بررسی کران‌های ضریب همبستگی پیرسون و کاربرد آن در تحلیل خسارت‌های بیمه‌ای

رحیم محمودوند[۱]

تاریخ دریافت: ۱۴۰۲/۰۷/۰۲
تاریخ پذیرش: ۱۴۰۲/۱۲/۲۳

چکیده:

در پژوهش‌های بیم‌سنجی، خسارت‌های بیمه‌ای را با توزیع‌های احتمالی مناسب مدل‌بندی می‌کنند. ازآنجاکه خسارت‌ها، پس از ارزیابی، با واحدهای پولی معین می‌شوند از توزیع‌هایی که مقادیر مثبت را اختیار می‌کنند برای مدل‌بندی آن‌ها استفاده می‌شود. علاوه بر این، با توجه به قراردادهای بیمه‌ای، خسارت‌ها در یک محدوده کران‌دار قرار می‌گیرند که بایستی در مدل‌بندی لحاظ شوند. این ویژگی‌ها در حالت یک متغیره دشواری و محدودیت چندانی ایجاد نمی‌کند؛ اما در حالت چند متغیره مسئله قدری پیچیده‌تر می‌شود. در چنین شرایطی مفصل‌ها می‌توانند مفید واقع شوند. بااین‌وجود بررسی همبستگی بین متغیرها، به‌عنوان نخستین گام در تحلیل چندمتغیره، نقش مهمی ایفا می‌کند. بر این اساس بررسی تأثیر کران‌دار بودن خسارت‌ها بر همبستگی بین آن‌ها مسئله‌ای است که در این مقاله موردتوجه قرارگرفته است. در این راستا ضریب همبستگی پیرسون، به‌عنوان متداول‌ترین شاخص برای بررسی رابطه بین متغیرها، مورداستفاده قرارگرفته است. ابتدا مسئله بر پایه ضریب همبستگی بین دو متغیر تصادفی بررسی‌شده و در ادامه بررسی‌ها بر روی برآورد گشتاوری ضریب همبستگی پیرسون انجام‌شده است. داده‌های مربوط به خسارت‌های مالی و جانی بیمه‌نامه‌های شخص ثالث در یکی از شرکت‌های بیمه ایرانی به‌عنوان یک مطالعه موردی بررسی‌شده است. کران‌های پائینی و بالایی برای پارامتر ضریب همبستگی پیرسون و برآورد گشتاوری آن به‌دست‌آمده است. کران‌های مربوط به پارامتر ضریب همبستگی با توجه به تابع مفصل به‌دست‌آمده است درحالی‌که برای برآورد ضریب همبستگی از آماره‌های ترتیبی استفاده‌شده است. علاوه بر این، با توجه به ماهیت داده‌ها، ضریب همبستگی بین خسارت‌های مالی و جانی به دو صورت محاسبه و با یکدیگر مقایسه شده‌اند. مقایسه کران‌های به‌دست‌آمده نشان می‌دهد که کران‌های $+1$ و -1 برای ضریب همبستگی پیرسون در خسارت‌های بیمه‌ای در دسترس نیست و کران‌های باریک‌تری برای این ضریب قابل ترسیم است.

واژه‌های کلیدی: آماره‌های مرتب، تحدید، گشتاور

۱ مقدمه

در ماده یک قانون بیمه مصوب سال ۱۳۱۶ هجری شمسی چنین ذکرشده است که «بیمه عقدی است که به‌موجب آن یک‌طرف تعهد می‌کند در ازای پرداخت وجه یا وجوهی از طرف دیگر در صورت وقوع یا بروز حادثه، خسارت وارده بر او را جبران نموده یا وجه معینی بپردازد. متعهد را بیمه‌گر، طرف تعهد را بیمه‌گذار، وجهی را که بیمه‌گذار به بیمه‌گر می‌پردازد حق بیمه و آنچه را که بیمه می‌شود، موضوع بیمه نامند». بر اساس این تعریف، بیمه‌گر بایستی با توجه به موضوع بیمه، حق بیمه را معین کند. این مسئله، یکی از مهم‌ترین مسائل بیمه‌گران است و برای حل آن به دانش متخصصانی چون بیم‌سنجان و آماردانان نیاز دارند. این متخصصان، به‌طورمعمول از ابزارهای احتمالی برای پاسخ دادن به این مسئله استفاده می‌کنند. بر این اساس در ادبیات حوزه بیم‌سنجی خسارت‌های مرتبط با موضوع بیمه را به‌عنوان متغیرهای تصادفی در نظر می‌گیرند و از نظریه توزیع‌ها و قضایای حدی، چون قانون قوی اعداد بزرگ، کمک می‌گیرند ([۲۴] و [۲۳]). بااین‌وجود در عمل چند نکته مهم وجود دارد که دستمایه این نوشتار شده است:

(۱) چندوجهی بودن موضوع بیمه؛

(۲) محدودیت در پرداخت خسارت از سوی بیمه‌گر؛

(۳) مثبت بودن خسارت‌ها.

بیمه‌گران برای هر قرارداد بیمه‌ای، ریسک‌های تحت پوشش را تعریف می‌کنند و معمولاً بیش از یک ریسک، در اغلب قراردادهای بیمه‌ای، تحت پوشش قرار می‌گیرد. این بدان معناست که خسارت‌های یک

[۱] گروه آمار، دانشکده علوم پایه، دانشگاه بوعلی سینا (نویسنده مسئول: r.mahmodvand@gmail.com)

قرارداد بیمه‌ای می‌تواند ناشی از چند منبع باشد که در بسیاری از موارد روی همدیگر تأثیرگذار هستند. به‌عنوان‌مثال، در پالایشگاه‌ها و صنایع فرآیندی اگر درکنار یک مخزن، آتش‌سوزی رخ دهد امکان رخداد انفجار هم وجود دارد و برعکس. پس دو ریسک انفجار و آتش‌سوزی دارای وابستگی ذاتی هستند (برای کارهای نظری در این زمینه [۱۳]-[۱۵] را ببینید). بیمه‌گران با توجه به شرایط بازار، توانگری مالی و ارزیابی ریسکی که انجام می‌دهند در قراردادهای بیمه‌ای سقف تعهدات خود را مشخص می‌کنند. این موضوع به‌طور ملموسی در بیمه‌های درمان و اموال وجود دارد و بیمه‌گذاران با آن روبرو می‌شوند؛ یعنی خسارت‌ها از لحاظ کمیت می‌توانند هر مقدار مثبتی باشند، اما از لحاظ پرداخت، از سقف معینی بیشتر نخواهند شد. گاهی نیز خسارت‌ها از پایین هم دارای کف هستند. موضوعاتی چون فرانشیز از این منظر قابل‌بحث هستند (برای توضیح نظری بیشتر کلاگمن و همکاران [۲۴] را ببینید). وجود وابستگی بین ریسک‌های تحت پوشش در یک قرارداد بیمه‌ای موجب مختل شدن انعطاف گسترده‌ای است که در به‌کارگیری ابزارهای احتمالی برای متغیرهای تصادفی مستقل وجود دارد. آماردانان و بیم‌سنجان برای غلبه بر این محدودیت تلاش‌های گسترده‌ای کردند که منجر به معرفی ابزاری به نام مفصل شد. این ابزار به‌طور رسمی در مقاله اسکلار [۲۹] معرفی شد و پس‌ازآن کارهای بسیار زیادی برای به‌کارگیری آن در حوزه‌های مختلف صورت گرفت. مرور خوبی از برخی کارهای مرتبط در مطالعه جو [۲۲] و نیلسن [۲۶] ارائه‌شده است. پژوهش‌های متعددی هم در حوزه مطالعه‌های بیمه‌ای انجام‌شده است که از مفصل برای در نظر گرفتن وابستگی بین ریسک‌ها استفاده کرده‌اند. به‌عنوان‌مثال فریز و والدز [۱۶] منبع خوبی برای آشنایی با کاربردهای مفصل برای بیم‌سنجان است. علاوه بر این پژوهش‌های متعددی توسط امبرخت و همکاران او در دهه ۹۰ میلادی و پس‌ازآن انجام‌گرفته است که در بخش بعدی به برخی از آن‌ها اشاره خواهد شد. در مطالعه گائو و لی [۱۷] می‌توان فهرست به‌روزتری از پژوهش‌هایی که از مفصل در حوزه بیمه استفاده کرده‌اند را مشاهده کرد. بااین‌وجود معرفی مفصل، بر پایه تابع احتمال صورت گرفته است و مستلزم به‌کارگیری یک تابع ریاضی مشخص است. درکنار مفصل، بررسی ارتباط بین متغیرها معمولاً با استفاده از ابزارهای توصیفی ساده، چون نمودارها و شاخص‌های ساده‌ای از قبیل ضرایب همبستگی، هم موردتوجه قرار می‌گیرد. تفاوت رویکرد در آن است که در اینجا مبنای بررسی، داده‌هایی است که از پدیده‌ها در اختیار داریم. کاربرد ضریب همبستگی پیرسون در حوزه مطالعه‌های بیمه‌ای بسیار گسترده است. در حوزه بیم‌سنجی، پیش‌پردازش‌ها برای مدل‌سازی بر پایه این شاخص قابل انجام است. علاوه بر این در مطالعه‌های از نوع مدیریتی که به دنبال بررسی عوامل مؤثر بر بخش‌های مختلف بیمه هستند هم از ضریب همبستگی خطی استفاده می‌شود. به‌عنوان‌مثال صحت و همکاران [۱] بر پایه ضریب همبستگی پیرسون به بررسی ارتباط بین نوآوری سازمانی و مزیت رقابتی در شرکت‌های بیمه پرداخته است. در مطالعه دیگری مظلومی و هاشمی [۲] برای بررسی ارتباط قابلیت اجرای راهبردها با اثربخشی آن‌ها از ضریب همبستگی پیرسون کمک گرفته است. آوانزی و همکاران [۴] مطالعه کاربردی جالبی درباره وجود همبستگی بین رشته‌های مختلف بیمه‌ای انجام داده و درباره واقعی یا موهومی بودن چنین ارتباط‌هایی بحث کرده است. در پژوهش دیگری برای بررسی ارتباط بین تقلبات در بیمه سلامت و ابتلا به کووید ۱۹ از ضریب همبستگی پیرسون استفاده‌شده است [۳۰]. اهمیت این ضریب در قیمت‌گذاری مشتقات در پژوهش‌های دیگری چون بویر و همکاران [۸] بررسی‌شده است. در برخی از مطالعات هم به موضوع وابستگی ریسک‌ها به‌صورت کلی‌تری پرداخته‌اند و از این منظر ضریب همبستگی پیرسون را به‌عنوان یک کاندید موردبررسی قرار داده‌اند ([۷] و [۲۰]). در این مقاله می‌خواهیم ببینیم که آیا مثبت و کران‌دار بودن خسارت‌های بیمه‌ای می‌تواند بر میزان برآورد شاخص ضریب همبستگی پیرسون تأثیر داشته باشد. در واقع همان‌طور که برخی از پژوهشگران هم اشاره‌کرده‌اند علیرغم کاربردهای گسترده ضریب همبستگی پیرسون با دو مسئله جدی درباره این شاخص روبرو هستیم ([۶] و [۷]):

- استفاده نابجا: عموماً مربوط به بررسی ارتباط‌های ذاتاً غیرخطی با استفاده از این شاخص است؛
- تفسیر نادرست: به ماهیت توزیع داده‌ها بستگی دارد و گاه نادیده گرفته می‌شود.

شایان یاد است که موضوع محدوده باریک‌تر از بازه $[-۱,+۱]$ برای ضریب همبستگی پیرسون بین دو متغیر تصادفی، پیش‌تر در پژوهش‌های دیگری هم در نظر گرفته‌شده است ([۵] و [۹])؛ اما در خصوص محدوده برآورد این شاخص صحبتی نشده است.

۲ نتایج نظری

فرض کنید X و Y متغیرهای تصادفی مثبت با تابع‌های توزیع F_X و F_Y و ضریب همبستگی پیرسون $\rho(X,Y)$ و تابع مفصل C باشند. آنگاه دنوئیت (۲۰۰۵) نشان دادند که برای متغیر تصادفی یکنواخت U

داریم:

$$-1 \le \frac{C(F_X^{-1}(U), F_Y^{-1}(1-U))}{\sqrt{Var(X)Var(Y)}} \le \rho(X,Y) \le \frac{C(F_X^{-1}(U), F_Y^{-1}(U))}{\sqrt{Var(X)Var(Y)}} \le 1.$$

امبرخت و همکاران [۱۲] و امبرخت [۱۱] بر این اساس بحث می‌کند که به‌عنوان‌مثال، اگر متغیر تصادفی $X \sim LN(0,1)$ و $Y \sim LN(0,2)$ باشند، آنگاه نمی‌توان یک مدل توأم با این حاشیه‌ای‌ها با هر مقدار ضریب همبستگی در بازه $[-1,+1]$ داشت. برای این حالت خاص نشان داده‌شده است که

$$-1 < \frac{e^{-\sigma}-1}{\sqrt{(e-1)(e^{\sigma^2}-1)}} \le \rho(X,Y) \le \frac{e^{\sigma}-1}{\sqrt{((e-1)(e^{\sigma^2}-1)}} < 1.$$

البته این حالت خاص پیش‌تر توسط دی‌ویوس [۱۰] به‌دست‌آمده و در مطالعه‌های مختلفی مانند رومانو و سیگل [۲۷] و شی و هوانگ [۲۸] هم به آن اشاره‌شده است. با این تفاوت که در مطالعه‌های پیشین این نتیجه با توجه به برابری زیر از هوفدینگ [۲۱] برای ضریب همبستگی

$$\rho(X,Y) = \frac{\int\int (F_{X,Y}(x,y) - F_X(x)F_Y(y))\,dxdy}{\sqrt{Var(X)Var(Y)}},$$

حاصل شد. علاوه بر این کران‌های همبستگی بین یک متغیر نرمال و یک متغیر برنولی توسط گراداشتاین [۱۸] بررسی‌شده است. همچنین باربیرو [۵] درباره کران‌های ضریب همبستگی پیرسون بین دو متغیر رتبه‌ای با یک سری محدودیت و تأثیر آن‌ها بر کران‌های این شاخص بحث کرده است. این نتایج نشان می‌دهند که می‌توان وابستگی بسیار قوی بین متغیرهای تصادفی داشت اما میزان ضریب همبستگی خطی نزدیک به صفر باشد. البته این نتیجه عجیبی نیست و به شکل‌های مختلف دیده‌شده است. به‌عنوان‌مثال اگر $X_1,\ldots,X_n$ متغیرهای تصادفی با واریانس یکسان σ^2 باشند و فرض کنیم برای هر زوج i و j رابطه $cov(X_i,X_j)=\rho$ برقرار باشد، آنگاه از نامنفی بودن واریانس مجموع این متغیرها به‌سادگی می‌توان دید که $\rho \in [-1/(n-1),1]$. علاوه بر این اگر $X_1,\ldots,X_n$ متغیرهای تصادفی مستقل با واریانس یکسان σ^2 باشند و $Y_1,\ldots,Y_n$ متغیرهای تصادفی مستقل با واریانس یکسان σ^2 باشند و فرض کنیم برای هر زوج i و j رابطه $cov(X_i,Y_j)=\sigma^2\rho$ برقرار است آنگاه نتیجه می‌شود که $\rho \in [-1/n,1/n]$. این تحدید با توجه به برقراری روابط (۱) است:

$$var\left(\sum_{i=1}^{n} X_i \pm \sum_{i=1}^{n} Y_i\right) \ge 0. \qquad (1)$$

هرچند این نتایج با در نظر گرفتن شرایط خاصی بر روی متغیرها به‌دست‌آمده‌اند، در بسیاری مواقع در عمل فراموش می‌کنیم که ممکن است شرایط مشابهی باعث تحدید بازه مقادیر ممکن ضریب همبستگی شود. همان‌طورکه گفته شد در برخی از شرایط عملی، مانند خسارت‌های بیمه‌ای، مقادیر متغیرها دارای محدودیت‌هایی هستند که ممکن است باعث ایجاد ویژگی‌های نظری خاصی بر روی شاخص‌های آماری شوند. اکنون اجازه دهید درباره برآورد گشتاوری ضریب همبستگی پیرسون بحث کنیم. می‌دانیم که اگر یک نمونه n تایی از (x_i,y_i) ها داشته باشیم، آنگاه برآورد به روش گشتاوری ضریب همبستگی پیرسون به‌صورت زیر است

$$r = \frac{s_{xy}}{s_x.s_y} = \frac{\sum_{i=1}^{n} x_iy_i - n\bar{x}\bar{y}}{(n-1)s_x.s_y}, \qquad (2)$$

که در آن s_x و s_y به ترتیب انحراف معیار نمونه‌ای مؤلفه اول و مؤلفه دوم این زوج نمونه‌ها هستند و s_{xy} کوواریانس نمونه‌ای بین این دو مؤلفه را نشان می‌دهد. محمودوند و حسنی [۲۵] نشان دادند که برای مشاهده‌های نامنفی ضریب تغییرات نمونه‌ای همواره کوچک‌تر از جذر تعداد مشاهده‌هاست. با توجه به نتیجه ایشان برای مشاهده‌های نامنفی داریم:

$$\frac{s_x}{\bar{x}} \le \sqrt{n} \ , \ \frac{s_y}{\bar{y}} \le \sqrt{n} \implies \frac{1}{s_x.s_y} \ge \frac{1}{n\bar{x}\bar{y}}.$$

با ضرب طرفین این رابطه در عبارت $|\sum_{i=1}^{n} x_iy_i - n\bar{x}\bar{y}|$، همواره برای مشاهده‌های مثبت داریم:

$$|r| = \left|\frac{\sum_{i=1}^{n} x_iy_i - n\bar{x}\bar{y}}{(n-1)s_xs_x)}\right| \ge \left|\frac{\sum_{i=1}^{n} x_iy_i}{n(n-1)\bar{x}.\bar{y}} - \frac{1}{n-1}\right| = \left|\frac{n}{n-1}\frac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n} x_iy_i + \sum_{i\neq j} x_iy_j} - \frac{1}{n-1}\right|. \qquad (3)$$

حال اگر فرض کنیم برای همه مقادیر نمونه‌ها شرط‌های $x_i < m_x$ و $y_i < m_y$ که در آنها m_x و m_y مقادیر معینی هستند، برقرار باشد آنگاه داریم

$$\frac{m_xm_y}{m_xm_y + (n-1)\max_{i\neq j} x_iy_j} \le \frac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n} x_iy_i + \sum_{i\neq j} x_iy_j} \le \frac{m_xm_y}{m_xm_y + (n-1)\min_{i\neq j} x_iy_j}, \qquad (4)$$

که با بزرگ شدن n طرفین این نابرابری به صفر همگرا می‌شوند. با فرض مثبت بودن مشاهده‌ها و به‌کارگیری نامساوی $n\bar{x}\bar{y} > s_x.s_y$ در صورت تعریف r و رابطه (۲) می‌توان دید که اگر $s_{xy} > 0$ آنگاه:

$$\frac{\sum_{i=1}^{n} x_iy_i}{(n-1)n\bar{x}\bar{y}} - \frac{1}{n-1} \le r \le \frac{\sum_{i=1}^{n} x_iy_i}{(n-1)s_xs_y} - \frac{1}{n-1} \qquad (5)$$

که به‌صورت زیر نیز قابل بازنویسی است:

$$\frac{n}{n-1}\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_iy_i+\sum_{i\neq j}x_iy_j}-\frac{1}{n-1}\leq r \leq \frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)-\sum_{i\neq j}x_iy_j}{(n-1)s_xs_y}-\frac{1}{n-1}. \quad (۶)$$

رابطه (۶) نشان می‌دهد که هرچقدر عبارت $\sum_{i\neq j}x_iy_j$ بزرگ‌تر باشد، مقدار برآورد ضریب همبستگی از کرانگین‌های ۱+ و ۱− فاصله می‌گیرد.

لم ۱.۲. *فرض کنید $w_1^{(x)},\ldots,w_n^{(x)}$ و $w_1^{(y)},\ldots,w_n^{(y)}$ متغیرهای برنولی باشند. به‌علاوه فرض کنید $w_{(1)}^{(x)}\leq\ldots\leq w_{(n)}^{(x)}$ و $w_{(1)}^{(y)}\leq\ldots\leq w_{(n)}^{(y)}$ آنگاه داریم:*

$$\sum_{i=1}^{n}w_i^{(x)}w_i^{(y)}\leq\sum_{i=1}^{n}w_{(i)}^{(x)}w_{(i)}^{(y)}.$$

اثبات: با توجه به مقادیر متغیرها واضح است که

$$\sum_{i=1}^{n}w_i^{(x)}w_i^{(y)}\leq\sum_{i=1}^{n}w_i^{(x)}$$
$$\sum_{i=1}^{n}w_i^{(x)}w_i^{(y)}\leq\sum_{i=1}^{n}w_i^{(y)}.$$

در نتیجه داریم:

$$\sum_{i=1}^{n}w_i^{(x)}w_i^{(y)}\leq\min\left\{\sum_{i=1}^{n}w_i^{(x)},\sum_{i=1}^{n}w_i^{(y)}\right\}.$$

از طرفی اگر فرض کنیم در مجموعه $w_1^{(x)},\ldots,w_n^{(x)}$ تعداد p مقدار یک و $n-p$ صفر وجود داشته باشد و در مجموعه $w_1^{(y)},\ldots,w_n^{(y)}$ تعداد q مقدار ۱ و $n-q$ مقدار صفر وجود داشته باشد آنگاه:

$$\sum_{i=1}^{n}w_i^{(x)}=p \quad \sum_{i=1}^{n}w_i^{(y)}=q.$$

علاوه بر این به‌سادگی داریم:

$$\sum_{i=1}^{n}w_{(i)}^{(x)}w_{(i)}^{(y)}=\min\{p,q\}.$$

در نتیجه داریم:

$$\sum_{i=1}^{n}w_i^{(x)}w_i^{(y)}\leq\min\left\{\sum_{i=1}^{n}w_i^{(x)},\sum_{i=1}^{n}w_i^{(y)}\right\}=\sum_{i=1}^{n}w_{(i)}^{(x)}w_{(i)}^{(y)}.$$

□

لم ۲.۲. *فرض کنید $w_1^{(x)},\ldots,w_n^{(x)}$ و $w_1^{(y)},\ldots,w_n^{(y)}$ متغیرهای برنولی باشند. به‌علاوه فرض کنید $w_{(1)}^{(x)}\leq\ldots\leq w_{(n)}^{(x)}$ و $w_{(1)}^{(y)}\leq\ldots\leq w_{(n)}^{(y)}$ آنگاه داریم:*

$$\sum_{i=1}^{n}w_i^{(x)}w_i^{(y)}\geq\sum_{i=1}^{n}w_{(i)}^{(x)}w_{(n-i+1)}^{(y)}.$$

اثبات: فرض کنید $v_i^{(y)}=1-w_i^{(y)}$. در این صورت با توجه به لم ۲.۱ واضح است که

$$\sum_{i=1}^{n}w_i^{(x)}v_i^{(y)}\leq\sum_{i=1}^{n}w_{(i)}^{(x)}v_{(i)}^{(y)}.$$

می‌دانیم که $v_{(i)}^{(y)}=1-w_{(n-i+1)}^{(y)}$ در نتیجه داریم:

$$\sum_{i=1}^{n}w_i^{(x)}\left(1-w_i^{(y)}\right)\leq\sum_{i=1}^{n}w_{(i)}^{(x)}\left(1-w_{(n-i+1)}^{(y)}\right),$$

که اثبات لم ۲.۲ را کامل می‌کند. □

قضیهٔ ۳.۲. *فرض کنید $x_{(1)}<x_{(2)}<\ldots<x_{(n)}$ آماره‌های مرتب متناظر با نمونه‌های $x_1,\ldots,x_n$ و $y_{(1)}<y_{(2)}<\ldots<y_{(n)}$ آماره‌های مرتب متناظر با نمونه‌های $y_1,\ldots,y_n$ باشند، آنگاه:*

$$\sum_{i=1}^{n}x_{(i)}y_{(n-i+1)}\leq\sum_{i=1}^{n}x_iy_i\leq\sum_{i=1}^{n}x_{(i)}y_{(i)}, \quad (۷)$$

$$n\sum_{i=1}^{n}x_{(i)}y_{(n-i+1)}\leq\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\leq n\sum_{i=1}^{n}x_{(i)}y_{(i)}. \quad (۸)$$

اثبات: توضیح مختصری درباره اثبات بخشی از این قضیه در هاردی و همکاران [۱۹] آمده است که در اینجا سعی شده است جزئیات کامل‌تری از اثبات ارائه شود. ابتدا کران بالای رابطه (۷) را بررسی می‌کنیم. توجه داریم که هر یک از متغیرهای مرتب را می‌توان به‌صورت ترکیب خطی با ضرایب صفر و یک از متغیرهای $x_{(j)}-x_{(j-1)}$ نوشت:

$$\begin{aligned}
x_{(1)} &= x_{(1)}\times 1+\left(x_{(2)}-x_{(1)}\right)\times 0+\cdots+\left(x_{(n)}-x_{(n-1)}\right)\times 0\\
x_{(2)} &= x_{(1)}\times 1+\left(x_{(2)}-x_{(1)}\right)\times 1+\cdots+\left(x_{(n)}-x_{(n-1)}\right)\times 0\\
&\vdots\\
x_{(n)} &= x_{(1)}\times 1+\left(x_{(2)}-x_{(1)}\right)\times 1+\cdots+\left(x_{(n)}-x_{(n-1)}\right)\times 1
\end{aligned}$$

اگر به نوشتار بالا توجه کنید ضرایب متغیر مرتب iام به‌صورت $(0,\ldots,0,1,\ldots,1)$ است که دارای i مقدار یک در ابتدای بردار و $n-i$ صفر در ادامه است؛ اما x_i یکی از این معادله‌هاست و با فرض آنکه $w_1^{(x_i)},\ldots,w_n^{(x_i)}$ ضرایب آن را نشان دهد، داریم:

$$x_i=\sum_{j=1}^{n}w_j^{(x_i)}\left(x_{(j)}-x_{(j-1)}\right),$$

که در آن $x_{(0)}=0$. توجه داریم که به‌عنوان نمونه اگر $n=3$ و $x_2>x_1$ و $x_3>x_1$ آنگاه برای برابری‌های:

$$\begin{aligned}
x_1 &= x_{(1)}\times 1+\left(x_{(2)}-x_{(1)}\right)\times 1+\left(x_{(3)}-x_{(2)}\right)\times 0,\\
x_2 &= x_{(1)}\times 1+\left(x_{(2)}-x_{(1)}\right)\times 0+\left(x_{(3)}-x_{(2)}\right)\times 0,\\
x_3 &= x_{(1)}\times 1+\left(x_{(2)}-x_{(1)}\right)\times 1+\left(x_{(3)}-x_{(2)}\right)\times 1,
\end{aligned}$$

می‌توانیم ضرایب صفر و یک را به‌صورت ماتریسی بنویسیم:

$$\begin{pmatrix} ۱ & ۱ & ۰ \\ ۱ & ۰ & ۰ \\ ۱ & ۱ & ۱ \end{pmatrix}.$$

حالا اگر هر ستون را به‌صورت صعودی مرتب کنیم به ماتریس زیر می‌رسیم:

$$\begin{pmatrix} ۱ & ۰ & ۰ \\ ۱ & ۱ & ۰ \\ ۱ & ۱ & ۱ \end{pmatrix}.$$

معادلات مرتبط با این ماتریس ضرایب به‌صورت زیر خواهد بود:

$$\begin{aligned} x_{(۱)} &= x_{(۱)} \times ۱ + \left(x_{(۲)} - x_{(۱)}\right) \times ۰ + \left(x_{(۳)} - x_{(۲)}\right) \times ۰, \\ x_{(۲)} &= x_{(۱)} \times ۱ + \left(x_{(۲)} - x_{(۱)}\right) \times ۱ + \left(x_{(۳)} - x_{(۲)}\right) \times ۰, \\ x_{(۳)} &= x_{(۱)} \times ۱ + \left(x_{(۲)} - x_{(۱)}\right) \times ۱ + \left(x_{(۳)} - x_{(۲)}\right) \times ۱. \end{aligned}$$

بنابراین می‌توان نتیجه گرفت که

$$x_{(i)} = \sum_{j=۱}^{n} w_j^{(x_{(i)})} \left(x_{(j)} - x_{(j-۱)}\right).$$

با این قابلیت بازچینش ضرایب و معادلات داریم:

$$\begin{aligned} \sum_{i=۱}^{n} x_i y_i &= \sum_{i=۱}^{n} \sum_{j=۱}^{n} w_j^{(x_i)} \left(x_{(j)} - x_{(j-۱)}\right) \sum_{k=۱}^{n} w_k^{(y_i)} \left(y_{(k)} - y_{(k-۱)}\right), \\ \sum_{i=۱}^{n} x_i y_i &= \sum_{j=۱}^{n} \left(x_{(j)} - x_{(j-۱)}\right) \sum_{k=۱}^{n} \left(y_{(k)} - y_{(k-۱)}\right) \sum_{i=۱}^{n} w_j^{(x_i)} w_k^{(y_i)}. \end{aligned}$$

با استفاده از لم ۲.۱ در آخرین مجموع در رابطه اخیر نتیجه می‌شود که:

$$\begin{aligned} \sum_{i=۱}^{n} x_i y_i &\leq \sum_{j=۱}^{n} \left(x_{(j)} - x_{(j-۱)}\right) \sum_{k=۱}^{n} \left(y_{(k)} - y_{(k-۱)}\right) \sum_{i=۱}^{n} w_j^{(x_{(i)})} w_k^{(y_{(i)})}, \\ \sum_{i=۱}^{n} x_i y_i &\leq \sum_{i=۱}^{n} \sum_{j=۱}^{n} w_j^{(x_{(i)})} \left(x_{(j)} - x_{(j-۱)}\right) \sum_{k=۱}^{n} w_k^{(y_{(i)})} \left(y_{(k)} - y_{(k-۱)}\right), \end{aligned}$$

که با توجه به برابری

$$x_{(i)} = \sum_{j=۱}^{n} w_j^{(x_{(i)})} \left(x_{(j)} - x_{(j-۱)}\right),$$

اثبات کران بالای رابطه (۷) در قضیه ۲.۳ را کامل می‌کند. درستی کران پائین با توجه به لم ۲.۲ به‌سادگی قابل‌بررسی است. برای رابطه (۸) دقت داریم که همواره:

$$\sum_{i=۱}^{n} \sum_{j=۱}^{n} \left(x_{(i)} - x_{(j)}\right) \left(y_{(i)} - y_{(j)}\right) \geq ۰.$$

زیرا اگر $i < j$ باشد هر دو پرانتز منفی هستند و اگر $i > j$ هر دو پرانتز مثبت هستند. با ساده کردن این عبارت، درستی کران بالایی رابطه (۸) به اثبات می‌رسد. به‌طور مشابه کران پائینی از رابطه زیر حاصل می‌شود و اثبات قضیه کامل است:

$$\sum_{i=۱}^{n} \sum_{j=۱}^{n} \left(x_{(i)} - x_{(j)}\right) \left(y_{(n-i+۱)} - y_{(j)}\right) \leq ۰. \quad \square$$

با استفاده از رابطه (۷) در قضیه ۱ می‌توان برای ضریب همبستگی پیرسون کران‌های زیر را تعریف کرد:

$$\frac{\sum_{i=۱}^{n} x_{(i)} y_{(n-i+۱)} - n\overline{xy}}{(n-۱) s_x s_y} \leq r \leq \frac{\sum_{i=۱}^{n} x_{(i)} y_{(i)} - n\overline{xy}}{(n-۱) s_x s_y}. \quad (۹)$$

توجه داریم که کران پائینی در رابطه (۹) برابر ضریب همبستگی زوج‌های $(x_{(۱)}, y_{(n)}), \ldots (x_{(n)}, y_{(۱)})$ است و کران بالایی برابر ضریب همبستگی زوج‌های $(x_{(۱)}, y_{(۱)}), \ldots (x_{(n)}, y_{(n)})$ است. از لحاظ نظری مقادیر نمونه‌های $(x_۱, y_۱), \ldots, (x_n, y_n)$ می‌تواند منطبق بر یکی از موارد $(x_{(۱)}, y_{(n)}), \ldots (x_{(n)}, y_{(۱)})$ یا $(x_{(۱)}, y_{(۱)}), \ldots (x_{(n)}, y_{(n)})$ باشد. طبق قضیه ۱ واضح است که کران سمت چپ در رابطه (۹) منفی و کران بالایی مثبت است. با استفاده از رابطه (۹) و با فرض مثبت بودن مشاهده‌ها و به‌کارگیری نامساوی $n\bar{x}\bar{y} > s_x . s_y$ در صورت تعریف r و رابطه (۳) می‌توان دید که اگر $s_{xy} < ۰$ آنگاه:

$$\begin{aligned} \frac{\sum_{i=۱}^{n} x_{(i)} y_{(n-i+۱)} - n\overline{xy}}{(n-۱) s_x s_y} &\leq r \\ &\leq \frac{n}{n-۱} \frac{\sum_{i=۱}^{n} x_i y_i}{\sum_{i=۱}^{n} x_i y_i + \sum_{i \neq j} x_i y_j} - \frac{۱}{n-۱}. \end{aligned}$$

بنابراین فارغ از اینکه مقدار s_{xy} مثبت یا منفی باشد، از تلفیق روابط (۶) و (۹) برای داده‌های مثبت داریم:

$$\begin{aligned} \frac{\sum_{i=۱}^{n} x_{(i)} y_{(n-i+۱)} - n\overline{xy}}{(n-۱) s_x s_y} &\leq r \\ \leq \frac{\left(\sum_{i=۱}^{n} x_i\right) \left(\sum_{i=۱}^{n} y_i\right) - \sum_{i \neq j} x_i y_j}{(n-۱) s_x s_y} &- \frac{۱}{n-۱}. \quad (۱۰) \end{aligned}$$

علاوه بر این از تلفیق روابط (۷) و (۸) در قضیه ۱ می‌توان نتیجه گرفت که

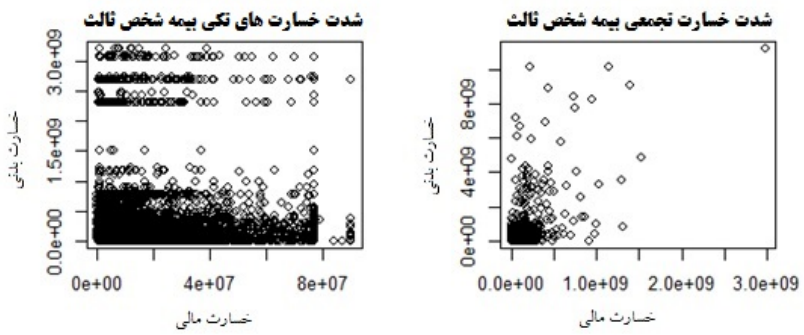
$$\frac{\sum_{i=1}^{n} x_{(i)}y_{(n-i+1)} - \sum_{i=1}^{n} x_{(i)}y_{(i)}}{(n-1)s_x s_y} \leq r \leq \frac{\sum_{i=1}^{n} x_{(i)}y_{(i)} - \sum_{i=1}^{n} x_{(i)}y_{(n-i+1)}}{(n-1)s_x s_y}. \quad (۱۱)$$

رابطه (۱۱) را می‌توان به‌صورت زیر نیز نوشت:

$$|r| \leq \frac{\sum_{i=1}^{n} \left(x_{(i)} - x_{(n-i+1)}\right)\left(y_{(i)} - y_{(n-i+1)}\right)}{2(n-1)s_x s_y}. \quad (۱۲)$$

۳ یک مثال واقعی

تعداد ۷۹۱۸ ادعای خسارت مربوط به بیمه شخص ثالث اتومبیل که منجر به پرداخت خسارت در یک شرکت بیمه‌شده‌اند موردبررسی قرار گرفت. سوابق خسارت‌ها در فایل داده‌های این شرکت در دو دسته «جانی» و «مالی» که ما آن‌ها را با X و Y نشان می‌دهیم، دسته‌بندی شده بود. هر پرونده در یکی از سه نوع (۱) فقط ادعای خسارت جانی (۲) فقط ادعای خسارت مالی و (۳) هر دو نوع ادعا، قرار داشت. بر این اساس بردارهای مربوط به یافته‌های متغیرهای (X,Y) مربوط به پرونده خسارتی i به یکی از سه صورت $(x_i,0)$، $(0,y_i)$ و (x_i,y_i) می‌باشد که در آن‌ها $x_i>0$، $y_i>0$. میزان ضریب همبستگی پیرسون بین دو نوع خسارت برابر ۰٫۰۰۱۴۳− است. توجه داریم که با به‌کارگیری فرمول (۹) می‌توان نتیجه گرفت که مقدار ضریب همبستگی پیرسون برای این داده‌ها می‌تواند در بازه (۹۶۰۱۳٪ و ۳۴۱۵۹٪−) قرار بگیرد. علاوه بر این، بر اساس فرمول (۱۰) نیز بازه (۳۷۶۲۶٪ و ۳۴۱۵۹٪−) به دست می‌آید. ناهمگونی بین دو کران ممکن است این آگاهی را به ما منتقل کند که بررسی بهتری بین خسارت‌های مالی و جانی بایستی صورت بگیرد. در واقع با توجه به آنکه منشأ هر دو ریسک جانی و مالی در بیمه شخص ثالث، در اغلب موارد، یکسان است یک راهکار استفاده از تجمیع خسارت‌ها برای هر بیمه‌گذار و محاسبه همبستگی بین متغیرهای تجمیعی است. برای تجمع خسارت‌ها از شماره بیمه‌نامه‌ها که یک کد یکتا است، استفاده شد. مجدداً یافته‌های خسارت‌های تجمیعی برای هر بیمه‌نامه در یکی از سه فرم $(\sum x_i, 0)$، $(0, \sum y_i)$ و $(\sum x_i, \sum y_i)$ قرار گرفت. در شکل ۲ نمودارهای پراکندگی برای هر دو حالت، پرونده‌های خسارت و مجموع خسارت‌های بیمه‌گذاران نمایش داده‌شده است. همان‌طور که ملاحظه می‌کنیم در نمودار تجمیعی ارتباط شفاف‌تری دیده می‌شود. مقدار ضریب همبستگی برای متغیرهای تجمیعی برابر ۵۶۱۲۶٪ است که بسیار بزرگ‌تر از حالتی است که مشاهده‌ها مربوط به پرونده‌های خسارتی هستند؛ مانند بالا بر پایه فرمول (۹) می‌توان دید که مقدار ضریب همبستگی برای این داده‌ها می‌تواند در بازه (۹۵۵۸۸٪ و ۲۹۷۱۲٪−) قرار بگیرد. همچنین، بر اساس فرمول (۱۰) نیز بازه (۸۸۸۷۹٪ و ۲۹۷۱۲٪−) به دست می‌آید که نشان می‌دهد دو فرمول نتایج تقریباً مشابهی را ارائه می‌دهند. توجه داریم که تجمیع خسارت‌ها برای هر بیمه‌نامه به سقف‌های پرداخت نزدیک‌تر می‌شود و با نتایج نظری این مقاله همخوانی بیشتری پیدا می‌کند.



شکل ۱. نمودار پراکندگی مربوط به خسارت‌های جانی و مالی بیمه شخص ثالث در مطالعه تحت بررسی

۴ بحث و نتیجه‌گیری

ضریب همبستگی پیرسون به‌عنوان یک گزینه متداول در بررسی ارتباط بین متغیرها همواره مورداستفاده قرار می‌گیرد. یکی از ویژگی‌های این ضریب، آن است که محدود به بازه $[-1, 1]$ است. این ویژگی برای برآورد گشتاوری این ضریب هم برقرار است. پیش‌تر درباره تحدید این بازه در توزیع‌های احتمالی مختلف بحث شده و بر اساس مفصل یک بازه باریک‌تر برای ضریب همبستگی پیرسون ارائه‌شده است. در این مقاله، موضوع برای برآورد گشتاوری ضریب همبستگی پیرسون موردبررسی قرار گرفت و نشان داده شد که تحت کران‌داری مقادیر یافته‌ها می‌توان کران‌های محدودتری برای برآورد گشتاوری ضریب همبستگی پیرسون ارائه کرد. این شرایط در مورد خسارت‌های بیمه‌ای برقرار است و لذا به‌عنوان یک کاربرد، مثالی از داده‌های بیمه‌ای مطرح و بررسی شد. نتایج نشان داد که توجه به این موضوع می‌تواند در تفسیر و نتیجه‌گیری از داده‌ها تغییرات قابل‌توجهی ایجاد کند.

مراجع

[۱] صحت، سعید.؛ مظلومی، نادر. فخیمی محمد پور، حمید. (۱۳۹۴). رابطه بین نوآوری سازمانی و مزیت رقابتی در شرکت‌های بیمه. *پژوهشنامه بیمه*، شماره ۱۱۸، صص ۳۴-۱.

[۲] مظلومی، نادر. هاشمی سید. علی‌رضا. (۱۳۹۳). الگوی رابطه قابلیت اجرا با اثربخشی راهبرد (موردمطالعه: صنعت بیمه ایران). *پژوهشنامه بیمه*، دوره ۳، شماره ۴، صص ۵۳۵-۵۲۰.

[3] Altman, D. G., and Bland, J. M. (1991). Improving doctors' understanding of statistics. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **154(2)**, 223-248.

[4] Avanzi, B., Taylor, G., and Wong, B. (2016). Correlations between insurance lines of business: An illusion or a real phenomenon? Some methodological considerations. *ASTIN Bulletin: The Journal of the IAA,* **46(2)**, 225-263.

[5] Barbiero, A. (2021). Inducing a desired value of correlation between two point-scale variables: a two-step procedure using copulas. *AStA Advances in Statistical Analysis,* **105(2)**, 307-334.

[6] Bland, J. M., and Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet,* **327(8476)**, 307-310.

[7] Britt, S., and Napoli, A. (2005). Linear correlation as a measure of dependency. *In XVth General Insurance Seminar*, Institute of Actuaries of Australia.

[8] Boyer, B. H., Gibson, M. S., and Loretan, M. (1999). Pitfalls in tests for changes in correlations. *International Finance Discussion Papers*.

[9] Denuit, M., Dhaene, J., Goovaerts, M. J., and Kaas, R. (2005). *Actuarial Theory for Dependent Risks: Measures, Orders, and Models*. John Wiley and Sons.

[10] de Veaux, D. (1976). Tight upper and lower bounds for correlation of bivariate distribution arising in air pollution modeling. *Technical report No. 5. Study on statistics and environmental factors in health (No. COO-2874-12)*. Stanford Univ., CA (USA). Dept. of Statistics

[11] Embrechts, P. (2009). Copulas: A personal view. *Journal of Risk and Insurance,* **76(3)**, 639-650.

[12] Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, **1**, 176-223. Cambridge: Cambridge University Press.

[13] Esmaeili, H., and Klüppelberg, C. (2010). Parameter estimation of a bivariate compound Poisson process. *Insurance: Mathematics and Economics,* **47(2)**, 224-233.

[14] Esmaeili, H., and Klüppelberg, C. (2011). Parametric estimation of a bivariate stable Lévy process. *Journal of Multivariate Analysis*, **102(5)**, 918-930.

[15] Esmaeili, H., and Klüppelberg, C. (2013). Two−Step Estimation of a Multi−Variate Lévy Process. *Journal of Time Series Analysis,* **34(6)**, 668-690.

[16] Frees, E. W., and Valdez, E. A. (1998). Understanding relationships using copulas. *North American actuarial Journal,* **2(1)**, 1-25.

[17] Gao, G., and Li, J. (2023). Dependence modeling of frequency-severity of insurance claims using waiting time. *Insurance: Mathematics and Economics,* **109**, 29-51.

[18] Gradstein, M. (1986). Maximal correlation between normal and dichotomous variables. *Journal of Educational Statistics,* **11(4)**, 259-261.

[19] Hardy, G. H., Littlewood, J. E., Pólya, G., and Pólya, G. (1934). *Inequalities.* Cambridge university press.

[20] Hashemi, S. J., Ahmed, S., and Khan, F. I. (2015). Correlation and dependency in multivariate process risk assessment. *IFAC-PapersOnLine,* **48(21)**, 1339-1344.

[21] Hoeffding, W. (1940). *MaBstabvariante Korrelationstheorie*. Schriften Math. Inst. Univ. Berlin 5, 181-233.

[22] Joe, H. (2015). *Dependence Modeling with Copulas.* Monographs on Statistics and Applied Probability 134. CRC Press, Boca Raton, FL.

[23] Kaas, R., Goovaerts, M., Dhaene, J., and Denuit, M. (2008). *Modern actuarial risk theory: using R*. Springer Science and Business Media.

[24] Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). *Loss models: from data to decisions.* John Wiley and Sons.

[25] Mahmoudvand, R., and Hassani, H. (2009), Two new confidence intervals for the coefficient of variation in a normal distribution. *Journal of Applied Statistics.* **36(4)**, 429-442.

[26] Nelsen, R. B. (2006). *An Introduction to Copulas.* 2nd ed. Springer, New York.

[27] Romano, J. P., and Siegel, A. F. (1986). *Counterexamples in Probability and Statistics*, Wadsworth and Brooks.

[28] Shih, W. J., and Huang, W. M. (1992). Evaluating correlation with proper bounds. *Biometrics*, 1207-1213.

[29] Sklar, A., (1959). *Fonctions de répartition à n dimensions et leurs marges*. Publications de l'Institut de Statistique de l'UniversitÚ de Paris, pp. 229-231.

[30] Yashraj Gupta, R., Sai Mudigonda, S., Baruah, P. K., and Krishna Kandala, P. (2021). Implementation of Correlation and Regression Models for Health Insurance Fraud in Covid-19 Environment using Actuarial and Data Science Techniques. arXiv e-prints, arXiv-2102.

Exploring limits for the Pearson Correlation Coefficient and its Application for Study of Insurance Losses

Rahim Mahmoudvand[1]

Abstract:

In actuarial studies, insurance losses are treated as random variables, and researchers seek appropriate probabilistic models to represent them. Since losses are evaluated in terms of a unity amount, distributions with positive support are typically used to model them. While this poses no issue for univariate cases, it becomes more complicated in multivariate scenarios. While copulas can be helpful in such situations, studying correlation is a crucial initial step. The Pearson correlation coefficient, widely used in statistical analysis, measures the strength and direction of the linear relationship between two variables. Furthermore, we analyze a real-world dataset from an Iranian insurance company, including losses due to physical damage and bodily injury, as covered by third-party liability insurance. Upper and lower limits for both the Pearson correlation coefficient and its estimator were derived from the analysis. Furthermore, two methods were used to determine the correlation between physical damage and bodily injury, and then the results were compared. Instead, our analysis reveals that narrower bounds can be established for the Pearson correlation coefficient in such cases. The results of this study provide important insights into modeling insurance losses in multivariate cases and have practical implications for risk management and pricing decisions in the insurance industry.

Keywords: Order Statistics, Moment, Confinement.

[1] **Department of Statistics, Faculty of Science, Bu-Ali Sina University**