

آزمون نیکویی برازش توزیع نمایی بر مبنای برآوردگرهای جدید آنتروپی

احسان زمانزاده

گروه آمار، دانشگاه اصفهان

تاریخ دریافت: ۱۳۹۲/۲/۲۷ تاریخ آخرین بازنگری: ۱۳۹۲/۵/۲۱

چکیده: در این مقاله، ابتدا دو برآوردگر جدید آنتروپی معرفی می شود. سپس آزمون نیکویی برازش فرضیه نمایی بودن توزیع جامعه بر مبنای برآوردگرهای جدید معرفی می شود و توان آن‌ها با توان سایر آزمون‌های بر مبنای آنتروپی توزیع نمایی مورد مقایسه قرار می گیرد. نتایج مطالعات شبیه سازی نشان می دهد که برآوردگرهای پیشنهادی عموماً عملکرد بهتری در مقایسه با سایر برآوردگرها در برآورد آنتروپی و آزمون نیکویی برازش دارند.

واژه‌های کلیدی: آنتروپی، آزمون نیکویی برازش، فاصله کولبک لایبلر.

۱ مقدمه

فرض کنید X یک متغیر تصادفی پیوسته، با تابع چگالی احتمال $f(x)$ باشد، شانون (۱۹۴۸) آنتروپی این متغیر تصادفی را به صورت

$$H(f) = - \int_{-\infty}^{\infty} \log(f(x)) f(x) dx \quad (1)$$

آدرس الکترونیک مسئول مقاله: احسان زمانزاده، E.Zamanzade@sci.ui.ac.ir
کد موضوع بندی ریاضی (۲۰۱۰): ۶۲G۱۰، ۶۲B۱۰

۶۲ آزمون نیکویی برازش توزیع نمایی بر مبنای برآوردهای آنتروپی

تعریف کرد. از آنجا که مفهوم آنتروپی کاربردهای فراوانی در مباحث آماری نظیر نظریه اطلاع و آزمون نیکویی برازش دارد، مسئله برآورد $H(f)$ توسط محققین زیادی مورد بررسی قرار گرفته است. واسیچک (۱۹۷۶) برای نخستین بار، آزمون نیکویی برازش برای توزیع نرمال را بر مبنای برآوردهای جدیدی از آنتروپی پیشنهاد داد. نتایج شبیه سازی واسیچک نشان می داد که آزمون بر مبنای آنتروپی، علی رغم سادگی در محاسبه دارای توان های بسیار خوبی در مقایسه با آزمون پرتوان شاپیرو-ویلک (۱۹۶۵) برای توزیع نرمال است. از آن به بعد محققین زیادی سعی در بهبود برآوردهای آزمون پیشنهاد شده توسط واسیچک (۱۹۷۶) و تعمیم آن به سایر توزیع های آماری داشته اند.

واسیچک (۱۹۷۶) نشان داد که می توان رابطه (۱) را به صورت

$$H(f) = - \int_0^1 \log\left(\frac{d}{dp} F^{-1}(p)\right) dp$$

بازنویسی کرد و با جایگزینی تابع توزیع F با توزیع تجربی F_n و استفاده از عملگر تفصل به جای عملگر مشتق، برآوردهای آنتروپی خود را به صورت زیر معرفی کرد. فرض کنید X_1, \dots, X_n نمونه ای تصادفی از توزیع پیوسته F باشد، در این صورت برآوردهای آنتروپی واسیچک (۱۹۷۶) عبارتست از

$$HV = \frac{1}{n} \sum_{i=1}^n \log\left(\frac{n}{c_i} (X_{(i+m)} - X_{(i-m)})\right) \quad (2)$$

که در آن $X_{(1)}, \dots, X_{(n)}$ آماره های مرتب، $c_i = 2m$ و m یک عدد مثبت کوچکتر یا مساوی $\frac{n}{2}$ است. ضمناً به ازای $i < 1$ $X_{(i)} = X_{(1)}$ و به ازای $i > n$ $X_{(i)} = X_{(n)}$ می باشد.

واضح است که عبارت داخل لگاریتم در (۲)، هنگامی که $i \leq m$ یا $i \geq n - m$ تقریب مناسبی برای برآورد $\frac{d}{dp} F^{-1}(p)$ نیست. ابراهیمی و همکاران (۱۹۹۴) برآوردهای پیشنهادی خود را بر مبنای اصلاح ضرایب c_i هنگامی که $i \leq m$ یا $i \geq n - m$ است، به صورت

$$HE = \frac{1}{n} \sum_{i=1}^n \log\left(\frac{n}{c_i} (X_{(i+m)} - X_{(i-m)})\right) \quad (3)$$

ارائه کردند، که در آن

$$c_i = \begin{cases} m + i - 1, & 1 \leq i \leq m \\ 2m, & m + 1 \leq i \leq n - m \\ m + n - i, & n - m + 1 \leq i \leq n \end{cases}$$

به ازای $i < 1$ ، $X_{(i)} = X_{(1)}$ و به ازای $i > n$ ، $X_{(i)} = X_{(n)}$ است.

نتایج شبیه‌سازی ابراهیمی و همکاران (۱۹۹۴) برای سه توزیع نرمال، نمایی و یکنواخت نشان داد که برآوردگر پیشنهادی آن‌ها آریبی و جذر میانگین مربع خطای کمتری نسبت به برآورد پیشنهادی واسیچک (۱۹۷۶) به ازای حجم نمونه 10^0 ، 2^0 و 3^0 و مقادیر مختلف m دارد.

کوریا (۱۹۹۵) اصلاح دیگری از برآوردگر واسیچک (۱۹۷۶) ارائه کرد که میانگین مربع خطای کمتری نسبت به برآورد پیشنهادی توسط واسیچک (۱۹۷۶) داشت. وی توجه کرد که می‌توان رابطه (۲) را به صورت

$$HV_{nm} = -\frac{1}{n} \sum_{i=1}^n \log\left(\frac{\frac{i+m}{n} - \frac{i-m}{n}}{X_{(i+m)} - X_{(i-m)}}\right) \quad (4)$$

بازنویسی کرد. اما عبارت داخل لگاریتم در رابطه (۴) در واقع شیب خطی است که نقاط $(\hat{F}(X_{(i+m)}), X_{(i+m)})$ و $(\hat{F}(X_{(i-m)}), X_{(i-m)})$ را به یکدیگر متصل می‌کند. او پیشنهاد داد که این شیب را با استفاده از رگرسیون خطی موضعی برحسب $\{X_{(i-m)}, \dots, X_{(i+m)}\}$ و با استفاده از تمام $2m + 1$ نقطه، به جای تنها دو نقطه، برآورد کنیم. لذا با در نظر گرفتن رابطه

$$F(X_{(j)}) = \alpha + \beta X_{(j)} + \epsilon$$

و برآورد β با روش کمترین توان‌های دوم، برآورد خود را به صورت

$$HC_{nm} = -\frac{1}{n} \sum_{i=1}^n \log(b_i) \quad (5)$$

ارائه کرد، که در آن

$$b_i = \frac{\sum_{j=i-m}^{i+m} (X_{(j)} - \bar{X}_{(i)}) \left(\frac{j-i}{n}\right)}{\sum_{j=i-m}^{i+m} (X_{(j)} - \bar{X}_{(i)})^2} \quad (6)$$

۶۴.....آزمون نیکویی برازش توزیع نمایی بر مبنای برآوردهای آنتروپی

$$X_{(i)} = X_{(n)}, i > n \text{ و به ازای } X_{(i)} = X_{(1)}, i < 1 \text{ و } \bar{X}_{(i)} = \sum_{j=i-m}^{i+m} \frac{X_{(j)}}{r_{m+1}} \text{ می باشد.}$$

مطالعه شبیه سازی برآوردهای کوریا (۱۹۹۵) و واسیچک (۱۹۷۶) نشان داد که این برآوردها عموماً از میانگین توان دوم خطای کمتری نسبت به برآوردها واسیچک (۱۹۷۶) در سه توزیع نرمال، نمایی و یکنواخت برخوردار است. کوریا (۱۹۹۵) هیچ مقایسه ای میان برآوردها پیشنهادی خود، با برآوردها پیشنهادی ابراهیمی و همکاران (۱۹۹۴) انجام نداده است.

یوسف زاده و ارقامی (۲۰۰۸) برآوردها از آنتروپی بر مبنای برآوردها جدیدی از تابع توزیع جامعه ارائه دادند. آنها ابتدا برآوردها از تابع توزیع پیوسته به صورت

$$\hat{F}_y(t) = \begin{cases} \frac{n-1}{n(n+1)} \left(\frac{n}{n-1} + \frac{t-X_{(r)}}{X_{(r)}-X_{(1)}} + \frac{t-X_{(1)}}{X_{(r)}-X_{(1)}} \right), & X_{(1)} \leq t \leq X_{(r)} \\ \frac{n-1}{n(n+1)} \left(i + \frac{1}{n-1} + \frac{t-X_{(i-1)}}{X_{(i+1)}-X_{(i-1)}} + \frac{t-X_{(i)}}{X_{(i+r)}-X_{(i)}} \right) & X_{(i)} \leq t \leq X_{(i+1)} \\ & i = 1, \dots, n-2 \\ \frac{n-1}{n(n+1)} \left(n-1 + \frac{1}{n-1} + \frac{t-X_{(n-r)}}{X_{(n)}-X_{(n-r)}} + \frac{t-X_{(n-1)}}{X_{(n+1)}-X_{(n-1)}} \right) & X_{(n-1)} \leq t \leq X_{(n)} \end{cases}$$

معرفی کردند، که در آن $X_{(n+1)} = X_{(1)}$ و $X_{(n)} = X_{(1)} - \frac{n}{n-1}(X_{(r)} - X_{(1)})$ است. آنگاه برآوردها

$$HY = \sum_{i=1}^n \log \left(\frac{X_{(i+m)} - X_{(i-m)}}{\hat{F}_y(X_{(i+m)}) - \hat{F}_y(X_{(i-m)})} \right) \frac{\hat{F}_y(X_{(i+m)}) - \hat{F}_y(X_{(i-m)})}{\sum_{i=1}^n (\hat{F}_y(X_{(i+m)}) - \hat{F}_y(X_{(i-m)}))} \quad (7)$$

را برای برآوردها آنتروپی یک متغیر تصادفی پیوسته پیشنهاد دادند، که به ازای $i < 1$ ، $X_{(i)} = X_{(1)}$ و به ازای $i > n$ ، $X_{(i)} = X_{(n)}$ می باشد. سپس در مطالعه شبیه سازی نشان دادند برآوردها پیشنهادی آنان، دارای اریبی و میانگین توان دوم خطای کمتری نسبت به برآوردهای واسیچک (۱۹۷۶) و ابراهیمی و همکاران (۲۰۰۴) در سه توزیع نرمال، نمایی و یکنواخت است.

زمان زاده و ارقامی (۱۳۸۷)، دو اصلاح مختلف از برآوردها کوریا (۱۹۹۵) پیشنهاد دادند: اصلاح اول بر مبنای این ایده بود که در محاسبه مقدار b_i در رابطه (۶)، هنگامی که $i \leq m$ یا $i \geq n - m + 1$ است تقریب مناسبی برای شیب خط نیست زیرا در این نقاط بیش از یک بار از $X_{(1)}$ و $X_{(n)}$ استفاده شده است. لذا آنان

اصلاح

$$HCZ_1 = -\frac{1}{n} \sum_{i=1}^n \log(b_i^*), \quad (8)$$

را پیشنهاد دادند، که در آن

$$b_i^* = \frac{\sum_{j=K_1(i)}^{K_2(i)} (X_{(j)} - \tilde{X}_{(i)}) (\hat{F}_n(X_{(j)}) - \tilde{F}_n(j))}{\sum_{j=K_1(i)}^{K_2(i)} (X_{(j)} - \tilde{X}_{(i)})^2}, \quad i = 1, \dots, n$$

$$k_1(i) = \begin{cases} 1, & 1 \leq i \leq m \\ i - m, & i \geq m + 1 \end{cases}, \quad k_2(i) = \begin{cases} i + m, & 1 \leq i \leq n - m \\ n, & i \geq n - m + 1 \end{cases}$$

تابع توزیع تجربی است. $\tilde{X}_{(i)} = \sum_{j=k_1(i)}^{k_2(i)} \frac{X_{(j)}}{k_2(i) - k_1(i) + 1}$ و $\tilde{F}_n(i) = \sum_{j=1}^{k_1(i)} \frac{\hat{F}_n(X_{(j)})}{k_2(i) - k_1(i) + 1}$ برآوردگر

برای دومین اصلاح چون در به دست آوردن b_i از تعداد مساوی مشاهده استفاده شده است، همه b_i ها از وزنهای مساوی برخوردارند. اما در محاسبه HCZ_1 وقتی که $i \leq m$ یا $i \geq n - m + 1$ از مشاهدات کمتری برای به دست آوردن b_i ها استفاده شده است، لذا منطقی به نظر می رسد که وزنهای کمتری به این مشاهدات اختصاص داده شود. بر مبنای این ایده، اصلاح دوم به صورت

$$HCZ_2 = -\sum_{i=1}^n w_i \log(b_i^*) \quad (9)$$

ارائه شد، که در آن $w_i = \frac{\hat{F}_n(X_{(i+m)}) - \hat{F}_n(X_{(i-m)})}{\sum_{i=1}^n (\hat{F}_n(X_{(i+m)}) - \hat{F}_n(X_{(i-m)}))}$ است. مطالعه شبیه سازی

زمانزاده و ارقامی (۱۳۸۷) نشان داد که این برآوردگرها عملکرد خوبی در مقایسه با برآوردگرهای آنتروپی و آسیچک (۱۹۷۶)، ابراهیمی و همکاران (۱۹۴۴) و کوریا (۱۹۹۵) در برآورد آنتروپی و آزمون نیکویی برازش دارند. علیزاده نوقابی و علیزاده نوقابی (۱۳۸۷) به مقایسه آزمون نیکویی برازش با سایر روشها پرداختند و حییبی راد و ارقامی (۱۳۸۶) آزمون تقارن توزیع را بر مبنای آنتروپی پیشنهاد دادند.

کولبک و لایبلر (۱۹۵۱)، تابع اطلاع تشخیص^۱ را معرفی کردند، که یک معیار

^۱ Discrimination information function

۶۶آزمون نیکویی برازش توزیع نمایی بر مبنای برآوردهای آنتروپی

مهم برای مقایسه دو توزیع است. به این تابع که براساس آنتروپی نیز می توان آن را بازنویسی کرد، آنتروپی متقاطع^۲ هم گفته می شود. فرض کنید بر مبنای مشاهده $X = x$ ، یکی از دو توزیع F یا G را به عنوان توزیع متغیر تصادفی X ، با تکیه گاه S مورد نظر باشد. در این صورت با توجه به قضیه بیز می توان نوشت:

$$\log\left(\frac{f(x)}{g(x)}\right) = \log\left(\frac{P(F|x)}{P(G|x)}\right) - \log\left(\frac{P(F)}{P(G)}\right)$$

که در آن $f(\cdot)$ و $g(\cdot)$ توابع چگالی احتمال، $P(\cdot|x)$ احتمال پسین و $P(\cdot)$ احتمال پیشین است، بنابراین $\log\left(\frac{f(x)}{g(x)}\right)$ را که در واقع تفاضل لگاریتم $\frac{P(F|x)}{P(G|x)}$ از لگاریتم $\frac{P(F)}{P(G)}$ است، می توان به عنوان میزان اطلاع موجود در مشاهده $X = x$ ، در پشتیبانی از F در برابر G یا لگاریتم نسبت بخت های F نسبت به G تفسیر کرد.

حال اگر مشاهده $X = x$ متعلق به تکیه گاه S باشد، آنگاه متوسط اطلاع موجود در این مشاهده، در حمایت از F در مقابل G عبارتست از:

$$D(f, g) = \int \log\left(\frac{f(x)}{g(x)}\right) dF(x).$$

همچنین می توان نشان داد که $D(f, g) \geq 0$ ، و تساوی رخ می دهد اگر و تنها اگر $f \stackrel{a.s.}{=} g$.

در بخش ۲ برآوردهای جدید آنتروپی به عنوان اصلاح برآوردهای یوسفزاده و ارقامی (۲۰۰۸) معرفی می شود. سپس در مطالعه ای شبیه سازی برآوردهای پیشنهادی با سایر برآوردهای آنتروپی مقایسه خواهند شد. در بخش ۳ آزمون نیکویی برازش برای فرضیه های نمایی بودن توزیع جامعه بر اساس برآوردهای جدید آنتروپی معرفی می شود و توان این آزمون ها با سایر آزمون های نمایی بودن توزیع جامعه بر مبنای برآوردهای آنتروپی و اسیچک (۱۹۷۶)، کوریا (۱۹۹۵)، یوسفزاده و ارقامی (۲۰۰۸) و زمانزاده و ارقامی (۱۳۸۷) مقایسه خواهد شد. بحث و نتیجه گیری نهایی در بخش ۴ آورده شده است.

^۲ Cross entropy

۲ برآوردگرهای جدید آنتروپی

همان‌طور که در رابطه (۷) ملاحظه می‌شود، در برآوردگر یوسف‌زاده و ارقامی (۲۰۰۸)، همانند برآوردگر واسیچک (۱۹۷۶)، تنها از نقاط $(\hat{F}_y(X_{(i+m)}), X_{(i+m)})$ و $(\hat{F}_y(X_{(i-m)}), X_{(i-m)})$ برای برآورد $\frac{d}{dp}F^{-1}(p)$ استفاده می‌شود و در واقع در رابطه $\frac{X_{(i+m)} - X_{(i-m)}}{\hat{F}_y(X_{(i+m)}) - \hat{F}_y(X_{(i-m)})}$ شیب خطی است که این نقاط را به یکدیگر متصل می‌کند. اما همان‌طور که در برآوردگر کوریا (۱۹۹۵)، استفاده از تمامی نقاط $\{X_{(i-m)}, \dots, X_{(i+m)}\}$ به جای دو نقطه و کاربرد رگرسیون خطی موضعی، منجر به بهبود برآوردگر واسیچک (۱۹۷۶) شد، انتظار می‌رود که این روش، در اینجا نیز منجر به بهبود برآوردگر پیشنهادی یوسف‌زاده و ارقامی (۲۰۰۸) شود. لذا بر مبنای ایده زمانزاده و ارقامی (۱۳۸۷) بر مبنای اصلاح برآوردگر کوریا (۱۹۹۵)، برآوردگرهای جدید به صورت

$$HYZ_1 = -\frac{1}{n} \sum_{i=1}^n \log(b_i^{**}), \quad (10)$$

$$HYZ_2 = -\sum_{i=1}^n w_i \log(b_i^{**}), \quad (11)$$

ارئه می‌شوند، که در آن وزن‌های به کار رفته در HCZ_2 ، b_i^{**} ضرایب معرفی شده در (۹) است با این تفاوت که از \hat{F}_y معرفی شده در رابطه (۷) به جای برآوردگر تابع توزیع تجربی استفاده شده است.

قضیه ۱: فرض کنید X_1, \dots, X_n نمونه‌ای تصادفی از جامعه‌ای با آنتروپی $H^X(f)$ باشد و $Y_i = kX_i$ که در آن $k > 0$ و $i = 1, \dots, n$. همچنین فرض کنید که HYZ_i^{kX} و HYZ_i^X ($i = 1, 2$) به ترتیب برآوردهای آنتروپی $H^{kX}(g)$ و $H^X(f)$ باشند، که در آن g تابع چگالی kX است. در این صورت، روابط زیر برقرارند:

الف- $E(HYZ_i^{kX}) = E(HYZ_i^X) + \log(k)$ ، $i = 1, 2$

ب- $V(HYZ_i^{kX}) = V(HYZ_i^X)$ ، $i = 1, 2$

ج- $MSE(HYZ_i^{kX}) = MSE(HYZ_i^X)$ ، $i = 1, 2$

۶۸ آزمون نیکویی برازش توزیع نمایی بر مبنای برآوردهای آنتروپی

برهان: به سادگی می توان نشان داد $b_i^{**}(kX_i) = \frac{b_i^{**}(X_i)}{k}$. بنابراین با توجه به $\sum_{i=1}^n w_i = 1$ نتیجه می شود که $HY Z_i^k X = HY Z_i^X + \log(k)$, $i = 1, 2$ و برهان قضیه کامل است.

در ادامه با مطالعه شبیه سازی، عملکرد برآوردهای پیشنهادی اصلاح شده با برآوردهای واسیچک (۱۹۷۶)، ابراهیمی و همکاران (۱۹۹۴)، کوریا (۱۹۹۵)، یوسفزاده و ارقامی (۲۰۰۸) و زمانزاده و ارقامی (۱۳۸۷) برحسب جذر میانگین توان های دوم خطا مورد مقایسه قرار می گیرد.

شایان ذکر است، انتخاب مقدار بهینه m ، که به ازای آن میانگین توان های دوم مینیمم شود، علاوه بر حجم نمونه (n) ، به توزیع جامعه نیز بستگی دارد که در عمل نامعلوم است. ویزورکوسکی و گورزسکی (۱۹۹۹)، مقدار $m = [\sqrt{n} + 0.5]$ را برای برآورد آنتروپی پیشنهاد دادند که در آن $[\cdot]$ نماد جز صحیح است.

برای مقایسه برآوردهای مختلف آنتروپی، به ازای $m = 10, 20, 30$ تعداد ۱۰۰۰۰ نمونه از توزیع های نرمال استاندارد، نمایی با میانگین ۱ و یکنواخت (۱، ۰) تولید و مقدار جذر میانگین توان های دوم خطای برآوردها به ازای $m = [\sqrt{n} + 0.5]$ محاسبه شده و نتایج شبیه سازی در جدول ۱ ارائه گردیده اند. در هر ستون، اعداد ستاره دار نشان دهنده آن است که برآوردها مزبور، دارای کمترین مقدار جذر میانگین توان های دوم خطا است.

همان طور که ملاحظه شود، برآوردهای پیشنهادی به طور یکنواخت عملکرد بهتری نسبت به برآوردهای واسیچک (۱۹۷۶)؛ ابراهیمی و همکاران (۱۹۹۴) و کوریا (۱۹۹۵) در هر سه توزیع و حجم نمونه در نظر گرفته شده دارند. در توزیع نرمال، برآوردهای پیشنهادی عملکرد بهتری نسبت به سایر برآوردها دارند. در توزیع نمایی و یکنواخت، برآوردها اول پیشنهادی و برآوردها یوسفزاده و ارقامی (۲۰۰۸) عملکرد نسبتاً مشابهی دارند. علاوه بر این به طور کلی، در حجم نمونه ۱۰، برآوردها اول پیشنهادی و برآوردها یوسفزاده و ارقامی (۲۰۰۸) بهترین هستند، در حالی که در حجم های نمونه ۲۰ و ۳۰ برآوردها اول سایر برآوردها نسبتاً بهتر است. همچنین از مقایسه برآوردها یوسفزاده و ارقامی (۲۰۰۸) با برآوردهای واسیچک (۱۹۷۶) و ابراهیمی و همکاران (۱۹۹۴) ملاحظه می شود که برآوردها

جدول ۱: جذر میانگین توان‌های دوم خطای برآوردگرهای مختلف آن‌تروپی

HYZ_2	HYZ_1	HCZ_2	HCZ_1	HY	TC	HE	HV	توزیع	n
۰/۲۸۸*	۰/۲۹۴	۰/۳۵۲	۰/۳۴۵	۰/۳۱۹	۰/۴۶۳	۰/۴۰۶	۰/۶۲۰	$N(0, 1)$	۱۰
۰/۳۶۰	۰/۳۶۶	۰/۳۷۸	۰/۳۸۳	۰/۳۵۷*	۰/۴۴۳	۰/۴۶۶	۰/۵۶۶	$Exp(1)$	
۰/۱۷۶	۰/۱۶۸	۰/۱۹۰	۰/۱۹۰	۰/۱۶۶*	۰/۲۹۹	۰/۴۳۲	۰/۴۵۰	$U(0, 1)$	
۰/۲۷۶*	۰/۲۷۶*	۰/۳۰۶	۰/۳۰۶	۰/۲۸۰	۰/۴۰۱	۰/۴۳۴	۰/۵۴۵	میانگین	
۰/۱۹۴	۰/۱۸۸*	۰/۲۰۳	۰/۲۶۴	۰/۲۱۷	۰/۲۶۴	۰/۲۴۹	۰/۳۶۷	$N(0, 1)$	۲۰
۰/۲۳۶*	۰/۲۴۴	۰/۲۴۵	۰/۲۴۷	۰/۲۴۲	۰/۲۷۱	۰/۲۵۲	۰/۳۵۲	$Exp(1)$	
۰/۰۹۸	۰/۰۹۱*	۰/۰۹۲	۰/۰۹۳	۰/۰۹۱*	۰/۱۵۵	۰/۱۳۳	۰/۲۷۳	$U(0, 1)$	
۰/۱۷۶	۰/۱۷۴*	۰/۱۸۰	۰/۲۰۱	۰/۱۸۳	۰/۲۳۰	۰/۲۱۱	۰/۳۳۳	میانگین	
۰/۱۵۷	۰/۱۴۶*	۰/۱۵۴	۰/۱۹۴	۰/۱۷۷	۰/۱۹۴	۰/۲۰۴	۰/۲۷۹	$N(0, 1)$	۳۰
۰/۱۹۵*	۰/۱۹۸	۰/۱۹۶	۰/۱۹۸	۰/۱۹۶	۰/۲۱۱	۰/۲۰۴	۰/۲۹۱	$Exp(1)$	
۰/۰۷۳	۰/۰۶۸	۰/۰۶۵	۰/۰۶۴*	۰/۰۶۶	۰/۱۱۳	۰/۰۹۶	۰/۲۱۰	$U(0, 1)$	
۰/۱۴۱	۰/۱۳۷*	۰/۱۳۸	۰/۱۵۲	۰/۱۴۶	۰/۱۷۲	۰/۱۶۸	۰/۲۶۰	میانگین	

یوسف‌زاده و ارقامی (۲۰۰۸) به‌طور یک‌سواخت عمل‌کرد بهتری نسبت به این برآوردگرها دارد که این مطلب را می‌توان به دلیل استفاده از برآوردگر بهتر از تابع توزیع، توجیه کرد.

۳ آزمون‌نمایی بودن جامعه

فرض کنید X_1, \dots, X_n نمونه‌ای تصادفی با تابع چگالی f باشند. آزمون فرضیه

$$H_0 : f = f_\theta$$

را در نظر بگیرید، که در آن f_θ تابع چگالی توزیع‌نمایی، با میانگین مثبت θ است. فاصله نامتقارن کولبک-لایبر f از f_θ عبارت است از

$$D(f, f_\theta) = \int_{-\infty}^{\infty} f(x) \log\left(\frac{f(x)}{f_\theta(x)}\right) dx,$$

که می‌توان آن را به‌سادگی به‌صورت

$$D(f, f_\theta) = -H(X) + \log(\theta) + \frac{1}{\theta} E_f(X), \quad (12)$$

۷۰.....آزمون نیکویی برازش توزیع نمایی بر مبنای برآوردهای آنتروپی

بازنویسی کرد، که مینیمم (۱۲) نسبت به θ ، به ازای $\theta = E_f(X)$ حاصل می شود، لذا داریم:

$$D_{Inf} = Inf_{\theta} D(f, f_{\theta}) = -H(X) + \log(E_f(X)) + 1,$$

از طرفی $D_{Inf} = Inf_{\theta} D(f, f_{\theta}) = 0$ ، اگر و تنها اگر فرضیه نمایی بودن جامعه درست باشد. بنابراین می توان آماره های آزمون نمایی بودن جامعه را به صورت

$$TYZ_1 = -HYZ_1 + \log(\bar{X}) + 1, \quad TYZ_2 = -HYZ_2 + \log(\bar{X}) + 1,$$

پیشنهاد کرد و فرضیه نمایی بودن جامعه را به ازای مقادیر بزرگ آماره های آزمون رد کرد. توزیع آماره های آزمون به روش تحلیلی قابل محاسبه نیستند، بنابراین برای به دست آوردن نقاط بحرانی آزمون ها از روش شبیه سازی مونت کارلو استفاده می شود. شایان ذکر است که با اندکی محاسبه می توان نشان داد که آماره های آزمون فوق نسبت به تبدیلات مقیاسی ناوردا هستند. بنابراین نقاط بحرانی آن ها به پارامتر مجهول بستگی ندارد. برای محاسبه نقاط بحرانی ابتدا از توزیع نمایی با میانگین یک، نمونه ای به حجم n تولید و مقدار آماره آزمون محاسبه می شود. این کار ۱۰۰۰۰ دفعه تکرار می شود. مقادیر بحرانی آزمون با استفاده از چندک $1 - \alpha$ ام توزیع تجربی آماره آزمون به دست می آید.

برای مقایسه توان آزمون های مختلف آماره های آزمون به صورت

$$TV = -HV + \log(\bar{X}) + 1, \quad TC = -HC + \log(\bar{X}) + 1,$$

$$TY = -HY + \log(\bar{X}) + 1, \quad TCZ_1 = -HCZ_1 + \log(\bar{X}) + 1,$$

$$TCZ_2 = -HCZ_2 + \log(\bar{X}) + 1,$$

محاسبه می شوند. با شبیه سازی مونت کارلو و تعداد ۱۰۰۰۰ دفعه تکرار، توان آزمون های مختلف نمایی بودن توزیع جامعه تحت ۹ توزیع جانشین مختلف به دست آورده می شوند. برای این کار ابتدا تعداد ۱۰۰۰۰ نمونه به حجم های $n = 10, 20, 40$ تحت هر یک از توزیع های جانشین تولید شده و توان هر یک از آزمون ها از طریق، نسبت تعداد دفعاتی که آماره آزمون از مقدار بحرانی آن بیشتر است به کل تعدد تکرارها، برآورد می شوند.

برای مقایسه توان‌ها، از توزیع‌های زیر به‌عنوان توزیع جانشین استفاده کرده‌ایم:
الف- توزیع وایبول با تابع چگالی:

$$f(x; \lambda, \beta) = \beta \lambda^\beta x^{\beta-1} \exp\{-(\lambda x)^\beta\}, \quad \beta > 0, \lambda > 0, x \geq 0;$$

ب- توزیع گاما با تابع چگالی:

$$f(x; \lambda, \beta) = \frac{\lambda^\beta x^{\beta-1} \exp\{-\lambda x\}}{\Gamma(\beta)}, \quad \beta > 0, \lambda > 0, x \geq 0;$$

ج- توزیع لگ‌نرمال با تابع چگالی:

$$f(x; \nu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\log(x) - \nu)^2\right\}, \quad -\infty < \nu < \infty, \sigma, x > 0;$$

به‌علاوه از توزیع‌نمایی مختلط برای مقایسه توان آزمون‌های مختلف استفاده شد اما توان کلیه آزمون‌ها در این خانواده بسیار کم حاصل شد.

توان آزمون‌های بر مبنای آنتروپی، به مقدار پارامتر m بستگی دارد، و مقدار بهینه m ، یعنی مقداری که به‌ازای آن توان آزمون ماکسیمم شود، علاوه بر حجم نمونه، به توزیع جانشین نیز وابسته است و چون توزیع جانشین در عمل نامعلوم است، نمی‌توان مقداری برای m پیشنهاد داد که به‌ازای آن، توان آزمون برای تمام توزیع‌های جانشینی ماکسیمم شود. به‌همین علت، مقادیر m برای n های مختلف پیشنهاد داده می‌شود که آزمون‌های پیشنهادی دارای توان نسبتاً خوبی در تمام توزیع‌های جانشین باشند. این مقادیر به این صورت به‌دست آمده‌اند که توان آزمون‌های پیشنهادی به‌ازای تمامی مقادیر $m \leq \frac{n}{4}$ و توزیع‌های جانشین شبیه‌سازی شده‌اند و مقادیری انتخاب شده‌اند که آزمون پیشنهادی دارای توان‌های خوبی برای تمامی توزیع‌های جانشین بوده است. از مقادیر m که در جدول ۲ ارائه شده‌اند برای آزمون‌های پیشنهادی و مقایسه آزمون‌های مختلف استفاده شده است. نتایج شبیه‌سازی مقایسه توان‌های آزمون‌های مختلف‌نمایی بودن توزیع جامعه در جدول ۲ ارائه شده‌اند. مقادیر ستاره‌دار در هر سطر نشانه آن است که آزمون مزبور، بیشترین توان را در مقایسه با سایر آزمون‌های آن سطر دارد.

۷۲ آزمون نیکویی برازش توزیع نمایی بر مبنای برآوردگرهای آنتروپی

ابراهیمی و همکاران (۱۹۹۴) نشان دادند که رابطه

$$HE = HV + \frac{2}{n} \left\{ m \log(m) - \log\left(\frac{(m-1)!}{(2m-1)!}\right) \right\}$$

بین برآوردگر پیشنهادی آنان با برآوردگر واسیچک (۱۹۷۶) برقرار است. لذا توان آزمون بر مبنای برآوردگر آنان با برآوردگر واسیچک (۱۹۷۶) برابر است.

جدول ۲: مقادیر پیشنهادی m برای آزمون نمایی بودن توزیع جامعه بر مبنای TYZ_1 و TYZ_2 ، به ازای مقادیر مختلف n

m	n
۲	$n \leq 8$
۳	$9 \leq n \leq 15$
۴	$16 \leq n \leq 35$
۵	$36 \leq n \leq 60$
۶	$61 \leq n \leq 100$
۷	$101 \leq n \leq 200$

همانطور که در جدول ۳ ملاحظه می شود. هیچ یک از هفت آزمون به ازای هر سه توزیع جانشین توان بیشتری نسبت به سایرین ندارند. اما آزمونهای مبتنی بر برآوردگر واسیچک (۱۹۷۶)، برآوردگر دوم زمانزاده و ارقامی (۱۳۸۷) و برآوردگر دوم پیشنهادی توانهای بهتری نسبت به سایر آزمونها دارد. در حجم نمونه ۱۰، در توزیعهای جانشین از خانواده گاما و وایبول TCZ_2 و TYZ_2 بهترین آماره های آزمون هستند و در حالی که در خانواده توزیع لگ نرمال، آماره آزمون واسیچک بهترین است. در حجمهای نمونه بزرگتر هنگامی که توزیع جانشین، از خانواده گاما و وایبول است، آزمون مبتنی بر TYZ_2 عملکرد بهتری نسبت به سایر آزمونها دارد در حالی که اگر توزیع جانشین متعلق به خانواده لگ نرمال باشد، آزمون مبتنی بر آماره TV بهترین است. به هر حال، در حجم نمونه کم آزمون مبتنی بر TCZ_2 و در حجمهای نمونه متوسط و بزرگ آزمون مبتنی بر TYZ_2 در بیشتر حالات عملکرد بهتری نسبت به سایر آزمونها دارد. بنابراین هیچ اطلاعاتی نسبت به توزیع جانشین نداشته باشیم، استفاده از این آزمونها در عمل توصیه می شود. نکته جالب توجه این است که با اینکه برآوردگر واسیچک (۱۹۷۶) عملکرد خوبی در برآورد آنتروپی در

جدول ۳: توان آزمون‌های نمایی بودن به‌ازای $\alpha = 0/05$

TYZ_2	TYZ_1	TY	TCZ_2	TCZ_1	TC	TV	توزیع جانشین	n
0/343	0/329	0/333	0/350*	0/319	0/332	0/325	$G(2, 2)$	
0/673*	0/646	0/657	0/669	0/616	0/636	0/634	$G(3, 3)$	
0/858	0/829	0/843	0/858	0/811	0/830	0/872*	$G(4, 4)$	
0/726	0/701	0/707	0/732*	0/682	0/702	0/690	$W(2, \Gamma(1 + \frac{1}{2}))$	10
0/990*	0/984	0/985	0/987	0/977	0/983	0/982	$W(2, \Gamma(1 + \frac{1}{3}))$	
0/999*	0/999*	0/999*	0/999*	0/999*	0/999*	0/999*	$W(2, \Gamma(1 + \frac{1}{4}))$	
0/136	0/144	0/175	0/133	0/137	0/141	0/204*	$LN(-2, 2)$	
0/447	0/492	0/516	0/517	0/499	0/501	0/591*	$LN(-\frac{2}{3}, 3)$	
0/714	0/758	0/767	0/717	0/757	0/757	0/815*	$LN(-8, 4)$	
0/618*	0/518	0/583	0/578	0/477	0/497	0/504	$G(2, 2)$	
0/945*	0/891	0/935	0/928	0/858	0/877	0/890	$G(3, 3)$	
0/995*	0/983	0/993	0/991	0/970	0/977	0/983	$G(4, 4)$	
0/964*	0/932	0/956	0/961	0/918	0/929	0/932	$W(2, \Gamma(1 + \frac{1}{2}))$	20
1/000*	1/000*	1/000*	1/000*	0/999	0/999	0/999	$W(2, \Gamma(1 + \frac{1}{3}))$	
1/000*	1/000*	1/000*	1/000*	1/000*	1/000*	1/000*	$W(2, \Gamma(1 + \frac{1}{4}))$	
0/612	0/580	0/665	0/612	0/578	0/583	0/654*	$LN(-2, 2)$	
0/949	0/950	0/965	0/948	0/949	0/948	0/967*	$LN(-\frac{2}{3}, 3)$	
0/994	0/995	0/997	0/993	0/995	0/995	0/999*	$LN(-8, 4)$	
0/849*	0/754	0/844	0/844	0/746	0/731	0/755	$G(2, 2)$	
0/998*	0/991	0/998*	0/997	0/990	0/988	0/991	$G(3, 3)$	
1/000*	1/000*	1/000*	1/000*	1/000*	0/999	1/000*	$G(4, 4)$	
0/999*	0/998	0/999*	0/999*	0/997	0/997	0/997	$W(2, \Gamma(1 + \frac{1}{2}))$	30
1/000*	1/000*	1/000*	1/000*	1/000*	1/000*	1/000*	$W(2, \Gamma(1 + \frac{1}{3}))$	
1/000*	1/000*	1/000*	1/000*	1/000*	1/000*	1/000*	$W(2, \Gamma(1 + \frac{1}{4}))$	
0/964	0/953	0/971*	0/962	0/952	0/953	0/967	$LN(-2, 2)$	
0/999	0/999	1/000*	0/999	0/999	0/999	1/000*	$LN(-\frac{2}{3}, 3)$	
1/000*	1/000*	1/000*	1/000*	1/000*	1/000*	1/000*	$LN(-8, 4)$	

مقایسه با سایر برآوردگرهای آنتروپی ندارد اما آزمون نیکویی برازش توزیع بر مبنای این برآوردگر آنتروپی توان‌های نسبتاً خوبی دارد و در برخی از موارد (توزیع لگ‌نرمال) بهترین است.

بحث و نتیجه‌گیری

دو برآوردگر جدید آنتروپی بر مبنای اصلاح برآوردگر یوسف‌زاده و ارقامی (۲۰۰۸) پیشنهاد گردید. در مطالعه شبیه‌سازی مقایسه عملکرد برآوردگرها نشان داد که برآوردگرهای پیشنهادی به‌طور یکنواخت از برآوردگرهای واسیچک (۱۹۷۶)،

۷۴ آزمون نیکویی برازش توزیع نمایی بر مبنای برآوردگرهای آنتروپی

ابراهیمی و همکاران (۱۹۹۴) و کوریا (۱۹۹۵) بهتر هستند و همچنین عملکرد خوبی نسبت به سایر برآوردگرهای آنتروپی دارند. مقایسه توان آزمون‌های نمایی بودن جامعه بیانگر آن است که آزمون‌های برمبنای برآوردگر واسیچک (۱۹۷۶)، برآوردگر دوم زمان‌زاده و ارقامی (۱۳۸۷) و برآوردگر دوم پیشنهادی در توزیع‌های مختلف جانشین بهترین آزمون‌ها هستند، هرچند به‌طور کلی بر حجم نمونه کوچک آزمون مبینی بر TCZ_2 و در حجم‌های نمونه متوسط و بزرگ آزمون مبینی بر TYZ_2 بهتر از سایر آماره‌های آزمون است.

لازم به ذکر است که انتخاب مقدار بهینه پارامتر m در برآورد آنتروپی و آزمون نیکویی برازش مرتبط با آن و بررسی رفتار مجانبی برآوردگر کوریا و اصلاحات آن نیازمند مطالعات جدید است.

تقدیر و تشکر

نویسندگان از پیشنهادات داوران محترم که باعث اصلاحات سازنده در این مقاله شده است، کمال تشکر و سپاسگزاری را دارند.

مراجع

حبیبی‌راد، آ.، ارقامی، ن. ر. (۱۳۸۶)، آزمون متقارن بودن توزیع براساس آنتروپی، مجله علوم آماری، جلد ۱، ۱۰۹-۱۲۰.

زمان‌زاده، ا.، ارقامی، ن. ر. (۱۳۸۷)، آزمون نیکویی برازش توزیع‌های نرمال و نمایی برمبنای برآوردگرهای جدید آنتروپی، مجله علوم آماری، جلد ۲، ۱۷۹-۲۰۰.

علیزاده نوقابی، ه.، علیزاده نوقابی، ر. (۱۳۸۷)، مقایسه توان آزمون‌های نیکویی برازش برمبنای آنتروپی با سایر روش‌ها، مجله علوم آماری، جلد ۲، ۹۷-۱۱۳.

Ebrahimi, N., Habibullah, M. and Soofi, E. (1994), Two Measures of Sample Entropy, *Statistics and Probability Letters*, **20**, 225-234.

- Correa, J. C. (1995), A New Estimator of Entropy, *Communications in Statistics Theory and Methods*, **24**, 2439-2449.
- Kullback, S. and Leibler, R. A. (1951), On Information and Sufficiency, *The Annals of Mathematical Statistics*, **22**, 79-86.
- Shanon, C. E. (1948), A Mathematical Theory of Communicatios, *Bell System Thechnical Journal*, **27**, 379-423; 623-656.
- Shapiro, S. S. and Wild, M. B. (1965), An Analysis of Variance Test for Normality (Complete Sample), *Biometrika*, **52**, 591-611.
- Vasicek, O. (1976), A Test fo Normality Based on Sample Entropy, *Journal of Royal Statistical Society, B*, **38**, 54-59.
- Wieczorkowski, P. and Grzegorzewsky, P. (1999), Entropy Estimator Improvements and Comparisons, *Communications in Statistics-Computation and Simulation*, **28**, 541-567.
- Yousefzadeh, F. and Arghami, N. R. (2008), Testing Exponentiality Based on Type II Censored Data and a New cdf Estimator, *Communications in Statistics-Simulation and Computation*, **37**, 1479-1499.