مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۳، شماره‌ی ۱، بهار و تابستان ۱۳۹۱، صص ۱۰۳ – ۱۲۷

براورد واریانس به‌روش جک‌نایف در آمارگیری‌های دوچارچوبی

مرجان نورینی

مرکز آمار ایران

چکیده: روش‌های معمول براورد واریانس از قبیل روش خطی‌سازی سری تیلور (روش دلتا)[1] در آمارگیری‌های چندچارچوبی[2] عموماً مستلزم محاسبه‌ی مشتق‌های جزئی بوده و این محاسبات با افزایش تعداد چارچوب‌ها پیچیده‌تر می‌شود. براورد واریانس به روش جک‌نایف[3] روش دیگری است که ضمن سهولت در محاسبه، موجب کاهش چشم‌گیری در اریبی براوردگر می‌شود. در این مقاله ابتدا به معرفی براوردگرهای چندچارچوبی مجموع جامعه و سپس استفاده از روش جک‌نایف در براورد واریانس براوردگرهای دوچارچوبی و مقایسه‌ی آن با روش خطی‌سازی سری تیلور طی یک مطالعه‌ی شبیه‌سازی می‌پردازیم.

واژگان کلیدی: براورد واریانس جک‌نایف؛ براورد واریانس خطی‌سازی؛ چارچوب‌های چندگانه؛ روش‌های بازنمونه‌گیری.

۱- مقدمه

چارچوب آماری یا فهرست واحدهای آمارگیری، اساس و مبنای یک طرح آمارگیری نمونه‌ای را تشکیل می‌دهد. گاهی ممکن است چارچوبی که تمامی واحدهای جامعه‌ی مورد مطالعه را پوشش دهد در دسترس نباشد، اما امکان دستیابی به پوشش کامل، با تلفیقی از دو یا چند چارچوب فراهم شود. در چنین حالتی به‌منظور دسترسی به پوشش مناسب، از دو چارچوب یا بیش‌تر به‌طور هم‌زمان استفاده می‌شود. گاهی نیز ممکن است یک چارچوب، پوشش کامل را برای جامعه‌ی مورد مطالعه فراهم کند، اما چارچوب ناقص دیگری موجود باشد که هزینه‌ی آمارگیری از آن کم‌تر از هزینه‌ی آمارگیری از چارچوب

دریافت: ۱۳۹۱/۷/۱۵، پذیرش: ۱۳۹۲/۴/۹.

کامل باشد. در این شرایط به‌دلیل پایین‌تر بودن هزینه‌ی آمارگیری از این چارچوب، می‌توان با هزینه‌ای مشخص و ثابت، از دو چارچوب استفاده کرده و اندازه‌ی نمونه را بزرگ‌تر و کارایی را افزایش داد. گاهی نیز ممکن است یک چارچوب فهرستی کامل، در دسترس باشد اما عملاً با گذشت زمانی نسبتاً طولانی به‌دلیل بروز تغییرات فراوان در آن، منبعی برای بروز خطاهای غیر نمونه‌گیری شود. از آن‌جا که یک فهرست ناحیه‌ای، کم‌تر در معرض تغییرات می‌باشد، ترکیب آن با یک چارچوب از اعضای جامعه که احتمالاً ناقص باشد، می‌تواند نتایج مفیدی را حاصل نماید. چنین آمارگیری‌هایی تحت عنوان آمارگیری‌های چندچارچوبی به کار می‌روند.

شاید بتوان گفت اولین شالوده‌ی آمارگیری‌های چندچارچوبی در سال ۱۹۴۹ با آمارگیری از فروشگاه‌های خرده‌فروشی که مجری آن دفتر سرشماری آمریکا بود گذاشته شد. سپس هارتلی [۵] نظریه‌ی مقدماتی چارچوب‌های چندگانه را توسعه داد. او با فرض این‌که اجتماع چارچوب‌ها، جامعه را پوشش می‌دهد، واحدهای جامعه را به زیرمجموعه‌های دو به دو ناسازگار شامل اجتماع‌ها و اشتراک‌های چارچوب‌های مختلف تقسیم کرد. بعد از وی لاند [۹]، فولر و بورمیستر [۴]، وگل [۱۹]، فوردوبوسکر [۳]، بانکیر [۲] ، کالتون و اندرسون [۷]، اسکینر [۱۶] و اسکینر و رائو [۱۷] مقالاتی را در این زمینه ارائه نموده‌اند.

در آمارگیری‌های چندچارچوبی، براوردگرهای مختلفی برای براورد مجموع جامعه پیش‌نهاد شده است. این براوردگرها عموماً توابعی ناخطی از مقادیر نمونه‌ای می‌باشند. برای براورد واریانس این‌گونه براوردگرها، روش‌های مختلفی از جمله خطی‌سازی سری تیلور پیش‌نهاد شده است.

اسکینر و رائو [۱۷] روش خطی‌سازی سری تیلور را پیش‌نهاد دادند. استفاده از این روش مستلزم محاسبه‌ی مشتق‌های جزئی بوده و این محاسبات با افزایش تعداد چارچوب‌ها پیچیده‌تر می‌شود. لیکن براورد واریانس به‌روش جک‌نایف ضمن سهولت در محاسبه، موجب کاهش چشمگیری در اریبی براوردگر می‌شود. در این مقاله، ضمن استفاده از روش جک‌نایف در براورد واریانس، به مقایسه‌ی آن با روش بسط سری تیلور می‌پردازیم. به این منظور در بخش ۲، براوردگرهای چندچارچوبی معرفی می‌شوند. در بخش ۳، ابتدا شرح مختصری از روش جک‌نایف ارایه شده و سپس به کاربرد آن در آمارگیری‌های چندچارچوبی می‌پردازیم. در بخش ۴ نتایج حاصل از مقایسه و بررسی براوردگرهای

چندچارچوبی و براورد واریانس آن‌ها با استفاده از شبیه‌سازی ارایه می‌شود و در نهایت در بخش ۵ نتیجه‌گیری بیان می‌شود.

۲- براوردگرهای مجموع جامعه

برای سادگی فرض کنید دو چارچوب A و B موجود است که هر دو ناقص و دارای واحدهای مشترک با یکدیگر هستند به‌طوری که مجموع آن‌ها روی هم، کل جامعه‌ی مورد مطالعه را پوشش می‌دهد. از چارچوب‌های A و B، سه $(۱-۲^۲)$ حوزه‌ی دو به دو ناسازگار به دست می‌آید.

حوزه‌ی a: شامل واحدهایی است که فقط در چارچوب A می‌باشند. $a = A \bigcap B^c$

حوزه‌ی b: شامل واحدهایی است که فقط در چارچوب B می‌باشند. $b = A^c \bigcap B$

حوزه‌ی ab: شامل واحدهایی است که در هر دو چارچوب می‌باشند. $ab = A \bigcap B$

(c نشان‌دهنده‌ی مکمل مجموعه می‌باشد).

نمادهای N_A، N_B، N_a، N_b و N_{ab} به‌ترتیب تعداد واحدها در چارچوب‌های A و B و حوزه‌های a، b و ab می‌باشند.

دو نمونه‌ی مستقل s_A و s_B بر اساس طرح‌های نمونه‌گیری احتمالاتی $p_A(s_A)$ و $p_B(s_B)$ به اندازه‌ی n_A و n_B از دو چارچوب فوق گرفته می‌شود، به‌طوری که احتمال شمول نمونه‌ی حاصل از چارچوب A برابر است با $\pi_i^A = p\{i \in s_A\}$ و احتمال شمول نمونه‌ی حاصل از چارچوب B برابر است با $\pi_i^B = p\{i \in s_B\}$، ([۸]).

براساس نمونه‌های مستقل فوق، n_a و n_{ab}^A، تعداد واحدهای نمونه‌گیری حاصل از چارچوب A می‌باشند که به‌ترتیب در حوزه‌های a و ab قرار دارند. به‌همین ترتیب n_b و n_{ab}^B نیز تعداد واحدهای نمونه‌گیری حاصل از چارچوب B می‌باشند که به‌ترتیب در حوزه‌های b و ab هستند.

با فرض این که Y_a، Y_b و Y_{ab} به‌ترتیب مجموع جامعه در حوزه‌های a، b و ab باشند، داریم:

$$Y = Y_a + Y_{ab} + Y_b \qquad (۱)$$

هدف براورد Y است.

چندین براوردگر نقطه‌ای تحت عنوان براوردگر دو چارچوبی برای براورد Y پیشنهاد

شده است که همه‌ی آن‌ها به فرم $\hat{Y} = \hat{Y}_a + \hat{Y}_{ab} + \hat{Y}_b$ می‌باشند. هر یک از این براوردگرها، بسته به این که اطلاعات حاصل از دو نمونه برای براورد Y چگونه با هم ترکیب می‌شوند با هم تفاوت‌هایی دارند.

با فرض این که N_A و N_B معلوم بوده و $N_a > ۰$ و $N_b > ۰$، وزن‌های نمونه‌گیری w_i^A و w_i^B عبارت‌اند از:

$$w_i^A = N_A[\pi_i^A \sum_{j\in s_A}(\frac{۱}{\pi_j^A})]^{-۱}, \qquad w_i^B = N_B[\pi_i^B \sum_{j\in s_B}(\frac{۱}{\pi_j^B})]^{-۱}$$

دو متغیر نشانگر زیر را برای هر یک از چارچوب‌های A و B در نظر می‌گیریم:

$$\delta_i^A = \begin{cases} ۱ & \text{واحد } i \text{ متعلق به چارچوب } A \text{ باشد} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

$$\delta_i^B = \begin{cases} ۱ & \text{واحد } i \text{ متعلق به چارچوب } B \text{ باشد} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

در نتیجه براوردگر مجموع در سه حوزه‌ی a و b و ab به‌صورت زیر است:

$$\hat{Y}_a^A = \sum_{i\in s_A} w_i^A \left(۱ - \delta_i^B\right) y_i, \qquad \hat{Y}_{ab}^A = \sum_{i\in s_A} w_i^A \delta_i^B y_i$$
$$\hat{Y}_b^B = \sum_{i\in s_B} w_i^B \left(۱ - \delta_i^A\right) y_i, \qquad \hat{Y}_{ab}^B = \sum_{i\in s_B} w_i^B \delta_i^A y_i$$

براوردگر اندازه‌ی هر یک از حوزه‌ها نیز به‌صورتی مشابه با قراردادن $y_i = ۱$ در تعاریف $\hat{Y}_a^A$ ، $\hat{Y}_{ab}^A$ ، $\hat{Y}_b^B$ ، $\hat{Y}_{ab}^B$ به دست می‌آید. همچنین دو براوردگر دیگر را نیز به‌صورت زیر تعریف می‌کنیم:

$$\hat{Y}_{ab}(\theta) = \theta\, \hat{Y}_{ab}^A + (۱-\theta)\, \hat{Y}_{ab}^B$$
$$\hat{N}_{ab}(\theta) = \theta\, \hat{N}_{ab}^A + (۱-\theta)\, \hat{N}_{ab}^B$$

که مقدار بهینه‌ی θ واریانس $\hat{Y}_{ab}(\theta)$ را مینیمم می‌کند.

۱-۲- براوردگرهای هارتلی و فولر- بورمیستر

هارتلی [۵] براوردگر دوچارچوبی زیر را برای براورد مجموع جامعه پیش‌نهاد کرد:

$$\hat{Y}_H(\theta) = \hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}(\theta) \qquad (۲)$$

فولر و بورمیستر [۴] هم براوردگر زیر را پیش‌نهاد دادند:

$$\hat{Y}_{FB}\left(\beta_1, \beta_2\right) = \hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}\left(\beta_1\right) + \beta_2(\hat{N}_{ab}^A - \hat{N}_{ab}^B) \qquad (۳)$$

مقادیر بهینه‌ی پارامترهای θ، β_1 و β_2 به ترتیب واریانس براوردگرهای $\hat{Y}_H(\theta)$ و $\hat{Y}_{FB}(\beta_1, \beta_2)$ را مینیمم کرده و بنا بر این به کوواریانس‌های $\hat{Y}_{ab}^A$ و $\hat{Y}_{ab}^B$ بستگی دارند. اما در عمل این کوواریانس‌ها مجهول بوده و باید از روی نمونه براورد شوند. در نتیجه $\hat{Y}_H(\hat{\theta}_H)$ و $\hat{Y}_{FB}(\hat{\beta}_{FB})$ به‌طور کلی توابعی خطی از y نخواهند بود و برای هر متغیر پاسخ باید به‌طور جداگانه محاسبه شوند. برای مثال اگر $\hat{Y}_{H_1}$، $\hat{Y}_{H_2}$ و $\hat{Y}_{H_3}$ براوردگرهای هارتلی برای تعداد کل بیمارانی که دچار تنگی نفس در گروه‌های سنی ۰- ۱۶ و ۱۷- ۴۵ و ۴۵ به بالا باشد، آن گاه $\hat{Y}_{H_1} + \hat{Y}_{H_2} + \hat{Y}_{H_3}$ لزوماً مساوی براورد هارتلی تعداد کل بیماران دچار تنگی نفس در جامعه نخواهد بود. در نتیجه فولرو بورمیستر [۴] بر آن شدند براوردگری ارایه دهند که از یک مجموعه وزن‌های یکسان برای همه‌ی متغیرهای پاسخ استفاده کند. آن‌ها $\hat{Y}_{FB,srs}$ را برای براورد Y معرفی کردند که این براوردگر، تحت طرح نمونه‌گیری تصادفی ساده طراحی شده بود و نمی‌توانست مستقیماً با طرح‌های نمونه‌گیری پیچیده به کار رود. چون عموماً به‌صورتی ناسازگار با این نوع نمونه‌گیری‌ها طراحی شده بود.

۲-۲- براوردگر ماکسیمم درستنمایی‌نما[۴]

اسکینر [۱۶] از $\hat{Y}_{FB,srs}$ تفسیری به‌صورت یک براوردگر ماکسیمم درستنمایی داشت. چون این براوردگر از مجموعه‌ی یکسانی از وزن‌ها استفاده می‌کرد، به براوردگر فولر- بورمیستر ارجحیت داشت. ولی این براوردگر، مستقیماً نمی‌توانست با طرح‌های نمونه- گیری پیچیده به کار رود. چون به‌طور کلی به‌صورتی ناسازگار با این نوع طرح‌های نمونه- گیری طراحی شده بود.

اسکینر و رائو [۱۸] تصمیم گرفتند این براوردگر را طوری اصلاح کنند که به براوردگری سازگار تحت طرح‌های نمونه‌گیری پیچیده نائل آیند. آن‌ها در این حالت، براوردگر ماکسیمم درستنمایی‌نما را برای براورد مجموع پیش‌نهاد کردند. این براوردگر، اصلاح‌شده‌ی یک براوردگر ماکسیمم درستنمایی تحت طرح نمونه‌گیری تصادفی ساده، برای رسیدن به براوردگری سازگار با طرح‌های نمونه‌گیری پیچیده بود.
آن‌ها براوردگر ماکسیمم درستنمایی‌نما را به‌صورت زیر ارایه دادند:

$$\hat{Y}_{PML}(\theta)=\frac{N_A-\hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a^A}\ \hat{Y}_a^A+\hat{N}_{ab}^{PML}(\theta)\ \frac{\hat{Y}_{ab}(\theta)}{\hat{N}_{ab}(\theta)}+\frac{N_B-\hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b^B}\hat{Y}_b^B \qquad (۴)$$

به‌طوری که $\hat{N}_{ab}^{PML}(\theta)$ تابعی از $\hat{N}_{ab}^{A}$ ، $\hat{N}_{ab}^{B}$ و θ بوده و کوچک‌ترین ریشه‌ی معادله‌ی درجه‌ی دو، به‌صورت زیر است:

$$(\frac{\theta}{N_B}+\frac{۱-\theta}{N_A})x^۲-(۱+\frac{\theta}{N_B}\hat{N}_{ab}^{A}+\frac{۱-\theta}{N_A}\hat{N}_{ab}^{B})x+\hat{N}_{ab}(\theta)=۰$$

این براوردگر بر خلاف براوردگرهای هارتلی و فولر- بورمیستر، تابعی خطی از y است. اسکینر و رائو [۱۷] انتخاب $\theta=\theta_P$ را برای مینیمم کردن واریانس مجانبی $\hat{N}_{ab}^{PML}(\theta)$ با

$$\theta_p=\frac{N_aN_BV(\hat{N}_{ab}^{B})}{N_aN_BV(\hat{N}_{ab}^{B})+N_bN_AV(\hat{N}_{ab}^{A})} \qquad (۵)$$

پیش‌نهاد کردند. در عمل N_a و N_b و واریانس‌ها در رابطه‌ی (۵) مجهول بوده و باید از روی داده‌ها براورد شوند.

۳-۲- براوردگرهای تک‌چارچوبی[۵]

بانکیر [۲]، کالتون و اندرسون [۷] و اسکینر [۱۶] برای براورد مجموع جامعه با در نظر گرفتن مشاهدات به‌گونه‌ای که آن‌ها از یک چارچوب با وزن‌های اصلاح‌شده برای مشاهدات در حوزه‌ی متداخل ab نمونه‌گیری شده‌اند، براورگری ارایه دادند. وزن‌های اصلاح‌شده برای براوردگرهای تک چارچوبی اسکینر [۱۶] و کالتون و اندرسون [۷] به

شناسایی واحدهای مشابه در نمونه‌ها نیازی ندارد. این وزن‌ها عبارت‌اند از:

$$w_i = \begin{cases} (\pi_i^A + \pi_i^B)^{-1} & i \in ab \\ (\pi_i^A)^{-1} & i \in a \\ (\pi_i^B)^{-1} & i \in b \end{cases}$$

آن‌ها براوردگر زیر را برای براورد مجموع پیش‌نهاد کردند:

$$\hat{Y}_{SF} = \sum_{i\in s_A} w_i y_i + \sum_{i\in s_B} w_i y_i \qquad (۶)$$

این براوردگر قابل تعمیم به بیش از دو چارچوب می‌باشد و از هیچ گونه اطلاعات کمکی در ارتباط با N_A و N_B استفاده نمی‌کند. از آن جایی‌که استفاده از اطلاعات کمکی، باعث کاهش واریانس و افزایش دقت براوردگر می‌شود [۲]، اسکینر و رائو [۱۷] دو روش برای تصحیح این براوردگر ارایه دادند که عبارت‌اند از:

۱) روش تعدیل نسبتی [۲]
۲) روش رگرسیونی

اسکینرو رائو [۱۷] نشان دادند که فرایند تعدیل به براوردگر تجربی زیر همگرا است:

$$\hat{Y}_{SFrake} = \frac{N_A - \hat{N}_{ab}^{rake}}{\hat{N}_a}\hat{Y}_a + \frac{\hat{N}_{ab}^{rake}}{\hat{N}_{abS}}\hat{Y}_{abS} + \frac{N_B - \hat{N}_{ab}^{rake}}{\hat{N}_b}\hat{Y}_b \qquad (۷)$$

که در آن:

$$\hat{Y}_{abS} = \sum_{s_A} w_i \delta_i^B y_i + \sum_{s_B} w_i \delta_i^A y_i$$
$$\hat{N}_{abS} = \sum_{s_A} w_i \delta_i^B + \sum_{s_B} w_i \delta_i^A$$

و $\hat{N}_{ab}^{rake}$ کوچک‌ترین ریشه‌ی معادله‌ی درجه‌ی دو به‌صورت زیر است:

$$\hat{N}_{abS}x^2 - \left[\hat{N}_{abS}(N_A + N_B) + \hat{N}_{aS}^A \hat{N}_{bS}^B\right]x + \hat{N}_{abS}N_A N_B = 0 \qquad (۸)$$

که $\hat{N}_{as}^A = \sum_{s_A} w_i\, \delta_i^A$ و $\hat{N}_{bs}^B = \sum_{s_B} w_i\, \delta_i^B$.

جزئیات محاسبه‌ی این براوردگر در [۱] آورده شده است.

تصحیح با N_A و N_B معلوم به‌عنوان دو متغیر کمکی مورد استفاده که به‌ترتیب با $\hat{N}_{AS}$ و $\hat{N}_{BS}$ براورد می‌شوند، از طریق روش رگرسیونی، براوردگر زیر را نتیجه می‌دهد:

$$\hat{Y}_{SFreg} = \hat{Y}_S + \hat{\beta}_{S۲}^T \left[N_A - \hat{N}_{AS}, N_B - \hat{N}_{BS}\right] = ۰ \qquad (۹)$$

به‌طوری که مقدار بهینه‌ی $\hat{\beta}_{s۲}$ برابر است با:

$$\hat{\beta}_{S۲} = -Cov\{\left[\frac{\hat{N}_{AS}}{V(\hat{N}_{AS})}, \frac{\hat{N}_{BS}}{V(\hat{N}_{BS})}\right], \hat{Y}_S\}$$

۳- براورد واریانس

روش جک‌نایف براوردهایی از پارامتر مورد مطالعه را از هر یک از زیرنمونه‌های نمونه‌ی اصلی به دست آورده و سپس براورد واریانس براوردگر نمونه‌ی اصلی را از تغییرپذیری بین براوردگرهای این زیر نمونه‌ها نتیجه می‌دهد.

نخستین بار کنولی [۱۰]، جک‌نایف را به‌عنوان روشی بر اساس حذف هر بار یک مشاهده از مجموعه داده‌های اولیه و محاسبه‌ی مجدد براوردگر با استفاده از بقیه‌ی داده‌ها برای کاهش اریبی براوردگر ضریب همبستگی پیاپی معرفی نمود. در مقاله‌ای دیگر [۱۱] این روش را تعمیم داده و خواص کلی کاهش اریبی آن را در یک جامعه‌ی متناهی بیان کرد. سپس توکی [۱۸] پیش‌نهاد کرد که این روش علاوه بر کاهش اریبی، می‌تواند برای براورد واریانس نیز استفاده شود.

فرض می‌کنیم $X_۱,.....,X_n$ متغیرهای تصادفی مستقل و هم توزیع با مقادیر حقیقی از یک توزیع نامعلوم F باشند و آماره‌ی $\hat{\theta}(X_۱,.....,X_n)$، براوردگری برای پارامتر نامعلوم θ بوده که از روی نمونه‌ی اصلی محاسبه می‌شود. $\hat{\theta}_i$ براوردگری برای $\hat{\theta}$ بر اساس نمونه‌ای به اندازه‌ی $n-۱$ است که بعد از حذف i امین واحد نمونه به دست می‌آید. آن‌گاه براوردگر جک‌نایف $\hat{\theta}$ این نمونه عبارت است از:

$$\hat{\theta}_J = (n-۱)^{-۱} \sum_{i=۱}^{n} \left[n\hat{\theta} - (n-۱)\hat{\theta}_i\right]$$

براورد واریانس جک‌نایف براوردگر بالا عبارت است از:

$$v_J(\hat{\theta}) = n^{-1}\left(n-1\right)\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta})^2$$

همان‌طور که ملاحظه می‌شود روش جک‌نایف به فرضیات مدل بستگی کم‌تری داشته و نیازی به فرمول‌های نظری که در روش سنتی می‌باشد، ندارد و در مجموع به‌عنوان یک روش ناپارامتری در تحلیل‌های آماری شناخته می‌شود. با این حال، جک‌نایف نیاز به محاسبه‌ی مکرر آماره به تعداد n مرتبه دارد که واقعاً این کار در قدیم امکان‌پذیر نبود.

در این بخش، برآورد جک‌نایف واریانس براوردگرهای دوچارچوبی ارایه شده و سپس هم‌ارزی مجانبی آن با براوردگر خطی‌سازی واریانس (براوردگری که اسکینر و رائو با استفاده از روش خطی‌سازی سری تیلور به دست آوردند [۱۷]) اثبات می‌شود.

فرض کنید چارچوب‌های A و B به‌ترتیب دارای H و L طبقه باشند به‌طوری که طبقات h و l در هر یک از آن‌ها به‌ترتیب شامل N_h^A و N_l^B واحد بوده و در مجموع $\widetilde{N}_h^A$ و $\widetilde{N}_l^B$ واحد نمونه‌گیری اولیه (psu) دارند که به‌ترتیب $\widetilde{n}_h^A$ و $\widetilde{n}_l^B$ واحد از آن‌ها نمونه‌گیری می‌شود. به‌طوری که $\widetilde{n}_A = \sum_{h=1}^{H}\widetilde{n}_h^A$ و $\widetilde{n}_B = \sum_{l=1}^{L}\widetilde{n}_l^B$ کل نمونه‌ای است که به‌ترتیب از چارچوب‌های A و B انتخاب می‌شود. وزن طبقات h و l در چارچوب‌های A و B به‌ترتیب با $W_h^A = N_h^A/N_A$ و $W_l^B = N_l^B/N_B$ نشان داده می‌شود [۸]. لازم به یادآوری است که اغلب در نمونه‌گیری‌های چندمرحله‌ای، نمونه‌ی بدون جایگذاری از واحدهای نمونه‌گیری اولیه با احتمال متناسب با اندازه (به‌دلیل کارایی بالای نمونه‌گیری بدون جایگذاری نسبت به نمونه‌گیری با جایگذاری) به عمل آمده و سپس در مرحله‌ی برآورد واریانس، برای صرفه‌جویی در زمان و هزینه و اجتناب از محاسبات بالقوه پرزحمت احتمال‌های شمول توأم، نمونه‌گیری را به‌گونه‌ای در نظر می‌گیرند که گویی با جایگذاری انتخاب شده است [۱۵].

این تقریب عموماً منجر به بیش براورد شدن واریانس براوردگر مجموع می‌شود اما در صورتی که کسرهای نمونه‌گیری مرحله‌ی اول بزرگ نباشد، اریبی نسبی بزرگ نخواهد بود. براوردگر واریانس جک‌نایف و دیگر براوردگرهای واریانس بازنمونه‌گیری از این روش در برآورد واریانس استفاده می‌کنند.

حال فرض کنید که نمونه‌هایی به اندازه‌ی (≥ 2) $\widetilde{n}_h^A$ و $\widetilde{n}_l^B$ به‌ترتیب با احتمال‌های

شمول متناسب با اندازه، $\tilde{\pi}_{hi}^{A}=\tilde{n}_{h}^{A}\,p_{hi}^{A}$ و $\tilde{\pi}_{lj}^{B}=\tilde{n}_{l}^{B}\,p_{lj}^{B}$ از چارچوب‌های A و B انتخاب می‌شود، به‌طوری که p_{hi}^{A} و p_{lj}^{B} به‌ترتیب برابر با احتمال انتخاب واحد iام و jام در طبقات h و l متناسب با اندازه‌ی psuها بوده و $\sum_{i}p_{hi}^{A}=1$ و $\sum_{j}p_{lj}^{B}=1$ ($p_{hi}^{A}=u_i\Big/\sum_{A}u_i$ و $p_{lj}^{B}=u_j\Big/\sum_{B}u_j$ که u_i اندازه‌ی تقریبی iامین psu است).

با در نظر گرفتن دو بردار $A=(Y_a\,,\,Y_{ab}\,,N_{ab}\,,N_A)^T$ و $B=(Y_b\,,\,Y_{ab}\,,N_{ab}\,,N_B)^T$، آن‌گاه براورد A برابر است با:

$$\hat{A}=\sum_{h=1}^{H}\sum_{i=1}^{\tilde{n}_h^A}\frac{\hat{A}_{hi}}{\tilde{\pi}_{hi}^{A}}=\sum_{h=1}^{H}\sum_{i=1}^{\tilde{n}_h^A}\frac{\hat{A}_{hi}}{\tilde{n}_h^A p_{hi}^{A}}=\sum_{h=1}^{H}\sum_{i=1}^{\tilde{n}_h^A}\frac{N_h^A a_{hi}}{\tilde{n}_h^A}=\sum_{h=1}^{H}\sum_{i=1}^{\tilde{n}_h^A}N_h^A\bar{a}_h$$

که $\hat{A}_{hi}$ براوردگر ناااریب بردار مجموع A_{hi} بر اساس نمونه‌گیری در مرحله‌ی دوم و مراحل بعدی بوده و $a_{hi}=\hat{A}_{hi}\big/(N_h^A\,p_{hi}^{A})$ می‌باشد. براوردگر $\hat{B}=\sum_{l=1}^{L}\sum_{j=1}^{\tilde{n}_l^B}N_l^B\,b_{lj}\big/\tilde{n}_l^B$ نیز به‌صورتی مشابه تعریف می‌شود. تحت فرضیات نمونه‌گیری باجایگذاری، a_{hi}ها، براوردگرهای ناااریب مستقل و هم‌توزیع میانگین جامعه در طبقه‌ی hام چارچوب A با $E(a_{hi})=\overline{A}_h=A_h\big/N_h^A$ می‌باشند در حالی‌که برای $h\neq h'$، a_{hi} و $a_{h'i'}$ مستقل بوده اما لزوماً هم‌توزیع نیستند. به‌طور مشابه b_{lj}ها هم براوردگرهای ناااریب مستقل و هم توزیع بردار میانگین جامعه در طبقه‌ی lام چارچوب B با $E(b_{lj})=\overline{B}_l=B_l\big/N_l^B$ می‌باشند.

پارامتری به‌صورت $\tau=g(\overline{A},\overline{B})$ را که تابعی از میانگین‌های جامعه، $\overline{A}=A/N$ و $\overline{B}=B/N$ است در نظر می‌گیریم. میانگین‌های بالا به‌ترتیب با $\hat{\overline{A}}=\sum_{h=1}^{H}w_h^A\bar{a}_h$ و $\hat{\overline{B}}=\sum_{l=1}^{L}w_l^B\bar{b}_l$ براورد شده و $\hat{\tau}$ با $\hat{\tau}=g(\hat{\overline{A}},\hat{\overline{B}})$ براورد می‌شود [۸].

براوردگرهای دوچارچوبی می‌توانند در قالب $\hat{\tau}$ با در نظر گرفتن مجموع جامعه به‌صورت $Y=N\overline{Y}=N\,g(\overline{A},\overline{B})$ بیان شوند. در این صورت میانگین $\overline{Y}$ برابر است با:

$$(۱۰) \quad g(\bar{A},\bar{B}) = \bar{Y} = \left(\frac{N_A}{N}\right)\left(\bar{A}_۱ + \theta\bar{A}_۲\right) + \left(\frac{N_B}{N}\right)\left\{\bar{B}_۱ + (۱-\theta)\bar{B}_۲\right\}$$

در این صورت براوردگر هارتلی برابر با $\hat{Y}_H(\theta_H) = N\, g(\hat{\bar{A}},\hat{\bar{B}})$ خواهد بود [۱].

ماتریس واریانس بردار $\hat{\bar{A}}$ برابر با $\Sigma_A = \sum_{h=۱}^{H}(W_h^A)^۲\, \Sigma_h^A/\tilde{n}_h^A$ بوده که با $S_A = \sum_{h=۱}^{H}(W_h^A)^۲\, S_h^A/\tilde{n}_h^A$ براورد می‌شود. به‌طور مشابه، ماتریس واریانس $\hat{\bar{B}}$ یعنی Σ_B نیز با $S_B = \sum_{l=۱}^{L}(W_l^B)^۲\, S_l^B/\tilde{n}_l^B$ براورد می‌شود که S_h^A و S_l^B به‌ترتیب براوردگرهای ماتریس واریانس $\sqrt{\tilde{n}_h^A}\,\bar{a}_h$ و $\sqrt{\tilde{n}_l^B}\,\bar{b}_l$ می‌باشند و عبارت‌اند از:

$$S_h^A = (\tilde{n}_h^A - ۱)\sum_{i=۱}^{\tilde{n}_h^A}(a_{hi}-\bar{a}_h)(a_{hi}-\bar{a}_h)^T$$

$$S_l^B = (\tilde{n}_l^B - ۱)\sum_{j=۱}^{\tilde{n}_l^B}(b_{lj}-\bar{b}_l)(b_{lj}-\bar{b}_l)^T$$

برای اثبات سازگاری براوردگر جک‌نایف واریانس و همچنین هم‌ارزی مجانبی آن با براوردگر خطی‌سازی واریانس، نیاز به شرط‌های زیر است:

آ) فرض کنیم که وقتی $\tilde{n}_A, \tilde{n}_B \to \infty$ آن‌گاه[۶]: $W_h^A\,\tilde{n}_A/\tilde{n}_h^A = \mathrm{O}(۱)$ و $W_l^B\,\tilde{n}_B/\tilde{n}_l^B = \mathrm{O}(۱)$.

ب) فرض می‌کنیم $g_A(\tilde{a},\tilde{b})$ برداری q بعدی از مشتق‌های اول تابع $g(\cdot)$ نسبت به مؤلفه‌های a، $g_B(\tilde{a},\tilde{b})$ برداری r بعدی از مشتق‌های اول تابع $g(\cdot)$ نسبت به مؤلفه‌های b باشد که در $\tilde{a}$ و $\tilde{b}$ مقدار دهی می‌شوند. همچنین فرض می‌کنیم $g_A''(\tilde{a},\tilde{b})$ یک ماتریس q از مشتق‌های دوم، $\partial^۲ g(\partial a_j \partial a_k)$ و $g_B''(\tilde{a},\tilde{b})$ یک ماتریس r بعدی از مشتق‌های دوم $\partial^۲ g(\partial b_j \partial b_k)$ در $\tilde{a}$ و $\tilde{b}$ باشند و g_A'' و g_B'' در یک همسایگی از $(\bar{A},\bar{B})$ پیوسته و کراندار باشند.

ج) $\tilde{n} = \tilde{n}_A + \tilde{n}_B$ به‌طوری که $\tilde{n}_A/\tilde{n} \to k \in (۰,۱)$.

شرط‌های بالا توسط رائو و وو [۱۲] برای تحقیق و بررسی خواص براوردگرهای

واریانس در طرح نمونه‌گیری طبقه‌بندی چندمرحله‌ای مورد استفاده قرار گرفته است.

شرط آ اذعان می‌دارد که تعداد طبقات در چارچوب‌ها می‌تواند کراندار یا بی‌کران باشد. اگر اندازه‌ی نمونه‌ی حاصل از طبقات در چارچوب‌های A و B کراندار باشد یعنی $\tilde{n}_h^A = \mathrm{O}(1)$ و $\tilde{n}_l^B = \mathrm{O}(1)$ آن گاه شرط آ هم‌ارز است با $W_h^A = \mathrm{O}(\tilde{n}_A^{-1})$ و $W_h^B = \mathrm{O}(\tilde{n}_B^{-1})$.

در مجموع فرض می‌شود $\sum_h W_h^A \Sigma_h^A = \mathrm{O}(1)$ و $\sum_l W_l^B \Sigma_l^B = \mathrm{O}(1)$ یعنی متوسط وزنی کواریانس درون طبقات کراندار باشد.

شرط ج نیز تضمین می‌کند که نمونه‌ی حاصل از یک چارچوب، به‌طور مجانبی روی نمونه‌ی چارچوب دیگر هیچ تأثیری نمی‌گذارد.

قضیه‌ی ۱. فرض می‌کنیم شروط آ تا ج برقرار باشند. آن‌گاه:

$$Var(\hat{\tau}) = g_A^T(\bar{A},\bar{B})\Sigma_A g_A(\bar{A},\bar{B}) + g_B^T(\bar{A},\bar{B})\Sigma_B g_B(\bar{A},\bar{B}) + o\left(\tilde{n}^{-1}\right) \quad (۱۱)$$

($o(\tilde{n}^{-1})$ عبارت از جملاتی از مرتبه‌ی کوچک‌تری از $\tilde{n}^{-1}$ است)

براوردگر خطی‌سازی واریانس بالا عبارت است از:

$$(۱۲) \quad v_L(\hat{\tau}) = g_A^T\left(\hat{\bar{A}},\hat{\bar{B}}\right) S_A g_A\left(\hat{\bar{A}},\hat{\bar{B}}\right) + g_B^T\left(\hat{\bar{A}},\hat{\bar{B}}\right) S_B g_B\left(\hat{\bar{A}},\hat{\bar{B}}\right).$$

اثبات قضیه و همچنین اثبات سازگاری براوردگر $v_L(\hat{\tau})$ در [۱] آمده است.

فرض می‌کنیم $\hat{\tau}_{(hi)}^A$ براوردگری به‌صورت $\hat{\tau}$ باشد که بعد از حذف مشاهده‌ی i امین psu نمونه‌ی طبقه‌ی h ام به دست آمده و $\hat{\tau}_{(hi)}^A = g(\hat{\bar{A}}_{(hi)}, \hat{\bar{B}})$ که $\hat{\bar{A}}_{(hi)}$ نیز براوردگر $\bar{A}$ بعد از حذف i امین psu نمونه‌ی طبقه‌ی h ام در چارچوب A می‌باشد. به‌طور مشابه $\hat{\tau}_{(lj)}^B = g(\hat{\bar{A}}, \hat{\bar{B}}_{(lj)})$ که $\hat{\bar{B}}_{(lj)}$ نیز براوردگر $\bar{B}$ بعد از حذف j امین psu نمونه‌ی طبقه‌ی l ام در چارچوب B است. براوردگر جک‌نایف واریانس $\hat{\tau}$ (با توجه به استقلال نمونه‌ها) عبارت است از (جزئیات در [۱] دیده شود):

$$(۱۳) \quad \upsilon_J(\hat{\tau}) = \sum_{h=۱}^{H} \frac{\tilde{n}_h^A - ۱}{\tilde{n}_h^A} \sum_{i=۱}^{\tilde{n}_h^A} (\hat{\tau}_{(hi)}^A - \hat{\tau})^۲ + \sum_{l=۱}^{L} \frac{\tilde{n}_l^B - ۱}{\tilde{n}_l^B} \sum_{j=۱}^{\tilde{n}_l^B} (\hat{\tau}_{(lj)}^B - \hat{\tau})^۲$$

رائو و لوهر [۱۴] در قضیه‌ی ۲ (قضیه‌ی زیر) ثابت کردند که $\upsilon_J(\hat{\tau})$ و $\upsilon_L(\hat{\tau})$ به‌طور مجانبی هم‌ارز هستند.

قضیه‌ی ۲. با فرض این که شروط آ- ج برقرار باشند. آن‌گاه:

$$(۱۴) \qquad \upsilon_J(\hat{\tau}) = \upsilon_L(\hat{\tau}) + o_p(\tilde{n}^{-۱})$$

($o_p(\tilde{n}^{-۱})$ عبارت است از جملاتی از مرتبه‌ی کوچک‌تری از $\tilde{n}^{-۱}$ در احتمال)[۷].

براوردگر جک‌نایف واریانس بالا در حالتی که $\hat{\tau}$ تابع همواری از $\hat{\bar{A}}$ و $\hat{\bar{B}}$ باشد، مورد بررسی قرار گرفته و سازگاری و نیز هم ارزیش با براوردگر خطی‌سازی واریانس، طی قضیه‌های ۱ و ۲ ثابت می‌شود [۱]. در بین براوردگرهای دوچارچوبی، براوردگرهای SF و SF_{rake} به‌صورت توابع همواری از میانگین جوامع، قابل بیان هستند. دیگر براوردگرها نیز مادامی که پارامترهای θ_p ، θ_H ، β_{FB} و $\beta_{S۲}$ ثابت بوده و از روی داده‌ها براورد نشوند، نیز توابع همواری از میانگین جوامع خواهند بود. بنا بر این قضیه‌ی ۲ بیان می‌دارد که براوردگر جک‌نایف واریانس برای شکل بهینه‌ی هر براوردگر سازگار است [۱۴].

۱-۳- جک‌نایف کامل و اصلاح‌شده

براوردگرهای $\hat{\theta}_p$ ، $\hat{\theta}_H$ ، $\hat{\beta}_{FB}$ و $\hat{\beta}_{S۲}$ همگی توابعی از ماتریس‌های واریانس S_A و S_B هستند و به‌طور کلی نمی‌توانند به‌صورت توابع همواری از میانگین در نمونه‌گیری طبقه‌بندی بیان شوند (واریانس‌ها معمولاً نامعلوم بوده و باید از روی نمونه براورد شوند). بنا بر این قضیه‌ی ۲ همیشه نمی‌تواند به‌طور مستقیم با براوردگرهایی که توابعی از S_A و S_B هستند به کار رود. مشکل دیگر که در استفاده از روش جک‌نایف در خیلی از نمونه‌های طبقه‌بندی رخ می‌دهد، هنگام محاسبه‌ی $S_{(hi)}^A$ و $S_{(lj)}^B$ است، چون

$$(۱۵)\quad S^{A}_{(hi)} = S_A + \frac{(W^{A}_{h})^{۲}}{\tilde{n}^{A}_{h}-۲}\left[\frac{۲S^{A}_{h}}{\tilde{n}^{A}_{h}} - \frac{\tilde{n}^{A}_{h}}{\left(\tilde{n}^{A}_{h}-۱\right)^{۲}}(a_{hi}-\bar{a}_h)(a_{hi}-\bar{a}_h)^T\right]$$

در صورتی که $\tilde{n}^{A}_{h}=۲$ باشد رابطه‌ی (۱۵) قابل محاسبه نیست. به‌طور مشابه $S^{B}_{(lj)}$ نیز وقتی که $\tilde{n}^{B}_{l}=۲$ قابل محاسبه نیست. برای رفع این مشکل رائو و لوهر [۱۴] یک براورد جک‌نایف اصلاح‌شده در حالتی که $\tilde{n}^{A}_{h}=۲$ و یا $\tilde{n}^{B}_{l}=۲$ ارایه دادند.

جک‌نایف (جک‌نایف اصلاح‌شده برای طرح نمونه‌گیری با ۲ واحد نمونه‌گیری اولیه در هر طبقه که بعداً به معرفی آن می‌پردازیم) حتی زمانی که براوردگرها به S_A و S_B بستگی دارند، براورد سازگاری از واریانس می‌دهد [۱۴].

در $\hat{\tau}$ فرض بر این است که پارامترهای θ_p، θ_H، β_{FB} و $\beta_{S۲}$ ثابت بوده و از روی نمونه براورد می‌شوند. اما اگر این پارامتر ثابت نبوده و به S_A و S_B بستگی داشته باشند باید اثر این براوردگرها در $\hat{\tau}$ منظور شود. فرض می‌کنیم که

$$\hat{\tau} = g\left(\hat{\bar{A}},\hat{\bar{B}}\right) = f\left(\hat{\bar{A}},\hat{\bar{B}},\beta\right)$$
$$\beta = [E_A + E_B]^{-۱}[e_A + e_B]$$

که هر عنصر ماتریس p بعدی E_A و هر عنصر بردار r بعدی e_A ترکیب خطی از عناصر $\sum_A$ هستند و این ویژگی برای E_B و e_B برقرار است. آن‌گاه براوردگرهای $\hat{Y}_{PML}(\hat{\theta}_p)$، $\hat{Y}_H(\hat{\theta}_H)$، $\hat{Y}_{FB}(\hat{\beta}_{FB})$ و $\hat{Y}_{SFreg}(\hat{\beta}_{S۲})$ همگی به‌صورت $N\hat{\varsigma} = N f(\hat{\bar{A}},\hat{\bar{B}},\hat{\beta})$ می‌باشند به‌طوری که $\hat{\beta}$ براورد S_A و S_B را در β جایگزین می‌کند. برای مثال اگر بردارهای A و B را به‌صورت زیر در نظر بگیریم:

$$A = (Y_a, Y_{ab}, N_{ab}, N_A)^T \qquad B = (Y_a, Y_{ab}, N_{ab}, N_B)^T$$

آن گاه براوردگر هارتلی $\hat{Y}_H(\hat{\theta}_H)$ را می‌توان به‌صورت زیر نوشت:

$$\hat{Y}_H\left(\hat{\theta}_H\right) = \hat{Y}_H(\theta_H) + (\hat{\beta}_H - \beta_H)\left[N_A(\hat{\bar{A}}_۲ - \bar{A}_۲) - N_B(\hat{\bar{B}}_۲ - \bar{B}_۲)\right]$$

به‌طوری که مقدار بهینه‌ی β_H برابر است با:

$$\beta_H = -\frac{k^{۲}\sum_A(۱,۲) - \sum_B(۱,۲) - \sum_B(۲,۲)}{k^{۲}\sum_A(۲,۲) + \sum_B(۲,۲)}, \qquad k = \frac{N_A}{N_B}$$

$\Sigma_A(۱,۲)$ عبارت است از عنصر $(۱,۲)$ امین Σ_A. برای براوردگر ماکسیمم درستنمایی‌نما، پارامتر مناسب عبارت است از $\beta_p = k^۲ \Sigma_A(۳,۳)/\Sigma_B(۳,۳)$.

فرض کنید $\hat{\varsigma}^A_{(hi)}$، براوردگری به‌صورت $\hat{\varsigma}$ باشد که بعد از حذف مشاهدات i امین psu نمونه‌ی طبقه‌ی h ام حاصل شده است. آنگاه $\hat{\varsigma}^A_{(hi)} = f(\hat{\bar{A}}_{(hi)}, \hat{\bar{B}}, \hat{\beta}^A_{(hi)})$ که $\hat{\beta}^A_{(hi)}$ نیز براوردگر β با استفاده از $S^A_{(hi)}$ و S_B با فرض $\tilde{n}^A_h > ۲$ می‌باشد. به‌طور مشابه $\hat{\varsigma}^B_{(lj)} = f(\hat{\bar{A}}, \hat{\bar{B}}_{(lj)}, \hat{\beta}^B_{(lj)})$ است به‌شرط این که $\tilde{n}^B_l > ۲$ باشد. آنگاه براوردگر جک‌نایف واریانس $\hat{\varsigma}$ عبارت است از:

$$v_J(\hat{\varsigma}) = \sum_{h=۱}^{H} \frac{\tilde{n}^A_h - ۱}{\tilde{n}^A_h} \sum_{i=۱}^{\tilde{n}^A_h} (\hat{\varsigma}_{(hi)} - \hat{\varsigma})^۲ + \sum_{l=۱}^{L} \frac{\tilde{n}^B_l - ۱}{\tilde{n}^B_l} \sum_{j=۱}^{\tilde{n}^B_l} (\hat{\varsigma}_{(lj)} - \hat{\varsigma})^۲$$

(۱۶)

براوردگر جک‌نایف اصلاح شده‌ی واریانس، $v_{MJ}(\hat{\varsigma})$ نیز به همان صورت $v_J(\hat{\varsigma})$ است با این تفاوت که به‌جای $\hat{\beta}^A_{(hi)}$ یا $\hat{\beta}^B_{(lj)}$ از $\hat{\beta}$ استفاده کرده و قابل کاربرد با $\tilde{n}^A_h = ۲$ یا $\tilde{n}^B_l = ۲$ می‌باشد.

برای به دست آوردن براوردگر خطی واریانس ς با استفاده از خطی‌سازی سری تیلور آن با فرض این که f تابعی مشتق‌پذیر با مشتقات جزئی مرتبه‌ی دوم پیوسته و کراندار در یک همسایگی از $f(\bar{A}, \bar{B}, \beta)$ است و $\partial f(\bar{A}, \bar{B}, \beta)/\partial\beta$ و همچنین تحت شرط ج داریم:

$$Var(\hat{\varsigma}) = E(\hat{\varsigma} - \varsigma)^۲ = g^T_A(\bar{A}, \bar{B}) \Sigma_A g_A(\bar{A}, \bar{B}) + g^T_B(\bar{A}, \bar{B}) \Sigma_B g_B(\bar{A}, \bar{B}) + o(\tilde{n}^{-۱})$$

و براوردگر خطی‌سازی واریانس $\hat{\varsigma}$ عبارت است از:

$$v_L(\hat{\varsigma}) = \hat{g}^T_A\left(\hat{\bar{A}}, \hat{\bar{B}}\right) S_A \hat{g}_A\left(\hat{\bar{A}}, \hat{\bar{B}}\right) + \hat{g}^T_B\left(\hat{\bar{A}}, \hat{\bar{B}}\right) S_B \hat{g}_B\left(\hat{\bar{A}}, \hat{\bar{B}}\right) \quad (۱۷)$$

که $\hat{g}_A$ و $\hat{g}_B$ از $\hat{\beta}$ به‌جای β استفاده می‌کنند.

قضیه‌ی ۳. هر سه براوردگر واریانس $v_J(\hat{\varsigma})$، $v_{MJ}(\hat{\varsigma})$ و $v_L(\hat{\varsigma})$ به‌طور مجانبی در حالتی که اختلاف بین هر جفت از آن‌ها از مرتبه‌ی $o_p(\tilde{n}^{-۱})$ باشد هم‌ارزند. (برای اثبات

[۱۴] دیده شود).

براورگرهای جک‌نایف واریانس مذکور برای حالت نمونه‌گیری طبقه‌بندی بدون جایگذاری از واحدهای نمونه‌گیری اولیه با فرض ناچیز بودن کسرهای نمونه‌گیری در دو چارچوب، مطرح شده و قابل کاربرد با طرح‌های نمونه‌گیری طبقه‌بندی چندمرحله‌ای بدون جایگذاری نیز می‌باشند. در حالتی که نمونه‌گیری با جایگذاری از واحدهای نمونه‌گیری اولیه به‌عمل آید نیز همین فرمول‌های واریانس صدق می‌کنند. یک حالت خاص ممکن است زمانی رخ دهد که در یکی از چارچوب‌ها مثلاً چارچوب فهرستی B ، نمونه‌گیری تصادفی ساده‌ی طبقه‌بندی انجام شده و کسرهای نمونه‌گیری ناچیز باشد. در این حالت b_{lj} برداری از مقادیر مرتبط با j امین واحد در طبقه‌ی l ام چارچوب B و $\tilde{n}_l^B = n_l^B$ عبارت از تعداد واحدهای نمونه‌گیری‌شده از N_l^B واحد در طبقه‌ی l ام چارچوب B می‌باشد. در صورتی که کسر نمونه‌گیری n_l^B / N_l^B ناچیز نباشد، به‌طور مشابه با روش جک‌نایف در حالت تک چارچوبی ([۱۹])، $(n_l^B - ۱)/n_l^B$ با $((n_l^B - ۱)/n_l^B)(۱ - n_l^B / N_l^B)$ جایگزین می‌شود.

۲-۳- براورد جک‌نایف واریانس براوردگرهای دوچارچوبی

تمام براوردگرهای دو چارچوبی می‌توانند به‌صورت زیر بیان شوند:

$$\hat{Y} = \sum_{t \in S_A} \tilde{w}_t^A y_t + \sum_{t \in S_B} \tilde{w}_t^B y_t \qquad (۱۸)$$

که از وزن‌های اصلاح‌شده‌ی $\tilde{w}_t^A$ و $\tilde{w}_t^B$ استفاده می‌کند. برای مثال در $\hat{Y}_{PML}(\theta)$ از وزن‌های زیر استفاده می‌شود:

$$\tilde{w}_t^A = \begin{cases} w_t^A \dfrac{N_A - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a^A} & t \in a \\ \dfrac{w_t^A \theta \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_{ab}(\theta)} & t \in ab \end{cases}$$

$$\tilde{w}_t^B = \begin{cases} w_t^B \dfrac{N_B - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b^B} & t \in b \\ \dfrac{w_t^B (1-\theta)\hat{N}_{ab}^{PML}(\theta)}{\hat{N}_{ab}(\theta)} & t \in ab \end{cases}$$

برای محاسبه‌ی براورد جک‌نایف $\hat{Y}_{(hi)}^A$ به‌سادگی وزن‌های w_t^A با $w_{t(hi)}^A$ جایگزین می‌شود. اگر واحد t در خوشه‌ی kام طبقه‌ی gام در چارچوب A باشد آن‌گاه

$$w_{t(hi)}^A = \begin{cases} 0 & hi = gk \quad \text{اگر} \\ \dfrac{\tilde{n}_h^A}{(\tilde{n}_h^A - 1)} w_t^A & h = g \quad i \neq k \quad \text{اگر} \\ w_t^A & h \neq g \quad \text{اگر} \end{cases}$$

به‌طور مشابه $\hat{Y}_{(lj)}^B$ با استفاده از وزن‌های $w_{t(lj)}^B$ محاسبه شده و سپس براورد جک‌نایف واریانس $\hat{Y}$ از رابطه‌ی زیر به دست می‌آید.

$$v_J(\hat{Y}) = \sum_{h=1}^{H} \frac{\tilde{n}_h^A - 1}{\tilde{n}_h^A} \sum_{i=1}^{\tilde{n}_h^A} (\hat{Y}_{(hi)}^A - \hat{Y})^2 + \sum_{l=1}^{L} \frac{\tilde{n}_l^B - 1}{\tilde{n}_l^B} \sum_{j=1}^{\tilde{n}_l^B} (\hat{Y}_{(lj)}^B - \hat{Y})^2$$

اگر وزن‌های اصلاح شده‌ی $\tilde{w}_t^A$ و $\tilde{w}_t^B$ به عناصر S_A و S_B بستگی داشته باشند، برای محاسبه‌ی براورد جک‌نایف $\hat{Y}_{(hi)}^A$ احتیاج به محاسبه‌ی $S_{(hi)}^A$ است. ماتریس $S_{(hi)}^A$ به ازای $\tilde{n}_h^A \geq 3$ می‌تواند با استفاده از یک جک‌نایف جداگانه درون هر تکرار جک‌نایف، این بار با مجموعه داده‌ای که مشاهدات iامین psu طبقه‌ی hام چارچوب A حذف شده است، به دست آید.

۴- مطالعه‌ی شبیه‌سازی

در این بخش پس از معرفی نحوه‌ی اجرای شبیه‌سازی، به بررسی ویژگی‌های تجربی و مقایسه‌ی براوردگرهای دوچارچوبی و همچنین مقایسه‌ی براوردگرهای واریانس (خطی‌سازی، جک‌نایف و جک‌نایف اصلاح‌شده) با استفاده از شبیه‌سازی می‌پردازیم.

یک نمونه‌ی خوشه‌ای دومرحله‌ای با $\tilde{n}^A$ خوشه و m عنصر از هر خوشه به‌عنوان

نمونه‌ی حاصل از چارچوب A $(n_A = \tilde{n}_A.m)$ و یک نمونه‌ی تصادفی ساده با n_B مشاهده به‌عنوان نمونه‌ی حاصل از چارچوب B تولید می‌شود. با فرض نامتناهی بودن جامعه‌ی تحت مطالعه، N_a/N و N_b/N را با γ_a و γ_b جایگزین می‌کنیم. نمونه‌ی حاصل از چارچوب A شامل مقادیر

$$\left\{(y_{ij}, m_{ai}),\quad i = ۱,\cdots,\tilde{n}_A \qquad j = ۱,\cdots,m\right\}$$

می‌باشد که m_{ai} عبارت از تعداد عناصر نمونه‌گیری‌شده از i امین خوشه‌ی نمونه‌ای، متعلق به حوزه‌ی a و y_{ij} نیز مقدار مرتبط با j امین عنصر نمونه‌ای در i امین خوشه‌ی نمونه می‌باشد. نمونه‌ی حاصل از چارچوب B شامل مقادیر نمونه‌ای $\left\{(y_j, n_b), j = ۱,\cdots,n_B\right\}$ است که در آن n_b، تعداد عناصر نمونه‌ای متعلق به حوزه‌ی b و y_j، مقدار مرتبط با j امین عنصر نمونه‌ای است (جزئیات شبیه‌سازی در [۱] آمده است).

برای شبیه‌سازی، نمونه‌گیری را ۱۰۰۰۰ بار تکرار کرده و از هر نمونه، براوردگرهای $\hat{Y}_H(\hat{\theta}_H)$، $\hat{Y}_{FB}(\hat{\beta}_{FB})$ و $\hat{Y}_{PML}(\hat{\theta}_P)$ و براوردگرهای تک چارچوبی $\hat{Y}_{SF}$ و $\hat{Y}_{SFrake}$ و $\hat{Y}_{SFreg}(\hat{\beta}_{s۲})$ با استفاده از مقادیر بهینه‌ی $\hat{\theta}_H$، $\hat{\beta}_{FB}$، $\hat{\theta}_P$ و $\hat{\beta}_{s۲}$ و به ازای مقادیر مختلف γ_a، γ_b، $\tilde{n}_A$ و n_B محاسبه می‌کنیم و سپس میانگین توان دوم خطای تجربی هر براوردگر (متوسط توان دوم انحراف براوردگر از مقدار واقعی آن) را به‌صورت زیر محاسبه می‌کنیم:

$$EMSE = \frac{۱}{R}\sum_{r=۱}^{R}\left(\hat{Y}_r - \bar{\hat{Y}}\right)^۲$$

که $\hat{Y}_r$ عبارت است از براورد $\hat{Y}$ برای r امین تکرار شبیه‌سازی، $\bar{\hat{Y}} = \frac{۱}{R}\sum_{r=۱}^{R}\hat{y}_r$ و R، تعداد تکرارهای شبیه‌سازی است.

برای هر ۶ براوردگر Y، سه براوردگر خطی‌سازی واریانس (L)، جک‌نایف (J) و جک‌نایف اصلاح‌شده (MJ) را محاسبه می‌کنیم. در محاسبه‌ی واریانس جک‌نایف به‌دلیل این‌که از مقادیر براوردشده‌ی پارامترها استفاده می‌شود و این مقادیر، خود به ماتریس‌های S_A و S_B وابسته‌اند، $S^A_{(hi)}$ و $S^B_{(lj)}$ با استفاده از یک جک‌نایف جداگانه درون هر تکرار

جک‌نایف محاسبه می‌شوند. درصد اریبی نسبی (RB) هر سه براوردگر واریانس را با استفاده از رابطه‌ی زیر محاسبه می‌کنیم:

$$RB = \frac{۱۰۰(EV - EMSE)}{EMSE}$$

که EV عبارت است از میانگین ۱۰۰۰۰ براورد واریانس در هر روش. همچنین با استفاده از بازه‌ی اطمینان $(۱-\alpha)$ درصد توزیع نرمال:

$$\text{براورد} \pm z_{(۱-\frac{\alpha}{۲})} SE\ (\text{براورد})$$

و نیز صدک توزیع t با $(\tilde{n}_A - ۱)$ درجه‌ی آزادی، احتمال پوشش تجربی برای روش‌های خطی‌سازی، جک‌نایف اصلاح‌شده و جک‌نایف محاسبه شده است. در این شبیه‌سازی، تمام محاسبات با استفاده از نرم افزار Splus۲۰۰۰ صورت گرفته است.

نتایج شبیه‌سازی در جدول‌های ۱ و ۲ (در جدول ۱، $\frac{n_A}{n_B} = ۳$ و در جدول ۲ ، $\frac{n_A}{n_B} = ۶$) نشان می‌دهد که براوردگر ماکسیمم درستنمایی‌نما، همواره کاهشی کوچک در $EMSE$ نسبت به براوردگرهای هارتلی و فولر- بورمیستر در اثر براورد $\hat{\theta}_H$ و $\hat{\beta}_{FB}$ نشان می‌دهد. در مقایسه با براوردگرهای دو چارچوبی، براوردگر تک چارچوبی، افزایش قابل ملاحظه‌ای را در $EMSE$ وقتی $\gamma_a \le \gamma_b$ یا n_A خیلی بزرگ‌تر از n_B باشد نشان می‌دهد. اریبی نسبی هر سه براوردگر واریانس با افزایش اندازه‌ی نمونه میل به کاهش دارد. با این حال برای نمونه‌های کوچک‌تر، روش‌های خطی‌سازی و جک‌نایف اصلاح‌شده، $EMSE$ را کم براورد می‌کنند. (نتایج به ازای مقادیر دیگر پارامترهای γ_a و γ_b و دیگر اندازه‌های نمونه از دو چارچوب به غیر از آن‌چه در جدول‌های ۱ و ۲ است در [۱] آمده است).

نتایج حاصل از بازه‌ی اطمینان نشان می‌دهد که روش جک‌نایف کامل به وضوح احتمال پوشش بالاتری نسبت به دو روش دیگر دارد در حالی که دو روش دیگر به وضوح احتمال پوشش کم‌تری دارند که البته با افزایش اندازه‌ی نمونه، احتمال پوشش افزایش می‌یابد.

۵- بحث و نتیجه‌گیری

نتایج حاصل از بررسی خواص براوردگرهای دوچارچوبی نشان می‌دهد که براوردگر ماکسیمم درستنمایی‌نما به‌دلیل این که از وزن‌های یکسانی برای همه‌ی متغیرها استفاده می‌کند نسبت به براوردگر هارتلی و فولر- بورمیستر در آمارگیری‌هایی با متغیرهای متعدد ارجحیت دارد. در صورتی که $\frac{N_a}{N} \cong \frac{N_b}{N}$ $(\gamma_a \cong \gamma_b)$ و دو طرح نمونه‌گیری مشابه باشند، براوردگر تک‌چارچوبی با تصحیح نسبتی، نسبت به دیگر براوردگرها از کارایی بالاتری برخوردار است. براوردگر جک‌نایف دارای اریبی کوچک‌تری نسبت به براوردگر واریانس خطی‌سازی است. اگر چه احتمال‌های پوشش تجربی برای هر سه روش تقریباً مشابه هستند، اما وقتی که صدک توزیع t در محاسبه‌ی بازه‌ی اطمینان استفاده می‌شود، روش جک‌نایف احتمال پوشش خیلی بالاتری را نسبت به دو براوردگر دیگر نشان می‌دهد. جک‌نایف معمولاً با توابع ناخطی نظیر نسبت مجموع دو جامعه به کار می‌رود. در این حالت مشتق‌های جزئی که در محاسبه‌ی براورد خطی‌سازی چنین مقادیر ناخطی استفاده می‌شود، در آمارگیری‌های دوچارچوبی پیچیده‌تر از براوردگر تک‌چارچوبی است. در صورتی که با استفاده از جک‌نایف می‌توان از این محاسبات پیچیده اجتناب کرد.

روش‌های بازنمونه‌گیری دیگر براورد واریانس نظیر تکرار مکرر متعادل (BRR)[۸] و خودگردان[۹] نیز می‌توانند به موازات جک‌نایف در این‌گونه طرح‌ها مورد استفاده قرار گیرند. مزیت این روش‌ها این است که بر خلاف جک‌نایف با توابع ناهموار نظیر میانه نیز قابل استفاده هستند.

جدول ۱- میانگین توام دوم خطای براوردگرهای چارچوب دوکان، درصد اریبی نسبی و احتمال پوشش سه براوردگر واریانس برای $\frac{n_A}{n_B} = ۳$

براوردگر	$EMSE$	درصد اریبی نسبی			احتمال پوشش					
					نرمال					
		L	MJ	J	L	MJ	J	L	MJ	J
		$\gamma_b = ۰/۲$			$\tilde{n}_A = ۱۰, m = ۳۰ (n = \tilde{n}_A \times m), n_B = ۱۰۰, \gamma_a = ۰/۱,$					
H	۵/۰۵	-۱۲/۱	-۱۰/۱	۱۱/۳	۰/۹۲۰	۰/۹۲۶	۰/۹۲۵	۰/۹۲۹	۰/۹۳۷	۰/۹۵۱
FB	۵/۱۵	-۱۶/۷۴	-۱۵/۷۴	۱۴/۸۷	۰/۸۸۵	۰/۹	۰/۹۰۵	۰/۹۲۴	۰/۹۴۶	۰/۹۴۹
PML	۴/۵۷	-۱۳/۰۷	-۱۲/۰۷	۷/۴	۰/۹۱۱	۰/۹۱۸	۰/۹۴۲	۰/۹۴۲	۰/۹۴۵	۰/۹۵۰
SF	۱۱/۴۲	۳/۵۹	۳/۵۹	۳/۵۹	۰/۹۵۴	۰/۹۵۲	۰/۹۵۲	۰/۹۲۶	۰/۹۳۷	۰/۹۴۹
SF_{reg}	۴/۸۸	-۱۳/۱۳	۱۱/۱۳	۱۱/۲۷	۰/۹۳۹	۰/۹۴۱	۰/۹۵	۰/۹۵۸	۰/۹۶۲	۰/۹۷
SF_{rake}	۴/۵۳	-۳/۴	-۳/۱	۴/۳۲	۰/۹۳۰	۰/۹۴۹	۰/۹۵۰	۰/۹۶۷	۰/۹۷۰	۰/۹۵۷

جدول ۲- میانگین توام دوم خطای براوردگرهای چارچوب دوگان، درصد اریبی نسبی و احتمال پوشش سه براوردگر واریانس برای $\frac{n_A}{n_B} = ۶$

براوردگر	$EMSE$	درصد اریبی نسبی			احتمال پوشش: نرمال			احتمال پوشش		
		L	MJ	J	L	MJ	J	L	MJ	J
		$\tilde{n}_A = ۲۰, m = ۳۰ (n = \tilde{n}_A \times m), n_B = ۱۰۰, \gamma_a = ۰/۱,$ $\gamma_b = ۰/۲$								
H	۲/۳۷	-۱۱/۳۶	-۱۱/۰۶	-۰/۶۳	۰/۹۲۲	۰/۹۲۲	۰/۹۵	۰/۹۳۷	۰/۹۳۷	۰/۹۵۲
FB	۲/۴۳	-۱۷/۶۶	-۱۶/۶۶	۱/۸	۰/۸۸۵	۰/۹۰۸	۰/۹۱۸	۰/۹۰۲	۰/۹۲۰	۰/۹۴۸
PML	۲/۳	-۸/۲	-۷/۸۱	-۱/۱۰	۰/۹۳۰	۰/۹۳۹	۰/۹۴۳	۰/۹۴۶	۰/۹۴۹	۰/۹۴۷
SF	۵/۴۵	-۷/۹۹	-۷/۹۹	-۷/۹۹	۰/۹۳۴	۰/۹۳۶	۰/۹۵۸	۰/۹۳۸	۰/۹۳۷	۰/۹۷۲
SF_{reg}	۲/۳۸	-۱۲/۶۹	-۱۲/۰۵	۰/۰۸	۰/۹۲۱	۰/۹۲۱	۰/۹۴۷	۰/۹۱۴	۰/۹۳۲	۰/۹۵۲
SF_{rake}	۲/۳۲	-۱/۷	-۰/۷	-۳/۰۷	۰/۹۵	۰/۹۰۵	۰/۹۴۵	۰/۹۴۰	۰/۹۴۸	۰/۹۵۷

توضیحات

1. Delta Method
2. Multiple frame surveys
3. Jackknife
4. Psedo-maximum likelihood estimator
5. Single frame estimator

۶. برای دو دنباله از اعداد $\{a_n\}$ و $\{b_n\}$، $a_n = \mathrm{O}(b_n)$ اگر به ازای هر n و یک c ثابت $|a_n| \le c|b_n|$؛ و $a_n = \mathrm{o}(b_n)$ اگر وقتی $n \to \infty$، $\frac{a_n}{b_n} \to 0$.

۷. برای دو دنباله از متغیرهای تصادفی $\{X_n\}$ و $\{Y_n\}$، $X_n = \mathrm{O}_p(Y_n)$ اگر برای هر $\varepsilon > 0$، یک ثابت $C_\varepsilon > 0$ وجود داشته باشد به‌طوری که $p\{|X_n| \ge C_\varepsilon |Y_n|\}$ برای هر n؛ $X_n = \mathrm{o}_p(Y_n)$ اگر وقتی $n \to \infty$، $\frac{X_n}{Y_n} \xrightarrow{p} 0$.

8. Balanced repeated replication (BRR)
9. Bootstrap

مرجع‌ها

[۱] نورینی، مرجان (۱۳۸۲). طرح‌های نمونه‌گیری چندچارچوبی. پایان‌نامه‌ی کارشناسی ارشد. دانشگاه صنعتی اصفهان، اصفهان.

[2] Bankier, M.D. (1986). Estimators based on several stratified samples with application to multiple frame surveys, *Journal of the American Statistical Association,* **81,** 1074-1079.

[3] Ford, B.L. and Bosecker, R.R. (1979). Multiple frame estimation with stratified overlap domain, in *Proceedings of the Social Statistics Section, American Statistical Association*, 219-224.

[4] Fuller, W.A. and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames, *in Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.

[5] Hartley, H.O. (1962). Multiple frame surveys, *in Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

[6] Hartely, H.O. (1974). Multiple frame methodology and selected

application, *Sankhya, Ser. C*, **36**, 203-206.

[7] Kalton, G. and Anderson, D.W. (1986). Sampling rare populations, *Journal of the Royal statistical Society, Ser. A*, **149**, 65-82.

[8] Lohr, S. and Rao, J.N.K. (1997). Jackknife variance estimation in multiple frame surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 552-557

[9] Lund, R.E. (1968). Estimators in multiple frame surveys, in *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.

[10] Quenoulle, M.H. (1949). Approximation tests of correlation in time Series, *Journal of Royal Stat. Soc.*, **B 11**, 68-84

[11] Quenoulle, M.H. (1956). Notes on bias in estimation, *Bimetrika*, **43**, 353-360,

[12] Rao, J.N.K. and Wu, C.F J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics, *Journal of the American Statistical Association*, **80**, 620-630.

[13] Rao, J.N.K. and Lohr, S.L. (1997). *Jackknife Variance Estimation in Dual Frame Surveys*, Technical Report, Department of Mathematics and Statistics, Carleton University.

[14] Rao, J.N.K. and Lohr, S.L. (2000). Inference from dual frame survey, *Journal of the American Statistical Association*, **95**, 271-280.

[15] Sarndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

[16] Skinner, C.J. (1991). On the efficiency of raking ratio estimators for multiple frame surveys, *Journal of the American Statistical Association*, **86**, 779-784.

[17] Skinner, C.J. and Rao, j.N.K. (1996). Estimation in dual frame survey with complex design, *Journal of the American Statistical Association*, **91**, 349-356.

[18] Tukey, J.W. (1958). Bias and Confidence in Note Quite Large Samples

(Abstract), *Annals of Mathematical Statistics*, **29**, 614.

[19] Vogel, F.A. (1975). Survey with overlapping frame-problems in application, in Proceedings of the Social Statistics Section, *American Statistical Association*, 694-699.

[20] Wolter, K.M. (1985). *Introduction to Variance Estimation*, Springer-Verlag, New York.

مرجان نورینی
فوق لیسانس آمار
تهران، خیابان فاطمی، نبش خیابان رهی معیری، مرکز آمار ایران.
رایانشانی: Mar_noorini@yahoo.com