© Statistical Research and Training Center
مجلّه‌ی پژوهش‌های آماری ایران
پژوهشکده‌ی آمار
دوره‌ی ۲، شماره‌ی ۲، پاییز و زمستان ۱۳۸۴، صص ۱۴۱-۱۵۸
J. Statist. Res. Iran **2** (2005): 167-184

یک تقریب جدید برای توزیع صفر آماره‌ی آزمون نسبت درست‌نمایی برای بررسی k مشاهده‌ی دورافتاده در یک نمونه‌ی نرمال

رحیم محمودوند[†,*] و حسین حسنی[‡]

[†]دانشگاه پیام نور مرکز تویسرکان

[‡]بانک مرکزی جمهوری اسلامی ایران و دانشگاه کاردیف

چکیده. در این مقاله یک تقریب جدید برای توزیع صفرِ آماره‌ی آزمون نسبت درست‌نمایی برای آزمون k مشاهده‌ی دورافتاده‌ی پایینی یا بالایی در یک نمونه‌ی نرمال معرفی شده است. برای بررسی دقت تقریب جدید، مقدارهای بحرانیِ به دست آمده از تقریب جدید با مقدارهایی که توسط توزیع دقیق برای حالت‌های $k = ۱, ۲$ به دست آمده‌اند، مقایسه شده‌اند. همچنین نتایج با یک روش تقریبی دیگر ــ که به‌وسیله‌ی بارنت و لوییس (۱۹۹۴) ارائه شده است ــ برای حالت‌های $k = ۳, ۴$ مقایسه شده‌اند.

واژگان کلیدی. مشاهده‌ی دورافتاده؛ توزیع نرمال؛ آزمون نسبت درست‌نمایی؛ تقریب.

۱ مقدمه

برای انجام دادن آزمون‌های آماری یا روش‌های برآوردیابی، معمولاً فرض می‌شود که مشاهده‌های موجود در یک نمونه‌ی تصادفی، مستقل و هم‌توزیع‌اند. در برخی موارد، فرضی دیگر ــ که در آزمون‌های آماری، نقش

[*] نویسنده‌ی عهده‌دار مکاتبات

مهمی را ایفا می‌کند – به دو فرض پیشین افزوده شده و آن این است که مشاهده‌ها از یک جامعه‌ی نرمال استخراج شده‌اند. گاهی در یک نمونه‌ی موجود، دیده می‌شود که یک مشاهده یا مجموعه‌ای از مشاهده‌ها با بقیه ناسازگار است. به‌عنوان مثال، فرض کنید مشاهده‌ها بر اساس اندازه مرتب شده‌اند و برخی از آن‌ها بیش از حد معمول از مشاهده‌های دیگر بزرگ‌تر یا کوچک‌تر باشند، یا از الگوی توزیع مشاهده‌ها تبعیت نکنند. چنین مشاهده‌هایی، معمولاً دورافتاده نامیده می‌شوند. در این مورد، تعریف‌های مختلفی ارائه شده است (به‌عنوان مثال، هوکینز (۱۹۸۰)، روسو و لی‌روآ (۱۹۸۷)، بارنت و لوییس (۱۹۹۴)، مور و مکاب (۱۹۹۹) و آگاروال و یو (۲۰۰۱) را ببینید).

یکی از مراحل مهم در هر تحلیل آماری (پیش از آغاز تحلیل)، بررسی وجود مشاهده‌های دورافتاده است. بسیاری از فن‌های آماری به وجود مشاهده‌ی دورافتاده در نمونه حساس‌اند و در صورت وجود چنین مشاهده‌هایی، قابلیت چندانی ندارند. اما شایان ذکر است که روش‌هایی نیز وجود دارند که کم‌تر تحت تأثیر مشاهده‌های دورافتاده قرار دارند (به‌عنوان مثال، هامپل و دیگران (۱۹۸۶) را ببینید). مشاهده‌های دورافتاده می‌توانند حاصل از خطا باشند، که در آن صورت باید از مجموعه‌ی مشاهده‌ها کنار گذاشته شوند؛ ولی چنان‌چه شواهدی بر این واقعیت وجود نداشته باشد که غیر عادی بودن مشاهده(ها) به‌دلیل وجود خطا است، باید با آن‌ها محتاط عمل کرد و تحلیل‌ها را با وجود چنین مشاهده‌هایی انجام داد، اما با تأثیر کم‌تری نسبت به بقیه‌ی مشاهده‌ها (هوبر، ۱۹۸۱).

به‌طور خلاصه در خصوص مشاهده‌های دورافتاده می‌توان به سه نکته‌ی کلّی اشاره کرد: اوّل این‌که مشاهده‌های دورافتاده از کجا آمده‌اند، دوم این‌که چرا باید مراقب حضور این مشاهده‌ها بود، و سوم این‌که در صورت وجود مشاهده‌های دورافتاده، باید چگونه عمل کرد. روش‌های بسیار زیادی در مورد شناخت مشاهده‌های دورافتاده در یک نمونه‌ی تصادفی بیان شده است، که در ادامه به چند مورد در حد اختصار اشاره می‌شود.

برای تشخیص مشاهده‌های دورافتاده می‌توان از روش‌های عینی – که یک بینش اوّلیه در مورد مشاهده‌ها ارائه می‌کنند – استفاده کرد. در این‌گونه روش‌ها معمولاً از روش‌های نموداری (مانند هیستوگرام، جعبه‌ای، احتمال نرمال) یا استاندارد کردن مشاهده‌ها (مقایسه‌ی مقادیر استاندارد شده با چندک‌های توزیع نرمال) استفاده می‌کنند (چمبرز و دیگران، ۱۹۸۳)؛ اما بدیهی است که این‌گونه قضاوت از دقت کافی برخوردار نیست و متکی به سلیقه‌ی تحلیل‌گر در خصوص انتخاب مشاهده‌ی دورافتاده است.

علاوه بر این روش‌ها می‌توان از آزمون‌های آماری نیز برای تشخیص مشاهده‌ی دورافتاده استفاده کرد. از این آزمون‌ها می‌توان برای تشخیص چولگی توزیع مشاهده‌ها در حالتی نیز که اندازه‌ی نمونه کوچک است، استفاده کرد (دیوید، ۱۹۸۱). همان‌طور که پیش‌تر بیان شد، اهمیت وجود توزیع نرمال و استفاده از آن در آزمون‌های آماری بسیار مهم است. در این خصوص، آزمون‌هایی به‌منظور کشف مشاهده‌های

دورافتاده (هنگامی که نمونه‌ی استخراج شده از توزیع نرمال باشد) به‌وسیله‌ی فرگوسن (۱۹۶۱)، گرابس (۱۹۶۹)، جانسون و هانت (۱۹۷۹) و تود و دیگران (۱۹۸۳) ارائه شده است. آزمون‌ها و مقادیر بحرانی متفاوتی مربوط به کشف مشاهده‌های دورافتاده مورد استفاده قرار گرفته، که برخی از آن‌ها را می‌توان در مورد توزیع نرمال نیز به کار گرفت. آزمون‌هایی که در مورد توزیع نرمال ارائه شده، جداول مقادیر بحرانی برای انواع آزمون‌ها را در حالتی که یک یا هر دو پارامتر توزیع نرمال (میانگین و واریانس) مجهول باشند، شامل می‌شوند (به‌عنوان مثال، تود (۱۹۸۵) را ببینید).

در این نوشتار فقط بررسی شده است که آیا تعداد مشخصی از مشاهده‌ها در یک نمونه‌ی نرمال (مثلاً kتای آن‌ها) نسبت به بقیه‌ی مشاهده‌ها دورافتاده هستند یا خیر. خواننده‌ی علاقه‌مند به مطالب مربوط به توزیع‌های دیگر می‌تواند به بارنت و لوییس (۱۹۹۴) و تود (۲۰۰۲) مراجعه کند. به‌عنوان یک روش برای بررسی این‌که چه تعداد از مشاهده‌ها در یک نمونه از توزیع نرمال، دورافتاده هستند، می‌توان از آزمون نسبت درست‌نمایی استفاده کرد. برای حالت‌های $k = ۱, ۲$ توزیع دقیق آماره‌ی آزمون نسبت درست‌نمایی مشخص است، اما زمانی که تعداد مشاهده‌های دورافتاده بیش از دو مورد باشد ($k > ۲$) هنوز توزیع دقیق این آماره مشخص نیست و این مسئله هنوز بدون جواب باقی مانده است (ژانگ و یو، ۲۰۰۶). هر چند به دست آوردن توزیع دقیق، امری ضروری و لازم به نظر می‌رسد، علی‌رغم تلاش برای یافتن توزیع دقیق، استفاده از تقریب‌های مطلوب نیز روش دیگری است که مناسب به نظر می‌رسد.

در این مقاله، ضمن بیان توزیع تقریبیِ ارائه‌شده از پیش، برای آماره‌ی آزمون نسبت درست‌نمایی (در حالتی که تعداد مشاهده‌های دورافتاده از دو مورد بیش‌تر باشد) یک تقریب جدید برای توزیع این آماره در حالت کلّی مطرح و با تقریب بارنت و لوییس (۱۹۹۴) مقایسه شده است. به‌منظور جامعیت مقایسه‌ی تقریب ارائه‌شده در این مقاله، از دو رویکرد سود جسته‌ایم: یکی این‌که مقادیر حاصل از تقریبِ ارائه‌شده در این مقاله را با مقادیری که از توزیع دقیق برای حالت‌های $k = ۱, ۲$ حاصل شده‌اند تطابق داده‌ایم، و دیگری این‌که مقادیر حاصل از این تقریب را با تقریب بارنت و لوییس برای موارد $k = ۳, ۴$ مقایسه کرده‌ایم.

۲ آزمون نسبت درست‌نمایی برای تشخیص k مشاهده‌ی دورافتاده‌ی بالایی (پایینی) در یک نمونه‌ی نرمال

فرض کنید $x_۱, \ldots, x_n$ نمونه‌ای تصادفی از یک جامعه‌ی نرمال، و $x_{(۱)}, \ldots, x_{(n)}$ مقادیر مرتب‌شده‌ی آن‌ها باشد. برای تشخیص k مقدار دورافتاده‌ی بالایی در این نمونه، فرض صفر به‌صورت

$H_0 : x_1, \ldots, x_n \sim \mathrm{N}(\mu, \sigma^2)$ و فرض مقابل به‌صورت زیر است:

$$H_1 : \begin{cases} x_{(1)}, \ldots, x_{(n-k)} \sim \mathrm{N}(\mu, \sigma^2) \\ x_{(n-k+1)}, \ldots, x_{(n)} \sim \mathrm{N}(\mu + a, \sigma^2) \end{cases} \qquad (1)$$

که در این فرض‌ها پارامترهای a، μ و σ^2 نامعلوم‌اند. برای بررسی ناسازگاری k مشاهده‌ی دورافتاده‌ی بالایی با بقیه‌ی مشاهده‌ها، می‌توان آزمون زیر را انجام داد:

$$\begin{cases} H_0 : a = 0 \\ H_1 : a > 0 \end{cases} \qquad (2)$$

آماره‌ی آزمون نسبت درست‌نمایی برای بررسی k مشاهده‌ی دورافتاده‌ی بالایی ($T_n^{(k)}$) و k مشاهده‌ی دورافتاده‌ی پایینی ($T_{n,(k)}$) به‌صورت زیر است (بارنت و لوییس، ۱۹۹۴):

$$T_n^{(k)} = \frac{x_{(n)} + \cdots + x_{(n-k+1)} - k\bar{x}}{s}, \qquad T_{n,(k)} = \frac{k\bar{x} - x_{(1)} - \cdots - x_{(k)}}{s},$$

که در آن، $\bar{x}$ و s به ترتیب میانگین و انحراف معیار نمونه هستند. پس بنا بر آزمون نسبت درست‌نمایی، هرگاه $T_n^{(k)} \geqslant C_\alpha$ ، فرض صفر رد می‌شود، که در آن C_α (α سطح آزمون است) باید با توجه به توزیع $T_n^{(k)}$ مشخص شود.

۳ توزیع $T_n^{(k)}$ و $T_{n,(k)}$

ابتدا توجه کنید که $(x_{(n)}, \ldots, x_{(1)}) \stackrel{d}{=} (-x_{(1)}, \ldots, -x_{(n)})$ و بنا بر این $T_n^{(k)} \stackrel{d}{=} T_{n,(k)}$. در نتیجه بدون از دست دادن کلّیتِ مسئله، کافی است فقط در باره‌ی توزیع $T_n^{(k)}$ صحبت کنیم. همچنین به‌سادگی می‌توان دید که به‌ازای $k = 1, \ldots, n-1$ همواره داریم

$$T_n^{(k)} \stackrel{d}{=} T_n^{(n-k)}, \qquad T_n^{(k)} \stackrel{d}{=} T_{n,(n-k)}, \qquad T_{n,(k)} \stackrel{d}{=} T_{n,(n-k)}. \qquad (3)$$

توزیع $T_n^{(k)}$ به‌طور دقیق برای حالتی که فقط یک مشاهده‌ی دورافتاده وجود داشته باشد ($k = 1$) با استفاده از روابط بازگشتی به دست آمده است (بارنت و لوییس، ۱۹۹۴). با تعمیم روش در حالتی که فقط یک مشاهده‌ی دورافتاده وجود دارد، توزیع $T_n^{(k)}$ با استفاده از یک رابطه‌ی بازگشتی مشابه برای حالتی که فقط دو مشاهده‌ی دورافتاده ($k = 2$) در نمونه وجود داشته باشد نیز به دست آمده است (ژانگ و یو، ۲۰۰۶). اما در حالت کلّی، هنوز توزیع این آماره در صورت وجود بیش از دو مشاهده‌ی دورافتاده مشخص نیست.

۳٫۱ توزیع دقیق $T_n^{(1)}$ و $T_n^{(2)}$

ابتدا توزیع $T_n^{(k)}$ را برای $k = 1, 2$ بیان می‌کنیم. در حالتی که $n = 2$ است، به‌راحتی می‌توان دید که

$$x_{(2)} = \frac{x_1 + x_2 + |x_1 - x_2|}{2}, \qquad s = \frac{|x_1 - x_2|}{\sqrt{2}},$$

و در نتیجه،

$$T_2^{(1)} = \frac{x_{(2)} - \bar{x}}{s} = \frac{1}{\sqrt{2}}.$$

بنا بر این، برای $n = 2$ توزیع $T_n^{(1)}$ به‌صورت زیر است.

$$P\left(T_n^{(1)} \leqslant t\right) = \begin{cases} 0, & t < \frac{1}{\sqrt{2}} \\ 1, & t \geqslant \frac{1}{\sqrt{2}} \end{cases}$$

ژانگ و یو (۲۰۰۶) نشان داده‌اند که توزیع $T_n^{(1)}$ برای $n = 3, 4, \ldots$ بر اساس رابطه‌ی زیر به دست می‌آید.

$$P\left(T_n^{(1)} \leqslant t\right) = n \int_{\frac{1}{\sqrt{n}}}^{t} P\left(T_{n-1}^{(1)} \leqslant g_n(x)\right) f_{T_n}(x)dx, \qquad (۴)$$

که در این رابطه،

$$f_{T_n}(x) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)} \cdot \frac{\sqrt{n}}{n-1} \left(1 - \frac{n}{(n-1)^2}x^2\right)^{\frac{n-2}{2}-1},$$

$$g_n(x) = \frac{nx}{(n-1)\sqrt{\frac{n-1}{n-2}\left\{1 - \frac{nx^2}{(n-1)^2}\right\}}},$$

$$\frac{1}{\sqrt{n}} \leqslant t \leqslant \frac{(n-1)}{\sqrt{n}}.$$

به همین ترتیب، رابطه‌ی زیر برای توزیع $T_n^{(2)}$ به دست می‌آید.

$$P\left(T_n^{(2)} \leqslant t\right) = P\left(T_n^{(1)} \leqslant \frac{t}{2}\right) + n \int_{\frac{t}{2}}^{\frac{n-1}{\sqrt{n}}} P\left(T_{n-1}^{(1)} \leqslant \tilde{g}_n(x)\right) f_{T_n}(x)dx, \qquad (۵)$$

که در آن،

$$\tilde{g}_n(x) = \frac{t - \frac{n-1}{n-2}x}{(n-1)\sqrt{\frac{n-1}{n-2}\left\{1 - \frac{nx^2}{(n-1)^2}\right\}}},$$

$$\frac{2}{\sqrt{n}} \leqslant t \leqslant \sqrt{\frac{2(n-1)(n-2)}{n}}.$$

۳٫۲ تقریب‌هایی برای توزیع $T_n^{(k)}$

محاسبه‌ی توزیع دقیق $T_n^{(k)}$ با استفاده از روابط بازگشتی یاد شده در بخش ۳٫۱ کار ساده‌ای نیست و همان‌طورکه پیش‌تر بیان شد، هنوزکاری جدّی در این زمینه صورت نگرفته است. در این مقاله دو روش تقریبی برای توزیع $T_n^{(k)}$ مطرح می‌شود؛ یکی از آن‌ها به‌وسیله‌ی بارنت و لوییس (۱۹۹۴) و مورد دیگر برای اوّلین بار در این مقاله ارائه شده است. ابتدا فرض می‌کنیم مجموعه‌ی اندیس‌های $I = \{1, 2, \ldots, n\}$ به دو قسمت، یکی $[I]_k = \{i_1, \ldots, i_k\}$ با شرط $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n$ و دیگری $(I)_k = I \backslash [I]_k$ افراز شده است. همچنین نماد $T_{[I]_k}$ را به‌صورت زیر تعریف می‌کنیم.

$$T_{[I]_k} = \frac{\sum_{i \in [I]_k} X_i - k\overline{X}}{S}.$$

با تعریف این نماد، می‌توان نوشت

$$T_n^{(k)} = \max_{[I]_k} T_{[I]_k}.$$

پس می‌توان گفت که تابع توزیع تجمعی هر یک از $\binom{n}{k}$ متغیر تصادفی $T_{[I]_k}$، یک کران بالا برای تابع توزیع تجمعی $T_n^{(k)}$ است.

قضیه‌ی ۱ توزیع متغیر تصادفی $T_n^{(k)}$ به‌صورت زیر است.

$$f_{T_{[I]_k}}(x) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)} \sqrt{\frac{n}{k(n-k)(n-1)}} \left(1 - \frac{n}{k(n-k)(n-1)}x^2\right)^{\frac{n-2}{2}-1},$$

که در آن،

$$|x| \leqslant \sqrt{\frac{k(n-k)(n-1)}{n}}.$$

برهان. ابتدا توجه کنید که

$$T_{[I]_k}=\frac{\sqrt{n-1}\ \frac{k(n-k)}{n}\left(\bar{X}_k-\bar{X}_{n-k}\right)}{\sqrt{\sum_{i\in[I]_k}\left(X_i-\bar{X}_k\right)^2+\sum_{i\in(I)_k}\left(X_i-\bar{X}_{n-k}\right)^2+\frac{k(n-k)}{n}\left(\bar{X}_k-\bar{X}_{n-k}\right)^2}}.$$

در نتیجه،

$$\frac{n}{k(n-k)(n-1)}T^2_{[I]_k}=\left\{1+\frac{\sum_{i\in[I]_k}\left(X_i-\bar{X}_k\right)^2+\sum_{i\in(I)_k}\left(X_i-\bar{X}_{n-k}\right)^2}{\frac{k(n-k)}{n}\left(\bar{X}_k-\bar{X}_{n-k}\right)^2}\right\}^{-1}, \qquad (۶)$$

که در روابط بالا $\overline{X}_k=\sum_{i\in[I]_k}X_i/k$ و $\overline{X}_{n-k}=\sum_{i\in(I)_k}X_i/(n-k)$. می‌توان دید که مقدار ماکسیمم طرف راست (۶) برابر با ۱ است و در نتیجه،

$$-\sqrt{\frac{k(n-k)(n-1)}{n}}\leqslant T_{[I]_k}\leqslant\sqrt{\frac{k(n-k)(n-1)}{n}},$$

که کران بالا برای $X_i=a$، $(i\in[I]_k)$ و $X_i=b\neq a$، $(i\in(I)_k)$، (a و b دو عدد حقیقی دلخواه هستند) حاصل می‌شود. همچنین کران پایینی برای $X_i=a$، $(i\in[I]_k)$ و $X_i=2a$، $(i\in(I)_k)$ حاصل می‌شود. واضح است که عبارت مخرج در سمت راست (۶) مستقل از صورت است و در ضمن، صورت و مخرج کسر دارای توزیع خی‌دو، به‌ترتیب با درجه‌های آزادی $n-2$ و ۱ هستند. بنا بر این، نسبت آن‌ها دارای توزیع فیشر با درجه‌های آزادی $n-2$ و ۱ است و بنا بر این، کلّ عبارت دارای توزیع بتا است؛ یعنی،

$$\frac{n}{k(n-k)(n-1)}T^2_{[I]_k}\sim\text{Beta}\left(\frac{1}{2},\frac{n-2}{2}\right).$$

نتیجه با یک تبدیل مناسب به دست می‌آید.

۳٫۲٫۱ تقریب بارنِت و لوییس

ابتدا توجه کنید که اگر A_1، A_2، . . .، و A_m پیشامدهای دلخواهی باشند، همواره رابطه‌ی زیر برای آن‌ها برقرار است.

$$P\left(\bigcap_{k=1}^{m}A_i\right)\leqslant P(A_k),\qquad k=1,2,\ldots,m$$

با استفاده از این نابرابری به‌راحتی می‌توان دید که

$$P\left(T_n^{(k)} \leqslant t\right) = P\left(\max_{[I]_k} T_{[I]_k} \leqslant t\right) = P\left(\bigcap_{[I]_k} \{T_{[I]_k} \leqslant t\}\right) \leqslant P\left(T_{[I]_k} \leqslant t\right)$$

و بنا بر این، یک کران بالا برای توزیع $T_n^{(k)}$ به‌صورت زیر است.

$$P\left(T_n^{(k)} \leqslant t\right) \leqslant \int_L^t f_{T_{[I]_k}}(x)\ dx, \qquad (۷)$$

که در آن،

$$L = -\sqrt{\frac{k(n-k)(n-۱)}{n}}, \qquad |t| \leqslant -L$$

همچنین اگر $A_۱$، $A_۲$، ...، و A_m پیشامدهای دلخواهی باشند، بنا بر نابرابری بون‌فرونی نتیجه می‌شود که

$$P\left(\bigcap_{k=۱}^{m} A_k\right) \geqslant \sum_{k=۱}^{m} P(A_k) - m + ۱$$

و لذا یک کران پایین برای توزیع $T_n^{(k)}$ به‌صورت زیر به دست می‌آید.

$$P\left(T_n^{(k)} \leqslant t\right) \geqslant m \int_L^t f_{T_{[I]_k}}(x)\ dx - m + ۱, \quad m = \binom{n}{k} \qquad (۸)$$

این کران می‌تواند به‌عنوان یک تقریب برای توزیع $T_n^{(k)}$ معرفی شود. این کران، معادل با تقریبی است که برای توزیع $T_n^{(k)}$ در بارنت و لوییس (۱۹۹۴) معرفی شده است. این کران، یک مشکل بزرگ دارد و آن هم این است که یک تابع احتمال نیست و بنا بر این، برای برخی از مقادیر t مناسب نیست. به‌عنوان مثال، اگر $P\left(T_{[I]_k} \leqslant t\right) \leqslant (m-۱)/m$ باشد، مقدار این کران منفی می‌شود و بنا بر این، استفاده از این تقریب بی‌معنا است.

۳٫۲٫۲ تقریب جدید

برای معرفی تقریب جدید، نخست دو مسئله به‌صورت مجزّا بررسی می‌شود. در ابتدا بررسی می‌شود که آیا توزیع مجانبی $T_{[I]_k}$ها نرمال است و سپس بررسی می‌شود که آیا همبستگی بین $T_{[I]_k}$ها نیز (هنگامی که اندازه‌ی نمونه زیاد می‌شود) به صفر گرایش دارد. نتیجه‌ی توأم این دو نکته، بررسی فرض استقلال

$T_{[I]_k}$هاست. برای مورد اوّل (نرمال بودن توزیع $T_{[I]_k}$ها) نتیجه‌ی زیر به‌راحتی به دست می‌آید.

$$\lim_{n\to\infty} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)} \sqrt{\frac{n}{k(n-k)(n-1)}} \left(1 - \frac{n}{k(n-k)(n-1)}x^2\right)^{\frac{n-4}{2}-1} = \frac{1}{\sqrt{2\pi k}} \exp\left\{-\frac{x^2}{2k}\right\}.$$

برای توضیح همبستگی بین $T_{[I]_k}$ها نخست توجه کنید که

$$E\left\{T_{[I]_k}\right\} = 0, \qquad E\left(\frac{n}{k(n-k)(n-1)} T^2_{[I]_k}\right) = \frac{1}{n-1}.$$

در نتیجه،

$$\mathrm{var}\left(T_{[I]_k}\right) = E\left(T^2_{[I]_k}\right) = \frac{k(n-k)}{n}.$$

همچنین

$$\sum_{[I]_k} T_{[I]_k} = 0, \qquad T_{[I]_k} = T_{\{i_1,\ldots,i_k\}} = \sum_{j=1}^{k} T_{i_j}.$$

فرض کنید $k = 1$. در این حالت نتیجه می‌شود که

$$\mathrm{var}\left(\sum_{i=1}^{n} T_i\right) = n - 1 + n(n-1)\mathrm{cov}\left(T_{\{i\}}, T_{\{j\}}\right) = 0.$$

در نتیجه،

$$\rho\left(T_{\{i\}}, T_{\{j\}}\right) = \frac{-1}{n-1}, \qquad i \neq j$$

با توجه به این رابطه، با افزایش n، همبستگی بین $T_{[I]_k}$ها به صفر نزدیک می‌شود. اما همه‌ی همبستگی‌های بین $T_{[I]_k}$ها برای kهای بزرگ‌تر از ۱ لزوماً منفی نیستند. برای حالت $k = 2$ نتیجه می‌شود که

$$\mathrm{cov}\left(T_{\{i_1,i_2\}}, T_{\{j_1,j_2\}}\right) = \mathrm{cov}\left(T_{\{i_1\}} + T_{\{i_2\}},\ T_{\{j_1\}} + T_{\{j_2\}}\right) = \mathrm{cov}\left(T_{\{i_1\}}, T_{j_1}\right) + \mathrm{cov}\left(T_{\{i_1\}}, T_{\{j_2\}}\right) + \mathrm{cov}\left(T_{\{i_2\}}, T_{\{j_2\}}\right) + \mathrm{cov}\left(T_{\{i_2\}}, T_{\{j_2\}}\right)$$

$$= \begin{cases} -\frac{۴}{n}, & i_۱ \neq i_۲ \neq j_۱ \neq j_۲ \\ ۱ - \frac{۴}{n}, & \begin{cases} i_۱ = j_۱ \neq i_۲ \neq j_۲ \\ i_۱ = j_۲ \neq i_۲ \neq j_۱ \\ i_۲ = j_۱ \neq i_۱ \neq j_۲ \\ i_۲ = j_۲ \neq i_۲ \neq j_۱ \end{cases} \end{cases}$$

بنا بر این، ضریب همبستگی به‌صورت زیر است.

$$\rho\left(T_{\{i_۱,i_۲\}}, T_{\{j_۱,j_۲\}}\right) = \begin{cases} -\frac{۲}{n-۱}, & i_۱ \neq i_۲ \neq j_۱ \neq j_۲ \\ \frac{n-۴}{۲n-۴}, & \begin{cases} i_۱ = j_۱ \neq i_۲ \neq j_۲ \\ i_۱ = j_۲ \neq i_۲ \neq j_۱ \\ i_۲ = j_۱ \neq i_۱ \neq j_۲ \\ i_۲ = j_۲ \neq i_۲ \neq j_۱ \end{cases} \end{cases}$$

برای حالت کلّی، فرض کنید z نشان‌دهنده‌ی تعداد اندیس‌های برابر در زوج $\left(T_{[I]_k}, T_{[J]_k}\right)$ یا مانند نمادگذاری قبلی، در زوج $\left(T_{\{i_۱,\ldots,i_k\}}, T_{\{j_۱,\ldots j_k\}}\right)$ باشد. در این صورت، ضریب همبستگی بین $T_{\{i_۱,\ldots,i_k\}}$ و $T_{\{j_۱,\ldots j_k\}}$ از رابطه‌ی زیر به دست می‌آید.

$$\rho\left(T_{\{i_۱,\ldots,i_k\}}, T_{\{j_۱,\ldots,j_k\}}\right) = \frac{zn - k^۲}{kn - k}, \qquad z = ۰, \ldots, k-۱ \tag{۹}$$

در نتیجه،

$$\lim_{n\to\infty} \rho\left(T_{\{i_۱,\ldots,i_k\}}, T_{\{j_۱,\ldots,j_k\}}\right) = \frac{z}{k}, \qquad z = ۰, \ldots, k-۱ \tag{۱۰}$$

از طرفی، توزیع فراوانی z در جدول زیر داده شده است.

z	۰	۱	۲	$\ldots$	$k-۱$	کل
f_i	$\binom{n}{k}\binom{k}{۰}\binom{n-k}{k}$	$\binom{n}{k}\binom{k}{۱}\binom{n-k}{k-۱}$	$\binom{n}{k}\binom{k}{۲}\binom{n-k}{k-۲}$	$\ldots$	$\binom{n}{k}\binom{k}{k-۱}\binom{n-k}{۱}$	$\binom{n}{k}^۲ - \binom{n}{k}$

با توجه به جدول توزیع فراوانی z، نسبت مقادیر z برابر با صفر، به کل تعداد زوج‌ها، بر اساس رابطه‌ی زیر به سمت ۱ میل می‌کند.

$$\frac{\binom{n}{k}\binom{n-k}{k}}{\binom{n}{k}^۲ - \binom{n}{k}} = \frac{\binom{n-k}{k}}{\binom{n}{k} - ۱} \geqslant \frac{\binom{n-k}{k}}{\binom{n}{k}} = \frac{(n-k)!(n-k)!}{n!(n-۲k)!} \to ۱$$

در نتیجه $P(Z=0) \to 1$ و این بدان معنا است که با افزایش n، متغیرهای تصادفی $\{T_{[I]_k}\}$ به سمت ناهمبستگی و نرمال بودن میل می‌کنند و لذا فرض استقلال متغیرهای تصادفی $\{T_{[I]_k}\}$ تا اندازه‌ای قابل قبول است.

با توجه به آنچه بیان شد، ما تقریب زیر را برای توزیع $T_n^{(k)}$ در حالت کلّی پیشنهاد می‌کنیم.

$$P\left(T_n^{(k)} \leqslant t\right) = P\left(\max_{[I]_k} T_{[I]_k} \leqslant t\right) = P\left(\bigcap_{[I]_k} \{T_{[I]_k} \leqslant t\}\right)$$

$$\approx \left\{\int_L^t f_{T_{[I]_k}}(x)\, dx\right\}^{\binom{n}{k}}, \qquad k = 1, 2, \ldots, n-1 \tag{11}$$

۳٫۳ ویژگی‌های تقریب جدید و بررسی دقت آن

به‌منظور بررسی دقت تقریبِ پیشنهاد شده، ذکر چند نکته لازم به نظر می‌رسد. یکی این‌که تقریبِ ارائه‌شده خاصیت هم‌توزیعی‌ای را که برای حالت دقیق، طبق رابطه‌ی (۳) وجود دارد، دارا است؛ زیرا به‌راحتی می‌توان دید که $T_{[I]_k} \stackrel{d}{=} T_{[I]_{n-k}}$ و در نتیجه،

$$P\left(T_n^{(n-k)} \leqslant t\right) = P\left(\max_{[I]_{n-k}} T_{[I]_{n-k}} \leqslant t\right) = P\left(\bigcap_{[I]_{n-k}} \{T_{[I]_{n-k}} \leqslant t\}\right)$$

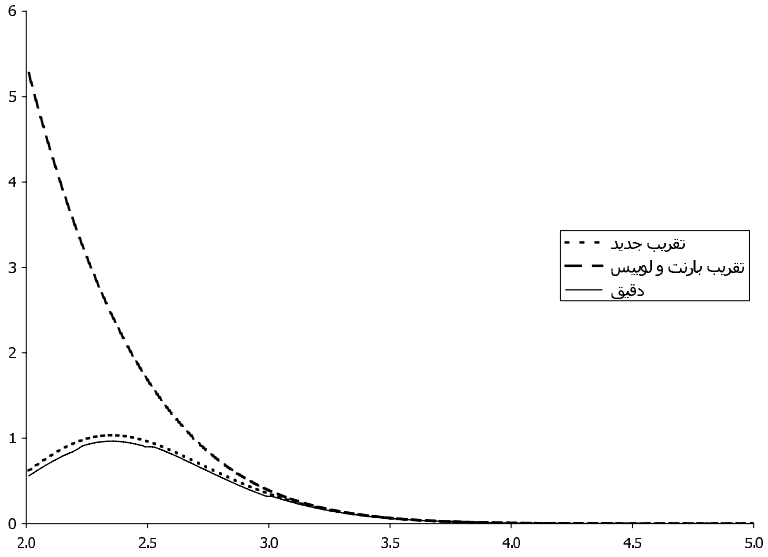
$$\approx \left\{\int_L^t f_{T_{[I]_{n-k}}}(x)\, dx\right\}^{\binom{n}{n-k}}$$

$$= \left\{\int_L^t f_{T_{[I]_k}}(x)\, dx\right\}^{\binom{n}{k}}, \qquad k = 1, \ldots, n-1. \tag{12}$$

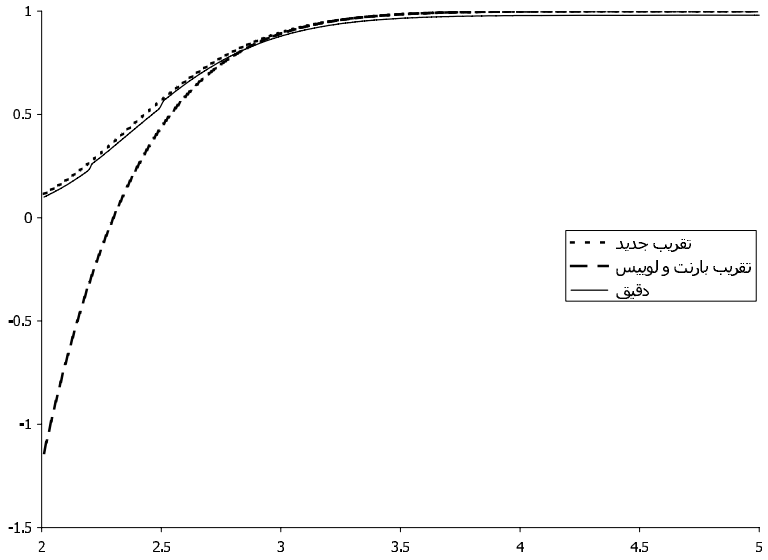
از (۱۱) و (۱۲) نتیجه می‌شود که توزیع‌های تقریبیِ معرفی‌شده برای $T_n^{(k)}$ و $T_n^{(n-k)}$ یکسان‌اند. همچنین تقریب پیشنهادی در این مقاله، عیب تقریبِ مطرح‌شده در بارنت و لوییس (۱۹۹۴) را ندارد؛ یعنی، یک تابع احتمال است. به‌عنوان مثال، با استفاده از نرم‌افزار Maple نمودارهایی برای مقایسه‌ی تابع توزیع تجمعی و تابع چگالی احتمال رسم شده است (شکل‌های ۱ و ۲).

در شکل ۱، بخشی از تابع چگالی احتمال $T_{100}^{(1)}$ برای حالت دقیق و با استفاده از تقریب بارنت و لوییس و تقریب جدید رسم شده است. در این شکل، محور افقی مقدار t را نشان می‌دهد و محور عمودی مقدار تابع چگالی احتمال $f_{T_{100}^{(1)}}(t)$ را.

مطابق با شکل ۱ دیده می‌شود که تقریب جدید، بسیار نزدیک‌تر به تابع چگالی دقیق است تا تقریب



شکل ۱. تابع‌های چگالی دقیق و تقریبی برای $T_{۱۰۰}^{(۱)}$



شکل ۲. تابع‌های توزیع تجمعی دقیق و تقریبی برای $T_{۱۰۰}^{(۱)}$

بارنت و لوییس. همچنین دقت تقریب‌ها با نزدیک شدن به دنباله‌ی توزیع تقریباً برابر می‌شود. اما واضح است که در حالت کلّی، دقت تقریب بارنت و لوییس در برآورد تابع چگالی احتمال، نسبت به تقریب جدید بسیار کم‌تر است.

در شکل ۲، بخشی از تابع توزیع تجمعی $T^{(1)}_{100}$ برای حالت دقیق و با استفاده از تقریب بارنت و لوییس و تقریب جدید رسم شده است. در این شکل، محور افقی مقدار t را نشان می‌دهد و محور عمودی مقدار تابع توزیع تجمعی $P\left(T^{(1)}_{100} \leqslant t\right)$ را.

در شکل ۲ دیده می‌شود که تقریب جدید برای tهای کوچک، بسیار نزدیک‌تر به تابع توزیع تجمعی دقیق است تا تقریب بارنت و لوییس. به‌تدریج با بزرگ شدن مقدار t، تفاوت سه منحنی کم می‌شود. اما واضح است که در حالت کلّی، دقت تقریب بارنت و لوییس در برآورد تابع توزیع تجمعی، نسبت به تقریب جدید بسیار کم‌تر است.

نکته‌ی دوم، دقت یا اختلاف مقادیر بحرانیِ به دست آمده از تقریب بارنت و لوییس و تقریب جدید با مقادیری است که از روش دقیق به دست آمده‌اند. با توجه به وجود توزیع دقیق $T^{(k)}_n$ برای $k = 1, 2$، مقدارهای بحرانی به‌صورت دقیق محاسبه شده‌اند (ژانگ و یو، ۲۰۰۶) و همین مقدارها با استفاده از روش‌های تقریبی محاسبه شده‌اند؛ لذا می‌توان آن‌ها را با هم مقایسه و دقت تقریب بارنت و لوییس و تقریب جدید را بررسی کرد. در جدول‌های ۱ و ۲ برخی از مقدارهای بحرانی مربوط به توزیع $T^{(k)}_n$ (در حالت‌های $k = 1, 2$) که حاصل روش‌های تقریبی‌اند، همراه با مقدارهای دقیق محاسبه‌شده‌ی آن‌ها ــ با استفاده از روابط (۱۱) و (۱۲) ــ ارائه شده است. در توصیف نتایج مندرج در این دو جدول، چهار نکته‌ی زیر جالب توجه است.

آ) تقریب جدید از لحاظ دقت، بهتر از تقریب بارنت و لوییس است؛ هر چند در برخی موارد، اختلاف آن‌ها ناچیز است. البته باید متذکر شد چنان‌چه $y = \int_L^t f_{T_{[I]_k}}(x)\, dx$ باشد، به‌راحتی می‌توان دید که

$$0 \leqslant y \leqslant 1 \quad \Rightarrow \quad y^m \leqslant y, \quad y^m \geqslant my - m + 1$$

و بنا بر این، دقت تقریب جدید، بیش‌تر از دقت تقریب بارنت و لوییس است.

ب) دقت تقریب جدید برای حالت $k = 1$ بیش‌تر است و قطعاً بدان دلیل است که همبستگی‌ها در این حالت به صفر نزدیک‌ترند و سریع‌تر همگرا به صفر می‌شوند.

جدول ۱. مقدارهای بحرانی برای توزیع $T_n^{(k)}$ در سطح $\alpha = ۰/۰۱$

	n					
	۵	۱۰	۲۰	۳۰	۵۰	۱۰۰
$k = ۱$						
دقیق	۱/۷۴۹	۲/۴۱۰	۲/۸۸۴	۳/۱۰۳	۳/۳۳۷	۳/۶۰۰
تقریب بارنت و لوییس	۱/۷۵۱	۲/۴۱۳	۲/۸۸۶	۳/۱۰۶	۳/۳۴۰	۳/۶۵۰
تقریب جدید	۱/۷۴۹	۲/۴۱۰	۲/۸۸۳	۳/۱۰۲	۳/۳۳۵	۳/۶۰۰
$k = ۲$						
دقیق	۲/۱۶۰	۳/۴۰۲	۴/۴۳۷	۴/۹۴۶	۵/۴۹۷	۶/۱۱۸
تقریب بارنت و لوییس	۲/۱۶۴	۳/۴۰۶	۴/۴۶۵	۴/۹۸۰	۵/۵۵۲	۶/۱۹۳
تقریب جدید	۲/۱۶۰	۳/۴۰۲	۴/۴۳۵	۴/۹۵۱	۵/۵۱۶	۶/۱۳۶

جدول ۲. مقدارهای بحرانی برای توزیع $T_n^{(k)}$ در سطح $\alpha = ۰/۰۵$

	n					
	۵	۱۰	۲۰	۳۰	۵۰	۱۰۰
$k = ۱$						
دقیق	۱/۶۷۱	۲/۱۷۶	۲/۵۵۷	۲/۷۴۵	۲/۹۵۶	۳/۲۰۷
تقریب بارنت و لوییس	۱/۶۷۲	۲/۱۷۹	۲/۵۵۹	۲/۷۴۹	۲/۹۶۷	۳/۲۲۵
تقریب جدید	۱/۶۷۰	۲/۱۷۳	۲/۵۵۳	۲/۷۴۳	۲/۹۵۵	۳/۲۰۷
$k = ۲$						
دقیق	۲/۱۰۱	۳/۱۹۷	۴/۱۱۰	۴/۵۶۱	۵/۰۵۸	۵/۶۳۸
تقریب بارنت و لوییس	۲/۱۰۳	۳/۱۹۸	۴/۱۲۳	۴/۶۰۰	۵/۱۲۳	۵/۷۴۵
تقریب جدید	۲/۱۰۰	۳/۱۹۳	۴/۱۱۳	۴/۵۸۴	۵/۰۹۴	۵/۶۵۷

پ) دقت هر دو تقریب برای nهای کوچک، بسیار زیاد است و برای nهای بزرگ‌تر، تقریب جدید دقیق‌تر است.

ت) با افزایش α دقت تقریب‌ها کم می‌شود تا جایی که همان‌طور که قبلاً مطرح شد، برای برخی از مقدارهای α تقریب بارنت و لوییس غیر قابل استفاده است.

در جدول‌های ۳ و ۴ مقدارهای بحرانیِ شبیه‌سازی‌شده (با استفاده از نمونه‌های به اندازه‌ی ۱۰,۰۰۰، حاصل از برنامه‌ی S-Plus، که ضمیمه‌ی مقاله است) برای $k = ۳, ۴$، همراه با مقدارهای محاسبه‌شده از

جدول ۳. مقدارهای بحرانی برای توزیع $T_n^{(k)}$ در سطح $\alpha = ۰/۰۱$

	n				
	۱۰	۲۰	۳۰	۵۰	۱۰۰
$k = ۳$					
شبیه‌سازی	۳/۹۹۷	۵/۶۱۲	۶/۴۳۱	۷/۳۲۹	۸/۳۸۸
تقریب بارنت و لوییس	۴/۰۰۴	۵/۶۳۰	۶/۴۶۷	۷/۳۹۹	۸/۴۹۳
تقریب جدید	۳/۹۹۸	۵/۶۱۴	۶/۴۵۱	۷/۳۸۸	۸/۴۷۴
$k = ۴$					
شبیه‌سازی	۴/۳۲۳	۶/۵۳۰	۷/۶۶۰	۸/۹۳۵	۱۰/۳۰۹
تقریب بارنت و لوییس	۴/۳۳۱	۶/۵۴۴	۷/۷۱۷	۹/۰۵۲	۱۰/۶۲۱
تقریب جدید	۴/۳۲۳	۶/۵۲۹	۷/۷۰۰	۹/۰۳۵	۱۰/۵۹۹

جدول ۴. مقدارهای بحرانی برای توزیع $T_n^{(k)}$ در سطح $\alpha = ۰/۰۵$

	n				
	۱۰	۲۰	۳۰	۵۰	۱۰۰
$k = ۳$					
شبیه‌سازی	۳/۸۱۳	۵/۳۱۱	۶/۰۵۱	۶/۸۷۱	۷/۸۵۵
تقریب بارنت و لوییس	۳/۸۱۸	۵/۳۲۱	۶/۱۱۱	۶/۹۹۹	۸/۰۵۶
تقریب جدید	۳/۸۱۴	۵/۳۱۴	۶/۱۰۰	۶/۹۹۲	۸/۰۴۴
$k = ۴$					
شبیه‌سازی	۴/۱۵۵	۶/۲۴۹	۷/۲۳۵	۸/۴۰۸	۹/۷۷۲
تقریب بارنت و لوییس	۴/۱۶۱	۶/۲۶۱	۷/۳۷۹	۸/۶۶۱	۱۰/۱۸۲
تقریب جدید	۴/۱۵۵	۶/۲۵۳	۷/۷۳۰	۸/۶۵۱	۱۰/۱۷۲

روش‌های تقریبی ارائه شده است. در مقایسه‌ی مقدارهای بحرانی در جدول‌های ۳ و ۴ به نظر می‌رسد که با افزایش n، مقدارهای بحرانی تا حدودی نسبت به مقدارهای شبیه‌سازی‌شده بزرگ‌تر باشند، در حالی که تفاوت آن‌ها برای نمونه‌های دارای اندازه‌ی کم‌تر از ۳۰ بسیار ناچیز است. البته باید توجه داشت که در یک تعداد تکرار ثابت، برای nهای کوچک‌تر، روش شبیه‌سازی دقیق‌تر عمل می‌کند. به هر ترتیب به نظر می‌رسد که تقریب بارنت و لوییس و تقریب جدید، برای nهای کوچک نیز بسیار مطلوب‌اند و نسبت به روش دقیق، نیاز به محاسباتی به مراتب کم‌تر دارند. علاوه بر این، مانند حالت $k = ۱, ۲$ دقت تقریب جدید، بیش‌تر از دقت تقریب بارنت و لوییس است و با افزایش n، این تفاوت بیش‌تر می‌شود.

۴ بحث و نتیجه‌گیری

در حدود یک قرن است که تشخیص مشاهده‌های دورافتاده مورد توجه آمارشناسان و دیگر دانشمندان است. محققان در طی سالیان متمادی روش‌های متعددی را برای تشخیص مشاهده‌های دورافتاده معرفی کرده‌اند. به‌دلیل اهمیت توزیع نرمال در مباحث آماری‌کاربردی، این مسئله برای توزیع نرمال نیز مطرح شده است. در این مقاله برای تشخیص مشاهده‌های دورافتاده در یک نمونه‌ی استخراج‌شده از توزیع نرمال، آزمون نسبت درست‌نمایی مورد استفاده قرارگرفت. توزیع دقیق آماره‌ی آزمون نسبت درست‌نماییِ به دست آمده، هنوز در نوشتگان آماری به‌جز برای یک و دو مشاهده‌ی دورافتاده نامشخص است. در این مقاله برای توزیع صفرِ آماره‌ی آزمون نسبت درست‌نمایی در حالت کلّی، یک فرمول تقریبی جدید (رابطه‌ی (۱۱) را ببینید) معرفی و ویژگی‌های آن بررسی شد. از مقایسه‌ی مقدارهای حاصل‌شده از تقریب پیشنهادی، تقریب بارنت و لوییس، و توزیع دقیق برای یک و دو مشاهده‌ی دورافتاده دیده شد (جدول‌های ۱ و ۲ را ببینید) که دقت تقریبِ پیشنهاد شده در این مقاله بسیار زیاد است. همچنین در مقایسه با مقدارهای شبیه‌سازی‌شده (که در حال حاضر، مورد استفاده‌ی عملی است) در حالتی که ۳ یا ۴ مشاهده‌ی دورافتاده در نمونه وجود داشته باشد، دیده شد (جدول‌های ۳ و ۴ را ببینید) که دقت تقریب بارنت و لوییس و تقریب جدید، زیاد است و البته برای اندازه‌ی نمونه‌ی بزرگ‌تر، مقدارهای محاسبه‌شده از روش‌های تقریبی، اندکی بزرگ‌تر می‌شوند. همچنین دیده شد (جدول‌های ۱ تا ۴ را ببینید) که دقت تقریب‌ها با حرکت به سمت دنباله‌ی راست توزیع $T_n^{(k)}$ (یعنی برای αهای کوچک‌تر) رو به افزایش است. اما با وجود این‌که به نظر می‌رسد دقت تقریب بارنت و لوییس و تقریب جدید، چندان متفاوت نیست، دیده شد که در برخی موارد، تقریب بارنت و لوییس منجر به نتایج نادرست می‌شود. در واقع، همان‌طور که پیش‌تر اشاره شد، تقریب بارنت و لوییس، تابع احتمال نیست و نمی‌توان از آن برای برآورد تابع چگالی احتمال استفاده کرد؛ در حالی که با استفاده از تقریب جدید، این امر نیز میسّر است. نکته‌ی دیگری که در باره‌ی این تقریب‌ها می‌توان مطرح کرد، ویژگی محاسباتی آن‌ها است که برای حالت‌های $k = ۱, ۲$ به‌مراتب منعطف‌تر و کم‌تر از روش دقیق است.

مرجع‌ها

Aggarwal, C.C.; Yu, P.S. (2001). Outlier detection for high dimentional data. In Proc. ACM SIGMOD.

Barnett, V.; Lewis, T. (1994). *Outliers in Statistical Data*, 3rd ed. Wiley, Chichester.

Chambers, J.; Cleveland, W.; Kleiner, B.; Tukey, P. (1983). *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA.

David, H.A. (1981). *Order Statistics*, 2nd ed. Wiley, New York.

Ferguson, T.S. (1961). On the rejection of outliers. Proc. Berkeley Symp. Math. Statist. and Prob. 1 Univ. of Cla. Press, Berkeley, pp. 253-285.

Grubbs, F. (1969). Procedures for detecting outlying observation in samples. *Technometrics* **11**, 1-12.

Hampel, F.R.; Ronchetti, E.; Rousseeuw, P.J.; Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Hawkins, D.M. (1980). *The Identification of Outliers*. Chapman, London.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Johnson, B.A.; Hunt, H.H. (1979). Performance characteristics for certain tests to detect outliers. Proceeding of statistical computing section, Annual meeting of the american statistical association, Washington, B.C.

Moore, D.S.; McCabe, G.P. (1999). *Introduction to the Practice of Statistics*, 3rd ed. New York.

Rousseeuw, P.J.; Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Thode, H.C. Jr. (1985). Power of absolute moment tests against symetric non-normal alternatives. Ph.D. Dissertation, University Microfilms, Ann Arbor, Michigan.

Thode, H.C. Jr. (2002). *Testing for Normality*. Marcel Dekker, New York.

Thode, H.C. Jr.; Smith, L.A.; Finch, S.J. (1983). Power of test of normality for detecting scale contaminated normal samples. *Commun. Stat. Simul. Comput.* **12**, 675-695.

Zhang, J.; Yu, k. (2006). The null distribution of the likelihood ratio test for one or two outliers in a normal sample. *Test* **15**, 141-150.

رحیم محمودوند
گروه آمار،
دانشگاه پیام نور مرکز تویسرکان،
تویسرکان، ایران.
پیام‌نگار: *r_mahmodvand@yahoo.com*

حسین حسنی
بانک مرکزی جمهوری اسلامی ایران،
خیابان پاسداران، ساختمان شماره ۲۰۷/۱،
تهران، ایران.

گروه آمار، دانشکده‌ی ریاضی،
دانشگاه کاردیف،
کاردیف، انگلستان.
پیام‌نگار: *hassanih@cf.ac.uk*

پیوست

برنامه‌ی زیر به‌راحتی می‌تواند برای شبیه‌سازی مقدارهای برخی از چندک‌های توزیع $T_n^{(k)}$ مورد استفاده قرارگیرد. در این برنامه، `rept` تعداد تکرارها برای شبیه‌سازی، `start` کم‌ترین مقدار ممکن n، و `end` مقدار پایانی مورد نظر برای n است.

```
quanoutnk<-function(rept,k,start,end){
z<-array(dim=c(end-start+1,6))
header<-c("alpha=0.20","alpha=0.10","alpha=0.05","alpha=0.01","alpha=0.005",
          "alpha=0.001")
lmargin<-c(paste("n="start:end))
dimnames(z)<-list(lmargin,header)
for(n in start : end){
y<-array(dim=rept)
for(i in 1 : rept){
x<-sort(c(rnorm(n)))
y[i]<-(sum(x[(n-k+1) : n])-k*mean(x))/sqrt(var(x))
z[n-start+1,1 :6]<-quantile(y,c(0.8,0.9,0.95,0.99,0.995,0.999))}}
z}
```