آمار و احتمالات مهندسی

تمرین هفتم - آمار

حسام و علیرضا

تاریخ تحویل ۱۴۰۱/۱۰/۲۲

سؤال ۱.

فرض کنید برنامه‌ای با $n = 100$ فایل کد داریم. بر اثر حمله‌ای که به سرور حاوی فایل‌های برنامه انجام شده، تعدادی باگ به هر فایل اضافه شده است. فرض کنید متغیر X_i تعداد باگ در فایل iاُم است که از توزیع پواسون با میانگین ۱ پیروی می‌کند. همچنین تعداد باگ در هر فایل مستقل از فایل‌های دیگر است. احتمال اینکه تعداد کل باگ‌های اضافه شده به برنامه کمتر از ۹۰ باشد را تخمین بزنید.

پاسخ.

با استفاده از قضیه حد مرکزی داریم:

$$\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

همچنین می‌دانیم در توزیع پواسون داریم:

$$\mu = E[X_i] = \lambda = 1$$
$$\sigma^2 = Var[X_i] = \lambda$$

بنابراین می‌توان گفت: $\bar{X}_n \sim \mathcal{N}(\lambda, \frac{\lambda}{n})$

تعداد کل باگ‌های اضافه شده به برنامه به صورت $Y = \sum_{i=1}^{n} X_i$ محاسبه می‌شود.

$$Y = \sum_{i=1}^{n} X_i = n\bar{X}_n \rightarrow P(Y < 90) = P(n\bar{X}_n < 90) = P(\bar{X}_n < 0.9)$$

$$= P(\frac{\bar{X}_n - \lambda}{\frac{\sqrt{\lambda}}{\sqrt{n}}} < \frac{0.9 - \lambda}{\frac{\sqrt{\lambda}}{\sqrt{n}}}) = \Phi(\frac{0.9 - \lambda}{\frac{\sqrt{\lambda}}{\sqrt{n}}}) = \Phi(-1) = 1 - \Phi(1) = 0.159$$

سؤال ۲.

فرض کنید انتخابات شورای شهر پیش رو است که در آن ۲ کاندیدا با هم به رقابت خواهند پرداخت. یک شرکت نظرسنجی تصمیم گرفته با کمک شما نتیجه انتخابات را پیش‌بینی کند. این شرکت قصد دارد طوری نظرسنجی خود را انجام دهد که با ۹۶ درصد اطمینان، خطای پیش‌بینی کمتر از ۱ درصد باشد.

الف) حداقل از چند نفر باید نظرسنجی کنیم تا دقت خواسته شده توسط شرکت به دست آید؟

ب) اگر تعداد افراد واجد شرایط رای دادن در شهر کمتر از تعداد افرادی باشد که در قسمت الف به دست آمد، چه اتفاقی خواهد افتاد؟ این موضوع را تفسیر کنید.

پاسخ .

الف) برای iاُمین نفری که به طور اتفاقی انتخاب شده تعریف می‌کنیم:

$$X_i = \begin{cases} ۰ & \text{برای کاندید A} \\ ۱ & \text{برای کاندید B} \end{cases}$$

میانگین جامعه را به دست می‌آوریم:

$$\bar{X}_n = \frac{X_۱ + X_۲ + \ldots + X_n}{n}$$

خواسته‌ی شرکت این است که با اطمینان ۹۶ درصد خطای پیش‌بینی کمتر از ۱ درصد باشد:

$$P(|\bar{X}_n - \mu| \geq ۰/۰۱) \leq ۰/۰۴$$

با استفاده از نامساوی چبیشف داریم:

$$P(\mu - \epsilon < X < \mu + \epsilon) \geq ۱ - \frac{\sigma^۲}{\epsilon^۲} \rightarrow P(|x - \mu| \geq \epsilon) \leq \frac{\sigma^۲}{\epsilon^۲}$$

$$P(|\bar{X}_n - \mu| \geq ۰/۰۱) \leq \frac{\sigma^۲_{\bar{X}_n}}{(۰/۰۱)^۲} = \frac{\sigma^۲_x}{n(۰/۰۱)^۲}$$

متغیر تصادفی X دارای توزیع برنولی است. طبق اصل ناکافی بودن دلیل می‌دانیم که اگر یک آزمایش تصادفی دارای N نتیجه ممکن باشد و ما هیچ اطلاعی درباره نحوه وقوع آن نداشته باشیم، باید احتمال وقوع هرکدام از آن‌ها را مساوی فرض کنیم. پس با توجه به اصل ناکافی بودن دلیل می‌توانیم فرض کنیم که احتمال رای دادن شخص به هردو کاندید مساوی است (چون صورت سوال هیچ اطلاعاتی در مورد درصد محبوبیت کاندیدها یا احتمال انتخاب هر کاندید توسط اشخاص به ما نداده) پس داریم:

$$X \sim Bernoulli(۰/۵) \rightarrow \sigma^۲_X = p(۱ - p) = \frac{۱}{۴}$$

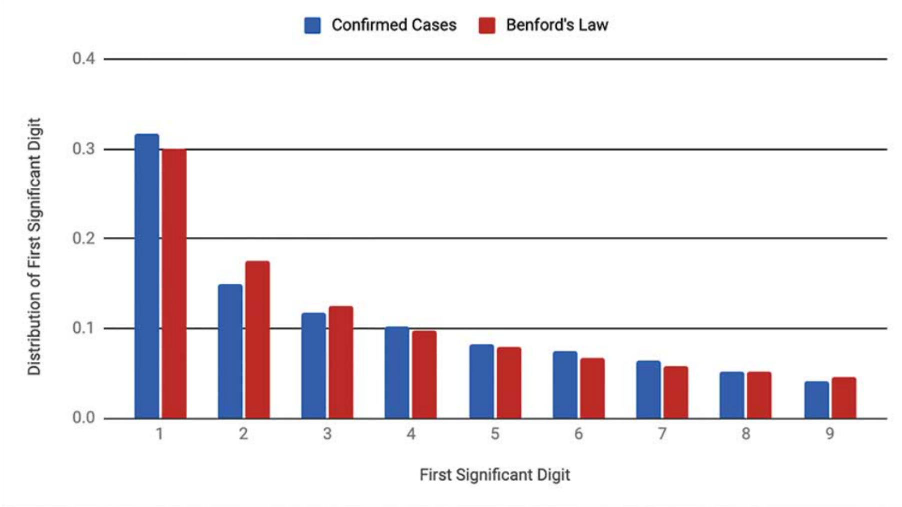
$$\frac{۱}{۴n(۰/۰۱)^۲} = ۰/۰۴ \rightarrow n = \frac{۱}{۴ \times ۰/۰۴(۰/۰۱)^۲} = ۶۲۵۰۰$$

در نتیجه اگر ۶۲۵۰۰ نفر (که به صورت تصادفی و با روش sampling درست انتخاب شده‌اند) در نظرسنجی شرکت کنند، دقت خواسته شده توسط شرکت به دست خواهد آمد.

ب) اگر تعداد کل افراد واجد شرایط رای دادن، کمتر از ۶۲۵۰۰ نفر باشد، کران به دست آمده توسط نامساوی چبیشف بی‌معنی خواهد بود، چرا که این بدان معنی است که باید از کل جمعیت واجد شرایط رای دادن نظرسنجی کنیم تا با دقت ۹۶ درصد اطمینان داشته باشیم که خطای پیش‌بینی ما کمتر از ۱ درصد خواهد بود. در حالی که می‌دانیم نظرسنجی از کل افراد به معنی اطمینان ۱۰۰ درصدی از پیش‌بینی است. این یکی از ضعف‌های کران به دست آمده توسط نامساوی چبیشف است. دقت کنید که اهمیت نامساوی چبیشف در عمومیت آن است. یعنی هیچ فرض و محدودیتی درباره متغیر تصادفی برای برقراری این نامساوی نیاز نیست (البته به جز متناهی بودن واریانس). هرگاه آگاهی بیشتری (افزون بر میانگین و واریانس) درباره‌ی متغیر تصادفی داشته باشیم (مثلا بیشترین مقدار n در مثال بالا) می‌توانیم کران‌های دقیق‌تری نسبت به کران چبیشف به دست آوریم.

سؤال ۳.

نمودار زیر، نشان‌دهنده توزیع پرارزش‌ترین رقم میزان مبتلایان روزانه به کووید ۱۹ در آمریکا از ژانویه تا اکتبر سال ۲۰۲۰ (به مدت ۲۹۰ روز) می‌باشد. مطابق نمودار، پرارزش‌ترین رقم تعداد افراد مبتلا شده در ماه‌های مختلف، از قانون بنفورد پیروی می‌کند! شما به عنوان متخصص آمار،



با فرض صحیح بودن آمار ابتلای کووید در کشور آمریکا قصد دارید تا با استفاده از قانون بنفورد، صحت آمار دیتاست دیگری که به شما داده شده است را بدست آورید.

مطابق با قانون بنفورد، به احتمال ۰/۴۷ می‌بایست پرارزش‌ترین رقم تعداد مبتلایان در هر روز برابر با عدد ۱ یا ۲ باشد. یک دیتاست از میزان مبتلایان در جهان در یک بازه ۴۳ روزه به شما داده شده است. در دیتاست داده شده، در ۱۲ روز از ۴۳ روز، پرارزش‌ترین رقم تعداد مبتلایان برابر با ۱ یا ۲ می‌باشد. احتمال رخ دادن حالت ذکر شده برای پرارزش‌ترین رقم داده‌ها را در دیتاست داده شده بدست آورید، سپس با توجه به قانون بنفورد، صحت دیتاست داده شده را بررسی کنید. (راهنمایی: از توزیع دوجمله‌ای استفاده کنید. برای بررسی صحت دیتاست داده شده، کافی‌است تا درصد اختلاف حالت ذکر شده را با پیش‌بینی قانون بنفورد مقایسه کنید.)

پاسخ .

متغیر تصادفی X را برابر با تعداد روزهایی در نظر می‌گیریم که پرارزش‌ترین رقم آنها برابر با ۱ یا ۲ می‌باشد.
احتمال آنکه پرارزش‌ترین رقم داده‌ها برابر با ۱ یا ۲ باشد را میتوان با توزیع دوجمله‌ای با پارامترهای n و p بیان کرد. پس داریم:

$$\binom{۴۳}{۱۲} \times (۰/۴۷)^{۱۲} \times (۰/۵۳)^{۳۱} = ۰/۰۵$$

حال به بررسی صحت دیتای داده شده میپردازیم. آزمون فرض را به صورت مقابل مینویسیم:

$H_۰$: واقعی بودن دیتاست داده شده

H_A: غیرواقعی بودن دیتاست داده شده

مطابق با قانون بنفورد، پرارزش‌ترین رقم اعداد داده شده با احتمال ۰/۴۷ برابر با ۱ یا ۲ می‌باشد. پس به طور متوسط در یک دیتاست صحیح، تعداد $۲۰ \approx ۴۳ \times ۰/۴۷$ عدد از داده‌ها دارای پرارزش‌ترین رقمی برابر با ۱ یا ۲ می‌باشند.
بنابراین برای صحت‌سنجی داده‌های داده شده، اختلاف تعداد روزهایی که دارای رقم پرارزش ۱ یا ۲ هستند را با ۲۰ مقایسه کرده و صحت آن را می‌سنجیم:

$$\textbf{Difference Rate} = \frac{|۱۲ - ۲۰|}{۲۰} = ۰/۴$$

دیتای فوق ۴۰ درصد با میزان بدست آمده توسط قانون بنفورد تفاوت دارد، پس می‌توانیم نتیجه بگیریم که دیتاست داده شده دارای خطا است و واقعی نیست و فرض صفر را رد میکنیم.

سؤال ۴.

فرض کنید $X_1, X_2, \ldots, X_n$ یک نمونه تصادفی nتایی از توزیعی با چگالی احتمال زیر باشد.

الف) برآورد گشتاوری پارامتر θ را به دست آورید. سپس واریانس این توزیع را برحسب پارامتر تخمین‌زده شده بنویسید.

ب) همچنین برای نمونه‌ی ۱۰ تایی به دست آمده از توزیع زیر، پارامتر به دست آمده توسط تخمین‌گر Maximum-Likelihood را محاسبه کنید.

$$f_X(x) = \begin{cases} \theta x^{\theta-1} & 0 < x < 1 \\ 0 & o.w. \end{cases}$$

$$Sample = \{0/0710,\ 0/2110,\ 0/4698,\ 0/3675,\ 0/5991,\ 0/9513,\ 0/6049,\ 0/9917,\ 0/1551,\ 0/2154\}$$

پاسخ .

الف) به دلیل اینکه دو پارامتر مجهول داریم، کافی است $E[X]$ و $E[X^2]$ را محاسبه کنیم:

$$E[X] = \int x f_X(x)dx = \int_0^1 \theta x^{\theta} dx = [\frac{\theta x^{\theta+1}}{\theta+1}]_0^1 = \frac{\theta}{\theta+1}$$

$$E[X^2] = \int x^2 f_X(x)dx = \int_0^1 \theta x^{\theta+1} dx = [\frac{\theta x^{\theta+2}}{\theta+2}]_0^1 = \frac{\theta}{\theta+2}$$

همچنین داریم:

$$Var[X] = E[X^2] - E[X]^2 = \frac{\theta}{\theta+2} - (\frac{\theta}{\theta+1})^2 = \frac{\theta}{(\theta+2)(\theta+1)^2}$$

با استفاده از روش گشتاورها داریم:

$$\hat{M}_1 = E[X] \implies \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\theta}{\theta+1} \implies \theta = \frac{\bar{X}}{1-\bar{X}}$$

بنابراین برآوردکننده گشتاوری میانگین و واریانس به صورت زیر خواهد بود:

$$\mu = \frac{\theta}{\theta+1} \rightarrow \mu = \bar{X}$$

$$Var[X] = \frac{\theta}{(\theta+2)(\theta+1)^2} = \frac{\frac{\bar{X}}{1-\bar{X}}}{(\frac{2-\bar{X}}{1-\bar{X}})(\frac{1}{1-\bar{X}})^2} = \frac{\bar{X}(1-\bar{X})^2}{2-\bar{X}}$$

ب) تابع likelihood به صورت زیر تعریف می‌شود:

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n (\prod_{i=1}^{n} x_i)^{\theta-1}$$

حال برای تابع log-likelihood داریم:

$$LL(\theta) = n \ln\theta + (\theta-1)\sum_{i=1}^{n} \ln x_i$$

برای یافتن تخمین Maximum-Likelihood کافی‌ست مشتق تابع log-likelihood را گرفته و برابر صفر قرار دهیم:

$$\frac{dLL(\theta)}{d\theta} = \frac{n}{\theta} + \sum_{i=۱}^{n} \ln x_i = ۰$$

$$\hat{\theta}_{MLE} = -\frac{n}{\sum_{i=۱}^{n} \ln x_i}$$

برای نمونه داده‌شده خواهیم داشت:

$$\begin{cases} n = ۱۰ \\ \sum_{i=۱}^{n} \ln x_i = -۱۰/۴۳ \end{cases} \implies \hat{\theta}_{MLE} = -\frac{۱۰}{-۱۰/۴۳} \approx ۰/۹۶$$

سؤال ۵.

توزیع مقابل را در نظر بگیرید.

$$f(x|\mu, \sigma) = \frac{۱}{۲\sigma} e^{-\frac{|x-\mu|}{\sigma}}$$

با استفاده از روش گشتاور، امید ریاضی و واریانس توزیع فوق را بدست آورید.

پاسخ .

توزیع فوق با نام توزیع لاپلاس معروف است و از آن در تحلیل ریسک سرمایه‌گذاری و تحلیل آمارهای اقتصادی استفاده می‌شود.
از تغییر متغیر $u = x - \mu$ استفاده می‌کنیم:

$$\mathrm{E}[X] = m_۱ = \frac{۱}{۲\sigma} \int_{-\infty}^{\infty} u\, e^{-\frac{|u|}{\sigma}}\, du + \mu = \underbrace{\frac{۱}{۲\sigma} \int_{-\infty}^{۰} u\, e^{\frac{u}{\sigma}}\, du}_{\text{set } u=-u} + \frac{۱}{۲\sigma} \int_{۰}^{\infty} u\, e^{-\frac{u}{\sigma}}\, du + \mu$$

$$= -\frac{۱}{۲\sigma} \int_{۰}^{\infty} u\, e^{-\frac{u}{\sigma}}\, du + \frac{۱}{۲\sigma} \int_{۰}^{\infty} u\, e^{-\frac{u}{\sigma}}\, du + \mu = \mu \implies$$ **ممان مرتبه اول- امید ریاضی**

$$\mathrm{E}\left[X^۲\right] = m_۲ = \mathrm{E}\left[(U+\mu)^۲\right] = \mathrm{E}\left[U^۲\right] + ۲\mu\, \mathrm{E}\left[U\right] + \mu^۲$$

$$= \mathrm{E}\left[U^۲\right] + ۲\mu\, \mathrm{E}\left[X - \mu\right] + \mu^۲ = \frac{۱}{۲\sigma} \int_{-\infty}^{\infty} u^۲ e^{-\frac{|u|}{\sigma}}\, du + \mu^۲$$

$$= \underbrace{\frac{۱}{۲\sigma} \int_{-\infty}^{۰} u^۲ e^{\frac{u}{\sigma}}\, du}_{\text{set } u=-u} + \frac{۱}{۲\sigma} \int_{۰}^{\infty} u^۲ e^{-\frac{u}{\sigma}}\, du + \mu^۲$$

$$= \frac{۱}{۲\sigma} \int_{۰}^{\infty} u^۲ e^{-\frac{u}{\sigma}}\, du + \frac{۱}{۲\sigma} \int_{۰}^{\infty} u^۲ e^{-\frac{u}{\sigma}}\, du + \mu^۲$$

$$= \underbrace{\frac{۱}{\sigma} \int_{۰}^{\infty} u^۲ e^{-\frac{u}{\sigma}}\, du}_{\text{set } v=\frac{u}{\sigma}} + \mu^۲ = \sigma^۲ \int_{۰}^{\infty} v^۲ e^{-v}\, dv + \mu^۲ = ۲\sigma^۲ + \mu^۲ \implies$$ **ممان مرتبه دوم**

$$\mathrm{Var}[X] = \mathrm{E}\left[X^۲\right] - (\mathrm{E}[X])^۲ = ۲\sigma^۲$$

سؤال ۶.

یک تحلیل‌گر مسابقات بوکس، قصد دارد تا احتمال برخورد موفق ضربات یک بوکسور را بسنجد. او ابتدا با بررسی مسابقات آن بوکسور، نمونه‌ای ۲۰۰ تایی از ضربات او در مسابقات مختلف را بدست می‌آورد. در این نمونه، ۱۳۰ ضربه از ۲۰۰ ضربه به صورت کامل با حریف برخورد کرده است و انحراف معیار نمونه برابر با $\sqrt{۲}$ می‌باشد.(ضربات بوکسور نسبت به هم مستقل هستند) مربی او ادعا دارد که بوکسور در هر ۲۰۰ ضربه، حداقل ۱۵۰ ضربه را با موفقیت به حریف وارد می‌کند، درستی ادعای وی را بررسی کنید.

پاسخ .

ابتدا فرضیات آزمون فرض را مشخص می‌کنیم:

$H_۰$: ادعای مربی صحیح است و بوکسور بیش از ۱۵۰ ضربه را با موفقیت وارد می‌کند.
H_A: ادعای مربی صحیح نیست و نمونه مشاهده شده نمی‌تواند تصادفی باشد.

حال تخمین نقطه‌ای $\hat{p}$ را بدست می‌آوریم:

$$H_۰ = p = \frac{۱۵۰}{۲۰۰} = ۰/۷۵$$
$$H_A = \hat{p} = \frac{۱۳۰}{۲۰۰} = ۰/۶۵$$

حال شرایط لازم برای استفاده از تقریب نرمال را بررسی می‌کنیم. ضربات بوکسور طبق فرض به نسبت به یکدیگر مستقل هستند.

$$np = ۲۰۰ \times ۰/۶۵ = ۱۳۰ > ۱۰ \quad \checkmark$$
$$n(۱-p) = ۲۰۰ \times ۷۰ > ۱۰ \quad \checkmark$$

حال به محاسبه P-value می‌پردازیم: متغیر تصادفی X را برابر تعداد ضربات موفق قرار می‌دهیم.

$$X \sim N(\mu = \hat{p} = ۰/۶۵, \sigma_{\hat{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{۲}}{\sqrt{۲} \times ۱۰} = ۰/۱$$

$$\textbf{P-value} = P(X > ۰/۷۵) = P(Z > \frac{۰/۷۵ - ۰/۶۵}{۰/۱}) = P(Z > ۱)$$
$$= ۱ - P(Z < ۱) = ۱ - ۰/۸۴ = ۰/۱۶ > ۰/۰۵$$

از آنجایی که P-value بزرگ‌تر از ۰/۰۵ است، با توجه به دیتای داده شده نمی‌توان صحت ادعای مربی را رد کرد.

سؤال ۷.

علی در تابستان گذشته در زمینه بلاکچین مطالعاتی را انجام داده است و هم‌اکنون قصد دارد تا با استفاده از شبکه بیت‌کوین، مقداری پول را به حساب یکی از دوستانش بریزد. او هنگام ثبت تراکنش با هزینه‌ای به نام کارمزد تراکنش روبرو شده است. کارمزد تراکنش یک مقدار نسبی است که هر کاربر بابت ثبت شدن تراکنش خود می‌بایست پرداخت کند. اگر مقدار کارمزد تراکنش بسیار کم باشد، تراکنش ثبت نشده و اگر بیشتر از حد معمول باشد، با وجود ثبت شدن تراکنش کاربر پول اضافه‌ای را از دست می‌دهد و ضرر می‌کند. از آنجایی که کارمزد تراکنش یک مبلغ نسبی است، علی با جمع‌آوری آخرین کارمزدهای پرداخت شده در شبکه‌ی بیت‌کوین، سعی دارد تا هزینه تقریبی کارمزد مورد نیاز را تخمین بزند.

علی ۱۰۰ کارمزد را بررسی کرده و میانگین آن‌ها برابر با ۱٫۰۸ بدست آمد. انحراف معیار تمامی کارمزدهای شبکه طبق آمارها برابر با ۲٫۳ می‌باشد.

الف) بازه اطمینان ۹۵ و ۹۹ درصد را برای میزان کارمزد بهینه‌ای که علی باید بپردازد را بدست آورید .

ب) عرض بازه‌های بدست آمده در قسمت الف را با یکدیگر مقایسه کنید و ارتباط میان میزان دقت و عرض بازه را مشخص کنید. شرح دهید که آیا همواره دقت بیشتر در بازه اطمینان برای ما مناسب است یا خیر.

پاسخ .

الف)

$$\overline{X} - \frac{Z \times \sigma}{\sqrt{۱۰۰}} < \mu < \overline{X} + \frac{Z \times \sigma}{\sqrt{۱۰۰}}$$

بازه اطمینان %۹۵ :

$$\overline{X} - \frac{۱٫۹۶ \times ۲٫۳}{\sqrt{۱۰۰}} < \mu < \overline{X} + \frac{۱٫۹۶ \times ۲٫۳}{\sqrt{۱۰۰}}$$
$$\Longrightarrow ۱٫۰۸ - ۰٫۴۰۵۸ < \mu < ۱٫۰۸ + ۰٫۴۰۵۸$$

بازه اطمینان %۹۹ :

$$\overline{X} - \frac{۲٫۵۷ \times ۲٫۳}{\sqrt{۱۰۰}} < \mu < \overline{X} + \frac{۲٫۵۷ \times ۲٫۳}{\sqrt{۱۰۰}}$$
$$\Longrightarrow ۱٫۰۸ - ۰٫۵۹۱۱ < \mu < ۱٫۰۸ + ۰٫۵۹۱۱$$

ب)

بازه اطمینان ٪۹۵ کوتاه‌تر(به بیان دیگر دقیق‌تر) از بازه اطمینان ٪۹۹ است. در واقع، بازه‌ای که سطح اطمینان بالاتری دارد، خطای برآورد بیشتری دارد. با توجه به این نکته، استفاده از دقت بسیار بالا می‌تواند منجر به خطای برآورد زیادی شود.