آزمون فرض نمایی بودن بر اساس اطلاع کولبک ـ لیبلر [1]

مهدی زارعی [2]

چکیده

در این مقاله آزمونی برای فرض نمایی بودن بر اساس تابع اطلاع کولبک ـ لیبلر معرفی می‌شود. آزمون از برآورد آنتروپی واسیچک استفاده کرده و مقادیر بحرانی متناظر با استفاده از شبیه سازی مونت کارلو[3] محاسبه شده‌اند. در پایان استفاده از آزمون پیشنهاد شده در یک مثال تشریحی بررسی شده است.

۱. مقدمه

بسیاری از نتایج در آزمونهای طول عمر، بر فرض نمایی بودن توزیع عمر محصول استوارند. به منظور انجام آزمون فرض نمایی بودن، تاکنون آماره‌های آزمون مختلفی توسط برخی از آماردانان پیشنهاد شده‌اند. آمارهٔ آزمونی که در این مقاله استفاده می‌شود، بر اساس برآورد آنتروپی است.

در بخش دوم، آنتروپی و برخی از خواص آن و نیز آنتروپی نسبی یا تابع کولبک ـ لیبلر بررسی می‌شود. در بخش سوم بسط تابع اطلاع کولبک ـ لیبلر و معرفی یک آمارهٔ آزمون بر اساس برآورد آنتروپی مورد نظر است. در بخش چهارم نحوهٔ انجام آزمون و به دست آوردن نقاط بحرانی مطرح شده و در نهایت یک مثال عددی بررسی می‌شود.

۲. آنتروپی و آنتروپی نسبی

ابتدا مفهوم آنتروپی را برای حالتی که X یک متغیر تصادفی گسسته است، بررسی می کنیم. متغیر تصادفی X را با تابع جرم احتمال $P(x)=P(X=x)$ و برآمدهای $\{x_j, j=1,...,k\}$ در نظر بگیرید. عدم قطعیت مورد انتظار مرتبط با برآمد یک مشاهده از متغیر تصادفی X را آنتروپی متغیر تصادفی X گفته و با نماد $H(X)$ نشان می‌دهند [5] که عبارت است از:

$$H(X)=-\sum_{x}P(x)\log P(x) \qquad (1)$$

در بیشتر موارد معمولا از پایه‌های ۲ یا e برای log استفاده می‌شود که آنتروپی در پایهٔ ۲ برحسب بیت[4] و در پایهٔ e برحسب نت[5] سنجیده می‌شود [1].

مشاهده می‌شود که آنتروپی تابعی از احتمالهای $P(x_k),...,P(x_1)$ بوده و به مقادیر مشاهده شدهٔ x_j بستگی ندارد. برای حالت $P(x)=0$ ، $P(x)\log P(x)$ به طور قراردادی صفر در نظر گرفته می‌شود. می‌توان نشان داد که در حالت گسسته $0 \le H(X) \le K$ است و برابری $H(X)=\log K$ ، وقتی اتفاق می‌افتد که احتمالها، یعنی $P(x_j)$ ها برابر $1/K$ باشند. وقتی که همهٔ پیشامدها هم احتمال اند، بیشترین عدم قطعیت،از این نظر که کدام پیشامد رخ خواهد داد، پیش می‌آید. بنابراین داشتن بیشترین مقدار آنتروپی در این حالت رضایتبخش است. این واقعیت که وقتی پیشامدها به طور یکسان غیر حتمی‌اند، $H(X)$ مقدار بیشتری است،

---

[1] Kullback-leibler

[2] کارشناس ارشد آمار، دانشگاه آزاد اسلامی واحد شهرکرد

[3] Mont Carlo

[4] Bit

[5] Nat

ولی وقتی حتمیتی وجـود داشته باشد، $H(X)$ صفر است و انتخـاب آنتروپی را به عنوان اندازه‌ای برای بیان عدم قطعیت توجیه می‌کند. اکنون برای روشن شدن مفهوم عدم قطعیت، متغیر تصادفی X را در دو مثال زیر بررسی و مقایسه می‌کنیم.

مثال ۱: فرض کنید متغیر X چهار برآمد a، b، c و d را با احتمالهای برابر $\frac{1}{4}$ اختیار کند. با توجه به (۱) آنتروپی متغیر تصادفی X برابر است با:

$$H(X)=-\sum_{x}P(x)\log_{2}P(x)=2bits$$

البته تفسیر عدد به دست آمده و واحد آن از نظر کاربرد در نظریهٔ ارتباطات و کدگذاری حائز اهمیت است. برای توجیه بهتر این مقدار مثال دیگری را بررسی کرده، نتایج را مقایسه می‌کنیم.

مثال ۲: فرض کنید متغیر تصادفی مثال ۱ چهار برآمد a، b، c و d را به ترتیب با احتمالهای $\frac{1}{64}$، $\frac{1}{64}$، $\frac{1}{64}$ و $\frac{61}{64}$ اختیار کند. در این صورت آنتروپی X به صورت زیر به دست می‌آید:

$$H(X)=-\sum_{x}P(x)\log_{2}P(x)=0/215bits$$

با مقایسهٔ این دو مثال می‌توان به سادگی دریافت که در مثال ۱ با وجود برآمدهای با احتمالهای برابر، در حقیقت با عدم قطعیت در مورد برآمد X (بیشترین عدم قطعیت) مواجه هستیم و به همین دلیل آنتروپی نسبت به آنتروپی در مثال ۲ که در آن با عدم قطعیت کمتر (و اطلاع بیشتر) در مورد برآمدهای X مواجهیم، مقدار بزرگتری دارد.

اکنون پس از روشن شدن مفهوم عدم قطعیت و رابطهٔ آن با اطلاع، آنتروپی را برای حالتی که X یک متغیر تصادفی پیوسته با تابع چگالی $f(x)$ است، تحت عنوان آنتروپی پیوسته به صورت زیر تعریف می‌کنیم:

$$H(X)=H(f)=-\int_{-\infty}^{\infty}f(x)\ln f(x)dx \qquad (2)$$

برخلاف آنتروپی در حالت گسسته، آنتروپی پیوسته می‌تواند منفی باشد و نیز حالت $H(X)=\infty$ امکان پذیر است.

مثال ۳: فرض کنید متغیر تصادفی X دارای توزیع یکنواخت روی $(0,b)$ باشد. بنابراین چگالی آن به صورت زیر است:

$$f_X(x)=\begin{cases}\frac{1}{b} & ;0\le x\le b\\ 0 & ;\text{سایر جاها}\end{cases}$$

طبق تعریف آنتروپی پیوسته، آنتروپی X به صورت زیر محاسبه می‌شود:

$$H(X)=-\int_{0}^{b}\frac{1}{b}\ln\frac{1}{b}dx=\ln b\ .$$

واضح است که برای $b<1$ ، $\ln b<0$ بوده و در نتیجه آنتروپی منفی است. همچنین با بزرگ شدن مقدار b، آنتروپی افزایش می‌یابد که این افزایش با توجه به بزرگ شدن پراکندگی و در نتیجه افزایش میزان عدم قطعیت قابل توجیه است.

یکی از موضوعات مورد علاقه، آنتروپی چگالی f نسبت به چگالی g است که آن را تحت عنوان آنتروپی نسبی یا تابع اطلاع کولبک ـ لیبلر به صورت

$$K(f:g)=\int_{-\infty}^{\infty}f(x)\ln\frac{f(x)}{g(x)}dx=E_f\ln\frac{f(X)}{g(X)}$$

معرفی کرده، فرض می‌کنیم چگالیهای f و g به گونه‌ای باشند که انتگرال وجود داشته باشد. در واقع تابع کولبک ـ لیبلر اندازه‌ای از فاصله یا اختلاف بین دو توزیع است[۲]. یکی از خواص مهم تابع کولبک ـ لیبلر که برای دو حالت گسسته و پیوسته یکسان است، بیان می‌کند که اگر f و g دو تابع چگالی باشند، آنگاه $K(f:g)\ge 0$ و تساوی وقتی برقرار است که برای همهٔ xها، $f(x)=g(x)$ [٤].

۳. آزمون فرض نمایی بودن

فرض کنید نمونهٔ تصادفی $X_1,\dots,X_n$ از یک توزیع احتمال F با چگالی $f(x)$ روی تکیه‌گاهی نامنفی با میانگین $\mu<\infty$ گرفته شده باشد. می‌خواهیم فرض

$$H_0:f(x)=f_0(x;\lambda)=\lambda\exp(-\lambda x) \qquad (3)$$

را که در آن $\lambda=\frac{1}{\mu}$ معلوم (یا مجهول) است، در برابر فرض

$$H_1:f(x)\ne f_0(x;\lambda) \qquad (4)$$

آزمون کنیم. برای انجام تشخیص بین دو فرض (۳) و (٤) از تابع کولبک ـ لیبلر بین دو توزیع به صورت

$$K(f:f_0)=\int_0^\infty f(x)\ln\frac{f(x)}{f_0(x)}dx \qquad (۵)$$

استفاده می‌کنیم. بنابر آنچه گفته شد، $K(f:f_0)\geq 0$ و برابری تنها وقتی برقرار است که $f(x)=f_0(x)$ باشد.

واضح است که تحت فرض صفر، $K(f:f_0)=0$ بوده و مقدار بزرگ $K(f:f_0)$ فرض H_1 را تأیید می‌کند. با بسط عبارت (۵) داریم:

$$K(f:f_0)=\int_0^\infty f(x)\ln f(x)dx-\int_0^\infty f(x)\ln f_0(x)dx$$

با توجه به این که $f_0(x)=\lambda\exp(-\lambda x)$ و نیز تعریف آنتروپی می‌توان نوشت:

$$K(f:f_0)=-H[f(x)]-\ln\lambda+1 \qquad (۶)$$

واسیچک[۶] ([۶]) برآوردگری برای آنتروپی به صورت

$$H_{mn}=\frac{1}{n}\sum_{i=1}^{n}\ln\{\frac{n}{2m}(X_{(i+m)}-X_{(i-m)})\} \qquad (۷)$$

ارائه کرد. اندازهٔ پنجره‌ای m، یک عدد صحیح مثبت کمتر از $\frac{n}{2}$ بوده، $X_{(j)}=X_{(1)}$ اگر $j<1$، $X_{(j)}=X_{(n)}$ اگر $j>n$ و $X_{(1)},\ldots,X_{(n)}$ آماره‌های ترتیبی بر اساس نمونه‌ای تصادفی به حجم n هستند. حال برای برآورد (۶) از برآوردگر H_{mn} استفاده می‌کنیم. در صورتی که λ مجهول باشد، برآورد آن را بر اساس میانگین نمونه، به صورت

$$\hat{\lambda}=\frac{1}{\hat{\mu}}=\frac{1}{\bar{x}}$$

به کار می‌بریم. بنابراین برآورد تابع کولبک ـ لیبلر به صورت زیر به دست می‌آید:

$$K_{mn}=-H_{mn}+\ln\bar{x}+1$$

آمارهٔ K_{mn} اطلاع کولبک ـ لیبلر را بین توزیع اصلی داده‌ها و مدل نمایی برآورد می‌کند. مقدار بزرگ K_{mn} بیانگر غیر نمایی بودن توزیع نمونه است. با به کارگیری تبدیل یکنوای زیر در مورد K_{mn} به صورت

$$I_{mn}=\exp(-K_{mn})=\exp(H_{mn}-\ln\bar{x}-1) \qquad (۸)$$

می‌توانیم از I_{mn} برای انجام تشخیص بین H_0 و H_1 استفاده کنیم. مشاهده می‌شود که $0<I_{mn}<1$ بوده و مقدار کوچک I_{mn} فرض H_0 را تأیید می‌کند.

پس فرض H_0 را در سطح معنی‌دار بودن α رد می‌کنیم اگر $I_{mn}\leq C_{m,n}(\alpha)$، که در آن $C_{m,n}(\alpha)$ نقطهٔ بحرانی فرض صفر است و در ادامه نحوهٔ محاسبهٔ آن را شرح می‌دهیم.

نکته ـ اگر پارامتر نمایی تحت فرض H_0 مقدار معلوم $\lambda=\lambda_0$ باشد، بنابراین

$$I_{mn}(\lambda_0)=\exp(H_{mn}+\ln\lambda_0-1)$$

که $I_{mn}(\lambda_0)$ دارای تمام خواص I_{mn} است.

۴. به دست آوردن نقاط بحرانی $C_{m,n}(\alpha)$

محاسبهٔ آمارهٔ I_{mn} ساده است ولی به دست آوردن توزیع نمونه‌ای آن کاری دشوار است. بنابراین با استفاده از روش شبیه سازی مونت کارلو مقدار بحرانی $C_{m,n}(\alpha)$ را برای α در سطوح مختلف محاسبه می‌کنیم. بر همین اساس برای $n\leq 120$، تعداد ۵۰۰۰ نمونهٔ نمایی به حجم n تولید کرده و برای هر $m<\frac{n}{2}$، مقدار I_{mn} را محاسبه می‌کنیم. برای هر m و n با استفاده از توزیع تجربی I_{mn}، مقدار $C_{m,n}(\alpha)$ را برای α در سطوح ۰/۱، ۰/۰۱، ۰/۰۵ و ۰/۰۲۵ به دست می‌آوریم. برای هر n، مقداری از m که بیشترین $C_{m,n}(\alpha)$ را به دست می‌دهد، آزمونی را با کمترین محافظه کاری (در مورد رد H_0) مهیا می‌سازد که به طور یکنواخت توان بیشتری را در مقایسه با دیگر آزمونها نشان می‌دهد [۳].

جدول (۱) اندازهٔ پنجره‌ای مربوط به کمترین محافظه کاری را مطابق با مقادیر گوناگون n نشان داده و جدول (۲) بیشترین مقدار $C_{m,n}(\alpha)$ را برای اندازه‌های مختلف نمونهٔ n در سطوح مختلف α نشان می‌دهد.

برای اندازه نمونهٔ n که $120<n<300$، مقادیر بحرانی را می‌توان با استفاده از درونیابی خطی

$$C_{m,n}(\alpha)\approx a(\alpha)+0/5\times10^{-2}n \qquad (۹)$$

---

[۶] Vasicek

تقریب زد که $a(0/01)=0/81$ ، $a(0/025)=0/82$ ، $a(0/05)=0/83$ و $a(0/1)=0/84$.

بنابر آنچه گفته شد مراحل انجام آزمون را می‌توان به صورت زیر خلاصه کرد:

الف) ابتدا با استفاده از جدول (۱) اندازهٔ پنجره‌ای m را بر اساس اندازهٔ نمونهٔ n پیدا کرده، I_{mn} را محاسبه می‌کنیم.

ب) با استفاده از جدول (۲) نقطهٔ بحرانی $C_{m,n}(\alpha)$ را به دست آورده، برای مقادیر $300 > n > 120$ از تقریب (۹) استفاده می‌کنیم.

ج) فرض H_0 را در سطح معنی دار بودن α رد می‌کنیم اگر $I_{mn} < C_{m,n}(\alpha)$.

مثال ٤: داده‌های زیر مسافت پیموده شده (برحسب کیلومتر) برای ۱۹ تن از افراد متصدی حمل و نقل در ارتش را که در انجام خدمت، کوتاهی کرده‌اند، نشان می‌دهد:

۷۸۸، ۷۰٦، ٦۲۹، ۵۳۹، ۵۰۸، ۳۹۳، ۳۲۰، ۲۷۱، ۲۰۰، ۱٦۲

۲۸۸۰، ۲۳۵۵، ۱۹۸٤، ۱٦۰۳، ۱٤٦۲، ۱۱۸۲، ۱۱۰۱، ۱۰۰۳، ۸۸٤

می‌خواهیم فرض نمایی بودن توزیع این داده‌ها را در سطح معنی دار بودن ۰/۱ بررسی کنیم. ابتدا با استفاده از جدول (۱) اندازهٔ پنجره‌ای $m=4$ را به دست می‌آوریم، آنگاه با استفاده از روابط (۷) و (۸) مقدار $I_{mn}=0/72$ را محاسبه می‌کنیم. برای $n=19$ ، $m=4$ و $\alpha=0/1$ جدول (۲) مقدار بحرانی $C_{4,19}(0/1)=0/6937$ را به دست می‌دهد و از آنجایی که $I_{mn}=0/72 > C_{4,19}(0/1)=0/6937$ ، فرض H_0 یعنی نمایی بودن توزیع داده‌ها را نمی‌توان رد کرد.

جدول ۱: مقادیر m مطابق با بیشترین مقدار نقاط بحرانی

اندازهٔ پنجره‌ای m	اندازهٔ نمونهٔ n
۱	۳ − ٤
۲	۵ − ۷
۳	۸ − ۱٤
٤	۱۵ − ۲٤
۵	۲۵ − ۳۵
٦	۳٦ − ۵۰
۷	۵۱ − ۷۰
۸	۷۱ − ۸۰
۹	۸۱ − ۱۰۰
۱۰	۱۰۱ − ۱۲۰
۱۱	۱۲۱ − ۱۵۰
۱۲	۱۵۱ − ۲۰۰
۱۳	> ۲۰۰

جدول ۲: مقادیر بحرانی $C_{m,n}(\alpha)$

مقادیر α				
۰/۱۰	۰/۰۵	۰/۰۲۵	۰/۰۱	n
۰/۳۷۲۰	۰/۳۰۹۰	۰/۲۶۸۰	۰/۲۰۶۰	۵
۰/۵۵۷۲	۰/۵۱۷۸	۰/۴۷۹۰	۰/۴۳۹۲	۱۰
۰/۵۷۸۹	۰/۵۲۳۷	۰/۴۸۳۷	۰/۴۵۷۲	۱۱
۰/۵۹۱۲	۰/۵۵۴۹	۰/۵۰۸۶	۰/۴۶۴۷	۱۲
۰/۶۲۰۹	۰/۵۸۲۲	۰/۵۳۹۹	۰/۴۹۸۷	۱۳
۰/۵۱۱۰	۰/۶۱۱۲	۰/۵۷۷۴	۰/۵۰۱۵	۱۴
۰/۶۸۱۵	۰/۶۱۱۸	۰/۵۸۳۰	۰/۵۴۹۴	۱۵
۰/۶۶۳۰	۰/۶۳۳۹	۰/۵۹۸۴	۰/۵۷۱۴	۱۶
۰/۶۸۱۵	۰/۶۵۴۰	۰/۶۰۶۰	۰/۵۷۷۲	۱۷
۰/۶۸۶۲	۰/۶۵۹۰	۰/۶۱۶۸	۰/۵۸۰۲	۱۸
۰/۶۹۳۸	۰/۶۶۰۲	۰/۶۳۲۰	۰/۵۹۶۰	۱۹
۰/۷۰۴۵	۰/۶۷۹۹	۰/۶۵۵۰	۰/۶۰۷۶	۲۰
۰/۷۴۸۸	۰/۷۱۵۹	۰/۶۹۱۷	۰/۶۶۷۵	۲۵
۰/۷۷۶۲	۰/۷۵۳۵	۰/۷۲۷۷	۰/۷۰۰۳	۳۰
۰/۸۱۴۴	۰/۷۹۲۹	۰/۷۷۰۸	۰/۷۴۲۸	۴۰
۰/۸۳۳۷	۰/۸۱۴۲	۰/۸۰۲۰	۰/۷۷۶۲	۵۰
۰/۸۹۸۱	۰/۸۸۸۳	۰/۸۸۲۰	۰/۸۷۴۰	۱۰۰
۰/۹۰۸۱	۰/۹۰۲۳	۰/۸۹۲۰	۰/۸۸۶۴	۱۲۰

منابع

[۱] د. س. جونز، نظریهٔ مقدماتی اطلاع، ترجمهٔ ناصر رضا ارقامی، محمد علی پور عبدالله نژاد، تهران، مرکز نشر دانشگاهی،۱۳۷۷.

[2] Cover, T. M. & J. A. Thomas, 1991, *Element of Information Theory,* New York, Wiley.

[3] Ebrahimi, N.et al, 1992, *Testing Exponentiality Based on Kullback-Leibler Information,* J. R. Stat. Soc. B., 54, 739-748.

[4] Kullback, S., 1959, *Information Theory and Statistics,* New York,Wiley.

[5] Soofi, E. S., 1994, *Capturing The Intangible Cincept of Information,* JASA, 89, 1243-1254.

[6] Vasicek, O., 1976, *A Test for Normality Based on Sample Entropy,* J. R. Stat. Soc. B., 38, 54-59.