ و ۲ بافتنگار بوت‌استرپ میانگین و میانهٔ مرحلهٔ 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 ۳ـ۱ نیز نشان داده شده است. در بوت‌استرپ کردن میانهٔ بخش ۳ـ۱ شرط نزدیکی بافتنگار بوت‌استرپ به توزیع نرمال را پیوستگی مشتق f در همسایگی m ذکر کردیم که واضح است در نمونهٔ گسسته $(x_1, \ldots, x_7)$ این شرط برقرار نیست و در نتیجه شکل ۲ گسسته است.



شکل ۱ ـ بافتنگار بوت‌استرپ میانگین.



شکل ۲ ـ بافتنگار بوت‌استرپ میانه.

فاصلهٔ اطمینان ۰٫۹۵ میانگین جامعه با استفاده از توزیع tـاستودنت بدون استفاده از بوت‌استرپ به صورت زیر است:

$$\mu \in [\bar{x} \pm t_{n-1,0.025}\frac{s}{\sqrt{n}}] = [86.86 \pm 2.447(\frac{66.77}{\sqrt{7}})]$$
$$= [25.11, 148.61].$$

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

$$\mu \in [\bar{x}^*_{(۲۵)}, \bar{x}^*_{(۱۷۵)}] = [۴۱٫۷۵, ۱۳۰٫۸۶].$$

فاصلهٔ اطمینان صدکی بوت‌استرپ کوتاهتر از فاصلهٔ اطمینان tـاستودنت است، در نتیجه بادقت بیشتری μ را شامل می‌شود. فاصلهٔ اطمینان بوت‌استرپ برای میانهٔ جامعه را با استفاده از محاسبات مثال ۴ به صورت زیر ارائه می‌کنیم. چون

$$P^*\{x_{(۱)} < \tilde{X}^* < x_{(۷)}\} = ۱ - p_۱ - p_۷ \cong ۰٫۹۸,$$

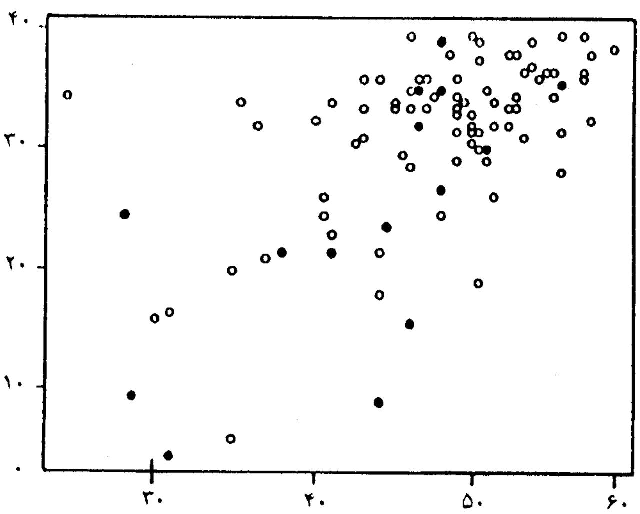
$$P^*\{x_{(۲)} < \tilde{X}^* < x_{(۶)}\} = p_۲ - p_۲ - p_۵ \cong ۰٫۷۸$$

درنتیجه فاصله اطمینان بوت‌استرپ ۰٫۹۸ و ۰٫۷۸ برای میانهٔ جامعه به صورت [۲۳, ۹۹] و [۳۸, ۹۹]است.

۲.۵ حالت دو متغیره، ضریب همبستگی

در زمستان سال ۱۳۷۳ تحقیقی با همکاری آقای دکترمشیری استاد درس زبان‌انگلیسی در دانشکدهٔ الکترونیک دانشگاه صنعتی‌شریف انجام گردید که در این مثال و مثال بعد از اطلاعات این تحقیق استفاده شده است.

از بین $N = ۹۸$ دانشجوی رشتهٔ الکترونیک دانشگاه صنعتی‌شریف، برای مطالعهٔ ارتباط نمرات درک مطلب (c) و دستور زبان (g) درس زبان انگلیسی، نمونه‌ای تصادفی به حجم $n = ۱۵$ انتخاب کرده‌ایم. نمرات جامعه با $(\circ + \bullet)$ و نمونه با $(\bullet)$ در شکل ۳ نشان داده شده است، محور افقی نمرات درک مطلب و محور عمودی نمرات دستور زبان دانشجویان است.



شکل ۳ـ نمرات درس زبان انگلیسی دانشجویان.

ضریب همبستگی جامعه و نمونهٔ نمرات به صورت زیر است:

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

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

مرحلهٔ۱ـ نمونهٔ تصادفی مشاهده شده را به صورت زوجی $x_i = (c_i, g_i), i = ۱, \ldots, ۱۵$ در نظر می‌گیریم. از نمونهٔ زوجی $(x_۱, \ldots, x_{۱۵})$ به روش نمونه‌گیری تصادفی سادهٔ با جایگذاری، نمونهٔ زوجی بوت‌استرپ $(x^*_۱, \ldots, x^*_{۱۵})$ را به‌دست می‌آوریم.

مرحلهٔ۲ـ ضریب همبستگی نمونهٔ زوجی بوت استرپ $(x^*_۱, \ldots, x^*_{۱۵})$ را به صورت $r^* = Corr^*(c^*, g^*)$ محاسبه می‌کنیم.

مرحلهٔ۳ـ مراحل ۱و۲ را B بارتکرار کرده و B ضریب همبستگی بوت‌استرپ $r^*_۱, \ldots, r^*_B$ را به‌دست می‌آوریم. اریبی و انحراف معیار ضریب همبستگی نمونه را با استفاده از روش مونت‌کارلو به صورت زیر تقریب می‌کنیم:

$$\widehat{bias}^*(r^*) = \frac{۱}{B}\sum_{b=۱}^{B} r^*_b - r,$$

$$\widehat{sd}^*(r^*) = \{\frac{۱}{B-۱}\sum_{b=۱}^{B}(r^*_b - \frac{۱}{B}\sum_{j=۱}^{B} r^*_j)^۲\}^{۱/۲}$$

جدول ۲ برآورد بوت‌استرپ اریبی و انحراف معیار ضریب همبستگی را به‌ازای مقادیر مختلف B نشان می‌دهد.

B	۵۰	۱۰۰	۲۵۰	۵۰۰	۱۰۰۰	۱۰۰۰$R.S.$
$\widehat{bias}^*(r^*)$	−۰٫۰۰۱	−۰٫۰۱۴	۰٫۰۱۵	۰٫۰۰۲	−۰٫۰۰۱	−۰٫۰۱۲
$\widehat{sd}^*(r^*)$	۰٫۱۴۷	۰٫۱۴۶	۰٫۱۴۴	۰٫۱۴۹	۰٫۱۵۲	۰٫۲۰۲

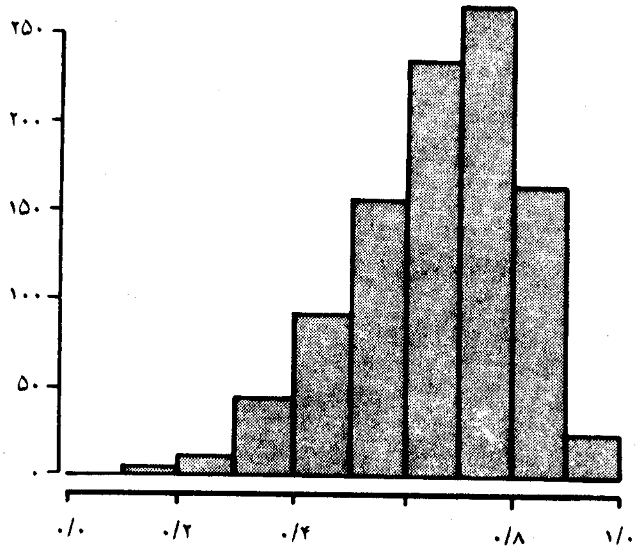
جدول ۲ـ برآورد بوت‌استرپ اریبی و انحراف معیار ضریب همبستگی.

در این مثال چون نمرات جامعهٔ $N = ۹۸$ دانشجو را در اختیار داریم با استفاده از روش و نمونه‌گیری مونت‌کارلو، ۱۰۰۰ بار نمونهٔ تصادفی زوجی به حجم $n = ۱۵$ را تولید و در نتیجه ۱۰۰۰ ضریب همبستگی نمونهٔ $r_۱, \ldots, r_{۱۰۰۰}$ را محاسبه می‌کنیم. در نتیجه با استفاده از روش مونت‌کارلو اریبی و خطای معیار ضریب همبستگی را به صورت زیر برآورد می‌کنیم(ستون آخر جدول ۲ مقادیر زیر هستند):

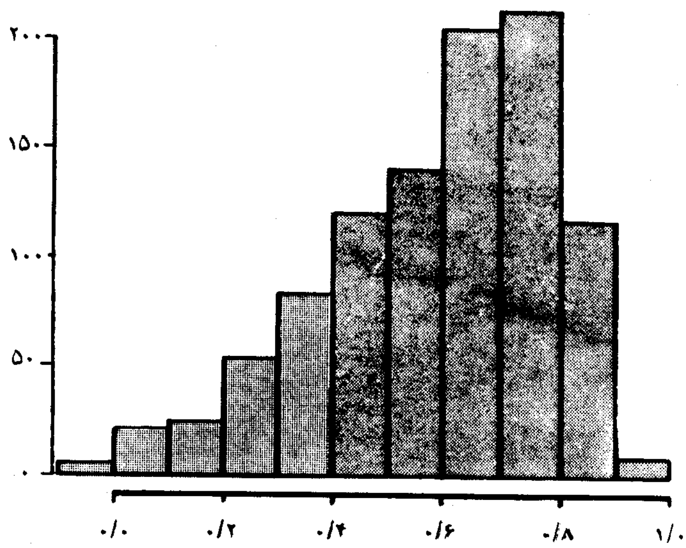
$$\widehat{bias}(r) = \frac{۱}{۱۰۰۰}\sum_{i=۱}^{۱۰۰۰} r_i - \rho = -۰٫۰۱۲,$$

$$\widehat{SE} = \{\frac{۱}{۹۹۹}\sum_{i=۱}^{۱۰۰۰}(r_i - \frac{۱}{۱۰۰۰}\sum_{j=۱}^{۱۰۰۰} r_j)^۲\}^{۱/۲} = ۰٫۲۰۲.$$

شکل ۴ بافتنگار بوت‌استرپ ضریب همبستگی حاصل از ۱۰۰۰ تکرار $r_1^*,\ldots,r_{1000}^*$ مرحلهٔ ۳ را نشان می‌دهد. شکل ۵ بافتنگار شبیه‌سازی ضریب همبستگی حاصل از ۱۰۰۰ تکرار $r_1,\ldots,r_{1000}$ نمونهٔ تصادفی است. واضح است که شکل ۴ تقریب بسیار خوبی از شکل ۵ است.



شکل ۴ـ بافتنگار بوت‌استرپ ضریب همبستگی.



شکل ۵ـ بافتنگار شبیه‌سازی ضریب همبستگی.

فاصلهٔ اطمینان ۰٫۹۵ ضریب همبستگی جامعه را با استفاده از فاصلهٔ اطمینان صدکی بوت‌استرپ محاسبه می‌کنیم. یعنی صدکهای ۰٫۰۲۵ و ۰٫۹۷۵ بافتنگار شکل ۴ را به‌دست می‌آوریم، به‌عبارت دیگر اگر ۱۰۰۰ آمارهٔ بوت‌استرپ $r_1^*,\ldots,r_{1000}^*$ را مرتب کنیم، فاصلهٔ اطمینان صدکی بوت‌استرپ ۰٫۹۵ میانگین جامعه به صورت زیر است:

$$[r^*_{(25)}, r^*_{(975)}] = [۰٫۳۲۲, ۰٫۸۹۹].$$

دیده می‌شود که ضریب همبستگی جامعهٔ $\rho = ۰٫۵۹۷$ به فاصلهٔ اطمینان بالا تعلق دارد.

۳.۵ حالت چند متغیره، بزرگترین مقدار ویژهٔ ماتریس کواریانس

از بین دانشجویان دانشگاه صنعتی شریف $n = ۹۸$ دانشجو به تصادف انتخاب و نمرات میان‌ترم و پایان ترم درس زبان انگلیسی شامل سه قسمت درک مطلب، لغت و دستور زبان را ثبت کرده‌ایم.

مجموعهٔ داده‌ها یک ماتریس ۶×۹۸ است که هر سطر آن برداری به صورت $x_i = (x_{i1},\ldots,x_{i6}), (i = ۱, ۲, \ldots, ۹۸)$ به ترتیب شامل نمرات درک مطلب، لغت و دستور زبان میان‌ترم 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 $\sum$ است به صورت زیر محاسبه کرده‌ایم:

$$\bar{x} = (۱۴٫۷۸, ۱۶٫۳۷, ۱۴٫۲۷, ۱۶٫۳۵, ۱۳٫۳۵, ۱۵٫۸۱).$$

$$S = \begin{pmatrix} ۷٫۴۲ & & & & & \\ ۱٫۶۱ & ۷٫۰۵ & & & & \\ ۵٫۴۲ & ۵٫۲۷ & ۲۵٫۶۲ & & & \\ ۳٫۹۴ & ۲٫۲۱ & ۶٫۲۱ & ۸٫۱۲ & & \\ ۴٫۳۸ & ۵٫۵۵ & ۹٫۵۵ & ۵٫۷۷ & ۱۳٫۹۹ & \\ ۴٫۹۲ & ۴٫۲۵ & ۱۳٫۶۱ & ۵٫۷۵ & ۸٫۲۹ & ۱۳٫۴۸ \end{pmatrix}$$

همچنین مقادیر ویژه ماتریس S به صورت زیر است:

$$(\hat\lambda_1,\ldots,\hat\lambda_6) = (۴۷٫۴۹, ۱۰٫۳۶, ۶٫۷۵, ۳٫۸۶, ۳٫۷۶, ۳٫۴۶).$$

پارامتر $\theta = \lambda_1 / \sum_{i=1}^{6} \lambda_i$ را با برآورد $\hat\theta = \hat\lambda_1 / \sum_{i=1}^{6} \hat\lambda_i = ۰٫۶۲۷۵$ در نظر می‌گیریم. هدف برآورد اریبی و انحراف معیار آمارهٔ $\hat\theta$ و همچنین محاسبهٔ فاصلهٔ اطمینان برای θ جامعه با استفاده از الگوریتم بوت‌استرپ است.

مرحلهٔ۱ـ از نمونهٔ تصادفی شش متغیرهٔ مشاهده شدهٔ $(x_1,\ldots,x_{98})$، به روش نمونه‌گیری تصادفی سادهٔ با جایگذاری نمونهٔ بوت‌استرپ شش متغیرهٔ $(x_1^*,\ldots,x_{98}^*)$ را به دست می‌آوریم.

مرحلهٔ۲ـ ماتریس کواریانس نمونهٔ بوت‌استرپ $(x_1^*,\ldots,x_{98}^*)$ را به‌دست آورده و مقادیر ویژهٔ $\lambda_1^*,\ldots,\lambda_6^*$ ماتریس S^* را محاسبه می‌کنیم و آنگاه $\theta^* = \lambda_1^* / \sum_{i=1}^{6} \lambda_i^*$ را به‌دست می‌آوریم.

مرحلهٔ۳ـ مراحل ۱ و ۲ را $B = ۱۰۰۰$ بار تکرار کرده $\theta_1^*,\ldots,\theta_{1000}^*$ را به‌دست می‌آوریم. اریبی و انحراف معیار برآورد کنندهٔ $\hat\theta$ را با استفاده از روش مونت‌کارلو به صورت زیر تقریب می‌کنیم:

$$\widehat{bias}^*(\theta^*) = \frac{۱}{۱۰۰۰}\sum_{b=۱}^{۱۰۰۰}\theta^*_b - \hat{\theta}$$
$$= ۰٫۶۲۷۹ - ۰٫۶۲۷۵ = ۰٫۰۰۰۴ \simeq ۰,$$

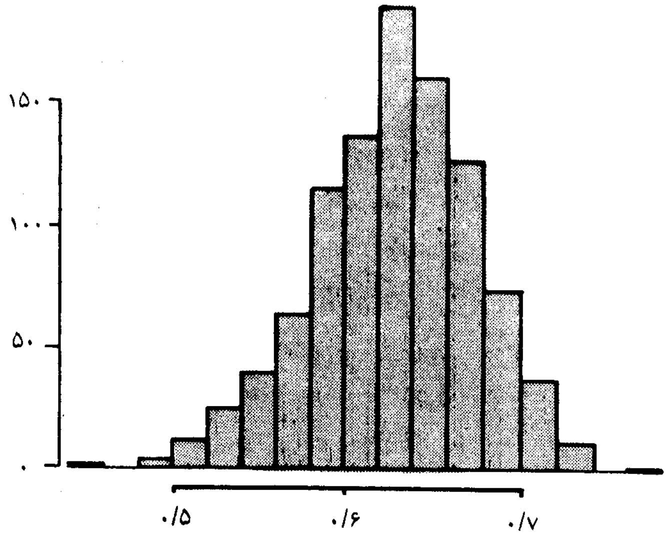
$$\widehat{sd}^*(\theta^*) = \{\frac{۱}{۹۹۹}\sum_{b=۱}^{۱۰۰۰}(\theta^*_b - \frac{۱}{۱۰۰۰}\sum_{j=۱}^{۱۰۰۰}\theta^*_j)^۲\}^{۱/۲} = ۰٫۰۴۶.$$

شکل ۶ بافتنگار بوت‌استرپ θ^* حاصل از ۱۰۰۰ تکرار $\theta^*_۱, \ldots, \theta^*_{۱۰۰۰}$ مرحلهٔ ۳ را نشان می‌دهد، واضح است که شکل ۶ به توزیع نرمال بسیار نزدیک است. در نتیجه فاصلهٔ اطمینان ۰٫۹۵ برای θ را باتوجه به توزیع نرمال θ^* می‌توان به صورت زیر محاسبه کرد:

$$\theta \in [\hat{\theta} \pm z_{۰٫۰۲۵}\widehat{sd}^*(\theta^*)] = [۰٫۶۲۸ \pm ۱٫۹۶(۰٫۰۴۶)]$$
$$= [۰٫۵۳۸, ۰٫۷۱۸].$$

همچنین فاصلهٔ اطمینان ۰٫۹۵ برای θ را با استفاده از فاصلهٔ اطمینان صدکی بوت‌استرپ یعنی با استفاده از ۱۰۰۰ تکرار $\theta^*_۱, \ldots, \theta^*_{۱۰۰۰}$ مرحلهٔ ۳ به صورت زیر محاسبه می‌کنیم:

$$\theta \in [\theta^*_{(۲۵)}, \theta^*_{(۹۷۵)}] = [۰٫۵۲۷, ۰٫۷۰۹].$$



شکل ۶ـ بافتنگار بوت‌استرپ θ^*.

درآنالیز چند متغیره با فرض اینکه $X \sim N_۶(\mu, \sum)$ است، فاصلهٔ اطمینان ۰٫۹۵ تقریبی برای θ به‌صورت زیر محاسبه می‌شود:

$$\theta \in [\hat{\theta}(۱ \pm z_{۰٫۰۲۵}\sqrt{\frac{۲}{n}})] = [۰٫۶۲۸(۱ \pm ۱٫۹۶\sqrt{\frac{۲}{۹۸}})]$$
$$= [۰٫۴۵۲, ۰٫۸۰۴].$$

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

مراجع

[۱] ایران پناه، نصراله (۱۳۷۵). الگوریتم بوت‌استرپ در سریهای زمانی و مشاهدات وابسته، رساله کارشناسی ارشد، گروه آمار، موسسه‌ی 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] Diciccio, 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.