تبدیل مشخصه‌ای توزیع نرمال

نلی کریمانی
گروه آمار، دانشگاه پیام نور، ایران.
neli_ karimani@yahoo.com

حسین جباری خامنه‌ای
گروه آمار، دانشگاه تبریز
h_ jabbari@tabrizu.ac.ir

لیدر نوایی
گروه آمار، دانشگاه پیام نور، ایران.
leadernavaei@yahoo.com

چکیده

در این مقاله ما یك تبدیل جدید از متغیرهای تصادفی RVs که توزیع نرمال را مشخص می‌کند، پیشنهاد می‌کنیم. آن تبدیل به ما اجازه می‌دهد که n تا متغیر تصادفی نرمال هم‌توزیع و مستقل iid که دارای میانگین و واریانس نامعلوم هستند را به $(n-۲)$ تا متغیر نرمال جدید هم‌توزیع و مستقل با میانگین صفر زمانی که همان واریانس نامعلوم را در خود نگه دارد، تبدیل کنیم. این متعلق به کلاسی از تبدیلات است که برای کاهش تعدادی از پارامترهای نامعلوم یا از بین بردن تمامی آنها طراحی شده است. چند اظهار نظر قدیمی درباره روش‌هایی برای از بین بردن پارامترها در توزیع نرمال داده شده‌اند و دو کاربرد ممکن از تبدیل جدید شرح داده شده است.

واژگان کلیدی : آزمون نرمال بودن؛ پارامترهای مزاحم؛ تبدیلات متغیر تصادفی؛ مشخص‌سازی نرمال بودن.

۱. پیش‌گفتار

اولین مسئله مورد توجه، آزمون نرمال بودن می‌باشد. بیشترین مسائل نیکویی برازش که به شیوه واقعی به دست می‌آید شامل پارامترهای مزاحم می‌باشد. همانطور که فرض صفر مورد آزمون، فرض‌های مرکب هستند. آزمون‌های کلوموگروف، اندرسون-دارلینگ (۱۹۵۲) و سایر آزمون‌هایی که بر اساس انحراف EDF (تابع چگالی تجربی) هستند آزمون‌هایی خوب و دقیق می‌باشند که فقط نیاز به معلومات پارامتر دارند. تعدیل‌های بیشتری به این آزمون‌ها برای رفع این عیب (نیاز به معلومات) پیشنهاد داده شده‌اند و نیز برای بیشتر آزمون‌های دیگری که بر اساس انحراف EDF نیستند، فرمول‌بندی شده‌اند. به هر حال، تاکنون اکثریت آزمون‌های به کار رفته برای نرمال بودن (برای مثال، آزمون های جارکو-برا[۷]؛ لی لی فورس[۸]؛ شاپیرو-ویلک[۹]؛ اندرسون-دارلینگ تعدیل یافته [۱]؛ استیفن [۱۰] ؛ گریجو و همکاران [۳] و غیره) راه حل های پیشنهادی را که درست باشند را تضمین نمی‌کنند. بنابرین، یک راه طبیعی برای رسیدن به این نیاز، کاهش فرض مرکب که معادل یک فرض ساده است می‌باشد(بوندسون [۲]؛ سورگو-سیشادری [۴]). اولین مرحله برای بدست آوردن این نیاز،ب ه کاربردن تبدیل پیشنهاد شده اینجا بر روی داده های اصلی است. بدیهی است بعد از این تبدیل، یك تیمار اضافی از داده ها بایستی به منظور از بین بردن پارامترهای مقیاسی همچنان انجام داده شود. دومین استفاده ممکن از تبدیل مان (در اولین مرحله از تحلیل) به مسئله معروف بئرنس-فیشر مربوط می شود. تصور اصلی از تبدیل پیشنهاد شده آن است که به ما اجازه می‌دهد که یك مجموعه داده‌های نرمال تبدیل شده برای هر نمونه اصلی به دست آوریم که با میانگین صفر و همان واریانس‌های اصلی است، حتی اگر میانگین‌های دو نمونه اصلی به شدت متفاوت باشند. اولین کاهش به همان میانگین‌ها در مدتی که واریانس‌های اصلی به دست آمده است را حفظ می کند، ممکن است که روش‌های خودگردان و جایگشتی برای به دست آوردن توزیع دقیق صفر از هر آماره را به کار برد. همان طور که این تکنیك هنوز تحت بررسی است، ما این وضع را در این مقاله فعلی بیشتر بحث نمی‌کنیم.

۲. نتایج اصلی

تا آنجایی که می‌دانیم اولین مثال از یك تکنیك برای از بین بردن پارامتر مکانی مزاحم در هلمرت [۵، ۶] بود، که " تبدیل هلمرت " نامیده می‌شود. هنگام روبرو شدن با نمونه تصادفی به اندازه n از هر توزیع نرمال، آن تبدیل به ما اجازه می‌دهد که $(n-۱)$ مقادیر جدید هم‌توزیع و مستقل به طور نرمال توزیع شده را به دست آوریم که با میانگین صفر و همانند داده های اصلی دارای همان واریانس نامعلوم هستند. اگر $X_۱,\ldots,X_n$ ، n تا متغیرهای تصادفی هم توزیع و مستقل از توزیع نرمال با میانگین μ واریانس نامعلوم

σ^2 باشد $X \sim N(\mu, \sigma^2)$ متغیرهای تبدیل یافته هلمرت $Z_1, Z_2, \ldots Z_n$ همانند زیر تعریف شده‌اند:

$$Z_1 = (X_1 - X_2)/\sqrt{2}$$
$$\vdots$$
$$Z_i = (X_1 + \ldots + X_2 + iX_{i+1})/\sqrt{i(i+1)}$$
$$Z_{n-1} = (X_1 + \ldots + X_2 + (n-1)X_n)/\sqrt{n(n-1)}$$

که به اندازه $(n-1)$ هم‌توزیع و مستقل هستند که $Z \sim N(0, \sigma^2)$. این تبدیل در نوع خودش مهم است، زیرا منجر می‌شود که بدانیم زمانی که $\sum_i^{n-1} Z_i^2 = \sum_i^n (X_i - \overline{X})^2$، واریانس نمونه s^2 برابر است با: $s^2 = \sum_i^n (X_i - \overline{X})^2 / n-1$ که از توزیع کای دو $\sigma^2 \chi^2_{n-1} / (n-1)$ پیروی می‌کند و مستقل از میانگین نمونه $\overline{X}$ می‌باشد.

قضیه ۱.۲. فرض کنید $X_1, \ldots, X_n$ ، برای $n > 3$ متغیرهای تصادفی مستقل وهم‌توزیع باشند. تبدیل زیر را در نظر بگیرید:

$$Y_i = (X_i - \overline{X}) + (X_c - X_k)/\sqrt{2(n-2)}, \qquad i = 1, 2, \ldots, n, \qquad i \neq c, k \tag{1}$$

جایی که $i \leq c \neq k \leq n$ به صورت دو عدد صحیح دلخواه انتخاب شده‌اند و $\overline{X} = \sum_{i \neq (c,k)}^n X_i / (n-2)$ که $(n-2)$ تا متغیرهای تصادفی تبدیل شده به صورت $i.i.d$ هستند: $Y_i \sim N(0, \sigma^2), \quad i = 1, 2, \ldots, n, \quad i \neq c, k$ اگر و تنها اگر $X_i \sim N(\mu, \sigma^2)$.

اثبات. برای قسمت شرط لازم، کافی است معادله (۱) به روش زیر دوباره مرتب‌سازی شود:

$$Y_i = \frac{(n-3)X_i - sum^n_{j \neq (i,c,k)} X_j}{n-2} + \frac{(X_c - X_k)}{\sqrt{2(n-2)}} \tag{2}$$

از این فرمول‌سازی مجدد واضح است که معادله (۱) یك ترکیب خطی از متغیرهای تصادفی مستقل که مجموع آنها به صورت نرمال توزیع شده است زیرا مؤلفه‌های آن به صورت نرمال مستقل هستند. به هر حال، اگر Y_i عمومی به صورت نرمال توزیع شده باشد، بر اساس قضیه‌ی لوکاکس (۱۹۵۶) که به نو به خود نتیجه‌ی گذشته به دست آمده توسط کرامر(۱۹۳۶) راتعمیم می‌دهد، که مؤلفه‌های آن بایستی همچنان به صورت نرمال توزیع شده باشد.
برای قسمت شرط کافی، بایستی نشان داده شود که

الف) میانگین : $E(Yi) = 0$ ؛ ب) واریانس: $Var(Yi) = \sigma^2$؛ ج) Y_i ها به طور تصادفی مستقل هستند.

در (الف) بوسیله ساختاری که داریم: $\sum_{i \neq (c,k)}^n (X_i - \overline{X}) = 0$ و چون $E(X_i - \overline{X}) = E(X_j - \overline{X})$، نتیجه می‌شود که

$$E\left[\sum_{i \neq (c,k)}^{n} (X_i - \overline{X})\right] = (n-2)E(X_i - \overline{X}) = 0$$

و بنابرین $E(X_i - \overline{X}) = 0$. علاوه بر این $E\left[\frac{(X_c - X_k)}{\sqrt{2(n-2)}}\right] = \frac{E(X_c) - E(X_k)}{\sqrt{2(n-2)}} = 0$

در (ب) متغیرهای تصادفی $\sum_{i \neq (c,k)}^n (X_i - \overline{X})^2 \sim \sigma^2 \chi^2_{n-3}$، بنابراین

$$Var(Y_i) = Var(X_i - \overline{X}) + Var\left[\frac{(X_c - X_k)}{\sqrt{2(n-2)}}\right] = \sigma^2$$

در (ج) ابتدا بایستی ثابت شود که هر دو متغیر تصادفی عمومی Y_i ، Y_j و $i \neq j$ ناهمبسته هستند. با امتحان کردن رفتار $(n-2)$ تا متغیرهای تصادفی $(X_i - \overline{X})$ برای بررسی اینکه آنها به هم وابسته هستند، راحت است. چون $\sum_{i \neq (c,k)}^n (X_i - \overline{X}) = 0$ بنابرین ضرب دومی در $(X_i - \overline{X})$ و با به کار بردن عامل میانگین، به دست می‌آوریم:

$$E\left[(X_1 - \overline{X})(X_i - \overline{X}) + \ldots + (X_i - \overline{X})^2 + (X_n - \overline{X})(X_i - \overline{X})\right] = 0$$
$$E\left[(X_1 - \overline{X})(X_i - \overline{X})\right] = E\left[(X_j - \overline{X})(X_i - \overline{X})\right], \quad \forall i \neq j \qquad \textbf{و همچنین}$$
$$(n-3)E\left[(X_j - \overline{X})(X_i - \overline{X})\right] + E\left[(X_i - \overline{X})^2\right] = 0 \qquad \textbf{که نتیجه می‌دهد:}$$

بنابرین، $E\left[(X_j - \overline{X})(X_i - \overline{X})\right] = -\sigma^2/(n-2)$ آنگاه:

$$Cov(Y_i, Y_j) = E\left\{\left[(X_i - \overline{X}) + \frac{(X_c - X_k)}{\sqrt{2(n-2)}}\right]\left[(X_j - \overline{X}) + \frac{(X_c - X_k)}{\sqrt{2(n-2)}}\right]\right\}$$
$$= -\frac{\sigma^2}{(n-2)} + \frac{\sigma^2}{(n-2)} = 0$$

دوماً، برای اثبات استقلال تصادفی هر زوج $i \neq j, (Y_i, Y_j)$ باید نشان داده شود که متغیرهای تصادفی به صورت نرمال دو متغیره توزیع شده‌اند. در واقع اگر توزیع‌های حاشیه ای از نرمال دو متغیره ناهمبسته باشند، آنگاه آنها از هم مستقل هستند. برای نشان دادن این استقلال کافی است، نشان دهیم که متغیر تصادفی عمومی $\zeta = aY_i + bY_j$ برای هر زوج $(a,b) \in [R^۲ - (۰,۰)]$ به صورت نرمال توزیع شده‌اند. با استفاده از رابطه‌ی (۲) داریم:

$$\begin{aligned}\zeta &= a\left[\frac{(n-۳)X_i - \sum_{m\neq(i,c,k)}^{n} X_m}{n-۲} + \frac{(X_c - X_k)}{\sqrt{۲(n-۲)}}\right] \\ &+ b\left[\frac{(n-۳)X_j - \sum_{m\neq(i,c,k)}^{n} X_m}{n-۲} + \frac{(X_c - X_k)}{\sqrt{۲(n-۲)}}\right] \\ &= \frac{[(n-۳)a-b]X_i + [(n-۳)b-a]X_j - (a+b)\sum_{m\neq(i,c,k)}^{n} X_m}{n-۲} + \frac{(a+b)(X_c - X_k)}{\sqrt{۲(n-۲)}}\end{aligned}$$

حال ما بایستی بررسی کنیم که آیا در ادامه سه معادله‌ی وابسته به ضرایب در (a,b) ، برای حداقل یك جفت از اعداد حقیقی بواسطه $(۰,۰)$ به طور همزمان برقرار هستند: $(n-۳)a-b=۰; (n-۳)b-a=۰; a+b=۰$ چون $n>۳$ ، بررسی اینکه تنها زوج $(۰,۰)$ به طور مشترك در سه معادله صدق می‌کند، آسان است. این برهان قضیه را به اتمام می‌رساند. □

توجه ۱.۲. میانگین نمونه تبدیل شده Y_i ها عبارتند از: $\overline{Y} = frac(X_c - X_k)\sqrt{۲(n-۲)}$ این نتیجه‌ی ساده می‌تواند هنگام روبرو شدن با نمونه‌گیری مجدد و روش‌های جایگشتی مفید باشد. ارزش مشاهده آن را دارد که عمل ثابت بودن دوعدد صحیح دلخواه c و k معادل است با معرفی یك عنصر خارجی از تصادف بودن.

۳. مقایسه توان بین تبدیلات Y و Z

اندازه های دو نمونه $n = ۲۰, ۵۰$ در نظر گرفته شده‌اند و خطای نوع اول برای داشتن یک مقایسه شفاف‌تر در $(\alpha = ۰/۲۰)$ ثابت در نظر گرفته شده است. برای هر توزیع و اندازه نمونه، نمونه های تصادفی ۵۰۰ تایی استخراج شده‌اند و دو تبدیل برای هر نمونه بکاربرده شده و آماره آزمون کولموگروف استفاده شده است. نتیجه‌ها نبایستی به عنوان عملکردی از یک آزمون واقعی برای نرمال بودن تفسیر شوند، بلکه تنها به عنوان یک روش برای مقایسه تبدیل Z در مقابل تبدیل Y می باشند. در واقع، پارامتر مقیاسی معمولاً مجهول می‌باشد و برای وقتی که آماره آزمون در جایی که پارامترها در $H_۰$ باید به صورت کاملاً معلوم استفاده شده باشند، یک تیمار اضافی برای آزمون نرمال بودن لازم است. نتیجه‌های شبیه‌سازی توصیف شده در جدول ۱ ارائه شده‌اند. برای $n = ۲۰$ می‌تواند مشاهده شود که Y به طور واضح و یکنواخت قوی‌تر از Z است؛ با $n = ۵۰$، Y هنوز نسبت به Z قوی‌تر است اما اختلاف بین دو تبدیل تمایل به کاهش دارد. بنابرین نتیجه‌ها در موافقت با فرضیات قرار داده شده بالا ظاهر می‌شوند. بعضی از نویسندگان (به طور مثال ،سورگو و همکاران [۴]) در نظر گرفته‌اند که تبدیل Z اولین قدم برای آزمون نرمال بودن می‌باشد. چند نتیجه بهتر می‌تواند احتمالاً بوسیله جایگزینی Z با تبدیل Y ها به دست آید.

جدول ۱: مقایسه توان بین تبدیلات Y و Z تحت غیر نرمال بودن با استفاده از آزمون کولموگروف $(\alpha = ۰/۲۰)$

	n=20		*n=50*	
	Ys	*Zs*	*Ys*	*Zs*
یکنواخت	*28/7*	*23/2*	*33/0*	*29/1*
نمایی	*58/7*	*45/1*	*90/3*	*81/7*
لاپلاس	*42/0*	*30/6*	*45/5*	*42/0*
استیودنت $t(٤)$	*34/6*	*26/6*	*40/3*	*37/1*
گاما(۲)	*46/0*	*32/4*	*78/0*	*72/6*
بتا (۲,۱)	*32/8*	*27/2*	*39/3*	*34/3*

مراجع

[1] T. W. Anderson and D. A. Darling, *Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes.*, Ann. Mathemat. Statist. **23** (1952), 193–212.

[2] L. Bondesson, *To reduce a composite hypothesis to a simple one by sampling from the structural or conditional distribution*, Scand. J. Statist. **9(3)** (1982), 129–138.

[3] S. Grego, I. Pegoraro and F. Pesarin, *Problemi di verifica d'ipotesi di normalità: approfondimenti ed innovazioni*, Unpublished dissertation, Padua University: Padua, Italy (1997).

[4] M. Csorgo and V. Seshadri, *On the problem of replacing composite hypotheses by equivalent simple ones*, Rev. Int. Statist. Instit. **38** (1970), 351–68.

[5] F. R. Helmert, *Über die Berechnung des wahrscheinlichen Fehlers aus einenendlichen Anzahl wahrer Beobochtungsfehler*, Zeitschrift fuer angewandte Mathematik und Physik **20** (1875), 300–303.

[6] F. R. Helmert, *Über die Wahrscheinlichkeit der Potenzsummen der Beobachtungsfehler und über einige damit im Zusammenh änge stehende Fragen*, Zeitschrift fuer angewandte Mathematik und Physik **21** (1876), 192–218.

[7] C. M. Jarque and A. K. Bera, *A test for normality of observations and regression residuals*, Rev. Int. Statist. Instit. **55** (1987), 163–172.

[8] H. Lilliefors, *On the Kolmogorov–Smirnov test for normality with mean and variance unknown*, J. Amer. Statist. Assoc. **62** (1967), 399–402.

[9] S. S. Shapiro and M. B. Wilk, *An analysis of variance test for normality (complete samples)*, Biometrika **52** (1965), 591–611.

[10] M. A. Stephens, *EDF statistics for goodness of fit and some comparisons*, J. Amer. Statist. Assoc. **69** (1974), 730–737.