

رویکرد آنتروپی به تنظیم جدول عمر، مطالعه موردی: جدول عمر ایران

رضوان رضائی، غلامحسین یاری*
دانشگاه علم و صنعت ایران، دانشکده ریاضی
دریافت ۹۸/۰۲/۲۵ پذیرش ۹۸/۰۸/۱۳

چکیده

تحلیل بقا و به‌ویژه برآورد توزیع بقا از موضوعات مهم در علوم آماری است. روش‌های پارامتری و ناپارامتری مختلفی برای برآورد توزیع بقا مطرح شده‌اند. در این ارتباط توزیع‌های بقای تئوریک مشخص شده‌اند و پارامترهایشان به کمک روش‌هایی مانند برآوردگر حداکثر درست‌نمایی و برآوردگر بیزی به دست می‌آیند. از جمله روش‌های ناپارامتری نیز می‌توان به روش کاپلان‌مایر، رگرسیون کاکس و جدول عمر اشاره کرد. علاوه بر این، یکی دیگر از مباحثی که در تحلیل بقا اهمیت زیادی دارد است طبقه‌بندی داده‌ها است که همواری و نیکویی برازش دو نیاز اساسی برای آن محسوب می‌شوند. از سوی دیگر در نظریه احتمال بر اساس عبارت تعریف‌شده پایای آنتروپی، دو مدل بهینه‌سازی یکی بر اساس اصل بیشینه آنتروپی (ME) و دیگری بر پایه اصل کمینه معیار کولیک-لیبلر (MKL) به منظور برآورد توزیع احتمال ارائه شده است. در این مقاله، رویکرد دو مدل بهینه‌سازی فوق را به برآورد توزیع بقا و توزیع احتمال به‌ویژه برای داده‌های طبقه‌بندی شده بررسی می‌کنیم. در این پژوهش‌ها علاوه بر بررسی مدل‌های پارامتریک، روش ناپارامتری جدیدی که یک تابع هدف ترکیب شده از دو اصل ME و MKL و یک ضریب برای اطمینان از درجه نیکویی برازش و هموارسازی برآوردها که نشان‌دهنده اولویت این دو شاخص در طبقه‌بندی داده‌ها است را به کار می‌بریم. ما از این روش برای برآورد توزیع احتمال مرگ و میر سن مشخص (ستون Q_x) در جدول عمر، استفاده می‌کنیم. در پایان به کمک این روش جدول عمر زنان و مردان ایران در سال ۱۳۹۰ (ش.ه) را تنظیم می‌کنیم.

واژه‌های کلیدی: تحلیل بقا، نظریه اطلاع، اصل بیشینه آنتروپی، اصل کمینه کولیک-لیبلر، جدول عمر.

مفاهیم و مقدمات

تحلیل بقا در آمار زیستی و عمل‌گر اهمیت زیادی دارد. در این مقاله قصد داریم توزیع طول عمر که از موضوعات مهم تحلیل بقا است را به کمک نظریه اطلاع برآورد کنیم و سپس جدول عمر زنان و مردان ایران در سال ۱۳۹۰ (ش.ه) را به این وسیله تنظیم کنیم. برای این منظور ابتدا برخی از مفاهیم مقدماتی مربوط به نظریه اطلاع و جدول عمر را ارائه می‌کنیم.

تعریف ۱. فرض کنید X یک متغیر تصادفی پیوسته با تابع چگالی $f_X(x)$ و با تکیه‌گاه S_X باشد آن‌گاه آنتروپی شانون X به صورت $h(X) = -\int_{S_X} f(x) \log f(x) dx$ تعریف می‌شود [۱۱].

معیار فوق که میزان عدم قطعیت متغیر تصادفی X را نشان می‌دهد اولین بار به وسیله شانون^۱ معرفی شد. پس از آن برخی تعمیم‌های این تعریف و همین‌طور شاخص‌های مرتبط دیگر از جمله شاخص آنتروپی نسبی ارائه شد. این شاخص به وسیله کولیک و لیبلر^۲ با نام آنتروپی نسبی^۳ معروف و بدین صورت تعریف می‌شود:

*نویسنده مسئول yari@iust.ac.ir

1. Shannon

2. Kullback & Leibler

۳. این معیار به عنوان یک شاخص برای بررسی نیکویی برازش یک توزیع احتمال $f(x)$ در مقابل توزیع احتمال مرجع $g(x)$ استفاده می‌شود. به بیان دقیق‌تر این نام‌گذاری به دلیل کارایی این معیار در تعیین میزان افتراق (تفاوت) بین دو تابع چگالی احتمال است.

تعریف ۲. برای هر دو متغیر تصادفی پیوسته X و Y با توابع چگالی به ترتیب f و g ، آنتروپی نسبی که معیار کولبک-لیبلر نیز نامیده می‌شود از رابطه (۱) به دست می‌آید [۹].

$$KL(f(x)||g(x)) = \int_R f(x) \ln(f(x)/g(x)) dx, \quad (۱)$$

تعمیم‌هایی از شاخص KL نیز معرفی شده است، از جمله می‌توان به تعریفی که به وسیله لئو^۱ ارائه شده است اشاره کرد که در ادامه آن را مطرح می‌کنیم.

تعریف ۳. آنتروپی نسبی بر اساس تابع بقا (KLS) X_1, X_2, \dots یک دنباله مثبت، مستقل و هم توزیع از متغیرهای تصادفی با تابع بقای غیرصعودی $\bar{F}(x; \theta) = P_\theta(X > x)$ که θ بردار پارامترهای نامعلوم و $\bar{G}(x) = \sum_{i=0}^{n-1} (1-i/n) I_{[X_{(i)}, X_{(i+1)})}(x)$ تابع بقای تجربی بر اساس یک نمونه تصادفی n تایی از $\bar{F}(x; \theta)$ که I تابع مشخصه و $0 = X_{(0)} \leq X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ نمونه مرتب شده است، را در نظر بگیرید. در این صورت معیار آنتروپی نسبی یا کولبک-لیبلر بر اساس توابع بقای $\bar{F}(x; \theta)$ و $\bar{G}(x)$ به صورت رابطه (۲) محاسبه می‌شود [۱۰].

$$KLS(\bar{G}_n(x)||\bar{F}(x)) = \int_0^\infty [\bar{G}_n(x) \ln(\bar{G}_n(x)/\bar{F}(x))] - [\bar{G}_n(x) - \bar{F}(x)] dx, \quad (۲)$$

در رابطه (۲) $\bar{F}(x; \theta)$ توزیع مرجع است و KLS فاصله $\bar{G}(x)$ تا $\bar{F}(x; \theta)$ را می‌سنجد. رابطه (۲) را می‌توان بدین صورت بیان کرد [۱۴]:

$$KLS(\bar{G}_n(x)||\bar{F}(x)) = \int_0^\infty [\bar{G}_n(x) \ln(\bar{G}_n(x)/\bar{F}(x))] - [\bar{X} - E(X)] dx, \quad (۳)$$

(تعاریف بالا برای متغیر تصادفی گسسته X با تابع جرم احتمال $P(X=x)$ نیز با جابه‌جایی \int, Σ برقرار است. هم‌چنین برای مطالعه جزئیات بیش‌تر به [۱۴] مراجعه کنید.)

از سوی دیگر در جمعیت‌شناسی پیش‌بینی احتمال مرگ افراد جامعه برای بررسی ساختار بقای جامعه اهمیت خاصی دارد. این پیش‌بینی تنها بر اساس محاسبه احتمال مرگ افراد پیشین در شرایط نسبتاً همگن و با استفاده از اطلاعات جمعیتی جمع‌آوری شده، امکان‌پذیر است. علاوه بر این در اکچوئری تعهدات یک شرکت بیمه عمر به طول عمر بیمه‌شدگان وابسته است. جدول عمر^۲ یکی از مهم‌ترین کشفیات جمعیت‌شناسی و از نخستین ابزارهای آماری است که برای بررسی حوزه‌های اصلی جمعیت چون مرگ‌ومیر، امید زندگی^۳، مهاجرت^۴، باروری^۵ و هم‌چنین ساختار رشد جمعیت به کار گرفته شده است. ادموندهالی و جان‌گرانت^۶ اولین دانشمندانی بودند که (به صورت جداگانه) جدول عمرهایی را برای لهستان و انگلیس ابداع کردند [۱]. جدول عمر، جدول مرگ‌ومیر^۷ نیز نامیده می‌شود و دارای انواع مختلفی است: جدول عمر کامل، جدول عمر خلاصه، جدول عمر مقطعی و جدول عمر نسلی [۱]. در ایران تاکنون جدول مرگ‌ومیری که به صورت رسمی منتشر شود و مبنای محاسبات جمعیتی و بیمه‌ای باشد، انتشار نیافته است. مشکل اساسی، در دست نبودن آمار کل و دقیق فوت‌شدگان برحسب سن در زمان فوت است. با این حال در سال‌های اخیر کوشش‌هایی از طرف بعضی سازمان‌ها و پژوهش‌گران انجام شده است. شاید از اولین کارهایی که می‌توان در زمینه ساخت جدول عمر در ایران نام برد، تلاشی است که نه‌آپتیان و خزانه در سال ۱۳۵۲ در طرحی با عنوان «بررسی و

1. Lio

2. Life table

3. Life Expectancy

4. Migration

5. Fertility

6. J. Grant and Edmund Halley

7. Mortality table

برآورد میزان‌های اختصاصی سنی مرگ و میر و باروری در ایران» داشتند. آنها جدول عمر کشور را در بازه‌های ۵ ساله و به تفکیک شهر و روستا و جنس تنظیم کردند. بعد از آن افرادی از جمله ک.ال کهلی، شمس، دکتر حسین ملک افضلی و محمودی، دکتر حسین ملک افضلی و همکاران، نوراللهی، زنجانی و کوششی، میرزایی، کوششی و ناصری، نوراللهی و نجاتیان جداول عمر با روش‌های متفاوت و با اهداف بعضاً متفاوتی، طراحی و تنظیم کرده‌اند [۱]. برخی از ستون‌های مهم این جدول بدین‌قرار هستند:

x و $x+d$ سن شروع، $x+d$ سن پایان دوره و d فاصله آن دو سن به سال تمام است برای مثال در جدول عمر کامل $d=1$ است)، l_x (تعداد بازماندگان تا سن درست x که در این جدول‌ها با 10^5 نفر شروع شده است)، d_x (تعداد مرگ‌ومیرها از سن x تا $x+d$)، e_x (امید زندگی در سن x است که $e_x = \int_0^\infty l_a da$)، q_x (احتمال مرگ‌ومیر در هر بازه x تا $x+d$ است که در تشکیل یک جدول نقش کلیدی ایفا می‌کند) و M_x (میزان مرگ‌ومیر در وسط گروه سن x و $x+d$ را نشان می‌دهد و معمولاً با m_x مشخص می‌شود. با فرض ثبات نرخ لحظه‌ای مرگ و میر در بازه سنی x تا $x+d$ ، رابطه $q_x = 1 - \exp(-m_x)$ برقرار است [۱۲]. برای جزئیات بیش‌تر در مورد سایر ستون‌های جدول عمر و روابط بین آن‌ها به [۱] مراجعه کنید.

معیار KLS و برآورد پارامترهای توزیع گمپرتز

در بررسی‌های بقای سیستم‌ها به منظور مدل‌سازی داده‌های مرگ و میر و طول عمر آن‌ها، برخی توزیع‌های آماری استفاده می‌شود. از جمله این توزیع‌ها، توزیع گمپرتز، گمپرتز-مکهم، وایبل، هلیگمن پولارد نوع ۱، ۲ و ۳ و سیلر و هم‌چنین تعمیم‌های از آنها مانند توزیع‌های BGG^1 ، $OLGG^2$ و $EGWG^3$ هستند.

در اینجا از معیار KLS برای بررسی توزیع‌های پارامتریک مرگ و میر استفاده می‌کنیم. برای این منظور کافی است تابع بقا و امید ریاضی توزیع موردنظر را در رابطه (۳) جای‌گذاری کرده و سپس نسبت به هرکدام از پارامترها با در نظر گرفتن ثابت بودن سایر پارامترها از عبارت KLS به‌دست آمده مشتق گرفته و آنها را برابر صفر قرار داده و سپس دستگاه به‌دست‌آمده را حل کرد [۱۴]. در اینجا برای جلوگیری از دور شدن هدف این بخش از ارائه جزئیات بیش‌تر این روش صرف‌نظر می‌کنیم و خواننده را به بررسی [۱۴] دعوت می‌کنیم و فقط نتایج مربوط به استفاده از این روش را برای توزیع گمپرتز با تابع بقا $\bar{F}(X) = \exp(-(\lambda/c)\exp(cx-1))$ و امید ریاضی $E(X) = (1/c)\exp(\lambda/c)[(\lambda/c) - \ln(\lambda/c) - \gamma]$ که c و λ پارامترهای این توزیع و γ ثابت اوایلر است، ارائه می‌کنیم. بر اساس آن‌چه بالا گفته شد با جای‌گذاری $\bar{F}(X)$ و $E(X)$ توزیع گمپرتز در رابطه (۳) داریم:

$$KLS(\bar{G}_n(x) \parallel \bar{F}(x)) = \sum_{k=1}^n \left(1 - \frac{k-1}{n}\right) \ln\left(1 - \frac{k-1}{n}\right) \Delta x_k + -(\bar{x} - (1/c)\exp(\lambda/c)[(\lambda/c) - \ln(\lambda/c) - \gamma]) \\ + \sum_{k=1}^n \left(1 - \frac{k-1}{n}\right) \frac{\lambda}{c} \left(\frac{1}{c} e^{cx_{(k)}} - x_{(k)} - \frac{1}{c} e^{cx_{(k-1)}} + x_{(k-1)}\right).$$

سپس با مشتق‌گیری از مقدار KLS فوق، به‌ترتیب نسبت به پارامتر λ (با فرض ثابت بودن پارامتر c) و c (با فرض ثابت بودن پارامتر λ) و دستگاه معادله زیر به‌دست می‌آید.

1. The beta generalized Gompertz distribution
2. The Odd Log-Logistic Generalized Gompertz Distribution
3. The Exponentiated Generalized Weibull-Gompertz Distribution

$$\left\{ \begin{aligned} & \frac{1}{c} \sum_{k=1}^n \left(1 - \frac{k-1}{n}\right) \left(\frac{1}{c} (e^{cx_k} - e^{cx_{(k-1)}}) - \Delta x_{(k)}\right) + \frac{e^{\frac{\lambda}{c}}}{c^2} \left(\frac{\lambda}{c} - \ln\left(\frac{\lambda}{c}\right) - \gamma\right) + \frac{e^{\frac{\lambda}{c}}}{c^2} \frac{\lambda - c}{c\lambda} = 0, \\ & \left(1 + \frac{\lambda}{c}\right) \frac{-e^{\frac{\lambda}{c}}}{c^2} \left(\frac{\lambda}{c} - \ln\left(\frac{\lambda}{c}\right) - \gamma\right) + \frac{e^{\frac{\lambda}{c}}}{c} \left(\frac{1}{c} - \frac{\lambda}{c}\right) + \sum_{k=1}^n \left[\left(1 - \frac{k-1}{n}\right) \left(\frac{-\lambda}{c}\right) \left(\frac{1}{c} (e^{cx_k} - e^{cx_{(k-1)}}) - \Delta x_{(k)}\right)\right] + \\ & \sum_{k=1}^n \left[\left(1 - \frac{k-1}{n}\right) \left(\frac{\lambda}{c}\right) \left(\frac{-1}{c^2} (e^{cx_k} - e^{cx_{(k-1)}}) + \frac{1}{c} (x_{(k)} e^{cx_{(k)}} - x_{(k-1)} e^{cx_{(k-1)}})\right)\right] = 0. \end{aligned} \right.$$

چنان‌که ملاحظه می‌شود جواب‌های این دستگاه به‌طور سر راست قابل محاسبه نیست و باید از روش‌های عددی مانند نیوتن، شبه نیوتن، روش‌های تکرار و از نرم‌افزارهای مربوطه مانند MATLAB کمک گرفت. به‌طور مثال در بخش‌های بعد از این روش برای مدل‌سازی مرگ و میر ۲۰۸ موش استفاده می‌کنیم و نتایج را در جدول ۱ ارائه می‌کنیم. در ادامه، رویکرد آنتروپی به داده‌های طبقه‌بندی شده به‌ویژه برای تنظیم یک جدول عمر را مورد مطالعه و بررسی قرار می‌دهیم. قبل از آن به‌طور خلاصه اصول بهینه‌سازی آنتروپی را شرح می‌دهیم.

بررسی اجمالی اصول بهینه‌سازی آنتروپی

آنتروپی نظریه اطلاع، به‌وسیله شانون و اصل بهینه‌سازی آنتروپی تحت عنوان اصل بیشینه آنتروپی به‌وسیله جینز ارائه شده است [۱۱]. کولیک و همکاران نیز با معرفی اصل کمینه معیار کولیک-لیبلر شرایط استفاده از آنتروپی را وسعت دادند و آن را از یک شاخص نظریه اطلاع به یک ابزار آمار استنباطی تبدیل کردند [۹]. به بیان دقیق‌تر، بهینه‌سازی آنتروپی شامل دو اصل بیشینه آنتروپی (ME^1) و اصل کمینه معیار کولیک-لیبلر (MKL^2) است و این دو اصل برای برآورد توزیع احتمال با استفاده از مفهوم آنتروپی به‌کار می‌روند. اولی یک توزیع احتمال را تنها بر پایه اطلاعات معلوم، بدون اضافه کردن هرگونه اطلاعات ذهنی دیگری، برآورد می‌کند و دومی برای برآورد کردن یک توزیع احتمال که نزدیک‌ترین توزیع به توزیع پیشین است، استفاده می‌شود.

جینز اعتقاد داشت که با داشتن میانگین، معمولاً بی‌شمار توزیع سازگار وجود دارد. ME ما را تشویق به انتخاب توزیعی که آنتروپی شانون را به حداکثر رسانده و به‌طور هم‌زمان سازگار با محدودیت میانگین است، می‌کند. برای این منظور فرض کنید Θ یک متغیر تصادفی گسسته در فضای احتمال (Ω, F, P) باشد که در آن $\Omega = \{\theta_1, \theta_2, \dots, \theta_n\}$ و مقدار نامعلوم $P(\Omega = \theta_i) = p_i, i = 1, 2, \dots, n$ است و به کمک برخی اطلاعات معلوم $g_j(\theta) (j = 1, 2, \dots, m)$ مانند میانگین، واریانس، j امین گشتاور و غیره برآورد می‌شود. ME بدین‌صورت توصیف می‌شود:

$$\max \quad H(p) = -\sum_{i=1}^n p_i \ln p_i \quad s.t. \quad \left\{ \begin{aligned} & \sum_{i=1}^n p_i = 1 \\ & \sum_{i=1}^n p_i g_j(\theta) = E_j, j = 1, 2, \dots, m \\ & p_i \geq 0, i = 1, 2, \dots, n \end{aligned} \right.$$

که در آن $H(p)$ آنتروپی توزیع احتمال روی Ω است.

همچنین اگر یک توزیع پیشین برای Θ یعنی $Q(\Theta = \theta_i) = q_i$ وجود داشته باشد، آن گاه MKL می تواند برای به دست آوردن توزیع برآورد شده دیگری که از نظر آماری نزدیک ترین به توزیع پیشین تحت همان قیود ME است، استفاده شود. MKL بدین صورت مدل بندی می شود:

$$\min K(P, Q) = -\sum_{i=1}^n p_i \ln(p_i/q_i) \quad s.t. \begin{cases} \sum_{i=1}^n p_i = 1 \\ \sum_{i=1}^n p_i g_j(\theta) = E_j, j=1, 2, \dots, m \\ p_i \geq 0, i=1, 2, \dots, n \end{cases}$$

که $K(P, Q)$ معیار کولبک لیبلر است.

دو مدل مذکور به کمک روش لاگرانژ [۸] می توانند حل شوند که جواب ها بدین صورت است:

$$p_i^{(\max)} = \exp[-\alpha_0 - \sum_{i=1}^m \alpha_i g_j(\theta)], \quad (4)$$

$$p_i^{(\min)} = q_i \exp[-\beta_0 - \sum_{i=0}^m \beta_i g_j(\theta)] \quad (5)$$

و $\alpha_1, \alpha_2, \dots, \alpha_n$ و $\beta_1, \beta_2, \dots, \beta_n$ ضرایب لاگرانژ مدل های مذکور هستند.

«طبقه بندی داده ها» با استفاده از اصول بهینه سازی آنتروپی

یک گروه زنده که جمعیت اولیه آن را با l_0 نشان می دهیم، در نظر بگیرید. در زمان $x \geq 0$ جامعه ی زنده را با $l_x (x = 0, 1, \dots, T)$ مشخص و فرض می کنیم که $l_T = 0$ باشد. همچنین $d_x (x = 0, 1, \dots, T)$ را تعداد مرگ و میر جامعه از x تا $x+1$ در نظر بگیرید آن گاه

$$d_x = l_x - l_{x+1}. \quad (6)$$

مقدار d_x معمولاً از مشاهدات به دست می آید. در تحلیل بقا q_x برای نشان دادن احتمال مرگ و میر که به معنای احتمال مرگ یک فرد در سال بعد است، استفاده می شود. احتمال مرگ و میر یک پارامتر اساسی برای ساختن یک جدول عمر است و اهمیت بسیار زیادی در علم بیمه عمر دارد. در شرایط واقعی این احتمال از داده های نمونه می تواند به دست آید و به صورت $\hat{q}_x = d_x/l_x$ مشخص می شود. بررسی و چرایی این که $\sum_{x=0}^T \hat{q}_x \neq 1$ آسان است. به هر حال، \hat{q}_x برآوردی از q_x است که در شرایط واقعی نزدیک ترین مقدار ممکن را برای احتمال مرگ و میر ارائه می کند. این فرایند در بیمه عمر «طبقه بندی داده» نامیده می شود. علاوه بر این، برآورد q_x بر اساس اطلاعات نمونه است؛ بنابراین ما باید از اطلاعات موجود در نمونه به طور کامل استفاده کنیم و سعی کنیم بهترین اطلاعات اضافی ممکن را وارد مسئله کنیم. این امر یک پایه و اساس منطقی برای استفاده ME و MKL را می تواند فراهم کند. حال به منظور استفاده از ME و MKL، احتمال مرگ و میر را به صورت (۷) تعریف می کنیم.

$$\tilde{q}_x = d_x/l_0, x=0, 1, \dots, T, \quad (7)$$

(نسبتی از احتمال مرگ جامعه در گروه های سنی مختلف به جامعه اولیه). \tilde{q}_x نیاز توزیع احتمال را می تواند فراهم کند (زیرا در اصول احتمال کلموگروف صدق می کند: $\forall x: 0 \leq \tilde{q}_x \leq 1$ و $\sum_{x=0}^T \tilde{q}_x = 1$) و روابط

مشخص و برآورد آن به روش ME از حل رابطه (۸) حاصل می‌شود.

$$\max H = -\sum_{i=1}^n \hat{q}_x \ln \hat{q}_x \quad s.t. \begin{cases} \sum_{i=1}^n \hat{q}_x = 1 \\ \sum_{i=1}^n x \hat{q}_x = E_1 \\ \sum_{i=1}^n (x - E_1)^j \hat{q}_x = E_j, j = 2, \dots, k \\ \hat{q}_x \geq 0, x = 1, 2, \dots, T \end{cases} \quad (8)$$

که در آن \hat{q}_x برآوردی از \tilde{q}_x ، E_1 میانگین داده‌های نمونه و E_j گشتاور j ام است، از سوی دیگر \tilde{q}_x مانند یک توزیع پیشین برای احتمال مرگ می‌تواند در نظر گرفته شود، در نتیجه یک مدل MKL برای برآورد کردن \tilde{q}_x به صورت (۹) است.

$$\min K = -\sum_{i=1}^n \hat{q}_x \ln(\hat{q}_x / \tilde{q}_x) \quad s.t. \begin{cases} \sum_{i=1}^n \hat{q}_x = 1 \\ \sum_{i=1}^n x \hat{q}_x = E_1 \\ \sum_{i=1}^n (x - E_1)^j \hat{q}_x = E_j, j = 2, \dots, k \\ \hat{q}_x \geq 0, x = 1, 2, \dots, T \end{cases} \quad (9)$$

حل معادلات (۸) و (۹) به کمک روش لاگرانژ با ضرایب α_j, β_j به صورت روابط (۱۰) و (۱۱) است.

$$\hat{q}_x^{(\max)} = \exp[-\sum_{i=0}^m \alpha_j (x - E_1)^j], \quad (10)$$

$$\hat{q}_x^{(\min)} = \tilde{q}_x \exp[-\sum_{i=0}^m \beta_j (x - E_1)^j] \quad (11)$$

داده‌های طبقه‌بندی شده به وسیله ترکیب دو روش ME و MKL (MEMKL)

چنان‌که در بخش‌های قبل گفته شد همواری و نیکویی برازش همیشه مهم‌ترین موضوعات مورد توجه در طبقه‌بندی داده‌ها بوده‌اند هر چند تکنیک‌های بسیار دیگری نیز توسعه یافته‌اند. به عنوان مثال، گسترده‌ترین روش مورد استفاده تا به حال، به خصوص به وسیله بیمه‌دانان آمریکای شمالی برای ساخت جداول عمر، روش ویتاکر-هندرسون^۱ از طبقه‌بندی است [۲]. این روش در کار بولمن^۲ و ویتکار^۳ معرفی شد و به نظریه هندرسون کمک کرد [۴]، [۵]، [۶]، [۱۳]. روش ویتاکر-هندرسون مقادیر طبقه‌بندی شده را با کمینه کردن مقدار

$$M = F + hS = \sum_{x=1}^n \omega_x (v_x - u_x)^2 + h \sum_{x=1}^{n-z} [\Delta^z v_x]^2 \quad (12)$$

1. Whittaker-Henderson

2. Bohlmann

3. Whittaker

به دست می‌آورد که F و S به ترتیب مقادیر وزندار شده نیکویی برازش برای داده‌های اولیه و ضریب همواری هستند. u_x توزیع پیشین مرگ و میر و v_x مرگ و میر برآورد شده است، یعنی نتیجه طبقه‌بندی داده و $\Delta^x v_x$ از z^{th} اختلاف v_x (معمولاً $z = 2, 3$ یا بیش‌تر) است. ω_x ضریب وزنی و h عامل تعدیل مثبت بین نیکویی برازش و همواری است. این روش به‌طور گسترده‌ای استفاده شده و تبدیل به یک منطق اساسی در طبقه‌بندی داده و دیگر روش‌های جاری در این زمینه شده است.

به‌طور کلی، مشخصه‌سازی روش‌های مختلف طبقه‌بندی داده ممکن است از دو جهت باشد، یکی به‌کار بردن نیکویی برازش و همواری و دیگری چگونگی اندازه‌گیری همواری و نیکویی برازش؛ بنابراین، طبقه‌بندی داده را به‌عنوان یک موضوع دو هدفه می‌توان نگاه کرد. از یک طرف، نتایج طبقه‌بندی داده‌ها باید هموار باشد و از سوی دیگر باید به داده‌های اصلی نزدیک باشند. در ادامه یک رویکرد جدید از طبقه‌بندی داده که هر دو روش ME و MKL را در مدل زیر می‌تواند ترکیب کند را مطرح می‌کنیم [۷]:

$$\min \quad G = \mu \sum_{x=1}^T \hat{q}_x \ln \hat{q}_x + (1 - \mu) \sum_{x=1}^T \hat{q}_x \ln(\hat{q}_x / \tilde{q}_x) \quad s.t. \begin{cases} \sum_{x=1}^T \hat{q}_x = 1 \\ \sum_{x=1}^T x \hat{q}_x = E_1 \\ \sum_{x=1}^T (x - E_1)^j \hat{q}_x = E_j, j = 2, \dots, k \\ \hat{q}_x \geq 0, x = 1, 2, \dots, T \end{cases} \quad (13)$$

که در آن $0 \leq \mu \leq 1$ عامل تعدیل داده شده بین همواری و نیکویی برازش، E_1 امید ریاضی و E_j ها گشتاورهای مرکزی مرتبه j ام متغیر تصادفی X هستند. (در معادله (۱۳) ملاحظه می‌شود زمانی که μ را ۰ قرار می‌دهیم تنها روش MKL و زمانی که آن را برابر ۱ قرار می‌دهیم تنها روش ME، یعنی به‌ترتیب معادلات (۹) و (۸)، برقرار می‌ماند.) از آن‌جا که معادلات (۸) و (۹) قابل حل هستند، به‌آسانی می‌توان نتیجه گرفت که معادله (۱۳) نیز قابل حل است.

در مدل بالا، ME به‌عنوان شاخص همواری و MKL به‌عنوان شاخص نیکویی برازش استفاده می‌شوند. از این‌رو، این دو شاخص با یک ضریب خطی μ برای انعکاس دادن وزن‌های مختلف همواری و نیکویی برازش، یک‌پارچه می‌شوند. دلیل اتخاذ یک ترکیب محدب از همواری و نیکویی برازش، مطمئن شدن از تحذب تابع هدف و قابل حل بودن مدل پیشنهاد شده است. در واقع، چگونگی تصمیم گرفتن وزن مناسب بین همواری و نیکویی برازش موضوعی بسیار بحث برانگیز است و معمولاً با روش تجربی تعیین می‌شود. در ادامه این بخش از این روش ترکیبی برای طبقه‌بندی داده‌های مرگ ۲۰۸ موش استفاده می‌کنیم.

مثال: داده‌های جدول (۱) مربوط به بررسی طول عمر ۲۰۸ موش است که تا ستون ۵ برگرفته از [۳] است. در این جدول مقدار X معرف طول عمر موش‌ها است و مقادیرش از ۱ تا ۱۴ سال تغییر می‌کند. در ستون چهارم، پنجم و ششم جدول (۱) مقادیر برآورد شده هر طبقه به‌ترتیب به‌کمک روش بیشینه درست‌نمایی (MLE)، روش بیزی (BE) و روش کولبک-لیبلر بر اساس تابع بقا (KLS) محاسبه و ارائه شده است. مقادیر پارامترهای برآورد شده λ ، c ، به‌روش KLS با نقطه شروع ۰/۱۵ و ۰/۱۵ به‌کمک دستور fsolve در نرم‌افزار MATLAB به‌ترتیب ۰/۰۷۵۷ و ۰/۲۸۰۷ هستند. در پایان این مدل‌ها به‌کمک شاخص‌های χ^2 و $[\Delta^4 \tilde{q}_x]^2$ که به‌ترتیب مربوط به نیکویی برازش و همواری مدل هستند با هم مقایسه شدند. چنان‌که ملاحظه می‌شود روش KLS برای برازش مدل احتمالاتی در مقابل دو روش

MLE و BE مناسب نیست چرا که شاید از نظر نیکویی برازش نسبت به دو روش دیگر بهتر عمل کرده اما از نظر همواری بسیار ضعیف عمل کرده است. (برای جزئیات بیش‌تر به [۷] مراجعه کنید).

در ادامه، در جدول (۲) مقادیر \tilde{q}_x به کمک روش MEMKL برای مقادیر مختلفی از μ تحت قیود مطرح شده در رابطه (۱۳) با k برابر ۵ محاسبه و مدل‌های به‌دست‌آمده به کمک شاخص‌های χ^2 و $[\Delta^4 \tilde{q}_x]^2$ با هم مقایسه شده‌اند. همه محاسبات این جدول مربوط به حل رابطه (۱۳) هستند که به کمک دستور fmincon در نرم‌افزار MATLAB به‌دست‌آمده‌اند و مقدار Exit Flag همه‌ی مدل‌ها ۱ است. نتایج این جدول نشان می‌دهد مدل‌های مربوط به μ برابر ۱ و ۹/۰ بیش‌ترین مقدار χ^2 و کم‌ترین مقدار $[\Delta^4 \tilde{q}_x]^2$ و مدل‌های مربوط به μ برابر ۰ و ۰/۱ کم‌ترین مقدار χ^2 و بیش‌ترین مقدار $[\Delta^4 \tilde{q}_x]^2$ را در بین مدل‌ها دارند.

جدول ۱. داده‌های مربوط به طول عمر ۲۰۸ موش

X	d_x	\tilde{q}_x	MLE	BE	KLS
۱	۳	۰/۰۱۴۴	۰/۰۳۱۱	۰/۰۳۲۱	۰/۰۹۱۹
۲	۳	۰/۰۱۴۴	۰/۰۱۶۸	۰/۰۱۷۳	۰/۱۰۸۴
۳	۶	۰/۰۲۸۸	۰/۰۲۴۴	۰/۰۲۴۹	۰/۱۲۳۱
۴	۶	۰/۰۲۸۸	۰/۰۳۴۹	۰/۰۳۵۴	۰/۱۳۳۰
۵	۱۶	۰/۰۷۶۹	۰/۰۴۹۱	۰/۰۴۹۶	۰/۱۳۴۶
۶	۱۴	۰/۰۶۷۳	۰/۰۶۷۶	۰/۰۶۸	۰/۱۲۴۹
۷	۲۵	۰/۱۲۰۲	۰/۰۹۰۱	۰/۰۹۰۲	۰/۱۰۳۲
۸	۲۰	۰/۰۹۶۲	۰/۱۱۴۴	۰/۱۱۴	۰/۰۷۳۳
۹	۳۲	۰/۱۵۳	۰/۱۳۵۱	۰/۱۳۴۲	۰/۰۴۲۵
۱۰	۲۵	۰/۱۲۰۲	۰/۱۴۳۷	۰/۱۴۲۳	۰/۰۱۸۸
۱۱	۲۷	۰/۱۲۹۸	۰/۱۳۱	۰/۱۲۹۷	۰/۰۱۵۹
۱۲	۱۳	۰/۰۶۲۵	۰/۰۹۵۳	۰/۰۹۴۸	۰/۰۱۱۱
۱۳	۱۱	۰/۰۵۲۹	۰/۰۵۰۱	۰/۰۵۰۵	۰/۰۱۰۱
۱۴	۷	۰/۰۳۳۷	۰/۰۱۶۵	۰/۰۱۷	۰/۰۱۰۰
χ^2			۰/۰۷۵۶	۰/۰۷۳۷	۰/۰۰۲۳
$[\Delta^4 \tilde{q}_x]^2$			۰/۰۰۰۸	۰/۰۰۰۷	۲/۴۸۷۷

توجه داریم که برای برازش بر داده‌ها، مدلی مناسب‌تر است که مقدار $G = \mu[\Delta^4 \tilde{q}_x]^2 + (1-\mu)\chi^2$ کم‌تری داشته باشد. برای انتخاب مدل مناسب، دو مدل اول و آخر را با وجود این‌که هر دو از کم‌ترین مقدار G برخوردارند از مقایسات حذف می‌کنیم زیرا که صرفاً به‌ترتیب فقط به نیکویی برازش و همواری مدل توجه دارند. سپس مدل‌های مربوط به μ برابر ۰/۸ و ۰/۹ را برای برازش بر داده‌ها مناسب معرفی می‌کنیم زیرا کم‌ترین مقدار G را در بین سایر مدل‌ها به خود اختصاص داده‌اند؛ اما با توجه به این‌که در مدل مربوط به μ برابر ۰/۹ نیز بیش‌ترین توجه به نیکویی برازش است مدل مربوط به μ برابر ۰/۸ را که توازن مناسب‌تری بین دو شاخص مذکور برقرار کرده است را مناسب‌تر معرفی می‌کنیم. به‌طور کلی از نتایج بالا، می‌توان نتیجه گرفت که روش پیشنهادی کاراست و ضریب تعدیل نقش مهمی در مبادله کردن همواری و نیکویی برازش ایفا می‌کند که باعث ایجاد انعطاف‌پذیری در طبقه‌بندی داده‌ها است.

تنظیم جدول عمر خلاصه ایران در سال ۱۳۹۰ به کمک روش MEMKL

در این بخش قصد داریم جدول عمر زنان و مردان ایران را در سال ۱۳۹۰ به کمک روش MEMKL تنظیم کنیم. برای این منظور از داده‌های [۱] استفاده می‌کنیم که در جدول ۳ ارائه شده‌اند. نتایج مربوط به استفاده از روش مذکور برای این داده‌ها در جدول ۴ تنظیم شده است. این جدول گویای این موضوع است که در بیش‌تر مواقع به‌ازای تعداد قیود ثابت، مقدار G مدل به‌ازای μ از ۰ تا ۰/۵ افزایش و سپس به‌ازای μ از ۰/۵ تا ۱ کاهش می‌یابد. هم‌چنین واضح

است مدل‌ها بسیار به مقدار μ وابسته هستند از این‌رو، بهتر است به‌طور تجربی انتخاب شود. علاوه بر این، زمانی که μ برابر ۵ است یعنی زمانی که شاخص نیکویی برازش اهمیت کاملی در مقابل شاخص همواری مدل دارد مقادیر G بسیار کوچک‌تر از مواقع دیگر است.

جدول ۲. \tilde{q}_x مربوط به جدول ۱ به کمک روش MEMKL برای مقادیر مختلفی از μ

μ x	۰	۰,۱	۰,۲	۰,۳	۰,۴	۰,۵	۰,۶	۰,۷	۰,۸	۰,۹	۱
۱	۰/۰۱۴۶	۰/۰۱۴۴	۰/۰۱۴۲	۰/۰۱۴	۰/۰۱۳۹	۰/۰۱۳۷	۰/۰۱۳۵	۰/۰۱۳۳	۰/۰۱۳۲	۰/۰۱۳	۰/۰۱۲۸
۲	۰/۰۱۴۲	۰/۰۱۴۵	۰/۰۱۴۹	۰/۰۱۵۳	۰/۰۱۵۶	۰/۰۱۶	۰/۰۱۶۳	۰/۰۱۶۷	۰/۰۱۷۱	۰/۰۱۷۴	۰/۰۱۷۸
۳	۰/۰۲۴۸	۰/۰۲۸۲	۰/۰۲۸۱	۰/۰۲۷۸	۰/۰۲۷۶	۰/۰۲۷۴	۰/۰۲۷۲	۰/۰۲۶۹	۰/۰۲۶۶	۰/۰۲۶۳	۰/۰۲۶۱
۴	۰/۰۲۸۸/۰	۰/۰۲۹۷	۰/۰۳۰۷	۰/۰۳۱۷	۰/۰۳۲۶	۰/۰۳۳۷	۰/۰۳۴۷	۰/۰۳۵۷	۰/۰۳۶۷	۰/۰۳۷۸	۰/۰۳۸۸
۵	۰/۰۷۷۴	۰/۰۷۵۵	۰/۰۷۳۷	۰/۰۷۱۲	۰/۰۶۹۱	۰/۰۶۷	۰/۰۶۵	۰/۰۶۳	۰/۰۶۱۱	۰/۰۵۹۱	۰/۰۵۷۲
۶	۰/۰۶۸۲	۰/۰۶۹۵	۰/۰۷۰۸	۰/۰۷۲۱	۰/۰۷۳۶	۰/۰۷۴۹	۰/۰۷۵۹	۰/۰۷۲۲	۰/۰۷۸۳	۰/۰۷۹۵	۰/۰۸۰۷
۷	۰/۱۲۱۶	۰/۱۲	۰/۱۱۸۷	۰/۱۱۷۲	۰/۱۱۵۵	۰/۱۱۴	۰/۱۱۲۶	۰/۱۱۰۸	۰/۱۰۹۳	۰/۱۰۷۷	۰/۱۰۶۱
۸	۰/۰۹۶۷	۰/۰۹۹۶	۰/۱۰۲۴	۰/۱۰۵۲	۰/۱۰۸۲	۰/۱۱۱۲	۰/۱۱۴۲	۰/۱۱۷۴	۰/۱۲۰۵	۰/۱۲۳۷	۰/۱۲۶۷
۹	۰/۱۵۲۸	۰/۱۵۱۲	۰/۱۴۹۵	۰/۱۴۷۸	۰/۱۴۶۱	۰/۱۴۴۴	۰/۱۴۲۷	۰/۱۴۱	۰/۱۳۹۲	۰/۱۳۷۵	۰/۱۳۵۸
۱۰	۰/۱۱۹۵	۰/۱۲۰۳	۰/۱۲۱۳	۰/۱۲۲۴	۰/۱۲۳۳	۰/۱۲۴۱	۰/۱۲۵	۰/۱۲۵۷	۰/۱۲۶۷	۰/۱۲۷۵	۰/۱۲۸۳
۱۱	۰/۱۲۸۸	۰/۱۲۶۷	۰/۱۲۴۴	۰/۱۲۲	۰/۱۱۹۹	۰/۱۱۷۷	۰/۱۱۵۴	۰/۱۱۳۲	۰/۱۱۱۱	۰/۱۰۸۹	۰/۱۰۶۷
۱۲	۰/۰۶۲۳	۰/۰۶۳۸	۰/۰۶۵۵	۰/۰۶۷	۰/۰۶۸۷	۰/۰۷۰۳	۰/۰۷۲	۰/۰۷۳۷	۰/۰۷۵۳	۰/۰۷۷	۰/۰۷۸۷
۱۳	۰/۰۵۳	۰/۰۵۳۱	۰/۰۵۳	۰/۰۵۳	۰/۰۵۲	۰/۰۵۲۸	۰/۰۵۲۸	۰/۰۵۲۶	۰/۰۵۲۵	۰/۰۵۲۴	۰/۰۵۲۲
۱۴	۰/۰۳۳	۰/۰۳۳۶	۰/۰۳۳۴	۰/۰۳۳	۰/۰۳۳	۰/۰۳۲۹	۰/۰۳۲۷	۰/۰۳۲۶	۰/۰۳۲۴	۰/۰۳۲۳	۰/۰۳۲۲
χ^2	$\hat{e}-۰.۵$	۰/۰۰۰۴	۰/۰۰۱۴	۰/۰۰۳	۰/۰۰۵۴	۰/۰۰۸۳	۰/۰۱۱۹	۰/۰۱۶۳	۰/۰۲۱۱	۰/۰۲۶۸	۰/۰۳۳۱/۰
$[\Delta^4 \tilde{q}_x]^2$	۰/۶۱۳۱	۰/۵۰۱	۰/۳۹۶۵	۰/۳۰۴۳	۰/۲۲۲۴	۰/۱۵۴۹	۰/۱۰۰۷	۰/۰۵۵۹	۰/۰۲۵۵	۰/۰۰۶۴	۰/۰۰۰۱
G	$\hat{e}-۰.۵$	۰/۰۵۰۴	۰/۰۸۰۴	۰/۰۹۳۴	۰/۰۹۲۲	۰/۰۸۱۶	۰/۰۶۵۲	۰/۰۴۴	۰/۰۲۶۴	۰/۰۰۸۵	۰/۰۰۰۱

جدول ۳. داده‌های مربوط به مرگ‌ومیر زنان و مردان ایران سال ۱۳۹۰

جنسیت	مردان		زنان	
x	d_x	\tilde{q}_x	d_x	\tilde{q}_x
۰	۲۳۸۹	۰/۰۲۳۸۹	۲۰۱۹	۰/۰۲۰۱۹
۱	۳۹۸	۰/۰۰۳۹۸	۳۷۷	۰/۰۰۳۷۷
۵	۲۴۷	۰/۰۰۲۴۷	۱۸۴	۰/۰۰۱۸۴
۱۰	۲۰۷	۰/۰۰۲۰۷	۱۵۰	۰/۰۰۱۵
۱۵	۳۹۸	۰/۰۰۳۹۸	۲۴۷	۰/۰۰۲۴۷
۲۰	۵۵۰	۰/۰۰۵۵	۳۴۸	۰/۰۰۳۴۸
۲۵	۵۴۲	۰/۰۰۵۴۲	۴۳۳	۰/۰۰۴۳۳
۳۰	۶۰۸	۰/۰۰۶۰۸	۵۴۴	۰/۰۰۵۴۴
۳۵	۷۹۰	۰/۰۰۷۹	۷۳۴	۰/۰۰۷۳۴
۴۰	۱۱۹۵	۰/۰۱۱۹۵	۱۰۶۱	۰/۰۱۰۶۱
۴۵	۱۹۸۶	۰/۰۱۹۸۶	۱۶۴۴	۰/۰۱۶۴۴
۵۰	۳۲۱۲	۰/۰۳۲۱۲	۲۴۶۸	۰/۰۲۴۶۸
۵۵	۵۱۶۹	۰/۰۵۱۶۹	۳۷۰۲	۰/۰۳۷۰۲
۶۰	۷۶۹۵	۰/۰۷۶۹۵	۵۶۶۴	۰/۰۵۶۶۴
۶۵	۱۰۹۷۱	۰/۱۰۹۷۱	۸۹۰۴	۰/۰۸۹۰۴
۷۰	۱۴۵۶۵	۰/۱۴۵۶۵	۱۳۲۸۸	۰/۱۳۲۸۸
۷۵	۱۶۹۶۱	۰/۱۶۹۶۱	۱۷۵۵۴	۰/۱۷۵۵۴
۸۰	۳۲۱۱۶	۰/۳۲۱۱۶	۴۰۶۷۹	۰/۴۰۶۷۹

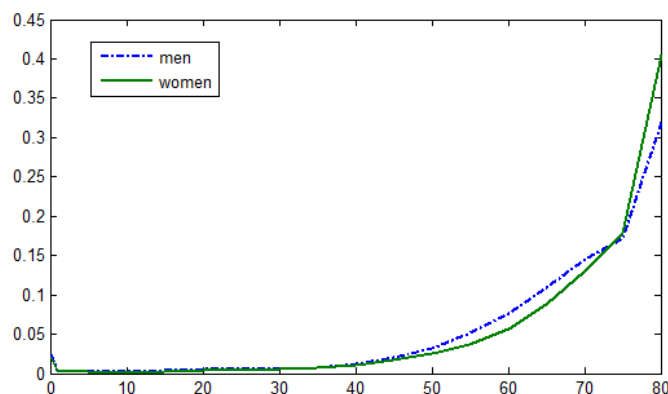
در این جا با حذف دو مدل ME و MKL (به تنهایی) و مقایسه مقادیر G باقی مانده، مقدار k و μ را به ترتیب ۴ و ۰/۱ انتخاب می‌کنیم. در پایان بر اساس مدل منتخب، جدول عمر سال ۱۳۹۰ زنان و مردان ایران را در جدول ۵ تنظیم و روند مرگ و میر سنین مشخص شده آنها را در نمودار شکل ۱ رسم کرده‌ایم. چنان که در این شکل ملاحظه می‌شود مقادیر برآورد شده احتمالات مرگ و میر مردان و زنان ایرانی بسیار نزدیک به هم است و هر دو روند صعودی دارند و در سنین ۷۵ تا ۸۰ سالگی احتمال مرگ و میر زنان بیش‌تر از مردان تخمین زده شده است.

جدول ۴. $G \times 10^{-6}$ مدل‌های مربوط به داده‌های جدول ۳ به کمک روش MEMKL برای مقادیر مختلفی از k و μ

جنسیت	مردان					زنان				
μ	k					k				
	۱	۲	۳	۴	۵	۱	۲	۳	۴	۵
۰	۱	۰	۱	۳۵	۱۱۳۷۵۴	۰	۰	۱	۱۳	۱۸۶۶۶۶
۱/۰	۳۵۰۲۹	۲۶۱۶۷۶	۲۳۵۷	۲۳۴۸	۱۰۳۵۷۰	۵۴۵۴	۳۸۳۱	۳۷۵۸	۳۴۵۷	۲۵۹۶۴۹
۲/۰	۸۶۱۶۷	۵۵۹۴۹۶	۴۴۳۱	۴۲۸۴	۸۶۵۲۱	۱۲۷۹۵	۷۲۳۴	۶۸۵۳	۶۴۲۸	۲۱۷۷۸۲
۳/۰	۱۴۶۸۰۴	۸۴۹۲۸۴	۶۲۱۷	۴۹۸۵	۹۲۴۶۰	۲۱۵۱۱	۹۹۳۵	۹۰۹۶	۸۰۰۸	۲۱۴۲۸۵
۴/۰	۲۱۱۰۵۶	۱۰۹۹۶۹	۷۳۴۶	۶۳۶۷	۷۵۴۶۳	۳۰۸۸۲	۱۱۷۱۶	۱۰۵۲۱	۱۰۲۰۳	۱۵۴۳۹۱
۵/۰	۲۷۴۴۹۲	۱۲۹۸۱۲	۸۰۳۴	۶۷۲۶	۵۹۲۵۸	۴۰۴۹۸	۱۲۸۱۱	۱۰۹۲۴	۱۰۰۷۷	۱۳۱۲۰۵
۶/۰	۳۲۸۱۳۲	۱۳۹۷۹۶	۸۰۲۷	۶۶۵۶	۴۷۱۳۷	۴۹۱۸۱	۱۲۸۹۹	۱۰۶۰۱	۱۰۳۴۸	۱۰۱۸۱
۷/۰	۳۵۶۸۱۳	۱۳۶۴۷۷	۷۳۷۵	۵۸۸۷	۶۳۱۲	۵۴۵۷۹	۱۱۹۳۴	۹۲۵۷	۹۵۲۴	۹۲۴۵
۸/۰	۳۳۶۷۸۶	۱۱۷۴۲۵	۵۸۸۰	۴۷۷۰	۵۱۹۲	۵۴۳۹۹	۹۵۷۱	۷۱۴۹	۸۱۲۴	۷۸۶۶
۹/۰	۲۳۷۰۱۲	۷۴۹۱۳۹	۳۴۸۹	۳۰۱۹	۳۳۶۰	۴۰۳۹۰	۶۰۱۲	۴۱۰۴	۶۰۹۴	۵۸۲۸
۱	۵	۱۱۳	۵۱	۶۱۹	۹۱۸	۲۰	۵۹۲	۱۰۵	۳۴۹۳	۲۲۹۱۱

نتیجه‌گیری

در این مقاله از معیارهای نظریه اطلاع برای بررسی توزیع بقا و توزیع احتمال مرگ و میر کمک گرفتیم. از جمله این معیارها اصل بیشینه آنتروپی شانون و اصل کمینه آنتروپی کولبک-لیبلر هستند. در بررسی مرگ و میر یکی از ابزارهای موجود طبقه‌بندی اطلاعات، جدول عمر است و در طبقه‌بندی داده‌ها همواری و نیکویی برازش مدل برآورد شده اهمیت ویژه‌ای دارد. این دو معیار به ترتیب معادل با اصل بیشینه آنتروپی شانون و اصل کمینه آنتروپی کولبک لیبلر هستند. ما با کمینه کردن مدل ترکیبی دو اصل فوق، توزیع احتمالاتی مرگ و میر را برآورد کردیم و از این روش برای تنظیم جدول عمر استفاده کردیم. بر این اساس جدول عمر زنان و مردان ایران در سال ۱۳۹۱ تنظیم شده است. به‌طور کلی، می‌توان نتیجه گرفت که روش پیشنهادی کاراست و ضریب تعدیل نقش مهمی در مبادله کردن همواری و نیکویی برازش ایفا می‌کند که باعث ایجاد انعطاف‌پذیری در طبقه‌بندی داده‌ها است.



شکل ۱. نمودار روند احتمال مرگ و میر زنان و مردان ایران در سال ۱۳۹۰، برآورد شده به کمک روش MEMKL

جدول ۵. جدول عمر زنان و مردان ایران ۱۳۹۰ به روش MEMKL

جنسیت	مردان					زنان				
x	l_x	d_x	\hat{q}_x	${}_d m_x$	e_x	l_x	d_x	\hat{q}_x	${}_d m_x$	e_x
۰	۱۰۰۰۰۰	۲۳۸۹	۰.۲۳۲۸۴/۰	۰.۲۳۵۵۹/۰	۶۵/۷۰	۱۰۰۰۰۰	۲۰۱۹	۰.۱۹۷۳/۰	۰.۱۹۹۲۷/۰	۴۱/۷۳
۱	۹۷۶۱۱	۳۹۸	۰.۰۴۵۳/۰	۰.۰۴۵۴/۰	۳۸/۷۱	۹۷۹۸۱	۳۷۷	۰.۰۴۱۷۷/۰	۰.۰۴۱۸۶/۰	۹۲/۷۳
۵	۹۷۲۱۳	۲۴۷	۰.۰۲۷۰۷/۰	۰.۰۲۷۱۱/۰	۶۶/۶۷	۹۷۶۰۴	۱۸۴	۰.۰۱۹۸۶/۰	۰.۰۱۹۸۸/۰	۲۰/۷۰
۱۰	۹۶۹۶۶	۲۰۷	۰.۰۲۱۵۹/۰	۰.۰۲۱۶۱/۰	۸۳/۶۲	۹۷۴۲۰	۱۵۰	۰.۰۱۵۹۶/۰	۰.۰۱۵۹۷/۰	۳۳/۶۵
۱۵	۹۶۷۵۲	۳۹۸	۰.۰۳۸۴۱/۰	۰.۰۳۸۴۸/۰	۹۶/۵۷	۹۷۲۷۰	۲۴۷	۰.۰۲۴۱۸/۰	۰.۰۲۴۲۱/۰	۴۳/۶۰
۲۰	۹۶۳۶۱	۵۵۰	۰.۰۵۱۹/۰	۰.۰۵۲۰۴/۰	۱۹/۵۳	۹۷۰۲۲	۳۴۸	۰.۰۳۳۵۶/۰	۰.۰۳۳۶۲/۰	۵۷/۵۵
۲۵	۹۵۸۱۰	۵۴۲	۰.۰۵۲۸/۰	۰.۰۵۲۹۴/۰	۴۸/۴۸	۹۶۶۷۵	۴۳۳	۰.۰۴۱۷۵/۰	۰.۰۴۱۸۴/۰	۷۶/۵۰
۳۰	۹۵۲۶۸	۶۰۸	۰.۰۶۰۶۳/۰	۰.۰۶۰۸۱/۰	۷۴/۴۳	۹۶۲۴۱	۵۴۴	۰.۰۵۳۱۴/۰	۰.۰۵۳۲۸/۰	۹۸/۴۵
۳۵	۹۴۶۶۰	۷۹۰	۰.۰۸۰۲۴/۰	۰.۰۸۰۵۶/۰	۰۰/۳۹	۹۵۶۹۸	۷۳۴	۰.۰۷۲۹/۰	۰.۰۷۳۱۷/۰	۲۳/۴۱
۴۰	۹۳۸۷۰	۱۱۹۵	۰.۱۲۲۱۳/۰	۰.۱۲۲۸۸/۰	۳۱/۳۴	۹۴۹۶۳	۱۰۶۱	۰.۱۰۶۹۹/۰	۰.۱۰۷۵۷/۰	۵۲/۳۶
۴۵	۹۲۶۷۵	۱۹۸۶	۰.۲۰۰۷۸/۰	۰.۲۰۲۸۲/۰	۷۲/۲۹	۹۳۹۰۲	۱۶۴۴	۰.۱۶۷۴۱/۰	۰.۱۶۸۸۳/۰	۹۱/۳۱
۵۰	۹۰۶۸۹	۳۲۱۲	۰.۳۲۴۸۲/۰	۰.۳۳۰۲۱/۰	۳۱/۲۵	۹۲۲۵۸	۲۴۶۸	۰.۲۵۰۳۳/۰	۰.۲۵۳۵۲/۰	۴۳/۲۷
۵۵	۸۷۴۷۷	۵۱۶۹	۰.۵۱۴۹۳/۰	۰.۵۲۸۶۶/۰	۱۴/۲۱	۸۹۷۹۱	۳۷۰۲	۰.۳۷۴۸۸/۰	۰.۳۸۲۰۹/۰	۱۱/۲۳
۶۰	۸۲۳۰۹	۷۶۹۵	۰.۷۶۸۳۴/۰	۰.۷۹۹۴۶/۰	۳۰/۱۷	۸۶۰۸۹	۵۶۶۴	۰.۵۶۰۹۴/۰	۰.۵۷۷۲۹/۰	۹۹/۱۸
۶۵	۷۴۶۱۳	۱۰۹۷۱	۱.۰۸۶۵/۰	۱.۱۵۰۱۸/۰	۸۱/۱۳	۸۰۴۲۵	۸۹۰۴	۰.۸۸۳۴۶/۰	۰.۹۲۴۹۵/۰	۱۴/۱۵
۷۰	۶۳۶۴۲	۱۴۵۶۵	۱.۴۵۱۶۸/۰	۱.۵۶۸۵/۰	۷۳/۱۰	۷۱۵۲۱	۱۳۲۸۸	۱.۳۱۲۳۱/۰	۱.۴۰۶۷۸/۰	۶۹/۱۱
۷۵	۴۹۰۷۷	۱۶۹۶۱	۱.۷۱۸۹۱/۰	۱.۸۸۶۱/۰	۱۵/۸	۵۸۲۳۳	۱۷۵۵۴	۱.۷۸۸/۰	۱.۹۶۹۸۹/۰	۷۵/۸
۸۰	۳۲۱۱۶	۳۲۱۱۶	۳۲۰.۱۱۳/۰	۳۸۵۸۲۹/۰	۱۱/۶	۴۰۶۷۲	۴۰۶۷۹	۴۰۵۵۲۶/۰	۵۲۰.۰۷۸/۰	۴۱/۶

منابع

۱. طه حسینی، "ساخت جدول عمر سالانه برای ایران"، مرکز آمار ایران، پژوهشکده آمار (۱۳۹۲).
2. Alicja S. N., William F. S., "An extension of the Whittaker-Henderson method of graduation", Scand. Actuar J, 1 (2012) 70-79.
3. Ananda M. M, Dalpatadu R. J., Singh A. K., "estimating parameters of the force of mortality in actuarial studies", Actuar. Res. Clear. House, 1 (1993) 129-141.
4. Bohlmann G., "Ein Ausgleichungs ProblME", In Nachrichten von der Königl Gesellschaft der Wissenschaften zu Göttingen, MathMEatisch-physikalische Klasse; Horstmann, L., Ed., Commissionsverlag der Dieterich'schen Universitätsbuchhandlung: Göttingen, Germany, (1899) 260-271.
5. Henderson R., "A new method of graduation", Trans. Actuar. Soc. Am., 25 (1924) 29-53.
6. Henderson R., "Further rMEarks on graduation", Trans. Actuar. Soc. Am., 26 (1925) 52-74.
7. He D., Huang Q. Gao J., "A New Entropy Optimization Model for Graduation of Data in Survival Analysis", Entropy, 14 (2012) 1306-1316.
8. Kapur J. N., Kesavan H. K., "Entropy optimization Principles with Applications", AcadMEic Press Inc.: San Diego, CA, USA, (1992).
9. Kullback S., Leibler R. A., "on Information and Sufficiency", Annals of MathMEatical Statistics, 22 (1) (1951) 79-86.
10. Liu J., "Information theoretic content and probability", Ph.D.Thesis, University of Florida, USA, (2007) 25-26.

11. M. Cover T., A. Thomas J., "Elements of information theory", Second Edition, Published by John Wiley & Sons, Inc., Hoboken, New Jersey (2006).
12. Schoen R., "Modeling multigroup populations", plenum press, New York, (1899).
13. Whittaker E. T., "On a new method of graduation", Proc. Edinb. Math. Soc, 41 (1923) 63-75.
14. Yari G. H., Mirhabibi A., Saghafi A., "Estimation of the Weibull parameters by Kullback Leibler divergence of Survival functions", Appl. Math. Inf. Sci, 7, No. 1 (2013) 187-192.