آزمون خوبی برازندگی توزیع با استفاده از نوعی نامساوی چبیشف

حسین نگارستانی [۱]، زهره شیشه بر [۲]، مینا توحیدی [۳]

چکیده:
تشخیص توزیع داده‌ها یکی از مسائل مهم و کاربردی در انجام استنباط آماری می‌باشد. آزمون‌های زیادی جهت بررسی توزیع داده‌ها وجود دارد. در این مقاله یک آزمون جدید بر پایه نوعی از نامساوی چبیشف ارائه می‌گردد. نامساوی ذکر شده، بهبود یافته‌ی نامساوی چبیشف می‌باشد. مقایسه توان این آزمون با دیگر آزمون‌ها بیانگر برتری این آزمون می‌باشد.

واژه‌های کلیدی: آزمون خوبی برازندگی توزیع، نامساوی چبیشف.

۱ مقدمه

صورت کلاسیک نامساوی چبیشف به‌صورت زیر است:

$$P(|X-E(X)|\geq t)\leq \frac{Var(X)}{t^{2}}$$

این نامساوی برای هر متغیر تصادفی X با واریانس متناهی برقرار است. در سال ۲۰۰۶ گارود و همکارانش [۲] بر اساس نامساوی فوق، آزمونی را برای خوبی برازندگی توزیع ارائه کردند. به منظور یافتن آزمونی پرتوان‌تر، نیاز به کران دقیق‌تری برای نامساوی چبیشف داریم. یعنی نیاز به کوچکترین C_α داریم که در رابطه زیر صدق کند:

$$P(|X|\geq t)\leq C_\alpha.\frac{E|X|^\alpha}{t^\alpha}$$

که در آن α و t اعداد حقیقی مثبت هستند. C_α ثابت چبیشف نامیده می‌شود. سیسزار و همکارانش [۱] قضیه ای را برای به‌دست آوردن ثابت چبیشف در توزیع‌های آمیخته مقیاسی ارائه کردند. ما در اینجا به بیان این قضیه می‌پردازیم:

قضیه ۱: فرض کنید F یک تابع توزیع متقارن با تابع چگالی f باشد، به‌طوری‌که روی یک بازه متناهی یا نامتناهی مثبت و پیوسته و گشتاور مرتبه $a-$ ام آن متناهی باشد. اگر X یک متغیر تصادفی با توزیع آمیخته مقیاسی از تابع توزیع F باشد ℓ به صورت زیر تعریف شود:

$$\ell=\left(\frac{E|X|^a}{M_a}\right)^{\frac{1}{a}}$$

که درآن :

$$M_a=\int_{-\infty}^{\infty}|z|^a dF(z)$$

آن‌گاه:

$$\max P(|X|\geq t\ell)=\begin{cases} 2\bar{F}(z_a)z_a^a & t\geq z_a\\ 2\bar{F}(t) & \text{در سایر نقاط}\end{cases}$$

---
[۱] گروه آمار– دانشگاه شیراز
[۲] گروه آمار– دانشگاه شیراز
[۳] گروه آمار– دانشگاه شیراز

وقتی‌که z_a را کوچکترین ریشه مثبت معادله زیر در نظر گرفته شود:

$$\frac{zf(z)}{\overline{F}(z)} = a$$

که در آن $\overline{F}(z) = ۱ - F(z)$. در نتیجه ثابت چبیشف برابر است با:

$$C_a = \frac{۲\overline{F}(z_a)z_a^a}{M_a}$$

پیامد ۱: برای یک متغیر تصادفی X با تابع توزیع F، با در نظر گرفتن $\sigma = ۱$ در قضیه بالا ثابت چبیشف برابر است با:

$$D_a = \frac{۲\overline{F}(z_a)z_a^a}{E|X|^a}$$

در نتیجه یک کران بالا برای $P(|X| \geq t)$ به صورت زیر خواهد بود:

$$\frac{D_a}{t^a} = \frac{۲\overline{F}(z_a)z_a^a}{t^a}$$

۲ آزمون خوبی برازندگی چبیشف

هدف از انجام این آزمون این است که آیا یک مجموعه از داده‌ها را می‌توان با یک تابع توزیع مشخص $(F_۰)$ توصیف نمود. فرض کنید $Y_۱, Y_۲, \ldots, Y_n$ یک نمونه تصادفی n تایی از جامعه‌ای با توزیع F باشند. می‌خواهیم آزمون زیر را انجام دهیم:

$$\begin{cases} H_۰ : F = F_۰ \\ H_۱ : F \neq F_۰ \end{cases}$$

برای انجام این آزمون ابتدا بایستی داده‌های بین نقاط حداقل و حداکثر را به m بازه مساوی به طول h تقسیم کنیم. طول پیشنهادی بازه ها توسط اسکات به صورت زیر ارائه شد[۲]:

$$h = \frac{Y_{max} - Y_{min}}{۱ + \log_۲^n}$$

پس از این تقسیم بندی، فرض صفر را در هر بازه مورد بررسی قرار می‌دهیم. منطقی است که اگر در حداقل یک بازه اختلافی با توزیع فرض صفر وجود داشت، آن‌گاه فرض صفر را رد کنیم. یعنی داده ها از توزیع فرض صفر پیروی نمی کنند. در نتیجه رد فرض صفر در حداقل یک بازه معادل با رد فرض صفر کلی می‌باشد. احتمال این که یک مشاهده در بازه i —ام $(۱ \leq i \leq m)$ قرار گیرد را با p_i نمایش می‌دهیم.

فرض کنید n تعداد کل مشاهدات نمونه باشد. متغیر تصادفی X_i را تعداد مشاهدات در بازه i — ام در نظر می گیریم. همچنین آماره آزمون را در بازه i — ام به صورت زیر تعریف می‌شود[۲]:

$$\tau_i = \frac{X_i - E(X_i)}{\sqrt{Var(X_i)}}$$

فرض کنید که مشاهدات مستقل از یکدیگر باشند. در نتیجه متغیر تصادفی X_i دارای توزیع دوجمله‌ای با پارامترهای n و p_i است. از این رو $E(X_i) = np_i$ و $Var(X_i) = np_i(۱ - p_i)$ می‌باشد. p_i احتمال این است که یک مشاهده تحت فرض صفر در بازه i —ام قرار گیرد و به فرض صفر آزمون بستگی دارد.

حال با استفاده از قضیه حد مرکزی وقتی که $n \to \infty$ می‌توان توزیع متغیر تصادفی X_i که دوجمله‌ای می‌باشد را با توزیع نرمال تقریب زد (در عمل این تقریب برای نمونه‌هایی با حجم نمونه بیش از ۳۰ قابل استفاده

می‌باشد). در نتیجه با استفاده از نامساوی بهبود یافته چبیشف و قضیه (۱) برای توزیع نرمال در حالت $a = ۲$ خواهیم داشت :

$$\begin{aligned} P(|\tau_i| \geq k) &= P\left(\frac{|X_i - E(X_i)|}{\sqrt{Var(X_i)}} \geq k\right) \\ \cdots &= P(|Z| \geq k) \leq \frac{۲\overline{\Phi}(z_۲)z_۲^۲}{k^۲} \end{aligned}$$

که در آن Z دارای توزیع نرمال استاندارد و $\Phi(.)$ تابع توزیع مربوط به Z است .

در توزیع نرمال، برای حالت $a = ۲$ به وسیله محاسبات عددی مقدار $z_۲$ برابر ۱/۱۹ به‌دست آمده و طرف راست نامساوی فوق برابر است با:

$$\frac{۲\overline{\Phi}(z_۲)z_۲^۲}{k^۲} = \frac{\phi(z_۲)z_۲^۳}{k^۲} = \frac{۰/۳۳۱۴۳}{k^۲}$$

وقتی‌که $\phi(.)$ تابع چگالی مربوط به توزیع نرمال استاندارد باشد. همان‌طورکه اشاره شد رد فرض صفر در حداقل یک بازه $(|\tau_i| \geq k)$ معادل با رد فرض صفر کلی $(H_۰ : F = F_۰)$ می‌باشد. بنابراین می‌توان برای سهولت در انجام آزمون، آماره آزمون را به صورت زیر در نظر گرفت:

$$|\tau| = \max |\tau_i|$$

حال ناحیه پذیرش کلی [۴] را در نظر بگیرید:

$$\bigcap_{i=۱}^{m} (|\tau_i| < k) = (\max |\tau_i| < k)$$

از آنجا که ناحیه رد متمم ناحیه پذیرش می‌باشد، پس ناحیه رد [۵] به‌صورت زیر است:

$$\bigcup_{i=۱}^{m} (|\tau_i| \geq k) = (\max |\tau_i| \geq k) = (|\tau| \geq k)$$

[۴] Acceptance region
[۵] Rejection region

برای این‌که آزمون در سطح α باشد، باید رابطه زیر برقرار باشد:

$$P_{H_۰}(rejectH_۰) \leq \alpha.$$

از طرفی با استفاده از نامساوی بول داریم:

$$\begin{aligned} P_{H_۰}(Reject\ H_۰) &= P_{H_۰}(|\tau| \geq k) \\ &= P_{H_۰}\{\bigcup_{i=۱}^{m}(|\tau_i| \geq k)\} \\ &\leq \sum_{i=۱}^{m} P_{H_۰}(|\tau_i| \geq k) \leq \frac{mD_۲}{k^۲}. \end{aligned}$$

این آزمون در سطح α می‌باشد اگر مقدار $\frac{mD_۲}{k^۲}$ برابر α در نظر گرفته شود. در نهایت استراتژی آزمون جدید به‌صورت زیر اریه می‌شود:

در سطح معنی داری α فرض $H_۰$ رد می‌شود اگر و تنها اگر:

$$|\tau| \geq \sqrt{\frac{۰/۳۳۱۴۳m}{\alpha}}$$

این آزمون، آزمون چبیشف نامیده می‌شود.

۳ شبیه سازی

برای بررسی برتری آزمون جدید نسبت به دیگر آزمون‌ها توان این آزمون‌ها را با هم مقایسه می‌کنیم. برای مقایسه توان آزمون‌ها نیاز به مشخص کردن توزیع تحت فرض مقابل داریم . در فرض مقابل از توزیع‌های زیر استفاده می‌کنیم:

توزیع یکنواخت (۲ ،۲−)

توزیع t − استیودنت با سه درجه آزادی

توزیع لاپلاس

توزیع نمایی با میانگین $\frac{1}{3}$

توزیع کوشی استاندارد

توزیع لگ نرمال استاندارد

شبیه سازی به روش مونت کارلو در سطح ۰/۰۵ و با استفاده از نرم افزار $Minitab$ انجام شده است.

آزمون توزیع نرمال[6]:

برای مقایسه توان آزمون جدید و آزمون‌های معروفی همچون آندرسن − دارلینگ (AD)، کولموگروف − اسمیرنوف (KS) و خی − د[7] و برای آزمون کردن توزیع نرمال، ۱۰۰۰ نمونه ٦۰ تایی از توزیع‌های فرض مقابل تولید کرده و توان آزمون های ذکر شده را محاسبه نمودیم. نتایج در جدول شماره (۱) آمده است.

جدول ۱. توان آزمون‌های توزیع نرمال برای فرض‌های مقابل مختلف در سطح معنی داری ۰/۰۵.

$Chebyshev$	$Chi-Square$	KS	AD	فرض مقابل
۸۹	۷۲	۳۳	۴۴	$u(-۲,۲)$
٦۰	۵۵	۱۱	۳۴	$t(۳)$
۴۳	۳۳	٦	۳٦	$laplace$
۱۰۰	۹۸	۱۰۰	۱۰۰	$exp(۳)$
۹۸	۹۷	۵٦	۴٦	$cauchy$
۱۰۰	۹۹	۱۰۰	۱۰۰	$lognormal$

همان‌طور که در جدول (۱) آمده است توان آزمون جدید از دیگر آزمون‌ها بیشتر می‌باشد. پس این آزمون جهت بررسی نرمال بودن داده‌ها در حالتی که حجم نمونه بیشتر از ۵۰ است، توصیه می‌شود.

مراجع

[1] Csiszar, V., Mori, T. F. and Szekely, G. J. (2005), Chebyshev-type inequalities for scale mixtures,*Statist. Probab. Lett.*, 71, 323-335.

[2] Garrod, N., Pirkovic, S. R. and Valentincic A. (2006), Testing for discontinuity or type of distribution,*Math. Comput. Simul.*, 71, 9-15.

[3] Scott, D. (1992), *Multivariante Density Estimation*, John Wiley, New York.

[6] Normality test
[7] Chi-square