



آمار و احتمال مهندسی

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

تمرین هفتم – مقدمه‌ای بر برآوردیابی، آزمون فرض و بازه اطمینان

طراح: امیرمهدی فرزانه

سوپروایزر: سارا معصومی

تاریخ تحویل: ۲۳ دی ۱۴۰۳

بیشتر بدانیم: خطای دادستان

شرح مسئله

خطای دادستان^۱ یک مسئله معروف در آمار و احتمالات است که نشان می‌دهد احتمال یک واقعه چگونه می‌تواند اشتباه تفسیر شده و نتایج نادرستی از آن گرفته شود.

فرض کنید شما یک وکیل مدافع هستید و موکل شما (آقای x) به قتل عمد متهم شده است. با این حال شما می‌دانید موکل‌تان بی‌گناه است و می‌خواهید بی‌گناهی او را ثابت کنید. در جلسه دادگاه دادستان اعلام می‌کند که دی‌ان‌ای موکل شما با دی‌ان‌ای پیدا شده روی سلاح قتل مطابقت دارد و به همین دلیل آقای x متهم شماره یک پرونده شناخته شده است.

شما به این موضوع اعتراض می‌کنید و به دادستان می‌گویید که ممکن است تصادفاً و به اشتباه دی‌ان‌ای آقای x با دی‌ان‌ای موجود روی سلاح منطبق شده باشد. دادستان اعتراض شما را نمی‌پذیرد و می‌گوید: "احتمال این که دی‌ان‌ای یک فرد بی‌گناه با دی‌ان‌ای مجرم منطبق شود ۱ در ۱ میلیون است. بدون شک آقای x گناهکار است و باید مجازات شود." استدلال دادستان ممکن است در نگاه اول منطقی به نظر برسد ولی اگر بیشتر به آن فکر کنیم متوجه مشکل بزرگ این استدلال خواهیم شد. به گفته دادستان،

$$P(\text{بی‌گناه بودن فرد} \mid \text{منطبق شدن دی‌ان‌ای فرد با دی‌ان‌ای مجرم}) \approx \frac{1}{1000000}$$

یا به بیان دیگر،

$$P(\text{بی‌گناه بودن فرد} \mid \text{گناهکار شناخته شدن فرد}) \approx \frac{1}{1000000}$$

اما آیا ناچیز بودن این احتمال ثابت می‌کند که آقای x گناهکار است؟ برای بررسی این موضوع فرض کنید شهری که این جرم در آن رخ داده است ۴ میلیون نفر جمعیت دارد. پس طبق ادعای دادستان، به طور میانگین ۴ نفر در این شهر وجود دارند که بی‌گناه‌اند ولی دی‌ان‌ای آن‌ها با دی‌ان‌ای مجرم تطابق پیدا می‌کند و گناهکار شناخته می‌شوند. همچنین یک مجرم نیز در این شهر وجود دارد. پس دی‌ان‌ای ۵ نفر می‌تواند با دی‌ان‌ای مجرم مطابقت داشته باشد. حال اگر یک فرد در دادگاه گناهکار شناخته شود احتمال بی‌گناه بودن آن چقدر خواهد بود؟ با اطلاعاتی که در دست داریم به سادگی می‌توانیم این احتمال را محاسبه کنیم،

$$\begin{aligned} P(\text{بی‌گناه بودن یک فرد و گناهکار شناخته شدن او}) &= \frac{P(\text{گناهکار شناخته شدن فرد} \mid \text{بی‌گناه بودن فرد})}{P(\text{گناهکار شناخته شدن یک فرد})} \\ &= \frac{\frac{4}{1000000}}{\frac{5}{1000000}} = 0.8 \end{aligned}$$

^۱The Prosecutor's Fallacy

در نتیجه احتمال این که آقای x به اشتباه گناهکار شناخته شده باشد ۰/۸ است که احتمال قابل توجهی است!!!

اشتباه استدلال دادستان این است که او فرض کرده آقای x بی‌گناه است و می‌گوید در اینصورت احتمال گناهکار شناخته شدن او ۱ در ۱ میلیون می‌باشد. در حالی که بی‌گناه بودن آقای x یک فرض نیست، بلکه دقیقاً موضوعی است که باید بررسی شود. پس روش درست برخورد با این مسئله این است که بررسی کنیم در صورت گناهکار شناخته شدن آقای x احتمال بی‌گناه بودن آن چقدر خواهد بود. اکنون ممکن است خطای دادستان احمقانه به نظر برسد، اما متأسفانه این یک خطای رایج و مکرر است. برای مثال می‌توان به پرونده غم‌انگیز **سالی کلارک^۲** اشاره کرد. او در سال ۱۹۹۸ به اشتباه به قتل دو فرزندش متهم و به حبس ابد محکوم شد.

آزمون فرض و ماتریس درهم‌ریختگی

اکنون می‌خواهیم برای بررسی اتهام موکل‌تان یک آزمون فرض طراحی کنیم. آزمون زیر را در نظر بگیرید،

H_0 : آقای x بی‌گناه است

H_1 : آقای x گناهکار است

در جدول زیر می‌توانید اطلاعاتی که از این آزمون در دست داریم را مشاهده کنید،

	تطبیق DNA فرد با DNA مجرم	عدم تطبیق DNA فرد با DNA مجرم
فرد مجرم باشد	1	0
فرد بی‌گناه باشد	4	4,000,000 – 4

از جدول بالا کاملاً مشخص است که،

$$P(\text{بی‌گناه} | \text{مجرم}) = \frac{4}{4,000,000} = \frac{1}{1,000,000}$$

$$P(\text{مجرم} | \text{بی‌گناه}) = \frac{4}{5} = 0.8$$

جدول بالا در واقع یک **ماتریس درهم‌ریختگی^۳** است و معمولاً آن را به صورت زیر نشان می‌دهند،

	(تطبیق DNA) H_0 رد	(عدم تطبیق DNA) H_0 عدم رد
H_0 نادرست است (مجرم)	1 True Positive	0 False Negative
H_0 درست است (بی‌گناه)	4 False Positive	4,000,000 – 4 True Negative

ماتریس درهم‌ریختگی سطح معناداری آزمون فرض را نیز نشان می‌دهد. جالب است بدانید که عبارت زیر معادل با α یا همان سطح معناداری آزمون فرض است،

$$P(\text{بی‌گناه} | \text{معجرم}) = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} = \alpha$$

به α خطای نوع اول آزمون^۴ نیز می‌گویند. خطای نوع اول آزمون در واقع نشان‌دهنده احتمال این است که فرض صفر برقرار باشد و ما به اشتباه آن را رد کنیم. در مسئله خطای دادستان دیدیم که خطای نوع اول آزمون یا همان α برابر با $\frac{1}{10000}$ بود که مقدار بسیار ناچیزی است. پس نتیجه می‌گیریم پایین بودن مقدار خطای نوع اول آزمون لزوماً به معنای دقیق بودن نتیجه آزمون نیست. در این‌جا می‌توانید بیشتر در مورد خطاهای آزمون فرض بخوانید.

۱. رژیم غذایی

۱۵ نمره

یک پزشک ورزشی یک برنامه غذایی تهیه کرده است و برای بررسی کارکرد آن، تعدادی داوطلب را به صورت تصادفی انتخاب کرده و از آن‌ها خواسته است که از این برنامه غذایی استفاده کنند. نتایج تغییر وزن داوطلبان به این صورت است:

۲۸ نفر به‌طور میانگین $0.3kg$ کاهش وزن داشته‌اند، وزن ۲۴ نفر هیچ تغییری نداشته است و ۸ نفر در مجموع $2/4kg$ افزایش وزن داشته‌اند. همچنین گشتاور مرتبه دوم تغییر وزن این داوطلبان برابر $5/4kg^2$ می‌باشد.

آ. یک‌بار با استفاده از آزمون فرض (با استفاده از z-test و به روش p-value)، و بار دیگر با روش بازه اطمینان به این پزشک کمک کنید تا بررسی کند آیا برنامه غذایی او باعث تغییر وزن داوطلبان شده است یا خیر. فرض کنید $\alpha = 0.05$. (۱۰ نمره)

ب. در قسمت آ، اگر تعداد داوطلبان انتخاب شده ۷ نفر بود نمی‌توانستیم برای انجام آزمون فرض از z-test استفاده کنیم. علت این موضوع را بیان کنید و سپس روشی جایگزین را نام ببرید که برای انجام آزمون فرض در این حالت مناسب باشد. (۵ نمره)

۲. انتخابات

۱۵ نمره

در انتخابات استانی، در صورتی که یک نامزد در بیشتر از نیمی از استان‌ها بیشترین رأی را آورده باشد برنده انتخابات خواهد بود. جدول زیر تعداد رأی‌های مربوط به ۵ استان را نشان می‌دهد. تعداد نمونه جمع‌آوری شده از هر استان ۲۰۰ تا می‌باشد.

شماره استان	رأی‌دهندگان به نامزد X	رأی‌دهندگان به نامزد Y
۱	۱۶۱	۳۹
۲	۸۹	۱۱۱
۳	۱۱۰	۹۰
۴	۱۸۸	۱۲
۵	۱۰۶	۹۴

با توجه به نتایج این نظرسنجی، آیا X می‌تواند نتیجه‌گیری کند که او برنده انتخابات خواهد بود؟ (از سطح اطمینان $\alpha = 0.05$ استفاده کنید.)

۳. اطلاعات از دست‌رفته بانک

۲۵ نمره

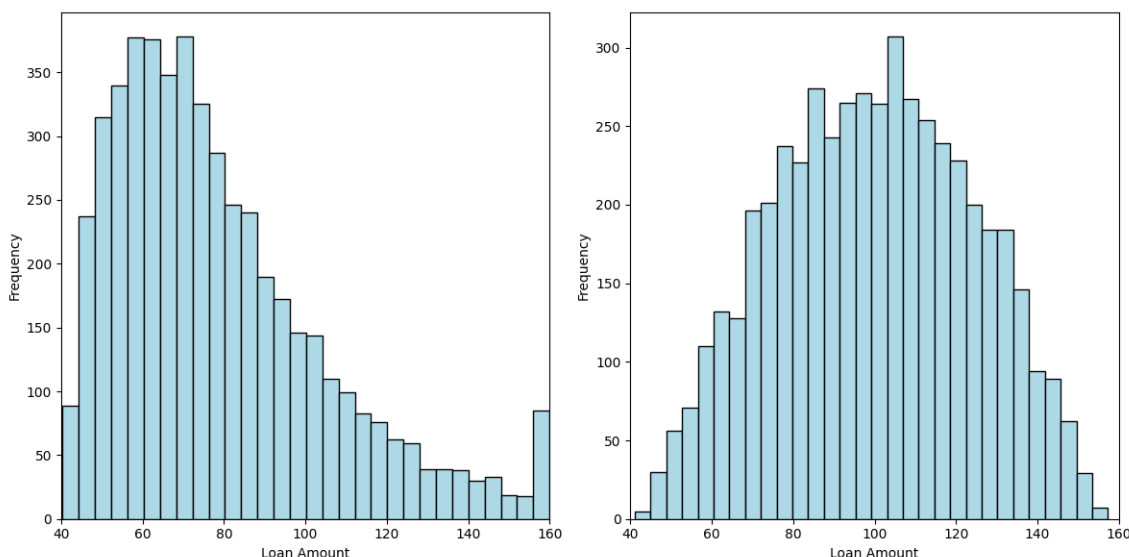
پس از حمله هکرها، بخشی از اطلاعات حساب مشتریان یک بانک از دست رفته است. از شما خواسته شده تا با استفاده از روش‌هایی که آموخته‌اید برخی از این اطلاعات از دست رفته را محاسبه کنید.

^۴Type ۱ error

آ. فرض کنید بعد از حمله هکری فقط اطلاعات حساب سپرده ۱۰۰ مشتری در دیتابیس بانک باقی مانده است. می‌دانیم با توجه به این نمونه، طول بازه اطمینان ۹۰ درصد برای میانگین میزان سپرده‌گذاری مشتریان (به میلیون تومان) برابر با ۱۸ است. از شما خواسته شده است با توجه به این اطلاعات انحراف معیار میزان سپرده‌گذاری افراد در این بانک را محاسبه کنید تا بدانیم پراکندگی نقدینگی حساب‌های سپرده در این بانک چقدر است. (۱۰ نمره)

ب. می‌دانیم نسبت افرادی که حساب سپرده خود را در سال آینده نخواهند بست حداکثر برابر با ۰/۴ است. حال می‌خواهیم از هکرها اطلاعات حساب سپرده برخی از مشتریان را به صورت تصادفی پس بگیریم. حداقل تعداد این مشتریان باید چقدر باشد تا مطمئن باشیم طول بازه اطمینان ۹۹ درصد نسبت افرادی که در سال آینده حساب خود را نمی‌بندند حداکثر ۰/۴ است؟ (۱۰ نمره)

ج. دو نمودار زیر نشانگر میزان وام دریافتی توسط دو گروه متفاوت از مشتریان هستند. از آنجایی که میانگین وام دریافتی هر دو گروه را از دست داده‌ایم، می‌خواهیم بازه اطمینان‌های ۹۵ درصد این دو میانگین را محاسبه کنیم. به این منظور کدام یک از این دو گروه نیاز به نمونه‌برداری با اندازه بزرگ‌تری دارد؟ توضیح دهید. (۵ نمره)



۲۵ نمره

۴. وب‌سایت خوب یا بد؟

یک برنامه‌نویس فرانت می‌خواهد بررسی کند در طراحی سایت چقدر عملکرد خوبی داشته است و برای این کار به کمک شما احتیاج دارد. فرض کنید زمان حضور کاربران در صفحه این سایت (به دقیقه) از توزیع ویبول^۵ پیروی می‌کند. تابع چگالی احتمال این توزیع بدین شکل است،

$$f(x; k, \lambda) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{(k-1)} e^{-(x/\lambda)^k}, \quad x \geq 0$$

آ. فرض کنید مقدار پارامتر k معلوم است. حال با استفاده از روش درست‌نمایی بیشینه^۶ برآوردگر پارامتر λ را به دست آورید. (۱۰ نمره)

ب. این بار با استفاده از روش گشتاوری^۷ برآوردگر پارامتر λ را محاسبه کنید. به ازای چه مقداری از k برآوردگر این قسمت معادل با برآوردگر قسمت قبل خواهد بود؟ (۵ نمره)

^۵Weibull Distribution

^۶Maximum likelihood estimation

^۷Method of moments

ج. نمونه جدیدی از زمان حضور ۱۶۰۰ کاربر در این وبسایت در اختیار داریم که میانگین آن ۸ است. اگر قبل از استخدام این برنامه‌نویس میانگین حضور کاربران در این وبسایت ۱۰ دقیقه بوده باشد، آیا این برنامه‌نویس عملکرد مناسبی در طراحی سایت و جذب کاربران داشته است یا خیر؟ $(\sigma = 0.8, \alpha = 0.01)$ (۱۰ نمره)

۵. دستگاه ونتیلاتور

۱۰ + ۱۰ نمره

در سال ۲۰۲۰، با شیوع گسترده‌ی COVID-19، بیمارستان‌های سراسر جهان نیاز مبرمی به دستگاه‌های ونتیلاتور^۸ برای بیماران مبتلا به مشکلات تنفسی شدید پیدا کردند. یکی از چالش‌های مهم، مدیریت زمان استفاده از این دستگاه‌ها و پیش‌بینی زمان خرابی آن‌ها بود. فرض کنید X_1, X_2, \dots, X_n یک نمونه تصادفی از توزیعی باشد که نسبت زمانی که نوعی ونتیلاتور بدون خرابی کار می‌کند به طول عمر مفید آن را نشان می‌دهد. تابع چگالی احتمال این توزیع به صورت زیر است،

$$f(x; \theta) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \quad 0 < \theta < \infty$$

آ. برآوردگر درست‌نمایی بیشینه θ را به دست آورید.

ب. آیا برآوردگر به دست آمده یک برآوردگر پایدار^۹ است؟ (امتیازی)
راهنمایی: $Y_i = \ln X_i \sim E(\frac{1}{\theta})$ و در نتیجه $\ln X_i \sim \text{Gamma}(n, \frac{1}{\theta})$ $T = -\sum_{i=1}^n \ln X_i$

۶. برآورد فاصله‌ای واریانس

۱۰ نمره

فرض کنید $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$. یک فاصله اطمینان $100(1 - \alpha)\%$ برای σ^2 به دست آورید.

راهنمایی: برای حل این سوال باید از توزیع کای-دو^{۱۰} کمک بگیریم. توزیع کای-دو با k درجه آزادی^{۱۱} $(\chi^2_{(k)})$ در واقع توزیع مجموع مربعات k متغیر تصادفی مستقل از توزیع نرمال استاندارد است. یعنی می‌توان گفت:

$$X_1, \dots, X_k \stackrel{\text{iid}}{\sim} N(0, 1) \longrightarrow Y = \sum_{i=1}^k X_i^2 \sim \chi^2_{(k)}$$

همچنین از این توزیع برای به دست آوردن توزیع واریانس نمونه‌ای استفاده می‌شود. می‌تواند نشان داد اگر یک نمونه تصادفی n تایی از یک توزیع نرمال با واریانس σ^2 داشته باشیم و S^2 واریانس این نمونه باشد خواهیم داشت،

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

اثبات کامل این قضیه را می‌توانید در این‌جا بخوانید.

می‌دانیم (a, b) یک بازه اطمینان $100(1 - \alpha)\%$ برای σ^2 است اگر داشته باشیم، $P(a < \sigma^2 < b) = 1 - \alpha$. حال با توجه به توضیحات داده شده این بازه اطمینان را پیدا کنید.

۷. توزیع نمایی دوتایی (امتیازی)

۱۰ نمره

فرض کنید X_1, X_2, \dots, X_n یک نمونه تصادفی از توزیعی با تابع چگالی زیر باشد.

$$f(x; \theta) = \frac{1}{\theta} e^{-|x-\theta|}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$$

با روش درست‌نمایی بیشینه برآوردگر پارامتر θ را به دست آورید.

^۸ Ventilator

^۹ Consistent estimator

^{۱۰} Chi-squared Distribution

^{۱۱} degrees of freedom