مجله علوم آماری، بهار و تابستان ۱۳۹۲
جلد ۷، شماره ۱، ص ۶۱-۷۵

آزمون نیکویی برازش توزیع نمایی برمبنای 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 می‌شود. سپس آزمون نیکویی برازش فرضیه نمایی بودن توزیع جامعه برمبنای برآوردگرهای جدید معرفی می‌شود و توان آن‌ها با توان سایر آزمون‌های برمبنای آنتروپی توزیع نمایی مورد مقایسه قرار می‌گیرد. نتایج مطالعات شبیه‌سازی نشان می‌دهد که برآوردگرهای پیشنهادی عموماً عملکرد بهتری در مقایسه با سایر برآوردگرها در برآورد آنتروپی و آزمون نیکویی برازش دارند.

واژه‌های کلیدی : آ نتروپی، 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 به‌صورت

$$H(f) = -\int_{-\infty}^{\infty} \log(f(x))f(x)dx \qquad (۱)$$

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

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 برمبنای برآوردگر جدیدی از آنتروپی پیشنهاد داد. نتایج شبیه‌سازی واسیچک نشان می‌داد که آزمون برمبنای آنتروپی، علی رغم سادگی در محاسبه دارای توان‌های بسیار خوبی در مقایسه با آزمون پرتوان شاپیرو–ویلک (۱۹۶۵) برای توزیع نرمال است. از آن به بعد محققین زیادی سعی در بهبود برآوردگر و آزمون پیشنهاد شده توسط واسیچک (۱۹۷۶) و تعمیم آن به سایر توزیع‌های آماری داشته‌اند.

واسیچک (۱۹۷۶) نشان داد که می‌توان رابطه (۱) را به‌صورت

$$H(f) = -\int_{0}^{1} \log(\frac{d}{dp}F^{-1}(p))dp$$

بازنویسی کرد و با جایگزینی تابع توزیع F با توزیع تجربی F_n و استفاده از عملگر تفاضل به جای عملگر مشتق، برآوردگر آنتروپی خود را به‌صورت زیر معرفی کرد.

فرض کنید $X_1, \ldots, X_n$ نمونه‌ای تصادفی از توزیع پیوسته F باشد، در این صورت برآوردگر آنتروپی واسیچک (۱۹۷۶) عبارتست از

$$HV = \frac{1}{n}\sum_{i=1}^{n} \log(\frac{n}{c_i}(X_{(i+m)} - X_{(i-m)})) \qquad (۲)$$

که در آن $X_{(1)}, \ldots, X_{(n)}$ آماره‌های مرتب، $c_i = 2m$ و m یک عدد مثبت کوچکتر یا مساوی با $\frac{n}{2}$ است. ضمناً به‌ازای $i < 1$، $X_{(i)} = X_{(1)}$ و به‌ازای $i > n$، $X_{(i)} = X_{(n)}$ می‌باشد.

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) برآوردگر پیشنهادی خود را برمبنای اصلاح ضرایب c_i، هنگامی که $i \leq m$ یا $i \geq n - m$ است، به‌صورت

$$HE = \frac{1}{n}\sum_{i=1}^{n} \log(\frac{n}{c_i}(X_{(i+m)} - X_{(i-m)})) \qquad (۳)$$

ارائه کردند، که در آن

$$c_i = \begin{cases} m+i-۱, & ۱ \le i \le m \\ ۲m, & m+۱ \le i \le n-m \\ m+n-i, & n-m+۱ \le i \le n \end{cases}$$

به‌ازای $i < ۱$، $X_{(i)} = X_{(۱)}$ و به‌ازای $i > n$، $X_{(i)} = X_{(n)}$ است.

نتایج شبیه‌سازی ابراهیمی و همکاران (۱۹۹۴) برای سه توزیع نرمال، نمایی و یکنواخت نشان داد که برآوردگر پیشنهادی آن‌ها اریبی و جذر میانگین مربع خطای کمتری نسبت به برآورد پیشنهادی واسیچک (۱۹۷۶) به‌ازای حجم نمونه ۱۰، ۲۰ و ۳۰ 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 به‌صورت

$$HV_{nm} = -\frac{۱}{n}\sum_{i=۱}^{n}\log\left(\frac{\frac{i+m}{n} - \frac{i-m}{n}}{X_{(i+m)} - X_{(i-m)}}\right) \qquad (۴)$$

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 $\{X_{(i-m)}, \ldots, X_{(i+m)}\}$ و با استفاده از تمام $۲m+۱$ نقطه، به‌جای 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 به‌صورت

$$HC_{nm} = -\frac{۱}{n}\sum_{i=۱}^{n}\log(b_i) \qquad (۵)$$

ارائه کرد، که در آن

$$b_i = \frac{\sum_{j=i-m}^{i+m}(X_{(j)} - \bar{X}_{(i)})(\frac{j-i}{n})}{\sum_{j=i-m}^{i+m}(X_{(j)} - \bar{X}_{(i)})^{۲}} \qquad (۶)$$

و $\bar{X}_{(i)} = \sum_{j=i-m}^{i+m} \frac{X_{(j)}}{۲m+۱}$ و به‌ازای $i < ۱$، $X_{(i)} = X_{(۱)}$ و به‌ازای $i > n$، $X_{(i)} = X_{(n)}$ می‌باشد.

مطالعه شبیه‌سازی برآوردگرهای کوریا (۱۹۹۵) و واسیچک (۱۹۷۶) نشان داد که این برآوردگر عموماً از میانگین توان دوم خطای کمتری نسبت به برآودگر واسیچک (۱۹۷۶) در سه توزیع نرمال، نمایی و یکنواخت برخوردار است. کوریا (۱۹۹۵) هیچ مقایسه‌ای میان برآوردگر پیشنهادی خود، با برآوردگر پیشنهادی ابراهیمی و همکاران (۱۹۹۴) انجام نداده است.

یوسف‌زاده و ارقامی (۲۰۰۸) برآوردگری از آنتروپی برمبنای برآوردگر جدیدی از تابع توزیع جامعه ارائه دادند. آنها ابتدا برآوردگری از تابع توزیع پیوسته به‌صورت

$$\hat{F}_y(t) = \begin{cases} \frac{n-۱}{n(n+۱)}(\frac{n}{n-۱} + \frac{t-X_{.}}{X_{(۲)}-X_{.}} + \frac{t-X_{(۱)}}{X_{(۳)}-X_{(۱)}}), & X_{(۱)} \le t \le X_{(۲)} \\ \frac{n-۱}{n(n+۱)}(i + \frac{۱}{n-۱} + \frac{t-X_{(i-۱)}}{X_{(i+۱)}-X_{(i-۱)}} + \frac{t-X_{(i)}}{X_{(i+۲)}-X_{(i)}}) & X_{(i)} \le t \le X_{(i+۱)} \\ & i = ۱, \ldots, n-۲ \\ \frac{n-۱}{n(n+۱)}(n-۱ + \frac{۱}{n-۱} + \frac{t-X_{(n-۲)}}{X_{(n)}-X_{(n-۲)}} + \frac{t-X_{(n-۱)}}{X_{(n+۱)}-X_{(n-۱)}}) & X_{(n-۱)} \le t \le X_{(n)} \end{cases}$$

معرفی کردند، که در آن $X_{.} = X_{(۱)} - \frac{n}{n-۱}(X_{(۲)} - X_{(۱)})$ و $X_{(n+۱)} = X_{(n)} + \frac{n}{n-۱}(X_{(n)} - X_{(n-۱)})$ است. آنگاه برآوردگر

$$HY = \sum_{i=۱}^{n} \log(\frac{X_{(i+m)} - X_{(i-m)}}{\hat{F}_y(X_{(i+m)}) - \hat{F}_y(X_{(i-m)})}) \frac{\hat{F}_y(X_{(i+m)}) - \hat{F}_y(X_{(i-m)})}{\sum_{i=۱}^{n}(\hat{F}_y(X_{(i+m)}) - \hat{F}_y(X_{(i-m)}))} \quad (۷)$$

را برای برآورد آنتروپی یک متغیر تصادفی پیوسته پیشنهاد دادند، که به‌ازای $i < ۱$، $X_{(i)} = X_{(۱)}$ و به‌ازای $i > n$، $X_{(i)} = X_{(n)}$ می‌باشد. سپس در مطالعه شبیه‌سازی نشان دادند برآوردگر پیشنهادی آنان، دارای اریبی و میانگین توان دوم خطای کمتری نسبت به برآوردگرهای واسیچک (۱۹۷۶) و ابراهیمی و همکاران (۲۰۰۴) در سه توزیع نرمال، نمایی و یکنواخت است.

زمان‌زاده و ارقامی (۱۳۸۷)، دو اصلاح مختلف از برآوردگر کوریا (۱۹۹۵) پیشنهاد دادند: اصلاح اول برمبنای این ایده بود که در محاسبه مقدار b_i در رابطه (۶)، هنگامی که $i \le m$ یا $i \ge n - m + ۱$، است تقریب مناسبی برای شیب خط نیست زیرا در این نقاط بیش از یک بار از $X_{(۱)}$ و $X_{(n)}$ استفاده شده است. لذا آنان

اصلاح

$$HCZ_۱ = -\frac{۱}{n}\sum_{i=۱}^{n}\log(b_i^*), \qquad (۸)$$

را پیشنهاد دادند، که در آن

$$b_i^* = \frac{\sum_{j=K_۱(i)}^{K_۲(i)}(X_{(j)} - \tilde{X}_{(i)})(\hat{F}_n(X_{(j)}) - \tilde{\hat{F}}_n(j))}{\sum_{j=K_۱(i)}^{K_۲(i)}(X_{(j)} - \tilde{X}_{(i)})^۲}, \quad i = ۱,\ldots,n$$

$$k_۱(i) = \begin{cases} ۱, & ۱ \le i \le m \\ i-m, & i \ge m+۱ \end{cases}, \quad k_۲(i) = \begin{cases} i+m, & ۱ \le i \le n-m \\ n, & i \ge n-m+۱ \end{cases}$$

$\tilde{X}_{(i)} = \sum_{j=k_۱(i)}^{k_۲(i)} \frac{X_{(j)}}{k_۲(i)-k_۱(i)+۱}$ و $\tilde{\hat{F}}_n(i) = \sum_{j=۱k_۱(i)}^{k_۲(i)} \frac{\hat{F}_n(X_{(j)})}{k_۲(i)-k_۱(i)+۱}$ برآوردگر $\hat{F}_n$ تابع توزیع تجربی است.

برای دومین اصلاح چون در به‌دست آوردن b_i، از تعداد مساوی مشاهده استفاده شده است، همه b_iها از وزن‌های مساوی برخوردارند. اما در محاسبه $HCZ_۱$ وقتی که $i \le m$ و یا $i \ge n-m+۱$، از مشاهدات کمتری برای به‌دست آوردن b_iها استفاده شده است، لذا منطقی به‌نظر می‌رسد که وزن‌های کمتری به این مشاهدات اختصاص داده شود. برمبنای این ایده، اصلاح دوم به‌صورت

$$HCZ۲ = -\sum_{i=۱}^{n} w_i \log(b_i^*) \qquad (۹)$$

ارائه شد، که در آن $w_i = \frac{\hat{F}(X_{(i+m)}) - \hat{F}_n(X_{(i-m)})}{\sum_{i=۱}^{n}(\hat{F}_n(X_{(i+m)}) - \hat{F}_n(X_{(i-m)}))}$ است. مطالعه شبیه‌سازی زمان‌زاده و ارقامی (۱۳۸۷) نشان داد که این برآوردگرها عملکرد خوبی در مقایسه با برآوردگرهای آنتروپی واسیچک (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 حبیبی‌راد و ارقامی (۱۳۸۶) آزمون تقارن توزیع را برمبنای آنتروپی پیشنهاد دادند.

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

[۱] Discrimination information function

مهم برای مقایسه دو توزیع است. به این تابع که براساس آنتروپی نیز می‌توان آن را بازنویسی کرد، آنتروپی متقاطع[۲] هم گفته می‌شود. فرض کنید برمبنای مشاهده $X = x$، یکی از دو توزیع F یا G را به‌عنوان توزیع متغیر تصادفی X، با تکیه‌گاه S مورد نظر باشد. در این صورت با توجه به قضیه بیز می‌توان نوشت:

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

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

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

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

همچنین می‌توان نشان داد که $D(f, g) \geq 0$، و تساوی رخ می‌دهد اگر و تنها اگر $f \overset{a.s}{=} g$.

در بخش ۲ برآوردگرهای جدید آنتروپی به‌عنوان اصلاح برآوردگر یوسف‌زاده و ارقامی (۲۰۰۸) معرفی می‌شود. سپس در مطالعه‌ای شبیه‌سازی برآوردگرهای پیشنهادی با سایر برآوردگرهای آنتروپی مقایسه خواهند شد. در بخش ۳ آزمون نیکویی برازش برای فرضیه‌های نمایی بودن توزیع جامعه بر اساس برآوردگرهای جدید آنتروپی معرفی می‌شود و توان این آزمون‌ها با سایر آزمون‌های نمایی بودن توزیع جامعه برمبنای برآوردگرهای آنتروپی واسیچک (۱۹۷۶) ، کوریا (۱۹۹۵)، یوسف‌زاده و ارقامی (۲۰۰۸) و زمان‌زاده و ارقامی (۱۳۸۷) مقایسه خواهد شد. بحث و نتیجه‌گیری 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 $\frac{X_{(i+m)}-X_{(i-m)}}{\hat{F}_y(X_{(i+m)})-\hat{F}_y(X_{(i-m)})}$ در رابطه (۷) شیب خطی است که این نقاط را به یکدیگر متصل می‌کند. اما همان‌طور که در برآوردگر کوریا (۱۹۹۵)، استفاده از تمامی نقاط $\{X_{(i-m)}, \ldots, X_{(i+m)}\}$ به‌جای 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 برمبنای ایده زمان‌زاده و ارقامی (۱۳۸۷) برمبنای اصلاح برآوردگر کوریا (۱۹۹۵)، برآوردگرهای جدید به‌صورت

$$HYZ_1 = -\frac{1}{n}\sum_{i=1}^{n}\log(b_i^{**}), \qquad (۱۰)$$

$$HYZ_2 = -\sum_{i=1}^{n} w_i \log(b_i^{**}), \qquad (۱۱)$$

ارئه می‌شوند، که در آن w_i وزن‌های به‌کار رفته در $HCZ2$، b_i^{**} ضرایب معرفی شده در (۹) است با این تفاوت که از $\hat{F}_y$ معرفی شده در رابطه (۷) به‌جای برآوردگر تابع توزیع تجربی استفاده شده است.

قضیه ۱ : فرض کنید $X_1, \ldots, X_n$ نمونه‌ای تصادفی از جامعه‌ای با آنتروپی $H^X(f)$ باشد و $Y_i = kX_i$، که در آن $k > 0$ و $i = 1, \ldots, n$. همچنین فرض کنید که HYZ_i^X و HYZ_i^{kX} $(i = 1, 2)$ به ترتیب برآوردهای آنتروپی $H^X(f)$ و $H^{kX}(g)$ 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- $\mathrm{MSE}(HYZ_i^{kX}) = \mathrm{MSE}(HYZ_i^X)$، $i = 1, 2$.

برهان : به‌سادگی می‌توان نشان داد $b_i^{**}(kX_i) = \frac{b_i^{**}(X_i)}{k}$. بنابراین با توجه به $\sum_{i=1}^{n} w_i = 1$ نتیجه می‌شود که $HYZ_i^{kX} = HYZ_i^X + \log(k),\ i = 1, 2$ و برهان قضیه کامل است.

در ادامه با مطالعه شبیه‌سازی، عملکرد برآوردهای پیشنهادی اصلاح شده با برآوردگرهای واسیچک (۱۹۷۶)، ابراهیمی و همکاران (۱۹۹۴)، کوریا (۱۹۹۵)، یوسف‌زاده و ارقامی (۲۰۰۸) و زمان‌زاده و ارقامی (۱۳۸۷) برحسب جذر میانگین توان‌های دوم خطا مورد مقایسه قرار می‌گیرد.

شایان ذکر است، انتخاب مقدار بهینه m، که به‌ازای آن میانگین توان‌های 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 برآوردگرهای مختلف آنتروپی، به‌ازای $n = 10, 20, 30$؛ تعداد ۱۰۰۰۰ نمونه از توزیع‌های نرمال استاندارد، نمایی با میانگین ۱ و یکنواخت $(0, 1)$ تولید و مقدار جذر میانگین توان‌های دوم خطای برآوردگرها به‌ازای $m = [\sqrt{n} + 0/5]$، محاسبه شده و نتایج شبیه‌سازی در جدول ۱ ارائه گردیده‌اند. در هر ستون، اعداد ستاره‌دار نشان دهنده آن است که برآوردگر مزبور، دارای کمترین مقدار جذر میانگین توان‌های دوم خطا است.

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

جدول ۱: جذر میانگین توان‌های دوم خطای برآوردگرهای مختلف آنتروپی

$HYZ_۲$	$HYZ_۱$	$HCZ_۲$	$HCZ_۱$	HY	TC	HE	HV	توزیع	n
۰/۲۸۸*	۰/۲۹۴	۰/۳۵۲	۰/۳۴۵	۰/۳۱۹	۰/۴۶۳	۰/۴۰۶	۰/۶۲۰	$N(۰,۱)$	
۰/۳۶۰	۰/۳۶۶	۰/۳۷۸	۰/۳۸۳	۰/۳۵۷*	۰/۴۴۳	۰/۴۶۶	۰/۵۶۶	$Exp(۱)$	
۰/۱۷۶	۰/۱۶۸	۰/۱۹۰	۰/۱۹۰	۰/۱۶۶*	۰/۲۹۹	۰/۴۳۲	۰/۴۵۰	$U(۰,۱)$	۱۰
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	$N(۰,۱)$	
۰/۲۳۶*	۰/۲۴۴	۰/۲۴۵	۰/۲۴۷	۰/۲۴۲	۰/۲۷۱	۰/۲۵۲	۰/۳۵۲	$Exp(۱)$	
۰/۰۹۸	۰/۰۹۱*	۰/۰۹۲	۰/۰۹۳	۰/۰۹۱*	۰/۱۵۵	۰/۱۳۳	۰/۲۷۳	$U(۰,۱)$	۲۰
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	$N(۰,۱)$	
۰/۱۹۵*	۰/۱۹۸	۰/۱۹۶	۰/۱۹۸	۰/۱۹۶	۰/۲۱۱	۰/۲۰۴	۰/۲۹۱	$Exp(۱)$	
۰/۰۷۳	۰/۰۶۸	۰/۰۶۵	۰/۰۶۴*	۰/۰۶۶	۰/۱۱۳	۰/۰۹۶	۰/۲۱۰	$U(۰,۱)$	۳۰
۰/۱۴۱	۰/۱۳۷*	۰/۱۳۸	۰/۱۵۲	۰/۱۴۶	۰/۱۷۲	۰/۱۶۸	۰/۲۶۰	میانگین	

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

۳ آزمون نمایی بودن جامعه

فرض کنید $X_۱,\ldots,X_n$ نمونه‌ای تصادفی با تابع چگالی f باشند. آزمون فرضیه

$$H_۰ : f = f_\theta$$

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

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

که می‌توان آن را به‌سادگی به‌صورت

$$D(f,f_\theta)=-H(X)+\log(\theta)+\frac{۱}{\theta}E_f(X), \qquad (۱۲)$$

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

$$D_{Inf} = Inf_\theta D(f, f_\theta) = -H(X) + \log(E_f(X)) + ۱,$$

از طرفی $D_{Inf} = Inf_\theta D(f, f_\theta) = ۰$، اگر و تنها اگر فرضیه نمایی بودن جامعه درست باشد. بنابراین می‌توان آماره‌های آزمون نمایی بودن جامعه را به‌صورت

$$TYZ_۱ = -HYZ_۱ + \log(\bar{X}) + ۱, \quad TYZ_۲ = -HYZ_۲ + \log(\bar{X}) + ۱,$$

پیشنهاد کرد و فرضیه نمایی بودن جامعه را به‌ازای مقادیر بزرگ آماره‌های آزمون رد کرد. توزیع آماره‌های آزمون به روش تحلیلی قابل محاسبه نیستند، بنابراین برای به‌دست آوردن نقاط بحرانی آزمون‌ها از روش شبیه‌سازی مونت‌کارلو استفاده می‌شود. شایان ذکر است که با اندکی محاسبه می‌توان نشان داد که آماره‌های آزمون فوق نسبت به تبدیلات مقیاسی ناوردا هستند. بنابراین نقاط بحرانی آن‌ها به پارامتر مجهول بستگی ندارد. برای محاسبه نقاط بحرانی ابتدا از توزیع نمایی با میانگین یک، نمونه‌ای به حجم n تولید و مقدار آماره آزمون محاسبه می‌شود. این کار ۱۰۰۰۰ دفعه تکرار می‌شود. مقادیر بحرانی آزمون با استفاده از چندک $\alpha - ۱$ام توزیع تجربی آماره آزمون به‌دست می‌آید.

برای مقایسه توان آزمون‌های مختلف آماره‌های آزمون به‌صورت

$$TV = -HV + \log(\bar{X}) + ۱, \quad TC = -HC + \log(\bar{X}) + ۱,$$
$$TY = -HY + \log(\bar{X}) + ۱, \quad TCZ_۱ = -HCZ_۱ + \log(\bar{X}) + ۱,$$
$$TCZ_۲ = -HCZ_۲ + \log(\bar{X}) + ۱,$$

محاسبه می‌شوند. با شبیه‌سازی مونت‌کارلو و تعداد ۱۰۰۰۰ دفعه تکرار، توان آزمون‌های مختلف نمایی بودن توزیع جامعه تحت ۹ توزیع جانشین مختلف به‌دست آورده می‌شوند. برای این کار ابتدا تعداد ۱۰۰۰۰ نمونه به حجم‌های $n = ۱۰, ۲۰, ۴۰$ تحت هر یک از توزیع‌های جانشین تولید شده و توان هر یک از آزمون‌ها از طریق، نسبت تعداد دفعاتی که آماره آزمون از مقدار بحرانی آن بیشتر است به کل تعداد تکرارها، برآورد می‌شوند.

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:

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

ج- توزیع لگ‌نرمال با تابع چگالی:

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

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

توان آزمون‌های برمبنای 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 nهای مختلف پیشنهاد داده می‌شود که آزمون‌های پشنهادی دارای توان نسبتاً خوبی در تمام توزیع‌های جانشین باشند. این مقادیر به این صورت به‌دست آمده‌اند که توان آزمون‌های پیشنهادی به‌ازای تمامی مقادیر $m\leq\frac{n}{2}$ 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 مختلف نمایی 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.

ابراهیمی و همکاران (۱۹۹۴) نشان دادند که رابطه

$$HE = HV + \frac{۲}{n}\{m\log(m) - \log(\frac{(m-۱)!}{(۲m-۱)!})\}$$

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

جدول ۲: مقادیر پیشنهادی m برای آزمون نمایی بودن توزیع جامعه برمبنای $TYZ_۱$ و $TYZ_۲$، به‌ازای مقادیر مختلف n

m	n
۲	$n \leq ۸$
۳	$۹ \leq n \leq ۱۵$
۴	$۱۶ \leq n \leq ۳۵$
۵	$۳۶ \leq n \leq ۶۰$
۶	$۶۱ \leq n \leq ۱۰۰$
۷	$۱۰۱ \leq n \leq ۲۰۰$

همان‌طور که در جدول ۳ ملاحظه می‌شود. هیچ یک از هفت آزمون به‌ازای 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 $TCZ_۲$ و $TYZ_۲$ بهترین آماره‌های آزمون هستند و درحالی که در خانواده توزیع لگ‌نرمال، آماره آزمون واسیچک بهترین است. در حجم‌های نمونه بزرگ‌تر هنگامی که توزیع جانشین، از خانواده گاما و وایبول است، آزمون مبتنی بر $TYZ_۲$ عملکرد بهتری نسبت به سایر آزمون‌ها دارد در حالی که اگر توزیع جانشین متعلق به خانواده لگ‌نرمال باشد، آزمون مبتنی بر آماره TV بهترین است. به‌هر حال، در حجم نمونه کم آزمون مبتنی بر $TCZ_۲$ و در حجم‌های نمونه متوسط و بزرگ آزمون مبتنی بر $TYZ_۲$ در بیشتر حالات عملکر بهتری نسبت به سایر آزمون‌ها دارد. بنابراین هیچ اطلاعاتی نسبت به توزیع جانشین نداشته باشیم، استفاده از این آزمون‌ها در عمل توصیه می‌شود. نکته جالب توجه این است که با اینکه برآوردگر واسیچک (۱۹۷۶) عملکرد خوبی در برآورد آنتروپی در

جدول ۳: توان آزمون‌های نمایی بودن به‌ازای $\alpha = ۰/۰۵$

n	توزیع جانشین	TV	TC	$TCZ_۱$	$TCZ_۲$	TY	$TYZ_۱$	$TYZ_۲$
	$G(۲,۲)$	۰/۳۲۵	۰/۳۳۲	۰/۳۱۹	۰/۳۵۰*	۰/۳۳۳	۰/۳۲۹	۰/۳۴۳
	$G(۳,۳)$	۰/۶۳۴	۰/۶۳۶	۰/۶۱۶	۰/۶۶۹	۰/۶۵۷	۰/۶۴۶	۰/۶۷۳*
	$G(۴,۴)$	۰/۸۷۲*	۰/۸۳۰	۰/۸۱۱	۰/۸۵۸	۰/۸۴۳	۰/۸۲۹	۰/۸۵۸
	$W(۲,\Gamma(۱+\frac{۱}{۲}))$	۰/۶۹۰	۰/۷۰۲	۰/۶۸۲	۰/۷۳۲*	۰/۷۰۷	۰/۷۰۱	۰/۷۲۶
۱۰	$W(۲,\Gamma(۱+\frac{۱}{۳}))$	۰/۹۸۲	۰/۹۸۳	۰/۹۷۷	۰/۹۸۷	۰/۹۸۵	۰/۹۸۴	۰/۹۹۰*
	$W(۲,\Gamma(۱+\frac{۱}{۴}))$	۰/۹۹۹*	۰/۹۹۹*	۰/۹۹۹*	۰/۹۹۹*	۰/۹۹۹*	۰/۹۹۹*	۰/۹۹۹*
	$LN(-۲,۲)$	۰/۲۰۴*	۰/۱۴۱	۰/۱۳۷	۰/۱۳۳	۰/۱۷۵	۰/۱۴۴	۰/۱۳۶
	$LN(-\frac{۹}{۴},۳)$	۰/۵۹۱*	۰/۵۰۱	۰/۴۹۹	۰/۵۱۷	۰/۵۱۶	۰/۴۹۲	۰/۴۴۷
	$LN(-۸,۴)$	۰/۸۱۵*	۰/۷۵۷	۰/۷۵۷	۰/۷۱۷	۰/۷۶۷	۰/۷۵۸	۰/۷۱۴
	$G(۲,۲)$	۰/۵۰۴	۰/۴۹۷	۰/۴۷۷	۰/۵۷۸	۰/۵۸۳	۰/۵۱۸	۰/۶۱۸*
	$G(۳,۳)$	۰/۸۹۰	۰/۸۷۷	۰/۸۵۸	۰/۹۲۸	۰/۹۳۵	۰/۸۹۱	۰/۹۴۵*
	$G(۴,۴)$	۰/۹۸۳	۰/۹۷۷	۰/۹۷۰	۰/۹۹۱	۰/۹۹۳	۰/۹۸۳	۰/۹۹۵*
	$W(۲,\Gamma(۱+\frac{۱}{۲}))$	۰/۹۳۲	۰/۹۲۹	۰/۹۱۸	۰/۹۶۱	۰/۹۵۶	۰/۹۳۲	۰/۹۶۴*
۲۰	$W(۲,\Gamma(۱+\frac{۱}{۳}))$	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*
	$W(۲,\Gamma(۱+\frac{۱}{۴}))$	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*
	$LN(-۲,۲)$	۰/۶۵۴*	۰/۵۸۳	۰/۵۷۸	۰/۶۱۲	۰/۶۶۵	۰/۵۸۰	۰/۶۱۲
	$LN(-\frac{۹}{۴},۳)$	۰/۹۶۷*	۰/۹۴۸	۰/۹۴۹	۰/۹۴۸	۰/۹۶۵	۰/۹۵۰	۰/۹۴۹
	$LN(-۸,۴)$	۰/۹۹۹*	۰/۹۹۵	۰/۹۹۵	۰/۹۹۳	۰/۹۹۷	۰/۹۹۵	۰/۹۹۴
	$G(۲,۲)$	۰/۷۵۵	۰/۷۳۱	۰/۷۴۶	۰/۸۴۴	۰/۸۴۴	۰/۷۵۴	۰/۸۴۹*
	$G(۳,۳)$	۰/۹۹۱	۰/۹۸۸	۰/۹۹۰	۰/۹۹۷	۰/۹۹۸*	۰/۹۹۱	۰/۹۹۸*
	$G(۴,۴)$	۱/۰۰۰*	۰/۹۹۹	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*
	$W(۲,\Gamma(۱+\frac{۱}{۲}))$	۰/۹۹۷	۰/۹۹۷	۰/۹۹۷	۰/۹۹۹*	۰/۹۹۹*	۰/۹۹۸	۰/۹۹۹*
۳۰	$W(۲,\Gamma(۱+\frac{۱}{۳}))$	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*
	$W(۲,\Gamma(۱+\frac{۱}{۴}))$	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*
	$LN(-۲,۲)$	۰/۹۶۷	۰/۹۵۳	۰/۹۵۲	۰/۹۶۲	۰/۹۷۱*	۰/۹۵۳	۰/۹۶۴
	$LN(-\frac{۹}{۴},۳)$	۱/۰۰۰*	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹	۱/۰۰۰*	۰/۹۹۹	۰/۹۹۹
	$LN(-۸,۴)$	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*	۱/۰۰۰*

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

بحث و نتیجه‌گیری

دو برآوردگر جدید آنتروپی برمبنای 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)،

ابراهیمی و همکارارن (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 مبتنی بر $TCZ_{۲}$ و در حجم‌های نمونه متوسط و بزرگ آزمون مبتنی بر $TYZ_{۲}$ 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.