Statistical Research and Training Center
©(2005) J. Stat. Res. Iran **2** 94-106

مجلّه‌ی پژوهش‌های آماری ایران
دوره‌ی ۲، شماره‌ی ۱، بهار و تابستان ۱۳۸۴، صص ۱-۱۳

یک آزمون جدید نیکویی برازش با استفاده از تابع مشخصه‌ی تجربی

مینا توحیدی* و مهدی سلمان‌پور

دانشگاه شیراز

چکیده. تابع مشخصه در تعیین تابع توزیع احتمال نقش مهم و کلیدی دارد و به‌طور منحصر به فردی تابع احتمال یک متغیر تصادفی به‌وسیله‌ی آن تعیین می‌شود. اگر $c(t)$ و $c_n(t)$ به‌ترتیب تابع مشخصه‌ی توابع توزیع $F(\cdot)$ و $F_n(\cdot)$ باشند، آنگاه فرض صفر $F(x) = F_n(x), \forall x \in \mathbb{R}$ را می‌توان به فرض $c(t) = c_n(t), \forall t \in \mathbb{R}$ تبدیل کرد و از تابع مشخصه‌ی تجربی، $c_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp\{itX_j\}$، در آزمون نیکویی برازش توزیع استفاده نمود. به همین دلیل، بین سال‌های ۱۹۷۲ تا ۱۹۹۳ بسیاری از پژوهشگران از تابع مشخصه‌ی تجربی برای آزمون فرض‌های مختلف آماری استفاده کردند. در بیش‌تر آزمون‌های معرفی شده، مقایسه‌هایی بین $c_n(t)$ و $c_0(t)$، برای تعداد کمی از مقادیر t، صورت گرفته است و این امر موجب سازگار نبودن آزمون‌ها شده است. ما در این مقاله سعی کرده‌ایم که مقایسه‌ی بین تابع مشخصه‌ی تجربی و تابع مشخصه‌ی توزیع جامعه تحت فرض صفر را برای تعداد زیادی از مقادیر t انجام دهیم و بدین ترتیب آزمونی را معرفی کرده‌ایم که بسیار پرتوان‌تر از آزمون‌های ناپارامتری قبلی بوده است.

واژگان کلیدی. آزمون نیکویی برازش؛ آزمون سازگار؛ بردارهای ویژه؛ تابع مشخصه؛ تابع مشخصه‌ی تجربی؛ روش مؤلفه‌های اصلی؛ قضیه‌ی حد مرکزی چندمتغیره؛ مقادیر ویژه.

* نویسنده‌ی عهده‌دار مکاتبات.

[DOI: 10.18869/acadpub.jsri.2.1.1]

۱ مقدمه

فرض کنید که X متغیری تصادفی در فضای احتمال (Ω, A, P) با تابع توزیع $F(x)$ باشد و بخواهیم فرض X دارای توزیع شناخته شده‌ی $F_0(x)$ را آزمون کنیم. توزیع $F_0(x)$ را توزیع برازنده بر X (یا توزیع جور با یافته‌های X) می‌گویند. بر اساس n مشاهده‌ی مستقل و هم‌توزیع ناشناخته‌ی $F(x)$، آزمون‌هایی را که برای آزمودن فرض $H_0 : F(x) = F_0(x),\ \forall x \in \mathbb{R}$ در مقابل فرض $H_1 : F(x) \neq F_0(x),\ \exists x \in \mathbb{R}$ به‌کار برده می‌شوند، آزمون‌های نیکویی برازش توزیع می‌نامند.

تاکنون آزمون‌های ناپارامتری بسیاری برای آزمودن فرض‌های بالا معرفی شده است که از جمله می‌توان به آزمون‌های کولموگروف-اسمیرنوف، کرامر-ون میسز، اندرسن-دارلینگ و واتسون اشاره کرد. با این وجود، آماردانان در مسائل مختلف آماری، همواره به دنبال یافتن آزمون‌های پرتوان‌تری بوده‌اند. همچنین در بعضی موارد، صورت بسته‌ای برای تابع توزیع $F_0(x)$ وجود ندارد و در نتیجه انجام آزمون‌های معروف یاد شده، امکان‌پذیر نیست. این مشکلات آماردانان را بر آن داشت که از تابع مشخصه‌ی تجربی در آزمون فرض‌های آماری استفاده کنند.

می‌دانیم که تابع مشخصه در تعیین تابع توزیع احتمال نقش مهم و کلیدی دارد و به‌وسیله‌ی آن می‌توان تابع احتمال یک متغیر تصادفی را تعیین کرد. به‌همین دلیل، بین سال‌های ۱۹۷۲ تا ۱۹۹۳ پژوهشگران در مقاله‌های گوناگون، با استفاده از تابع مشخصه‌ی تجربی، آزمون‌های نیکویی برازش را با در نظر گرفتن توزیع‌های مشخص $F_0(x)$ ارائه کرده‌اند.

هدف ما در این مقاله، طراحی یک آزمون نیکویی برازش کلی (با در نظر گرفتن هر توزیع مشخص $F_0(x)$) بر اساس تابع مشخصه‌ی تجربی است که پرتوان‌تر از آزمون‌های ارائه شده‌ی قبلی باشد.

در بخش دوم، به معرفی تابع مشخصه‌ی تجربی و برخی از ویژگی‌های آن خواهیم پرداخت. آماره‌ی کلّی آزمون و توزیع حدی آن را در بخش سوم ارائه خواهیم کرد و در بخش چهارم بر اساس تابع مشخصه‌ی تجربی و استفاده از روش مؤلفه‌های اصلی آماره‌ی جدیدی را معرفی خواهیم کرد. در پایان در بخش پنجم، با استفاده از شبیه‌سازی مونت کارلو در یک مثال، پرتوان‌تر بودن آزمون جدید را نسبت به آزمون‌های قبلی نشان خواهیم داد.

۲ تابع مشخصه‌ی تجربی و برخی ویژگی‌های آن

همان‌طور که می‌دانید تابع مشخصه‌ی یک متغیر تصادفی یک‌بعدی با تابع توزیع $F(x)$ به‌صورت زیر تعریف می‌شود:

[DOI: 10.18869/acadpub.jsri.2.1.1]

$$c(t) = E[\exp\{itX\}] = \int \exp\{itx\} dF(x),$$

و به‌طور منحصر به‌فردی تابع توزیع توسط تابع مشخصه تعیین می‌شود. این تابع دارای دو خاصیت مهم زیر است:

آ) $c(t) = E\{\cos(tX)\} + iE\{\sin(tX)\}$ که در آن $E\{\cos(tX)\}$ را قسمت حقیقی ($\mathrm{Re}\ c(t)$) و $E\{\sin(tX)\}$ را قسمت موهومی ($\mathrm{Im}\ c(t)$) تابع مشخصه می‌نامیم.

ب) تابع مشخصه برای هر متغیر تصادفی X موجود است و برای هر $t \in \mathbb{R}$، $|c(t)| \leqslant ۱$.

یک برآوردگر سازگار برای $c(t)$، با در دست داشتن یک نمونه‌ی تصادفی $X_۱$، ...، X_n از توزیع $F(x)$، تابع مشخصه‌ی تجربی است که به‌صورت زیر تعریف می‌شود:

$$c_n(t) = \frac{۱}{n}\sum_{j=۱}^{n} \exp\{itX_j\} = \frac{۱}{n}\sum_{j=۱}^{n} \cos(tX_j) + i\left\{\frac{۱}{n}\sum_{j=۱}^{n} \sin(tX_j)\right\}.$$

یکی از مهم‌ترین خواص احتمالاتی تابع مشخصه‌ی تجربی در لم زیر آمده است:

لم ۱ اگر $c_n(t)$ تابع مشخصه‌ی تجربی نمونه‌ی تصادفی $X_۱$، ...، X_n از توزیع $F(x)$ با تابع مشخصه‌ی $c(t)$ باشد، آن‌گاه همگرایی‌های زیر با احتمال ۱ برقرار است:

$$c_n(t) \to c(t),$$
$$\mathrm{Re}\ c_n(t) \to \mathrm{Re}\ c(t),$$
$$\mathrm{Im}\ c_n(t) \to \mathrm{Im}\ c(t),$$

که در آن‌ها $\mathrm{Re}\ c_n(t) = \frac{۱}{n}\sum_{j=۱}^{n} \cos(tX_j)$ و $\mathrm{Im}\ c_n(t) = \frac{۱}{n}\sum_{j=۱}^{n} \sin(tX_j)$ است.

برهان. با استفاده از قانون قوی اعداد بزرگ برای متغیرهای تصادفی $\{\exp\{itX_۱\}$، ...، $\exp\{itX_n\}$ و با توجه به متناهی بودن $|c_n(t)|$ و $|c(t)|$، این لم به‌سادگی اثبات می‌شود.

لم اخیر بیان می‌کند که در صورت بزرگ بودن اندازه‌ی نمونه، مقدار $c_n(t)$ به $c(t)$ نزدیک خواهد بود. با توجه به این‌که فرض $H_۰ : F(x) = F_۰(x),\ \forall x$ را می‌توان به‌صورت $H_۰ : c(t) = c_۰(t),\ \forall t \in \mathbb{R}$ نوشت، پس می‌توان آماره‌ی آزمون نیکویی برازش توزیع را بر اساس تابع مشخصه‌ی تجربی $c_n(t)$، طرح‌ریزی کرد.

[DOI: 10.18869/acadpub.jsri.2.1.1]

۳ آزمون‌های نیکویی برازش بر اساس تابع مشخصه‌ی تجربی

برای بررسی آزمون‌های نیکویی برازش بر اساس تابع مشخصه‌ی تجربی ابتدا لازم است مفهوم سازگاری آزمون را روشن سازیم.

تعریف ۱ یک آزمون فرض را سازگار گویند، هرگاه با افزایش اندازه‌ی نمونه، توان آزمون به‌سمت یک میل کند. به‌عبارت دیگر توان مجانبی آزمون برابر یک شود.

فیگین و هتکت (۱۹۷۶) آزمون فرض H_0 را با مقایسه‌ی $\mathrm{Im}\ c_n(t)$ (یا $\mathrm{Re}\ c_n(t)$) با $\mathrm{Im}\ c_0(t)$ (یا $\mathrm{Re}\ c_0(t)$) برای یک مقدار ثابت t انجام دادند. آن‌ها با انتخاب یک t مناسب، به آزمونی پرتوان‌تر از آزمون کرامر-ون میسز دست یافتند. فیورورگر و موریکا (۱۹۷۷) یک آزمون را برای فرض تقارن ارائه کردند. آن‌ها آماره‌ی خود را بر اساس انتگرال توان دوم اختلاف بین $\mathrm{Im}\ c_n(t)$ و صفر بیان کردند (تابع مشخصه، یک تابع حقیقی است اگر و فقط اگر توزیع مربوط متقارن باشد) سپس هال و ولش (۱۹۸۳) آزمونی را برای آزمودن فرض نرمال بودن توزیع ارائه دادند. در بیش‌تر آزمون‌های معرفی شده در مقاله‌های پژوهشگران، مقایسه‌هایی بین $c_n(t)$ و $c_0(t)$ برای برخی مقادیر t انجام شده است، در نتیجه آزمون‌های ارائه شده سازگار نیستند. در سال ۱۹۸۸ بارینس و هنز آماره‌ای را بر اساس انتگرال توان دوم اختلاف بین $c_n(t)$ و $c_0(t)$ طراحی کردند و سازگاری آزمون مطرح شده را اثبات کردند، اما محاسبه‌ی مقدار آماره بسیار مشکل بود.

در سال ۱۹۹۷ فن برای رسیدن به یک آزمون سازگار، آماره‌ای را بر اساس توان دوم اختلاف بین $c_n(t)$ و $c(t)$، در نقاط $t_1, \ldots, t_m$ معرفی کرد و برای اثبات سازگاری آزمون، فرض کرد که $m \equiv m_n \to +\infty$ وقتی که $n \to +\infty$. برای توضیح چگونگی انجام این آزمون، فرض کنید بردار $\mathbf{t}_m = (t_1, \ldots, t_m)$ حاوی m نقطه‌ی یاد شده، باشد. با تعریف بردارهای $\mathbf{Z}_n(\mathbf{t}_m)$ و $\mathbf{Z}(\mathbf{t}_m)$ به شکل زیر:

$$\mathbf{Z}_n(\mathbf{t}_m) = \begin{pmatrix} \mathrm{Re}\ c_n(t_1) \\ \mathrm{Re}\ c_n(t_2) \\ \vdots \\ \mathrm{Re}\ c_n(t_m) \\ \mathrm{Im}\ c_n(t_1) \\ \vdots \\ \mathrm{Im}\ c_n(t_m) \end{pmatrix}, \qquad \mathbf{Z}(\mathbf{t}_m) = \begin{pmatrix} \mathrm{Re}\ c(t_1) \\ \vdots \\ \mathrm{Re}\ c(t_m) \\ \mathrm{Im}\ c(t_1) \\ \vdots \\ \mathrm{Im}\ c(t_m) \end{pmatrix}.$$

[DOI: 10.18869/acadpub.jsri.2.1.1]

به‌راحتی دیده می‌شود که $\mathbf{Z}_n(\mathbf{t}_m)$ برآوردگری ناااریب و سازگار برای $\mathbf{Z}(\mathbf{t}_m)$ است، اگر تابع مشخصه‌ی توزیع جامعه $c\,(t)$ باشد. بنا بر این، به‌منظور بررسی فرض صفر، باید بردارهای $\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}(\mathbf{t}_m)$ را مورد ارزیابی قرار داد. تفاضل این دو بردار برابر است با:

$$\mathbf{Z}_n\left(\mathbf{t}_m\right) - \mathbf{Z}\left(\mathbf{t}_m\right) = \frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} \cos(t_1 X_j) - E\left\{\cos(t_1 X_j)\right\} \\ \vdots \\ \cos(t_m X_j) - E\left\{\cos(t_m X_j)\right\} \\ \sin(t_1 X_j) - E\left\{\sin(t_1 X_j)\right\} \\ \vdots \\ \sin(t_m X_j) - E\left\{\sin(t_m X_j)\right\} \end{pmatrix} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{Y}_j\left(\mathbf{t}_m\right),$$

که در آن

$$\left(\mathbf{Y}_j(\mathbf{t}_m)'\right) = \left(\cos(t_1 X_j) - E\left\{\cos(t_1 X_j)\right\}, \ldots, \sin(t_m X_j) - E\left\{\sin(t_m X_j)\right\}\right).$$

تحت فرض صفر، امید ریاضی بردار تصادفی $\mathbf{Y}_j\left(\mathbf{t}_m\right)$ برابر با بردار صفر $2m$ بعدی است. اگر ماتریس واریانس-کوواریانس $\mathbf{Y}_j\left(\mathbf{t}_m\right)$ را با Ω_0 نشان دهیم، این ماتریس را می‌توان به‌صورت زیر افراز کرد:

$$\Omega_0 = \begin{pmatrix} \Omega_{011} & \Omega_{012} \\ \Omega_{021} & \Omega_{022} \end{pmatrix},$$

که درایه‌های این زیر ماتریس‌ها برابر است با:

$$\begin{aligned} (\Omega_{011})_{ij} &= \operatorname{cov}\left\{\cos(t_i X_1),\ \cos(t_j X_1)\right\} \\ &= E\left\{\cos(t_i X_1)\cos(t_j X_1)\right\} - E\left\{\cos(t_i X_1)\right\} E\left\{\cos(t_j X_1)\right\} \\ &= \frac{1}{2}\big[\operatorname{Re}\, c_0(t_i + t_j) + \operatorname{Re}\, c_0(t_i - t_j) \\ &\quad - 2\operatorname{Re}\, c_0(t_i)\operatorname{Re}\, c_0(t_j)\big], \qquad i, j = 1, \ldots, m,\ i \neq j \end{aligned}$$

[DOI: 10.18869/acadpub.jsri.2.1.1]

و به‌صورت مشابه:

$$
\begin{aligned}
(\Omega_{\circ ۱۱})_{ii} &= \operatorname{var}\{\cos(t_i X_۱)\} \\
&= \frac{۱}{۲}\left\{ ۱ + \operatorname{Re}\, c_\circ(۲t_i) - ۲\left(\operatorname{Re}\, c_\circ(t_i)\right)^۲ \right\}, \qquad i = ۱,\ldots,m \\
(\Omega_{\circ ۱۲})_{ij} &= \operatorname{cov}\{\cos(t_i X_۱),\ \sin(t_j X_۱)\} \\
&= \frac{۱}{۲}\big\{ \operatorname{Im}\, c_\circ(t_i + t_j) - \operatorname{Im}\, c_\circ(t_i - t_j) \\
&\quad - ۲\operatorname{Re}\, c_\circ(t_i)\operatorname{Im}\, c_\circ(t_j)\big\}, \qquad i,j = ۱,\ldots,m \\
(\Omega_{\circ ۲۲})_{ij} &= \operatorname{cov}\{\sin(t_i X_۱),\ \sin(t_j X_۱)\} \\
&= \frac{۱}{۲}\big\{ \operatorname{Re}\, c_\circ(t_i - t_j) - \operatorname{Re}\, c_\circ(t_i + t_j) \\
&\quad - ۲\operatorname{Im}\, c_\circ(t_i)\operatorname{Im}\, c_\circ(t_j)\big\}, \qquad i,j = ۱,\ldots,m \\
(\Omega_{\circ ۲۲})_{ii} &= \operatorname{var}\{\sin(t_i X_۱)\} \\
&= \frac{۱}{۲}\left\{ ۱ - \operatorname{Re}\, c_\circ(۲t_i) - ۲\left(\operatorname{Im}\, c_\circ(t_i)\right)^۲ \right\}, \qquad i = ۱,\ldots,m \\
\Omega_{\circ ۲۱} &= \Omega'_{\circ ۱۲}.
\end{aligned}
$$

اگر $\mathbf{Z}_\circ(\mathbf{t}_m)$ مقدار $\mathbf{Z}(\mathbf{t}_m)$ تحت فرض صفر باشد، آن‌گاه آماره‌ی آزمون که در حقیقت همان آماره‌ی والْد برای بردار تصادفی $\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_\circ(\mathbf{t}_m)$ می‌باشد، به‌صورت زیر معرفی می‌شود:

$$
(۱) \qquad T_n^\circ = \left(\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_\circ(\mathbf{t}_m)\right)' \Omega_\circ^{-۱} \left(\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_\circ(\mathbf{t}_m)\right).
$$

فن (۱۹۹۷) با تعریف ماتریس وزن قطری $W(\mathbf{t}_m, \theta)$ وقتی که اعضای روی قطر $(w_j, j = ۱,\ldots,۲m)$ غیر منفی باشند و W وابسته به m یا θ (نه لزوماً هر دو) بوده و $\theta = \theta_n \to \circ$ وقتی که $n \to +\infty$، آماره‌ی جدید T_n را معرفی کرد:

$$
(۲) \qquad T_n = \left(\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_\circ(\mathbf{t}_m)\right)' \Omega_\circ^{-\frac{۱}{۲}} W(\mathbf{t}_m, \theta)\, \Omega_\circ^{-\frac{۱}{۲}} \left(\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_\circ(\mathbf{t}_m)\right),
$$

و در قضیه‌ای توزیع حدی T_n را ارائه داد:

قضیه‌ی ۱ (فن (۱۹۹۷)) فرض کنید W ماتریسی قطری باشد که به θ وابسته نباشد و m عدد صحیح مثبت و متناهی باشد، در این صورت تحت فرض صفر nT_n دارای توزیع حدی $\sum_{j=۱}^{۲m} w_j \chi^۲_{(۱),j}$ خواهد

[DOI: 10.18869/acadpub.jsri.2.1.1]

بود که در آن $\chi^{۲}_{(۱),j}$ متغیرهای تصادفی مستقل $\chi^{۲}$ با یک درجه‌ی آزادی است یعنی:

$$nT_n \xrightarrow{d} \sum_{j=۱}^{۲m} w_j \chi^{۲}_{(۱),j},$$

و همچنین اگر W ماتریسی خودتوان با رتبه‌ی k باشد $(k \leqslant ۲m)$ آن‌گاه:

$$nT_n \xrightarrow{d} \chi^{۲}_{(k)}.$$

برهان. با استفاده از قضیه‌ی حد مرکزی چندمتغیره، بردار تصادفی $\sqrt{n}\{\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_۰(\mathbf{t}_m)\}$ دارای توزیع مجانبی نرمال با میانگین صفر و واریانس $\Omega_۰$ است و در نتیجه قضیه به‌سادگی اثبات می‌شود.

در انجام این آزمون، یافتن مقدار m مناسب و همچنین انتخاب نقاط $t_۱$، ...، t_m دارای اهمیت ویژه‌ای است. این نقاط باید به‌گونه‌ای انتخاب شوند که بررسی تابع مشخصه‌ی تجربی در نقاط $t_۱$،...، t_m ، بتواند تعیین‌کننده‌ی تابع توزیع مناسب باشد. هر چند نویسندگان بسیاری آزمون نیکویی برازش بر اساس تابع مشخصه‌ی تجربی را مورد بررسی قرار داده‌اند، با این وجود روش کاربردی روشنی برای انتخاب m و نقاط t_i ارائه نشده است و در بیش‌تر مقاله‌ها m را عدد یک یا دو انتخاب کرده‌اند!

اوبانک و لارسیا (۱۹۹۲) m را عددی در نظرگرفتند که $(R(m))$، حاصل میانگین توان دوم خطای برآوردگر تابع چگالی را مینیمم کند.

$$R(m) = -(n+۱)\sum_{j=۱}^{m} \frac{\tilde{a}^{۲}_{j,n}}{n} + \frac{m}{n-۱} + \sum_{j=۱}^{m} \frac{\tilde{a}_{۲j,n}}{n-۱},$$

که در آن $\tilde{a}_{j,n} = \frac{۱}{n}\sum_{k=۱}^{m} \sqrt{۲}\cos(j\pi F_۰(x_k))$ و $F_۰$ تابع توزیع تحت فرض صفر می‌باشد.

با وجود آن که محاسبات لازم برای یافتن m از طریق پیشنهاد بالا، به‌وسیله‌ی نرم‌افزارهای ریاضی چندان دشوار نیست، ولی mای که از این راه به‌دست می‌آید، غالباً بسیار بزرگ (تقریباً برابر با اندازه‌ی نمونه) و یا بسیار کوچک (تقریباً برابر با یک) خواهد بود. کوچک بودن m، دقت آزمون را کاهش می‌دهد و بزرگ بودن m، محاسبات مربوط به معکوس کردن ماتریس واریانس-کوواریانس $\Omega_۰$ را مشکل می‌کند. در بسیاری از موارد به حالت‌هایی برمی‌خوریم که به علت خطای محاسبات و خطای گرد کردن، ماتریس $\Omega_۰$ معکوس‌پذیر نیست.

فن (۱۹۹۶) روش یافتن m مناسب را در آزمون نیکویی برازش توزیع نرمال استاندارد بیان کرد اما این روش قابل تعمیم به سایر توزیع‌ها نیست.

وجود این مشکلات، ما را بر آن داشت که روشی برای یافتن m ابداع کنیم که بتواند در آزمون نیکویی برازش هر توزیعی به کار آید. ما این کار را با استفاده از روش مؤلفه‌های اصلی انجام داده‌ایم و آماره‌ی

[DOI: 10.18869/acadpub.jsri.2.1.1]

جدیدی برای آزمون فرض صفر ارائه داده‌ایم که در بخش‌های بعد در مورد آن صحبت خواهیم کرد.

۴ آماره‌ی مؤلفه‌های اصلی والْد در آزمون نیکویی برازش

در محاسبه‌ی آماره‌ی والْد که در رابطه‌ی (۱) معرفی شد، به معکوس ماتریس $۲m \times ۲m$ بعدی $\Omega_۰$ نیاز داریم. اگر m بزرگ باشد، در بیش‌تر موارد عملی، به دلیل وجود مقادیر بسیار کوچک در بین مقادیر ویژه ماتریس $\Omega_۰$، این ماتریس معکوس را حتا با نرم‌افزارهای جدید رایانه‌ای نمی‌توان به‌دست آورد. بنا بر این، ما توجه خود را به‌سوی متغیرهایی معطوف می‌کنیم که بیش‌ترین واریانس را دارند، زیرا متغیرهایی با واریانس پایین را می‌توان به‌عنوان متغیرهایی ثابت در نظر گرفت که در ماتریس واریانس-کوواریانس $\Omega_۰$ تأثیری ندارند. بدین وسیله می‌توانیم به‌راحتی مسئله‌ی خود را در یک زیرفضا با بعد کم‌تر مورد مطالعه قرار دهیم.

این کار را به‌کمک روش مؤلفه‌های اصلی می‌توان انجام داد. یعنی ترکیب‌های خطی از $\mathbf{Y}_j(\mathbf{t}_m)$ را در نظر می‌گیریم که بیش‌ترین تأثیر را در ماتریس واریانس-کوواریانس $\Omega_۰$ داشته باشند. اگر $\lambda_۱ \geqslant \lambda_۲ \geqslant \ldots \geqslant \lambda_{۲m}$ مقادیر ویژه‌ی ماتریس $\Omega_۰$ باشند، واریانس مؤلفه‌های اصلی (ترکیب‌های خطی از $\mathbf{Y}(\mathbf{t}_m)$) برابر با مقادیر λ_i خواهد بود. k را به‌گونه‌ای می‌یابیم که نسبت $\frac{\sum_{i=۱}^{k}\lambda_i}{\sum_{i=۱}^{۲m}\lambda_i}$ یعنی سهم k مؤلفه‌ی اصلی اول از واریانس کلّ جامعه، مقدار نسبتاً بالایی (حدود ۹۰ تا ۹۹ درصد) باشد. سپس در مطالعات خود فقط از k مؤلفه‌ی اصلی اول استفاده می‌کنیم. اگر $\lambda_۱ \geqslant \lambda_۲ \geqslant \ldots \geqslant \lambda_k$، بزرگ‌ترین k مقدار ویژه‌ی ماتریس $\Omega_۰$ بوده و $\boldsymbol{\beta}_۱$، ...، $\boldsymbol{\beta}_k$، k بردار ویژه‌ی مربوط به آن‌ها باشند، آن‌گاه آماره‌ی «مؤلفه‌های اصلی والْد» را به‌صورت زیر تعریف می‌کنیم:

$$T_n^* = \left(\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_۰(\mathbf{t}_m)\right)' B_k \Lambda_k^{-۱} B_k' \left(\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_۰(\mathbf{t}_m)\right), \qquad (۳)$$

که در آن Λ_k یک ماتریس قطری $k \times k$ است که اعضای روی قطر آن مقادیر ویژه‌ی $\lambda_۱ \geqslant \ldots \geqslant \lambda_k$ هستند. همچنین B_k ماتریسی $۲m \times k$ بعدی است که ستون‌های آن k بردار ویژه‌ی مربوط به $\lambda_۱, \ldots, \lambda_k$ هستند.

قضیه‌ی زیر توزیع مجانبی nT_n^* را مشخص می‌کند که می‌توان در مسائل عملی از آن استفاده کرد.

قضیه‌ی ۲ اگر بردار تصادفی $\mathbf{v}_n$ برابر با $B_k'\left(\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_۰(\mathbf{t}_m)\right)$ باشد آن‌گاه آماره‌ی مؤلفه‌های اصلی والْد برابر است با:

$$T_n^* = \mathbf{v}_n' \Lambda_k^{-۱} \mathbf{v}_n,$$

[DOI: 10.18869/acadpub.jsri.2.1.1]

و تحت فرض صفر توزیع حدی nT_n^*، توزیع $\chi^۲$ با k درجه‌ی آزادی است یعنی:

$$nT_n^* \xrightarrow{d} \chi^۲_{(k)}.$$

برهان. با توجه به خواص مقادیر ویژه و بردارهای ویژه‌ی یک ماتریس، می‌دانیم که برای $i = ۱, \ldots, k$، $\boldsymbol{\beta}_i' \Omega_۰ \boldsymbol{\beta}_i = \lambda_i$ و برای $i, j = ۱, \ldots, k, i \neq j$، $\boldsymbol{\beta}_i' \Omega_۰ \boldsymbol{\beta}_j = ۰$ با استفاده از قضیه‌ی حد مرکزی چندمتغیره، تحت فرض صفر، توزیع مجانبی $\sqrt{n}\mathbf{v}_n$ نرمال با میانگین صفر و ماتریس واریانس Λ_k می‌باشد یعنی:

$$\sqrt{n}\mathbf{v}_n \xrightarrow{d} N(\mathbf{0}, \Lambda_k),$$

یا

$$\sqrt{n}\Lambda_k^{-\frac{۱}{۲}}\mathbf{V}_n \xrightarrow{d} N(\mathbf{0}, I),$$

و در نتیجه:

$$nT_n^* = n\mathbf{v}_n'\Lambda_k^{-۱}\mathbf{v}_n \xrightarrow{d} \sum_{j=۱}^{k} \chi^۲_{(۱),j} = \chi^۲_{(k)}.$$

۵ شبیه‌سازی و مقایسه‌ی آزمون جدید با آزمون‌های پیشین

به‌طور خلاصه می‌توان الگوریتم به‌دست آوردن آماره‌ی بالا و انجام آزمون نیکویی برازش توزیع را با در دست داشتن n مشاهده از توزیع جامعه، به‌شکل زیر نوشت:

۱) با در نظر گرفتن $(F_۰(\cdot))$ به‌عنوان توزیع جامعه تحت فرض صفر مقادیر $t_۱$، ...، t_m را به‌وسیله‌ی رابطه‌های $t_i = F_۰^{-۱}\left(\frac{i}{m+۱}\right)$، $i = ۱, \ldots, m$ به‌دست می‌آوریم. (m را مساوی n در نظر می‌گیریم.)

۲) بردارهای $\mathbf{Z}_n(\mathbf{t}_m)$ و $\mathbf{Z}_۰(\mathbf{t}_m)$ را تشکیل می‌دهیم.

۳) بردار $۲m$ بعدی $\mathbf{Z}_n(\mathbf{t}_m) - \mathbf{Z}_۰(\mathbf{t}_m)$ را مشخص می‌کنیم.

۴) ماتریس $\Omega_۰$ را می‌سازیم و مقادیر ویژه‌ی مربوط به آن‌ها را به‌دست می‌آوریم.

۵) مقدارهای ویژه را به‌صورت نزولی مرتب می‌کنیم و مقدار k مناسب را به‌گونه‌ای می‌یابیم که

$$\frac{\sum_{i=۱}^{k}\lambda_i}{\sum_{i=۱}^{۲m}\lambda_i} \geqslant ۰/۹۹.$$

[DOI: 10.18869/acadpub.jsri.2.1.1]

(البته انتخاب کردن ۰٫۹۹ اختیاری است و جهت افزایش دقت می‌توان از کران‌های بالاتر نیز استفاده کرد.)

۶) ماتریس‌های B_k و Λ_k را تشکیل داده و با محاسبه‌ی بردار $\mathbf{v}_n$ مقدار آماره‌ی T_n^* را به‌دست می‌آوریم.

۷) p-مقدار آزمون را با استفاده از فرمول $p = P\left(\chi^2_{(k)} > nT_n^*\right)$ p-مقدار برای رد یا قبول فرض صفر محاسبه می‌کنیم.

اگرچه به‌دست آوردن p-مقدار و مقایسه‌ی آن با میزان با معنایی α راهی برای پذیرش یا رد فرض صفر می‌باشد، اما برای بررسی برتری آزمون ارائه شده نسبت به آزمون‌های ناپارامتری دیگر نیاز به محاسبه‌ی توان آزمون داریم. توان آزمون با مشخص شدن فرض مقابل، قابل دسترسی است. در نتیجه باید فرض صفر $H_0 : F(x) = F_0(x)$ را در مقابل فرض مشخص $H_1 : F(x) = F_1(x)$ آزمون کنیم که در آن $F_0(\cdot)$ و $F_1(\cdot)$ توابع توزیع کاملاً مشخصی هستند.

در جدول ۱ نتایج مقادیر توان آزمون جدید در آزمودن فرض صفر نرمال در مقابل توزیع‌های متفاوتی، بر اساس ۱۰۰۰ نمونه‌ی ۵۰ تایی آورده شده است. مقادیر توان برای آزمون‌های ناپارامتری کولموگروف-اسمیرنوف (KS)، کرامر-ون میسز (CM)، اندرسن-دارلینگ (AD) و واتسون (WA) محاسبه شده است.

همه‌ی آماره‌ی آزمون‌های بالا بر اساس تابع توزیع تجربی بوده و به‌صورت زیر تعریف شده اند. آزمون کولموگروف-اسمیرنوف:

$$\begin{aligned} KS &= \max\left\{D^+, D^-\right\}, \\ D^+ &= \max_{1\leqslant j\leqslant n}\left\{\frac{j}{n} - \hat{F}\left(X_{(j)}\right)\right\}, \\ D^- &= \max_{1\leqslant j\leqslant n}\left\{\hat{F}\left(X_{(j)}\right) - \frac{j-1}{n}\right\}. \end{aligned}$$

آزمون کرامر-ون میسز:

$$CM = \frac{1}{12n} + \sum_{j=1}^{n}\left(\hat{F}\left(X_{(j)}\right) - \frac{2j-1}{2n}\right)^2.$$

آزمون اندرسن-دارلینگ:

$$AD = -n - \frac{1}{n}\sum_{j=1}^{n}\left[(2j-1)\log\hat{F}\left(X_{(j)}\right) + \{2(n-j)+1\}\log\left\{1-\hat{F}\left(X_{(j)}\right)\right\}\right].$$

[DOI: 10.18869/acadpub.jsri.2.1.1]

آزمون واتسون:

$$WA = \frac{۱}{۱۲n} + \sum_{j=۱}^{n} \left\{ \hat{F}\left(X_{(j)}\right) - \frac{۲j-۱}{۲n} \right\}^{۲} - n \left\{ \sum_{j=۱}^{n} \frac{\hat{F}\left(X_{(j)}\right)}{n} - \frac{۱}{۲} \right\}^{۲},$$

که در تمام آن‌ها $\hat{F}(\cdot)$ تابع توزیع X و $X_{(j)}$، jامین آماره‌ی ترتیبی است.

شبیه‌سازی با روش مونت کارلو در سطح ۵٪ و با استفاده از نرم‌افزارS-pluse و Maple(Waterloo) انجام گرفته است.

جدول ۱. درصد توان ۵ آزمون ناپارامتری در سطح ۵٪ برای فرض نرمال در برابر فرض‌های مختلف آماری

	KS	CM	WA	AD	NEW
Alternative					
Cauchy	۵۶	۵۹	۷۹	۴۶	۹۹
Exp(۱)	*	*	*	*	*
t(۳)	۱۱	۱۲	۱۷	۳۴	۷۳
Logistic	۵۳	۵۷	۹۰	۷۳	۹۹
U(−۲, ۲)	۳۳	۲۸	۶۸	۴۴	۷۶
Lognormal	*	*	*	*	*
Laplace	۶	۶	۱۰	۳۶	۷۶
Gumbel	۶۶	۷۸	۲۲	۹۲	۹۰
Triangular	*	*	*	*	*
GH(۰/۲, ۰/۵)	۱۴	۱۴	۱۶	۵۲	۶۷

همچنین برای فرض مقابل به‌ترتیب توزیع‌های زیر در نظرگرفته شده است:

- توزیع کوشی استاندارد؛
- توزیع نمایی با پارامتر ۱؛
- توزیع t با سه درجه‌ی آزادی؛
- توزیع لوژستیک؛
- توزیع لگ نرمال؛
- توزیع نمایی دوگانه (لاپلاس) استاندارد؛
- توزیع گامبل (مقدار کرانگین)؛

[DOI: 10.18869/acadpub.jsri.2.1.1]

- توزیع یکنواخت در بازه‌ی $(-۲, ۲)$؛
- توزیع مثلثی در بازه‌ی $(-۱, ۱)$؛
- توزیع توکی g-h با نماد GH(g,h) و به‌صورت $X = \exp\left\{\frac{h\mathbf{Z}^۲}{۲}\right\}\frac{\exp\{g\mathbf{Z}\}-۱}{g}$ که در آن $\mathbf{Z} \sim \mathrm{N}(۰, ۱)$.

در جدول ۱، NEW نشان‌دهنده‌ی آزمون جدید و نماد * نشان‌دهنده‌ی توان ۱۰۰٪ است. نتایج اصلی برگرفته از جدول ۱ به‌شرح زیر است:

۱) در بین آزمون‌هایی که بر پایه‌ی توزیع تجربی بنا شده‌اند، آزمون واتسون و اندرسن-دارلینگ پرتوان‌تر از کولموگروف-اسمیرنوف هستند. این موضوع به‌خصوص در فرض‌های مقابل با دم‌های سنگین همچون کوشی، لوژستیک و گامبل مشهودتر است.

۲) در فرض مقابل گامبل آزمون جدید ارائه شده به‌خوبی آزمون اندرسن-دارلینگ عمل می‌کند و در فرض‌های مقابل دیگر، آزمون جدید پرتوان‌تر از آزمون‌های دیگری است که بر پایه‌ی تابع توزیع تجربی استوارند. این موضوع ادعای ما را مبنی بر بهینه بودن روش جدید نشان می‌دهد.

مرجع‌ها

Baringhaus, L.; Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function, *Metrika* **35**, 339-348.

Eubank, R.L.; LaRiccia, V.N. (1992). Asymptotic comparison of Cramer-Von Mises and nonparametric function estimation techniques for testing goodness-of-fit, *Ann. Statist.* **20**, 2071-2086.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation, *J. Amer. Statist. Assoc.* **91**, 674-688.

Fan, Y. (1997). Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function, *J. Multivariate Statist.* **62**, 36-63.

Feigin, P.D.; Heathcote, C.R. (1976). The empirical characteristic function and the Cramer-Von Mises statistics, *Sankhyā* **38**, 309-325.

Feuerverger, A.; Mureika, R.A. (1977). The empirical characteristic function and its applications, *Ann. Statist.* **5**, 88-97.

Hall, P.; Welsh, A.H. (1983). A test for normality based on the empirical characteristic function, *Biometrika* **70**, 723-726.

[DOI: 10.18869/acadpub.jsri.2.1.1]

Lukacs, E. (1970). Characteristic function, 2nd ed. Charies Griffin, London.

دریافت: ۲۱ تیر ۱۳۸۴
آخرین اصلاح: ۹ اردیبهشت ۱۳۸۵

مینا توحیدی
گروه آمار، دانشکده‌ی علوم،
دانشگاه شیراز،
چهارراه ادبیات،
شیراز، ایران.
پیام‌نگار: *mtowhidi@susc.ac.ir*

مهدی سلمان‌پور
گروه آمار، دانشکده‌ی علوم،
دانشگاه شیراز،
چهارراه ادبیات،
شیراز، ایران.
پیام‌نگار: *mhi-salman@hotmail.com*

[DOI: 10.18869/acadpub.jsri.2.1.1]