تصادفی کردن پاسخ سوالات حساس چند گزینه‌ای و برآورد نسبت گزینه‌های تقلب دانشجویان در دانشگاه شهید چمران اهواز

سید محمد رضا علوی [۱]

چکیده:

در نمونه گیری وقتی با سوال حساس مواجه هستیم پاسخگو معمولا پاسخ واقعی را بیان نمی‌کند. روش پاسخ‌های تصادفی شده با ادغام سوال دیگر در سوال حساس این امکان را فراهم می‌کند که پاسخگو اطمینان داشته باشد که پاسخ واقعی علیه او به کار نمی‌رود. با استفاده از پاسخ‌های تصادفی شده از یک نمونه گیری و کاربرد قاعده احتمال کل می‌توان نسبت گزینه‌های حساس را برآورد کرد. در این مطالعه برآورد نسبت گزینه‌های مختلف یک سوال چند گزینه‌ای با استفاده از تصادفی کردن پاسخ‌ها توسط سوال دو گزینه‌ای بررسی شده و جهت ارزیابی، برنامه‌هایی در محیط S-PLUS نوشته شده است. با استفاده از روش مذکور نسبت گزینه‌های سوال حساس در باره‌ی تقلب کردن دانشجویان در دانشگاه شهید چمران اهواز برآورد شده است.

واژه‌های کلیدی: سوال حساس چند گزینه‌ای، روش پاسخ‌های تصادفی شده، آزمایش برنولی، گزینه‌های تقلب دانشجویان.

۱ مقدمه

در بسیاری از آمارگیری‌های نمونه‌ای باید اطلاعاتی را که دارای ماهیت حساس هستند از افرادی که در نمونه انتخاب شده‌اند استخراج کرد. مثلا در بررسی‌های مربوط به برنامه تنظیم خانواده ممکن است پرسیدن سوالاتی درباره استفاده از داروها، راهکارهای جلوگیری از بارداری یا تاریخچه سقط جنین ضروری باشد. یا در بررسی‌های اجتماعی و فرهنگی ممکن است درصد معتادان، بزهکاران و سابقه سرقت مورد علاقه باشد، لذا پرسیدن سوالاتی که ماهیت حساس دارند ضروری به نظر می‌رسد. اما این قبیل سوالات ممکن است برای بعضی افراد ترساننده یا عذاب آور باشد و پاسخگو به راحتی پاسخ ندهد. برای اجتناب از بی پاسخی بیش از حد یا پاسخ‌های گمراه کننده روش‌هایی به نام پاسخ‌های تصادفی شده ابداع شده است[۴]. اولین بار یک نوع آن را وانر[۱۰] در سال ۱۹۶۵ معرفی کرد. دراین بخش روش تصادفی کردن پاسخ‌ها برای سوال چند گزینه‌ای و برآورد نسبت آنها مطالعه می‌شود در بخش دوم روش

[۱] گروه آمار دانشگاه شهید چمران اهواز

برای سوال پنج گزینه‌ای تقلب دانشجویان شرح داده شده و در بخش سوم با اجرای نتایج بخش قبل و جمع آوری داده‌های واقعی در دانشگاه شهید چمران نسبت گزینه‌های تقلب برآورد می‌شود.

۱.۱ روش پاسخ‌های تصادفی شده جهت استخراج اطلاعات حساس

فرض کنید در سوال حساس k گزینه‌ای، گزینه‌ها را با $A_۱, A_۲ \dots, A_k$ نشان دهیم و $\theta_۱, \theta_۲, \dots, \theta_k$ به ترتیب نسبت گزینه‌های سوال حساس باشند. فرض کنید جامعه U توسط گزینه‌ها به k زیر جامعه $U_۱, U_۲, \dots, U_k$ افراز شده باشد. واضح است که θ_i نسبت تعلق به U_i می‌باشد و $\sum_{i=۱}^{k} \theta_i = ۱$.

چون k پارامتر داریم که مجموع آنها یک است لذا برای برآورد این نسبت‌ها $k-۱$ ترفند تصادفی به کارمی‌بریم. در هر ترفند از پاسخگو تقاضا می‌شود فقط به یکی از سوال‌های $Q_۱, Q_۲, \dots, Q_k$ پاسخ دهد که Q_j به صورت زیر مطرح می‌شود:

Q_j: آیا شما به زیر جامعه jام تعلق دارید؟

تحت ترفند i، شانس این که سوال Q_j انتخاب شود q_{ij} است که j می‌تواند مقادیر $۱, ۲, \dots$ و k و i مقادیر $۱, ۲, \dots$ و $k-۱$ را اختیار کند و $\sum_{j=۱}^{k} q_{ij} = ۱$ و $۰ \leq q_{ij} \leq ۱$.

بنا به قاعده احتمال کل برای ترفند i احتمال پاسخ بله برابر است با

$$p_i = P(yes|i) = \sum_{j=۱}^{k} q_{ij}\theta_j \qquad i = ۱, ۲, \dots, k-۱$$

از حل $k-۱$ معادله بالا و معادله $\sum_{i=۱}^{k} \theta_i = ۱$ در صورتی که ماتریس ضرائب A معکوس داشته باشد، داریم:

$$\Theta = A^{-۱}B$$

که بردار B و ماتریس A به صورت زیر تعریف می‌شوند:

$$B = \{b_i\}, b_i = p_i - q_{ik} \quad , \quad A = \{q_{ij} - q_{ik}\}$$

$$i = ۱, ۲, \dots, k-۱, j = ۱, ۲, \dots, k$$

تحت ترفند i پاسخ فرد یک متغیر برنولی با احتمال پیروزی p_i است، لذا اگر یک نمونه تصادفی n_i از جامعه تحت ترفند i داشته باشیم نسبت پاسخ‌های بله در نمونه یعنی $\hat{p_i}$، یک برآورد نااریب نسبت پاسخ بله تحت ترفند i می‌باشد. در نتیجه یک برآورد نااریب برای بردار Θ به صورت زیر است:

$$\hat{\Theta} = A^{-۱}\hat{B} \qquad (۱)$$

که بردار $\hat{B}$ به صورت زیر تعریف می‌شود:

$$\hat{B} = \{\hat{b_i}\} \ , \ \hat{b_i} = \hat{p_i} - q_{ik} \quad i = ۱, ۲, \dots, k-۱ \qquad (۲)$$

ماتریس کواریانس $\hat{\Theta}$ وقتی که نمونه گیری تحت ترفندهای مختلف مستقل باشند برابر است با:

$$Cov(\hat{\Theta}) = A^{-۱} Diag(v_۱, v_۲, \dots, v_{k-۱}) A^{-۱} \qquad (۳)$$

که $v_i = \frac{p_i(۱-p_i)}{n_i}$، یک برآورد برای این ماتریس کواریانس به صورت زیر ارائه می‌شود:

$$\hat{Cov}(\hat{\Theta}) = A^{-۱} Diag(\hat{v}_۱, \hat{v}_۲, \dots, \hat{v}_{k-۱}) A^{-۱} \qquad (۴)$$

که در آن $\hat{v}_i = \frac{\hat{p}_i(۱-\hat{p}_i)}{n_i}$.

تذکر ۱) با توجه به این که q_{ij}ها معلوم و تحت اختیار طراح نمونه گیری هستند لذا باید مناسب انتخاب

شوند واضح است که اگر $q_{i1} = q_{i2} = \cdots = q_{ik}$ برای ترفندهای مختلف باشد در آن صورت ماتریس A معکوس نخواهد داشت. لذا توصیه می‌شود حداقل دو تا از آنها مختلف انتخاب شوند. در بخش دوم برای سادگی در اجرای نمونه گیری از مقادیر ۰/۴، ۰/۱۵، ۰/۱۵، ۰/۱۵، ۰/۱۵ و ۰/۱۵ جهت احتمال انتخاب سوالات در ترفندهای مختلف استفاده شد.

تذکر ۲) باید توجه داشت که ممکن است جواب $\hat{\Theta} = A^{-1}B$ غیر قابل قبول باشد و آن وقتی است که یکی از مولفه‌ها در خارج از بازه [۰,۱] باشد. به نظر می‌رسد که هر چه q_{ij} ها به هم نزدیک انتخاب شوند احتمال وقوع چنین حالتی زیاد می‌شود.

۲ برآورد گزینه‌های تقلب دانشجویان

تقلب در مراکز علمی و آموزشی خصوصا در دانشگاه‌ها از جمله آفت‌های علمی این مراکز است. لذا تحقیق و پژوهش در این مورد از اهمیت خاصی برخوردار است. متاسفانه در ایران تحقیقات مورد توجهی در این خصوص انجام نگرفته است. علوی و چینی پرداز[۱] نرخ تقلب دانشجویان را در کل دانشگاه شهید چمران اهواز به روش پاسخ‌های تصادفی شده با سوالات دو گزینه‌ای ۰/۰۳۷۸ $\pm$ ۰/۲۳۳ برآورد و در دانشکده‌های مختلف به روش مذکور نرخ تقلب را مقایسه کردند. در منابع خارجی تحقیقات مناسبی در این زمینه وجود دارد. اکثر این تحقیقات توسط توزیع پرسش‌نامه‌های بدون نام بین دانشجویان یا ارسال پستی به افراد صورت گرفته است.

بعضی از مطالعات کوشش می‌کنند که تقلب‌های ثابت و همیشگی را شناسایی کنند. لابف [۷] گزارش داد که حدود ۷٪ از دانشجویان تقلب کننده همیشگی هستند. مک کاب [۸] گزارش داد که ۱۹٪ از دانشجویان بیش از پنج بار تقلب کرده‌اند و اوی و برملی [۳] گزارش دادند که ۲۲٪ از دانشجویان بیش از پنج بار تقلب کرده‌اند. کولیسن [۲] درصد دانشجویانی را که اصرار در تقلب دارند از ۵٪ برای رشته‌های علوم تا ۵۰٪ برای رشته‌های علوم اقتصادی گزارش داد. شیرس و دایتون [۹] دریافتند که گزارشات برای پنج رفتار تقلب آمیز نرخی بین ۳۹٪ تا۸۳٪ را نشان می‌دهد و ادعا کردند که هنگامی که از روش پاسخ‌های تصادفی استفاده می‌شود می‌توان ریسک اشتباه را کاهش داد[۹]. کرکولیت و سیگموند [٦] نشان دادند که صحت اطلاعات روش پاسخ‌های تصادفی شده از روش معمولی بهتر است. هجدن و واندر گیلز [۵] نیز روش پاسخ‌های تصادفی شده را با روش معمولی مقایسه کردند.

فرض کنید می‌خواهیم پاسخ سوال پنج گزینه‌ای زیر را از دانشجویان دانشگاه اخذ کنیم: تا کنون چند بار در امتحانات دانشگاه تقلب کرده‌اید؟

۱)هیچ ۲)فقط یکبار ۳)فقط دو بار ۴)سه یا چهار بار ۵)پنج بار یا بیشتر

اگر θ_1 نسبت دانشجویانی باشد که تقلب نمی‌کنند و $\theta_2, \theta_3, \theta_4, \theta_5$ به ترتیب نسبت دانشجویانی باشند که فقط یک بار، فقط دو بار، سه یا چهار بار و پنج بار یا بیشتر تقلب می‌کنند، در آن صورت ممکن است علاقه‌مند به برآورد این نسبت‌ها در دانشگاه با استفاده از روش

پاسخ‌های تصادفی در یک طرح نمونه گیری به قرار زیر باشیم. ابتدا پنج سوال دو حالتی زیر را در نظر می‌گیریم:

$Q_۱$: آیا شما از دانشجویانی محسوب می‌شوید که تا کنون در امتحانات دانشگاه تقلب نکرده‌اند؟

$Q_۲$: آیا شما از دانشجویانی محسوب می‌شوید که تا کنون در امتحانات دانشگاه فقط یک بار تقلب کرده‌اند؟

$Q_۳$: آیا شما از دانشجویانی محسوب می‌شوید که تا کنون در امتحانات دانشگاه فقط دو بار تقلب کرده‌اند؟

$Q_۴$: آیا شما از دانشجویانی محسوب می‌شوید که تا کنون در امتحانات دانشگاه سه یا چهار بار تقلب کرده‌اند؟

$Q_۵$: آیا شما از دانشجویانی محسوب می‌شوید که تا کنون در امتحانات دانشگاه پنج بار یا بیشتر تقلب کرده‌اند؟

سپس چهار نمونه تصادفی با جایگذاری مستقل از هم هر کدام تحت ترفند خاصی از جامعه انتخاب کرده و با توجه به ترفند مورد نظر از دانشجویان تقاضا می‌شود فقط به یکی از پنج سوال فوق بدون این که کسی متوجه شود پاسخ داده و نتیجه را در پاسخنامه علامت زند. در ترفند اول دانشجو یک عدد تصادفی سه رقمی انتخاب می‌کند؛

اگر دو رقم سمت راست آن از ۰۰ تا ۳۹ باشد فقط به سوال $Q_۱$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۴۰ تا ۵۴ باشد فقط به سوال $Q_۲$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۵۵ تا ٦۹ باشد فقط به سوال $Q_۳$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۷۰ تا ۸۴ باشد فقط به سوال $Q_۴$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۸۵ تا ۹۹ باشد فقط به سوال $Q_۵$ پاسخ می‌دهد

در این ترفند $q_{۱۲} = q_{۱۳} = q_{۱۴} = q_{۱۵} = ۰/۱۵$ و $q_{۱۱} = ۰/۴$.

در ترفند دوم دانشجو یک عدد تصادفی سه رقمی انتخاب می‌کند؛

اگر دو رقم سمت راست آن از ۰۰ تا ۱۴ باشد فقط به سوال $Q_۱$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۱۵ تا ۵۴ باشد فقط به سوال $Q_۲$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۵۵ تا ٦۹ باشد فقط به سوال $Q_۳$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۷۰ تا ۸۴ باشد فقط به سوال $Q_۴$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۸۵ تا ۹۹ باشد فقط به سوال $Q_۵$ پاسخ می‌دهد

در این ترفند $q_{۱۱} = q_{۱۳} = q_{۱۴} = q_{۱۵} = ۰/۱۵$ و $q_{۱۲} = ۰/۴$.

در ترفند سوم دانشجو یک عدد تصادفی سه رقمی انتخاب می‌کند؛

اگر دو رقم سمت راست آن از ۰۰ تا ۱۴ باشد فقط به سوال $Q_۱$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۱۵ تا ۲۹ باشد فقط به سوال $Q_۲$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۳۰ تا ٦۹ باشد فقط به سوال $Q_۳$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۷۰ تا ۸۴ باشد فقط به سوال

سوال $Q_۴$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۸۵ تا ۹۹ باشد فقط به سوال $Q_۵$ پاسخ می‌دهد

در این ترفند $q_{۱۱} = q_{۱۲} = q_{۱۴} = q_{۱۵} = ۰/۱۵$ و $q_{۱۳} = ۰/۴$.

در ترفند چهارم دانشجو یک عدد تصادفی سه رقمی انتخاب می‌کند؛

اگر دو رقم سمت راست آن از ۰۰ تا ۱۴ باشد فقط به سوال $Q_۱$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۱۵ تا ۲۹ باشد فقط به سوال $Q_۲$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۳۰ تا ۴۴ باشد فقط به سوال $Q_۳$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۴۵ تا ۸۴ باشد فقط به سوال $Q_۴$ پاسخ می‌دهد

اگر دو رقم سمت راست آن از ۸۵ تا ۹۹ باشد فقط به سوال $Q_۵$ پاسخ می‌دهد

در این ترفند $q_{۱۱} = q_{۱۲} = q_{۱۳} = q_{۱۵} = ۰/۱۵$ و $q_{۱۴} = ۰/۴$.

در هر کدام از این نمونه‌ها پاسخگو فقط در پاسخنامه زیر علامت می‌زند.

پاسخنامه:
بله □
خیر □

بعد از جمع آوری پاسخنامه‌ها در هر ترفند نسبت پاسخ بله در آن ترفند یعنی $\hat{p}_i$ها را محاسبه می‌کنیم. با توجه به q_{ij}ها ماتریس A به صورت زیرمی‌باشد:

$$A = \{q_{ij} - q_{i۵}\} = ۰/۲۵ I_۴$$

که $I_۴$ ماتریس همانی مرتبه ۴ است. لذا

$$\hat{\Theta} = ۴\hat{B} = ۴\{\hat{b_i}\}$$

که $\hat{b_i} = \hat{p}_i - ۰/۱۵$ برای $i = ۱,۲,۳,۴$ در نتیجه داریم:

$$\hat{\theta_i} = ۴(\hat{p}_i - ۰/۱۵) i = ۱,۲,۳,۴ \quad (۵)$$
$$\hat{\theta_۵} = ۱ - \sum_{i=۱}^{۴} \hat{\theta}_i$$

برای مثال فرض کنید در یک نمونه تصادفی ۱۰۰ تایی تحت ترفند اول ۲۸ دانشجو پاسخ بله داده‌اند و در یک نمونه تصادفی ۱۰۰ تایی دیگر مستقل از نمونه قبلی تحت ترفند دوم ۲۰ دانشجو پاسخ بله داده و به همین ترتیب در نمونه‌های ۱۰۰ تایی دیگر مستقل از نمونه‌های قبلی تحت ترفندهای سوم و چهارم به ترتیب ۱۸ و ۱۷ دانشجو پاسخ بله داده باشند، در نتیجه بردار برآورد برابر است با:

$$(\hat{\theta_۱}, \hat{\theta_۲}, \hat{\theta_۳}, \hat{\theta_۴}, \hat{\theta_۵}) = (۰/۵۲, ۰/۲, ۰/۱۲, ۰/۰۸, ۰/۰۸)$$

یعنی ۵۲% دانشجویان اصلا تقلب نمی‌کنند، ۲۰% دانشجویان فقط یک بار، ۱۲%دانشجویان فقط دو بار،۸% دانشجویان ۳ یا ۴ بار و ۸% آنها ۵ بار یا بیشتر تقلب می‌کنند. جهت ارزیابی روش با استفاده از شبیه‌سازی از یک جامعه که پارامترهای آن معلوم باشند برنامه‌هایی در محیط S-PLUS نسخه ۲۰۰۰ نوشته شد تا احتمال منفی بودن حداقل یک برآورد را محاسبه کند. با استفاده از آن ملاحظه گردید که هرچه گزینه‌های سوال بیشتر و نیز هرچه اندازه نمونه کمتر باشد احتمال منفی بودن بیشتر می‌باشد.

۳ برآورد نسبت گزینه‌های تقلب دانشجویان در دانشگاه شهید چمران اهواز

در یک نمونه گیری در سطح دانشگاه شهید چمران که در سه ماهه اول (فروردین تا خرداد) سال ۸۴ به روش فوق برای سوال مطرح شده در بخش دوم صورت گرفت تحت ترفندهای اول تا چهارم داده‌ها جمع آوری شد و خلاصه نتایج در درجدول (۱) آمده است.

جدول ۱. خلاصه داده‌های نمونه گیری به روش پاسخ‌های تصادفی شده برای سوال پنج گزینه‌ای

	ترفند اول	ترفند دوم	ترفند سوم	ترفند چهارم
تعداد بله	۱۵۲	۷۰	۵۶	۶۳
اندازه نمونه	۴۳۵	۴۱۹	۳۶۲	۳۹۵
نسبت بله	۰/۳۹۴	۰/۱۶۷	۰/۱۵۵	۰/۱۵۹

با جایگذاری در روابط (۵) داریم:

$$\hat{\Theta}' = (\hat{\theta_۱}, \hat{\theta_۲}, \hat{\theta_۳}, \hat{\theta_۴})$$
$$= (۰/۷۹۸, ۰/۰۶۸۳, ۰/۰۱۸۸, ۰/۰۳۸۰)$$

و

$$\hat{\theta_۵} = ۱ - \sum_{i=۱}^{۴}(\hat{\theta_i} = ۰/۰۷۷۳$$

در نتیجه

$$\hat{\Theta}' = (\hat{\theta_۱}, \hat{\theta_۲}, \hat{\theta_۳}, \hat{\theta_۴}, \hat{\theta_۵})$$
$$= (۰/۷۹۸, ۰/۰۶۸۳, ۰/۰۱۸۸, ۰/۰۳۸۰, ۰/۰۷۷۳)$$

برآورد ماتریس واریانس کواریانس $\hat{\Theta}$ برابر است با:

$$\hat{Cov}(\hat{\Theta}) = A^{-۱} Diag(\hat{v}_۱, \hat{v}_۲, \ldots, \hat{v}_{k-۱}) A^{-۱}$$
$$= ۰/۰۱۶ Diag(۰/۵۲, ۰/۳۳, ۰/۳۶, ۰/۳۴)$$
$$= ۱۰^{-۳} Diag(۸/۴, ۵/۳, ۵/۸, ۵/۴)$$

عناصر روی قطر این ماتریس قطری برآورد واریانس برآوردهای متناظر را نشان می‌دهند. ملاحظه می‌شود که حدودا ۲۰٪ دانشجویان تقلب می کنند که موید نتیجه علوی و چینی پرداز[۱] است.

۴ نتایج و پیشنهادات

وقتی با سوال حساس چند گزینه‌ای مواجه هستیم توصیه می‌شود که از روش پاسخ‌های تصادفی شده استفاده شود. در دانشگاه شهید چمران طبق روش تصادفی شده برآورد شد که ۷۹/۸٪ دانشجویان اصلا تقلب نمی‌کنند، ۶/۸٪ فقط یک بار، ۱/۹٪ فقط دوبار، ۳/۸٪ سه یا چهار بار و ۷/۷٪ بیش از چهار بار تقلب می کنند.

تشکر: بعضی از نتایج این تحقیق مستخرج از طرح تحقیقاتی شماره ۵۱۲ دانشگاه شهید چمران (مرجع شماره [۱]) می‌باشد. از دانشگاه شهید چمران به خاطر حمایت مادی از این طرح تشکر می‌شود.

مراجع

[۱] علوی، م.ر. و چینی پرداز، ر. (۱۳۸۴). مقایسه تقلب دانشجویان در دانشکده‌ها با استفاده از روش پاسخ تصادفی شده، گزارش نهایی طرح تحقیقاتی شماره ۵۱۲ دانشگاه شهید چمران اهواز.

[2] Collison, M. (1990), Survey at rutgers suggests that cheating may be on the rise at large universities, *Chronicle of Higher Education*, A31-A32.

[3] Eve, R. and Bromley, D. (1981), Scholtastic dishonesty among college undergraduates: Parrallel tests of sociological explanations. *Youth and Society*, 13(1), 3-22.

[4] Hedayat, A.S. and Sinha Bikas K. (1991), *Design and Inference in Finite Population Sampling*, John Wiley and Sons Inc., 318.

[5] Heijden, PGM Vander, Gils, G. van, Bouts, J. and Hox, J.J. (2000), A comparison of randomized response, CASAQ, and direct questioning; Eliciting sensitive information in the context of welfare and unemployment benefit, *Sociological Methods and Research*, 28, 505-537.

[6] Kerkvliet, J., and Sigmund, C. L. (1999), Can we control cheating in the classroom?, *Journal of Economic Education*, 4, 331-343.

[7] Labeff, E., Clark, R., Hains, V. and Diekhoff, G. (1990), Situational ethics and college student cheating, *Sociological Inquiry*, 60, 190-198.

[8] McCabe, D. (1992), The influence of situational ethics on cheating among college students, *Sociological Inquiry*, 62(3), 365-373.

[9] Scheers, N.J. and Dayton, C.M. (1987), Improved estimation of academic cheating behavior using the randomized response technique, *Research in Higher Education*, 26(1), 61-69.

[10] Warner, S.L. (1965), Randomized Response: A survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, 60, 63-69.