آمار و احتمال مهندسی

نیم‌سال اول ۱۴۰۳ـ۱۴۰۲

مدرس: دکتر نجفی

دانشکده‌ی مهندسی کامپیوتر

موعد تحویل: ۱۶ دی

تمرین سری ششم

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

مسئله‌ی ۱. (۱۴ نمره)

مادربزرگ مهدی که چایخور قهاری است، ادعا می‌کند که می‌تواند تشخیص دهد که ابتدا چای در فنجان ریخته شده یا آب جوش! اما مهدی که به تازگی در کلاس آمار و احتمال مهندسی با آزمون فرض آشنا شده تصمیم می‌گیرد که ادعای مادربزرگ را بررسی کند. به همین منظور او ۸ فنجان چای آماده می‌کند و در ۴ تای آن‌ها ابتدا چای و سپس آب جوش، و در ۴ تای دیگر ابتدا آب جوش و سپس چای را اضافه می‌کند. حال او ۸ فنجان به طور تصادفی روبروی مادربزرگش قرار می‌دهد تا او ۴ فنجانی که ابتدا در آن‌ها چای ریخته شده است را تشخیص دهد.

آ) برای آزمایش طراحی شده فرض صفر و جایگزین را بنویسید.

ب) آزمون فیشر را اجرا کرده و مقادیر p-value را بنویسید.

حل. آ) فرض صفر: مادربزرگ نمی‌تواند فنجان‌هایی که ابتدا در آن‌ها چای و سپس آب جوش ریخته شده است را تشخیص دهد.

فرض جایگزین: مادربزرگ می‌تواند فنجان‌هایی که ابتدا در آن‌ها چای و سپس آب جوش ریخته شده است را تشخیص دهد.

ب) امکان دارد مادربزرگ ۰،۲،۴،۶ و یا ۸ فنجان را به درستی تشخیص دهد. احتمال هر یک از این حالات را بدست می‌آوریم:

$$p_0 = \frac{\binom{4}{0}\binom{4}{0}}{\binom{8}{4}} = \frac{1}{70}, p_2 = \frac{\binom{4}{1}\binom{4}{1}}{\binom{8}{4}} = \frac{16}{70}, p_4 = \frac{\binom{4}{2}\binom{4}{2}}{\binom{8}{4}} = \frac{36}{70}, p_6 = \frac{\binom{4}{3}\binom{4}{3}}{\binom{8}{4}} = \frac{16}{70}, p_8 = \frac{\binom{4}{4}\binom{4}{4}}{\binom{8}{4}} = \frac{1}{70}$$

حال مقادیر p-value را بدست می‌آوریم:

$$\text{p-value}_0 = p_0 + p_2 + p_4 + p_6 + p_8 = 1$$
$$\text{p-value}_2 = p_2 + p_4 + p_6 + p_8 = \frac{69}{70}$$
$$\text{p-value}_4 = p_4 + p_6 + p_8 = \frac{53}{70}$$
$$\text{p-value}_6 = p_6 + p_8 = \frac{17}{70}$$
$$\text{p-value}_8 = p_8 = \frac{1}{70}$$

▷

مسئله‌ی ۲. (۱۵ نمره)

اداره هواشناسی شهر تهران، ۴ دستگاه سنجش آلودگی هوا را در یک منطقه قرار داده است. فرض کنید شاخص آلودگی هوا در این منطقه ثابت است اما این دستگاه‌ها دقیق نیستند و شاخص آلودگی را با کمی نویز گزارش می‌دهند. در یک روز نسبتا آلوده، مقادیر گزارش شده توسط این ۴ دستگاه به شرح زیر است:

$$۱۵۳, ۱۴۸, ۱۵۲, ۱۵۳$$

آ) با کمک داده‌های جمع‌آوری شده، یک بازه اطمینان ۹۹ درصدی برای شاخص آلودگی هوا در این منطقه ارائه دهید.

ب) اگر شاخص آلودگی هوا بیش از ۱۵۰ باشد، هوا در شرایط ناسالم برای تمامی گروه‌ها قرار می‌گیرد و مدارس و دانشگاه‌ها تعطیل می‌شوند. برخی از صاحب نظران معتقند که از آنجایی که میانگین شاخص آلودگی هوا بیش از ۱۵۰ بوده، هوا ناسالم است و مدارس و دانشگاه‌ها می‌بایست تعطیل شوند. از طرفی گروهی دیگر از صاحب نظران بر این باورند که میانگین بدست آمده به طور محسوس و معناداری از ۱۵۰ بیشتر نبوده و مدارس و دانشگاه‌ها نباید تعطیل شوند. با توجه به اینکه p-value قابل قبول برای تعطیلی مدارس حداکثر ۰/۱ است، بررسی کنید که آیا مدارس تعطیل شوند یا خیر.

پ) برای کاهش خطای نوع اول و دوم باید مقدار $p-value$ قابل قبول را افزایش دهیم یا کاهش؟ چرا؟

حل.

آ) از آنجایی که توزیع عدد گزارش شده توسط هر دستگاه از توزیع نرمال با واریانس نامشخص می‌آید، می‌توانیم توزیع میانگین را Student’s t در نظر بگیریم. با توجه به این نکته برای بازه اطمینان خواهیم داشت:

$$-t_{/۰۰۵} \leqslant \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \leqslant t_{/۰۰۵} \to \bar{x} - \frac{s}{\sqrt{n}} t_{/۰۰۵} \leqslant \mu \leqslant \bar{x} + \frac{s}{\sqrt{n}} t_{/۰۰۵}$$

با توجه به اعداد گزارش شده و از آنجایی که $df = n - ۱ = ۳$ داریم:

$$\bar{x} = \frac{۱۵۳ + ۱۴۸ + ۱۵۲ + ۱۵۳}{۴} = ۱۵۱/۵, \quad s = \sqrt{\frac{۱/۵^۲ + ۳/۵^۲ + ۰/۵^۲ + ۱/۵^۲}{۳}} \approx ۲/۳۸۰۴$$

چون $t_{/۰۰۵} \approx ۵/۸۴۱$ است، می‌توان بازه اطمینان ٪۹۹ را بدست آورد:

$$۱۵۱/۵ \pm ۶/۹۵$$

ب) از آزمون تک نمونه‌ای t-test با فرض صفر $\mu = ۱۵۰$ و فرض مقابل $\mu > ۱۵۰$ استفاده می‌کنیم. مقدار t-value داده‌ها را به صورت زیر محاسبه می‌کنیم:

$$\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \approx ۱/۲۶۰۲ < t_{۰/۱} \approx ۱/۶۳۸$$

پس نمی‌توانیم فرض صفر را رد کنیم و در نتیجه مدارس و دانشگاه‌ها تعطیل نمی‌شوند.

پ) خطای نوع اول برابر با سطح اهمیت آزمون است بنابراین برای کاهش خطای نوع اول باید مقدار عددی سطح اهمیت را کاهش دهیم. خطای نوع دوم هنگامی اتفاق می‌افتد که فرض صفر غلط باشد و ما به اشتباه نتوانیم آن را رد کنیم. افزایش مقدار عددی سطح اهمیت رد کردن فرض صفر را ساده کرده و احتمال آن را بیشتر می‌کند (در حالی که تاثیری روی درست یا غلط بودن فرض صفر در عالم واقعیت ندارد!) بنابراین افزایش مقدار عددی سطح اهمیت باعث کاهش احتمال خطای نوع دوم می‌شود. ▷

مسئله‌ی ۳. (۱۴ نمره)

متغیر تصادفی پیوسته X با تابع چگالی احتمالاتی زیر را در نظر بگیرید:

$$f_X(x) = \begin{cases} ۳x^۲ & x \in [۰, ۱] \\ ۰ & o.w. \end{cases}$$

اگر $Y|X = x \sim \text{Geometric}(x)$ باشد، تخمین MAP متغیر تصادفی X به شرط $Y = ۳$ را بدست آورید.

حل. با توجه به تعریف توزیع هندسی برای $y = ۱, ۲, \cdots$ داریم:

$$P_{Y|X}(y|x) = x(۱ - x)^{y-۱}$$

پس با جایگزینی $y = ۳$ خواهیم داشت:

$$P_{Y|X}(۳|x) = x(۱ - x)^۲$$

اکنون باید به دنبال $x \in [۰, ۱]$ بگردیم که مقدار زیر را بیشینه کند:

$$P_{Y|X}(y|x)f_X(x) = x(۱ - x)^۲ \cdot ۳x^۲ = ۳x^۳(۱ - x)^۲$$

برای یافتن بیشینه این تابع از آن برحسب x مشتق گرفته و حاصل را برابر با صفر قرار می‌دهیم.

$$\frac{d}{dx}\left[۳x^۳(۱ - x)^۲\right] = ۹x^۲(۱ - x)^۲ - ۶x^۳(۱ - x) = ۰ \rightarrow \begin{cases} x = ۰ \\ x = \frac{۳}{۵} \quad \max \\ x = ۱ \end{cases}$$

با بررسی مقادیر فوق تخمینگر MAP را بدست می‌آوریم و داریم:

$$\hat{x}_{MAP} = \frac{۳}{۵}$$

$\triangleright$

مسئله‌ی ۴. (۲۰ نمره)

مسئله‌ی رگرسیون خطی ساده را در نظر بگیرید که در آن ورودی‌های $\{(x_i, y_i)\}_{i=۱}^n$ داده شده است. مقادیر $\{x_i\}_{i=۱}^n$ یقینی و مشخص هستند. اما به ازای هر x_i مقدار y_i از طریق رابطه‌ی زیر به دست می‌آید:

$$y_i = \beta_۰ + \beta_۱ x_i + \epsilon_i$$

که $\epsilon_i \sim \mathcal{N}(۰, \sigma^۲)$. اگر بدانیم که $\beta_۰$، $\beta_۱$ و σ نامنفی هستند، به پرسش‌های زیر پاسخ هید.

آ) اثبات کنید که تخمین بیشینه درست‌نمایی دو پارامتر $\beta_۰$ و $\beta_۱$ معادل انتخاب مقادیری برای $\beta_۰$ و $\beta_۱$ است که میانگین مربعات خطا را کمینه می‌کند.

ب) اثبات کنید که تخمین‌های بدست آمده در بخش پیشین نااُریب بوده و از توزیع‌های زیر پیروی می‌کنند:

$$\hat{\beta}_۱ \sim \mathcal{N}\left(\beta_۱, \frac{\sigma^۲}{\sum_i (x_i - \bar{x})^۲}\right), \quad \hat{\beta}_۰ \sim \mathcal{N}\left(\beta_۰, \frac{\sigma^۲ \sum_i x_i^Y}{n\sum_i (x_i - \bar{x})^۲}\right)$$

پ) بررسی کنید که آیا تخمینگر بیشینه درست‌نمایی عضوی از خانواده‌ی خطی تخمینگرهای زیر است یا نه؟ اگر هست رابطه‌ی γ_i را برحسب داده‌های ورودی بدست آورید.

$$\tilde{\beta_۱} = \frac{\sum \gamma_i y_i}{\sum \gamma_i x_i} \quad \text{such that} \quad \sum_i \gamma_i = ۰$$

ت) اثبات کنید هر تخمینگری که عضو خانواده فوق است نااُریب می‌باشد.

ث) اثبات کنید به ازای هر انتخابی از مقادیر γ_i در خانواده فوق داریم $Var(\hat{\beta_1}) \leqslant Var(\tilde{\beta_1})$. نتیجه بدست آمده را توضیح دهید.

حل.

آ)

$$f_L = f(y_1, \cdots, y_n|\beta_0, \beta_1) = \prod_{i=1}^{n} p(y_i|\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\Big)$$

$$\to f_L = (2\pi)^{\frac{-n}{2}} \sigma^{-n} \exp\Big(- \frac{\sum_i (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\Big)$$

$$\mathcal{L} = \log(f_L) = c - \frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial \mathcal{L}}{\partial \beta_0} \propto \sum (y_i - \beta_0 - \beta_1 x_i) = 0 \to \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\frac{\partial \mathcal{L}}{\partial \beta_1} \propto \sum (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \to \beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

ب) رابطه را برای $\hat{\beta_1}$ اثبات می‌کنیم و باقی روند برای $\hat{\beta_0}$ به همین صورت می‌باشد.

$$\hat{\beta_1} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum_i \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} y_i$$

از آنجایی که y_i خود یک توزیع گاوسی دارد، $\hat{\beta_1}$ نیز جمع تعدادی جمله‌ی گاوسی می‌شود و در نتیجه توزیع آن نیز گاوسی می‌شود. پس داریم:

$$\mathbb{E}[\hat{\beta_1}] = \sum_i \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i) = \beta_1$$

$$Var(\hat{\beta_1}) = \sum_i \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 Var(y_i) = \frac{\sigma^2}{\sum (x_i - \bar{x})}$$

پ) بدیهیست با جایگذاری $\gamma_i = x_i - \bar{x}$ به همیان تخمین کمینه‌ی مربعات خطا می‌رسیم.

ت)

$$\mathbb{E}[\tilde{\beta_1}] = \sum_i \frac{\gamma_i}{\sum \gamma_i x_i} \mathbb{E}[y_i] = \sum_i \frac{\gamma_i}{\sum \gamma_i x_i} (\beta_0 + \beta_1 x_i) = 0 + \frac{\sum \gamma_i x_i}{\gamma_i x_i} \beta_1 = \beta_1$$

ث)

$$\frac{\gamma_i}{\sum \gamma_i x_i} := d_i \to Var(\tilde{\beta_1}) = \sigma^2 \sum d_i^2$$

$$\sigma^2 \sum d_i^2 \sum (x_i - \bar{x})^2 \geqslant \sigma^2 \Big(\sum d_i (x_i - \bar{x}) \Big)^2 = \sigma^2$$

$$\to \sigma^2 \sum d_i^2 = Var(\tilde{\beta_1}) \geqslant \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = Var(\hat{\beta_1})$$

در این خانواده از تخمینگرها همه نااُریب هستند و تخمین بیشینه درست‌نمایی در میان آن‌ها از کمترین واریانس برخوردار است پس می‌توان ادعا کرد که این تخمین بهترین است. ▷

مسئله‌ی ۵. (۱۵ نمره)

یک مدل رگرسیون خطی به شکل $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ در شرایطی که خطای ان از توزیع نرمال $\epsilon \sim N(0, \sigma^2)$ پیروی میکند داریم.

MLE ضرایب $\beta_0, \beta_1, \beta_2$ را بیابید.

حل.

$$L(\beta_0, \beta_1, \beta_2, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2}{2\sigma^2}\right)$$

$$l(\beta_0, \beta_1, \beta_2, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0$$

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) x_{i1} = 0$$

$$\frac{\partial l}{\partial \beta_2} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) x_{i2} = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}1 - \hat{\beta}2\bar{x}2$$

$$\hat{\beta}1 = \frac{\sum i = 1^n (xi1 - \bar{x}1)(y_i - \bar{y})}{\sum i = 1^n (xi1 - \bar{x}1)^2}$$

$$\hat{\beta}2 = \frac{\sum i = 1^n (xi2 - \bar{x}2)(y_i - \bar{y})}{\sum i = 1^n (xi2 - \bar{x}_2)^2}$$

▷

مسئله‌ی ۶. (۱۰ نمره)

فرض کنید میانگین معدل همه ی دانشجویان فارغ‌التحصیل‌شده از دانشگاه صنعتی شریف در سال ۲۰۰۵ برابر ۳/۰۵ (بر اساس معیار GPA) بوده است. سازمان ثبت احوال قصد دارد سوابق ۱۰۰ دانشجوی فارغ التحصیل در سال ۲۰۲۲ را بررسی کند تا ببیند میانگین GPA تغییر کرده است یا خیر. در این نمونه‌ی ۱۰۰ تایی، میانگین معدل برابر ۲/۱۵ بوده و انحراف معیار برابر ۱/۵ می‌باشد.

الف. فرصیه های صفر و جایگزین را برای این تحقیق بیان کنید.

ب. قبول/رد فرضیه H_0 را در مقابل H_1 جایگزین را در سطح اهمیت $\alpha = 0.05$ بررسی کنید. فرض کنید حجم نمونه به اندازه‌ی کافی بزرگ بوده و توزیع میانگین نمونه از توزیع نرمال پیروی می‌کند. (منظور از سطح اهمیت، Significance Level می‌باشد.)

حل.

الف) فرض صفر: $H_0 : \mu = 3.05$. فرض جایگزین: $H_1 : \mu \neq 3.05$

ب)

$$t_{oss} = \sqrt{n} \times (\bar{x} - \mu_0)/\sigma = \sqrt{100} \times (2.15 - 3.05)/1.5 = -6$$

بنابراین

$$p - value = P(Z > |t_{oss}|) = 2P(Z > 6) \approx 0$$

چون $p - value < \alpha$ پس فرض صفر رد می شود. ▷

مسئله‌ی ۷. (۱۲ نمره)

داده $\{(w[i], x[i])\}_{i=0}^{N-1}$ مشاهده شده است. می‌دانیم $x[i] = A + w[i]$ که A یک پارامتر نامعلوم با توزیع چگالی احتمال پیشین زیر است.

$$p(A) = \begin{cases} \lambda \exp(-\lambda A) & A \geqslant 0 \\ 0 & A < 0 \end{cases}$$

که $\lambda > 0$ و $W[i]$ها نمونه‌هایی از توزیع نرمال با میانگین ۰ و واریانس σ^2 می‌باشند. همچنین مقادیر $w[i]$ از A مستقل هستند. برآوردگر MAP را روی A بیابید.

حل. ابتدا می‌دانیم

$$p(w[n]) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{w^2[n]}{2\sigma^2})$$

در نتیجه با توجه به اینکه $x[n] = A + w[n]$، خواهیم داشت:

$$p(x[n]|A) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x[n]-A)^2}{2\sigma^2})$$

برای محاسبه‌ی Liklihood خواهیم داشت

$$p(x|A) = \prod_{i=0}^{N-1} p(x[n]|A) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n]-A)^2)$$

اگر تخمین MAP متغیر A را $\hat{A}_{MAP}$ بنامیم، خواهیم داشت:

$$\hat{A}_{MAP} = \arg\max_A \; p(A|x) = \arg\max_A \; p(x|A)p(A) = \arg\max_A \; (\log p(x|A) + \log p(A))$$

اگر توزیع A و $x|A$ را در این رابطه قرار دهیم، خواهیم داشت:

$$\hat{A}_{MAP} = \arg\max_A \; -\frac{1}{2\sigma^2} \sum_{i=0}^{N-1} (x[n]-A)^2 - \lambda A$$

اگر تابع هدف را $f(A)$ نامیده و مشتق آن را برابر ۰ قرار دهیم، خواهیم داشت:

$$\frac{df(A)}{dA} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n]-A) - \lambda = 0 \leftrightarrow A = \bar{x} - \frac{\lambda\sigma^2}{N}$$

که $\bar{x} = \sum_{n=0}^{N-1} (x[n])$ میانگین نمونه است. از آن‌جایی که

$$\frac{d^2 f(A)}{dA^2} = -\frac{N}{\sigma^2} - \lambda < 0$$

مراحل بالا $f(A)$ را بیشینه می‌کند.

در نهایت با در نظر گرفتن اینکه احتمال رخداد مقادیر منفی برای A مساوی ۰ است، خواهیم داشت:

$$\hat{A}_{MAP} = max(0, \bar{x} - \frac{\lambda\sigma^2}{N})$$

▷

موفق باشید :)