حل سوالات امتحان پایان‌ترم

آمار و احتمالات مهندسی

تیر ۱۴۰۰

مدرس: امیر نجفی

سوال ۱ ـ سوالات پاسخ کوتاه (۴ نمره):

۱-۱ـ برای هر یک از آزمایش‌های زیر توزیع مناسبی که جهت توصیف متغیرهای تصادفی آن لازم است را بنویسید.

- نتیجه‌ی پرتاب یک سکه.
- تعداد دفعاتی که توپ بسکتبال را پیش از گل شدن باید پرتاب کنیم.
- تعداد تصادفات ماشینی که در هر روز در تهران رخ می‌دهد.
- تعداد دفعاتی که در ۱۰۰ بار پرتاب یک تاس عدد ۳ ظاهر می‌شود.

جواب:

- توزیع برنولی، چرا که مختص متغیرهای تصادفی دو حالتی است.
- توزیع هندسی، چون این توزیع مربوط به تعداد دفعاتی است که می‌بایست یک آزمون برنولی را به صورت مستقل انجام داد تا اولین موفقیت حاصل شود.
- توزیع پواسون، زیرا می‌توان وقوع/عدم وقوع هر تصادف را به صورت یک متغیر تصادفی برنولی تصور نموده که تعداد کل تصادفات برابر با جمع آنان است. توزیع حاصله مشابه با یک توزیع دوجمله‌ای (binomial) رفتار کرده که تعداد دفعات انجام آن (n) بسیار زیاد و احتمال وقوع هر یک از آنان p بسیار کوچک است که در حد به یک توزیع پواسون نزدیک می‌شود.
- توزیع دوجمله‌ای، چرا که ظاهر شدن/نشدن عدد ۳ در هر بار پرتاب تاس یک واقعه دودویی است که با یک متغیر تصادفی برنولی مدلسازی می‌شود. تعداد دفعات ظاهر شدن عدد ۳ در ۱۰۰ بار پرتاب معادل با جمع ۱۰۰ عدد از متغیرهای تصادفی برنولی است که یک توزیع binomial یا دوجمله‌ای را نتیجه می‌دهد.

۱-۲- اگر X_1 و X_2 دو متغیر تصادفی مستقل از هم باشند، به طوری که به ازای $k = 0, 1, 2, \ldots, \infty$ و $i = 1, 2$ داریم $\mathrm{P}(X_i = k) = (1-p)p^k$. آنگاه احتمال $\mathrm{P}(X_1 < X_2)$ را محاسبه نمایید.

جواب:

به دلیل تقارن داریم $\mathbb{P}(X_1 > X_2) = \mathbb{P}(X_2 > X_1)$. از طرفی می‌دانیم که بالاخره یکی از سه رخداد $X_1 < X_2,\ X_2 < X_1,\ X_1 = X_2$ اتفاق افتاده است و لذا جمع احتمال این سه می‌بایست برابر با یک شود. پس داریم:

$$\mathbb{P}(X_1 < X_2) = \frac{1 - \mathbb{P}(X_1 = X_2)}{2}$$

پس کافی است که $\mathbb{P}(X_1 = X_2)$ را محاسبه کنیم که برابر است با:

$$\mathbb{P}(X_1 = X_2) = \sum_{k=0}^{\infty} \mathbb{P}(X_1 = k)\,\mathbb{P}(X_2 = k) = \sum_{k=0}^{\infty} (1-p)^2 p^{2k} = \frac{(1-p)^2}{1-p^2} = \frac{1-p}{1+p}$$

در رابطه بالا از قانون احتمال کل و همچنین مستقل بودن X_1, X_2 استفاده شده است. پس از قرار دادن در رابطه اول و ساده‌سازی خواهیم داشت:

$$\mathbb{P}(X_1 < X_2) = \frac{1 - \frac{1-p}{1+p}}{2} = \frac{p}{1+p}$$

۱-۳- در صورتی که $f(x,y)$ به صورت زیر تعریف شده باشد، آیا x و y از هم مستقل هستند؟ با دلیل جواب دهید.

$$f(x,y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \le 1 \\ 0 & else \end{cases}$$

جواب:

واضح است که مستقل نیستند.

استدلال اول: برای مثال، در صورتی که بدانیم $x = 0.9$، آنگاه احتمال واقعه $|y| > 0.1$ برابر با صفر خواهد شد که نشان می‌دهد دانستن مقدار یکی از متغیرهای تصادفی در توزیع دومی تاثیرگذار است.

استدلال دوم: از تعریف استقلال استفاده می‌کنیم. ابتدا توزیع‌های حاشیه‌ای را پیدا کرده:

$$f_X(x) = \int_{\mathbb{R}} f_{XY}\mathrm{d}y = \frac{2\sqrt{1-x^2}}{\pi} \quad , \quad f_Y(y) = \int_{\mathbb{R}} f_{XY}\mathrm{d}x = \frac{2\sqrt{1-y^2}}{\pi}$$

و سپس می‌بینیم که:

$$f_X Y\left(x, y\right) \neq f_X\left(x\right) f_Y\left(y\right)$$

پس مستقل نیستند.

سوال ۲ (۴ نمره):

یک فرستنده مخابراتی قصد ارسال n بیت اطلاعات به یک گیرنده را دارد. داده‌ها از طریق یک کانال نویزی فرستاده شده و هر بیتی که توسط فرستنده ارسال می‌گردد با احتمال $p = 1/2$ و مستقل از ارسال‌های قبلی دچار خطا می‌شود.
به منظور اطمینان از صحت ارسال‌ها، در سمت فرستنده از یک کدگذار (coder) بهره می‌بریم. بدین شکل که به جای ارسال هر کدام از n بیت فوق، تعداد L بیت که به طریقی خاص انتخاب شده‌اند، به نمایندگی از آن ارسال می‌گردند. لذا، درنهایت به جای n بیت اولیه، تعداد کل nL بیت ارسال خواهند شد. این کار قرار است با چسباندن اطلاعات زائد (Redundancy) تعداد ارسال‌ها را افزایش داده، اما در عوض احتمال بروز خطا در آنان را کاهش دهد. خاصیت هر بلوک L بیتی که نماینده یکی از n بیت اولیه می‌باشد این است که تنها در صورتی بیت اصلی در سمت گیرنده دچار خطا خواهد شد که تمامی L بیتی که آن را نمایندگی می‌کنند در کانال خراب شوند. و حتی اگر صرفاً یکی از آنان سالم به مقصد برسد، بیت اصلی به درستی decode می‌گردد. در صورتیکه علاقه داشتید بدانید این کار چگونه امکان‌پذیر است، درس «تئوری اطلاعات» را در ترم‌های آینده بگیرید.
در این سوال قصد داریم تا احتمال بروز حداقل یک خطا در ارسال n بیت اصلی را بدست بیاوریم. در آخر نشان خواهیم داد که در صورت $O\left(n \log_2 n\right)$ بار استفاده از کانال به جای n بار، احتمال انتقال کاملاً صحیح کل داده‌ها برای $n \gg 1$ به سمت ۱ میل خواهد نمود.

الف) فرض کنید که اصلاً از کدگذاری استفاده نمی‌شد و در ارسال هر یک از n بیت اصلی، صرفاً همان بیت و به همان شکل اصلی خود ارسال می‌گشت (به عبارتی، داشتیم $L = 1$). احتمال اینکه از میان n بیت اصلی، حداقل یکی دچار خطا شود چقدر است؟
(به اختصار استدلال کنید که با افزایش n این احتمال به سمت یک میل خواهد کرد.)

ب) در صورت کدگذاری با طول بلوک‌های کد L، احتمال اینکه حداقل یکی از n بیت اصلی با خطا در گیرنده decode شود را حساب کنید.

ج) نشان دهید در صورتیکه طول بلوک‌های کد را به صورت $L = (1+\varepsilon)\log_2 n$ انتخاب کنیم (به ازای هر $\varepsilon > 0$)، احتمال بروز خطا با افزایش n به سمت صفر میل خواهد کرد.

جواب:

الف) احتمال بروز حداقل یک خطا در ارسال برابر یک منهای احتمال سالم رسیدن همگی بیت‌هاست. از آنجا که ارسال‌ها مستقل از یکدیگر و هر کدام تنها با احتمال ۱/۲ سالم می‌رسند، پس داریم:

$$\mathbb{P}(\text{Error}) = 1 - \frac{1}{2^n}$$

که به وضوح با افزایش n به صورت نمایی به ۱ نزدیک می‌شود. یعنی برای n های بزرگ احتمالاً حداقل یک خطا خواهیم داشت.

ب) جواب مشابه با بخش قبل است، با این تفاوت که این بار احتمال وقوع خطا دیگر ۱/۲ نیست. بلکه خطا در ارسال هر بیت زمانی رخ می‌دهد که تمامی L بیت نماینده آن دچار خطا شوند. پس احتمال سالم رسیدن هر بیت، مستقل از سایرین، برابر با $1 - 2^{-L}$ است که مشابه با استدلال بخش الف) و با توجه به استقلال در ارسال‌ها بدست آمده است. لذا، رابطه نهایی خواسته شده در بخش ب) برابر است با:

$$\mathbb{P}(\text{Error}) = 1 - \left(1 - \frac{1}{2^L}\right)^n$$

ج) صرفا می‌بایست $L = (1+\varepsilon)\log_2 n$ را در رابطه بخش دوم جایگذاری کنیم:

$$\lim_{n\to\infty} \mathbb{P}(\text{Error}) = \lim_{n\to\infty} 1 - \left(1 - \frac{1}{2^L}\right)^n = \lim_{n\to\infty} 1 - e^{-n2^{-L}} = \lim_{n\to\infty} 1 - e^{-\frac{n}{n^{1+\varepsilon}}} = \lim_{n\to\infty} 1 - e^{-n^{-\varepsilon}} = 0$$

که به ازای هر $\varepsilon > 0$، احتمال حتی یک ارسال خطادار با افزایش n به سمت صفر میل خواهد کرد.

سوال ۳ (۴ نمره):

یک سکه تصادفی با احتمال شیر یا خط نامعلوم را n بار به صورت مستقل پرتاب کرده و مشاهده می‌کنیم که k بار شیر ظاهر می‌شود.

الف) تخمین MLE از احتمال شیر آمدن سکه را محاسبه کنید.

حال با سازنده سکه ملاقات داشته و اطلاعاتی در مورد سکه کسب می‌کنیم. سازنده سکه می‌گوید که سکه‌های ساخت او به یک سمت بایاس دارند. وی معتقد است که احتمال شیر آمدن سکه‌هایش با احتمال α بین صفر تا ۱/۲ (یعنی در بازه $[0,0.5]$)، و با احتمال $1-\alpha$ بین ۱/۲ تا ۱ (یعنی بازه $(0.5,1]$) است. همچنین، در بازه صفر تا ۱/۲ میان مقادیر برتری نسبت به یکدیگر وجود ندارد. در بازه ۱/۲ تا ۱ نیز به همین شکل، برتری بین مقادیر نیست.
(بدون کاستن از کلیت مسئله، فرض کنید که $\alpha > 0.5$)

ب) با استفاده از این اطلاعات پیشین، تخمین MAP از احتمال شیر آمدن سکه را دوباره محاسبه کنید.

جواب:

الف) توزیع تعداد شیرها به شرط دانستن احتمال آنان p یک توزیع دوجمله‌ای است. با توجه به این واقعیت، ابتدا تابع درست‌نمایی را تشکیل داده و سپس آن را بیشینه می‌کنیم:

$$\mathbb{P}\left(k\,|\,p\right)=\binom{n}{k}p^{k}\left(1-p\right)^{n-k},\quad \log\mathbb{P}\left(k\,|\,p\right)=k\log p+\left(n-k\right)\log\left(1-p\right)+\text{const}$$

برای بیشینه‌سازی، ابتدا لازم است ببینیم که به ازای $p\to 0$ و $p\to 1$ مقدار درست‌نمایی به سمت صفر میل میکند. لذا در این بین می‌بایست یک بیشینه وجود داشته باشد. مقدار بیشینه کننده p^* را با مشتق‌گیری از تابع log-likelihood بدست می‌آوریم:

$$\frac{\partial}{\partial p}LL\left(p;k,n\right)=\frac{k}{p}-\frac{n-k}{1-p}=0\quad\Rightarrow\quad p^{*}=\frac{k}{n}$$

ب) به منظور یافتن تخمین بیشینه توزیع پسین (MAP) ابتدا لازم است که یک توزیع پیشین بر حسب اطلاعات داده شده مدلسازی شود. بر مبنای اطلاعات داده شده، تابع توزیع پیشین می‌بایست از دو بخش یکنواخت تشکیل شده باشد. بخش اول بین صفر و نیم (با احتمال مجموع α و لذا چگالی احتمال 2α) و بخش دوم بین نیم و ۱ با چگالی احتمال $2\left(1-\alpha\right)$. یعنی:

$$f_{p}\left(p\right)=\begin{cases}2\alpha & p\le 1/2\\ 2\left(1-\alpha\right) & p>1/2\end{cases}$$

حال بیشینه توزیع پسین که مطابق با زیر است را محاسبه می‌کنیم:

$$f\left(p\,|\,k\right)\propto\mathbb{P}\left(k\,|\,p,n\right)f_{p}\left(p\right)=p^{k}\left(1-p\right)^{n-k}f_{p}\left(p\right)$$

به دلیل خوشرفتار نبودن توزیع پیشین، بیشینه‌سازی را نمی‌توان با مشتق‌گیری انجام داد. از طرفی، از حل قسمت الف) می‌دانیم که بیشینه بخش درست‌نمایی از توزیع پسین در $p=k/n$ رخ می‌دهد. برای بیشینه‌سازی لازم است حالت‌بندی کنیم:

- (حالت اول) فرض کنید که داشته باشیم $k/n \leq 1/2$. در این صورت، تخمین MAP همواره همان تخمین MLE یعنی k/n می‌شود. چرا که مقدار $p^* = k/n$ هم بخش درست‌نمایی $\mathbb{P}\left(k \mid p, n\right)$ را بیشینه می‌سازد، و هم بخش توزیع پیشین $f_p\left(p\right)$ (دقت کنید که فرض کرده بودیم $\alpha > 1 - \alpha$).

- (حالت دوم) فرض کنید که داشته باشیم $k/n > 1/2$. در مورد تابع درست‌نمایی می‌دانیم که مقدار آن با افزایش p از صفر به سمت k/n افزایشی است و همچنین داریم $\mathbb{P}\left(p = k/n \mid k\right) > \mathbb{P}\left(p = 1/2 \mid k\right)$. اما در نقطه $p = 1/2$، مقدار توزیع پیشین یک افت ناگهانی داشته و مقدار آن $(1-\alpha)/\alpha$ برابر می‌شود و سپس ثابت می‌ماند. در این صورت لازم است که چک شود که شرط زیر برقرار است یا نه؟

$$(1-\alpha)\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} > \alpha\left(\frac{1}{2}\right)^k\left(1-\frac{1}{2}\right)^{n-k} \quad \text{or} \quad \left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} > \left(\frac{\alpha}{1-\alpha}\right)\frac{1}{2^n}$$

در صورتیکه برقرار بود، تخمین احتمال شیر آمدن کماکان همان k/n می‌شود چرا که شواهد درست‌نمایی توانسته بر اطلاعات پیشین فائق بیاید. در غیر این صورت، تخمین MAP برابر با $p^* = 1/2$ می‌شود.
به صورت خلاصه، می‌توان نوشت:

$$p^*_{\text{MAP}} = \begin{cases} \frac{k}{n} & k/n < 1/2 \quad \text{or} \quad \left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} > \left(\frac{\alpha}{1-\alpha}\right)\frac{1}{2^n} \\ \frac{1}{2} & o.w. \end{cases}$$

سوال ۴ (۴ نمره):

یک سازمان دولتی از شاخص عملکردی خود راضی نیست، و قصد دارد آن را بهبود ببخشد. شاخص عملکردی کل سازمان برابر با میانگین آماری شاخص عملکردی یکایک نیروهایش است، که می‌توان آنان را متغیرهای تصادفی هم‌توزیع ولی مستقل در نظر گرفت. سازمان یک بازرس انتخاب کرده تا نحوه انجام کار نیروهایش را بررسی کرده و شاخص عملکردی تعدادی از آنان را محاسبه کند.
بازرس تعداد ۱۰۰ نیرو را به صورت تصادفی انتخاب کرده و پس از بررسی سوابق کاری یکساله آنان شاخص‌های عملکردیشان را محاسبه نموده است. میانگین شاخص‌های عملکردی برای این جمعیت برابر با ۵۴٪ شده. همچنین بازرس انحراف معیار ۵٪ را کران بالایی برای انحراف معیار هر یک از ۱۰۰ اندازه‌گیری خود دانسته است.
سازمان به منظور بهبود عملکرد اخیراً تعدادی از مدیران اجرایی خود را تغییر داده. مدتی پس از این اتفاق، بازرس دوباره مشغول به کار شده و این بار ۱۰ نیرو را به صورت تصادفی انتخاب و میانگین شاخص عملکردی این جمعیت نمونه را

محاسبه کرده است. عدد بدست آمده این بار به ۵۷٪ افزایش پیدا کرده، و بازرس به دلیل دقت پایین‌تر در اثر کمبود وقت، این دفعه انحراف معیار ۸٪ را کران بالایی برای انحراف معیار اندازه‌گیری‌هایش گزارش کرده.
(دقت کنید که کران‌های بالای انحراف معیارها، یعنی ۵٪ و ۸٪، تخمین‌های شخصی خود بازرس هستند و مستقل از دادها مقداردهی شده‌اند).

حال، سازمان شما را به عنوان یک آماردان استخدام می‌کند و از شما می‌پرسد که آیا افزایش ۳ درصدی در شاخص عملکردی پس از تغییر مدیران اجرایی به لحاظ آماری «معنی‌دار» هست یا خیر؟

الف) فرضیه‌های H_0 و H_1 را در این حالت تعریف کرده و یک آماره مناسب برای انجام کار پیشنهاد دهید. همچنین توضیح دهید که آزمون فرضی که طراحی کرده‌اید دقیق (exact) است یا نادقیق.

ب) «میزان معنی‌دار بودن» افزایش شاخص عملکردی را بیابید. برای انجام این کار لازم است یک شاخص شناخته شده مانند p-value را محاسبه کنید.
(برای محاسبه دم یا tail نمودارها می‌توانید از اینترنت یا زبان‌های برنامه‌نویسی آماری مانند R استفاده کنید. همچنین برخی نامساوی‌ها در بخش راهنمایی‌ها نیز آمده که قابل استفاده هستند.)

فرض کنید که سازمان پس از مطالعه گزارش شما، اعلام می‌کند که p-value گزارش شده به اندازه کافی پایین نیست و لذا افزایش شاخص عملکردی سازمان هنوز معنی‌دار نشده است.

ج) اگر می‌توانستید اندازه فقط یکی از جمعیت‌های نمونه ۱۰۰ نفره (قبل از تعویض مدیران) و یا۱۰ نفره (بعد از آن) را به اندازه ۲۰ نفر افزایش داد تا مقدار p-value به بیشترین مقدار کمتر شود، شما پیشنهاد افزایش کدامیک را می‌دادید؟

جواب:

الف) قرار است ببینیم تغییر میانگین یک توزیع تصادفی معنی‌دار بوده یا نه. به عبارت دیگر، فرض کنید که شاخص‌های عملکردی نیروهای سازمان پیش از تغییر مدیران از توزیع f_1 با میانگین μ_1 و واریانس σ_1^2 پیروی کرده (در اینجا واریانس σ_1^2 مجموع واریانس ذاتی شاخص‌های عملکردی و دقت بازرس در اندازه‌گیری آنان است). این شاخص‌ها پس از تعویض مدیران از توزیع f_2 با میانگین و واریانس به ترتیب μ_2 و σ_2^2 پیروی می‌کنند.
سوال اصلی این است که آیا $\mu_1 = \mu_2$ یا خیر؟ لذا مطابق زیر المان‌های آزمون فرض را تعریف می‌کنیم:

- فرضیه H_0: میانگین‌ها برابرند، یعنی داریم $\mu_1 = \mu_2$.
- فرضیه H_1: میانگین‌ها برابر نیستند. یعنی داریم $\mu_1 < \mu_2$ یا $\mu_2 < \mu_1$.
- از آنجا برای انحراف معیارهای σ_1, σ_2 کران‌های بالای مستقل از داده داریم، می‌توانیم از آزمون آماری z-test استفاده کنیم. در صورتیکه مقادیر واریانس از روی داده‌ها محاسبه شده بود (و نه اطلاعات پیشین مستقل از دادگان) مجبور به استفاده از t-test بودیم.

آماره پیشنهادی بدین قرار است: باید آماره‌ای پیشنهاد داده شود که تحت فرضیه H_0 توزیع مشخصی داشته باشد. فرض کنید که $X_1, \ldots, X_{100}$ نمونه‌های i.i.d. متناظر با توزیع اولیه f_1، و نمونه‌های $Y_1, \ldots, Y_{10}$ مواردی مشابه و متناظر با توزیع f_2 باشند. فرض کنید که $\bar{X}_{100}$ و $\bar{Y}_{10}$ به ترتیب میانگین‌های نمونه‌ای متناظر با Xها و Yها باشند.
تحت فرضیه H_0 داریم:

$$\mathbb{E}\bar{X}_{100} = \mu_1 = \mu_2 \quad , \quad \mathrm{Var}\left(\bar{X}_{100}\right) = \frac{\sigma_1^2}{100} \leq \frac{5 \times 5}{100}$$
$$\mathbb{E}\bar{Y}_{10} = \mu_2 = \mu_1 \quad , \quad \mathrm{Var}\left(\bar{Y}_{10}\right) = \frac{\sigma_2^2}{10} \leq \frac{8 \times 8}{10}$$

با استفاده از تقریب گاوسی (تقریب قضیه حد مرکزی) در تعداد نمونه‌های بالا داریم برای $\bar{X}_{100}$ و $\bar{Y}_{10}$، و همچنین استقلال این دو از یکدیگر داریم:

$$Z \triangleq \frac{\bar{X}_{100} - \bar{Y}_{10}}{\sqrt{\frac{5^2}{100} + \frac{8^2}{10}}} \sim \mathcal{N}\left(0, \sigma^2\right), \quad \sigma < 1$$

که می‌بایست روی مقدار z-value فوق آزمون زده شود. در ضمن، به دلیل استفاده از تقریب گاوسی، آزمون فرض طراحی شده دقیق (exact) نیست و مقدار p-value به صورت دقیق محاسبه یا کراندار نمی‌گردد. مگر آنکه فرض کنیم خود توزیع‌های f_1 و f_2 گاوسی بوده باشند.

ب) به منظور چک کردن برقراری فرض باطل یا فرضیه جایگزین، می‌توان از یک آزمون دوطرفه ساده استفاده کرد. در این صورت لازم است تا با فرض $Z \sim \mathcal{N}\left(0, \sigma^2\right)$، احتمال زیر را محاسبه نمود:

$$\mathrm{p-value} = \mathbb{P}\left(|Z| > \frac{57 - 54}{\sqrt{\frac{25}{100} + \frac{64}{10}}} \right) \leq 0.245$$

که علامت کوچک‌تر یا مساوی به دلیل این است که واریانس z-value که با σ^2 نشان داده شده است با کران بالای خود یعنی ۱ تقریب خورده است.

ج) مقدار p-value برابر با 0.245 به وضوح بسیار بزرگ است و نشان می‌دهد افزایش شاخص عملکردی به لحاظ آماری معنی‌دار نشده است. به منظور بهبود z-value، با فرض ثابت ماندن درصدهای بدست آمده برای میانگین‌ها (یعنی ۵۷٪ و ۵۴٪) لازم است تا مخرج آن که مشتمل بر انحراف معیارهاست تا جای ممکن کوچک شود.

در صورت اضافه نمودن ۲۰ نفر جدید به جمعیت اولیه، مخرج کسر برابر با $\sqrt{\frac{25}{120}+\frac{64}{10}}$ و در صورت اضافه نمودن آنان به جمعیت دوم مخرج z-value برابر با $\sqrt{\frac{25}{100}+\frac{64}{30}}$ می‌گردد. واضح است که حالت دوم باعث افزایش چشمگیرتر مقدار z-value و لذا کاهش مقدار p-value خواهد گردید. مقدار کران بالای p-value در این حالت دوم برابر با 0.052 خواهد شد که کران بسیار بهتری است.

سوال ۵ (۴ نمره):

ملکول DNA در بدن موجودات زنده یک دنباله طولانی به طول N و متشکل از چهار الفبای اصلی A,C,G و T است. در اینجا، الفبای A,C,G,T بیانگر حروف ابتدایی از نام ۴ نوع منحصر به فرد از زیرملکولهایی به نام نوکلئوتید هستند.
فرض کنید که بتوانیم رشته DNA را به صورت تصادفی مدل‌سازی کنیم: فرض کنید هر یک از N حرف رشته مستقل از سایرین و با احتمال‌های یکسان 1/4 بتواند هر یک از اعضای الفبای A,C,G,T باشد (دقت کنید که در واقعیت اینطور نیست!). در این صورت، رشته DNA مورد بحث در این سوال، یک نمونه (یا realization) از این توزیع خواهد بود.

الف) قصد داریم بدانیم که یک زیررشته (substring) فرضی و ساختگی خاص به طول L (برای مثال، یک زیر رشته ۱۰۰تایی مانند $ACCGTATT\ldots GCC$) با چه احتمالی در حداقل یک جا از رشته DNA دیده خواهد شد. یک کران بالا برای احتمال خواسته شده بیابید.
(راهنمایی: می‌توانید از کران اجتماع یا Union Bound استفاده کنید)

ب) یک کران پایین نیز برای احتمال قسمت الف) پیدا کنید.
(راهنمایی: رشته DNA را به زیررشته‌های بدون همپوشانی با طول L تقسیم کنید)

ج) نشان دهید که در صورت انتخاب $L \geq (1+\varepsilon)\log_4 N$ احتمال دیده شدن یک زیررشته پیش فرض و ساختگی در رشته تصادفی DNA با افزایش N به سمت صفر میل می‌کند (به ازای هر $\epsilon > 0$). همچنین، در صورت انتخاب $L \leq (1-\varepsilon)\log_4 N$ احتمال مشاهده هر زیررشته ساختگی دلخواهی در رشته اصلی به سمت ۱ خواهد رفت.

د) (**امتیازی**) نشان‌دهید احتمال مشاهده زیررشته‌های خودتکرارشونده با طول $L \geq (2+\varepsilon)\log_4 N$ با افزایش N به سمت صفر میل خواهد کرد. رشته‌های خودتکرارشونده زیررشته‌هایی از DNA هستند که حداقل در یک جای دیگر از رشته نیز دوباره دیده می‌شوند.
(برای این بخش لازم است توضیحات مفصل ارائه دهید و صرف نوشتن روابط کافی نیست).

جواب:

الف) احتمال خواسته شده را با P نشان می‌دهیم. فرض کنید که رشته فرض مورد نظر را با S نشان دهیم. رخدادهای $A_1, A_2, \ldots, A_{N-L+1}$ را به صورت زیر تعریف می‌کنیم:

- رخداد A_i زمانی رخ می‌دهد که زیررشته شروع شده از اندیس i ام به طول L، تماماً با زیررشته پیش‌فرض S یکی باشد.؟

زیررشته پیش‌فرض S تصادفی نیست، ولی رشته اصلی به صورت تصادفی با احتمال‌های برابر ۱/۴ برای هر چهار عضو الفبا مدلسازی شده است. لذا احتمال یکی بودن هر یک از زیررشته‌های مورد نظر رخدادهای A_i با رشته ثابت S برابر با 4^{-L} خواهد بود. در اینجا از مستقل بودن اعضای رشته DNA نیز استفاده شده است.
مابقی ساده است. کافی است بنویسیم:

$$P = \mathbb{P}\left(A_1 \cup \ldots \cup A_{N-L+1}\right) \leq \sum_{i=1}^{N-L+1} \mathbb{P}\left(A_i\right) = (N-L+1)\,4^{-L} \leq N\left(\frac{1}{4}\right)^L$$

ب) دلیل استفاده از کران اجتماع در قسمت قبل این واقعیت بوده که رخدادهای A_i به دلیل همپوشانی‌های بالقوه از یکدیگر مستقل نبوده و امکان محاسبه اشتراک آنان بسیار سخت است. برای حل این مشکل، از کران پایین زیر استفاده می‌کنیم:

$$P = \mathbb{P}\left(A_1 \cup \ldots \cup A_{N-L+1}\right) \geq \mathbb{P}\left(A_1 \cup A_{L+1} \cup A_{2L+1} \cup \ldots \cup A_{L\lfloor N/L \rfloor - L+1}\right)$$

که معادل این است که رشته DNA به تعداد $\left\lfloor \frac{N}{L} \right\rfloor$ زیررشته با طول L تقسیم شده است که هیچ همپوشانی با یکدیگر ندارند. لذا رخدادهای متناظر با آنان از یکدیگر مستقل هستند. پس می‌توان نوشت:

$$\begin{aligned} P \geq \mathbb{P}\left(A_1 \cup A_{L+1} \cup A_{2L+1} \cup \ldots \cup A_{L\lfloor N/L \rfloor - L+1}\right) &= 1 - \mathbb{P}\left(A_1^c \cap A_{L+1}^c \cap \ldots \cap A_{L\lfloor N/L \rfloor - L+1}^c\right) \\ &= 1 - \prod_{i=0}^{\lfloor N/L \rfloor - 1} \mathbb{P}\left(A_{iL+1}^c\right) \end{aligned}$$

لذا خواهیم داشت:

$$P \geq 1 - \left(1 - \frac{1}{4^L}\right)^{\left\lfloor \frac{N}{L} \right\rfloor}$$

ج) کافی است که مقادیر اشاره شده برای L را در کران‌های بالا و پایین احتمال که در قسمت‌های الف) و ب) بدست آورده‌ایم جایگذاری کنیم. در مورد $L \geq (1+\varepsilon)\log_4 N$ داریم:

$$\lim_{N\to\infty} P \leq \lim_{N\to\infty} N \frac{1}{4^{(1+\varepsilon)\log_4 N}} = \lim_{N\to\infty} N^{-\varepsilon} = 0$$

و در مورد $L \leq (1-\varepsilon)\log_4 N$

$$\lim_{N\to\infty} P \geq \lim_{N\to\infty} 1-\left(1-\frac{1}{4^L}\right)^{\left\lfloor \frac{N}{L}\right\rfloor} = \lim_{N\to\infty} 1-e^{-\frac{N}{L4^L}} \geq \lim_{N\to\infty} 1-e^{-\frac{N}{N^{1-\varepsilon}\log N}} = \lim_{N\to\infty} 1-e^{-\frac{N^{\varepsilon}}{\log N}} = 1$$

د) (**امتیازی**) رخداد A_{ij} را برای $i,j = 1,2,\ldots,N-L+1$ (با فرض $i \neq j$) به این صورت تعریف می‌کنیم که زیررشته به طول L که از اندیس i شروع شده، با زیررشته‌ای با طول مشابه که از اندیس j شروع شده تماماً یکی باشد. در این صورت می‌توان نوشت:

$$\mathbb{P}\left(A_{ij}\right) = \prod_{k=0}^{L-1} \mathbb{P}\left(X_{i+k} = X_{j+k}\right) = \prod_{k=0}^{L-1} \left[\sum_{x\in\{A,C,G,T\}} \mathbb{P}\left(X_{i+k}=x\right)\mathbb{P}\left(X_{j+k}=x\right)\right] = 4^{-L}$$

مابقی کار ساده بوده و دوباره لازم است که یک کران اجتماع ساده در نظر گرفته شود:

$$\mathbb{P}\left(\text{self}-\text{repeat}\right) = \mathbb{P}\left(\bigcup_{i,j} A_{i,j}\right) \leq \binom{N-L+1}{2} 4^{-L} \leq N^2 4^{-L}$$

و در آخر با جایگذاری مقدار L مشابه با قسمت ج) خواهیم داشت:

$$\lim_{N\to\infty} \mathbb{P}\left(\text{self}-\text{repeat}\right) \leq \lim_{N\to\infty} N^2 \frac{1}{4^{(2+\varepsilon)\log_4 N}} = \lim_{N\to\infty} \frac{N^2}{N^2} N^{-\varepsilon} = 0$$

توضیحات و راهنمایی‌ها:

* سوالات امتیازی هر کدام ۱ نمره اضافه و مستقل از بارم‌بندی سوالات دارند.

* کران اجتماع: فرض کنید رویدادهای $A_1, \ldots, A_k$ زیرمجموعه‌هایی از فضای نمونه Ω (یا معادلاً اعضایی از مجموعه وقایع $\mathscr{F}$) باشند. در این صورت، همواره داریم:

$$\mathbb{P}\left(A_1 \cup \ldots \cup A_k\right) \leq \mathbb{P}\left(A_1\right) + \ldots + \mathbb{P}\left(A_k\right)$$

* در صورتیکه $X \sim \mathcal{N}(0,1)$، آنگاه کران زیر برای دم توزیع برقرار است: $\mathbb{P}(X > t) \leq \exp\left(-t^2/2\right)$.