آمار و احتمال مهندسی

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

تمرین کامپیوتری سوم ــ قضیه حد مرکزی و تخمین پارامتر

طراح: متین بذرافشان
سوپروایزر: مهدی جمال‌خواه

تاریخ تحویل: ۹ دی ۱۴۰۳

نکات

- هدف تمرین درک عمیق‌تر مفاهیم درس می‌باشد، در نتیجه زمان کافی برای تحلیل کردن نتایج اختصاص دهید.
- در ابتدای همه‌ی سوالات seed را سه رقم آخر شماره دانشجویی‌تان قرار دهید.
- پاسخ تمرین باید به صورت یک فایل زیپ با نام CA3_[Last-Name]_[Student-Id].zip بارگذاری شود. پاسخ سوالات تئوری و تحلیل نتایج‌ها باید به صورت Markdown در فایل Notebook یا در یک فایل pdf که شامل نمودارها و نتایج نیز هست، باشد.

بیشتر بدانیم: اثبات قضیه حد مرکزی

در این بخش می‌خواهیم قضیه حد مرکزی را بوسیله تابع مولد گشتاور اثبات کنیم. تابع مولد گشتاور همان طور که پیش‌تر هم دیده‌اید به صورت زیر تعریف می‌شود:

$$\Phi_Y(t) = \mathbb{E}[e^{ty}] = \int_{-\infty}^{\infty} e^{ty} f_Y(y)dy \tag{1}$$

بخش e^{ty} را با بسط تیلور آن جایگزین می‌کنیم:

$$e^{ty} = \sum_{n=0}^{\infty} \frac{(ty)^n}{n!} = \sum_{n=0}^{\infty} \frac{t^n y^n}{n!} \tag{2}$$

$$\Phi_Y(t) = \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{t^n y^n}{n!} f_Y(y)dy = \sum_{n=0}^{\infty} \frac{t^n}{n!} \int_{-\infty}^{\infty} y^n f_Y(y)dy = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}[y^n] \tag{3}$$

رابطه ۳ در واقع همان خاصیت معروف تابع مولد گشتاور هست، یعنی مشتق‌های آن بیانگر گشتاورهای متغیر تصادفی می‌باشد.

حال با این مقدمه، قضیه را اثبات می‌کنیم. رابطه قضیه حد مرکزی به صورت زیر می‌باشد:

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i \quad ; \quad X_i \overset{iid}{\sim} (\mu, \sigma^2) \tag{4}$$

$$\mathbb{E}[\bar{X}_N] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[X_i] = \frac{1}{N} N\mu = \mu$$

$$\mathrm{Var}(\bar{X}_N) = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}(X_i) = \frac{1}{N^2} N\mu = \frac{\sigma^2}{N}$$

متغیر تصادفی $\bar{X}_N$ را استاندارد می‌کنیم و با Z_N نمایش می‌دهیم:

$$Z_N = \frac{N\bar{X}_N - N\mu}{\sigma\sqrt{N}} = \frac{\sum_{i=1}^{N} X_i - N\mu}{\sqrt{N}\sigma} = \frac{\sum_{i=1}^{N}(X_i - \mu)}{\sqrt{N}}$$
$$= \sum_{i=1}^{N}\left(\frac{X_i - \mu}{\sqrt{N}}\right) = \sum_{i=1}^{N}\frac{Y_i}{\sqrt{N}} \quad ; \quad Y_i = \frac{X_i - \mu}{\sigma} \tag{5}$$

همان طور که از رابطه ۵ مشخص است Y_i نیز استاندارد شده‌ی X_i می‌باشد. بنابرین امید ریاضی و واریانس آن به ترتیب برابر صفر و یک می‌باشد، در نتیجه می‌توانیم تابع مولد گشتاور آن را از رابطه ۳ به صورت زیر بازنویسی کنیم:

$$\Phi_Y(t) = 1 + \frac{t^2}{2} + \sum_{i=3}^{N}\frac{t^n}{n!}\mathbb{E}[y^n] \tag{6}$$

حال با توجه به رابطه Z_N و Y_i در معادله ۵ و خواص تابع مولد گشتاور، می‌توانیم تابع مولد گشتاور Z_N را به صورت زیر بنویسیم:

$$\Phi_{Z_N}(t) = [\Phi_Y(\frac{t}{\sqrt{N}})]^N = \left[1 + \frac{t^2}{2N} + \sum_{n=3}^{N}\frac{t^n}{n!N^{n/2}}\mathbb{E}[y^n]\right]^N \tag{7}$$

از هر دو طرف لگاریتم می‌گیریم تا توان را حذف کنیم، چون کار با توان دشوار است:

$$\ln\Phi_{Z_N}(t) = N\ln\left(1 + \frac{t^2}{2N} + \sum_{n=3}^{N}\frac{t^n}{n!N^{n/2}}\mathbb{E}[y^n]\right) \tag{8}$$

بار دیگر از بسط تیلور کمک می‌گیریم تا لگاریتم را نیز حذف کنیم:

$$\ln(1+x) = \sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n} \tag{9}$$

$$\ln\Phi_{Z_N}(t) = N\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\left[\frac{t^2}{2N} + \sum_{k=3}^{N}\frac{t^k}{k!N^{k/2}}\mathbb{E}[y^k]\right]^n \tag{10}$$

حال طبق قضیه حد مرکزی باید N را به بی‌نهایت میل بدهیم. که در این صورت عبارت هایی که پس از ساده‌سازی در مخرج آن‌ها N ظاهر می‌شود، صفر خواهند شد. اگر به عبارت ۱۰ دقت کنید تمام عبارت ها به ازای $n \geq 2$ در مخرجشان N ظاهر می‌شود و به ازای $n = 1$ نیز تنها عبارت اول یعنی $\frac{t^2}{2N}$ هست که بعد سازی در آن اصلا N وجود نخواهد داشت و مابقی همگی در مخرجشان N ظاهر می‌شود. در نتیجه خواهیم داشت:

$$\lim_{n\to\infty}\ln\Phi_{Z_N}(t) = \frac{t^2}{2} \tag{11}$$

$$\Phi_{Z_N} = exp(\frac{t^2}{2}) \tag{12}$$

عبارت ۱۲ دقیقا تابع مولد گشتاور متغیر تصادفی نرمال استاندارد می‌باشد و بین متغیر تصادفی و تابع مولد گشتاور رابطه یک به یک برقرار است(اثبات آن خارج از حوصله این بحث است) در نتیجه Z_N متغیر تصادفی نرمال استاندارد است:

$$Z_N \sim N(0,1) \Rightarrow \bar{X}_N \sim N(\mu, \frac{\sigma^2}{N}) \tag{13}$$

خب، این همه ریاضیات رو اینجا آوردیم برای چی؟ دقیقا! برای اینکه با استفاده از آن و شهودی که نسبت به احتمال دارید به این فکر کنید که چرا چنین اتفاقی در دنیا خارج اتفاق می‌افتد؛ یعنی چرا مجموع یک سری متغیر تصادفی مستقل یک توزیع نرمال را تشکیل می‌دهند!؟

۱. نمونه برداری و حد مرکزی ۳۵ نمره

در این سوال می‌خواهیم قضیه حد مرکزی و ویژگی‌های آن را به صورت تجربی ببینیم و با دانسته‌های تئوری آن‌ها را تحلیل کنیم. به این منظور ابتدا سه توزیع زیر را درنظر بگیرید:

- توزیع پوآسون با نرخ ۱۰
- توزیع نمایی با نرخ $\frac{۱}{۲}$
- توزیع هندسی با احتمال موفقیت $\frac{۱}{۵}$

۱. برای هر یک از توزیع‌ها میانگین و واریانس آن‌ها را گزارش کنید.

۲. برای هر توزیع، به ازای سه اندازه نمونه ۳۰، ۳۰۰ و ۳۰۰۰ هرکدام ۱۰۰۰ بار نمونه برداری کنید و سپس هیستوگرام میانگین این نمونه‌ها را رسم کنید. (دقت کنید که طول گام محور افقی همه‌ی نمودارها یکسان باشند تا قابل مقایسه شوند.)

۳. برای هر توزیع بالا، میانگین و خطای استاندارد را گزارش کنید. با افزایش اندازه نمونه، چه تغییری در این پارامترها مشاهده می‌کنید؟

۴. به ازای هر توزیع به‌دست‌آمده، یک توزیع نرمال با میانگین و انحراف معیار همان توزیع رسم کنید و به نمودارهای بالا اضافه کنید. چه مشاهده می‌کنید؟ به نظر شما با افزایش سایز نمونه، توزیع های به‌دست‌آمده به چه تابعی میل خواهند کرد؟ پارامترهای این تابع وابسته به چه پارامترهای توزیع‌های اولیه می‌باشد؟

۵. حال فرض کنید که به احتمال $\frac{۱}{۳}$ یک توزیع را از سه توزیع بیان شده در ابتدای سوال انتخاب می‌کنیم و سپس از آن توزیع به اندازه سایز نمونه، نمونه برمی‌داریم. این کار را ۱۰۰۰ بار و به ازای سه تعداد نمونه اشاره شده انجام دهید و سپس توزیع میانگین نمونه‌ها را رسم کنید، چه مشاهده می‌کنید. میانگین و خطای استاندارد این میانیگین‌ها را گزارش کنید.

۶. آیا می‌توانید میانگین و انحراف معیار نمونه‌ها را پیش از نمونه برداری و با توجه به توزیع‌های اولیه بدست‌آورد؟ اگر بله فرمول آن را بنویسید و این کار را برای بخش قبل انجام دهید.

۷. به توزیع‌های بدست آمده در بخش ۵، یک توزیع نرمال با میانگین و واریانس آن توزیع اضافه کنید. با توجه به نتایج به‌نظر شما توزیع‌های حاصل از بخش ۴، مشترکا نرمال می‌باشند؟ آیا این قضیه برای همه‌ی توزیع های حاصل از قضیه حد مرکزی برقرار می‌باشد؟ تحلیل کنید.

۲. کاربردهای خطای میانگین مربعات ۳۰ نمره

در درس خواندید که با مشاهده متغیر تصادفی X می‌توان متغیر تصادفی Y را پیش‌بینی کرد. به طوری که اگر g(X) تابعی باشد که برای پیش‌بینی Y استفاده می‌شود، این تابع بهترین تخمینگر Y است اگر $E[(Y-g(X))^۲]$ کمینه باشد.

روش‌های مشابه این روش برای تخمین مجهول‌های یک مساله وجود دارد، که در این سوال با دو مورد دیگر آشنا می‌شوید:

- فرض کنید تابع $y=f(x)=ax+b$ وجود دارد، درحالی که ما مقادیر پارامترهای a و b را نمی‌دانیم. با مشاهده جفت نقطه‌های (x,y) می توان به تخمین خوبی از این تابع رسید. به این فرآیند رگرسیون خطی می‌گویند. برای تخمین این پارامترها، تابع خطای میانگین مربعات را به صورت زیر تعریف می‌کنیم:

$$\text{MSE}_{\text{regression}} = \frac{۱}{n}\sum_{i=۱}^{n}(y_i - f(x_i))^۲$$

به طوری که x_i و y_i مشاهده های ما می‌باشند.

- فرض کنید که می‌خواهید یک نقطه را به عنوان مرکز تعدادی نقطه در صفحه مختصات معرفی کنید، یک راه آن استفاده از خطای میانگین مربعات هست، که برای این مساله به صورت زیر بازنویسی می‌شود:

$$\text{MSE}_{\text{center}} = \frac{۱}{n}\sum(x_i - c_x)^۲ + (y_i - c_y)^۲$$

به طوری که x_i ها نقاط ما و c مرکز آن‌هاست، که می‌خواهیم بدست آوریم.

می‌دانیم که این توابع، توابعی درجه دو می‌باشند. مقدار کمینه آن‌ها در نقطه‌ای اتفاق می‌افتد که مشتق آن‌ها برابر صفر شود. برای بدست آوردن مقادیر مطلوب دو مساله عنوان شده، نوت‌بوک MSE.ipynb را دنبال کنید و بخش‌های مشخص شده را تکمیل کنید.

۳. کاربرد حد مرکزی در توزیع برنولی — ۳۵ نمره

یک آزمایش برنولی را با احتمال موفقیت p در نظر بگیرید. اگر X_i را خروجی i_امین آزمایش و $S_n = X_۱ + X_۲ + ... + X_n$ در نظر بگیریم، S_n یک توزیع دو جمله‌ای با پارامترهای n و p خواهد داشت.

۱. یک توزیع دو جمله‌ای با $n = ۲۷۰$ و $p = ۰/۳$ را در نظر بگیرید. نمودار این توزیع را رسم کنید. (دقت کنید در محور افقی، همه مقادیر ممکن را در نظر بگیرید.)

۲. برای داشتن یک نمودار معیار، این توزیع را استاندارد کنید. (این کار را با کم کردن مقادیر x از میانگین و تقسیم آنها بر انحراف معیار انجام دهید.)

۳. حال نمودار توزیع نرمالی با میانگین ۰ و انحراف معیار ۱ ایجاد کنید و به نمودار بالا اضافه کنید. چه مشاهده می‌کنید؟

۴. مجموع طول میله‌های نمودار چند جمله‌ای را محاسبه کنید. آیا این مقدار برابر ۱ می‌باشد؟ با این وجود به نظر شما چرا چنین چیزی در نمودار بالا مشاهده می‌شود؟

۵. برای حل این مشکل، باید تابع توزیع نرمال در یک ضریبی ضرب شود. برای بدست آوردن این ضریب از روش ریمان کمک بگیرید:

$$\int_a^b f(x)dx \approx \frac{b-a}{N}\sum_{i=۰}^{N} f(x_i) \quad ; \quad x_i = a + i\frac{b-a}{N}$$

ضریب به‌دست‌آمده را به همراه انحراف معیار توزیع بخش ۱ گزارش کنید. چه مشاهده می‌کنید؟ توضیح دهید.

توزیع دو جمله‌ای و نرمال را بعد از اعمال این ضریب در یک شکل رسم کنید. آیا مشکل برطرف شده است؟

با توجه به نتایج بالا می‌توان گفت که می‌شود از توزیع نرمال به عنوان یک تقریب برای توزیع دوجمله‌ای استفاده کرد. به بیان ریاضی داریم:

$$Binomial(n,p,k) \approx \frac{۱}{\sqrt{np(۱-p))}}\phi(\frac{k-np}{\sqrt{np(۱-p))}})$$

- $\phi(x)$ همان توزیع نرمال استاندارد می‌باشد.

حال در ادامه از این نتیجه بهره خواهیم برد.

۶. احتمال دقیقا ۵۵ بار رو آمدن یک سکه سالم را یک بار به کمک توزیع دوجمله‌ای و یک بار به کمک توزیع نرمال بدست آورید.

با تعمیم قضیه بالا داریم:

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) = \int_a^b \phi(x)\,dx$$

۷. حال احتمال تعداد ۴۰ الی ۶۰ بار رو آمدن سکه سالم را به کمک قضیه بالا بدست آورید.