آمار و احتمال مهندسی

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

پاسخ تمرین چهارم ــ 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:

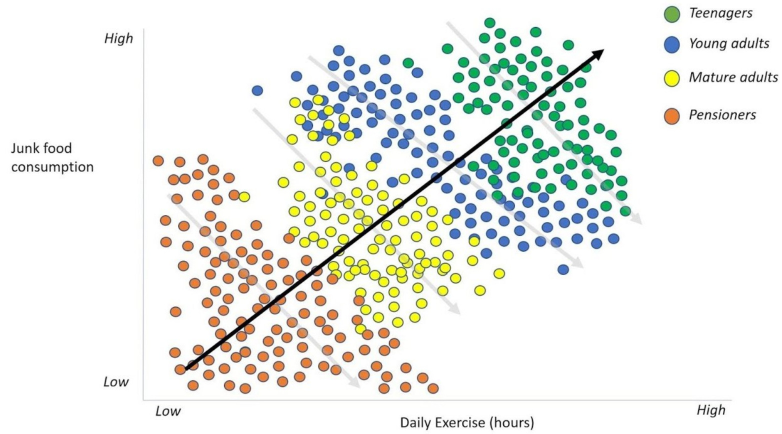
درمان	تعداد موفقیت‌ها	تعداد بیماران	درصد موفقیت کلی
درمان A	۲۵۰	۴۰۰	۶۲٫۵%
درمان B	۲۱۶	۴۰۰	۵۴%

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)

پاسخ :

الف) تمام ۸ حالت مختلف را برای ۳ پرتاب در نظر گرفته و طبق آن‌ها، جدول را تکمیل می‌کنیم:

$X\backslash Y$	0	1	2	$p(x_i)$
0	1/8	1/8	0	1/4
1	1/8	2/8	1/8	1/2
2	0	1/8	1/8	1/4
$p(y_j)$	1/4	1/2	1/4	1

ب)

$$\mathrm{P}(X < ۲ \mid Y = ۱) = \frac{\mathrm{P}(X < ۲ \wedge Y = ۱)}{\mathrm{P}(Y = ۱)} = \frac{\frac{۱}{۸} + \frac{۲}{۸}}{\frac{۱}{۲}} = \frac{\frac{۳}{۸}}{\frac{۴}{۸}} = \frac{۳}{۴}$$

$$\mathrm{P}(Y = ۲ \mid X = ۲) = \frac{\mathrm{P}(Y = ۲ \wedge X = ۲)}{\mathrm{P}(X = ۲)} = \frac{\frac{۱}{۸}}{\frac{۲}{۸}} = \frac{۱}{۲}$$

ج)

روش اول:

با توجه به تعریف کوواریانس داریم:

$$\mathrm{Cov}(X,Y) = E\left((X-E(X))(Y-E(Y))\right)$$

با توجه به جدول، داریم:

$$E(X) = \sum_{i=0}^{2} x_i \cdot p(x_i) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

$$E(Y) = \sum_{i=0}^{2} y_i \cdot p(y_i) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

رابطه‌ی کوواریانس را باز می‌کنیم:

$$\mathrm{Cov}(X,Y) = E\left((X-1)(Y-1)\right) = \sum_{i,j} (x_i-1)(y_j-1)\cdot p(x_i,y_j)$$

$$= (0-1)(0-1)\cdot\frac{1}{8} + (0-1)(1-1)\cdot\frac{1}{8} + (0-1)(2-1)\cdot 0 +$$
$$(1-1)(0-1)\cdot\frac{1}{8} + (1-1)(1-1)\cdot\frac{2}{8} + (1-1)(2-1)\cdot\frac{1}{8} +$$
$$(2-1)(0-1)\cdot 0 + (2-1)(1-1)\cdot\frac{1}{8} + (2-1)(2-1)\cdot\frac{1}{8}$$
$$= \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

روش دوم:

با توجه به تعریف کوواریانس داریم:

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

حالا با توجه به جدول، داریم:

$$E(XY) = \sum_{i,j} x_i y_j \cdot p(x_i,y_j) = 1\cdot 1\cdot\frac{2}{8} + 1\cdot 2\cdot\frac{1}{8} + 2\cdot 1\cdot\frac{1}{8} + 2\cdot 2\cdot\frac{1}{8} = \frac{10}{8} = \frac{5}{4}$$

$$\mathrm{Cov}(X,Y) = \frac{5}{4} - 1\cdot 1 = \frac{1}{4}$$

روش سوم:

اگر X_i را نتیجه‌ی پرتاب iــام بگذاریم، داریم:

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X_1+X_2, X_2+X_3)$$

با توجه به خواص کوواریانس و استقلال نتایج پرتاب‌های مختلف سکه، داریم:

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X_1,X_2) + \mathrm{Cov}(X_1,X_3) + \mathrm{Cov}(X_2,X_2) + \mathrm{Cov}(X_2,X_3) = \mathrm{Cov}(X_2,X_2) = \mathrm{Var}(X_2)$$

با توجه به اینکه تمامی X_i ها از توزیع $\mathrm{Bernoulli}(0.5)$ پیروی می‌کنند، داریم:

$$\mathrm{Cov}(X,Y) = \mathrm{Var}(X_2) = \frac{1}{4}$$

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 $X \sim \text{Poi}(\lambda p) \quad \text{و} \quad Y \sim \text{Poi}(\lambda(۱-p))$ (۱۰ نمره)

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

پاسخ :

الف) می‌دانیم هر باگی که در کدها بوجود می‌آید، به احتمال p اصلاح می‌شود و به احتمال $۱-p$ در بین کدها باقی می‌ماند. پس اگر تعداد کل باگ‌ها n باشد، تعداد باگ‌هایی که اصلاح می‌شوند از توزیع دوجمله‌ای با پارامترهای p و n پیروی می‌کنند:

$$X|N=n \sim \text{Bin}(n,p)$$

برای یافتن توزیع X و Y با استفاده از قانون احتمال کل و اینکه $X+Y=N$ داریم:

$$P(X=i,Y=j)=P(X=i,N=i+j)=P(X=i|N=i+j)\cdot P(N=i+j)$$

حالا با توجه به دوجمله‌ای بودن توزیع $X|N$ و پواسون بودن توزیع N داریم:

$$P(X=i,Y=j)=P(X=i|N=i+j)\cdot P(N=i+j)$$

$$=\frac{(i+j)!}{i!j!}p^i(۱-p)^j e^{-\lambda}\frac{\lambda^{i+j}}{(i+j)!}=\left(e^{-\lambda p}\frac{(\lambda p)^i}{i!}\right)\times\left(e^{-\lambda(۱-p)}\frac{(\lambda(۱-p))^j}{j!}\right)=P(X=i)\times P(Y=j)$$
$$(X\sim \text{Poi}(\lambda p) \quad \text{and} \quad Y\sim\text{Poi}(\lambda(۱-p)))$$

* توجه کنید فقط زمانی می‌توانیم از عبارت $X+Y=N$ وابسته بودن دو متغیر X و Y را نتیجه بگیریم که مقدار N مشخص باشد. اما در این سوال مقدار N مشخص نیست و خودش یک متغیر تصادفی است.

ب) با توجه به تعریف ضریب همبستگی داریم:

$$\rho(X,N)=\frac{Cov(X,N)}{\sqrt{Var(X)Var(N)}}$$

چون متغیرهای X و N توزیع پواسون دارند پس واریانس آن‌ها به‌ترتیب λp و λ می‌شود. برای محاسبه کوواریانس نیز داریم:

$$Cov(X,N) = Cov(X, X+Y) = E[X(X+Y)] - E[X]E[X+Y] = E[XX+XY] - E[X]E[X+Y]$$

با توجه به خاصیت خطی بودن امید ریاضی می‌توانیم بنویسیم:

$$E[XX+XY] - E[X]E[X+Y] = E[XX] + E[XY] - E[X](E[X]+E[Y]) =$$

$$(E[XX] - E[X]E[X]) + (E[XY] - E[Y]E[X]) = Cov(X,X) + Cov(X,Y)$$

با توجه به استقلال X و Y، کوواریانس آن‌ها صفر است و همچنین می‌دانیم $Cov(X,X) = Var(X)$. در نهایت با جایگذاری موارد بدست آمده ضریب همبستگی را حساب می‌کنیم:

$$\rho(X,N) = \frac{\lambda p}{\sqrt{\lambda p \times \lambda}} = \sqrt{p}$$

۳. از توام به شرطی　　　　۱۵ نمره

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

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

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)

پاسخ :

الف) تابع چگالی حاشیه‌ای $f_X(x)$ را محاسبه می‌کنیم:

$$f_X(x) = \int_x^{۱} ۱۰xy^{۲}\, dy$$

$$= ۱۰x\left[\frac{y^{۳}}{۳}\right]_x^{۱} = ۱۰x\left(\frac{۱}{۳} - \frac{x^{۳}}{۳}\right) = \frac{۱۰}{۳}x(۱-x^{۳})$$

سپس تابع چگالی شرطی $f_Y(y|x)$ را محاسبه می‌کنیم:

$$f_Y(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{۱۰xy^{۲}}{\frac{۱۰}{۳}x(۱-x^{۳})} = \frac{۳y^{۲}}{۱-x^{۳}} \quad : \quad ۰ \leq x \leq y \leq ۱$$

برای محاسبه‌ی $f_X(x|y)$ ابتدا $f_Y(y)$ را با انتگرال‌گیری از تابع چگالی مشترک نسبت به x محاسبه می‌کنیم:

$$f_Y(y) = \int_۰^{y} ۱۰xy^{۲}\, dx$$

$$= 10y^2 \int_0^y x\,dx = 10y^2 \left[\frac{x^2}{2}\right]_0^y = 10y^2 \times \frac{y^2}{2} = 5y^4$$

سپس تابع چگالی شرطی $f_X(x|y)$ را محاسبه می‌کنیم:

$$f_X(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{10xy^2}{5y^4} = 2\frac{x}{y^2} \quad : \quad 0 \le x \le y \le 1$$

به این ترتیب تابع چگالی شرطی $f_X(x|y)$ نیز به دست می‌آید.

ب) ابتدا $E[XY]$ را محاسبه می‌کنیم:

$$E[XY] = \int_0^1 \int_x^1 10xy^2.xy\,dy\,dx = \int_0^1 \int_x^1 10x^2y^3\,dy\,dx$$

در اینجا ابتدا انتگرال داخلی را نسبت به y حل می‌کنیم:

$$\int_x^1 10x^2y^3\,dy = 10x^2\left[\frac{y^4}{4}\right]_x^1 = 10x^2\left(\frac{1}{4} - \frac{x^4}{4}\right) = \frac{5}{2}x^2(1-x^4)$$

سپس انتگرال خارجی را نسبت به x محاسبه می‌کنیم:

$$E[XY] = \int_0^1 \frac{5}{2}x^2(1-x^4)\,dx = \frac{5}{2}\int_0^1 \left(x^2 - x^6\right)\,dx$$

$$= \frac{5}{2}\left[\frac{x^3}{3} - \frac{x^7}{7}\right]_0^1 = \frac{5}{2}\left(\frac{1}{3} - \frac{1}{7}\right) = \frac{5}{2} \times \frac{4}{21} = \frac{10}{21}$$

حالا $E[X]$ را محاسبه می‌کنیم:

$$E[X] = \int_0^1 f_X x\,dx = \frac{10}{3}\int_0^1 \left(x^2 - x^5\right)\,dx = \frac{10}{3}\left[\frac{x^3}{3} - \frac{x^6}{6}\right]_0^1 = \frac{10}{3}\left(\frac{1}{3} - \frac{1}{6}\right) = \frac{10}{3} \times \frac{1}{6} = \frac{5}{9}$$

سپس $E[Y]$ را محاسبه می‌کنیم:

$$E[Y] = \int_0^1 f_Y y\,dy = 5\int_0^1 y^5\,dy = 5\left[\frac{y^6}{6}\right]_0^1 = 5\left(\frac{1}{6} - 0\right) = \frac{5}{6}$$

در نهایت، کوواریانس $Cov(X,Y)$ را محاسبه می‌کنیم:

$$Cov(X,Y) = E[XY] - E[X]E[Y] = \frac{10}{21} - \frac{5}{9} \times \frac{5}{6} = \frac{10}{21} - \frac{25}{54}$$

$$= \frac{40}{84} - \frac{35}{84} = \frac{5}{84} = \frac{5}{378} \simeq 0/0132$$

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)

پاسخ :

الف)

$$E[N_۱N_۲] = E\left[\left(\sum_{i=۱}^{۵} S_i\right)\left(\sum_{j=۱}^{۵} T_j\right)\right] = \sum_{i=۱}^{۵}\sum_{j=۱}^{۵} E[S_iT_j] = \sum_{i=۱}^{۵}\sum_{j=۱}^{۵} \Pr(S_i = ۱, T_j = ۱)$$

از آنجا که یک دانشجو نمی‌تواند در دو گروه قرار بگیرد $\Pr(S_i = ۱, T_j = ۱)$ در حالت $i = j$ برابر صفر است. در حالت $i \neq j$ داریم:

$$\Pr(S_i = ۱, T_j = ۱) = \Pr(S_i = ۱)\Pr(T_j = ۱ | S_i = ۱)$$

احتمال قرارگیری دانشجوی صنایع iــام در هر یک از چهار گروه یکسان است بنابراین:

$$P(S_i = ۱) = \frac{۱}{۴}$$

اگر دانشجوی iـام در گروه اول قرار داشته باشد، برای دانشجوی jـام ۱۹ جای خالی باقی است، ۴ جا در گروه اول و ۵ جا در هر یک از سه گروه دیگر.
از این ۱۹ جا ۵ تا متعلق به گروه دوم است بنابراین:

$$P(T_j = 1 | S_i = 1) = \frac{5}{19}$$

از ۲۵ زوج (i, j)، ۲۰ زوج غیر یکسان هستند، پس داریم:

$$E[N_1 N_2] = 20 \times \frac{1}{4} \times \frac{5}{19} = \frac{25}{19}$$

$$E[N_1] = \sum_{i=1}^{5} E[S_i] = \frac{5}{4} = E[N_2]$$

پس با توجه به تعریف کواریانس داریم:

$$\mathrm{Cov}(N_1, N_2) = \frac{25}{19} - \frac{25}{16}$$

ب) متغیر تصادفی $A = N_1 + N_2$ را تعریف می‌کنیم. واضح است که $N_3 + N_4 = 5 - A$ بنابراین:

$$\rho(A,\ 5 - A) = \frac{\mathrm{Cov}(A,\ 5 - A)}{\sqrt{Var(A).Var(5 - A)}} = \frac{-\mathrm{Cov}(A,\ A)}{\sqrt{Var(A).Var(A)}} = -\frac{Var(A)}{Var(A)} = -1$$

یا مستقیما با استفاده از خواص ضریب همبستگی داریم:

$$\rho(A, 5 - A) = \rho(A, -A) = -1$$

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 $X = x_۱$ باشد.

ب) به کمک قانون بیز، تابع چگالی توزیع پسین P ($f_P(p|x_۱)$) 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:

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

$$Pr_X(x) = p^x(۱-p)^{۱-x}$$

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

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

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

$$Pr_X(x|p) = p^x(۱-p)^{۱-x}$$

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:

$$D = \{x_۱, x_۲, \dots, x_n\}$$

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

پاسخ :

الف) با توجه به روابطي که براي توزيع بتا داريم و اطلاعاتي که مهندس به ما داده است، داریم:

$$E(X) = \frac{\alpha}{\alpha+\beta} = \frac{4}{100}$$

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} = (0/2)^2$$

ب) با توجه به قانون بیز داریم: (دقت کنید که P از توزیع بتا است و پیوسته است و $X|P$ از توزیع برنولی است و گسسته است)

$$f_P(p|x_1) = \frac{\Pr(X=x_1|P=p)\cdot f_P(p)}{\Pr(X=x_1)} = \frac{p^{x_1}(1-p)^{(1-x_1)}\cdot \frac{1}{\rho(\alpha,\beta)}p^{(\alpha-1)}(1-p)^{(\beta-1)}}{\Pr(X=x_1)}$$

چون نیازی به محاسبه عباراتی که تابعی از p نیستند نداریم، پس:

$$f_P(p|x_1) = C\cdot p^{x_1}(1-p)^{(1-x_1)}\cdot p^{(\alpha-1)}(1-p)^{(\beta-1)} = C\cdot p^{(x_1+\alpha-1)}(1-p)^{(1-x_1+\beta-1)}$$

* دقت کنید با توجه به فرمی که پاسخ نهایی دارد می‌توان فهمید که از توزیع بتا است و پارامترهای آن را هم می‌توان یافت:

$$f_P(p|x_1) = C\cdot p^{(x_1+\alpha)-1}(1-p)^{(1-x_1+\beta)-1}$$

$$P|X=x_1 \sim Beta(x_1+\alpha, 1-x_1+\beta)$$

$$\Rightarrow P|X \sim Beta(X+\alpha, 1-X+\beta)$$

ج) مجدداً با توجه به قانون بیز داریم:

$$f_P(p|D) = \frac{\Pr_{\vec{X}(D|p)}\cdot f_{P(p)}}{\Pr_{\vec{X}(D)}} = \frac{\Pr(X_1=x_1,\dots,X_n=x_n|P=p)\cdot f_P(p)}{\Pr(X_1=x_1,\dots,X_n=x_n)}$$

از طرفی می‌دانیم که n متغیر تصادفی $X_1,\dots,X_n$ مستقل شرطی به شرط P هستند، پس داریم:

$$f_P(p|D) = \frac{\prod_{i=1}^n \Pr(X_i=x_i|P=p)\cdot f_P(p)}{\Pr(X_1=x_1,\dots,X_n=x_n)} = \frac{\prod_{i=1}^n p^{x_i}(1-p)^{(1-x_i)}\cdot \frac{1}{\rho(\alpha,\beta)}p^{(\alpha-1)}(1-p)^{(\beta-1)}}{\Pr(X_1=x_1,\dots,X_n=x_n)}$$

$$= \frac{p^{\sum_{i=1}^n x_i}(1-p)^{(n-\sum_{i=1}^n x_i)}\cdot \frac{1}{\rho(\alpha,\beta)}p^{(\alpha-1)}(1-p)^{(\beta-1)}}{\Pr(X_1=x_1,\dots,X_n=x_n)}$$

$$= C\cdot p^{\sum_{i=1}^n x_i}(1-p)^{(n-\sum_{i=1}^n x_i)}\cdot p^{(\alpha-1)}(1-p)^{(\beta-1)}$$

$$= C\cdot p^{\left(\sum_{i=1}^n x_i+\alpha\right)-1}(1-p)^{\left(n-\sum_{i=1}^n x_i+\beta\right)-1}$$

* دقت کنید با توجه به فرمی که پاسخ نهایی دارد، می‌توان فهمید که توزیع بتا است و پارامترهای آن هم می‌توان یافت:

$$f_P(p|D) = C\cdot p^{\left(\sum_{i=1}^n x_i+\alpha\right)-1}(1-p)^{\left(n-\sum_{i=1}^n x_i+\beta\right)-1}$$

$$\Rightarrow P|D \sim Beta\left(\sum_{i=1}^n x_i+\alpha, n-\sum_{i=1}^n x_i+\beta\right)$$

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 $\frac{۱}{۱۰}$ 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)

پاسخ :

الف) عددی که در هر پرتاب روی تاس دانشجوی برق ظاهر می‌شود را با متغیر تصادفی Xو عددی که در هر پرتاب روی تاس دانشجوی کامپیوتر ظاهر می‌شود را با متغیر تصادفی Y نشان می‌دهیم . از طرفی می‌دانیم که نهایتا در چالش برنده شده‌اند پس $X+Y$ را داریم و مساوی ۲۰ است، اما در طول راه حل، برای یافتن پاسخ کلی، آنرا $X+Y=n$ می‌گذاریم و در انتها عدد ۲۰ را جایگذاری می‌کنیم.

با توجه به مستقل بودن X و Y نسبت به همدیگر و تمام حالات ممکن برای آنها با حاصل جمع معین داریم:

$$P(X+Y=n)=\sum_{i=۱}^{n-۱}P(X=i,Y=n-i)=\sum_{i=۱}^{n-۱}P(X=i)P(Y=n-i)$$

از طرفی بدیهتا هر دوی X و Y از توزیع هندسی با پارامتر $\frac{۱}{۶}$ پیروي مي‌کنند که فعلا بجاي آن p مي‌گذاريم تا مسئله در حالت کلي حل شود و در نهايت $\frac{۱}{۶}$ را جايگذاري مي‌کنيم.

$$P(X+Y=n)=\sum_{i=۱}^{n-۱}p(۱-p)^{i-۱}p(۱-p)^{n-i-۱}=(n-۱)p^{۲}(۱-p)^{n-۲}$$

حالا مي‌توانيم توزيع احتمال تعداد دفعات تاس ريختن دانشجوي برق را با فرض دانستن تعداد کل تاس‌هاي ريخته شده محاسبه کنيم:

$$P(X=k\mid X+Y=n)=\frac{P(X=k,X+Y=n)}{P(X+Y=n)}=\frac{P(X=k,Y=n-k)}{P(X+Y=n)}$$

$$P(X=k\mid X+Y=n)=\frac{P(X=k)P(Y=n-k)}{(n-۱)p^{۲}(۱-p)^{n-۲}}=\frac{p(۱-p)^{k-۱}p(۱-p)^{n-k-۱}}{(n-۱)p^{۲}(۱-p)^{n-۲}}$$

$$P(X=k\mid X+Y=n)=\frac{۱}{n-۱}$$

پس متغير تصادفي $X = k|X + Y = n$ هيچ تابعيتي از k نداشت كه يعني داراي توزيع يكنواخت است
با جايگذاري مقدار n داريم:

$$P(X = k \mid X + Y = ۲۰) = \frac{۱}{۱۹}$$

پس اين متغير تصادفي، توزيع يكنواخت گسسته روي مقادير $X = ۱, ۲, ..., ۱۹$ دارد.

ب) خير، همانطور كه در راه حل بخش قبلي ديديم، در ابتدا احتمال ۶ آمدن تاس را p گذاشتيم كه در صورت لزوم جايگذاري كنيم، اما در پاسخ نهايي اثري از p نبود. پس توزيع احتمال تعداد دفعات تاس ريختن دانشجوي برق با فرض اينكه مي‌دانيم بعد از مجموعا ۲۰ بار تاس انداختن هر دو دانشجو ۶ آورده‌اند، مستقل از احتمال ۶ آمدن است و فقط كافيست كه احتمال ۶ آمدن مخالف صفر باشد تا به همان توزيع يكنواخت گسسته روي مقادير ۱ تا ۱۹ برسيم.

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 $T_A \sim \mathrm{Exp}(۱)$ توزیع طول عمر لامپ‌های خط تولید A است و $T_B \sim \mathrm{Exp}(۲)$ 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)

پاسخ :

الف) فرض كنيد يك قطعه را بصورت تصادفي از خروجي كارخانه برداشته‌ايم، متغير برنولی $X \sim Bernoulli(۰/۴)$ را به‌صورت زير تعريف مي‌كنيم:

$$X = \begin{cases} ۰ & \text{اگر قطعه در خط تولید A تولید شده باشد} \\ ۱ & \text{اگر قطعه در خط تولید B تولید شده باشد} \end{cases}$$

پس احتمال انتخاب هر خط توليد به شرح زير است:

$$P(X = ۰) = ۰/۶, \quad P(X = ۱) = ۰/۴$$

حال، برای یافتن تابع چگالی احتمال (PDF) طول عمر T یک قطعه به صورت کلی، از قانون احتمال کل استفاده می‌کنیم:

$$f_T(t) = f_{T|X}(t|۰)P_X(۰) + f_{T|X}(t|۱)P_X(۱)$$

با جایگذاری مقادیر داده شده:

$$f_T(t) = \left(e^{-t}\right) \times ۰/۶ + \left(۲e^{-۲t}\right) \times ۰/۴$$

با ساده‌سازی:

$$f_T(t) = ۰/۸e^{-۲t} + ۰/۶e^{-t}$$

بنابراین، تابع چگالی احتمال کلی طول عمر T برابر است با:

$$f_T(t) = ۰/۸e^{-۲t} + ۰/۶e^{-t}, \quad t \geq ۰$$

ب) فرض کنید لامپی که توسط مشتری خریداری شده است، بعد از مدت t خراب شده است. می‌خواهیم احتمال اینکه این لامپ توسط خط تولید B (دانشجوهای صنایع) تولید شده باشد را محاسبه کنیم.

با استفاده از قانون بیز داریم:

$$P_{X|T}(۱ \mid t) = \frac{f_{T|X}(t \mid ۱) \cdot P_X(۱)}{f_T(t)}$$

(از آنجایی که T پیوسته است، باید از چگالی احتمال T استفاده کنیم) با جایگذاری مقادیر:

$$\begin{aligned} P_{X|T}(۱ \mid t) &= \frac{۲e^{-۲t} \cdot ۰/۴}{۰/۸e^{-۲t} + ۰/۶e^{-t}} = \frac{۰/۸e^{-۲t}}{۰/۸e^{-۲t} + ۰/۶e^{-t}} \\ &= \frac{۰/۸}{۰/۸ + ۰/۶e^{t}}, \quad t \geq ۰ \end{aligned}$$

ج) برای محاسبه میانگین طول عمر $E[T]$، از تابع چگالی احتمال $f_T(t)$ استفاده می‌کنیم:

$$E[T] = \int_۰^\infty t \cdot f_T(t)\, dt$$

جایگذاری $f_T(t)$:

$$E[T] = \int_۰^\infty t \cdot \left(۰/۸e^{-۲t} + ۰/۶e^{-t}\right) dt$$

$$E[T] = ۰/۸ \int_۰^\infty t \cdot e^{-۲t}\, dt + ۰/۶ \int_۰^\infty t \cdot e^{-t}\, dt$$

انتگرال‌های فوق با انتگرال‌گیری جز به جز براحتی محاسبه می‌شوند:

$$E[T] = ۰/۸ \cdot \frac{۱}{۴} + ۰/۶ \cdot ۱ = ۰/۲ + ۰/۶ = ۰/۸$$

بنابراین، میانگین طول عمر $E[T] = ۰/۸$ است که کمتر از ۱ واحد می‌باشد. در نتیجه، شرکت توسط بازرس‌ها پلمپ خواهد شد و مصطفی دوباره بیکار خواهد شد! :(