آزمون پرت بودن گروهی از مشاهدات چند متغیره

اباذر خلجی [۱]، **منوچهر خردمندنیا** [۲]

چکیده:

فرض کنید m نمونه تصادفی n تایی از توزیع $N_p(\boldsymbol{\mu},\Sigma)$ داریم که مستقلند و می‌خواهیم فرض پرت بودن i امین نمونه n تایی را بیازمائیم. چنین آزمونی در بررسی‌های فاز اول کنترل فرآیند آماری چند متغیره با مشاهدات گروهی می‌تواند مفید باشد. امروزه مشهور است یک آماره بتا برای چنین آزمونی وجود دارد و از طریق رابطه توزیع بتا با توزیع F از آماره F می‌توان استفاده نمود. در متون آماری اثبات روشن و دقیقی برای توزیع آماره آزمون یافت نمی‌شود و در مواردی نیز، اثبات نادرست یا ناقص است. در این مقاله اثبات دقیق و نسبتاً روشن ارائه می‌دهیم و از طریق شبیه سازی قابلیت‌ها و ضعف‌های آزمون را مورد بررسی قرار می‌دهیم.

واژه‌های کلیدی: نقاط پرت، کنترل فرآیند چند متغیره ، فاز اول، T^2 ی هتلینگ

۱ مقدمه

در مدل سازی آماری و تحلیل داده‌ها اغلب این پرسش مطرح می‌شود که آیا در مجموعه داده‌ها نقاط پرت وجود دارد؟ نقاط پرت مشاهده‌هایی هستند که از الگوی اکثر مشاهدات پیروی نمی‌کنند. وجود نقاط پرت در نمونه‌های یک یا چند متغیره باعث تخریب برآوردگرها می‌شود و علاوه بر آن در یک فرایند کنترل وجود مشاهدات غیر معمول، میزان تغییرات را افزایش داده و همبستگی بین متغیرها را در هم می‌ریزد.

به نظر می‌رسد اولین کار مهم در زمینه شناسایی نقاط پرت را ویلکس (۱۹٦۳) انجام داده است [۹] . سریواستاوا و ون‌روسن (۱۹۹۸) در حالت مشاهدات تکی نرمال چند متغیره یک آماره بتا برای آزمون پرت بودن مشاهده i ام بدست آوردند و با استفاده از رابطه توریع بتا با توریع F یک آماره F معرفی کردند [۷] . تعمیم آزمون پرت بودن از حالت مشاهدات تکی به مشاهدات گروهی در متون کنترل فرآیند آماری چند متغیره مورد توجه بوده است. به نظر می‌رسد که اولین تلاش مهم در این راستا توسط آلت (۱۹۷۳) انجام شده ولی ایشان به جای توزیع بتا اشتباهاً به توریع T^2ی هتلینگ رسیده است [۳] . تریسی و همکاران (۱۹۹۲) به آماره درست بتا رسیده‌اند ولی اثباتی که ارائه کرده‌اند مختصر است [۸].

در مقاله حاضر اثباتی دقیق و نسبتاً روشن ارائه می‌دهیم.

یکی از کاربردهای گسترده آماره مورد مطالعه این مقاله، در کنترل فرآیندهای آماری در فاز اول است که در آن کنار گذاشتن مشاهدات غیر معمول به منظور اطمینان از تحت کنترل بودن فرآیند، مشابه برخورد با نقاط پرت می‌باشد [۱] .

۲ چند قضیه کلیدی

در این بخش ۳ قضیه کلیدی ارائه می‌شود که در بخش‌های بعد مورد استفاده قرار می‌گیرند.

قضیه ۱.۲. اگر $\overline{\boldsymbol{X}}$ و S به ترتیب بردار میانگین و ماتریس کوواریانس یک نمونه تصادفی n تایی از $N_p(\boldsymbol{\mu},\Sigma)$ باشد، آن‌گاه:

$$\overline{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu},\frac{1}{n}\Sigma) \quad \perp \quad (n-1)S \sim W_p(\Sigma,n-1)$$

که نماد $\perp$ بیانگر استقلال می‌باشد.

قضیه ۲.۲. اگر $F \sim F_{d_1,d_2}$ آنگاه

$$B=\frac{d_1F}{d_2+d_1F} \sim Beta(\frac{d_1}{2},\frac{d_2}{2})$$

قضیه ۳.۲. اگر $\boldsymbol{Y} \sim N_p(0,\Sigma)$ مستقل از $W \sim W_p(\Sigma,p)$ باشد، آن‌گاه:

$$T^2=k\boldsymbol{Y}^TW^{-1}\boldsymbol{Y} \sim \frac{kp}{k-p+1}F_{p,k-p+1}$$

[۱] دانشجو کارشناسی ارشد آمار دانشگاه اصفهان

[۲] استادیار گروه آمار دانشگاه اصفهان

[Downloaded from andisheyeamari.irstat.ir on 2025-05-13]

$$F = \frac{k-p+۱}{p} \boldsymbol{Y}^T W^{-۱} \boldsymbol{Y} \sim F_{p,k-p+۱}$$

برای اثبات قضایای فوق به کتاب خردمندنیا و علیرضایی [۲] مراجعه کنید.

قضیه ٤.۲. فرض کنید $B = \boldsymbol{Y}^T W^{-۱} \boldsymbol{Y}$. که در آن بردار $\boldsymbol{Y}$ دارای توزیع $N_p(۰, \Sigma)$ و ماتریس W دارای توزیع $W_p(\Sigma, k)$ می‌باشد. اگر بتوان نوشت $W = W_۱ + \boldsymbol{Y}\boldsymbol{Y}^T$ به طوری که $W_۱ \sim W_p(\Sigma, k-۱)$ و مستقل از بردار $\boldsymbol{Y}$ باشد، آنگاه [۲]:

$$B \sim Beta(\frac{p}{۲}, \frac{k-p}{۲})$$

اثبات. با استفاده از فرمول وارون جمع دو ماتریس (سیرل ۱۹۸۲[٦]) می‌توان نوشت:

$$\begin{aligned} B = \boldsymbol{Y}^T W^{-۱} \boldsymbol{Y} &= \boldsymbol{Y}^T (W_۱ + \boldsymbol{Y}\boldsymbol{Y}^T)^{-۱} \boldsymbol{Y} \\ &= \boldsymbol{Y}^T \left[W_۱^{-۱} - \frac{W_۱^{-۱} \boldsymbol{Y}\boldsymbol{Y}^T W_۱^{-۱}}{۱ + \boldsymbol{Y}^T W_۱^{-۱} \boldsymbol{Y}} \right] \boldsymbol{Y} \\ &= \frac{\boldsymbol{Y}^T W_۱^{-۱} \boldsymbol{Y}}{۱ + \boldsymbol{Y}^T W_۱^{-۱} \boldsymbol{Y}} \end{aligned}$$

با توجه قضیه ۳.۲ داریم:

$$F = \frac{k-p}{p} \boldsymbol{Y}^T W_۱^{-۱} \boldsymbol{Y} \sim F_{p,k-p}$$

بنابراین با استفاده از قضیه ۲.۲ می‌توان نوشت:

$$B = \frac{pF}{(k-p)+pF} \sim Beta(\frac{p}{۲}, \frac{k-p}{۲})$$

□

۳ آزمون پرت بودن یک مشاهده تکی

در این بخش ابتدا حالتی که مجموعه داده‌ها تنها دارای یک مشاهده پرت است، مورد مطالعه قرار می‌گیرد و علاقمند به بررسی پرت بودن مشاهده i ام هستیم لذا فرض‌های آماری زیر را در نظر گرفته آماره آزمون را همراه با اثبات دقیق ارائه می‌دهیم.

$$\begin{cases} H_۰ : \text{پرت نبودن مشاهده } i \text{ ام} \\ H_۱ : \text{پرت بودن مشاهده } i \text{ ام} \end{cases}$$

برای این منظور فرض می‌کنیم $\overline{\boldsymbol{X}}$ و S به ترتیب بردار میانگین و ماتریس کوواریانس یک نمونه تصادفی m تایی از $N_p(\boldsymbol{\mu}, \Sigma)$ باشد. فرض کنید $\overline{\boldsymbol{X}}_{(i)}$ و $S_{(i)}$ به ترتیب بردار میانگین و ماتریس کوواریانس پس از حذف مشاهده i ام باشد، می‌توان نوشت :

$$\overline{\boldsymbol{X}}_{(i)} = \frac{m}{m-۱} \overline{\boldsymbol{X}} - \frac{۱}{m-۱} \boldsymbol{X}_i$$

$$(m-۱)S = (m-۲)S_{(i)} + \frac{m}{m-۱} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T$$

بر اساس قضیه ۱.۲ $\overline{\boldsymbol{X}}_{(i)} \sim N_p(\boldsymbol{\mu}, \Sigma)$ و مستقل از $(m-۲)S_{(i)} \sim W_p(\Sigma, m-۲)$ است.

$$\left.\begin{array}{l} S_{(i)} \perp \overline{\boldsymbol{X}}_{(i)} \\ S_{(i)} \perp \boldsymbol{X}_i \end{array}\right\} \Rightarrow S_{(i)} \perp (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{(i)})$$

با توجه به اینکه $\sqrt{\frac{m}{m-۱}}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}) = \sqrt{\frac{m-۱}{m}}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{(i)})$ می‌توان نوشت :

$$(m-۲)S_{(i)} \perp \sqrt{\frac{m}{m-۱}} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})$$

اکنون تمامی شرایط برای استفاده از قضیه ٤.۲ بر قرار است. در واقع در این جا با علائم بکار رفته در قضیه ٤.۲ داریم $W = (m-۱)S$ ، $W_۱ = (m-۲)S_{(i)}$ و $\boldsymbol{Y} = \sqrt{\frac{m}{m-۱}}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})$ که $W_۱$ مستقل از $\boldsymbol{Y}$ است. حال با توجه به این قضیه می‌توان نوشت:

$$B_i = \frac{m}{(m-۱)^۲} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T S^{-۱} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})$$

$$B_i \sim Beta(\frac{p}{۲}, \frac{m-p-۱}{۲})$$

که $i = ۱, ..., m$. حال با توجه به قضیه ۲.۲ در نهایت داریم :

$$F_i = (\frac{m-p-۱}{p}) \frac{B_i}{۱ - B_i} \sim F_{p,m-p-۱}$$

اگر مقدار آماره F_i از $F_{\alpha,p,m-p-۱}$ بزرگتر باشد می‌توان نتیجه گرفت که فاصله $\boldsymbol{X}_i$ به طور معنی‌داری از میانگین مشاهدات دور است و لذا فرض $H_۰$ مبنی بر پرت نبودن مشاهده i ام رد می‌شود. لازم به ذکر است که آماره فوق توسط سریواستاوا و ون‌روسن (۱۹۹۸) به عنوان آماره‌ای برای شناسایی یک نقطه پرت در مجموعه داده‌های نرمال چند متغیره معرفی شده است [۷]. در واقع با توجه به مطالب فوق در این بخش قضیه زیر را اثبات کردیم.

قضیه ۱.۳. اگر $\boldsymbol{X}_۱، \boldsymbol{X}_۲، ..., \boldsymbol{X}_m$ یک نمونه تصادفی از $N_p(\boldsymbol{\mu}, \Sigma)$ باشد، آن‌گاه:

$$B_i = \frac{m}{(m-۱)^۲} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T S^{-۱} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})$$

[Downloaded from andisheyeamari.irstat.ir on 2025-05-13]

آن‌گاه:

$$B_i \sim Beta(\frac{p}{۲}, \frac{m-p-۱}{۲})$$

٤ آزمون پرت بودن گروهی از مشاهدات چند متغیره

فرض کنید m نمونه تصادفی n تایی از $N_p(\boldsymbol{\mu}, \Sigma)$ داریم که مستقلند و می‌خواهیم فرض‌های زیر را بیازمائیم :

$$\begin{cases} H_۰ : \text{نمونه } n \text{ تایی } i \text{ ام پرت نیست} \\ H_۱ : \text{نمونه } n \text{ تایی } i \text{ ام پرت است} \end{cases}$$

فرض کنید $\overline{\boldsymbol{X}}_i$ بردار میانگین i امین نمونه n تایی است $(i = ۱, ..., m)$ و قرار دهید:

$$\overline{\overline{\boldsymbol{X}}} = m^{-۱} \sum_{i=۱}^{m} \overline{\boldsymbol{X}}_i$$

$$S_{\overline{\boldsymbol{X}}} = (m-۱)^{-۱} \sum_{i=۱}^{m} (\overline{\boldsymbol{X}}_i - \overline{\overline{\boldsymbol{X}}})(\overline{\boldsymbol{X}}_i - \overline{\overline{\boldsymbol{X}}})^T$$

همچنین فرض کنید $\overline{\overline{\boldsymbol{X}}}_{(i)}$ و $S_{(i)}$ به ترتیب بردار میانگین و ماتریس کوواریانس میانگین‌های نمونه‌ای پس از حذف $\overline{\boldsymbol{X}}_i$ باشند، می‌توان نوشت:

$$\overline{\overline{\boldsymbol{X}}}_{(i)} = \frac{m}{m-۱}\overline{\overline{\boldsymbol{X}}} - \frac{۱}{m-۱}\overline{\boldsymbol{X}}_i$$

$$(m-۱)S_{\overline{\boldsymbol{X}}} = (m-۲)S_{(i)} + \frac{m}{m-۱}(\overline{\boldsymbol{X}}_i - \overline{\overline{\boldsymbol{X}}})(\overline{\boldsymbol{X}}_i - \overline{\overline{\boldsymbol{X}}})^T$$

طبق قضیه ۱.۲ $\overline{\overline{\boldsymbol{X}}}_{(i)} \sim N_p(\boldsymbol{\mu}, \frac{۱}{n(m-۱)}\Sigma)$ و مستقل از $(m-۲)S_{(i)} \sim W_p(\frac{۱}{n}\Sigma, m-۲)$ است. با کمی دقت ملاحظه می‌شود که شرایطی همانند قضیه ۱.۳ برقرار است. بنابراین بدون این که لازم باشد اثبات جدیدی ارائه شود، صرفاً براساس قضیه ۱.۳ می‌توان قضیه زیر را نتیجه گرفت.

قضیه ۱.٤. اگر $\overline{\boldsymbol{X}}_۱$، $\overline{\boldsymbol{X}}_۲$،...، $\overline{\boldsymbol{X}}_m$ یک نمونه تصادفی از $N_p(\boldsymbol{\mu}, \frac{۱}{n}\Sigma)$ باشد، آنگاه :

$$B_i = \frac{m}{(m-۱)^۲}(\overline{\boldsymbol{X}}_i - \overline{\overline{\boldsymbol{X}}})S_{\overline{\boldsymbol{X}}}^{-۱}(\overline{\boldsymbol{X}}_i - \overline{\overline{\boldsymbol{X}}})^T$$

آن‌گاه:

$$B_i \sim Beta(\frac{p}{۲}, \frac{m-p-۱}{۲})$$

حال با توجه به قضیه ۲.۲ در نهایت داریم:

$$F_i = (\frac{m-p-۱}{p})\frac{B_i}{۱-B_i} \sim F_{p,m-p-۱}$$

اگر مقدار مشاهده شده F_i بزرگتر از $F_{\alpha,p,m-p-۱}$ باشد، فرض پرت نبودن نمونه n تایی i ام در سطح α رد می‌شود.

مثال ۲.٤. داده‌های هاوکینز و همکاران (۱۹۸٤)

این مجموعه داده‌ها شامل ۷۵ مشاهده ۳ متغیره است. روسو و زومرن (۱۹۹۰) ادعا می‌کنند که ۱۵مشاهده اول این مجموعه داده‌ها، پرت هستند[۵]. برای دسترسی به داده‌های خام می‌توان به هاوکینز و همکاران مراجعه کرد[٤]. برای بررسی صحت این ادعا ما در این مثال مجموعه داده‌ها را به $m = ۵$ گروه $n = ۱۵$ تایی تقسیم کرده و میانگین هر گروه را محاسبه کردیم که بردارهای زیر حاصل می‌شود.

$$\overline{\boldsymbol{X}}_۱ = \begin{bmatrix} ۱۰/۰۱ \\ ۲۰/۴۹ \\ ۲۹/۴۳ \end{bmatrix}, \overline{\boldsymbol{X}}_۲ = \begin{bmatrix} ۱/۶۷ \\ ۱/۹۹ \\ ۱/۵۶ \end{bmatrix}, \overline{\boldsymbol{X}}_۳ = \begin{bmatrix} ۱/۴۳ \\ ۱/۶۱ \\ ۱/۶۵ \end{bmatrix}$$

$$\overline{\boldsymbol{X}}_۴ = \begin{bmatrix} ۱/۹۰ \\ ۱/۷۵ \\ ۱/۸۱ \end{bmatrix} \quad \overline{\boldsymbol{X}}_۵ = \begin{bmatrix} ۱/۰۳ \\ ۱/۶۹ \\ ۱/۷۰ \end{bmatrix}$$

براساس این بردارها آرایه‌های زیر حاصل می‌گردد.

$$\overline{\overline{\boldsymbol{X}}} = \begin{bmatrix} ۳/۱۲ \\ ۵/۶ \\ ۷/۲۳ \end{bmatrix}, S_{\overline{\boldsymbol{X}}} = \begin{bmatrix} ۳/۳۴ & ۲۸/۴۷ & ۴۱/۲۴ \\ ۲۸/۴۷ & ۶۷/۸۸ & ۹۴/۶۷ \\ ۴۱/۲۴ & ۹۴/۶۷ & ۱۳۷/۸۳ \end{bmatrix}$$

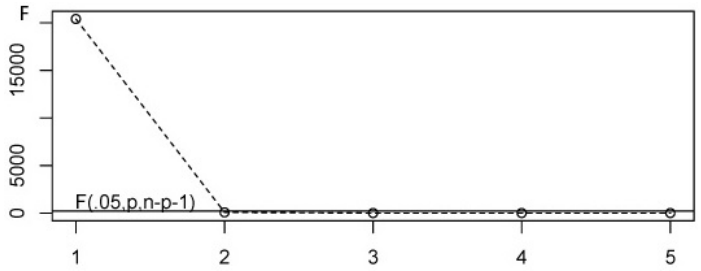
حال با استفاده از قضیه ۱.٤ آماره آزمون برای گروه اول برابر است با

$$B_۱ = \frac{۵}{(۵-۱)^۲}(\overline{\boldsymbol{X}}_۱ - \overline{\overline{\boldsymbol{X}}})S_{\overline{\boldsymbol{X}}}^{-۱}(\overline{\boldsymbol{X}}_۱ - \overline{\overline{\boldsymbol{X}}})^T = ۰/۹۹۹۹$$

$$F_۱ = (\frac{۵-۳-۱}{۳})\frac{۰/۹۹۹۹}{۰/۰۰۰۱} = ۲۰۳۹۶/۵۹$$

که از $F_{۰/۰۵,۳,۱} = ۲۱۵/۷۱$ بزرگتر می‌باشد و فرض صفر مبنی بر پرت نبودن گروه اول رد می‌شود. نتایج آزمون در سطح ۰/۰۵ در شکل ۱ آورده شده است که با توجه به شکل مقدار آماره آزمون برای گروه اول بالاتر از مقدار بحرانی (خط ممتد) قرار گرفته است و بیانگر این واقعیت است که آزمون صحت ادعای روسو و زومرن مبنی بر پرت بودن ۱۵ مشاهده اول را تایید می‌کند.

[Downloaded from andisheyeamari.irstat.ir on 2025-05-13]



شکل۱. نتایج آزمون F برای داده‌های گروه بندی شده هاوکینز و همکاران که در آن نقاط بیانگرمقدار آماره آزمون برای گروه‌ها و خط ممتد مقدار بحرانی آزمون می‌باشد

مثال ۳.٤. مطالعات شبیه سازی

در این مثال به کمک شبیه سازی و با استفاده از نرم افزار R عملکرد آزمون پرت بودن گروهی از مشاهدات را در آلودگی‌های مختلف مورد بررسی قرار می‌دهیم، به این صورت که گروه پرت را با میانگین، واریانس و همبستگی دیگری تولید می‌کنیم. درواقع در آلودگی واریانس و همبستگی ماتریس کوواریانس را آلوده می‌کنیم.

در هر اجرا m نمونه n تایی (با مقادیر مختلف m و n) از توزیع $N_۳(\boldsymbol{\mu},\Sigma)$ تولید می‌کنیم.

$$\boldsymbol{\mu}=\begin{bmatrix}۱\\۵\\۹\end{bmatrix} \qquad \Sigma=\begin{bmatrix}۱ & ۰/۹ & ۰/۹\\ ۰/۹ & ۱ & ۰/۹\\ ۰/۹ & ۰/۹ & ۱\end{bmatrix}$$

که نمونه n تایی m ام در همه موارد آلوده می‌باشد. در نهایت برای هر یک از سه حالت آلودگی میانگین، آلودگی واریانس و آلودگی همبستگی نتایج هزار تکرار آزمون را ارائه می‌دهیم.

۱. **آلودگی میانگین**

در این حالت $m-۱$ نمونه n تایی اول را از توزیع $N_۳(\boldsymbol{\mu},\Sigma)$ تولید کرده‌ایم ولی آخرین نمونه را با میانگین $\boldsymbol{\mu}_{outlier}=\boldsymbol{\mu}+\boldsymbol{a}_i$ تولید کرده‌ایم. در جدول ۱برای سه مقدار

$$\boldsymbol{a}_۱=\begin{bmatrix}۱ & ۱ & ۱\end{bmatrix}^T \qquad \boldsymbol{a}_۲=\begin{bmatrix}۱ & ۱ & ۰\end{bmatrix}^T$$
$$\boldsymbol{a}_۳=\begin{bmatrix}۱ & ۰ & ۰\end{bmatrix}^T$$

و برای مقادیر $m=۵,۱۰,۳۰,۱۰۰$ و $n=۴,۱۰,۲۰,۳۰$ نتایج هزار تکرار آزمون در سطح ۰/۰۵ ارائه شده است. در جدول ۱ درصد مواردی که آزمون نمونه n تایی پرت را بدرستی تایید کرده ارائه گردیده است.

جدول ۱: درصد تشخیص‌های صحیح در آلودگی میانگین

		n			
	m	٤	۱۰	۲۰	۳۰
$\boldsymbol{a}_۱$	۵	۸/۳	۱۱	۱٤/۲	۱۵/۸
	۱۰	۱۸/۷	۱۹/۲	۷٦/۵	۸۹/۹
	۳۰	۲۸/۱	۷٤/۱	۹٦	۹۹/۸
	۱۰۰	۳٦/۱	۷۷/۵	۹۸/۲	۹۹/۹
$\boldsymbol{a}_۲$	۵	۱٤/۹	۲۲/۱	۳۳/۷	۳۸/۷
	۱۰	۸۷/۱	۱۰۰	۱۰۰	۱۰۰
	۳۰	۹۸/۵	۱۰۰	۱۰۰	۱۰۰
	۱۰۰	۹۹/٦	۱۰۰	۱۰۰	۱۰۰
$\boldsymbol{a}_۳$	۵	۱۳/٤	۲۲/٦	۳۳/۳	٤۰/٤
	۱۰	۸٤/۵	۹۹/۵	۱۰۰	۱۰۰
	۳۰	۹۸/۵	۱۰۰	۱۰۰	۱۰۰
	۱۰۰	۹۹/۳	۱۰۰	۱۰۰	۱۰۰

با توجه به جدول ۱ با افزایش m و n عملکرد آزمون بهبود می‌یابد که افزایش m عملکرد آزمون را نسبت به n با نرخ سریع‌تری بهبود می‌بخشد. بردار $\boldsymbol{a}_۳$ نیز نسبت به دو بردار دیگر آلودگی بیشتری تولید می‌کند زیرا در این حالت با افزایش میانگین یک عضو همبستگی بین متغیرها نیز تحت

[Downloaded from andisheyeamari.irstat.ir on 2025-05-13]

تأثیر قرار می‌گیرد.

۲. آلودگی واریانس

در این حالت $m-۱$ نمونه n تایی اول را از توزیع $N_۳(\boldsymbol{\mu},\Sigma)$ و آخرین نمونه را از توزیع $N_۳(\boldsymbol{\mu},k\Sigma)$ برای مقادیر مختلف k تولید کرده‌ایم. در جدول ۲ برای مقادیر $k=۲,۵,۱۰$ ، $m=۵,۱۰,۳۰,۱۰۰$ و $n=۴,۱۰,۲۰,۳۰$ نتایج هزار تکرار آزمون در سطح ۰/۰۵ و درصد مواردی که آزمون نمونه n تایی پرت را بدرستی تایید کرده ارائه گردیده است.

جدول ۲: درصد تشخیص‌های صحیح در آلودگی واریانس

		n			
	m	٤	۱۰	۲۰	۳۰
$k=۲$	۵	۷/۱	۷/٤	۷/۹	۷/٦
	۱۰	۱۵/۵	۱۳/۸	۱٤/۹	۱۵
	۳۰	۲۲/٤	۲۲/۹	۲٤/۹	۲۳/۱
	۱۰۰	۲۳/۷	۲٤/۵	۲۷/۳	۲۳/۵
$k=۵$	۵	۹/٦	۹/۷	۸/٦	۸/٦
	۱۰	٤٤/۹	٤٤/۹	٤۷/۲	٤۱/۹
	۳۰	٦۰/۲	٦۱/٦	٦۰/۲	٦۲/۹
	۱۰۰	٦٤/۱	٦۳	٦٤/۱	٦۵
$k=۱۰$	۵	۱٤/۷	۱۲/۹	۱٤/۳	۱۳/۳
	۱۰	٦٦/٤	٦۷/۹	٦۸/۲	٦۹/۳
	۳۰	۸۲	۸۲/۵	۸۰	۸۳
	۱۰۰	۸۳/٤	۸٦	۸٤/۲	۸٦/۵

۳. آلودگی همبستگی

در این حالت نیز $m-۱$ نمونه n تایی اول را از توزیع $N_۳(\boldsymbol{\mu},\Sigma)$ تولید کرده‌ایم ولی آخرین نمونه را ماتریس کوواریانس دیگری از بین سه ماتریس کوواریانس زیر تولید کرده‌ایم.

$$\Sigma_۱=\begin{bmatrix} ۱ & ۰/۵ & ۰/۵ \\ ۰/۵ & ۱ & ۰/۵ \\ ۰/۵ & ۰/۵ & ۱ \end{bmatrix} \quad \Sigma_۲=\begin{bmatrix} ۱ & ۰ & ۰ \\ ۰ & ۱ & ۰ \\ ۰ & ۰ & ۱ \end{bmatrix}$$

$$\Sigma_۳=\begin{bmatrix} ۱ & -۰/۵ & -۰/۵ \\ -۰/۵ & ۱ & -۰/۵ \\ -۰/۵ & -۰/۵ & ۱ \end{bmatrix}$$

در جدول ۳ برای سه مقدار و برای مقادیر $m=۵,۱۰,۳۰,۱۰۰$ و $n=۴,۱۰,۲۰,۳۰$ نتایج هزار تکرار آزمون در سطح ۰/۰۵ ارائه شده است. در جدول ۳ درصد مواردی که آزمون نمونه n تایی پرت را بدرستی تایید کرده ارائه گردیده است.

جدول ۳: درصد تشخیص‌های صحیح در آلودگی همبستگی

		n			
	m	٤	۱۰	۲۰	۳۰
$\Sigma_۳$	۵	۱۲	۱۳/٤	۱۲/۸	۱٤/۳
	۱۰	٦۰/۲	٦۱/۱	٦۲	٦۳/۹
	۳۰	۷٤/۱	۷۲/۲	۷٦/۲	۷۳/۳
	۱۰۰	۷۲	۷٤/۲	۷٤/۷	۷۵/۷
$\Sigma_۲$	۵	۱۰/٤	۱۳	۱۱/۳	۱۱/٦
	۱۰	٤۹/۸	۵۲	۵۳/٤	٤۹/٤
	۳۰	٦۳	٦٦/۱	٦۷	٦٤/۲
	۱۰۰	٦۳/۱	٦٦/۵	٦۷	٦۹/٦
$\Sigma_۱$	۵	۸/۷	۸/٦	۹	۹/۲
	۱۰	۳۲/۳	۳۱/۱	۳۲/۹	۳۱/۲
	۳۰	٤۳/۹	٤۵/۲	٤۲/۲	٤۲/۳
	۱۰۰	٤۷/۵	٤٦/۱	۵۰	٤۹

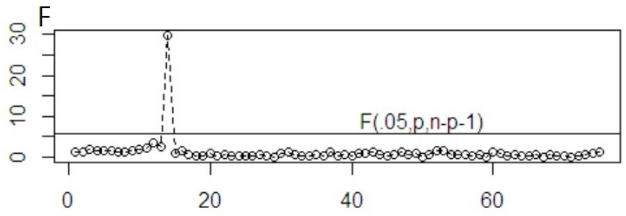
باتوجه به جداول فوق افزایش مقدار m و n موجب بهبود عملکرد آزمون می‌شود. ولی به نظر می‌رسد که حداقل در شرایط مثال

[Downloaded from andisheyeamari.irstat.ir on 2025-05-13]

حاضر، این آزمون در تشخیص آلودگی ضریب همبستگی ضعیف است ولی در تشخیص آلودگی در بردار میانگین عملکرد خوبی دارد. این نکته حائز اهمیت است که آلودگی‌های بکارگرفته شده در مثال حاضر بسیار خفیف می‌باشد که در عمل و در شرایط طبیعی با آلودگی‌های شدیدتری روبرو خواهیم بود.

۵ نتیجه گیری

داده‌های مثال۲.٤ را در نظر بگیرید، این مجموعه داده‌های مشهور به منظور بررسی روش‌های شناسایی نقاط پرت و آزمون آن‌ها در بسیاری از مقالات علمی و پژوهشی مورد استفاده قرار گرفته است که اکثر پژوهشگران پرت بودن ۱۵ مشاهده اول این مجموعه داده‌ها را تایید کرده‌اند. حال اگر از آزمون پرت بودن یک مشاهده تکی (بخش۳) را برای این مجموعه داده‌ها استفاده کنیم مشاهده می‌شود که این آزمون تنها پرت بودن مشاهده ١٤ را تایید می‌کند. نتیجه محاسبات آماره آزمون برای تک تک مشاهدات و مقدار بحرانی آزمون در شکل ۲ آورده شده است که تنها مشاهده ١٤ بالا تر از مقدار بحرانی (خط ممتد) قرار گرفته است به عبارت دیگر سایر مشاهدات پرت نیستند که خلاف واقعیت و نتایج مثال ۲.٤ می‌باشد.



شکل۲. نتایج آزمون مشاهدات تکی برای داده‌های هاوکینز

بنابراین زمانی که با چندین داده پرت در مجموعه داده‌ها مواجه هستیم بهتر است از آزمون پرت بودن گروهی از مشاهدات چند متغیره استفاده کنیم زیرا با این آزمون داده‌های پرت در یک گروه قرار گرفته و باعث می‌شوند که فاصله آن‌ها با میانگین داده حفظ شود و به تبع آن آماره آزمون از مقدار بحرانی بزرگتر شده و به نتیجه مطلوب برسیم.

مراجع

[۱] توانگر، م، طالبی، ه، علامت ساز، م . (۱۳۹۰)، خواص توزیعی آماره T^2 هتلینگ در کنترل کیفیت چند متغیره.اندیشه آماری

[۲] خردمندنیا، م، علیرضایی، م، مقدمه‌ای بر آمار چند متغیره، انتشارات نگارخانه، اصفهان، ۱۳۹۲. ،**(١٦)**،۲۷-۱۷.

[3] Alt, F. B. (1973). Aspects of Multivariate Control Charts. *Master of Science. Georgia In Stitvte of Technology*, **6**

[4] Hawkins, D. M. , Bradu, D. and Kass, G. V. (1984). Location of Several Outlier in Multiple Regression Data Using Elemental Sets. *technometrics*, **26**, 197-208.

[5] Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*. **85**, 633-639.

[6] Searle, S. (1928). *Matrix Algebra Useful for Statistics.* NewYork:John Wiley and Sone.

[7] Srivastava, M. S. and von Rosen, D. (1998). Outlier in Multivariate Regression Models, *J. Multivariate Anal.*, **65**, 195-208.

[8] Tracy, N. D. ;Young, J. C. and Mason, R. L. (1992). Multivariate Control Chart for Individual Observation. *Journal of Quality Technology*, **24**, 88-95.

[9] Wilks, S. S. (1963).Multivariate statistical outliers, *Sankhya: The Indian Journal of Statistics*, **25(4)**, 407-426.

[Downloaded from andisheyeamari.irstat.ir on 2025-05-13]