آمار و احتمال مهندسی

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

تمرین کامپیوتری اول ــ توزیع‌های آماری، توابع متغیر تصادفی

طراح: سروش اصفهانیان
سوپروایزر: مهدی جمال‌خواه

تاریخ تحویل: ۲۹ آبان ۱۴۰۳

نکات

- اگر پاسخ این تمرین با زبان برنامه نویسی R نوشته شود، ۱۰۰ درصد **نمره امتیازی** به آن تعلق می‌گیرد.
- 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 CA1_[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: پارادوکس سنت‌پترزبورگ

مساله سنت‌پترزبوگ به شرح زیر است:

- در ابتدا شما باید مبلغ اولیه‌ای برای شروع بازی پرداخت کنید.
- در هر مرحله از شما درخواست می‌شود تا یک سکه سالم را پرتاب کنید.
- اگر در پرتاب اول شیر آمد، شما ۲ دلار برنده می‌شوید.
- اگر در پرتاب دوم هم شیر آمد، ۴ دلار برنده می‌شوید.
- به این ترتیب با هر بار شیر آمدن سکه میزان برد ۲ برابر می‌شود.
- اولین باری که سکه خط بیاید بازی تمام می‌شود.
- شما حداکثر چه مبلغی را برای شروع بازی پرداخت کنید تا بازی برای شما منصفانه باشد؟

همان طور که در درس دیدید، اگر امید ریاضی برد از مبلغ پرداختی بیش‌تر باشد، بازی منصفانه است.
احتمال شیر آمدن n پرتاب اول $\frac{۱}{۲^n}$ است، پس با احتمال $\frac{۱}{۲^n}$، ۲^n دلار برنده می‌شوید:

$$E[X] = \sum_{i=۱}^{\infty} ۲^i(\frac{۱}{۲^i}) = ۱ + ۱ + ۱ + \cdots = \infty$$

این به این معنی است که فرد بازی کننده باید هر مقدار متناهی هزینه اولیه برای ورود به بازی را بپردازد. اما در واقعیت تقریبا هیچ کس حاضر نیست بیش تر از ۲۵ دلار برای ورود به چنین بازی‌ای بپردازد! پارادوکس سنت‌پترزبورگ در واقع همین اختلاف بین امید ریاضی

نامتناهی و مقدار پولی است که افراد حاضر به پرداخت برای انجام بازی می‌شوند. به نظر وقتی امید ریاضی به بی‌نهایت میل می‌کند، استفاده از آن به عنوان معیار منصفانه بودن منطقی نیست. مثلا بعضی می‌گویند اصلا بی‌نهایت پول در واقعیت وجود ندارد که بخواهیم به عنوان امیدریاضی برد در نظر بگیریم!؟

۱. توزیع فوق هندسی و دوجمله‌ای ۲۵ نمره

در بسیاری از انتخابات در سراسر دنیا، پس از اتمام رای‌گیری، فرآیندهای مختلفی برای اطمینان از صحت برگزاری انتخابات انجام می‌شود که در آن‌ها، از بین تمامی حوزه‌های موجود اخذ رای، تعدادی حوزه انتخاب می‌شوند و آرای آن‌ها مورد بررسی قرار می‌گیرند. حال، فرض کنید در انتخابات مورد نظر، در کل N=۱۰۰ حوزهٔ رای‌گیری وجود دارد که در k=۲۰ عدد از آن‌ها تقلب رخ داده است و همچنین، نهاد حسابرسی، m=۴۰ ناحیه را مورد بررسی قرار می‌دهد.

۱ــ توزیع فوق هندسی مربوط به فرآیند حسابرسی را با n=۱۵۰ نمونه شبیه‌سازی کنید و نمودار توزیع آن را رسم کنید.

۲ــ تابعی پیاده‌سازی کنید که به ازای تعداد نمونهٔ n=۱۰۰ تا n=۱۰۰۰۰، با افزایش‌های ۵۰ واحدی، مقادیر تئوری (با استفاده از روابط ریاضی) و عملیِ (با استفاده از توزیع شبیه‌سازی‌شده) میانگین و واریانس تعداد تقلبی که توسط نهاد حسابرسی یافت می‌شود را محاسبه کند و بازگرداند. (پارامترهای توزیع فوق هندسی همانند مقادیر گفته شده در صورت سوال است.)

۳ــ مقادیر عملی میانگین و واریانس به دست‌آمده برای مقادیر مختلف n را به همراه مقادیر تئوری در یک نمودار رسم کنید و آن‌ها را مقایسه کنید.

۴ــ به ازای تعداد نمونه ثابت $n = ۱۰۰۰$، به ازای مقادیر $m = [۴۰, ۶۰, ۸۰, ۱۰۰]$، توزیع‌های فوق هندسی مربوطه را شبیه‌سازی کنید و نمودار آن‌ها را در یک شکل رسم کنید. نمودارهای به دست‌آمده را تحلیل کنید.

در این قسمت، قصد داریم رابطه بین بین توزیع دوجمله‌ای و فوق هندسی را بررسی کنیم. توزیع فوق هندسی معمولا برای شبیه‌سازی نمونه‌گیری بدون جایگذاری استفاده می‌شود، در صورتی که در توزیع دوجمله‌ای، نمونه‌گیری با جایگذاری مناسب اتفاق می‌افتد. حال، هنگامی که مقدار جمعیت N در توزیع فوق هندسی بسیار بزرگ شود، اثر عدم جایگذاری بعد از نمونه گیری کاهش پیدا می‌کند و با میل کردن N به سمت بی‌نهایت، توزیع فوق هندسی به توزیع دوجمله‌ای میل می‌کند. اثبات تئوری این موضوع در این لینک قابل مشاهده است. در ادامه، این موضوع را به صورت عملی مشاهده خواهیم کرد.

۵ــ با استفاده از توابع پایه در پایتون، توابعی مجزا پیاده‌سازی کنید که پارامترهای هر کدام از توزیع‌های فوق هندسی و دوجمله‌ای را به همراه آرایه‌ای از مقادیر متغیر تصادفی، به عنوان ورودی دریافت کند و مقدار pmf توزیع مربوطه را برای مقادیر آن بازگرداند. (مقادیر متغیر تصادفی را از ۰ تا ۲۰ در نظر بگیرید.)

راهنمایی: می‌توانید از توابع **factorial** و **comb** از کتابخانه **math** برای پیاده‌سازی توابع کمک بگیرید.

۶ــ با استفاده از توابع پیاده‌سازی شده، منحنی pmf توزیع فوق هندسی و دوجمله‌ای را به ازای مقادیر $N = [۱۰۰, ۵۰۰, ۱۰۰۰, ۱۰۰۰۰]$ رسم کنید. (N برابر با تعداد حوزه‌های رای‌گیری است. در این بخش، برای هر مقدار N یک شکل جدا در نظر بگیرید و در هر شکل، توزیع دوجمله‌ای و فوق هندسی مربوطه را با هم رسم کنید.)

۷ــ نمودارهای به‌دست‌آمده توزیع فوق هندسی و دوجمله‌ای را برای مقادیر مختلف N (جمعیت) تحلیل کنید و نتیجهٔ به‌دست‌آمده را گزارش و با نتیجه تئوری مقایسه کنید.

۲. توزیع دوجمله‌ای، توزیع نرمال و خطای تصحیح ۴۰ نمره

شرکت‌های تولیدکنندهٔ مطرح حوزهٔ تکنولوژی، قبل از معرفی تولیدات جدید خود، معمولا برای تبلیغ هر چه بیشتر محصول، رویدادی با حضور افراد تأثیرگذار و مطرح دنیای تکنولوژی برگزار می‌کنند. حال، فرض کنید شما در گروه آنالیز دادهٔ یکی از این شرکت‌ها فعالیت می‌کنید. برای مراسم معرفی محصول جدید، گروه ارتباطات شرکت، تعداد ۱۰۰۰ ایمیل را به مبلغان حوزهٔ تکنولوژی در سراسر دنیا ارسال می‌کند. بر اساس دادهٔ پیشین، مخاطبان این جنس ایمیل‌ها، به حدود ۴۵ درصد از ایمیل‌ها پاسخ مثبت می‌دهند و در مراسم معرفی حضور پیدا می‌کنند. شرکتی که شما در آن مشغول به کار هستید، دچار بودجه‌ای محدود است و برای برگزاری مراسم، قادر است حداکثر سالنی با ظرفیت ۴۳۰ نفر تهیه کند. حال، از شما خواسته شده‌است احتمال این‌که حداکثر ۴۳۰ نفر از این ۱۰۰۰ نفر، به دعوت شرکت جواب مثبت بدهند را محاسبه کنید تا شرکت بر اساس آن تصمیم‌گیری کند.

۱ــ با استفاده از تابع pmf توزیع دوجمله‌ای که در سوال ۱ ، قسمت ۵، پیاده‌سازی کردید، تابع محاسبه CDF این توزیع را با استفاده از توابع پایه در پایتون، پیاده‌سازی کنید و احتمال خواسته‌شده را گزارش کنید.

به علت پیچیدگی محاسبات و محدود بودن منابع سخت‌افزاری، شرکت از شما می‌خواهد به جای استفاده از توزیع دوجمله‌ای، از توزیع نرمال برای تقریب مقدار احتمال خواسته‌شده استفاده کنید.

۲ــ با استفاده از CDF توزیع نرمال و بدون اعمال تصحیح پیوستگی، احتمال خواسته شده را تقریب بزنید و آن را به همراه مقدار خطا نسبت به احتمال حاصل‌شده در قسمت قبل گزارش کنید.

۳ــ بار دیگر با استفاده از CDF توزیع نرمال و با اعمال تصحیح پیوستگی، احتمال خوسته‌شده را تقریب بزنید و آن را به همراه مقدار خطا گزارش کنید.

۴ــ در این قسمت، ابتدا مقدار CDF برای مقادیر $X = ۰, ۱, ۲, ۳, ..., ۱۰۰۰$ را با هر سه روش بالا به دست‌آورید و مقادیر خطا را یکبار بین مقادیرِ CDF توزیع دوجمله‌ای و توزیع نرمال بدون تصحیح پیوستگی و یکبار بین مقادیر CDF توزیع دوجمله‌ای و توزیع نرمال با تصحیح پیوستگی به دست آورید. دو دسته خطای به‌دست‌آمده را بر حسب مقدار متغیر تصادفی در یک شکل رسم کنید. دو نمودار حاصل را تحلیل کنید. (X همان متغیر تصادفی است.)

۵ــ با استفاده از مقادیر CDF محاسبه شده، نمودار CDF بر حسب مقدار متغیر تصادفی را برای مقادیر بازهٔ $X \in [۴۴۰, ۴۶۰]$ برای هر سه روش مذکور، در یک شکل رسم کنید و نمودارها را مقایسه کنید.

۶ــ برای مقادیر $X = ۰, ۱, ۲, ۳, ..., ۱۰۰۰$ بار دیگر مقدار CDF را با دو روشِ توزیع دوجمله‌ای و توزیع نرمال با تصحیح پیوستگی محاسبه کنید و زمان سپری‌شده برای محاسبهٔ هر یک از مقادیر CDF را برای هر دو روش ذخیره کنید و بر حسب مقدار متغیر تصادفی در یک شکل رسم کنید. دو نمودار به دست‌آمده را تحلیل کنید.

راهنمایی: برای محاسبه زمان سپری شدهٔ یک عملیات با دقت بالا، از فرمت زیر استفاده کنید:

```python
import time
def get_elapsed_time(*args):
    start = time.perf_counter()
    # your function or operation
    end = time.perf_counter()
    return end - start
```

با توجه به محدود بودن منابع سخت‌افزای شرکت، از شما خواسته شده‌است یک استراتژی تعیین کنید که مشخص کند برای محاسبهٔ CDF چه مقادیری از X در بازهٔ $X \in [۴۰۰, ۶۰۰]$ از توزیع دوجمله‌ای و برای چه مقادیری، از توزیع نرمال با تصحیح پیوستگی استفاده کنیم. بدین منظور، فرض کنید هزینه پرداختی برای هر ثانیه افزایش زمان محاسبات برابر با ۱۰۰ واحد و هزینه پرداختی برای هر واحد خطا در محاسبه CDF برابر با $۱۰^۶$ واحد باشد.

۷ــ ابتدا برای استفاده از مقادیر به دست‌آمده در قسمت‌های ۴ و ۶، مقادیر افزایش زمان محاسبات هنگام استفاده از توزیع دوجمله‌ای نسبت به توزیع نرمال با تصحیح پیوستگی و خطای محاسبه CDF هنگام استفاده از توزیع نرمال با تصحیح پیوستگی را برای مقادیر بازهٔ $X \in [۴۰۰, ۶۰۰]$ به دست آورید. سپس تابعی پیاده‌سازی کنید که با استفاده از هزینه‌های فرض شده برای هر یک از دو مورد، اولین مقدار X را بازگرداند (با شروع از $X = ۴۰۰$ ودر بازهٔ $X = [۴۰۰, ۶۰۰]$) که محاسبهٔ CDF آن با استفاده از توزیع دوجمله‌ای مقرون به صرفه نیست. نقطه به دست آمده را گزارش کنید.

راهنمایی: محاسبه CDF با استفاده از توزیع دوجمله‌ای هنگامی مقرون به صرفه نیست که هزینه افزایش زمان محاسبات، از هزینه خطای محاسبهٔ CDF توسط توزیع نرمال با تصحیح پیوستگی، بیشتر شود.

۸ــ نمودار هزینه نهایی (اختلاف هزینه افزایش محاسبات و هزینه خطای محاسبات) بر حسب مقادیر $X \in [۴۰۰, ۶۰۰]$ را رسم کنید. با توجه به نمودار و نقطهٔ به دست آمده، یک استراتژی ساده برای محاسبه CDF برای مقادیر $X \in [۴۰۰, ۶۰۰]$ تعیین کنید و مشخص کنید آیا محاسبه CDF برای نقطه $X = ۴۳۰$ توسط توزیع دوجمله‌ای که در قسمت اول انجام شد، مقرون به صرفه بوده‌است یا خیر.

## ۳. بی‌حافظگی توزیع نمایی		۱۵ نمره

فرض کنید شما کارمند یک فروشگاه هستید که به طور متوسط، هر یک ربع، یک مشتری وارد آن می‌شود و همچنین، زمان بین ورود هر مشتری به فروشگاه، از توزیع نمایی پیروی می‌کند. یکی از همکاران شما ادعا می‌کند که با توجه به اینکه به طور میانگین، هر یک ربع، یک مشتری وارد فروشگاه می‌شود، اگر تا ۱۲ دقیقه مشتری جدیدی وارد فروشگاه نشده باشد، آنگاه به احتمال زیاد، به زودی یک مشتری جدید وارد فروشگاه می‌شود. در این سوال، قصد داریم شرایط مذکور را به صورت عملی پیاده‌سازی کنیم و در مورد صحت ادعای بیان‌شده نتیجه‌گیری کنیم. برای این‌کار، بازهٔ زمانی بین ورود مشتریان را شبیه‌سازی خواهیم کرد. فرض کنید ساعت کاری روزانهٔ فروشگاه ۸ ساعت است و بازه زمانی بین ورود مشتریان در یک روز، از یک توزیع نمایی پیروی می‌کند. (ورود هر مشتری به فروشگاه را یک آزمایش تصادفی یکسان در نظر بگیرید.)

۱ــ تابعی پیاده‌سازی کنید که پارامترهای یک توزیع نمایی را به همراه تعداد نمونه به عنوان ورودی دریافت کند و توزیع نمایی متناظر را بازگرداند.

۲ــ تابعی پیاده‌سازی کنید که مقادیر توزیع نمایی شبیه‌سازی‌شده را دریافت کند و در توزیع، برای بازه‌های زمانی بین ورود مشتریان که بیش از ۱۲ دقیقه است، مقدار زمان بعد از ۱۲ دقیقه تا ورود مشتری جدید را حساب کند و در یک آرایه واحد ذخیره کند.

- دقت کنید که در پیاده‌سازی تابع، نباید بازه‌هایی که زمان ورود مشتری از ۸ ساعت کاری گذشته است را در نظر بگیرید.

۳ــ با استفاده از توابع پیاده‌سازی‌شده، میانگین و هیستوگرام زمان ورود مشتری بعد از ۱۲ دقیقه را در کنار میانگین و هیستوگرام توزیع نمایی با پارامترهای صورت سوال و تعداد $۱۰۰ \times m$ نمونه به ازای $m = ۱۰, ۱۰۰, ۱۰۰۰$ رسم کنید و توزیع‌های به دست‌آمده را مقایسه کنید. (برای هر مقدار m، دو شکل در نظر بگیرید؛ یک شکل برای توزیع نمایی و یک شکل برای توزیع زمان‌های ورود مشتری بعد از ۱۲ دقیقه و دو شکل را در کنار هم به تصویر بکشید.)

۴ــ با توجه به نمودارهای حاصل و با توجه به خاصیت بی‌حافظگی توزیع نمایی، در مورد رد فرض فرد همکار در صورت سوال توضیح دهید و نتیجه‌گیری کنید.

۵ــ احتمال آنکه وقتی که بازه زمانی ورود بین دو مشتری از ۱۲ دقیقه بیشتر شده‌است، مشتری جدید ۱۵ دقیقه بعد از مشتری قبلی وارد فروشگاه شود را یک‌بار به صورت عملی و با استفاده از توزیع زمان ورود مشتریان بعد از ۱۲ دقیقه برای $m = ۱۰۰۰$ و یک‌بار به صورت تئوری، با استفاده از خاصیت بی‌حافظگی توزیع نمایی به‌ست آورید و مقادیر حاصل را مقایسه کنید.

## ۴. توابع متغیرهای تصادفی		۲۰ نمره

در این قسمت، با یک نمونه از کاربردهای توابع متغیرهای تصادفی یعنی تبدیل توزیع‌های آماری به یک دیگر آشنا خواهیم شد و دو نمونه از این تبدیل‌ها را به صورت عملی پیاده‌سازی خواهیم کرد.

تبدیل لگاریتمی یکی از مهم‌ترین و ساده‌ترین تبدیل‌ها به شمار می‌رود که در موارد مختلف مانند کاهش چولگی یک توزیع آماری کاربرد دارد. همچنین از این تبدیل، می‌توان برای تبدیل یک توزیع یکنواخت به توزیع نمایی استفاده کرد. برای مثال، اگر X یک متغیر تصادفی با توزیع یکنواخت $X \sim U(۰, ۱)$ باشد، و تابع Y را به صورت $Y = -۲ln(X)$ در نظر بگیریم، آنگاه برای توزیع Y خواهیم داشت:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|,\ \ f_X(x) = ۱ \text{ for } x \in [۰, ۱],\ \ \left|\frac{dx}{dy}\right| = \frac{۱}{۲}e^{-\frac{y}{۲}} \implies f_Y(y) = \frac{۱}{۲}e^{-\frac{y}{۲}} \implies Y \sim Exp(\lambda = \frac{۱}{۲})$$

بنابراین Y از یک توزیع نمایی با پارامتر $\lambda = \frac{۱}{۲}$ پیروی می‌کند. در ادامه، این موضوع را به صورت عملی اثبات می‌کنیم.

۱ــ توزیع یکنواخت مربوط به متغیر تصادفی $X \sim U(۰, ۱)$ را با $۱۰^۶$ نمونه شبیه‌سازی کنید.

۲ــ با استفاده از توابع موجود، متغیر $Y = -۲ln(X)$ را تشکیل دهید.

۳ــ تابع چگالی متغیر نمایی متناظر با Y را تشکیل دهید و در یک شکل، در کنار توزیع عملی متغیر Y ترسیم کنید و نتیجه را گزارش کنید.

همانطور که مشاهده شد، از تبدیل لگاریتمی می‌توان برای تبدیل توزیع یکنواخت به نمایی استفاده کرد. همچنین می‌توان از تبدیل‌های مختلفی برای تبدیل توزیع یکنواخت به توزیع نرمال نیز بهره برد. یکی از معروف‌ترینِ این تبدیل‌ها، **تبدیل باکس‌ــمولر** (Box–Muller transform) است. اگر متغیرهای تصادفی $U_۱$ و $U_۲$، دو متغیر تصادفی با توزیع یکنواخت $U(۰,۱)$ باشند، آن‌گاه، تبدیل باکس‌ــ‌مولر به صورت زیر خواهد بود:

$$Z_۱ = \sqrt{-۲\ln U_۱} \cdot \cos(۲\pi U_۲)$$

$$Z_۲ = \sqrt{-۲\ln U_۱} \cdot \sin(۲\pi U_۲)$$

که دو متغیر تصادفی حاصل‌شدهٔ $Z_۱$ و $Z_۲$، دو متغیر تصادفی **مستقل** با **توزیع نرمال استاندارد** هستند. در ادامه، این تبدیل را به صورت عملی بررسی می‌کنیم.

۴ــ توزیع‌های یکنواخت $U_۱ \sim U(۰,۱)$ و $U_۲ \sim U(۰,۱)$ را با $۱۰^۶$ نمونه شبیه‌سازی کنید.

۵ــ توزیع‌های $Z_۱$ و $Z_۲$ را با استفاده از روابط بیان‌شده تشکیل دهید.

راهنمایی: ابتدا دو متغیر $R = \sqrt{-۲ln(U_۱)}$ و $\theta = ۲\pi U_۲$ را تشکیل دهید و با استفاده از آن‌ها، $Z_۱$ و $Z_۲$ را بسازید.

۶ــ تابع چگالی متغیر نرمالِ استاندارد متناظر با $Z_۱$ و $Z_۲$ را تشکیل دهید و در دو شکل، در کنار توزیع عملی $Z_۱$ و $Z_۲$، رسم کنید و نتیجه را گزارش کنید.