مجله علوم آماری، بهار و تابستان ۱۳۹۹
جلد ۱۴، شماره ۱، ص ۲۳۳ – ۲۵۰
DOI: 10.29252/jss.14.1.233

آزمونی برای استقلال در داده‌های نرمال با بُعد بالا

داریوش نجارزاده

گروه آمار، دانشکده علوم ریاضی، دانشگاه تبریز

تاریخ دریافت: ۱۳۹۷/۱۰/۲۱ تاریخ پذیرش: ۱۳۹۸/۱۰/۲۰

چکیده: فرضیه استقلال کامل داده‌ها پیش‌نیاز بسیاری از استنباط‌های آماری است. روش‌های آزمون کلاسیک پاسخ‌گوی بررسی چنین فرضی در داده‌های با ابعاد بالا نیست. در این مقاله آماره آزمونی ساده برای وجوداستقلال کامل در داده‌های نرمال با بُعد بالا معرفی و با استفاده از نظریه مارتینگل‌ها مجانباً نرمال بودن توزیع این آماره ثابت شده است. به منظور ارزیابی عملکرد آزمون پیشنهادی و مقایسه آن با روش‌های موجود، مطالعه‌ای شبیه‌سازی انجام شده است. نتایج شبیه‌سازی نشان می‌دهد که آزمون پیشنهادی دارای خطای نوع اول تجربی با میانگین خطای نسبی کوچکتر نسبت به آزمون‌های موجود است. کاربردی از روش پیشنهادی بر مجموعه داده سطح بیان ژن تومور پروستات معرفی شده است.
واژه‌های کلیدی: آزمون استقلال کامل، توزیع نرمال چندمتغیره، داده‌های با بُعد بالا، نظریه مارتینگل.

۱ مقدمه

استنباط‌های آماری چند متغیره کلاسیک برای داده‌های با ابعاد بالا قابل استفاده نیستند (ساراناداسا و بای ، ۱۹۹۶؛ دمپستر ، ۱۹۵۸، ۱۹۶۰). چرا که بیشتر این آزمون‌ها بر این اساس طراحی شده‌اند که بُعد p ثابت و در مقایسه با اندازه نمونه n کوچک است. برای داده‌های با بُعد بالا این شرط دیگر برقرار نیست، چرا که در چنین داده‌هایی بُعد در مقایسه با اندازه نمونه به مراتب عدد بزرگتری است. نمونه‌هایی از چنین داده‌ها را می‌توان در داده‌های ریزآرایه[1]، زیستی، مالی، و چندرسانه‌ای یافت (ژیراد ، ۲۰۱۴). در

آدرس الکترونیکی نویسنده مسئول مقاله: داریوش نجارزاده، d_najarzadeh@tabrizu.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H15، 62H10، 60F05

[1] Microarray

ادبیات چندمتغیره با بُعد بالا، همواره آزمون استقلال بردارهای تصادفی دارای توزیع نرمال چندمتغیره از اهمیت خاصی برخوردار بوده است. چن و مودهولکار (۱۹۹۰) توزیع مجموع توان‌های دوم تبدیل‌های z فیشر برای آزمون فرضیه استقلال کامل را بررسی کردند. سیرواستاوا (۲۰۰۵، ۲۰۰۶) آزمون‌هایی در مورد ماتریس کوواریانس و از جمله آزمون استقلال در داده‌های نرمال با بُعد بالا را معرفی کرد. رید و سیرواستاوا (۲۰۱۲) آزمون‌هایی برای استقلال زیربردارهای یک بردار نرمال با بُعد بالا را مورد مطالعه قرار دادند. آزمون همزمان استقلال برای زیربردارهای چند بردار نرمال با بُعد نسبتاً بالا توسط نجارزاده (۱۳۹۸) معرفی شد. آزمون‌های نسبت‌درست‌نمایی در رابطه با ماتریس کوواریانس داده‌های نرمال با بُعد بالا توسط جیانگ و همکاران (۲۰۱۲) و جیانگ و کی (۲۰۱۵) معرفی شد. جیانگ و یانگ (۲۰۱۳) قضیه‌های حدیِ مرکزی برای آزمون‌های نسبت‌درست‌نمایی کلاسیک، از جمله آزمون استقلال، در جوامع نرمال با بُعد بالا را معرفی کردند.

این آزمون به صورت زیر معرفی می‌شود. فرض کنید $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ نمونه‌ای تصادفی از توزیع نرمال p-متغیره با ماتریس کوواریانس Σ است. در این حالت، فرضیه استقلال کامل به صورت

$$H_0 : \rho_{ij} = 0, \quad 1 \leq i < j \leq p, \tag{1}$$

تعریف می‌شود، که در آن ρ_{ij} درایه (i,j)ام ماتریس همبستگی

$$\boldsymbol{\rho} = (\mathrm{diag}\Sigma^{-\frac{1}{2}})\Sigma(\mathrm{diag}\Sigma^{-\frac{1}{2}}) = [\rho_{ij}]_{p\times p}$$

است، که در آن $\mathrm{diag}(\Sigma)$ ماتریس قطری با درایه‌های روی قطر اصلی ماتریس Σ است. فرض کنید S و $R = (\mathrm{diag}(S)^{-\frac{1}{2}})S(\mathrm{diag}(S)^{-\frac{1}{2}}) = [r_{ij}]_{p\times p}$ به ترتیب ماتریس‌های کوواریانس و همبستگی نمونه‌ای حاصل از $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ باشند. در این صورت، بر اساس آزمون نسبت درستنمایی فرضیه H_0 رد می‌شود اگر مقدار آماره

$$-\big(n - \frac{2p+11}{6}\big)\log(\det(R))$$

از صدک $100(1-\alpha)$ درصد توزیع $\chi^2_{p(p-1)/2}$ بزرگ‌تر باشد (مؤرهد ، ۲۰۰۵، قضیه ۵.۲.۱۱) که در آن $\det(R)$ دترمینان ماتریس R است. این آزمون برای داده‌های با ابعاد بالا که در آنها $p > n$، غیر قابل تعریف است، زیرا ماتریس همبستگی نمونه‌ای R به ازای $p > n$، ماتریسی تکین با $\det(R)$ صفر است. به منظور ارائه آزمونی مناسب در داده‌های با ابعاد بالا برای فرضیه (۱) پژوهش‌های انجام شده

است. در این راستا، سیرواستاوا (۲۰۰۵) ثابت کرد که تحت فرضیه (۱) وقتی p و n هر دو به بینهایت میل می‌کنند، آماره آزمون به صورت

$$T^* = \frac{(n-1)t_{np} - \frac{p(p-1)}{2}}{\sqrt{p(p-1)}} \xrightarrow{d} N(0,1),$$

است، که در آن $t_{np} = \sum_{i=2}^{p}\sum_{j=1}^{i-1} r_{ij}^2$ و $\xrightarrow{d}$ نماد همگرایی در توزیع است. سیرواستاوا (۲۰۰۵) نشان داد همگرایی توزیع T^* کند است، برای غلبه بر این مشکل، آماره آزمون

$$T = \frac{(n-1)(\hat{\gamma}_3 - 1)}{2\sqrt{1 - \frac{\hat{a}_{40}}{p\hat{a}_{20}^2}}}$$

را پیشنهاد داد که در آن

$$\hat{\gamma}_3 = \frac{(n-1)[tr(S^2) - tr(S)^2/(n-1)]}{(n-2)tr(\mathrm{diag}(S)^2)}, \qquad \hat{a}_{20} = \frac{n-1}{p(n+1)}tr(\mathrm{diag}(S)^2),$$

و $\hat{a}_{40} = \frac{1}{p}tr(\mathrm{diag}(S)^4)$. با فرض اینکه به ازای هر $i = 1,\ldots,8$ حد $\frac{tr(\Sigma^i)}{p}$ وقتی $p \to \infty$ موجود و مثبت است و همچنین $n = O(p^\delta)$ برای $0 < \delta \leq 1$، سیرواستاوا (۲۰۰۵) نشان داد که $T \xrightarrow{d} N(0,1)$ وقتی $p \to \infty$ و $n \to \infty$. از آنجا که T بر اساس ماتریس کوواریانس نمونه‌ای S ساخته می‌شود، تحت تبدیلات مقیاس از متغیرها ناوردا [2] نخواهد بود. اسکات (۲۰۰۵) با در نظر گرفتن آماره t_{np}، نشان داد که تحت فرضیه H_0 و این فرض که $\lim_{n,p\to\infty}(\frac{p}{n}) \to \gamma \in (0,\infty)$، همواره

$$t_{np}^* = \frac{t_{np} - \frac{p(p-1)}{2(n-1)}}{\sigma_{np}^2} \xrightarrow{d} N(0,1),$$

[2] Invariance

که در آن $\sigma_{np}^{۲}=\frac{p(p-۱)(n-۲)}{(n-۱)^{۲}(n+۱)}$. با الهام از آزمون اسکات (۲۰۰۵)، مائو (۲۰۱۴) آماره

$$T_{np}=\sum_{i=۲}^{p}\sum_{j=۱}^{i-۱}\frac{r_{ij}^{۲}}{۱-r_{ij}^{۲}}$$

را برای آزمون فرضیه (۱) معرفی کرده و نشان داد وقتی $\lim_{n,p\to\infty}(\frac{p}{n})\to\gamma\in(۰,\infty)$، آنگاه

$$T_{np}^{*}=\frac{T_{np}-\frac{p(p-۱)}{۲(n-۴)}}{\delta_{np}^{۲}}\xrightarrow{d}N(۰,۱),$$

که در آن $\delta_{np}^{۲}=\frac{p(p-۱)(n-۳)}{(n-۴)^{۲}(n-۶)}$. اخیراً، چانگ و کویی (۲۰۱۸) فرض $\lim_{n,p\to\infty}(\frac{p}{n})\to\gamma\in(۰,\infty)$ بیان شده در آزمون‌های اسکات (۲۰۰۵) و مائو (۲۰۱۴) را حذف و ثابت کردند بدون در نظر گرفتن این قید و تنها با وجود فرض‌های $p\to\infty$ و $n\to\infty$ نیز توزیع مجانبی t_{np}^{*} و T_{np}^{*} تحت فرضیه صفر همچنان نرمال خواهد بود. بعلاوه نشان دادند که تحت فرضیه (۱)، توزیع مجانبی آماره‌های

$$t_{np}^{c}=\sqrt{p(p-۱)}t_{np}^{*}+\frac{p(p-۱)}{۲},$$
$$T_{np}^{c}=\sqrt{p(p-۱)}T_{np}^{*}+\frac{p(p-۱)}{۲},$$

کای‌دو با $p(p-۱)/۲$ درجه آزادی خواهد بود.

در بخش ۲، آماره‌ای جدید برای آزمون فرضیه (۱) معرفی و ثابت می‌شود تحت فرضیه صفر وقتی $p=p_n\to\infty$ و $n\to\infty$، دارای توزیع نرمال است. این آماره که بر اساس ماتریس همبستگی طراحی شده ناوردای مکان-مقیاس[۳] است. در بخش ۳، مطالعه شبیه‌سازی به منظور بررسی و مقایسه خطای نوع اول و همچنین توان آزمون پیشنهادی با آزمون‌های بر مبنای آماره‌های t_{np}^{*}، T_{np}^{*}، t_{np}^{c} و T_{np}^{c} ارائه می‌شود. کاربردی از آزمون پیشنهادی بر مجموعه داده‌های حاصل از یک مطالعه واقعی در بخش ۴ معرفی می‌شود. در بخش ۵ بحث و نتیجه گیری ارائه می‌شود.

[3] Location-scale invariance

۲ آماره آزمون و توزیع مجانبی آن

اسکات (۲۰۰۵) آماره $t_{np}=\sum_{i=۲}^{p}\sum_{j=۱}^{i-۱} r_{ij}^{۲}$ و مائو (۲۰۱۴) آماره $T_{np}=\sum_{i=۲}^{p}\sum_{j=۱}^{i-۱}\frac{r_{ij}^{۲}}{۱-r_{ij}^{۲}}$ را برای آزمون فرضیه (۱) پیشنهاد دادند. شهود به کار رفته در آماره‌های فوق و این نکته که $\sum_{i=۲}^{p}\sum_{j=۱}^{i-۱}\sqrt{|\rho_{ij}|}=۰$ اگر و تنها اگر $\rho_{ij}=۰$، ما را به آماره‌ای دیگر برای آزمون فرضیه $H_۰$ به صورت

$$D_{np}=\sum_{i=۲}^{p}\sum_{j=۱}^{i-۱}\sqrt{|r_{ij}|}$$

رهنمون ساخت. توجه داریم که تحت فرضیه $\rho_{ij}=۰$، همواره

$$E|r_{ij}|^{k}=E\big[(r_{ij}^{۲})^{\frac{k}{۲}}\big]=\frac{\Gamma\big(\frac{n-۱}{۲}\big)\Gamma\big(\frac{k+۱}{۲}\big)}{\sqrt{\pi}\Gamma\big(\frac{k+n-۱}{۲}\big)}=\Gamma_n(k), \qquad (۲)$$

(مؤرهد ، ۲۰۰۵). بنابراین تحت فرضیه $H_۰$

$$ED_{np}=\sum_{i=۲}^{p}\sum_{j=۱}^{i-۱}E\sqrt{|r_{ij}|}=\frac{p(p-۱)}{۲}\Gamma_n(\tfrac{۱}{۲})=\mu_{np}.$$

و می‌توان نوشت $r_{ij}=\boldsymbol{U}_i^T\boldsymbol{U}_j$، (اسکات ، ۲۰۰۵)، که در آن $\boldsymbol{U}_۱,\ldots,\boldsymbol{U}_p$ بردارهای تصادفی مستقل با توزیع یکنواخت روی سطح گوی $(n-۱)$-بُعدی با شعاع یک، $S_{n-۱}(۱)$، و دارای تابع چگالی احتمال

$$f_{\boldsymbol{U}}(\boldsymbol{u})=\frac{\Gamma\big(\frac{n-۱}{۲}\big)}{۲\pi^{\frac{n-۱}{۲}}}I\big(\boldsymbol{u}^T\boldsymbol{u}-۱\big),\quad \boldsymbol{u}\in R^{n-۱},$$

هستند (فولند ، ۲۰۱۳)، که در آن $I(x)$ تابع نشانگر روی نقطه $x=۰$ است. بنابراین با توجه به نحوه نمایش $r_{ij}=\boldsymbol{U}_i^T\boldsymbol{U}_j$ برای $i_۱\neq i_۲$ و $j_۱\neq j_۲$ متغیرهای تصادفی $r_{i_۱j_۱}$ و $r_{i_۲j_۲}$ مستقل‌اند. در حالتی که تنها یکی از $i_۲\neq i_۱$ یا $j_۲\neq j_۱$ برقرار باشد، $r_{i_۱j_۱}$ و $r_{i_۲j_۲}$ حداقل به شکل مجانبی مستقل هستند (سیرواستاوا ، ۲۰۰۵). چون نتایج به صورت مجانبی بدست آمده است، فرض می‌شود برای حداقل یکی از دو حالت $i_۲\neq i_۱$ یا $j_۲\neq j_۱$ متغیرهای تصادفی $r_{i_۱j_۱}$ و $r_{i_۲j_۲}$ همواره مستقل هستند.

لم ۱: فرض کنید $\boldsymbol{U} = [U_۱, \ldots, U_m]^T$ یک بردار تصادفی با توزیع یکنواخت روی $S_m(۱)$ باشد. در این صورت برای هر بردار ثابت $\boldsymbol{v} \in S_m(۱)$، داریم $\Gamma_{m+۱}(۱) = \big[|\boldsymbol{v}^T\boldsymbol{U}|\big]$.

برهان: از آنجا که $\boldsymbol{v}$ و $\boldsymbol{U}$ هر دو بردارهایی روی $S_m(۱)$ هستند، $\boldsymbol{v}^T\boldsymbol{v} = \boldsymbol{U}^T\boldsymbol{U} = ۱$. در نتیجه

$$\boldsymbol{v}^T\boldsymbol{U} = (\boldsymbol{v}^T\boldsymbol{v})(\boldsymbol{U}^T\boldsymbol{U})\cos(\theta) = \cos(\theta),$$

که در آن θ زاویه بین دو بردار مذکور است. بنابراین، $\boldsymbol{v}^T\boldsymbol{U}$ تنها به زاویه بین $\boldsymbol{v}$ و $\boldsymbol{U}$ و نه به مکان آنها روی $S_m(۱)$ وابسته است. از این رو می‌توان فرض کرد $\boldsymbol{v} = [۱, ۰, \ldots, ۰]^T \in S_m(۱)$. از سوی دیگر در مختصات قطبی می‌توان نوشت:

$$U_j = \begin{cases} \cos(\theta_j)\sin(\theta_۱)\cdots\sin(\theta_{j-۱}) & j = ۱, \ldots, m-۱ \\ \sin(\theta_{m-۱})\sin(\theta_۱)\cdots\sin(\theta_{m-۲}) & j = m \end{cases}$$

که در آن $\theta_j \in [۰, \pi]$، $j = ۱, \ldots, m-۲$ و $\theta_{m-۱} \in [۰, ۲\pi]$ (سزبووسکی ، ۱۹۹۸). در نتیجه

$$\begin{aligned} E\big[|\boldsymbol{v}^T\boldsymbol{U}|\big] &= \int_{\mathbb{R}^m} |u_۱| f_{\boldsymbol{U}}(\boldsymbol{u}) d\boldsymbol{U}. \\ &= \int_۰^{\pi}\cdots\int_۰^{\pi}\int_۰^{۲\pi} |\cos(\theta_۱)| \prod_{j=۱}^{m-۲} \sin^{m-j-۱}(\theta_j) \frac{\Gamma\big(\frac{m}{۲}\big)}{۲\pi^{\frac{m}{۲}}} d\theta_{m-۱}\cdots d\theta_۱ \\ &= \frac{\Gamma\frac{m}{۲}}{\pi^{\frac{m}{۲}-۱}} \big[\prod_{j=۲}^{m-۲}\int_۰^{\pi} \sin^{m-j-۱}(\theta_j) d\theta_j\big] \int_۰^{\pi} |\cos(\theta_۱)| \sin^{m-۲}(\theta_۱) d\theta_۱. \quad (۳) \end{aligned}$$

با تغییر متغیر $u = sin^۲(\theta_j)$ داریم:

$$\int_۰^{\frac{\pi}{۲}} \sin^{m-j-۱}(\theta_j) d\theta_j = \frac{۱}{۲}\int_۰^{۱} u^{\frac{m-j}{۲}-۱}(۱-u)^{\frac{۱}{۲}-۱} du = \frac{۱}{۲}\frac{\Gamma(\frac{m-j}{۲})\sqrt{\pi}}{\Gamma(\frac{m-j+۱}{۲})}.$$

$$\int_{\frac{\pi}{۲}}^{\pi} \sin^{m-j-۱}(\theta_j) d\theta_j = \frac{۱}{۲}\frac{\Gamma(\frac{m-j}{۲})\sqrt{\pi}}{\Gamma(\frac{m-j+۱}{۲})}.$$

در نتیجه

$$\int_{۰}^{\pi} \sin^{m-j-۱}(\theta_j)d\theta_j = \frac{\Gamma\big(\frac{m-j}{۲}\big)\sqrt{\pi}}{\Gamma\big(\frac{m-j+۱}{۲}\big)}. \qquad (۴)$$

دوباره با تغییر متغیر $u = sin^{۲}(\theta_۱)$ داریم:

$$\int_{۰}^{\frac{\pi}{۲}} \cos(\theta_۱)\sin^{m-j-۱}(\theta_۱)d\theta_۱ = \frac{۱}{m-۱} \quad \text{و} \quad \int_{\frac{\pi}{۲}}^{\pi} \cos(\theta_۱)\sin^{m-j-۱}(\theta_۱)d\theta_۱ = -\frac{۱}{m-۱}.$$

پس

$$\int_{۰}^{\pi} |\cos(\theta_۱)|\sin^{m-۲}(\theta_۱)d\theta_۱ = \frac{۲}{m-۱}. \qquad (۵)$$

حال با جایگذاری (۴) و (۵) در معادله (۳) داریم:

$$E\big[|\boldsymbol{v}^T\boldsymbol{U}|\big] = \frac{\Gamma\big(\frac{m}{۲}\big)}{\sqrt{\pi}\Gamma\big(\frac{m+۱}{۲}\big)} = \Gamma_{m+۱}(۱).$$

در نتیجه با استفاده از لم ۱ و اینکه $r_{i_۱j_۱}$ و $r_{i_۲j_۲}$ برای حداقل یکی از دو حالت $i_۱ \neq i_۲$ یا $j_۱ \neq j_۲$ همواره مستقل هستند، تحت فرضیه (۱) می‌توان نوشت:

$$\begin{aligned} \mathrm{Var}D_{np} &= \sum_{i=۲}^{p}\sum_{j=۱}^{i-۱} \mathrm{Var}\sqrt{|r_{ij}|} \\ &= \sum_{i=۲}^{p}\sum_{j=۱}^{i-۱} \big[E|r_{ij}| - \Gamma_n^{۲}(\tfrac{۱}{۲})\big] \\ &= \frac{p(p-۱)}{۲}\big[\Gamma_n(۱) - \Gamma_n^{۲}(\tfrac{۱}{۲})\big] = \tau_{np}^{۲}. \end{aligned} \qquad (۶)$$

قضیه ۱: تحت فرضیه (۱)، وقتی p و n به ∞ همگرا می‌شوند، آنگاه

$$D_{np}^{*} = \frac{D_{np} - \mu_{np}}{\tau_{np}} \xrightarrow{d} N(۰, ۱).$$

برهان: فرض کنید

$$d_{n\ell} = D_{n\ell} - \mu_{n\ell} = \sum_{i=۲}^{\ell} \big(\sum_{j=۱}^{i-۱} \sqrt{|r_{ij}|} - (i-۱)\Gamma_n(\tfrac{۱}{۲})\big)$$
$$X_{n\ell} = d_{n\ell} - d_{n(\ell-۱)} = \sum_{j=۱}^{\ell-۱} \sqrt{|r_{\ell j}|} - (\ell-۱)\Gamma_n(\tfrac{۱}{۲}), \quad \ell = ۲, ..., p,$$

که در آن $d_{n۱} = ۰$. چون تحت فرضیه $H_۰$ می‌توان نوشت $r_{ij} = \boldsymbol{U}_i^T \boldsymbol{U}_j$، با فرض $F_{n,\ell-۱} = \{\boldsymbol{U}_۱, \ldots, \boldsymbol{U}_{\ell-۱}\}$ و بنابه لم داریم ۱،

$$E\big[|r_{\ell j}||F_{n,\ell-۱}\big] = E\big[|\boldsymbol{U}_l^T \boldsymbol{U}_j||F_{n,\ell-۱}\big] = \Gamma_n(۱), \quad j = ۱, \ldots, l-۱$$

که در آن $E\big[\,\cdot\,|\,\cdot\,\big]$ امید ریاضی شرطی است. در نتیجه

$$E\big[X_{n\ell}|F_{n,\ell-۱}\big] = E\big[d_{n\ell}|F_{n,\ell-۱}\big] - d_{n(\ell-۱)} = ۰$$

یا $E\big[d_{n\ell}|F_{n,\ell-۱}\big] = d_{n(\ell-۱)}$. دنباله $\{d_{n\ell}, \ell = ۲, \ldots, k\}$ یک مارتینگل (هال و هی‌دی ، ۱۹۸۰) با تفاضل‌های $X_{n۲}, \ldots, X_{nk}$ است. علاوه بر این چون τ_{np} مقداری ثابت است، به طور مشابه برای $n \geq ۲$، دنباله $\{\frac{d_{n\ell}}{\tau_{np}}, \ell = ۲, \ldots, k\}$ نیز یک مارتینگل با تفاضل‌های $Y_{n۲} = \frac{X_{n۲}}{\tau_{np}}, \ldots, Y_{nk} = \frac{X_{nk}}{\tau_{np}}$ است. حال بنابر قضیه ۲.۳ در مک‌لیش (۱۹۷۴) برای اینکه

$$D_{np}^* = \sum_{\ell=۲}^{p} Y_{n\ell} \xrightarrow{d} N(۰, ۱)$$

کافیست ثابت شود وقتی $p, n \to \infty$، آنگاه

$$\sum_{\ell=۲}^{p} E\big[Y_{n\ell}^۴\big] \to ۰ \quad \text{و} \quad E\big(\sum_{\ell=۲}^{p} Y_{n\ell}^۲ - ۱\big)^۲ \to ۰.$$

حال تحت فرضیه (۱) با استفاده از (۲) داریم:

$$\begin{aligned}
E[X_{n\ell}^{۲}] &= E\{[\sum_{j=۱}^{\ell-۱}\sqrt{|r_{\ell j}|-(\ell-۱)\Gamma_n(\tfrac{۱}{۲})}]^{۲}\} \\
&= \sum_{j=۱}^{\ell-۱}E|r_{\ell j|}+\sum_{j_۱=۱}^{\ell-۱}\sum_{\substack{j_۲=۱\\ j_۲\neq j_۱}}^{\ell-۱}E\sqrt{|r_{\ell j_۱}|}E\sqrt{|r_{\ell j_۲}|}-\big[(\ell-۱)\Gamma_n(\tfrac{۱}{۲})\big]^{۲} \\
&= (\ell-۱)\Gamma_n(۱)+\Gamma_n^{۲}(\tfrac{۱}{۲})(\ell-۱)(\ell-۲)-\big((l-۱)\Gamma_n(\tfrac{۱}{۲})\big)^{۲} \\
&= (\ell-۱)\big[\Gamma_n(۱)-\Gamma_n^{۲}(\tfrac{۱}{۲})\big]
\end{aligned}$$

که در آن برابری سوم از (۲) و رابطه

$$\begin{aligned}
\sum_{j_۱=۱}^{\ell-۱}\sum_{\substack{j_۲=۱\\ j_۲\neq j_۱}}^{\ell-۱}E\sqrt{|r_{\ell j_۱}|}E\sqrt{|r_{\ell j_۲}|} &= \sum_{j_۱=۱}^{\ell-۱}\sum_{j_۲=۱}^{\ell-۱}\underbrace{E\sqrt{|r_{\ell j_۱}|}E\sqrt{|r_{\ell j_۲}|}}_{\Gamma_n^{۲}(\frac{۱}{۲})}I(j_۲\neq j_۱) \\
&= \Gamma_n^{۲}(\tfrac{۱}{۲})\sum_{j_۱=۱}^{\ell-۱}\sum_{j_۲=۱}^{\ell-۱}(۱-I(j_۲=j_۱)) \\
&= \Gamma_n^{۲}(\tfrac{۱}{۲})\left[(\ell-۱)^{۲}-(\ell-۱)\right] \\
&= \Gamma_n^{۲}(\tfrac{۱}{۲})(\ell-۱)(\ell-۲).
\end{aligned}$$

بدست آمده است. پس $\sum_{\ell=۲}^{p}EX_{n\ell}^{۲}=\tau_{np}^{۲}$. فرض کنید $U_{\ell j}=\sqrt{|r_{\ell j}|-\Gamma_n(\frac{۱}{۲})}$، به گونه‌ای که $j=۱,\ \ldots,\ell-۱$ و $EU_{\ell j}=۰$. با توجه به آنچه که در مورد استقلال $r_{\ell j_۱}$ و $r_{\ell j_۲}$ برای $j_۱\neq j_۲$ تحت فرضیه $H_۰$ بیان شد، $U_{\ell j_۱}$ و $U_{\ell j_۲}$ حداقل مجانباً مستقل خواهند بود. حال از آنجا که $X_{n\ell}=\sum_{j=۱}^{\ell-۱}U_{\ell j}$ می‌توان نشان داد

$$X_{n\ell}^{۴}=\sum_{j=۱}^{\ell-۱}U_{\ell j^{۴}}+۶\sum_{j_۱=۲}^{\ell-۱}\sum_{j_۲=۱}^{j_۱-۱}U_{\ell j_۱}^{۲}U_{\ell j_۲}^{۲}+U_{۰},$$

که در آن $U_{\circ}$، جمله‌ای با $EU_{\circ} = \circ$ تحت فرضیه صفر است. پس

$$EX_{n\ell}^{۴} = \sum_{j=۱}^{\ell-۱} EU_{\ell j}^{۴} + ۶ \sum_{j_۱=۲}^{\ell-۱} \sum_{j_۲=۱}^{j_۱-۱} E\big[U_{\ell j_۱}^{۲}\big] E\big[U_{\ell j_۲}^{۲}\big]. \qquad (۷)$$

اما بر اساس رابطه (۲) می‌توان نوشت

$$\begin{aligned} EU_{\ell j}^{۴} &= E\big(\sqrt{|r_{\ell j}|} - \Gamma_n(\tfrac{۱}{۲})\big)^{۴} \\ &= E|r_{\ell j}|^{۲} - ۴\Gamma_n(\tfrac{۱}{۲}) E|r_{\ell j}|^{\frac{۳}{۲}} + ۶\Gamma_n^{۲}(\tfrac{۱}{۲}) E|r_{\ell j}| - ۴\Gamma_n^{۳}(\tfrac{۱}{۲}) E\sqrt{|r_{\ell j}|} + \Gamma_n^{۴}(\tfrac{۱}{۲}) \\ &= \Gamma_n(۲) - ۴\Gamma_n(\tfrac{۱}{۲})\Gamma_n(\tfrac{۳}{۲}) + ۶\Gamma_n^{۲}(\tfrac{۱}{۲})\Gamma_n(۱) - ۳\Gamma_n^{۴}(\tfrac{۱}{۲}) = c_{۴} \qquad (۸) \end{aligned}$$

$$\begin{aligned} EU_{\ell j}^{۲} &= E\big(\sqrt{|r_{\ell j}|} - \Gamma_n(\tfrac{۱}{۲})\big)^{۲} \\ &= E|r_{\ell j}| - ۲\Gamma_n(\tfrac{۱}{۲}) E\sqrt{|r_{\ell j}|} + \Gamma_n^{۲}(\tfrac{۱}{۲}) \\ &= \Gamma_n(۱) - \Gamma_n^{۲}(\tfrac{۱}{۲}) = c_{۲}. \qquad (۹) \end{aligned}$$

از جایگذاری (۸) و (۹) در رابطه (۷)

$$EX_{n\ell}^{۴} = (\ell - ۱)c_{۴} + ۳c_{۲}^{۲}(\ell - ۱)(\ell - ۲), \quad \ell = ۲, \ldots, k.$$

در نتیجه

$$\sum_{\ell=۲}^{p} E\big[X_{n\ell}^{۴}\big] = \frac{p(p-۱)}{۲}\big(c_{۴} + ۲(p-۲)c_{۲}^{۲}\big) = \frac{\big(c_{۴} + ۲(p-۲)c_{۲}^{۲}\big)}{\frac{p(p-۱)}{۲}\big[\Gamma_n(۱) - \Gamma_n^{۲}(\frac{۱}{۲})\big]^{۲}}.$$

حال بنابه لم ۱.۵ در جیانگ و یانگ (۲۰۱۳)

$$\frac{\Gamma\big(\frac{n-۱}{۲}\big)}{\Gamma\big(\frac{k+n-۱}{۲}\big)} = \big(\frac{n-۱}{۲}\big)^{-\frac{k}{۲}} + o(۱),$$

وقتی $n \to \infty$. در نتیجه،

$$\Gamma_n(k) = \frac{\Gamma\big(\frac{n-1}{2}\big)\Gamma\big(\frac{k+1}{2}\big)}{\sqrt{\pi}\Gamma\big(\frac{k+n-1}{2}\big)} = \frac{\Gamma\big(\frac{k+1}{2}\big)}{\sqrt{\pi}}\big(\frac{n-1}{2}\big)^{-\frac{k}{2}} + o(1), \qquad (10)$$

وقتی $n \to \infty$. حال با به کارگیری (۱۰) در روابط (۸) و (۹)، وقتی $n \to \infty$، داریم:

$$c_4 = k_0\big(\frac{n-1}{2}\big)^{-1} + o(1) \quad \text{و} \quad c_2 = k_1\big(\frac{n-1}{2}\big)^{-\frac{1}{2}} + o(1),$$

که در آن k_0 و k_1 اعدادی ثابت و متناهی هستند. پس،

$$\sum_{\ell=2}^{p} E\big[Y_{n\ell}^4\big] = \frac{k_0 + 2(p-2){k_1}^2}{\frac{p(p-1)}{2}\big(\frac{1}{\sqrt{\pi}} - \frac{\Gamma^2(\frac{3}{4})}{\pi}\big)^2} + o(1) \to 0. \qquad (11)$$

وقتی p و n به بینهایت همگرا می‌شود. همچنین، از استقلال $Y_{n\ell_1}$ و $Y_{n\ell_2}$ برای $\ell_1 \neq \ell_2$، نتیجه می‌شود

$$\begin{aligned} 0 \leq E\big[\big(\sum_{\ell=2}^{p} Y_{n\ell}^2 - 1\big)^2\big] &= \sum_{\ell=2}^{p} E\big[Y_{n\ell}^4\big] + \sum_{\ell_1=2}^{p}\sum_{\substack{\ell_2=2 \\ \ell_2\neq\ell_1}}^{p} E\big[Y_{n\ell_1}^2 Y_{n\ell_2}^2\big] - 1 \\ &= \sum_{\ell_1=2}^{p} E\big[Y_{n\ell_1}^4\big] + \sum_{\ell_1=2}^{p}\sum_{\substack{\ell_2=2 \\ \ell_2\neq\ell_1}}^{p} E\big[Y_{n\ell_1}^2\big]E\big[Y_{n\ell_2}^2\big] - 1 \\ &\leq \sum_{\ell=2}^{p} E\big[Y_{n\ell}^4\big] + \big(\sum_{\ell=2}^{p} E\big[Y_{n\ell}^2\big]\big)^2 - 1 \\ &= \sum_{\ell=2}^{p} E\big[Y_{n\ell}^4\big]. \qquad (12) \end{aligned}$$

بنابراین، از روابط (۱۱) و (۱۲) خواهیم داشت

$$E\big[\big(\sum_{\ell=2}^{p} Y_{n\ell}^2 - 1\big)^2\big] \to 0. \qquad (13)$$

در نهایت، بنابر قضیه ۲.۳ (مک‌لیش ، ۱۹۷۴)، وقتی $p, n \to \infty$، از روابط (۱۱) و (۱۳) نتیجه می‌شود

$$D^*_{np} = \frac{D_{np} - \mu_{np}}{\tau_{np}} \xrightarrow{d} N(۰, ۱).$$

۳ مطالعه شبیه‌سازی

در این بخش به منظور بررسی عملکرد آزمون D^*_{np} و همچنین مقایسه آن با آزمون‌های t^*_{np}، T^*_{np}، t^c_{np} و T^c_{np} مطالعه شبیه‌سازی انجام شده است. برآورد اندازه و توان برای هر یک از آزمون‌ها بر اساس ۱۰۰۰۰ تکرار نمونه‌گیری مستقل و در سطح معنی‌داری $\alpha = ۰/۰۵$ محاسبه شده است. در اینجا بُعد p و اندازه نمونه n مقادیری به ترتیب از مجموعه‌های $\{۵, ۵۰, ۵۰۰\}$ و $\{۶, ۶۰, ۳۰۰\}$ در نظر گرفته شده است. برای بررسی عملکرد آزمون‌ها از دیدگاه اندازه آزمون، معیار میانگین خطای نسبی به صورت

$$\text{ARE} = \frac{۱۰۰}{M\alpha} \sum_{j=۱}^{M} |\hat{\alpha}_j - \alpha|, \qquad (۱۴)$$

استفاده شده است، که در آن M برابر کلیه ترکیبات از مقادیر p و n (در اینجا، $M = ۹$) و $\hat{\alpha}_۱, \ldots, \hat{\alpha}_M$ مقادیر اندازه‌های تجربی یا برآوردشده متناظر با ترکیبات مد نظر است. مقادیر کوچک (۱۴) برای یک آزمون بیانگر عملکرد بهتر آن آزمون در حفظ سطح معنی‌داری α است. در این مطالعه به ازای هر مقدار از بُعد p، بردار میانگین برابر بردار صفر و ماتریس کوواریانس به صورت $\Sigma^{(\phi)} = (۱ - \phi)I_p + \phi J_p$ انتخاب شده است، که در آن $۰ < \phi < ۱$ مقداری ثابت، I_p ماتریس همانی $p \times p$ و $J_p = ۱_p ۱_p^T$ که ۱_p برابر بردار ستونی $p \times ۱$ از یک‌ها است. در حقیقت، مقدار پارامتر ϕ میزان گسیختگی از فرضیه $H_۰$ را نشان می‌دهد. در این مطالعه، ϕ عضوی از مجموعه $\{۰, ۰/۰۲, ۰/۰۵\}$ انتخاب شده است. برای هر ترکیب از مقادیر n، p و ϕ، فرایند نمونه‌گیری با اندازه نمونه n از توزیع نرمال p-متغیره $N_p(۰, \Sigma^{(\phi)})$ به تعداد ۱۰۰۰۰ بار تکرار و مقادیر اندازه و توان تجربی آزمون‌های t^*_{np}، T^*_{np}، t^c_{np}، T^c_{np} و D^*_{np} محاسبه شده است. برای نمونه‌های تحت فرضیه (۱) ($\phi = ۰$) اندازه‌های تجربی آزمون‌ها در جدول ۱ و برای نمونه‌های تحت فرضیه مقابل ($\phi = ۰/۰۲, ۰/۰۵$) توان‌های تجربی متناظر با این آزمون‌ها در جداول ۲ و ۳ آورده شده است. در حالتی که هر دو اندازه نمونه n و بُعد p اعداد بزرگی هستند، تمامی آزمون‌ها دارای اندازه‌های تجربی نزدیک به مقدار اسمی ۰/۰۵ هستند. برای داده‌های با بُعد بالا که در آن اندازه نمونه n

جدول ۱: اندازه‌های آزمون تجربی به ازای $\phi = ۰$

D^*_{np}	T^c_{np}	t^c_{np}	T^*_{np}	t^*_{np}	p	n
۰٫۰۵۱۱	۰٫۰۴۶۲	۰٫۰۴۶۵	۰٫۰۴۲۸	۰٫۰۴۴۶	۵	
۰٫۰۴۸۹	۰٫۱۰۷۴	۰٫۰۶۴۳	۰٫۱۳۴۳	۰٫۰۴۴۹	۵۰	۶
۰٫۰۴۷۳	۰٫۱۵۸۶	۰٫۰۶۵۶	۰٫۲۶۳۰	۰٫۰۴۱۴	۵۰۰	
۰٫۰۵۰۰	۰٫۰۵۳۱	۰٫۰۵۱۴	۰٫۰۴۵۶	۰٫۰۴۳۶	۵	
۰٫۰۵۱۲	۰٫۰۵۱۵	۰٫۰۵۱۱	۰٫۰۵۲۳	۰٫۰۵۰۳	۵۰	۶۰
۰٫۰۴۸۶	۰٫۰۵۲۰	۰٫۰۵۰۵	۰٫۰۵۱۴	۰٫۰۴۹۰	۵۰۰	
۰٫۰۴۹۸	۰٫۰۵۰۶	۰٫۰۴۹۶	۰٫۰۴۲۹	۰٫۰۴۲۲	۵	
۰٫۰۴۹۷	۰٫۰۵۴۷	۰٫۰۵۴۶	۰٫۰۵۰۹	۰٫۰۵۰۹	۵۰	۳۰۰
۰٫۰۵۱۰	۰٫۰۴۷۹	۰٫۰۴۷۰	۰٫۰۴۹۰	۰٫۰۴۹۱	۵۰۰	
۲٫۰۰۰۰	۴۰٫۸۴۴۴	۹٫۸۶۶۷	۷۱٫۴۶۶۷	۸٫۰۸۸۹	ARE	

در مقایسه با بُعد p بسیار کوچک است (به عنوان مثال، $n = ۶$ و $p = ۵۰۰$)، اندازه تجربی آزمون‌های T^*_{np} و T^c_{np} به ترتیب با مقادیر ۰٫۲۶۳۰ و ۰٫۱۵۸۶ بسیار بالاتر از سطح آزمون اسمی $\alpha = ۰٫۰۵$ است. این در حالی است که اندازه تجربی آزمون‌های t^*_{np} و به ویژه D^*_{np} نزدیک به مقدار اسمی $\alpha = ۰٫۰۵$ است. در این میان، برای ابعاد بالاتر، آزمون t^c_{np} در رابطه با حفظ سطح معنی‌داری اسمی ۰٫۰۵ نسبت به آزمون‌های t^*_{np} و D^*_{np} عملکرد ضعیف‌تر و نسبت به آزمون‌های T^*_{np} و T^c_{np} عملکرد بهتری دارد. با افزایش اندازه نمونه، هر پنج آزمون سطح معنی‌داری اسمی ۰٫۰۵ به خوبی حفظ نموده و اندازه‌های تجربی یکسانی را ارائه داده‌اند. به هر حال برای داده‌های با ابعاد بالا که در آنها اختلاف بُعد با اندازه نمونه اعدادی بسیار بزرگ است استفاده از آزمون‌های T^*_{np}، t^c_{np} و T^c_{np} توصیه نمی‌شود. همچنین با توجه به یافته‌های جدول ۱، روش آزمون معرفی‌شده در این مقاله مبنی بر آماره D^*_{np} در مقایسه با سایر آزمون‌ها از کمترین مقدار میانگین خطای نسبی با مقدار ARE برابر ۲ برخوردار است. به عبارت دیگر، اندازه آزمون تجربی روش معرفی شده در این مقاله در مقایسه با سایر روش‌های موجود به مقدار $\alpha = ۰٫۰۵$ نزدیک‌تر است.

در حالت کلی، با ملاک مقدار میانگین خطای نسبی، به لحاظ حفظ سطح معنی‌داری اسمی به ترتیب آزمون‌های D^*_{np}، t^*_{np}، t^c_{np}، T^c_{np} و T^*_{np} از اولویت برخوردارند. لازم به ذکر است که بر اساس شبیه‌سازی‌هایی که به دلیل حجم زیاد در این مقاله ذکر نشده‌اند نیز همواره اولویت‌بندی فوق حفظ شده است. با توجه به جداول ۲ و ۳ تمامی آزمون‌ها نسبت به انحرافات بسیار کوچک بسیار حساس‌اند به طوری که مثلا با انحرافی به اندازه $\phi = ۰٫۰۵$ از فرضیه صفر برای اندازه نمونه متوسط،

جدول ۲: توان‌های آزمون تجربی به ازای $\phi = ۰/۰۲$

D^*_{np}	T^c_{np}	t^c_{np}	T^*_{np}	t^*_{np}	p	n
۰/۰۵۲۹	۰/۰۴۶۱	۰/۰۴۷۰	۰/۰۴۲۷	۰/۰۴۴۵	۵	
۰/۰۵۱۶	۰/۱۱۳۳	۰/۰۷۰۲	۰/۱۴۰۱	۰/۰۴۹۸	۵۰	۶
۰/۰۹۹۸	۰/۲۰۲۸	۰/۱۴۰۵	۰/۲۶۹۶	۰/۱۰۵۷	۵۰۰	
۰/۰۵۱۰	۰/۰۶۰۵	۰/۰۵۸۳	۰/۰۵۳۱	۰/۰۵۱۵	۵	
۰/۰۹۲۸	۰/۱۵۸۲	۰/۱۵۵۷	۰/۱۱۴۷	۰/۱۱۲۴	۵۰	۶۰
۰/۹۲۳۲	۰/۹۷۳۵	۰/۹۷۲۶	۰/۹۶۲۴	۰/۹۶۱۳	۵۰۰	
۰/۰۶۷۴	۰/۰۹۱۱	۰/۰۹۰۱	۰/۰۸۰۰	۰/۰۷۸۸	۵	
۰/۶۳۶۱	۰/۸۱۶۷	۰/۸۱۶۵	۰/۷۵۱۳	۰/۷۵۱۰	۵۰	۳۰۰
۱	۱	۱	۱	۱	۵۰۰	

جدول ۳: توان‌های آزمون تجربی به ازای $\phi = ۰/۰۵$

D^*_{np}	T^c_{np}	t^c_{np}	T^*_{np}	t^*_{np}	p	n
۰/۰۵۲۱	۰/۰۴۶۳	۰/۰۵۱۳	۰/۰۴۳۶	۰/۰۴۶۵	۵	
۰/۰۷۸۹	۰/۱۳۸۷	۰/۱۱۱۰	۰/۱۴۷۶	۰/۰۷۹۳	۵۰	۶
۰/۳۸۳۰	۰/۴۱۸۹	۰/۴۵۸۸	۰/۴۳۱۰	۰/۴۰۸۳	۵۰۰	
۰/۰۷۴۳	۰/۱۰۸۱	۰/۱۰۴۱	۰/۰۹۷۷	۰/۰۹۵۱	۵	
۰/۷۰۲۴	۰/۸۳۰۹	۰/۸۳۰۲	۰/۷۷۹۷	۰/۷۷۸۳	۵۰	۶۰
۱	۱	۱	۱	۱	۵۰۰	
۰/۲۸۲۶	۰/۳۹۹۲	۰/۳۹۸۷	۰/۳۷۵۴	۰/۳۷۴۹	۵	
۱	۱	۱	۱	۱	۵۰	۳۰۰
۱	۱	۱	۱	۱	۵۰۰	

$n = ۶۰$، توان‌های تجربی نزدیک به یک برای ابعاد $p \geq ۵۰$ حاصل شده است. در ابعاد بالاتر به دلیل افزایش تعداد انحرافات هر چند کوچک از فرضیه صفر و در نتیجه بزرگ تر شدن مجموع این انحرافات توان‌های بالاتر از طرف تمامی آزمون‌ها به دست آمده است. به هر حال برای همین انحرافات کوچک، آزمون‌های T^c_{np} و T^*_{np} دارای توان بالاتری نسبت به آزمون‌های t^*_{np}، t^c_{np} و D^*_{np} هستند و در اکثر مواقع آماره آزمون معرفی شده دارای توان کمتری نسبت به سایر آزمون‌ها است. این امر با توجه به بالا بودن مقدار میانگین خطای نسبی سایر آزمون‌ها امری قابل پیش‌بینی است. در

حقیقت، آزمون‌هایی با اندازه‌های بزرگتر از سطح اسمی α همواره توان‌های بزرگ کاذب و فریبنده دارند (کریستنسن و رنچر ، ۱۹۹۷؛ جیانگ و یانگ ، ۲۰۱۳؛ ژانگ و همکاران ، ۲۰۱۷). با توجه به جداول ۲ و ۳، همواره افزایش اندازه نمونه یا بُعد یا هر دو منجر به توان‌های بالاتر از تمامی آزمون‌ها می‌شود.

۴ کاربرد روش پیشنهادی

این بخش به ارائه کاربردی از روش آزمون پیشنهادی بر مجموعه داده سطح بیان ژن تومور پروستات اختصاص یافته است[۴] (سینگ و همکاران ، ۲۰۰۲؛ دتلینگ و بوهمن ، ۲۰۰۲). این داده‌ها را می‌توان در بسته spls[۵] از نرم‌افزار R یافت. این مجموعه شامل اندازه‌گیری سطح بیان ۶۰۳۳ ژن به کمک فناوری اَفی‌متریکس[۶] روی دو نمونه با اندازه‌های $n_۱ = ۵۲$ و $n_۲ = ۵۰$ به ترتیب از بافت‌های تومور پروستات و بافت‌های سالم است. سینگ و همکاران (۲۰۰۲) بر اساس همبستگی‌های بین مقادیر سطوح بیان و همچنین همبستگی بین این سطوح با متغیرهای بالینی و آسیب‌شناختی (مانند سن بیمار، نتیجه آزمون‌های تشخیصی PSA و رتبه GS، غیره) به تشریح رفتار بالینی سرطان پروستات پرداختند.

در سطح $\alpha = ۰/۰۵$ استقلال همزمان $p = ۱۰۰۰$ سطح بیان ژن نخست از بین ۶۰۳۳ سطح بیان در هر دوجامعه بافت‌های سالم (I) و بافت‌های سرطانی (II) بررسی شده است. نتایج حاصل از آزمون‌های استقلال t^*_{np}، T^*_{np}، t^c_{np}، T^c_{np} و D^*_{np} روی دو جامعه در جدول ۴ آمده است. همان‌طور که ملاحظه

جدول ۴: آزمون استقلال در داده‌های تومور پرستات

مقدار آماره			
روش آزمون	بافت‌های سرطانی	بافت‌های سالم	مقدار بحرانی
$t^*_{n,p}$	۲۴۴۲/۶۰۲	۳۰۱۲/۱۰۴	۱/۹۵۹۹
$T^*_{n,p}$	$۹/۵۳۶۷ \times ۱۰^{۱۵}$	۳۶۵۷۵/۸۷	۱/۹۵۹۹
$t^c_{n,p}$	۱۳۴۴۸۲۹	۱۶۲۹۲۹۵	۱۲۵۵۷۲/۷
$T^c_{n,p}$	$۴/۷۶۳۶ \times ۱۰^{۱۸}$	۱۸۳۹۴۳۸۸	۱۲۵۵۷۲/۷
$D^*_{n,p}$	۷۳۸/۶۷۹۷	۸۳۲/۲۲۱۱	۱/۹۵۹۹

می‌شود فرضیه استقلال سطوح بیان این ۱۰۰۰ ژن در هر دو جامعه بافت‌های دارای تومور پروستات و بافت‌های سالم به شدت رد می‌شود و همبستگی خطی نسبتاً قوی بین سطوح بیان این ژن‌ها وجود دارد.

[4] Prostate Tumor Gene Expression dataset
[5] https://CRAN.R-project.org/package=spls
[6] Affymetrix technology

۵ بحث و نتیجه‌گیری

در این مقاله، آزمون استقلال در داده‌های نرمال با بُعد بالا بررسی و آزمونی جدید ارائه شد. با شبیه‌سازی، اندازه و توان آزمون پیشنهادی با آزمون‌های موجود محاسبه و نشان داده شد که آزمون معرفی شده به ویژه در حالت داده‌های با بُعد بالا، در مقایسه با دیگر آزمون‌ها دارای کمترین مقدار میانگین خطای نسبی است. در واقع خطای نوع اول واقعی سایر آزمون‌ها با مقدار اسمی α متفاوت است. چنین آزمون‌هایی به شکل کاذب دارای توان بالاتری هستند. آزمون پیشنهادی به دلیل اندازه آزمون واقعی نزدیک به سطح اسمی دارای کمترین مقدار توان آزمون در مقایسه با سایر آزمون‌ها بود. البته این اختلاف در توان با سایر آزمون‌ها، تنها برای انحرافات جزئی از فرضیه صفر بود و نه برای انحرافات میانگین و بالاتر که این حالت‌ها توان تمامی آزمون‌ها برابر می‌شد. برای داده‌های بُعد بالا، استفاده از آزمون پیشنهادی به دلیل مقدار میانگین خطای نسبی کم و به عبارت دیگر اندازه آزمون نزدیک به سطح اسمی α توصیه می‌شود.

مراجع

نجارزاده، د. (۱۳۹۸)، آزمون همزمان استقلال برای زیربردارهای چند بردار با بُعد نسبتاً بالای نرمال چندمتغیره، مجله علوم آماری ایران، **۱۳**، ۲۱۷-۲۳۳.

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis (3nd ed.)*, John Wiley & Sons, New York.

Bai, Z., and Saranadasa, H. (1996), Effect of High Dimension: by an Example of a Two Sample Problem, *Statistica Sinica*, **6**, 311-329.

Chang, S., and Qi, Y. (2018), On Schott's and Mao's Test Statistics for Independence of Normal Random Vectors, *Statistics & Probability Letters*, **140**, 132-141.

Chen, S., and Mudholkar, G. S. (1990), Null Distribution of the Sum of Squared Z-Transforms in Testing Complete Independence, *Annals of the Institute of Statistical Mathematics*, **42**, 149-155.

Christensen, W. F., and Rencher, A. C. (1997), A Comparison of Type I Error Rates and Power Levels for Seven Solutions to the Multivariate Behrens-Fisher Problem, *Communications in Statistics-Simulation and Computation*, **26**, 1251-1273.

Dempster, A. P. (1958), A High Dimensional Two Sample Significance Test, *The Annals of Mathematical Statistics*, 995-1010.

Dempster, A. P. (1960), A Significance Test for the Separation of Two Highly Multivariate Small Samples, *Biometrics*, **16**, 41-50.

Dettling, M., and Bühlmann, P. (2002). Supervised Clustering of Genes. *Genome biology* **3**, research0069-1.

Folland, G. B. (2013), *Real Analysis: Modern Techniques and Their Applications*, John Wiley & Sons, New York.

Giraud, C. (2014), *Introduction to High-Dimensional Statistics*, Chapman and Hall/CRC, Hoboken, New Jersey.

Gupta, A., and Song, D. (1997), Lp-Norm Spherical Distribution, *Journal of Statistical Planning and Inference*, **60**, 241-260.

Jiang, D., Jiang, T. and Yang, F. (2012), Likelihood Ratio Tests for Covariance Matrices of High-Dimensional Normal Distributions, *Journal of Statistical Planning and Inference*, **142**, 2241-2256.

Jiang, T., and Qi, Y. (2015), Likelihood Ratio Tests for High-Dimensional Normal Distributions, *Scandinavian Journal of Statistics*, **42**, 988-1009.

Jiang, T., and Yang, F. (2013), Central Limit Theorems for Classical Likelihood Ratio Tests for High-Dimensional Normal Distributions, *The Annals of Statistics,* **41**, 2029-2074.

Liang, J., Tang, M. L. and Chan, P. S. (2009), A Generalized Shapiro-Wilk W Statistic for Testing High-Dimensional Normality, *Computational Statistics & Data Analysis*, **53**, 3883-3891.

Mao, G. (2014), A New Test of Independence for High-Dimensional Data, *Statistics & Probability Letters*, **93**, 14-18.

McLeish, D. L. (1974), Dependent Central Limit Theorems and Invariance Principles, *The Annals of Probability*, **2**, 620-628.

Muirhead, R. J. (2005), *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, Inc, Hoboken, New Jersey.

Hall, P., and Heyde, C. C. (2014), *Martingale Limit Theory and Its Application*, Academic press, New York.

Schott, J. R. (2005), Testing for Complete Independence in High Dimensions, *Biometrika*, **92**, 951-956.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. and Lander, E. S. (2007). Gene Expression Correlates of Clinical Prostate Cancer Behavior, *Cancer cell*, **1**, 203-209.

Srivastava, M. S. (2005), Some Tests Concerning the Covariance Matrix in High Dimensional Data, *Journal of the Japan Statistical Society*, **35**, 251-272.

Srivastava, M. S. (2006), Some Tests Criteria for the Covariance Matrix with Fewer Observations than the Dimension, *Acta et Commentations Universitatis Tartuensis de Mathematica*, **10**, 77-93.

Srivastava, M. S., and Reid, N. (2012), Testing the Structure of the Covariance Matrix with Fewer Observations than the Dimension, *Journal of Multivariate Analysis*, **112**, 156-171.

Szabowski, P. (1998), Uniform Distributions on Spheres Infinite Dimensional and Their Generalizations, *Journal of Multivariate Analysis*, **64**, 103117.

Tan, M., Fang, H. B., Tian, G. L. and Wei, G. (2005), Testing Multivariate Normality in Incomplete Data of Small Sample Size, *Journal of Multivariate Analysis*, **93**, 164-179.

Zhang, J. T., Guo, J. and Zhou, B. (2017), Linear Hypothesis Testing in High-dimensional One-way MANOVA, *Journal of Multivariate Analysis*, **155**, 200-216.