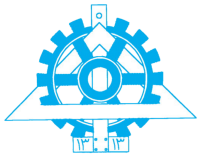

دانشگاه تهران، دانشکده مهندسی برق و کامپیوتر
آمار و احتمال مهندسی

پروژه ششم - قضیه حد مرکزی و برآورد پارامتر

طراح: یاسمن عموجعفری، پارسا بوکانی

شرح مجموعه پروژه‌ها

این تمرین شامل سه پروژه‌ی محاسباتی است که با هدف درک تجربی مفاهیم آماری طراحی شده‌اند:

- **پروژه‌ی اول - قضیه حد مرکزی:** مشاهده‌ی رفتاری نرمال‌مانندِ میانگین نمونه‌ها با افزایش اندازه‌ی نمونه، هم روی داده‌ی واقعی و هم روی توزیع‌های مشهور.
- **پروژه‌ی دوم - قانون بنفورد:** بررسی توزیع رقم اول داده‌ها و اعتبارسنجی قانون بنفورد با استفاده از داده‌های اقتصادی واقعی.
- **پروژه‌ی سوم - برآورد پارامتر:** تخمین پارامتر یک مدل پواسون با روش حداکثر درست‌نمایی و بررسی پایداری تخمین با افزایش تعداد مشاهدات.

۱. مشاهده‌ی قضیه حد مرکزی (CLT) در داده‌های واقعی و توزیع‌های مشهور

در این پروژه، هدف مشاهده‌ی تجربی قضیه حد مرکزی است: حتی اگر داده‌های خام نرمال نباشند، توزیعِ میانگینِ نمونه‌ها با افزایش اندازه‌ی نمونه n به شکل نرمال‌مانند نزدیک می‌شود. این پروژه شامل دو بخش اصلی است: (۱) تحلیل یک داده‌ی واقعی با توزیع راست‌کش، و (۲) تکرار همین ایده روی چند توزیع مشهور.

۱. بخش اول: داده‌ی واقعی `watch_time`

فرض کنید `watch_time` زمان تماشای روزانه‌ی ۵۰۰۰ کاربر (به دقیقه) را نشان می‌دهد و به دلیل وجود تعداد کم کاربران سنگین (تماشای زیاد)، داده‌ی خام معمولاً راست‌کش (Right-skewed) است.

(آ) میانگین‌های نمونه برای مقادیر مختلف n

برای مقادیر n داده‌شده در نوت‌بوک $(۲, ۵, ۱۰, ۳۰, ۵۰, ۷۰)$، این کارها را انجام دهید:

- $B = ۲۰۰۰$ بار نمونه‌گیری تصادفیِ بدون جایگذاری از `watch_time` انجام دهید و هر بار میانگین نمونه را محاسبه کنید.
- برای هر n، هیستوگرام میانگین‌های نمونه را رسم کنید.
- برای مقایسه‌ی بهتر میزان پراکندگی، از محور افقی (x-axis) یکسان در تمامی هیستوگرام‌ها استفاده کنید.
- توضیح دهید با افزایش n، شکل توزیع میانگین‌ها چه تغییری می‌کند و پراکندگی آن بیشتر می‌شود یا کمتر.

(ب) بازه‌ی ۹۵% برای میانگین نمونه: تئوری در برابر شبیه‌سازی

هدف این بخش مقایسه‌ی احتمال تئوری و احتمال تجربیِ قرار گرفتن $\bar{X}$ در یک بازه‌ی ۹۵% نرمال‌مانند است. با استفاده از $\hat{\mu}$ و $\hat{\sigma}$ (برآورد میانگین و انحراف معیار از داده)، یک n انتخاب کنید (مثلاً $n = ۳۶$) و مراحل زیر را انجام دهید:

- خطای معیار میانگین را محاسبه کنید:

$$\sigma_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}}.$$

- بازه‌ی ۹۵% را به شکل زیر بسازید:

$$[\hat{\mu} - ۲\sigma_{\bar{X}},\ \hat{\mu} + ۲\sigma_{\bar{X}}].$$

- احتمال تئوریِ $\bar{X}$ برای قرار گرفتن در این بازه را با فرض نرمال بودنِ $\bar{X}$ محاسبه کنید.
- با شبیه‌سازی (مثلاً ۱۰۰۰۰ بار) میانگین‌های نمونه را بسازید و احتمال تجربیِ افتادن $\bar{X}$ در بازه را به‌دست آورید.
- هیستوگرام میانگین‌های شبیه‌سازی را رسم کنید و دو مرز بازه را روی نمودار مشخص کنید.
- توضیح دهید مقدار تئوری و شبیه‌سازی چقدر نزدیک‌اند و با افزایش n چه تغییری انتظار دارید.

۲. بخش دوم: CLT روی توزیع‌های مشهور

در این بخش، همان ایده را برای چند توزیع مشهور (`Uniform`، `Exponential` و `Bernoulli`) بررسی کنید. برای هر توزیع و هر n در لیست:

- $B = ۲۰۰۰$ میانگین نمونه بسازید و هیستوگرام میانگین‌ها را رسم کنید.
- برای مقایسه‌ی بهتر پراکندگی میانگین‌ها، از محور افقی یکسان در هیستوگرام‌های مربوط به مقادیر مختلف n استفاده کنید.
- روی همان نمودار، منحنی نرمالِ نظری متناظر را نیز اضافه کنید (با میانگین و واریانسِ صحیحِ $\bar{X}$).
- برای هر توزیع، بیان کنید از چه مقداری از n به بعد، میانگین‌ها به نظر شما «به اندازه کافی» نرمال‌مانند می‌شوند.

۲. قانون بنفورد و قضیه حد مرکزی

در این پروژه، به بررسی *قانون بنفورد* و حضور آن در داده‌های دنیای واقعی با تکیه بر *قضیه حد مرکزی* (CLT) می‌پردازیم. قانون بنفورد بیان می‌کند که در بسیاری از مجموعه‌داده‌های طبیعی، توزیع رقم اول اعداد یکنواخت نیست و ارقام کوچک‌تر با احتمال بیشتری ظاهر می‌شوند.

در این پروژه از یک مجموعه‌داده‌ی اقتصادی واقعی استفاده کرده و توزیع رقم اول داده‌ها را با توزیع نظری قانون بنفورد مقایسه می‌کنید.

۱. **استخراج نظری قانون بنفورد**

احتمال رقم اول در قانون بنفورد را به‌صورت زیر استخراج کنید:

$$P(D = d) = \log_{۱۰}\left(۱ + \frac{۱}{d}\right), \quad d = ۱, \ldots, ۹.$$

سپس توضیح دهید که چگونه قضیه حد مرکزی می‌تواند ظهور قانون بنفورد را در داده‌های واقعی توجیه کند.

۲. **پیاده‌سازی قانون بنفورد**

تابعی بنویسید که احتمال نظری قانون بنفورد را برای یک رقم مشخص محاسبه کند. سپس با استفاده از این تابع، احتمالات مربوط به ارقام ۱ تا ۹ را محاسبه کرده و توزیع قانون بنفورد را رسم کنید.

۳. **بارگذاری داده و ساخت متغیر جدید**

مجموعه‌داده‌ی داده‌شده که شامل جمعیت و تولید ناخالص داخلی سرانه کشورهاست را بارگذاری کنید. با استفاده از رابطه‌ی زیر، متغیر جدید تولید ناخالص داخلی کل (GDP) را محاسبه نمایید:

$$\text{GDP} = \text{جمعیت} \times \text{تولید ناخالص داخلی سرانه}.$$

توضیح دهید که چرا متغیر GDP می‌تواند گزینه‌ی مناسبی برای بررسی قانون بنفورد باشد.

۴. **اعتبارسنجی تجربی قانون بنفورد**

رقم اول مقادیر GDP را استخراج کرده و توزیع تجربی ارقام اول را محاسبه کنید. سپس این توزیع را با توزیع نظری قانون بنفورد مقایسه کرده و با استفاده از نمودار، میزان تطابق داده‌ها با قانون بنفورد را تحلیل کنید.

۳. برآورد پارامتر: تخمین شلوغی یک کافه (مدل پواسون)

فرض کنید مسئول برنامه‌ریزی نیروی انسانی یک کافه‌ی کوچک نزدیک دانشگاه هستید. در هر ساعت، تعداد تصادفی‌ای از مشتریان وارد کافه می‌شوند. برخی ساعات خلوت هستند و برخی ساعات بسیار شلوغ.

تعداد مشتریان ورودی در هر ساعت با یک مدل پواسون به‌صورت زیر مدل می‌شود:

$$\text{Arrivals Hourly} \sim \text{Poisson}(\lambda)$$

که در آن λ میانگین تعداد مشتریان در هر ساعت است.

هدف این پروژه، تخمین λ و بررسی رفتار برآورد آن با افزایش حجم داده است.

۱. **یک نمونه مشاهده‌شده**

یک نمونه‌ی ۲۴ ساعته از تعداد مشتریان ساعتی تولید کنید و میانگین نمونه را محاسبه و گزارش دهید.

۲. **حداکثر درست‌نمایی به‌صورت یک منحنی**

بدون استفاده از فرمول بسته، برای مقادیر مختلف λ، میزان درست‌نمایی داده‌ی مشاهده‌شده را محاسبه کنید.

- منحنی لگاریتم درست‌نمایی را بر حسب λ رسم کنید.
- مقدار λ متناظر با بیشینه‌ی منحنی را به‌عنوان تخمین ML مشخص کنید.
- میانگین نمونه را روی همان نمودار علامت بزنید و در یک جمله مقایسه کنید.

۳. **تخمین ML از روی نمونه**

تخمین ML پارامتر λ را مستقیماً از نمونه محاسبه کنید و مقدار «میانگین تعداد مشتریان در هر ساعت» را گزارش دهید.

۴. **تأثیر افزایش تعداد مشاهدات**

برای طول‌های مختلف مشاهده (از چند ساعت تا چند روز):

- آزمایش را $B = ۱۰۰۰$ بار تکرار کنید و هر بار یک تخمین λ به‌دست آورید.
- تغییرپذیری تخمین‌ها را برای طول‌های مختلف مشاهده مقایسه کنید.
- با استفاده از نمودارهای مناسب (مانند هیستوگرام و جعبه‌ای)، نشان دهید با افزایش تعداد ساعات مشاهده، پراکندگی تخمین‌ها چگونه تغییر می‌کند.
- واریانس (یا انحراف معیار) تخمین‌ها را بر حسب تعداد ساعات مشاهده رسم کرده و روند آن را توضیح دهید.

۵. **مقایسه‌ی توزیع واقعی و توزیع تخمینی**

با استفاده از تخمین ML به‌دست‌آمده:

- تابع جرم احتمال پواسون با λ واقعی و $\hat{\lambda}$ را روی یک نمودار رسم کنید.
- مختصر توضیح دهید این دو توزیع چقدر به یکدیگر نزدیک هستند.