مجله علوم آماری، بهار و تابستان ۱۳۹۸
جلد ۱۳، شماره ۱، ص ۱۷۳ – ۱۸۴
DOI: 10.29252/jss.13.1.173

یک مدل پاسخ تصادفیده مرکب جدید

سید محمدرضا علوی، محمد جوهرزاده و رحیم چینی پرداز

گروه آمار، دانشگاه شهید چمران اهواز

تاریخ دریافت: ۱۳۹۶/۰۲/۲۵ تاریخ پذیرش: ۱۳۹۷/۰۷/۲۳

چکیده: معمولاً در بررسی‌های نمونه‌گیری هنگامی‌که سؤال حساس مستقیمی پرسیده شود، پاسخگو پاسخ واقعی را ارائه نمی‌کند. روش‌های پاسخ تصادفیده برای حفاظت از محرمانگی پاسخ‌ها مطرح شده‌اند. تمرکز این مقاله بر روش پاسخ تصادفیده در متغیرهای کیفی بر پایه روش سیمونس است. برای حفظ محرمانگی بیشتر با ترکیب دو روش جداگانه سیمونس، یک روش پاسخ تصادفیده مرکب جدید معرفی می‌شود و با استفاده از شبیه سازی در نرم افزار R کارایی روش تصادفیده پیشنهادی با روش سیمونس و روش تصادفیده علوی و تاج‌الدینی (۱۳۹۴) مقایسه می‌شود. با استفاده از روش تصادفیده پیشنهادی، نسبت تقلب دانشجویان در امتحانات دانشگاه شهید چمران اهواز برآورد شده‌است.

واژه‌های کلیدی: پاسخ تصادفیده، روش سیمونس، سؤال نامرتبط، پاسخ تصادفیده مرکب، نسبت حساس.

۱ مقدمه

در بسیاری از بررسی‌های نمونه‌گیری، آماردانان سعی دارند تا اطلاعاتی درباره ویژگی‌هایی از جوامع انسانی به‌دست آورند. اگر ویژگی‌های حساسی مستقیماً از افراد پرسیده شوند ممکن است با عدم پاس یا پاسخ‌های غیر واقعی مواجه شوند. زیرا این‌گونه سؤالات امکان دارد برای برخی افراد خجالت‌آور یا عذاب‌آور باشند. بعضی از موضوعات حساس مورد پرسش عبور از چراغ قرمز، قماربازی، فرار مالیاتی، مصرف مشروبات الکلی، استعمال مواد مخدر، تقلب در امتحانات و غیره است. به‌منظور حفظ محرمانگی و کاهش اریبی

آدرس الکترونیکی نویسنده مسئول مقاله: سید محمدرضا علوی، alavi_m@scu.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62D05،

در پاسخ‌ها، روش‌های پاسخ تصادفیده [۱] به‌عنوان جایگزین روش سؤالات مستقیم [۲] مطرح شده‌اند. ایده اصلی این روش‌ها نخستین بار توسط وارنر (۱۹۶۵) به‌منظور ترغیب به‌همکاری و بیان پاسخ‌های صادقانه مطرح شد. در این روش پاسخ تصادفیده، هر فرد با انجام یک آزمایش تصادفی مانند پرتاب سکه یا ریختن تاس ضمن محرمانه ماندن پاسخ، یکی از دو سؤال حساس یا مکمل حساس را برای پاسخگویی انتخاب می‌کند. در این پژوهش‌ها معمولاً هدف، برآورد نسبت حساس با استفاده از پاسخ‌های تصادفیده است. گرینبرگ و همکاران (۱۹۶۹) برای افزایش حفظ محرمانگی استفاده از سؤال نامرتبط غیرحساس را در پرسش دوم مطرح کردند که روش سیمونس نامیده می‌شود. مدل‌های تصادفیده زیادی توسط افرادی نظیر مورس (۱۹۷۱)، راجاوارائو (۱۹۷۸)، منگات و سینگ (۱۹۹۰)، کاک (۱۹۹۰)، منگات (۱۹۹۴)، سینگ (۲۰۰۲)، هانگ (۲۰۰۴)، چانگ و همکاران (۲۰۰۴)، کیم و وارده (۲۰۰۵)، جست وانگ و سینگ (۲۰۰۶)، حسین و شبیر (۲۰۰۷)، مهتا و همکاران (۲۰۱۲)، یزاری و علوی (۱۳۹۳) و علوی و تاج الدینی (۱۳۹۴ و ۲۰۱۶) معرفی شده‌اند. در این مقاله براساس روش سیمونس یک روش جدید پاسخ تصادفیده مرکب برای افزایش محرمانگی پاسخ پاسخگو، معرفی شده است. مقاله از پنج بخش تشکیل شده است. در بخش ۲ روش پاسخ تصادفیده سیمونس و گونه مکرر آن یعنی روش علوی و تاج‌الدینی (۱۳۹۴) بیان شده است. در بخش ۳ با ترکیب دو روش جداگانه سیمونس، یک روش پاسخ تصادفیده مرکب جدید معرفی و برآورد نسبت حساس توسط این روش پیشنهادی ارائه و در بخش ۴ کارایی [۳] این روش از طریق شبیه‌سازی با استفاده از بسته نرم‌افزاری R با روش سیمونس و روش علوی و تاج‌الدینی (۱۳۹۴) مقایسه شده است. سرانجام در بخش ۵ با استفاده از روش پیشنهادی نسبت تقلب دانشجویان در امتحانات دانشگاه شهید چمران اهواز آورده شده است..

۲ روش پاسخ تصادفیده سیمونس و گونه مکرر آن

در این بخش ابتدا روش پاسخ تصادفیده سیمونس شرح داده می‌شود و سپس گونه مکرر آن یعنی روش علوی و تاج‌الدینی (۱۳۹۴) بیان می‌شود.

---

[1] Randomized response techniques
[2] Direct questions
[3] Efficiency

۱.۲ روش پاسخ تصادفیده سیمونس

در روش سیمونس از فرد iام نمونه تقاضا می‌شود یک آزمایش برنولی با احتمال موفقیت p انجام داده در صورت پیروزی به‌سؤال حساس و در صورت شکست به یک سؤال نامرتبط غیر حساس پاسخ دهد. اگر متغیرهای مستقل برنولی Y_i ، X_i و T_i به‌ازای $i = ۱, ..., n$ به‌ترتیب معرف پاسخ سؤال حساس، پاسخ سؤال نامرتبط و نتیجه آزمایش برنولی با احتمال‌های به‌ترتیب θ ، π_B و p باشند، پاسخ تصادفیده به‌صورت

$$Z_i = Y_i T_i + X_i(۱ - T_i), \;\; i = ۱, ..., n \qquad (۱)$$

است. واضح است که Z_i یک متغیر تصادفی برنولی با احتمال پیروزی به‌صورت

$$\phi = E_P E_R(Z_i) = E_P E_R[Y_i T_i + X_i(۱ - T_i)] = p\theta + (۱ - p)\pi_B, \;\; i = ۱, ..., n$$

است، که در آن E_P بیانگر امید ریاضی روی تمام نمونه‌های ممکن و E_R امید ریاضی تحت عمل تصادفی کردن است. بر پایه یک نمونه تصادفی از پاسخ‌های تصادفیده، برآورد نااریب θ و واریانس آن به‌ترتیب به‌صورت

$$\hat{\theta} = \frac{\bar{Z} - (۱ - p)\pi_B}{p} \qquad (۲)$$

$$Var(\hat{\theta}) = \frac{\theta(۱ - \theta)}{n} + \frac{\theta(۱ - p)(۱ - \pi_B)}{np} + \frac{\pi_B(۱ - p)[۱ - \pi_B(۱ - p)(۱ - p)]}{np^۲}$$

هستند که در آن $\bar{Z} = \frac{\sum_{i=۱}^{n} Z_i}{n}$ میانگین پاسخ بله در نمونه تصادفی به حجم n یک برآورد نااریب برای ϕ است. در ادامه برآوردی برای $Var(\hat{\theta})$ به‌صورت

$$\hat{V}ar(\hat{\theta}) = \frac{Var(\bar{Z})}{p^۲} = \frac{\bar{Z}(۱ - \bar{Z})}{np^۲}$$

معرفی می‌شود.

۲.۲ روش پاسخ تصادفیده مکرر سیمونس

در صورتی‌که سؤال غیرمرتبط در روش سیمونس یک آزمایش تصادفی باشد که با تکرار آن نتایج بتوانند تغییر کنند، علوی و تاج‌الدینی (۱۳۹۴) با تکرار روش سیمونس یک روش جدید برای حفظ محرمانگی بیشتر معرفی کردند. لازم به ذکر است اگر سؤال غیرمرتبط غیرتصادفی باشد ممکن است پاسخگو احساس کند تکرار پاسخ تصادفیده سیمونس محرمانگی را حفظ نمی‌کند. در این روش از فرد i ام نمونه تقاضا می‌شود روش پاسخ تصادفیده سیمونس را f_i بار تکرار کند که f_i دارای توزیع اریب اندازه پواسن با میانگین دلخواه مثلاً ۳ است. اگر m_i نسبت تعداد بله پاسخ‌های تصادفیده فرد i ام نمونه باشد، در آن‌صورت یک برآورد نااریب برای نسبت حساس θ با استفاده از پاسخ‌های تصافیده فرد i ام به‌صورت

$$\hat{\theta}_i = \frac{m_i - (1-p)\pi_B}{p}, \quad i = 1, ..., n$$

است. پس یک برآورد ناریب براساس تمام پاسخ‌های تصادفیده نمونه از رابطه

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \hat{\theta}_i}{n} \qquad (۳)$$

به‌دست می‌آید (علوی و تاج‌الدینی، ۱۳۹۴).

۳ روش پاسخ تصادفیده مرکب جدید

در روش پاسخ تصادفیده از هر فرد در نمونه تصادفی تقاضا می‌شود اگر دارای صفت غیر حساس B است، روش پاسخ تصادفیده سیمونس اول و در غیر این‌صورت روش پاسخ تصادفیده سیمونس دوم را انجام دهد.

روش پاسخ تصادفیده سیمونس اول: اگر فرد دارای صفت غیرحساس B است یک آزمایش برنولی با احتمال موفقیت p_1 انجام داده در صورت پیروزی فقط به‌سؤال "آیا شما دارای صفت حساس A هستید؟" و در صورت شکست فقط به‌سؤال "آیا شما دارای صفت غیرحساس B هستید ؟" پاسخ داده و نتیجه را در پاسخ‌نامه ثبت کند.

روش پاسخ تصادفیده سیمونس دوم: اگر فرد دارای صفت غیرحساس B^c (مکمل B) است یک آزماش برنولی با احتمال موفقیت p_2 انجام داده در صورت پیروزی فقط به‌سؤال "آیا شما دارای صفت

حساس A هستید؟" و در صورت شکست فقط به‌سؤال "آیا شما دارای صفت غیرحساس B^c هستید ؟" پاسخ داده و نتیجه را در پاسخ‌نامه ثبت کند.
اگر متغیرهای مستقل برنولی Y، X، $T_۱$ و $T_۲$ به‌ترتیب پاسخ سؤال حساس A، پاسخ سؤال غیر حساس B، نتیجه آزمایش تصادفی روش اول و روش دوم با احتمال‌های به ترتیب θ، π_B، $p_۱$ و $p_۲$ باشند، آن‌گاه پاسخ تصادفیده Z به‌صورت

$$Z = X[T_۱Y + (۱ - T_۱)X] + (۱ - X)[T_۲Y + (۱ - T_۲)(۱ - X)]$$

است. با توجه به اینکه توان دوم یک متغیر برنولی معادل خودش است، Z به‌صورت

$$Z = [T_۱X + T_۲(۱ - X)]Y + (۱ - T_۱)X + (۱ - T_۲)(۱ - X) \qquad (۴)$$

ساده می‌شود. امید ریاضی پاسخ تصادفیده برابر

$$\begin{aligned} \phi &= E_P E_R[T_۱X + T_۲(۱ - X)]Y + (۱ - T_۱)X + (۱ - T_۲)(۱ - X) \\ &= [p_۱\pi_B + p_۲(۱ - \pi_B)\theta] + \pi_B(۱ - p_۱) + (۱ - \pi_B)(۱ - p_۲) \end{aligned}$$

است. درنتیجه اگر $Z_۱, ..., Z_n$ یک نمونه تصادفی از پاسخ‌های تصادفیده به روش پیشنهادی باشند، یک برآورد نااریب برای نسبت حساس θ به‌صورت

$$\hat{\theta} = \frac{\bar{Z} - [\pi_B(۱ - p_۱) + (۱ - \pi_B)(۱ - p_۲)]}{\pi_B p_۱ + (۱ - \pi_B)p_۲} \qquad (۵)$$

به‌دست می‌آید، که در آن $\bar{Z} = \frac{\sum_{i=۱}^{n} Z_i}{n}$ برآورد نااریب ϕ است. واریانس این برآوردگر برابر

$$Var(\hat{\theta}) = \frac{\phi(۱ - \phi)}{nk^۲} = \frac{(k\theta + d)k(۱ - \theta)}{nk^۲} = \frac{\theta(۱ - \theta)}{n} + \frac{d(۱ - \theta)}{nk}$$

است که در آن $k = \pi_B p_۱ + (۱ - \pi_B)p_۲$ و $d = \pi_B(۱ - p_۱) + (۱ - \pi_B)(۱ - p_۲)$.
برآورد این واریانس به‌صورت $\hat{V}ar(\hat{\theta}) = \frac{\bar{Z}(۱ - \bar{Z})}{nk^۲}$ است.

حالت های خاص مدل پیشنهادی:
(الف) عبارت (۵) به ازای $p_1 = 1 - p_2$، به‌صورت $\hat{\theta} = \frac{\bar{Z}-[\pi_B(1-2p_1)+p_1]}{\pi_B(2p_1-1)+(1-p_1)}$ است.
(ب) اگر $\pi_B = \frac{1}{2}$، آنگاه $\hat{\theta} = \frac{\bar{Z}-[1-\frac{p_1+p_2}{2}]}{\frac{p_1+p_2}{2}}$، که به ازای $p = \frac{p_1+p_2}{2}$ داریم $\hat{\theta} = \frac{\bar{Z}-[1-p]}{p}$
(پ) اگر $\pi_B = \frac{1}{2}$ و $p_1 = 1 - p_2$، آنگاه

$$\hat{\theta} = 2\bar{Z} - 1, \quad Var(\hat{\theta}) = \frac{4\phi(1-\phi)}{n} \tag{۶}$$

و یک برآورد برای این واریانس عبارتست از

$$\hat{V}ar(\hat{\theta}) = \frac{4\bar{Z}(1-\bar{Z})}{n}. \tag{۷}$$

۴ مطالعه شبیه‌سازی

با توجه به نااریب بودن برآورد نسبت حساس در هر سه روش، برای ارزیابی روش پیشنهادی و مقایسه کارایی آن با دو روش سیمونس و روش علوی و تاج الدینی (۱۳۹۴) بااستفاده از بسته نرم افزاری R سه شبیه‌سازی در نظر گرفته شده‌است.
برای شبیه سازی روش سیمونس مراحل زیر انجام شده است:
مرحله ۱: سه نمونه تصادفی به‌حجم n از متغیرهای تصادفی برنولی Y، X و T به‌ترتیب با احتمال‌های پیروزی معلوم θ، π_B و p تولید و سپس با استفاده از رابطه (۱) داده‌های تصادفیده Z به‌دست آورده و براساس رابطه (۲) برآوردهای $\hat{\theta}$ برای نمونه به‌حجم n محاسبه شده‌اند.
مرحله ۲: مرحله ۱، k بار تکرار، میانگین و واریانس این k بار به‌ترتیب به‌عنوان امید ریاضی و واریانس شبیه‌سازی شده $\hat{\theta}$ در نظر گرفته شدند.
برای شبیه‌سازی روش پیشنهادی مراحل زیر انجام شده است:
مرحله ۱: چهار نمونه تصادفی به‌حجم n از متغیرهای برنولی Y، X، T_1 و T_2 به‌ترتیب با احتمال‌های پیروزی معلوم θ، π_B، p_1 و p_2 تولید شدند. لازم به‌ذکر است که T_1 و T_2 بیانگر نتیجه آزمایش‌های برنولی مراحل اول و دوم روش پیشنهادی هستند.
حال با استفاده از رابطه (۴) داده‌های تصادفیده Z را به‌دست آورده و بر اساس رابطه (۵) ، برآوردهای θ در نمونه به‌حجم n محاسبه شدند.

مرحله ۲: مرحله ۱، k بار تکرار، سپس میانگین و واریانس این k بار به‌ترتیب به‌عنوان امید ریاضی و واریانس شبیه‌سازی شده $\hat{\theta}$ در نظر گرفته شدند.

برای شبیه‌سازی روش علوی و تاج‌الدینی (۱۳۹۴) مراحل زیر انجام شده است:

مرحله ۱: دو نمونه تصادفی به‌حجم n از متغیرهای تصادفی برنولی Y و X به‌ترتیب با احتمال‌های پیروزی معلوم θ و π_B تولید و سپس برای واحد iام، روش پاسخ تصادفیده سیمونس f_i بار تکرار شد که f_i دارای توزیع پواسن اریب اندازه با میانگین ۳ است. اگر m_i نسبت تعداد بله پاسخ‌های تصادفیده واحد iام باشد، در آن‌صورت یک برآورد نااریب برای نسبت حساس θ با استفاده از پاسخ‌های تصافیده واحد iام به‌صورت

$$\hat{\theta}_i = \frac{m_i - (۱ - p)\pi_B}{p}, \;\; i = ۱, ..., n$$

است، سپس یک برآورد ناریب براساس تمام پاسخ‌های تصادفیده نمونه از رابطه (۳) محاسبه شده است (علوی و تاج‌الدینی، ۱۳۹۴).

مرحله ۲: مرحله ۱، k بار تکرار و سپس میانگین و واریانس این k بار به‌ترتیب به‌عنوان امید ریاضی و واریانس شبیه‌سازی شده $\hat{\theta}$ در نظر گرفته شدند.

جدول ۱: واریانس روش‌های سیمونس، پیشنهادی و علوی و تاج الدینی به ازای $\pi_B = ۰/۳$

$Var(AT)$	$Var(P)$	$Var(S)$	$p_۲$	$p_۱$	p	θ	n
۰/۰۲۹	۰/۰۸۳	۰/۰۹۶	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۲۰
۰/۰۳۳	۰/۰۶۴	۰/۱۰۷	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۳۵	۰/۰۴۱	۰/۱۱۲	۰/۲۵	۰/۵	۰/۳۳	۰/۷	
۰/۰۱۴	۰/۰۳۲	۰/۰۳۸	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۵۰
۰/۰۱۶	۰/۰۲۵	۰/۰۴۲	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۱۷	۰/۰۱۶	۰/۰۴۵	۰/۲۵	۰/۵	۰/۳۳	۰/۷	
۰/۰۰۷	۰/۰۱۶	۰/۰۱۹	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۱۰۰
۰/۵	۰/۳۳	۰/۵	۰/۲۵	۰/۰۲۱	۰/۰۱۳	۰/۰۰۸	
۰/۰۰۸	۰/۰۰۸	۰/۰۲۲	۰/۲۵	۰/۵	۰/۳۳	۰/۷	
۰/۰۰۳	۰/۰۰۸	۰/۰۱۰	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۲۰۰
۰/۰۰۴	۰/۰۰۶	۰/۰۱۱	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۰۴	۰/۰۰۴	۰/۰۱۱	۰/۲۵	۰/۵	۰/۳۳	۰/۷	

جدول ۲: واریانس روش‌های سیمونس، پیشنهادی و علوی و تاج الدینی به ازای $\pi_B = ۰/۵$

$Var(AT)$	$Var(P)$	$Var(S)$	$p_۲$	$p_۱$	p	θ	n
۰/۰۳۹	۰/۰۶۹	۰/۱۱۴	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۲۰
۰/۰۴۱	۰/۰۵۵	۰/۱۱۷	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۴۰	۰/۰۳۶	۰/۱۱۴	۰/۲۵	۰/۵	۰/۳۳	۰/۷	
۰/۰۱۹	۰/۰۲۷	۰/۰۴۵	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۵۰
۰/۰۲۰	۰/۰۲۲	۰/۰۴۶	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۲۰	۰/۰۱۴	۰/۰۴۵	۰/۲۵	۰/۵	۰/۳۳	۰/۷	
۰/۰۰۹	۰/۰۱۴	۰/۰۲۲	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۱۰۰
۰/۰۱۰	۰/۰۱۱	۰/۰۲۳	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۱۰	۰/۰۰۷	۰/۰۲۲	۰/۲۵	۰/۵	۰/۳۳	۰/۷	
۰/۰۰۵	۰/۰۰۷	۰/۰۱۱	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۲۰۰
۰/۰۰۵	۰/۰۰۵	۰/۰۱۲	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۰۵	۰/۰۰۴	۰/۰۱۱	۰/۲۵	۰/۵	۰/۳۳	۰/۷	

جدول ۳: واریانس روش‌های سیمونس، پیشنهادی و علوی و تاج الدینی به ازای $\pi_B = ۰/۷$

$Var(AT)$	$Var(P)$	$Var(S)$	$p_۲$	$p_۱$	p	θ	n
۰/۰۳۱	۰/۰۵۷	۰/۱۱۰	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۲۰
۰/۰۳۳	۰/۰۴۶	۰/۱۰۹	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۳۴	۰/۰۳۱	۰/۰۹۸	۰/۲۵	۰/۵	۰/۳۳	۰/۷	
۰/۰۱۵	۰/۰۲۳	۰/۰۴۶	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۵۰
۰/۰۱۶	۰/۰۱۹	۰/۰۴۲	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۱۶	۰/۰۱۲	۰/۰۳۸	۰/۲۵	۰/۵	۰/۳۳	۰/۷	
۰/۰۰۸	۰/۰۱۲	۰/۰۲۲	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۱۰۰
۰/۰۰۸	۰/۰۰۹	۰/۰۲۱	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۰۳	۰/۰۰۶	۰/۰۱۱	۰/۲۵	۰/۵	۰/۳۳	۰/۷	
۰/۰۰۳	۰/۰۰۶	۰/۰۱۱	۰/۲۵	۰/۵	۰/۳۳	۰/۳	۲۰۰
۰/۰۰۴	۰/۰۰۵	۰/۰۱۱	۰/۲۵	۰/۵	۰/۳۳	۰/۵	
۰/۰۰۴	۰/۰۰۳	۰/۰۱۰	۰/۲۵	۰/۵	۰/۳۳	۰/۷	

جداول ۱ تا ۳ واریانس روش سیمونس ،(S) پیشنهادی (P) و علوی و تاج الدینی (AT) را تا سه رقم اعشار برای اندازه‌های نمونه ۲۰، ۵۰، ۱۰۰ و ۲۰۰ و مقادیر مختلف θ و π_B به‌ازای $k = ۱۰۰۰۰$ بار شبیه‌سازی، نشان می‌دهد. قابل ذکر است که θ احتمال نسبت حساس، π_B احتمال نسبت غیر حساس، p احتمال پیروزی در آزمایش برنولی به‌روش سیمونس و $p_۱$ و $p_۲$ احتمالات پیروزی در آزمایش‌های برنولی مرحله اول و دوم روش پیشنهادی هستند و مقادیر p، $p_۱$ و $p_۲$ جهت حفظ محرمانگی نباید خیلی کم و یا زیاد باشند، بنابراین در این شبیه‌سازی برای اختصار مقادیر ۰٫۳۳، ۰٫۵ و ۰٫۲۵ به‌ترتیب برای p، $p_۱$ و $p_۲$ در نظر گرفته شد. همان‌طور که از جداول ۱ تا ۳ ملاحظه می‌شود کارایی روش پیشنهادی بیشتر از سیمونس است. اما کارایی آن از روش علوی و تاج الدینی (۱۳۹۴) کمتر است. با افزایش حجم نمونه کارایی هر سه روش افزایش می‌یابد. به‌نظر می‌رسد با افزایش حجم نمونه کارایی روش پیشنهادی و روش علوی و تاج‌الدینی (۱۳۹۴) یکسان شود. با افزایش θ کارایی روش پیشنهادی افزایش می‌یابد.

۵ برآورد نسبت تقلب با روش پیشنهادی

تقلب از جمله روش‌های نامطلوبی است که باعث ایجاد رقابت ناسالم و گاه برتری کاذب متقلب بر دیگران می‌شود. از مصادیق تقلب در امتحانات می‌توان به استفاده از هر گونه یادداشت، نوشته، کتاب و جزوه غیر مجاز، رد و بدل کردن هر گونه اطلاعات کتبی یا شفاهی با سایر دانشجویان بدون هماهنگی با مراقبین و استفاده از تلفن همراه به هر دلیل اشاره کرد. یکی از آسیب‌های هر نظام آموزشی تقلب است که از جمله عوامل پدیدار شدن افت تحصیلی در محیط‌های آموزشی است. نسبت تقلب دانشجویان در امتحانات یکی از معیارهای محاسبه میزان سلامت اهداف آموزشی هر دانشگاه محسوب می‌شود. با توجه به‌این‌که اگر مستقیماً از دانشجویان درباره ارتکاب تقلب آن‌ها در امتحانات سؤال شود ممکن است پاسخ واقعی را ارائه نکنند، بنابراین در این بخش با استفاده از پاسخ‌های تصادفیده به روش پیشنهادی برای به‌دست آوردن نسبت تقلب دانشجویان در دانشگاه شهید چمران اهواز استفاده شده است.

جمع آوری داده‌ها در فروردین و اردیبهشت ماه سال ۱۳۹۵ با مراجعه تصادفی به دانشجویان مختلف صورت گرفته و مراحل انجام آزمایش برای آن‌ها شرح داده شده است. سؤال نامرتبط در پرسشنامه می‌بایست از دو ویژگی برخوردار باشد: یکی معلوم بودن احتمال پاسخ بله آن سؤال برای طراح نمونه‌گیری و دیگری مخفی ماندن پاسخ آن از مصاحبه کننده. از این‌رو سؤال نامرتبط، زوج یا فرد بودن یکان شماره شناسنامه دانشجو در نظر گرفته شد. به‌همراه پرسشنامه یک عدد سکه به دانشجو تحویل و از او درخواست شد روش تصادفیده مطرح شده در پرسشنامه را طبق دستورالعمل زیر در محل مورد نظر درج نماید:

درصورتی که یکان شماره شناسنامه شما زوج است یک سکه انداخته اگر شیرآمد، فقط به‌سؤال الف و اگر خط آمد فقط به‌سؤال ب زیر پاسخ داده و نتیجه را در پاسخ‌نامه گزارش نمایید.
الف) آیا شما در امتحانات رسمی دانشگاه تقلب کرده‌اید؟
ب) آیا رقم یکان شماره شناسنامه شما زوج است؟
اما درصورتی که یکان شماره شناسنامه شما فرد است یک سکه انداخته اگر شیرآمد فقط به‌سؤال الف و اگر خط آمد فقط به‌سؤال ب زیر پاسخ داده و نتیجه را در پاسخ‌نامه گزارش نمایید.
الف) آیا شما در امتحانات رسمی دانشگاه تقلب کرده‌اید؟
ب) آیا رقم یکان شماره شناسنامه شما فرد است؟
پاسخ‌نامه: بله □ خیر □

در این پرسشنامه چون از انداختن سکه برای انتخاب سؤال‌ها استفاده شده احتمال موفقیت آزمایش‌های برنولی در مرحله اول و دوم یعنی p_1 و p_2 برابر $\frac{1}{2}$ و چون از زوج یا فرد بودن یکان شناسنامه برای سؤال غیرحساس استفاده شده، π_B برابر با $\frac{1}{2}$ در نظر گرفته شده است. بنابراین از روابط (۶) و (۷) برای برآورد نسبت تقلب و برآورد واریانس آن استفاده شده است. در جدول ۴ خلاصه‌ای از نتایج نمونه‌گیری آمده است.

جدول ۴: نتایج نمونه‌گیری برای برآورد نسبت تقلب دانشجویی در دانشگاه

اندازه نمونه	تعداد بله	نسبت بله	برآورد نسبت تقلب	خطای معیار	بازه اطمینان ۹۵%
۱۷۰	۱۲۰	۰/۷۰۶	۰/۴۱۲	۰/۰۷۵	(۰/۲۶۵ و ۰/۵۵۹)

۶ بحث و نتیجه‌گیری

در این مقاله براساس روش پاسخ تصادفیده سیمونس، برای حفظ محرمانگی بیشتر یک روش پاسخ تصادفیده مرکب جدید برای برآورد نسبت حساس معرفی گردید و با شبیه سازی با استفاده از نرم‌افزار R کارایی آن با روش سیمونس و روش پاسخ تصادفیده علوی و تاج‌الدینی (۱۳۹۴) مقایسه شد. نشان داده شد که کارایی آن از روش سیمونس بیشتر اما از روش علوی و تاج‌الدینی کمتر است. با استفاده از داده‌های تصادفیده با این روش، نسبت تقلب دانشجویان در امتحانات دانشگاه شهید چمران اهواز در نیمسال دوم تحصیلی ۹۵-۱۳۹۴ بر اساس یک نمونه تصادفی به‌حجم ۱۷۰ دانشجو، ۰/۴۱۲ با خطای معیار ۰/۰۷۵ برآورد و فاصله اطمینان ۹۵% برای نسبت تقلب (۰/۲۶۵, ۰/۵۵۹) محاسبه شد.

تقدیر و تشکر

نویسندگان از پیشنهادات داوران و ویراستار محترم مجله که موجب ارائه بهتر مقاله شدند تشکر می‌نمایند.

مراجع

علوی، س. م. ر. و تاج الدینی، م. (۱۳۹۴)، یک روش پاسخ تصادفیده جدید و مقایسه آن با روش سیمونس، مجله علوم آماری ایران، ۹، ۲۲۷-۲۳۹.

یزاری، ز و علوی، س. م. ر . (۱۳۹۳)، مدل پاسخ تصادفیده ی کمّی اختیاری سه مرحله‌ای، مجله علوم آماری ایران، ۸، ۲۴۵-۲۶۰.

Alavi, S. M. R. and Tajodini, M. (2016), Maximum Likelihood Estimation of Sensitive Proportion Using Repeated Randomized Response Techniques, *Journal of Applied Statistics*, **43**, 563-571.

Chang, H., Wang, C. and Haung, K. (2005), On Estimating the Proportion of a Qualitative Sensitive Character Using Randomized Response Sampling, *Quality and Quantity*, **38**, 675–680.

Greenberg, B. G., Kuebler, R. R., Abernathy, J. R. and Horvitz, D. G. (1969), The Unrelated Question Randomized Response Model: Theoretical Framework, *Journal of the American Statistical Association*, **64**, 520–539.

Gjestvang, C. R. and Singh, S. (2006), A New Randomized Response Model, *Journal of the Royal Statistical Society*, B, **68**, 523-530.

Haung, K. (2004), A Survey Technique for Estimating the Proportion and Sensitivity in a Dichotomous Finite Population, *Statistica Neerlandica*, **58**, 75-82.

Hussain, Z. and Shabbir, J. (2007), Randomized Use of Warner's Randomized Response Model, InterStat: April 7, http://interstat.statjournals.net/INDEX/Apr07.html.

Kim, J. M., Warde, D. W. (2005), A Mixed Randomized Response Model, *Journal of Statistical Planning and Inference*, **133**, 211-221.

Kuk, A. Y. C. (1990), Asking Sensitive Questions Directly, *Biometrika*, **77**, 436-438.

Mangat , N. S. (1994), An Improved Randomized Response Strategy, Journal of the Royal Statistical Society, B, 56, 93-95.

Mangat, N. S. and Singh, R. (1990), An Alternative Randomized Response Procedure, *Biometrika*, **77**, 439-772.

Moors, J. J. A. (1971), Optimization of the Unrelated Question Randomized Response Model, *Journal of the American Statistical Association*, **66**, 627-629.

Raghavarao, D. (1978), On an Estimation Problem in Warner's Randomized Response Technique, *Biometrics*, **34**, 87-90.

Mehta, S., Dass, B. K., Shabbir, J. and Gupta, G. (2012), A Three Stage Optional Randomized Response Model, *Journal of Statistical Theory and Practice*, **6**, 417-427.

Singh, S. (2002), Randomized Response Model, *Metrika*, **56**, 131-142.

Warner, S. L. (1965), Randomized Response: a Survey Technique for Elim-inaiting Evasive Answer Bias, *Journal of the American Statistical Associ-ation*, **60**, 63-69.

مجله علوم آماری، بهار و تابستان ۱۳۹۸
جلد ۱۳، شماره ۱، ص ۱۸۵ – ۱۹۶
DOI: 10.29252/jss.13.1.185

مدل‌بندی رگرسیونی شکل از طریق مثلثی کردن

میثم مقیم بیگی، موسی گل‌علی‌زاده
گروه آمار، دانشگاه تربیت مدرس
تاریخ دریافت: ۱۳۹۵/۱۲/۱۹ تاریخ پذیرش: ۱۳۹۷/۰۷/۲۳

چکیده: با توجه به تعریف کندال از شکل به‌عنوان نقطه‌ای در فضای اَبَر کره، در این مقاله مدل‌بندی رگرسیونی شکل در این فضا مورد مطالعه قرار می‌گیرد. همچنین به منظور سهولت در مدل‌بندی، روش مثلث‌بندی شکل با استفاده از دو نقطه شاخص خاص پیشنهاد می‌شود که عملکرد مناسبی در مقایسه با رویکردهای دیگر دارد. مثلث‌بندی نه‌تنها مدل‌بندی رگرسیونی شکل را آسان می‌نماید بلکه توانایی بازسازی ساختار هندسی اشیاء با استفاده از ابزارهای ساده محاسباتی را دارد. نوآوری روش پیشنهادی مقاله حاضر در استفاده از متغیر تبینی مبتنی بر شکل اشیاء است که تغییرات هندسی متغیر پاسخ را به‌خوبی توصیف می‌کند. مقایسه و ارزیابی روش پیشنهادی با مدل انطباق پروکراستس کامل بر اساس معیار مجموع توان دوم خطا انجام و عملکرد دو مدل در تحلیل داده‌های پیکربندی جمجمه موش‌های آزمایشگاهی مورد بررسی قرار می‌گیرد.
واژه‌های کلیدی: رگرسیون کروی، آمار شکل، مثلث‌بندی، انطباق پروکراستس، فضای نااقلیدسی.

۱ مقدمه

گاهی اوقات شکل به‌عنوان منبع ارزشمندی از اطلاعات می‌تواند برای تحلیل آماری مورد استفاده قرار گیرد. انجام این مهم مستلزم حفظ ساختار هندسی آن است چراکه در غیر اینصورت بررسی تناظر بین اشیاء در یک قالب استاندارد ممکن نخواهد بود. در سالیان اخیر، موضوع مورد علاقه محققین، مدل‌بندی رگرسیونی داده‌های شکل بوده است. یکی از رویکردهای مرسوم برای نیل به این هدف، در نظر گرفتن

آدرس الکترونیکی نویسنده مسئول مقاله: موسی گل‌علی‌زاده، golalizadeh@modares.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J99، 62F30

مدل‌های معمول در فضای اقلیدسی برای مجموعه‌ای از اشکال است که با توجه به تعاریف ارائه شده از فضای شکل، به‌نظر می‌رسد این مدل‌ها از کارایی مطلوبی برخوردار نباشند. یکی از روش‌های مناسب جایگزین و تا حدی منطقی‌تر، انجام این تحلیل‌ها در فضای اصلی مورد مطالعه یعنی فضای شکل است.

بنا به کندال (۱۹۷۷) شکل، همه اطلاعات باقی‌مانده از یک پیکربندی پس از حذف اثرات مکان، مقیاس و دوران است. در دیدگاه ایشان فضای شکل یک فضای کروی با بعد بالا است. به‌طور دقیق‌تر و با پیروی از او، شکل یک مثلث معادل نقطه‌ای روی کره معمولی با بعد دو ($\mathbb{S}^2$)، و شکل یک مربع، هم ارز نقطه‌ای روی کره با بعد سه ($\mathbb{S}^3$) است. نوعی دیگری از نگاه به اشکال هندسی که توسط کندال (۱۹۸۴) و بوکشتین (۱۹۸۶) معرفی شد مثلث‌بندی[1] است. در این رویکرد با اختیار سه نقطه از هر شکل تمامی مثلث‌های ممکن به‌دست می‌آید و این مجموعه مثلث‌ها مبنای تحلیل آماری قرار می‌گیرد. می‌توان ملاحظه کرد که چنین روشی حالت خاصی از سنگفرش ورونی[2] است که به عنوان یکی از ابزارهای ریخت‌شناسی [3] در برخی از تحلیل‌های آماری ماتبط با مباحث فرایندهای تصادفی و الگوهای نقطه‌ای به‌کار گرفته شده است. جزئیاتی از این موضوع را می‌توان در استویان و استویان (۱۹۹۴) مطالعه کرد. حال اگر شخصی مثلث‌بندی روی یک شی را انجام و سپس مختصات مثلث‌های حاصل از طریق مختصات کندال (۱۹۸۴) بدست آورد او با تعدادی نقطه روی کره که بین آن‌ها همبستگی وجود دارد، روبه‌رو خواهد شد. آنگاه، مدل‌بندی رگرسیونی شکل بر اساس موقعیت این نقاط و سپس بازخوانی ارتباط اشکال موضوع تحقیق آماری است که مقاله حاضر سعی در پاسخ به آن دارد.

مدل‌های رگرسیونی بر روی کره در ابعاد بالا قبلاً توسط چانگ (۱۹۸۶) و ریوست (۱۹۸۹) ارائه شده‌اند. آن‌ها مدل‌های دورانی را روی کره معرفی کردند و به بررسی برخی از ویژگی‌های این مدل‌ها پرداختند. اخیراً، فیشبوق و همکاران (۲۰۱۳) مدلی رگرسیونی در فضای شکل که از تعمیم مدل رگرسیونی خطی ساده به‌دست می‌آید را معرفی کردند. آن‌ها با استفاده از مدل رگرسیون خطی و ژئودزیک شکل به برآورد پارامترهای مدل خود پرداختند. تقریباً با همین نگاه، کائو و همکاران (۲۰۱۴) استفاده از مدل رگرسیونی چندمتغیره برای داده‌های شکل روی صفحه را پیشنهاد کردند. آن‌ها با معرفی یک تابع رگرسیون برداری به پیش‌بینی شکل داده‌های پیکربندی صورت انسان پرداختند. هینکل و همکاران (۲۰۱۴) با گسترش رگرسیون چندجمله‌ای در منیفلد ریمانی به معرفی مدل‌های رگرسیونی برای داده‌های شکل پرداختند. آن‌ها همچنین به معرفی مدل ژئودزیک و اسپلاینی براساس چندجمله‌ای‌ها در منفیلد کره پرداخته و آن‌ها را با استفاده از داده‌های

[1] Triangulation
[2] Voronoi tessellation
[3] Morphology

واقعی مقایسه کردند.

بررسی مدل‌های رگرسیونی شکل که در آن متغیر پاسخ و پیشگو هر دو شکل (بر اساس تعریف کندال) هستند موضوع جالب توجه‌ای در مدل‌بندی آماری است. این مهم در طی سالیان گذشته مورد توجه محققان مختلف در مسائل کاربردی بوده است. به‌نظر می‌رسد مدل انطباق پروکراستس که کندال (۱۹۸۴) برای تصویر کردن یک پیکربندی بر روی پیکربندی دیگر به‌کار گرفت، خود نوعی از مدل رگرسیونی شکل باشد. لازم به اشاره است که مدل انطباق پروکراستس با تغییر مکان، دوران و تغییر مقیاس، یک پیکربندی را روی پیکربندی دیگر منطبق می‌کند. اما از آن‌جایی که چنین مدلی تغییرات موجود در شکل را مورد بررسی قرار نمی‌دهد کارایی لازم برای مدل‌بندی رگرسیونی شکل بر اساس شکل دیگری را ندارد. از این‌رو نیاز به معرفی یک مدل رگرسیونی کارا برای شکل یکی از دغدغه‌های محققین علاقه‌مند در این حوزه بوده و هست. پژوهش حاضر کوششی برای پاسخ به این نیاز است.

برای ارائه نتایج مقاله حاضر، در بخش ۲ به معرفی مدل رگرسیونی پروکراستس پرداخته می‌شود. در بخش ۳ روش مثلثی کردن شکل و مدل رگرسیون کروی مربوطه برای مثلث تشریح می‌شود. تحلیل یک مثال واقعی با رویکرد مورد اشاره در این مقاله در بخش ۴ می‌آید. علاوه بر این، مقایسه آماری مناسبی بین چند مدل نیز انجام خواهد شد.

۲ انطباق پروکراستس

یکی از راه‌های مشخص کردن شکل یک شئ استفاده کردن از نقاطی است که بر روی سطح خارجی شئ قرار می‌گیرند. انتخاب این نقاط باید به گونه‌ای باشد که تا حد امکان اطلاعات کلی موجود در شئ را بیان کنند. واضح است که هرچه تعداد این نقاط بیشتر باشد، اطلاعات بیشتری از شئ در دسترس خواهد بود. در آمار شکل به این مجموعه نقاط متناهی که هندسه شئ را مشخص می‌کنند، نقاط شاخص و به مجموعه تمام نقاط شاخصی که برای یک شئ در نظر گرفته می‌شود، پیکربندی آن شئ می‌گویند. به‌علاوه، به‌منظور مشخص کردن پیکربندی در روی صفحه با L نقطه شاخص، از یک ماتریس $L \times ۲$ بُعدی استفاده می‌کنند. با تعریف پیکربندی در مجموعه اعداد مختلط می‌توان این ماتریس را به‌صورت یک بردار مختلط $L \times ۱$ در نظر گرفت.

به‌ازای $i = ۱, \ldots, n$، مجموعه‌ای از پیکربندی‌های دو بُعدی مستقل $X_i = (x_۱, \ldots, x_k)^T$ و $Y_i = (y_۱, \ldots, y_k)^T$ در مجموعه اعداد مختلط با ویژگی $Y_i^* \mathbf{1}_k = ۰ = X_i^* \mathbf{1}_k$ که در آن Y_i^* ترانهاده مزدوج Y_i است را در نظر بگیرید. فرض کنید هدف، انطباق X_i بر روی Y_i یا به‌عبارتی دیگر یافتن یک

مدل رگرسیونی مناسب برای Y_i بر حسب X_i است. یک معیار مناسب برای انطباق X_i به روی Y_i استفاده از تبدیلات تشابه و تفاوت است. برای جزئیات بیشتر در این زمینه به فصل سوم درایدن و ماردیا (۱۹۹۸) مراجعه شود. با بازخوانی نمادگذاری اعداد مختلط و با پیروی از درایدن و ماردیا (۱۹۹۸)، یکی از معادلات رگرسیونی مناسب به‌صورت

$$\begin{aligned} Y_i &= (a+ib)\mathbf{1}_k + \beta e^{i\theta} X_i + \epsilon \qquad (1)\\ &= [\mathbf{1}_k, X_i]A + \epsilon_i \\ &= X_i^d A + \epsilon_i, \end{aligned}$$

است که در آن $A = (A_1, A_2)^T = (a+ib, \beta e^{i\theta})^T$ بردار دو بعدی از پارامترهای مختلط، $a+ib$ نمایانگر انتقال، $\beta > 0$ معرف مقیاس و $0 \le \theta < 2\pi$ نشانگر دوران هستند. به‌علاوه $X_i^d = [\mathbf{1}_k, X_i]$ یک ماتریس طرح $k \times 2$ و ϵ_i بردار خطای مختلط $k \times 1$ بُعدی است. به‌منظور ارائه یک انطباق مناسب، باید برآورد پارامتر A با می‌نیمم‌سازی تابع مجموع توان دوم خطاهای[۴]

$$\sum_{i=1}^{n} D^2(Y_i, X_i) = \sum_{i=1}^{n} \epsilon_i^* \epsilon_i = (Y_i - X_i^d A)^*(Y - X_i^d A)$$

به دست آید. در ادبیات آمار چندمتغیره، برآورد معادله رگرسیونی در این مسئله به انطباق کامل پروکراستس X_i روی Y_i معروف است. به عنوان مثال ماردیا و همکاران (۱۹۷۹) را ببینید. می‌توان نشان داد که برآورد ماتریس A که با $\hat{A}$ نمایش داده خواهد شد عبارتست از:

$$\hat{A} = (\hat{a}+i\hat{b}, \hat{\beta}e^{i\hat{\theta}})^T = \arg\inf \sum_{i=1}^{n} \epsilon_i^* \epsilon_i = \arg\inf \sum_{i=1}^{n} (Y_i - X_i^d A)^*(Y_i - X_i^d A).$$

به‌طور دقیق‌تر، برآورد پارامترها در انطباق کامل پروکراستس به‌صورت

$$\begin{aligned} &\hat{a} + i\hat{b} = 0, \\ &\hat{\theta} = \arg(\sum_{i=1}^{n} x_i^* y_i / n) = -\arg(\sum_{i=1}^{n} y_i^* x_i / n), \end{aligned}$$

[4] Sum of squared errors

$$\hat{\beta} = (\sum_{i=1}^{n} x_i^* y_i y_i^* x_i)^{1/2} / (\sum_{i=1}^{n} x_i^* x_i)$$

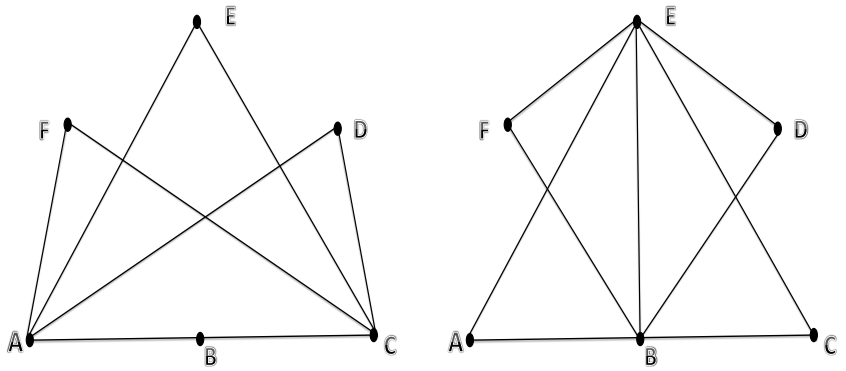
هستند (درایدن و ماردیا ، ۱۹۹۸). از آن که مدل رگرسیونی (۱) بر اساس پارامترهای مکان، دوران و مقیاس نوشته شده است، بنابراین نمی‌تواند تغییرات موجود در شکل‌ها را به‌درستی نشان دهد. به‌عبارت دیگر چنانچه تغییرات متغیر شکل پاسخ خیلی چشم‌گیرتر از تغییرات متغیر شکل پیش‌گو باشد، مدل رگرسیونی (۱) نمی‌تواند آن‌را به‌خوبی توصیف کند. به‌طور دقیق‌تر برآورد شکل متغیر پاسخ، شکلی است که از جابجایی، دوران و تغییر مقیاس شکل پیشگو به‌دست می‌آید و درنتیجه تا حد زیادی شبیه همان شکل پیشگو است. در خیلی از مواقع این موضوع مد نظر محقق نیست بلکه ایشان در پی مدلی است که تغییرات هر کدام از مولفه‌های درونی و بیرونی متغیر پاسخ را بر اساس چنین مولفه‌هایی در متغیر پیشگو توصیف کند. به‌نظر می‌رسد مثلث‌بندی پیکربندی‌ها و سپس توجه به هر یک از آن‌ها به‌عنوان افرازی از شکل شئ نهایی رویکرد مناسب‌تری برای نیل به موضوع مدل‌بندی رگرسیونی شکل باشد. چون شکل مثلث‌ها در دیدگاه کندال (۱۹۷۷) هم ارز نقاطی روی کره است مدل پیشنهادی را رگرسیون کروی نامیده‌ایم. در ادامه جزئیاتی از این مدل معرفی می‌شود.

۳ رگرسیون کروی

همانگونه که اشاره شد نمایش شکل (با ادبیاتی که کندال (۱۹۸۴) معرفی کرد) اشیاء با هر تعداد اضلاع روی $\mathbb{S}^2$ مستلزم مثلثی کردن پیکربندی است. بیان این نکته ضروری است که مثلثی کردن شکل به روش‌های مختلفی انجام می‌گیرد که برخی از آن‌ها در زلدیچ و همکاران (۱۹۸۹)، بوکشتین (۱۹۹۱) و رائو (۲۰۰۰) اشاره کرد. از روش‌های مثلثی کردن یک پیکربندی استفاده از یک محور اصلی است. در این روش دو نقطه شاخص روی یک پیکربندی به‌عنوان محور اصلی مثلث‌بندی انتخاب شده و نقطه شاخص سوم از میان سایر نقاط پیکربندی انتخاب می‌شود. دو نمونه از مثلثی کردن یک شکل بر اساس یک محور اصلی که برگرفته از بوکشتین (۱۹۹۱) است، در شکل ۱ به نمایش در آمده است. همان‌طور که ملاحظه می‌شود شکل ۱-ب بر اساس محور اصلی AC و شکل ۱-الف براساس محور اصلی BE مثلثی شده‌اند.

به‌عنوان یک ایده جدید می‌توان از روش مثلث‌بندی پیکربندی‌ها برای مدل‌بندی رگرسیونی اشیاء استفاده کرد. در این مقاله ما ابتدا پیکربندی‌ها را با استفاده از روش مثلث‌بندی براساس محور اصلی مثلث‌بندی کرده و سپس با استفاده از تبدیل کندال (۱۹۸۴) مثلث‌ها را به روی کره منتقل می‌کنیم. در

انتها نیز با بکارگیری مدل رگرسیونی کروی که در ادامه معرفی خواهیم کرد مثلث‌ها را مدل‌بندی می‌کنیم. مهمترین ویژگی روش مثلثی کردن اشیاء آن است که پس از پیش‌گویی شکل مثلث‌ها به‌راحتی می‌توان شکل کلی را با استفاده از عکس تبدیلات کندال بازسازی کرد. به‌طور دقیق‌تر، ابتدا شکل با استفاده از یک محور مثلثی شده سپس نقاط متناظر آن‌ها روی کره $\mathbb{S}^2$ مشخص می‌شوند. با چنین رویکردی به شکل اشیاء، می‌توان مدل رگرسیونی کروی که نمایانگر یک مدل رگرسیون مناسب برای بیان ارتباط بین مجموعه نقاطی از کره که متغیرهای پیشگو هستند با مجموعه دیگری از نقاط در همین کره که متغیرهای پاسخ خواهند بود، بنا کرد. از آن‌جایی که این روش به رگرسیون تک تک نقاط شکل به‌طور جداگانه می‌پردازد به‌نظر می‌رسد تغییرات شکل نیز در آن اعمال شود. از دیگر ویژگی‌های این روش آن است که چنانچه محقق علاقه‌مند به بررسی تغییرات مقیاس (اندازه) شکل باشد، می‌تواند چنین تغییراتی را از طریق اندازه محور اصلی در شکل اولیه شئ محاسبه کرد.



(ب) (الف)

شکل ۱: مثلثی کردن شکل با استفاده از یک محور اصلی الف: محور اصلی BE. ب: محور اصلی AC.

به‌طورکلی، به‌ازای $i = 1, \ldots, n$ و $d \geq 3$، برای بردار متغیرهای پیشگوی $\mathbf{x}_i = (x_{1i}, \ldots, x_{di})^T$ و متغیرهای پاسخ $\mathbf{y}_i = (y_{1i}, \ldots, y_{di})^T$، یک مدل رگرسیونی ساده مدلی به‌صورت $\mathbf{y}_i = \mu(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i$ است که در آن $\boldsymbol{\varepsilon}_i$ یک بردار از متغیر تصادفی با میانگین بردار صفر و ماتریس کوواریانس Σ_i است. همان‌طور که ملاحظه می‌شود میانگین متغیرهای پاسخ در واقع میانگین شرطی $\mu(\mathbf{x}_i) = E(\mathbf{y}_i|\mathbf{x}_i)$ است که می‌تواند صورت پارامتری، نیمه پارامتری یا حتی ناپارامتری داشته باشد (برای مشاهده جزئیات بیشتر به سبر و لی (۲۰۱۲) مراجعه شود). یکی از ساده‌ترین مدل‌های رگرسیونی روی کره مدل دورانی

است که $\mu(\mathbf{x}_i)$ صورت پارامتری دارد (چانگ، ۱۹۸۶). اگر $SO(d)$ مجموعه کلیه ماتریس‌های متعامد با دترمینان یک باشد، آنگاه در مدل دورانی پارامتری تابع $\mu(\mathbf{x}_i)$ به‌صورت $A\mathbf{x}_i$ در نظر گرفته می‌شود که در آن $A \in SO(d)$ ماتریسی $d \times d$ بعدی با مقادیر حقیقی متعلق به گروه متعامد یا یک زیر مجموعه از آن است. لذا مدل رگرسیونی کروی به‌صورت

$$\mathbf{y}_i = A\mathbf{x}_i + \boldsymbol{\varepsilon}_i \qquad (۲)$$

است، که در آن A یک دوران صلب بوده و $\boldsymbol{\varepsilon}_i$ متغیری تصادفی معرف خطا روی کره است. یکی از مناسب‌ترین توزیع‌های مورد استفاده روی کره که بتواند تغییرات $\boldsymbol{\varepsilon_i}$ها را توصیف کند توزیع فون میزس-فیشر است که اولین بار توسط ماردیا (۱۹۷۲) معرفی شد. نحوه برآورد پارامتر A در مدل رگرسیونی (۲) به‌وسیله مکنزی (۱۹۵۷) و استیفنز (۱۹۷۹) و با استفاده از روش کمترین توان‌های دوم خطا تشریح شد که جزئیاتی از آن در ادامه می‌آید.

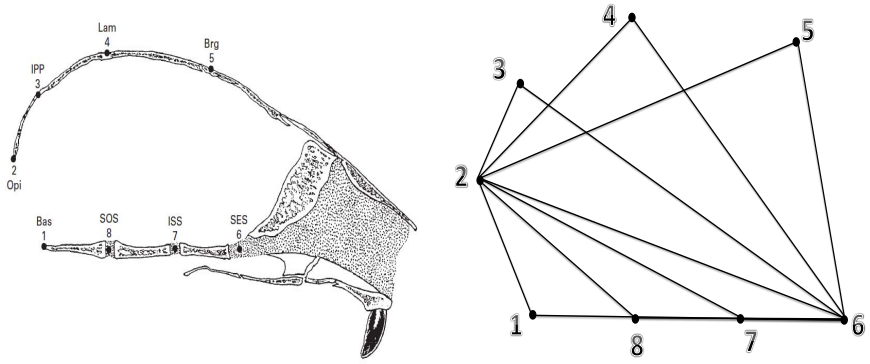
فرض کنید $C = E[\sum_{i=1}^{n} \mathbf{y}_i\mathbf{x}_i^T] \in \mathbb{R}^{d\times d}$. تجزیه ویژه مقدار پیراسته[5] این ماتریس به‌صورت $C = U\Lambda V^T$ است که $U, V \in SO(d)$ و Λ ماتریس قطری با درایه‌های $\lambda_1 \geq \ldots \geq \lambda_{d-1} \geq |\lambda_d|$ هستند. آنگاه یک دوران بهینه به‌صورت $\hat{A} = UV^T$ است. چانگ (۱۹۸۶) نشان داد که این برآوردگر، یک برآوردگر سازگار قوی بوده و دارای توزیع مجانبی نرمال است. همچنین او نشان داد که تبدیل نمایی معکوس $\hat{A}^T\hat{A}$ به‌طور مجانبی دارای توزیع نرمال چندمتغیره است. توجه شود که نتایج حاصل به این شرط وابسته است که ماتریس $\frac{1}{n}\sum_{i-1}^{n} \mathbf{x}_i\mathbf{x}_i^T$ به یک ماتریس تمام رتبه Σ همگرا باشد. شین و همکاران (۲۰۰۱) به این نکته اشاره کردند که رتبه Σ را می‌توان تا $d-1$ پایین آورد که در حقیقت یک شرط لازم برای سازگاری قوی $\hat{A}$ است.

۴ تحلیل داده‌های جمجمه موش

در بخش ۲ مدل رگرسیونی پروکراستس کامل، تشریح و در بخش ۳ روشی بر مبنای مدل‌بندی رگرسیونی شکل از طریق مثلث‌بندی پیکربندی و همچنین استفاده از رگرسیون کروی پیشنهاد شد. به‌کارگیری این دو مدل برای داده‌های واقعی و همچنین مقایسه آن‌ها می‌تواند معیاری برای ارزیابی عملکرد روش پیشنهادی باشد. به‌همین منظور، در این مقاله زیرمجموعه‌ای از داده‌های شکل جمجمه ۲۰ موش آزمایشگاهی که اولین

[5] Modified singular value decomposition

بار توسط بوکشتین (۱۹۹۱) مورد تحلیل قرار گرفت انتخاب می‌شود و مدل‌بندی رگرسیونی شکل بر اساس دو روش ارائه شده انجام خواهد گرفت. این اشکال مربوط به پیکربندی جمجمه موش‌ها در روز ۷-ام و ۱۵۰-ام از تولد آن‌ها است. نمونه‌ای از جمجمه موش‌ها در شکل ۲-ب به نمایش در آمده است.



(ب) (الف)

شکل ۲: الف: نحوه مثلث‌بندی با استفاده از نقاط شاخص ۲ و ۶. ب: نمونه‌ای از جمجمه موش‌های آزمایشگاهی.

بوکشتین (۱۹۹۱) توصیه کرد به‌منظور مثلثی کردن شکل با استفاده از یک محور، دو مسئله در نظر گرفته شود. نکات مورد اشاره ایشان عبارتند از: محور اصلی تا حد امکان از نقاط شاخص دور از هم انتخاب شود؛ محورهای انتخابی از درون پیکربندی عبور کند. با در نظر گرفتن این نکات یک انتخاب مناسب برای محور اصلی مثلث‌بندی، محور به‌دست آمده از خط واصل بین دو نقطه ۲ و ۶ است. با توجه به محور اصلی انتخاب شده نحوه مثلث‌بندی جمجمه‌ها در شکل ۲-الف نشان داده شده است. البته باید یادآوری کرد که رویکرد بوکشتین مثلث‌بندی پیکربندی‌ها روی صفحه و تحلیل آن در صفحه بوده است. با توجه به این‌که فضای شکل فضای ابر کره است لذا ما بر این باوریم که مثلث‌ها باید در فضای کره بررسی شوند. بنابراین با تلفیق روش بوکشتین (۱۹۹۱) و کندال (۱۹۸۴) و استفاده از مدل دورانی کروی به مدل‌بندی رگرسیونی داده‌های موش آزمایشگاهی می‌پردازیم.

به‌منظور اثبات ادعای بوکشتین در عملکرد مناسب نقاط دورتر در تحلیل‌های آماری، می‌توان تمامی مثلث‌بندی‌های ممکن از شکل را انجام و مدل رگرسیونی (۲) را برای مثلث‌ها به‌کار برد. بر اساس محورهای مختلف، مثلث‌بندی‌های گوناگونی به‌وجود می‌آید که مقدار مجموع توان دوم خطاهای آن‌ها در جدول ۱ ارائه شده است.

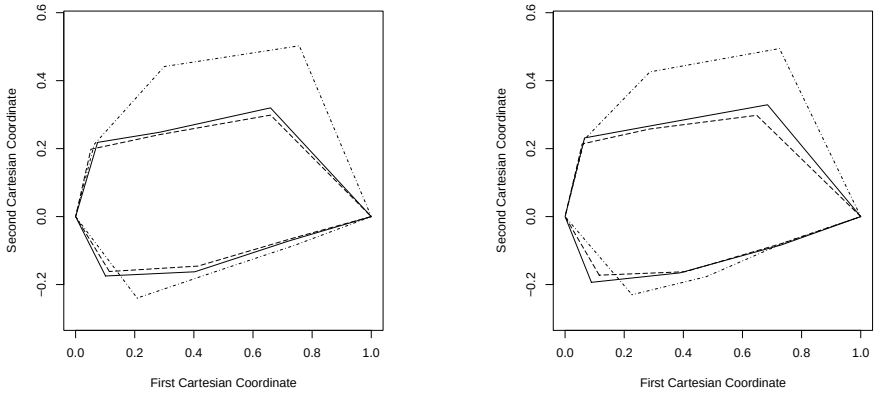
جدول ۱: مقدار مجموع توان دوم خطا با انتخاب جفت نقاط شاخص به‌عنوان محور اصلی مثلث‌سازی

نقاط شاخص	۱	۲	۳	۴	۵	۶	۷	۸
۱	—	۹۹٫۰۰	۱۳٫۷۶	۱۵٫۳۱	۳٫۶۰	۴٫۱۸	۱۰٫۵۲	۳۴٫۰۸
۲	—	—	۳۴٫۰۸	۲۰٫۹۱	۲٫۹۳	۱٫۵۷	۲٫۸۶	۵٫۴۰
۳	—	—	—	۱۳۷٫۲۶	۵٫۳۴	۱٫۶۲	۲٫۶۸	۳٫۸۷
۴	—	—	—	—	۱۱٫۲۲	۵٫۶۰	۸٫۴۴	۹٫۳۶
۵	—	—	—	—	—	۴۲٫۸۲	۳۰٫۶۰	۴٫۸۲
۶	—	—	—	—	—	—	۲۷٫۹۷	۸٫۱۷
۷	—	—	—	—	—	—	—	۴۵٫۸۵

همانگونه که از این جدول مشاهده می‌شود، مقدار مجموع توان دوم خطا برای مدل رگرسیونی (۲) بر اساس مثلث‌سازی شکل بر پایه محور اصلی و با استفاده از دو نقطه شاخص ۲ و ۶ کمترین و برابر ۱٫۵۷ است. پس از برازش مدل (۱) مقدار برآورد پارامترها برابر $\hat{\theta} = -۰٫۰۰۶$ رادیان و $\hat{\beta} = ۰٫۲۱$ به‌دست آمدند. همچنین مجموع توان دوم خطا برای مدل (۱) برابر ۱۳٫۹۰ است که بسیار بیشتر از مجموع توان دوم خطای مدل پیشنهادی است. پس از انجام رگرسیون برای داده‌های شکل، به‌منظور ارزیابی عملکرد روش پیشنهادی‌مان از آزمون گودال (۱۹۹۱) برای بررسی میزان انطباق مقادیر واقعی پاسخ (ساختار هندسی شکل جمجمه در روز ۱۵۰-ام) با مقادیر پیش‌گویی (حاصل از انجام مدل رگرسیون کروی شکل) متناظر آن‌ها در همان روز استفاده کردیم. مقدار آماره این آزمون برابر $۱٫۶ \times ۱۰^{-۴}$ و p-مقدار متناظر خیلی نزدیک به یک شد. بنابراین فرضیه برابری میانگین شکل مبتنی بر مقادیر واقعی و پیش‌گویی در سطح ۵٪ تایید می‌شود. همچنین این آزمون نشان می‌دهد که مدل رگرسیون کروی شکل پیشنهاد شده در این مقاله از دقت بالایی برخوردار است. واضح است که استفاده از معیارهای مرسوم رگرسیونی مانند ضریب تعیین برای این امر ترجیح داده می‌شود. اما به‌دلیل این‌که روش‌هایی برای محاسبه این معیار و کمیت‌های دیگر رگرسیونی تاکنون در حوزه رگرسیون شکل معرفی نشده است بررسی چنین امری به تحقیقات آتی در این حوزه موکول می‌شود.

برای نمایش کاربردی از مدل پیشنهادی‌مان در این مقاله پیش‌گویی ساختار هندسی جمجمه دو موش (شماره ۱ و ۲) در روز ۱۵۰-ام با استفاده اطلاعات موجود در متغیر تبینی (روز ۷-ام) را مدنظر قرار دادیم. اگرچه تغییر شکل جمجمه متأثر از عوامل متعددی می‌باشد اما انتظار می‌رود مدل رگرسیون کروی شکل پیشنهادی در این مقاله نیز بتواند ویژگی‌های هندسی مربوط به آن‌را توصیف کند. نمایشی از ساختار هندسی جمجمه دو موش انتخابی در روزهای ۷-ام و ۱۵۰-ام همراه با پیش‌گویی شکل جمجمه در روز

۱۵۰-ام در نمودار ۳ آمده است. همان‌طورکه ملاحظه می‌شود انطباق نسبتاً مناسبی بین مقادیر واقعی و پیش‌گویی (هر دو برای روز ۱۵۰-ام) وجود دارد.



(ب) (الف)

شکل ۳: شکل جمجمه دو موش شماره الف: ۱ و ب: ۲ در روز ۷-ام (نقطه-خط)، روز ۱۵۰-ام (خطوط بریده) و پیش‌گویی ساختار هندسی آن‌ها (خطوط ممتد).

بحث و نتیجه‌گیری

اغلب مدل‌بندی‌های رگرسیونی شکل در فضای مماسی یا روی صفحه انجام می‌گیرند. با توجه به این‌که فضای شکل یک ابر کره است، در نظر گرفتن مدل رگرسیونی مناسب در همین فضا ضروری به‌نظر می‌رسد. اگرچه مدل‌بندی رگرسیونی در فضای شکل به‌دلیل ماهیت غیراقلیدسی فضای مشاهدات با دشواری‌هایی روبه‌رو است، اما تسهیلات فراهم‌شده توسط این مدل قابلیت‌های منحصر بفردی خواهد داشت. در این مقاله پیشنهاد شد که یکی از رویکردهای ساده‌سازی صورت مسئله، مثلث‌بندی شکل هندسی اشیاء و استفاده از فضای کره دوبعدی است. این ایده کمک کرد که پس از انجام تحلیل‌ها بر روی مثلث‌ها، شکل به راحتی بازسازی شود. مدل رگرسیون کروی ویژگی بارز قرار داشتن روی کره تحت تبدیلات متعامد را داشت. عملکرد مدل پیشنهادی بر روی مجموعه داده واقعی اعتبار آن را در مقایسه با مدل‌های مرسوم قبلی نشان داد. با الگوبرداری از مقاله حاضر، تعمیم مدل رگرسیونی کره بر روی ابر کره‌های دیگر، یافتن مدل‌های

نیرومندتر و ارتباط رگرسیونی اشکال با متغیرهای دودویی و طبقه‌بندی‌شده می‌توانند موضوعات مناسبی برای تحقیقات آتی در حوزه مدل‌بندی رگرسیونی کروی باشند.

تقدیر و تشکر

نویسندگان کمال تشکر و قدردانی از داوران محترم مقاله و ویراستار مجله که با پیشنهادات ارزنده خود باعث بهبود مقاله شدند را دارند.

مراجع

Bookstein, F. L. (1986), Size and Shape Spaces for Landmark Data in Two Dimensions (With Discussion), *Statistical Sciences*, **1**, 181-242.

Bookstein, F. L. (1991), *Morphometric Tools for Landmark Data: Geometry and Biology*, Cambridge University Press, New York.

Cao, X., Wei, Y., Wen, F. and Sun, J. (2014), Face Alignment by Explicit Shape Rgression, *International Journal of Computer Vision*, **107**, 177-190.

Chang, T. (1986), Spherical Regression, *Annals of Statistics*, **14**, 907-924.

Dryden, I. L. and Mardia, K. V. (1998), *Statistical Shape Analysis*, John Wiley, Chichester.

Fishbaugh, J., Prastawa, M., Gerig, G. and Durrleman, S. (2013), Geodesic Shape Regression in the Framework of Currents, In Proceedings of *International Conference on Information Processing in Medical Imaging* , 718-729.

Goodall, C. (1991), Procrustes Methods in the Statistical Analysis of Shape, *Journal of Royal Statistical Society: Series B*, **53**, 285-339.

Hinkle, J., Fletcher, P.T. and Joshi, S. (2014), Intrinsic Polynomials for Regression on Riemannian Manifolds, *Journal of Mathematical Imaging and Vision*, **50**, 32-52.

Kendall, D. G. (1977), The Diffusion of Shape, *Advances in Applied Probability*, **9**, 428-430.

Kendall, D. G. (1984), Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces, *Bulletin of the London Mathematical Society*, **16**, 81-121.

MacKenzie, J. K. (1957), The Estimation of an Orientation Relationship, *Acta Cryst.*, **10**, 61-62.

Mardia, K.V., (1972), *Statistics of Directional Data*, Academic Press, London.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press, London.

Rao, C. R. (2000), A Note on Statistical Analysis of Shape Through Triangulation of Landmarks, *Proceedings of the National Academy of Sciences*, **97**, 2995-2998.

Rivest, L. P. (1989), Spherical Regression for Concentrated Fisher-von Mises Distribution, *Annals of Statistics*, **17**, 307-317.

Seber, G. A. and Lee, A. J. (2012), *Linear Regression Analysis*, John Wiley & Sons, New Jersey.

Shin, H. H., Takahara, G. K., and Murdoch, D. J. (2001), Uniqueness, Consistency and Optimality in Spherical Regression Experiments, *Statistics and Probability Letters*, **54**, 61-65.

Stephens, M. A. (1979), Vector Correlation, *Biometrika*, **66**, 41-48.

Stoyan, D., and Stoyan, H. (1994), *Fractals, Random Shapes, and Point Fields: Methods of Geometrical Statistics*, John Wiley & Sons, Chichester.

Zelditch, M. L., Debry, R. W., and Straney, D. O. (1989), Triangulation-Measurement Schemes in the Multivariate Analysis of Size and Shape, *Journal of Mammalogy*, **70**, 571-579.