آمار و احتمال مهندسی

اساتید: دکتر توسلی‌پور، دکتر وهابی
دانشکده مهندسی برق و کامپیوتر، دانشکدگان فنی، دانشگاه تهران

تمرین چهارم ـ توزیع‌های شرطی، توزیع بتا، کوواریانس و همبستگی

طراح: مصطفی کرمانی‌نیا
سوپروایزر: سالار صفردوست

تاریخ تحویل: ۱۸ آذر ۱۴۰۳

بیشتر بدانیم: پارادوکس سیمپسون

تصور کنید یک بیمارستان در حال مقایسه دو درمان A و B برای یک بیماری خاص است. آن‌ها داده‌هایی درباره‌ی نرخ موفقیت این دو درمان در دو گروه سنی مختلف بیماران جوان‌تر (زیر ۵۰ سال) و بیماران مسن‌تر (۵۰ سال و بالاتر) جمع‌آوری کرده‌اند:

گروه سنی	درصد موفقیت درمان A	درصد موفقیت درمان B
بیماران جوان‌تر	۳۰% (۵۰ از ۱۵۰)	۵۰% (۱۸۰ از ۳۶۰)
بیماران پیرتر	۸۰% (۲۰۰ از ۲۵۰)	۹۰% (۳۶ از ۴۰)

در جدول بالا مشاهده می‌شود که درمان B هم برای بیماران جوان‌تر و هم برای بیماران پیرتر درصد موفقیت بالاتری از درمان A دارد. حال سعی می‌کنیم از این جدول، جدول دیگری برای بررسی عملکرد درمان‌ها روی کل بیماران بنویسیم:

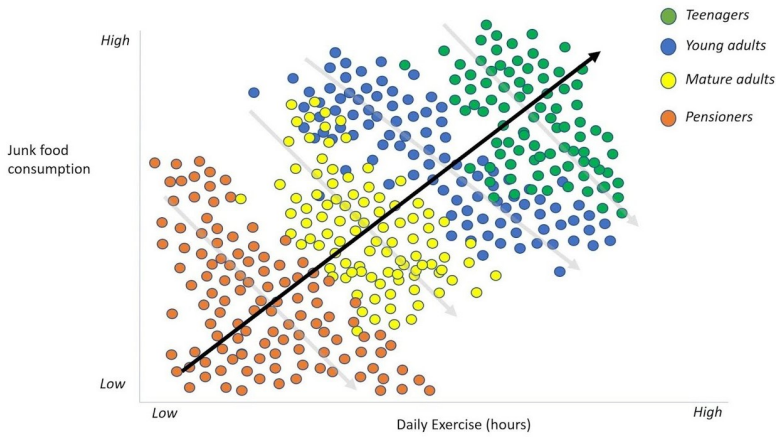
درمان	تعداد موفقیت‌ها	تعداد بیماران	درصد موفقیت کلی
درمان A	۲۵۰	۴۰۰	۶۲,۵%
درمان B	۲۱۶	۴۰۰	۵۴%

با بررسی درصد کلی موفقیت درمان‌ها دیده می‌شود که برعکس روند عملکرد درمان‌ها روی زیرگروه‌ها، درمان A در حالت کلی موفق‌تر از درمان B بوده است!

این وضعیت به ظاهر متناقض، نمونه‌ای از **پارادوکس سیمپسون** است که اولین بار توسط آمارشناس بریتانیایی ادوارد سیمپسون در سال ۱۹۵۱ توصیف شد. به طور کلی این پدیده زمانی رخ می‌دهد که یک روند مشاهده‌شده در چندین گروه مختلف، زمانی که داده‌ها با هم ترکیب می‌شوند، معکوس می‌شود. این پدیده نشان می‌دهد که برای انجام یک تحلیل صحیح و دقیق، باید به جزئیات داده‌ها توجه کرده و تأثیر متغیرهای مختلف را در نظر گرفت. نادیده گرفتن این متغیرهای پنهان و تأثیرگذار می‌تواند به نتایجی منجر شود که نه تنها نادرست هستند، بلکه می‌توانند به تصمیم‌گیری‌های اشتباه و پیامدهای جدی منجر شوند. این موضوع به ویژه در حوزه‌هایی مثل پزشکی، علوم اجتماعی و سیاست‌گذاری عمومی بسیار حیاتی است، چرا که اشتباهات آماری می‌توانند عواقب گسترده‌ای داشته باشند.

مثالی دیگر که می‌تواند باعث پیدا کردن شهود بیشتری نسبت به این پدیده شود در زیر توضیح داده شده است:
فرض کنید شکل نمایش داده شده حاصل نمونه‌برداری از جامعه‌ای خاص بوده باشد. دیده می‌شود که در بازه‌های سنی مختلف، میان مدت زمان ورزش انجام شده و میزان مصرف غذاهای ناسالم ارتباط(همبستگی) معکوسی وجود دارد. اما در صورتی که تمامی نمونه‌ها را با هم در نظر بگیریم، دیده می‌شود که برخلاف انتظار رابطه‌ی بین مدت زمان ورزش و مصرف غذای ناسالم به صورت مستقیم است. به همین دلیل اگر شخصی به خوبی از جزئیات مسئله آگاه نداشته باشد، محتمل است که نتیجه‌گیری اشتباهی از روی داده‌ها انجام دهد.

نکته: برای مثال اول نیز می‌توان مشابه این نمودار را با در نظر گرفتن پارامترهای نوع درمان و نتیجه‌ی درمان به عنوان متغیرهای محورهای نمودار رسم کرد، اما چون هر کدام از این متغیرها تنها دو حالت دارند، نمی‌توان از روی نمودار این پدیده را مشاهده کرد.



۱. سکه‌بازی ۲۰ نمره

یک سکه را سه بار پرتاب می‌کنیم. X را مساوی تعداد شیرها در دو پرتاب اول و Y را مساوی تعداد شیرها در دو پرتاب آخر می‌گذاریم.

الف) جدول احتمال را برای X و Y بکشید. (۷ نمره)

ب) به کمک جدول بخش **الف**، $P(X < ۲|Y = ۱)$ و $P(Y = ۲|X = ۱)$ را بنویسید. (۵ نمره)

ج) $\mathrm{Cov}(X, Y)$ را محاسبه کنید. (۸ نمره)

۲. مصطفی باگ‌زن! ۱۵ نمره

مصطفی در یک شرکت مشغول به کار شده و هر روز هزار خط کد می‌زند. اما چون در حین کار، با ربات همستر سکه جمع می‌کند، حواسش پرت می‌شود و همیشه کدهای باگ‌دار می‌زند. تعداد باگ‌های مصطفی در یک روز، متغیر تصادفی N با توزیع پواسون با پارامتر λ است.

خوشبختانه امیر هر روز کدهای مصطفی را خوانده و اصلاح می‌کند. فرض کنید امیر هر باگی را با احتمال p تشخیص داده و اصلاح کند. X را تعداد باگ‌های اصلاح شده توسط امیر در نظر بگیرید و Y را تعداد باگ‌هایی که امیر نتوانسته آنها را پیدا کند. (پس $X + Y = N$ است.)

الف) با کمک توزیع جرم احتمال مشترک، نشان دهید که X و Y از هم مستقل اند و $Y \sim \mathrm{Poi}(\lambda(۱ - p))$ و $X \sim \mathrm{Poi}(\lambda p)$ (۱۰ نمره)

ب) ضریب همبستگی بین X و N را بیابید. (۵ نمره امتیازی)

۳. از توام به شرطی ۱۵ نمره

تابع چگالی توام متغیرهای تصادفی X و Y بصورت زیر است:

$$f_{XY}(x,y)=\begin{cases}۱۰xy^{۲}, & ۰\le x\le y\le ۱\\ ۰, & \text{در غیر این صورت}\end{cases}$$

الف) توابع چگالی شرطی $f_{X|Y}(x|y)$ و $f_{Y|X}(y|x)$ را بدست آورید. (۸ نمره)

ب) $\mathrm{Cov}(X,Y)$ را حساب کنید. (۷ نمره)

۴. بازگشت مصطفی! ۱۵ نمره

مصطفی پس از اخراج شدن از کار قبلی‌اش، تصمیم گرفته است تا به‌کمک دوستانش استارت‌آپی راه‌اندازی کند. برای این استارت‌آپ لازم است ۴ دپارتمان مدیریت، مارکتینگ، امور اجرایی و منابع انسانی هر کدام با ۵ عضو تشکیل شوند. پس مصطفی دوستانش را که شامل ۵ نفر از دانشکده‌ی کامپیوتر، ۵ نفر از دانشکده‌ی برق، ۵ نفر از دانشکده‌ی صنایع و ۵ نفر از دانشکده‌ی مدیریت بودند، دور هم جمع کرده و بدون بررسی توانایی و حوزه‌ی کاری هر فرد، صرفا آن‌ها را بصورت تصادفی به داخل این ۴ دپارتمان ۵ نفره می‌فرستد.

اگر N_i تعداد دانشجویان صنایع در دپارتمان iام باشد:

الف) $\mathrm{Cov}(N_۱,N_۲)$ را حساب کنید. (۱۰ نمره)

راهنمایی:

برای حل سوال از متغیرهای شاخص S_i و T_i استفاده کنید که:

متغیر شاخص S_i برابر با ۱ است اگر دانشجوی صنایع iام در دپارتمان اول قرار گیرد:

$$N_۱=\sum_{i=۱}^{۵}S_i$$

و متغیر شاخص T_i برابر با ۱ است اگر دانشجوی صنایع iام در دپارتمان دوم قرار گیرد:

$$N_۲=\sum_{i=۱}^{۵}T_i$$

ب) ضریب همبستگی بین $N_۱+N_۲$ و $N_۳+N_۴$ را بیابید. (با محاسبات ساده می‌توان پاسخ را به دست آورد.) (۵ نمره)

۵. مهندس مشکوک! ۱۵ نمره

یک سکه‌ی ناشناخته داریم که توسط یک مهندس طراحی شده است. مهندس ادعا می‌کند که به طور میانگین، احتمال شیر آمدن سکه ۴ درصد و با انحراف معیار ۰.۲ می‌باشد.

الف) می‌خواهیم احتمال شیر آمدن سکه (P) را با توزیع بتا مدل‌سازی کنیم ($P \sim \text{Beta}(\alpha, \beta)$) . پارامترهای α و β این توزیع را پیدا کنید. (یافتن دو معادله‌ی دو مجهولی هم کافیست، در بخش‌های بعدی لازم به جایگذاری الفا و بتا نیست.) (۴ نمره)

مهندس از ادعای خود خیلی هم مطمئن نیست، به همین خاطر سکه را ۱ بار پرتاب می‌کند تا دیدگاه خودش از توزیع پیشین P را شفاف‌تر کند. فرض کنید X را نتیجه‌ی پرتاب سکه بنامیم و در این ۱ پرتاب، پیشامد مشاهده شده $X = x_1$ باشد.

ب) به کمک قانون بیز، تابع چگالی توزیع پسین P ($f_P(p|x_1)$) را بیابید. (لازم به محاسبه‌ی ثوابتی که تابعی از p نیستند، نیست) (۶ نمره)

راهنمایی:

در آزمایش پرتاب سکه، چون در حالت عادی احتمال شیر آمدن سکه (p) را می‌دانیم، اگر نتیجه‌ی پرتاب سکه را X بنامیم از توزیع برنولی پیروی کرده و داریم:

$$X \sim \text{Ber}(p)$$

$$Pr_X(x) = p^x(1-p)^{1-x}$$

اما در تعبیز بیزی، خود P را که احتمال شیرآمدن است با توزیع بتا با پارامترهای α و β مدل‌سازی می‌کنیم، پس خواهیم داشت:

$$P \sim \text{Beta}(\alpha, \beta)$$

$$X|P \sim \text{Ber}(P)$$

$$Pr_X(x|p) = p^x(1-p)^{1-x}$$

حالا فرض کنید مهندس از همان اول بجای یک بار پرتاب، سکه را n بار پرتاب کند و در این n پرتاب، مجموعه نتایج مشاهده شده روی سکه این‌ها باشند:

$$D = \{x_1, x_2, \ldots, x_n\}$$

ج) در این حالت هم تابع چگالی توزیع پسین P ($f_P(p|D)$) را بیابید. (۵ نمره امتیازی)

۶. نمره‌ی رایگان! ۱۵ نمره

گروهی از دانشجوهای دانشکده برق و کامپیوتر در یک چالش شرکت کرده‌اند. در این چالش یک تاس به نماینده‌ی دانشجوهای برق و یک تاس به نماینده‌ی دانشجوهای کامپیوتر داده می شود و هر کدام از آنها تا وقتی عدد ۶ بیاید، به تاس انداختن ادامه می‌دهند و اگر مجموع دفعاتی که این دو نفر تاس می اندازند تا هر دو ۶ بیاورند دقیقا ۲۰ بار بشود، تمام آن گروه دانشجوها در درس آمار و احتمال ۲۰ می‌گیرند. (مثلا اگر نماینده‌ی برق بعد از ۱۲ بار تاس انداختن و نماینده کامپیوتر بعد از ۸ بار تاس انداختن، ۶ بیاورند، برنده می‌شوند و همه ۲۰ می‌گیرند.)

الف) اگر بدانیم که این دانشجوها نهایتا در چالش برنده شده اند، توزیع احتمال تعداد دفعات تاس ریختن دانشجوی برق را بیابید. (۱۰ نمره)

ب) اگر بجای دو تاس معمولی، دو تاس ناعادلانه به آنها داده می شد که احتمال ۶ آمدن در آنها $\frac{1}{10}$ است و باز هم می‌دانستیم که در مسابقه برنده شده‌اند، آیا تغییری در توزیع احتمال تعداد دفعات تاس ریختن دانشجوی برق ایجاد می‌شد؟ توضیح دهید. (۵ نمره)

۷. تلاش آخر! ۱۵ نمره

مصطفی پس از ورشکست شدن استارت‌آپ‌اش، به‌همراه دوستانش یک کارخانه‌ی ساخت لوازم الکترونیکی راه انداخت. در این کارخانه دو خط تولید لامپ وجود دارد که خط تولید A را مهندسان برق، کنترل می‌کنند و ۶۰ درصد لامپ‌ها در همین خط تولید ساخته می‌شوند و خط تولید B را مهندسان صنایع کنترل می‌کنند و ۴۰ درصد لامپ‌ها در این خط تولید، ساخته می‌شوند.

از آنجایی که مهندسان برق دانش بیشتری در زمینه‌ی ساخت قطعات الکترونیکی دارند، طول عمر لامپ‌های خط تولید A از لامپ‌های خط تولید B بیشتر است. فرض کنید $T_A \sim \text{Exp}(۱)$ توزیع طول عمر لامپ‌های خط تولید A است و $T_B \sim \text{Exp}(۲)$ توزیع طول عمر لامپ‌های خط تولید B است. توجه کنید که در نهایت، تولیدات هر دو خط تولید با هم مخلوط شده و روانه‌ی بازار می‌شوند.

الف) اگر T را توزیع طول عمر نهایی محصولاتی که وارد بازار می‌شوند بگیریم، PDF آن را بیابید. (۵ نمره)

ب) یکی از مشتری‌های این کارخانه، لامپی خریده که بعد از مدتی خراب شده است. احتمال اینکه این لامپ توسط خط تولید شده‌های صنایع تولید شده باشد چقدر است؟ (۵ نمره)

ج) اگر میانگین طول عمر محصولات این کارخانه از ۱ واحد کمتر باشد، توسط بازرس‌ها پلمپ می‌شود. آیا این کارخانه هم پلمپ می‌شود و مصطفی دوباره بیکار خواهد شد؟ :) (۵ نمره)