روش های بازنمونه گیری در تحلیل داده های پیچیده

رضا مکرم [۱] ، وحید رنجبر [۲]

چکیده:

در تحلیل داده های پیچیده [۳]، معمولاً الگوهای نمونه گیری ساختاری غیر $i.i.d$ را برای داده ها در نظر می گیرند. لذا تکنیک های موجودی که برای برآورد نقطه‌ای و یا فواصل اطمینان وجود دارند معمولاً برای این نوع از داده ها قابل اجرا نمی باشند. برای حل این مشکل می توان از روش های نمونه گیری مانند نمونه گیری تکراری یا بازنمونه گیری [۴] استفاده کرد که در آنها برآورد پارامتر بدون نیاز به شرط $i.i.d$ بودن داده ها انجام گیرد. از این روش ها می توان با قدرت محاسباتی زیاد همچنین دوری جستن از تئوری های پیچیده برای برآورد پارامتر مورد استفاده قرار گیرند. این روش ها هر کدام دارای محاسن و همچنین کاستی هایی هستند که باعث ایجاد محدودیت در اجرای آنها می شوند. در این مقاله سعی شده است روش های مختلف بازنمونه گیری در الگوی نمونه گیری طبقه‌ای معرفی و سپس با یک مثال واقعی مورد مقایسه قرار گیرند.

واژه‌های کلیدی: بوت استرپ باجایگذاری، بوت استرپ بدون جایگذاری، جک نایف، نمونه گیری طبقه‌ای.

۱ مقدمه

روش های بازنمونه گیری، خصوصاً بوت استرپ و جک نایف، روش هایی هستند که بوسیله آنها می توان به راحتی برآوردگرهایی برای واریانس یافت و همچنین می توان فواصل اطمینان ناپاراتری را برای پارامترهای مورد نظر در جامعه به دست آورد. این روش ها از تئوری بسیار ساده‌ای نتیجه شده‌اند و استفاده از آنها غالباً به روش های برنامه نویسی کامپیوتر نیاز دارند. افرن [۴] برای اولین بار روشی جالب از بازنمونه گیری را برای پارامتر مورد نظر $\theta = \theta(F)$ بر اساس یک نمونه تصادفی n تایی $i.i.d$ از جامعه ای با توزیع نامشخص F معرفی نمود و آن را بوت استرپ نامید. او نشان داد که این روش می تواند از دیگر رقبای خود بهتر باشد اما در این باره بحث های متفاوتی نیز وجود دارد که برای توضیح بیشتر می توان به لو [۷] و وو [۱۳] مراجعه کرد.

هدف ما در این مقاله معرفی بسط روشهای بوت استرپ استاندارد (برای داده های $i.i.d$) به داده های پیچیده است. داده های پیچیده به داده های گفته می‌شود که برای آنها نمونه گیری به گونه ای است که خاصیت $i.i.d$ وجود ندارد. داده هایی که از روش های نمونه گیری طبقه‌ای، خوشه‌ای و نمونه گیری با احتمالات نابرابر

[۱] هیات علمی دانشگاه آزاد اسلامی واحد گرگان، Email: re612m@gmail.com
[۲] استادیار دانشگاه گلستان، گروه آمار، Email: v.ranjbar@gu.c.ir
[۳] Complex data
[۴] Resampling

حاصل می شوند این گونه اند.

روش هایی که در برآورد پارامتر و همچنین فواصل اطمینان برای داده های مختلط وجود دارند در عمل دارای مشکلاتی بوده و اغلب قابل تعمیم به مدلهای پیچیده نیستند و همچنین یافتن برآوردگرهای غیر خطی برای آنها مشکل است. سه روشی که عمدتاً برای این داده ها مورد استفاده قرار می گیرند عبارتند از روش خطی سازی[۵] (یا تیلور)، روش جک نایف و روش تکثیرهای متعادل تکرار شده[۶] (BRR). کرسکی و رائو [۶] به تفسیر و مقایسه این روش ها پرداخته و برای آنها نتایج مهمی را نیز به دست آوردند.

از مواردی که در بحث برآورد نقطه ای بسیار مورد توجه قرار دارد یافتن برآوردگری برای توابع غیر خطی از میانگین چند متغیر است. از جمله مهمترین این توابع ضریب همبستگی، نسبت و یا رگرسیون چند متغیر است. ممکن است یافتن این برآوردگرها در برخی مواقع چندان مشکل نباشد، اما می دانیم پس از انتخاب برآوردگر مسئله مهم تعیین میزان دقت و یا اصطلاحاً کارایی آن است. کارایی یک برآوردگر را به روش های مختلفی می توان تعیین کرد. از جمله ابزار مورد استفاده در تعیین کارایی، واریانس برآوردگر است. ممکن است یافتن واریانس توابع خطی از میانگین چند متغیر چندان کار سختی نباشد. اما در حالت غیر خطی مطمئناً این گونه نیست. به عنوان مثال فرم بسته ای برای یافتن واریانس ضریب همبستگی دو متغیر وجود ندارد. در این حالت می توان از روش های بازنمونه گیری استفاده کرد.

کرسکی و رائو [۶] نشان دادند که در این سه مدل برای توابع غیر خطی از میانگین چند متغیر مانند رگرسیون یا ضریب همبستگی می توان برآوردگرهایی پیدا کرد که به طور مجانبی سازگار هستند. با این حال این سه روش دارای نقاط ضعفی نیز هستند. به عنوان مثال روش خطی سازی نیازمند محاسبات تئوری زیادی بوده که باعث ایجاد مشکل در اجرای این روش می شود. از طرفی روشهای BRR و جک نایف فقط برای نمونه گیری طبقه ای که نمونه گیری اولیه باجایگذاری باشد تعریف شده اند. با این تفاصیل ما به روشی نیاز داریم که از محاسبات سخت تئوری دوری کرده و قابل تعمیم به مدلهای پیچیده نیز باشد.

در این مقاله بسط روش بوت استرپ به مدلهای پیچیده نمونه گیری که در آن نمونه اولیه به صورت بدون جایگذاری استخراج شده است و مشاهدات غیر $i.i.d$ هستند را توضیح داده خواهد شد. در قسمت دوم مقاله به نمونه گیری طبقه ای پرداخته می شود و در قسمت سوم روش های بوت استرپ موجود را برای این الگوی معرفی می کنیم که شامل روش بوت استرپ استاندارد (WR)، بوت استرپ باجایگذاری[۷] (BWR)، روش بازمقیاس گذاری بوت استرپ[۸] (RS) و بوت استرپ بدون جایگذاری[۹] (BWO)است. در قسمت چهارم مقاله روش $M.M$ را که شامل ویژگیهای مثبت روشهای BWO و BWR و قابل تعمیم به الگوهای

[۵] Linearization
[۶] Balanced Repeated Replication
[۷] With-Replacemant Bootstarp
[۸] Rescaling Bootstarp
[۹] Without-Replacemant Bootstarp

پیچیده باجایگذاری است را بیان می کنیم. درنهایت در قسمت پنجم با استفاده از یک مثال واقعی از جامعه ای با ساختار طبقه ای، پارامترهای میانگین و واریانس جامعه را توسط روش های بیان شده در این مقاله برآورد کرده و مقایسه می کنیم.

۲ نمونه گیری طبقه ای

در الگوی نمونه گیری طبقه ای، جامعه متشکل از N عضو بوده که به L طبقه مجزا تقسیم شده است. حجم جامعه در هر طبقه به ترتیب برابر با N_1، N_2، ... و N_L است. در این الگو نمونه گیری به این طریق بوده که از هر طبقه نمونه ای تصادفی ساده بدون جایگذاری (wtr) انتخاب شده که حجم نمونه ها در طبقات به ترتیب برابر است با n_1، n_2، ... و n_L، بطوریکه حجم کل نمونه برابر $n = n_1 + n_2 + ... + n_L$ خواهد بود.

فرض کنید y_{hi} اندازه‌ای از یک ویژگی از یک عضو در نمونه باشد که درآن اندیس h بیانگر شماره طبقه و زیرنویس i بیانگر iامین عضو در طبقه است. فرض کنید $\theta = \theta(S)$ پارمتر مورد نظر در جامعه بوده که در آن S مجموعه اعضای جامعه است، یعنی $S = \{Y_{hi} : h = 1, 2, ..., L; i = 1, 2, ..., N_h\}$. همچنین فرض کنید $\hat{\theta} = \hat{\theta}(s)$ برآوردگر پارامتر مورد نظر در نمونه باشد که در آن $s = \{y_{hi} : h = 1, 2, ..., L; i = 1, 2, ..., n_h\}$. میانگین جامعه را با μ نشان داده و برآوردگر نااریب آن را با $\bar{Y}_{st} = \sum_{h=1}^{L} W_h \bar{Y}_h$ نمایش می دهیم که در آن $\bar{Y}_h = \sum_{i=1}^{n_h} Y_{hi}/n_h$ و $W_h = N_h/N$. ککران [۳] برآوردگری نااریب برای $Var(\bar{Y}_{st})$ به صورت زیر ارائه داده است:

$$\widehat{Var(\bar{Y}_{st})} = \sum_{h=1}^{L} W_h^2 \frac{1 - f_h}{n_h} s_h^2. \qquad (1)$$

که درآن $f_h = n_h/N_h$ کسر نمونه گیری و s_h^2 برآورد واریانس نمونه ای است، یعنی

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2.$$

هدف در این مقاله یافتن برآوردگری برای رابطه ۱ به وسیله روش های بازنمونه گیری است.

۳ روش های بوت استرپ موجود

یکی از روش هایی که می توان به عنوان یک روش بازنمونه گیری در الگوی نمونه گیری طبقه ای در نظر گرفت اجرای روش بوت استرپ استاندارد برای نمونه در هر طبقه یعنی $\{y_{hi}\}_{i=1}^{n_h}$، است. این روش را با نماد (WR) نمایش می دهیم. در این صورت الگوریتم زیر حاصل می شود.

۱) نمونه گیری تصادفی با جایگذاری (WR) از نمونه اصلی یعنی $\{y_{hi}\}_{i=1}^{n_h}$ در هر یک از طبقه ها به طور مستقل تا حصول نمونه بوت استرپ یعنی $s^* = \{y_{hi}^* : h = 1, 2, ..., L; i = 1, 2, ..., n_h\}$، سپس محاسبه $\hat{\theta}^* = \hat{\theta}(s^*)$.

۲) B بار تکرار مرحله ۱ و به دست آوردن $\hat{\theta}_B^*, ..., \hat{\theta}_2^*, \hat{\theta}_1^*$.

۳) برآورد $Var(\hat{\theta})$ با مقدار زیر:

$$v_b = E_*(\hat{\theta}^* - E_*\hat{\theta}^*)^2.$$

یا تقریب مونت کارلوی آن یعنی:

$$\tilde{v}_B = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \overline{\hat{\theta}^*})^2 \quad ; \quad \overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*.$$

که درآن E_* امید ریاضی بازنمونه گیری می باشد. در معادلات فوق $E_*\bar{\theta}^*$ و$\overline{\hat{\theta}^*}$ را می توان با $\hat{\theta}$ جایگزین کرد.

نکته ۱ لازم به ذکر است که امید ریاضی بوت استرپ (E_*) با استفاده از اصل جایگزینی[10] یعنی با جایگزین کردن تابع توزیع تجربی به جای تابع توزیع جامعه حاصل می شود.

نکته ۲ الگوریتم های مختلف بازنمونه گیری فقط در شیوه انتخاب نمونه بوت استرپ باهم تفاوت داشته و بقیه مراحل آنها یکسان هستند.

قضیه ۱ : در ساده ترین حالت وقتی $\hat{\theta} = \bar{y}$ است می توان دید.

$$v_b = \sum_{h=1}^{L}\frac{W_h^2}{n_h}\left(\frac{n_h - 1}{n_h}\right)s_h^2. \qquad (2)$$

اثبات. الگوی بازنمونه گیری که در الگوریتم قبل بیان شد را می توان به صورت زیر تعبیر کرد: نمونه ای که در اختیار داریم با اجرای یک نمونه گیری طبقه ای بدون جایگذاری از جامعه اصلی که در آن حجم طبقه hام برابر N_h و حجم نمونه حاصل از آن طبقه برابر n_h است، بدست آمده است. حال این نمونه را به عنوان جامعه جدید در نظر می گیریم که جامعه ای است متشکل از L طبقه که حجم جامعه در طبقه hام برابر n_h می باشد. یک الگوی نمونه گیری طبقه ای باجایگذاری با حجم نمونه n_h در طبقه hام در این جامعه به اجرا در می آوریم. با این فرضیات و با استفاده از ویژگی های نمونه گیری طبقه ای با جایگذاری می توان واریانس میانگین نمونه را به صورت زیر تعریف کرد:

$$Var(\bar{Y}) = \sum_{i=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h}. \qquad (3)$$

که در آن

$$\sigma_h^2 = \frac{1}{n_h}\sum_{i=1}^{n_h}(y_{ih} - \bar{y}_h)^2.$$

از طرفی

$$\sigma_h^2 = \frac{n_h - 1}{n_h}s_h^2. \qquad (4)$$

با جایگذاری رابطه (؟؟) در رابطه (؟؟) اثبات قضیه کامل می شود. □

با مقایسه رابطه (؟؟) با رابطه (؟؟) می توان دریافت که برآوردگری که با استفاده از بوت استرپ استاندارد حاصل می‌شود یک برآوردگر اریب برای $Var(\bar{Y})$ است. نتیجه مهمتری که می توان به آن اشاره کرد این است که این برآوردگر برای واریانس سازگار نیست [۹]. البته این مشکل را می توان با درنظر گرفتن $n_h = n_\circ$ و $f_h = f$ به ازای هر h حل کرد. در این صورت عبارت زیر

$$\frac{n_\circ(1-f)v_b}{n_\circ - 1}$$

یک برآوردگر نااریب سازگار برای $Var(\bar{Y})$ خواهد بود. روش های دیگری نیز برای حل این مشکل ارائه شده است که در ادامه به طور مختصر به آنها اشاره کرده و به برخی از ویژگی های آنها اشاره می کنیم.

[10] Plog-in principle

۱.۳ بوت استرپ بدون جایگذاری

ابتدا این روش را در یک نمونه تصادفی ساده بدون جایگذاری بیان کرده سپس آنرا در نمونه گیری طبقه ای تعریف می کنیم. بوت استرپ بدون جایگذاری ابتدا توسط گروس [۵] برای برآورد واریانس معرفی شد. فرض کنید $N = kn$ باشد که k عددی صحیح است، الگوریتم روش BWO به صورت زیر است:

۱) تولید شبه جامعه ای توسط k بار تکرار نمونه اصلی.

۲) اجرای نمونه گیری بدون جایگذاری به حجم n از این جامعه تا حصول نمونه بوت استرپ $y_1^*, ..., y_n^*$.

۳) بدست آوردن $\hat{\theta}^* = \hat{\theta}(y_1^*, ..., y_n^*)$.

پس از محاسبه پارامتر بوت استرپ، بقیه الگوریتم شبیه به الگوریتم بوت استرپ استاندارد است.

قضیه ۲ : وقتی $\hat{\theta} = \bar{y}$ است در روش گروس می توان دید.

$$v_{bwo} = \frac{k(n-1)}{kn-1} \cdot \frac{1-f}{n} s^2. \qquad (۵)$$

اثبات. فرض کنید $x_1, ..., x_n$ نمونه اصلی و $y_1, ..., y_{kn}$ شبه جامعه حاصل از k بار تکرار نمونه باشد. همانطور که اشاره شد نمونه تصادفی بدون جایگذاری از جامعه استخراج می کنیم. لذا بر اساس نمونه گیری بدون جایگذاری میانگین و واریانس این نمونه عبارتست از:

$$\bar{y} = \frac{1}{kn} \sum_{i=1}^{kn} y_i = \frac{k}{kn} \sum_{i=1}^{n} x_i = \bar{x}.$$

$$\begin{aligned} Var(\bar{y}) &= \frac{1-f}{n} s_1^2 \\ &= \frac{1-f}{n} \cdot \frac{1}{kn-1} \sum_{i=1}^{kn} (y_i - \bar{x})^2 \\ &= \frac{1-f}{n} \cdot \frac{k}{kn-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \frac{1-f}{n} \cdot \frac{k(n-1)}{kn-1} s^2 \end{aligned}$$

که اثبات قضیه کامل می شود. □

همانطوری که ملاحظه می شود روش BWO حتی در ساده ترین حالت قادر به تولید برآوردگر نااریب برای واریانس میانگین نمونه نیست. البته در این حالت خاص نیز با استفاده از یک ضریب تصحیح می توان برآوردگر نااریب به دست آورد، اما این امر در الگوهای پیچیده تر نمونه گیری به راحتی انجام نمی گیرد. بیکل و فریدمن [۲] روشی را تحت عنوان بسط روش BWO در نمونه گیری طبقه ای ارائه دادند که بعدها توسط مک کارتی و اسنودن [۸] تکمیل گردید. روش آنها به این صورت است که فرض کنید برای هر طبقه رابطه $N_h = k_h n_h + r_h$ برقرار باشد که در آن $0 \le r_h \le n_h - 1$ و n_h و k_h اعداد صحیحی هستند. در این حالت دو شبه جامعه را طوری تشکیل می دهیم که یکی با k_h بار تکرار طبقه h ام و دیگری با $k_h + 1$ بار تکرار آن حاصل می شود. سپس از هر طبقه به طور مستقل نمونه ای بدون جایگذاری به حجم n_h با احتمال p_h از جامعه اول و با احتمال $1 - p_h$ از جامعه دوم استخراج می کنیم. که

$$p_h = \frac{(1-f_h)/(n_h-1) - a_{h2}}{a_{h1} - a_{h2}}.$$

که در آن

$$a_{hi} = \frac{k_h + i - 2}{(k_h + i - 1)n_h - 1}; \quad i = 1, 2.$$

مک کارتی و اسنودن [۸] ثابت کردند که در این حالت واریانس برآوردگر $\bar{y}^*_{st} = \sum_{h=1}^{L} W_h \bar{y}^*_h$ عبارتست از

$$v_{bwo}(\bar{y}^*_{st}) = \sum_{h=1}^{L} W_h^2 \frac{1 - f_h}{n_h} s_h^2.$$

که همان $Var(\bar{y}_{st})$ در نمونه گیری طبقه ای معمولی است. روش مک کارتی و اسنودن به راحتی قابل اجراست اما متاسفانه تحت شرایطی ممکن است p_h عددی منفی شود. مثالی از این حالت را مک کارتی و اسنودن [۸] برای $h = 1$ ارائه دادند. برای رفع این مشکل سیتر [۱۰] نیز روشی تحت عنوان بوت استرپ بدون جایگذاری در نمونه گیری طبقه ای ارائه داد که اساس کار او نیز مشابه روش مک کارتی و اسنودن [۸] تولید دو شبه جامعه است اما به شیوه ای متفاوت. روش سیتر اینگونه است که برای طبقه hام تعریف می کنیم:

$$n'_h = n_h - (1 - f_h),$$

$$k_h = \frac{N_h}{n_h}(1 - \frac{1 - f_h}{n_h}).$$

فرض کنید n'_h و k_h اعداد صحیحی باشند. نمونه های هر طبقه را مستقلاً k_h بار تکرار می کنیم تا شبه جامعه‌ای حاصل شود. سپس نمونه هایی به حجم n'_h بدون جایگذاری از هر طبقه استخراج می کنیم. پس از استخراج نمونه بوت استرپ بقیه مراحل مشابه قبل انجام می‌شود. تحت این شرایط سیتر ثابت کرده است که برآوردگر واریانس بوت استرپ برابر با واریانس نمونه در حالت معمولی خواهد بود. یعنی این روش برآوردگری نااریب برای واریانس میانگین نمونه تولید می کند.

به وضوح می توان دید که ممکن است n'_h و k_h اعداد صحیحی نباشند. در این صورت از رویه تصادفی کردن استفاده می شود. به این صورت که تعریف می کنیم $k_{h1} = \lceil k_h \rceil$، $k_{h2} = \lfloor k_h \rfloor$ و $n'_{h1} = n_h - 1$ و $n'_{h2} = n_h$ که در آن $\lfloor . \rfloor$ و $\lceil . \rceil$ به ترتیب جزء صحیح و جزء صحیح به اضافه یک هستند. حال در رویه گفته شده با احتمال p_h از زوج (n'_{h1}, k_{h1}) و با احتمال $1 - p_h$ از زوج (n'_{h2}, k_{h1}) استفاده می کنیم. که

$$p_h = \frac{(1 - f_h)/(n_h^2 - n_h) - a_{h2}}{a_{h1} - a_{h2}},$$

که در آن

$$a_{hi} = \frac{k_{hi}(1 - n'_{hi}/n_h.k_{hi})}{n'_{hi}(n_h.k_{hi} - 1)}; \qquad i = 1, 2.$$

در این حالت نیز سیتر ثابت کرده است که برآوردگر نااریب برای واریانس میانگین نمونه حاصل خواهد شد.

۲.۳ بازمقیاس گذاری بوت استرپ

بسیاری از آماره ها قابل نوشتن به صورت $\hat{\theta} = g(\bar{y})$ یعنی تابعی از میانگین نمونه هستند. ضرایب رگرسیون، نسبت میانگین ها و ضریب همبستگی از این نوع اند. رائو و وو [۹] برای برآورد واریانس یک برآوردگر مانند $\hat{\theta}$ روشی با الگوریتم زیر در نظر گرفتند:

۱) استخراج نمونه ای با جایگذاری از نمونه اصلی.

۲) بازمقیاس گذاری هر عضو بازنمونه گیری شده.

۳) محاسبه برآوردگر اصلی توسط بردار مقادیر بازمقیاس گذاری شده.

فاکتورهای بازمقیاس گذاری طوری انتخاب می شوند که واریانس بازنمونه گیری با واریانس معمولی در حالت

برآوردگرهای خطی برابر باشد.

در نمونه گیری طبقه ای اگر $\{y_{hi}, i = ۱, ..., n_h\}$ مقادیر نمونه اصلی در طبقه ام و $\{y^*_{hi}, i = ۱, ..., n^*_h\}$ نمونه ای تصادفی ساده از آن باشد، عامل بازمقیاس گذاری به صورت زیر تعریف می شود:

$$C_h = \sqrt{n^*_h}(n_h - ۱)^{-۱/۲}(۱ - f_h)^{۱/۲}.$$

با استفاده از این عامل مقادیر بازنمونه گیری به صورت زیر بازمقیاس گذاری می شوند:

$$\tilde{y}^*_h = \bar{y}_h + C_h(\bar{y}^*_h - \bar{y_h})$$

که در آن $\bar{y}_h$ و $\bar{y}^*_h$ به ترتیب میانگین نمونه اصلی و بازنمونه گیری در طبقه h ام هستند. پس از این مرحله میانگین بازمقیاس گذاری، یعنی $\tilde{y}^* = \sum_{h=۱}^{L} W_h \tilde{y}^*_h$، محاسبه شده و سپس برآوردگر اصلی با استفاده از این میانگین بدست می آید $(\hat{\theta}^* = g(\tilde{y}^*))$.

رائو و وو [۹] ثابت کردند که با این الگوریتم برآوردگر نااریب برای توابع خطی از میانگین نمونه قابل دستیابی است. همچنین آنها الگوریتم های مختلفی را برای کلاس بزرگی از مدل های نمونه گیری از جمله خوشه‌ای دو مرحله ای و نمونه گیری با احتمالات نابرابر ارائه دادند. البته این روش معایبی دارد که برخی از مهمترین آنها عبارتنداز:

۱) محاسبه فاکتور بازنمونه گیری در الگوهای پیچیده نمونه گیری می تواند مشکل ساز باشد.

۲) این روش حافظ دامنه تغییرات پارامتر نیست.

۳.۳ بوت استرپ باجایگذاری

این روش که دارای ساختار ساده ای است توسط مک کارتی و اسنودن [۸] معرفی شد. در الگوی نمونه گیری طبقه ای روش بوت استرپ باجایگذاری (BWR) به این صورت است که نمونه ای تصادفی باجایگذاری در هر طبقه به حجم نمونه اصلی در آن طبقه به طور مستقل انتخاب می شود. پس استخراج نمونه بوت استرپ بقیه مراحل مشابه قبل انجام می شود. از مزایای این روش می توان به سازگار بودن آن با ساختار بوت استرپ که نمونه ها به صورت باجایگذاری انتخاب می شوند اشاره کرد. این روش حالت خاصی از روش $M.M$[۱۱] است که بعدها توسط سیتر [۱۰] معرفی شد و در ادامه مقاله به آن پرداخته می شود. بنابر این تمام مواردی که برای روش $M.M$ گفته می شود در این الگو نیز صادق است.

۴ روش $M.M$

طراحی الگوی های بازنمونه گیری بهتر است به گونه ای باشد که با الگوی نمونه گیری اصلی داده ها متناسب باشند. به عبارتی اگر نمونه اصلی بدون جایگذاری انتخاب شده است بهتر است بازنمونه گیری هم طوری طراحی شود که نمونه ها به صورت بدون جایگذاری انتخاب شوند و به طور مشابه اگر نمونه اصلی باجایگذاری باشد منطقی است که بازنمونه گیری هم باجایگذاری باشد. این مسئله در مورد روش BWO صادق است چراکه در این روش از شبه جامعه تولید شده با کسر نمونه گیری برابر با کسر نمونه گیری داده های

[۱۱] Mirror-Match

اصلی نمونه هایی به حجم نمونه اصلی و به روش بدون جایگذاری انتخاب می شوند. این مطلب را می توان به عنوان یک مزیت برای این روش دانست. اما در برخی مدل های بازنمونه گیری مانند روش های BWR و یا بازمقیاس گذاری این گونه نیست چراکه در این الگوها نمونه گیری با جایگذاری انجام می شود. اما می توان گفت که این روش ها با ساختار اصلی بوت استرپ سازگار است. زیرا همان طور که می دانیم در روش بوت استرپ نمونه گیری به صورت با جایگذاری از نمونه اصلی صورت می گیرد. حال به نظر می رسد مدلی که این دو ویژگی را تواماً دارا باشد مدل مناسبی خواهد بود. روشی که در ادامه معرفی می شود دارای این ویژگی است.

۱.۴ روش $M.M$ در الگوی تصادفی ساده

فرض کنید از جامعه ای به حجم N نمونه ای تصادفی ساده بدون جایگذاری به حجم n استخراج شده باشد بطوریکه داشته باشیم $N = k.n$ یا به عبارتی $k = f^{-1}$. تعریف می کنیم $n' = f.n$، با فرض صحیح بودن n و k روش $M.M$ برای این الگو عبارتست از:

۱) نمونه گیری تصادفی ساده بدون جایگذاری به حجم n' از نمونه اصلی تا حصول $y^*_1, y^*_2, ..., y^*_{n'}$.

۲) ثبت و سپس جایگذاری این زیر نمونه در نمونه اصلی.

۳) k بار تکرار مستقل مراحل ۱ و ۲ تا حصول نمونه بوت استرپ n تایی $\underline{y}^* = (y^*_1, ..., y^*_n)$ و سپس محاسبه $\hat{\theta}^* = \theta(\underline{y}^*)$.

ثابت می شود که روش $M.M$ بر خلاف روش بازمقیاس گذاری حافظ دامنه تغییرات پارامتر است. (مراجعه شود به [۹، ۱۰]).

در ادامه قضیه ای در مورد برآوردگر بوت استرپ حاصل از روش فوق بیان و اثبات می کنیم. ابتدا امید ریاضی و واریانس بوت استرپ که در این قضیه مورد استفاده قرار می گیرند را تعریف می کنیم.

تعریف ۱ : فرض کنید $s^* = \{y^*_{ij}; i = 1, ...n, j = 1, ..., B\}$ یک نمونه بوت استرپ از نمونه ای به حجم n با B بار تکرار باشد. همچنین فرض کنید $\theta^*_j = \theta(y^*_j)$ تابعی از نمونه بوت استرپ jام باشد. امید ریاضی و واریانس بوت استرپ پارامتر θ^* را به ترتیب با نمادهای $E_*(\theta^*)$ و $Var_*(\theta^*)$ نشان داده و با استفاده از تعریف تابع توزیع تجربی به صورت زیر تعریف می شوند.

$$E_*(\theta^*) = \frac{1}{B}\sum_{b=1}^{B}\theta^*_b,$$

$$var_*(\theta^*) = \frac{1}{B}\sum_{b=1}^{B}(\theta^*_b - E_*(\theta^*))^2.$$

قضیه ۳ : در روش $M.M$ در الگوی نمونه گیری تصادفی ساده داریم:

الف) برآوردگر بوت استرپ یک برآوردگر ناااریب برای میانگین نمونه است یعنی $E_*(\bar{y}^*) = \bar{y}$.

ب) $Var_*(\bar{y}^*) = Var(\bar{y})$.

اثبات. بر اساس تعریف ۱ می توان نوشت

$$E_*(\bar{y}^*) = E_*\left(\frac{1}{n'k}\sum_{i=1}^{k}\sum_{j=1}^{n'}y^*_{ij}\right)$$

از آنجایی که استخراج k زیر نمونه به طور مستقل صورت می گیرد داریم:

$$E_*(\bar{y}^*) = \frac{1}{k}\sum_{i=1}^{k} E_*\left(\frac{1}{n'}\sum_{j=1}^{n'} y_{ij}^*\right) = \frac{1}{k}\sum_{i=1}^{k}\bar{y} = \bar{y}.$$

برای اثبات قسمت دوم قضیه بطور مشابه می توان نوشت:

$$\begin{aligned} Var_*(\bar{y}^*) &= Var_*\left(\frac{1}{n'k}\sum_{i=1}^{k}\sum_{j=1}^{n'} y_{ij}^*\right) \\ &= \frac{1}{k^2}\sum_{i=1}^{k} Var_*\left(\frac{1}{n'}\sum_{j=1}^{n'} y_{ij}^*\right) \\ &= \frac{1-f^*}{k.n'}s^2. \end{aligned}$$

چون کسر بازنمونه گیری با کسر نمونه گیری نمونه اصلی برابر است داریم:

$$Var_*(\bar{y}^*) = \frac{1-f}{n}s^2 = Var(\bar{y}).$$

که این اثبات قضیه را کامل می کند. □

شرطی که در استفاده از این رویه وجود دارد و شاید آنرا بتوان یک محدودیت دانست، لزوم صحیح بودن n' و برقراری شرط $n' = fn \geq 1$ است. البته راه حل هایی برای آن نیز ارائه شده است.

الف) فرض کنید $n' = fn \geq 1$ بوده ولی عدد صحیحی نباشد، آنگاه می‌توان با عمل تصادفی کردن حجم بازنمونه‌گیری را به این صورت انتخاب کرد که با احتمال p، n' را $[fn]$ در نظر گرفته و با احتمال $1-p$ آنرا $[fn]$ در نظر می گیریم. که در آن p طوری انتخاب می شود که نتیجه برآورد واریانس $\bar{y}$ برابر با برآوردگر نااریب آن در حالت معمولی باشد.

ب) حال اگر $0 < fn < 1$ باشد، n' را عددی بین ۱ و n انتخاب می کنیم. در این حالت الگوریتم زیر در نظر گرفته می شود:

۱) انتخاب $1 < n' < n$ و بازنمونه گیری به حجم n' تا حصول $y_{n'}^*, ..., y_2^*, y_1^*$.

۲) $k = [n(1-f^*)]/[n'(1-f)]$ بار (با شرط صحیح بودن k) تکرار مرحله ۱ به طور مستقل و جایگزین کردن نمونه های به حجم n' در هر مرحله تا حصول $y_{n^*}^*, ..., y_2^*, y_1^*$ که در آن $n^* = n'.k$.

به الگوی فوق روش $M.M$ اصلاح شده گفته می شود.

قضیه ۴ : تحت شرایط فوق (الف و ب) نیز نتایج قضیه ۳ در الگوی تصادفی ساده معتبر است.

اثبات. اثبات این قضیه به راحتی با نوشتن تعریف امید ریاضی و واریانس بوت استرپ نتیجه می شود. □

۲.۴ روش $M.M$ در الگوی طبقه ای

از آنجایی که الگوی نمونه گیری طبقه ای حالت کلی تر الگوی تصادفی ساده است یا به عبارتی نمونه تصادفی ساده همان نمونه گیری طبقه ای با $L = 1$ است، لذا تمام شرایطی که در قسمت قبل ارائه شد قابل تعمیم به حالت طبقه ای می باشد. فرض کنید در هر طبقه رابطه $N_h = n_h k_h$ (به عبارتی $k_h = f_h^{-1}$) برقرار باشد. تعریف می کنیم $n'_h = f_h n_h$ و فرض می کنیم k_h و n_h به ازای هر h اعداد صحیحی باشند. (در غیر این صورت از رویه تصادفی کردن استفاده می کنیم که در قسمت قبل اشاره شد.) آنگاه روش $M.M$ اصلاح شده را برای الگوی طبقه ای به صورت زیر می توان تعریف کرد:

۱) نمونه گیری تصادفی ساده بدون جایگذاری به حجم $n'_h < n_h$ از نمونه اصلی در هر طبقه تا حصول $y^*_{hn'_h}, ..., y^*_{h۲}, y^*_{h۱}$.

۲) ثبت و سپس جایگذاری این زیر نمونه در نمونه اصلی.

۳) $k_h = [n_h(۱ - f^*_h)]/[n'_h(۱ - f_h)]$ بار تکرار مستقل مراحل ۱ و ۲ تا حصول نمونه بوت استرپ $y^*_{hn_h}, ..., y^*_{h۲}, y^*_{h۱}$ و در نتیجه $s^* = \{y^*_{hi} : h = ۱, ۲, ..., L; i = ۱, ۲, ..., n_h\}$

۴) محاسبه $\hat{\theta}^* = \hat{\theta}(s^*)$.

باید توجه داشت که در قسمت نمونه گیری بدون جایگذاری کسر بازنمونه گیری برابر است با کسر نمونه گیری اصلی، به همین دلیل این مرحله را k_h بار تکرار می کنیم تا بردار بازنونه گیری به حجم n_h حاصل شود.

قضیه ۵ : در روش $M.M$ اصلاح شده در الگوی طبقه ای وقتی $\theta = \bar{Y}$ و $\hat{\theta} = \bar{y}_{st}$ باشند، برای برآوردگر بوت استرپ آن یعنی $\bar{y}^*_{st} = \sum_{h=۱}^{L} W_h \bar{y}^*_h$ که در آن $\bar{y}^*_h = \sum_{i=۱}^{k_h} \sum_{j=۱}^{n'_h} y^*_{hij}/n^*_h$ داریم:

الف) $\bar{y}^*_{st}$ برآوردگر ناریب $\bar{y}_{st}$ است، یعنی $E_*(\bar{y}^*_{st}) = \bar{y}_{st}$.

ب) واریانس بوت استرپ برآوردگر بوت استرپ $\bar{y}^*_{st}$ برابر است با واریانس برآوردگر معمولی $\bar{y}_{st}$. یعنی $Var_*(\bar{y}^*_{st}) = Var(\bar{y}_{st})$.

اثبات. می توان نوشت

$$E_*(\bar{y}^*_h) = E_*\left[\frac{۱}{n^*_h}\sum_{i=۱}^{k_h}\sum_{j=۱}^{n'_h} y^*_{hij}\right] = \frac{۱}{k_h}\sum_{i=۱}^{k_h} E_*\left[\frac{۱}{n'_h}\sum_{j=۱}^{n'_h} y^*_{hij}\right] = \frac{۱}{k_h}\sum_{i=۱}^{k_h}\bar{y}_h = \bar{y}_h.$$

پس

$$E_*(\bar{y}^*_{st}) = E_*\sum_{h=۱}^{L} W_h \bar{y}^*_h = \sum_{h=۱}^{L} W_h \bar{y}_h = \bar{y}_{st}.$$

که اثبات الف کامل می شود. برای اثبات قسمت دوم از آنجایی که نمونه گیری در طبقات مختلف به طور مستقل انجام می گیرد، داریم:

$$\begin{aligned} Var_*(\bar{y}^*_{st}) &= \sum_{h=۱}^{L} W_h^۲ Var_*\left(\frac{۱}{k_h n'_h}\sum_{i=۱}^{k_h}\sum_{j=۱}^{n'_h} y^*_{hij}\right) \\ &= \sum_{h=۱}^{L}\frac{W_h^۲}{h_h^۲}\sum_{i=۱}^{k_h}\left(\frac{۱ - f^*_h}{n'_h} s_h^۲\right) \\ &= \sum_{h=۱}^{L} W_h^۲ \frac{۱ - f_h}{n_h} s_h^۲ \end{aligned}$$

که اثبات قضیه کامل می شود. □

قضیه فوق برای هر n'_h و k_h برقرار است لذا می توان n'_h را طوری کرد که کسر بازنمونه گیری $(f^*_h = n'_h/n_h)$ برابر با کسر نمونه گیری اصلی $(f_h = n_h/n_h)$ باشد. در ادامه مزیت این انتخاب توضیح داده خواهد شد.

۱.۲.۴ انتخاب n'_h برای جور شدن گشتاور سوم

در الگوریتم قبل همانطوری که اشاره شد $E_*(\bar{y}^*) = \bar{y}$ و $Var_*(\bar{y}^*) = Var(\bar{y})$ است. یعنی دو گشتاور اول توزیع $\bar{y}^*$ برابر با برآوردگر ناریب دو گشتاور اول $\bar{y}$ هستند. در

این بخش هدف پیدا کردن n'_h است به طوری که گشتاور سوم بوت استرپ یعنی $E_*(\bar{y}^* - \bar{y})^3$ برابر با $\hat{\mu}_3(\bar{y})$ شود. ابتدا یادآور می شویم که برای $n_h \geq 2$ برآوردگر نااریب $E(\bar{y} - E(\bar{y}))^3$ برابر است با:

$$\hat{\mu}_3(\bar{y}) = \sum_{h=1}^{L} W_h^3 \frac{(1-f_h)(1-2f_h)m_{3h}}{(n_h-1)(n_h-2)},$$

که در آن

$$m_{3h} = \frac{1}{n_h}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^3.$$

گشتاور نمونه ای مرتبه سوم در طبقه hام است. با فرض صحیح بودن n'_h و k_h، حجم کل نمونه در هر طبقه عبارتست از

$$n^*_h = k_h n'_h = \frac{n_h(1-f^*_h)}{1-f_h}.$$

که در آن $f^*_h = n'_h/n_h$ کسر بازنمونه گیری است. با فرض $\varepsilon_h = W_h(\bar{y}^*_h - \bar{y}_h)$، می توان نوشت $E_*(\varepsilon_h) = 0$ و در نتیجه

$$\begin{aligned} E_*(\bar{y}^* - \bar{y})^3 &= E_*\left(\sum_{h=1}^{L}\varepsilon_h\right)^3 \\ &= \sum_{h=1}^{L} E_*(\varepsilon_h^3) \\ &= \sum_{h=1}^{L} W_h^3 E_*(\bar{y}^*_h - \bar{y}_h)^3. \quad (6) \end{aligned}$$

حال برای پیدا کردن $E_*(\bar{y}^*_h - \bar{y}_h)^3$، با صرف نظر کردن از زیرنویس h بنا به مستقل بودن طبقات می توان نوشت

$$\bar{y}^* = \frac{1}{k}\sum_{j=1}^{k}\bar{y}^*_j = \frac{1}{kn'}\sum_{j=1}^{k}\sum_{l=1}^{n'} y^*_{jl}.$$

که در آن زیر نویس های j و l به ترتیب مربوط به قسمت های باجایگذاری و بدون جایگذاری در الگوریتم هستند. حال داریم:

$$\begin{aligned} &E_*(\bar{y}^* - \bar{y})^3 \\ &= \frac{1}{k^3}\Big(\sum_{j=1}^{k} E_*(\bar{y}^*_j - \bar{y})^3 + 3\sum_{j\neq l}^{l} E_*(\bar{y}^*_j - \bar{y})E_*(\bar{y}^*_l - \bar{y})^2 \\ &\quad + \sum_{j\neq l\neq m}^{k} E_*(\bar{y}^*_j - \bar{y})E_*(\bar{y}^*_l - \bar{y})E_*(\bar{y}^*_m - \bar{y})\Big) \\ &= \frac{1}{k^3}\sum_{j=1}^{k} E_*(\bar{y}^*_j - \bar{y})^3 \\ &= \frac{n^2(1-f^*)(1-2f^*)m_3}{(kn'^2)(n-1)(n-2)}. \quad (7) \end{aligned}$$

رابطه اخیر از این حقیقت که $\bar{y}^*_j$ ها از یک نمونه تصادفی با جایگذاری به حجم n' از نمونه اصلی حاصل می شوند و از نتیجه مقاله سوخاتم و سوخاتم [۱۲] به دست می آید. داریم:

$$\begin{aligned} &\frac{n^2(1-f^*)(1-2f^*)m_3}{(kn'^2)(n-1)(n-2)} \\ &= \frac{(1-f_h)(1-2f_h)m_{3h}}{(n_h-1)(n_h-2)}[1-A_h]. \quad (8) \end{aligned}$$

که در آن

$$A_h = \frac{(f^*_h - f_h)}{(1-f^*_h)(1-2f_h)}.$$

با جایگذاری روابط (۷) و (۸) در رابطه (۶) داریم:

$$\begin{aligned} &E_*(\bar{y}^* - \bar{y})^3 \\ &= \sum_{h=1}^{L} W_h^3 \frac{(1-f_h)(1-2f_h)m_{3h}}{(n_h-1)(n_h-2)}[1-A_h]. \end{aligned}$$

A_h در رابطه قبل عامل خطای جورشدن می باشد که اگر کسر بازنمونه گیری با کسر نمونه گیری اصلی برابر باشد $A_h = 0$ خواهد شد. این یعنی اگر $n'_h = f_h = n_h$ باشد گشتاور سوم بازنمونه گیری با گشتاور سوم نمونه یکسان خواهد شد.

۲.۲.۴ تصادفی کردن انتخاب n'_h و k_h در حالت غیر صحیح

در ابتدا می توان به راحتی نشان داد که در الگوی نمونه گیری طبقه ای اگر$n'_h \le n_h/(۲-f_h) \le ۱$ باشد آنگاه

$$k_h = [n(۱-f_h^*)]/[n'(۱-f_h)] \ge ۱$$

خواهد بود. حال اگر k_h عدد صحیحی نباشد متغیر K_h را به صورت زیر تعریف می کنیم:

$$p_h = P(K_h = k_{h۱}) = \left(\frac{۱}{k_h} - \frac{۱}{k_{h۲}}\right) / \left(\frac{۱}{k_{h۱}} - \frac{۱}{k_{h۲}}\right)$$

و $P(K_h = k_{h۲}) = ۱ - p_h$.

که در آن $k_{h۱} = \lfloor k_h \rfloor$ و $k_{h۲} = \lceil k_h \rceil$ بوده که در شرط $۱ \le k_{k۱} \le k_h \le k_{h۲} \le n_h$ صدق می کنند. حال با الگوریتمی که در قسمت قبل ارائه شد و با استفاده از K_h، رویه را به طور مستقل در هر طبقه انجام می دهیم.

قضیه ۶ : با استفاده از روش فوق اگر $\hat{\theta} = \bar{y}$ باشد آنگاه $Var_*(\bar{y}^*) = Var(\bar{y})$ خواهد بود.

اثبات. می دانیم

$$Var_*(\bar{y}^*) = \sum_{h=۱}^{L} W_h^۲ Var(\bar{y}_h^*). \qquad (۹)$$

از آنجایی که نمونه گیری در طبقه ها به طور مستقل انجام می شود، در رابطه فوق $Var(\bar{y}_h^*)$ را برای یک طبقه به دست آورده و قضیه را ثابت می کنیم. لذا با در نظر نگرفتن زیرنویس h داریم:

$$Var_*(\bar{y}^*) = E_{۱*} var_{۲*}(\bar{y}^*) + Var_{۱*} E_{۲*}(\bar{y}^*). \quad (۱۰)$$

که در آن $E_{۱*}$ و $Var_{۱*}$ به ترتیب امید ریاضی و واریانس بازنمونه گیری و $E_{۱*}$ و $Var_{۱*}$ امید ریاضی و واریانس بازنمونه گیری به شرط K هستند. از طرفی

$$\begin{aligned} & E_{۱*} Var_{۲*}(\bar{y}^*) \\ &= p\, Var(\bar{y}^*|K=k_۱) + (۱-p) Var(\bar{y}^*|K=k_۲) \\ &= p \times \frac{۱-f^*}{k_۱ n'} s^۲ + (۱-p) \times \frac{۱-f^*}{k_۲ n'} s^۲ \\ &= \frac{۱-f^*}{n'} \left(\frac{p}{k_۱} + \frac{۱-p}{k_۲}\right) s^۲. \end{aligned}$$

با جایگذاری p در عبارت فوق داریم

$$E_{۱*} Var_{۲*}(\bar{y}^*) = \frac{۱-f^*}{kn'} s^۲.$$

از طرفی

$$Var_{۱*} E_{۲*}(\bar{y}^*) = Var_{۱*}(\bar{y}) = ۰.$$

که تساوی اخیر از استقلال $\bar{y}$ از k نتیجه می‌شود. با استفاده از این نتایج و روابط (؟؟) و(۱۰) اثبات قضیه کامل می شود. □

در روش فوق اگر از فرضیه $n'_h = f_h n_h$ استفاده شود احتمال این وجود دارد که هم n_h و هم k_h غیر صحیح باشند. در این حالت نیز از روش هایی برای تصادفی کردن این دو اختیار است تا تساوی $Var_*(\bar{y}^*) = Var(\bar{y})$ برقرار شود. یک روش تصادفی کردن k به سرط صحیح بودن n' در مرحله قبل اشاره شد. در این قسمت می خواهیم n' را نیز تصادفی در نظر بگیریم. یک روش به این صورت است که فرض کنید $n'_۱ = \lfloor n' \rfloor$، $n_۲ = \lceil n' \rceil$، $k_۱ = \lfloor k \rfloor$ و $k_۲ = \lceil k \rceil$ باشند، آنگاه جفت $(n'_۱, k_۱)$ را با احتمال p و $(n'_۲, k_۲)$ را با احتمال $۱-p$ انتخاب می کنیم که

$$p = \frac{\left(\frac{۱-f}{n} - \frac{۱-n'_۲/n}{k_۲ n'_۲}\right)}{\left(\frac{۱-n'_۱/n}{k_۱ n'_۱} - \frac{۱-n'_۲/n}{k_۲ n'_۲}\right)}.$$

با در نظر گرفتن رویه بیان شده در حالتی که $\hat{\theta} = \bar{y}$ به راحتی می توان نشان داد $Var_*(\bar{y}^*) = Var(\bar{y})$. با در نظر نگرفتن زیر نویس h داریم:

$$\begin{aligned} Var_*(\bar{y}^*) &= p\, Var_*(\bar{y}^*|K=k_1, n'=n'_1) \\ &\quad +(1-p)\; var_*(\bar{y}^*|K=k_2, n'=n'_2) \\ &= p \times \frac{1-n'_1/n}{k_1 n'_1} s^2 \\ &\quad +(1-p) \times \frac{1-n'_2/n}{k_2 n'_2} s^2. \end{aligned}$$

با جایگذاری p در عبارت فوق نتیجه لازم حاصل می شود.

۵ مقایسه عددی

در این قسمت دقت روش های مختلف بازنمونه گیری ارائه شده در این مقاله را برای یک مثال واقعی مقایسه می شوند. برای انجام این کار تولیدات علمی اساتید دانشگاه فردوسی مشهد در نظر گرفته شده اند. جامعه مورد مطالعه جامعه ای است شامل ۱۰ طبقه (دانشکده) و هدف تحقیق برآورد میانگین تولیدات علمی اساتید و تعیین میزان دقت این برآورد با استفاده از محک انحراف معیار است. تولیدات علمی اساتید بر مبنای فعالیت های علمی آنان مانند تألیف کتاب، تألیف مقاله، ترجمه کتاب و مقاله و نظیر آنها محاسبه می شود.

از آنجایی که محاسبه تولیدات علمی تمام اساتید دانشگاه امری وقت گیر است، با اجرای یک الگوی نمونه گیری طبقه ای با در نظر گرفتن دانشکده ها به عنوان طبقه، میانگین و انحراف معیار آن را برآورد می کنیم. در این جامعه $N = ۲۱۰$، $L = ۱۰$ و $N_i = \{۳۳, ۲۵, ۲۲, ۳۰, ۳۳, ۱۴, ۱۳, ۱۲, ۱۳, ۱۵\}$ است. با در نظر گرفتن کران خطای نمونه گیری (D) بر اساس رابطه زیر در تعیین حجم نمونه در الگوی طبقه ای، n را مشخص می کنیم.

$$n_0 = \frac{Z^2_{1-\alpha/2} S^2}{D^2}; \quad n = \frac{n_0}{1 + n_0/N}.$$

در رابطه فوق $Z^2_{1-\alpha/2}$ و S^2 به ترتیب چندک $1-\alpha/2$ توزیع نرمال و واریانس جامعه هستند. با فرض $\alpha = ۰/۰۵$ و $T = ۳۵$ و بر اساس نتایج حاصل از بررسی روی داده های این جامعه توسط آقای گنجی [۱] داریم $S^2 = ۱۲۸۴۰/۹$ و لذا خواهیم داشت $n = ۷۰$ و $f = ۰/۳۳$. با این فرضیات می توان حجم نمونه ها را با اعمال تخصیص متناسب به صورت زیر در نظر گرفت:

$$n_i = \{۱۱, ۸, ۷, ۱۰, ۱۱, ۵, ۴, ۴, ۴, ۶\}.$$

بر اساس شرایط فوق با استفاده از چهار روش بازنمونه گیری معرفی شده در این مقاله به برآورد میانگین جامعه و واریانس آن می پردازیم. نتایج حاصله به ازای $B = ۱۰۰۰$ در جدول ۱ ارائه شده است. لازم به ذکر است که نتایج با استفاده از نرم افزار R به دست آمده اند.

جدول ۱. جدول ارائه شده برای مقایسه روشهای بازنمونه گیری.

پارامتر	$\bar{y}_{st}$	$\bar{y}^*_{bwo}$	$\bar{y}^*_{mm}$	$\bar{y}^*_{wr}$	$\bar{y}^*_{rs}$
میانگین	۲۲۵/۱۸	۲۲۵/۱۳	۲۲۵/۰۱	۲۲۵/۱۰	۲۲۵/۸۳
واریانس	۱۰۸/۹۸	۱۰۶/۲۵	۱۱۰/۰۳	۱۰۲/۴۹	۱۴۷/۴۶

با مشاهده این جدول می توان دید که هر چهار روش بازنمونه گیری تقریباً برآورد یکسانی برای میانگین جامعه دارند و این برآورد ها بسیار به میانگین نمونه ($\bar{y}_{st}$) نزدیک هستند. همان طوری که قبلاً اشاره شد روش های بازنمونه گیری برآوردی برای آماره در نمونه اصلی هستند. به عنوان مثال در این جامعه میانگین برابر ۲۱۷/۵۵ است (بر اساس تحقیق آقای گنجی) و برآوردگر میانگین جامعه یعنی میانگین نمونه برابر ۲۲۵/۱۸. از جدول فوق پیداست که برآوردگرهای بازنونه گیری به میانگین نمونه ای نزدیکتر هستند تا به میانگین جامعه. یعنی آنها برآوردگرهای برای میانگین نمونه هستند. حال هرچه برآوردگر نمونه ای برای پارامتر جامعه دقیقتر باشد بازنمونه گیری نیز مطمئن تر خواهند بود.

در مورد واریانس میانگین نمونه هم بر اساس داده های جدول به نظر می رسد که روش های $M.M$ و BWO از دقت بیشتری برخوردار باشند.

۶ نتیجه گیری

در این مقاله روش های مختلف بازنمونه گیری در الگوی طبقه ای ارائه و مقایسه شده اند. با بررسی های انجام شده می توان گفت که روش $M.M$ کاراتر از دیگر روش های ارائه شده می باشد. این روش برای تمام مدل های نمونه گیری (طبقه ای؛ خوشه ای و نمونه گیری با احتمالات نابرابر) به راحتی قابل اجرا بوده و برای کلاس بزرگی از برآوردگرها از جمله واریانس توابع غیر خطی از میانگین و صدک ها معتبر است [۱۱].

مراجع

[۱] گنجی، م. (۱۳۸۱)، مقایسه تولیدات علمی اساتید دانشکده های دانشگاه فردوسی مشهد، رساله کارشناسی ارشد، دانشگاه فردوسی مشهد.

[2] Bickel, P.J. and Freedman, D.A. (1984), Asymptotic normality and the Bootstrap in stratified sampling., *Ann.Statist.* **12**, 470-482.

[3] Cochran, W.G. (1977), Sampling Techniques., *John Wiely, New York.*

[4] Efron, B. (1979), Bootstrap mothods: another look ot the jackknife., *Ann.Statist.* **7**, 1-26.

[5] Gross, S. (1980), Median Estimation in Sample Survey., *Proceedings of Section on Survey Research Methods, American Statistical Association.* , 181-184.

[6] Krewski, D., and Rao, J. N. K. (1981), Inference From Stratified Samples: Properties of the linearization, Jackknife and Balanced Reapeted Replication Methods., *The Annals of Statistics* **9**, 1010-1019.

[7] Loh, W.Y. (1987), Calibration Confidence Coefficients., *Journal of the American Statistical Assosiation* **82**, 155-162.

[8] McCarthy, P.J., and Snowden, C.B. (1985), The Bootstrap and finite population sampling., *Vital and Health Statistics* **2**, 1-23.

[9] Rao, J.N.K., and Wu, C.F.J. (1988), Resampling interfrence with complex survy data., *Journal of the Amrrican Statistical Association* **83**, 231-241.

[10] Sitter, R.R. (1992a), Comparing three Bootstrap methods for survy data., *The Canadian Journal of Statistics* **20**, 135-154.

[11] Sitter, R.R. (1992b), Resampling procedures for complex survy data., *Journal of the Amrrican Statistical Association* **87**, 135-154.

[12] Sukhatme, P.V. and Sukhatme, B.W. (1970), Sampling theory of survy with applications., *Iowa State University Press, Ames, IA*

[13] Wu, C.F.J. (1985), Variance Estimation for the Compined Ration and Combined Regression Estimators., *Journal of the Royal Statistical Society, Ser.B* **47**, 147-154.