به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

آمار و احتمال مهندسی

پروژه‌ی نهایی

طراح:	تابان سلیمانی
تاریخ ویرایش نسخه	۲۴ خرداد ۱۴۰۲
مهلت تحویل گزارش	۱۵ تیر ۱۴۰۲

فهرست مطالب

۱ تناقض تاریخ تولد

تصور کنید که با گروه n نفره از دوستان خود در یک دیدار مجدد پس از دبیرستان هستید. پس از آنکه بین مکالمه‌های عادی، کمی از هم جدا شده‌اید، شروع به فکر کردن درباره برخی از مسائل می‌کنید و ناگهان از خود می‌پرسید: احتمال اینکه در بین این افراد حداقل دو نفر در یک روز متولد شده باشند، چقدر است؟ (همانطوری که می‌دانید، طبق اصل لانه‌ی کبوتری با وجود ۳۶۶ نفر این احتمال برابر با ۱ است. بنابراین تعداد افراد را به صورت $n \leq 365$ و تعداد روزهای سال را ۳۶۵ روز در نظر بگیرید.) وقتی از عموم افراد راجع به حداقل مقدار n برای آنکه احتمال بالا برابر با ۰/۵ باشد سوال می‌شود، اغلب تمایل دارند تا به عنوان یک پاسخ شهودی ۱۸۳ یعنی نیمی از ۳۶۵ را انتخاب کنند. تفکر پشت این مقدار این است که با نصف کردن تعداد روزهای یک سال عادی، مقدار حداقل لازم به دست می‌آید تا احتمال برابر با ۰/۵ شود. مسئله تناقض تاریخ تولد یک واقعیت ریاضی است که در آن بر خلاف غرایز ما به طور غافل‌گیرکننده‌ای، ثابت می‌شود که تعداد افراد کمی لازمست تا حداقل دو نفر از آن‌ها در یک روز متولد شده باشند.

۱. به ازای مقادیر $n \leq 100$ ، احتمال اینکه حداقل دو نفر از افراد (از میان n نفر) در یک روز متولد شده باشند را بر حسب n رسم نمایید. حداقل مقدار n برای اینکه احتمال بالا برابر با ۰/۵ شود را به دست آورده و نیز گزارش کنید که به ازای چه مقدار n، احتمال بالا برابر با ۰/۹۹ خواهد شد.

۲. با استفاده از بسط سری تیلور برای تابع نمایی ($e^x = 1 + x + \frac{x^2}{2!} + \ldots \approx 1 + x$) مقادیر به دست آمده در قسمت قبل را توجیه نمایید. (راهنمایی: اگر فرض کنیم که احتمال خواسته شده بر حسب n به صورت $p_n(x)$ باشد، شما باید احتمال $\bar{p}_n(x)$ را با استفاده از بسط تیلور داده شده برحسب n تخمین بزنید.) تقریب به دست آمده برای احتمال خواسته شده به همراه مقدار واقعی آن را در قالب یک نمودار بر حسب n نمایش دهید.

۲ سکه‌بازی

این بار به دو بازیکن یک سکه برای بازی داده می‌شود. قاعده‌ی بازی به این شکل است که هریک، با مقدار معینی بودجه شروع به بازی می‌کنند، فرض کنید که بازیکن اول n_1 دلار و بازیکن دوم n_2 دلار در اختیار دارد. هر بار که سکه پرتاپ می‌شود، در صورتی که نتیجه شیر باشد بازیکن اول برنده‌ی این دور خواهد بود و در صورتیکه نتیجه خط باشد، بازیکن دوم برنده می‌شود. در هر دور فرد بازنده به برنده یک دلار می‌دهد و بازی تا جایی ادامه پیدا می‌کند که یکی از آن‌ها تمام بودجه‌ی خود را از دست دهد.

۱. به طور تئوری برای هر یک از دوحالت زیر، احتمال برد هریک از دو بازیکن را به دست آورید.

(آ) سکه سالم باشد.

(ب) احتمال شیر آمدن در هر بار پرتاب سکه برابر با p باشد.

۲. با استفاده از نتایج بخش قبلی استدلال نمایید که هریک از مقادیر p، n_1 و n_2 چگونه در برد هریک از بازیکن‌ها اثر می‌گذارد.

۳. با پیاده‌سازی سناریوی توصیف شده میانگین دورهای بازی برای حالتی که $p = 0.45$، $n_1 = 9$ و $n_2 = 1$ است را به دست آورید. (راهنمایی: در اینجا لازم است تا سناریوی بازی را چندین بار اجرا نمایید و درهربار از اجرای بازی، پرتاپ سکه را نهایتاً تا ۵۰ تکرار انجام دهید و سپس از نتایج به دست آمده در هر بار اجرا میانگین گرفته و آن را به عنوان میانگین دورهای بازی گزارش کنید.)

۳ محاسبه مساحت

یک روش آماری براي محاسبه مساحت زیرمجموعه S از مربع واحد این است که به طور تصادفی، یکنواخت و مستقل از یکدیگر دنباله‌ای از نقاط را در فضای مربع واحد $[0,\ 1] \times [0,\ 1]$ انتخاب کنیم. اگر نقطه i به زیرمجموعه S تعلق داشته باشد، مقدار متغیر تصادفی X_i برابر یک و در غیر این صورت صفر در نظر گرفته می‌شود. فرض کنید که $X_1,\ X_2,\ \ldots$ دنباله‌ای از این متغیرهای تصادفی باشد و برای هر مقدار n، S_n به صورت زیر تعریف شود.

$$S_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

۱. ابتدا به صورت تئوری نشان دهید که $E[S_n]$ با مساحت زیرمجموعه S برابر است و $var(S_n)$ با افزایش n به صفر میل می‌کند.

۲. یک رابطه‌ی بازگشتی برای محاسبه S_n، تنها با داشتن دو مقدار S_{n-1} و X_n ارائه دهید. (در واقع، در هر مرحله برای محاسبه S_n نیازی به ذخیره $X_k,\ k = 1,\ 2, \ldots,\ n-1$ نیست و تنها با داشتن دو مقدار S_{n-1} و X_n، محاسبه صورت می‌گیرد.)

۳. فرض کنید که زیرمجموعه‌ی S بیانگر دایره‌ی محاط به مربع واحد باشد. برای مقادیر $n = 1,\ 2,\ \ldots, 10000$، با تولید مختصات $(x,\ y)$ به صورت تصادفی و به روش شرح داده شده در قسمت قبلی، S_n را محاسبه کنید. توضیح دهید که چگونه با این روش می‌توانید به صورت تجربی مقدار عدد π را محاسبه کنید.

۴. مانند قسمت قبل، با محاسبه S_n برای مقادیر $n = 1,\ 2,\ \ldots, 10000$، مساحت ناحیه‌ی محدود به $0 \leq \cos \pi x + \sin \pi y \leq 1$ را به طور تقریبی محاسبه کنید.

۴ کارخانه گلاب‌گیری

فرض کنید شما در کاشان در یک کارخانه گلاب‌گیری مشغول به کار هستید. دوست شما ویلیام گاست، مایل‌ها دورتر در دوبلین مشغول کاری مشابه شماست اما هنوز مقاله معروف خود یعنی t-student را منتشر نکرده است و شما اطلاعی از توزیع t ندارید.

فرض کنید مقدار گلابی که از یک گل محمدی استخراج می‌شود از توزیع نرمال با میانگین ۱۵۰ و انحراف معیار ۱۰ پیروی می‌کند. و گل‌های باغ شما تعداد زیادی گل محمدی دارد که هیچ‌گاه تمام نمی‌شود.

شما که در آمار و احتمال دستی بر آتش دارید، با قضیه حد مرکزی آشنا هستید و می‌دانید که در صورتی که n عدد بزرگی باشد و $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$ داریم:

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z$$

که در آن X_iها مستقل و از توزیع یکسان با میانگین μ و واریانس σ^2 هستند.

شما کنجکاو می‌شوید که اگر نمونه‌ای که از جامعه داریم محدود باشد (به هر حال همه که مانند شما چنین باغ بزرگی ندارند!) و واریانس جامعه را نداشته باشیم و به جای آن از واریانس نمونه استفاده کنیم، چه اتفاقی خواهد افتاد؟ به عبارتی اگر x_iها نمونه کوچک ما باشد و داشته باشیم:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}$$

در آن صورت $t = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ از چه توزیعی تبعیت خواهد کرد؟

در نتیجه خودتان دست به کار می‌شوید تا با استفاده از شرایطی که برایتان مهیاست جواب سوالتان را پیدا کنید.

۱. در ابتدا تابعی بنویسید که با گرفتن نمونه‌ها و میانگین جامعه در ورودی مقدار t را محاسبه کند. حال شما برای اینکه توزیع t را به دست بیاورید باید به دفعات متعددی از گل‌هایتان نمونه‌گیری انجام داده و مقادیر t آن‌ها را به دست بیاورید.در ادامه نیز تابعی بنویسید که با گرفتن تعداد نمونه‌گیری‌ها و اندازه نمونه از گل‌های باغتان نمونه‌گیری را انجام دهد و با استفاده از تابعی که در قسمت قبل نوشتید مقادیر t متناظر آنها را برگرداند. با استفاده از تابعی که نوشتید به تعداد ۱۰۰۰۰ بار و هر با اندازه ۲ نمونه‌گیری را انجام دهید و نتیجه را در قالب یک نمودار نمایش دهید.

۲. بعد از کشیدن نمودار و مشاهده آن به شباهتش با توزیع نرمال استاندارد پی می‌برید و خوشحال می‌شوید که قضیه حد مرکزی برای این شرایط نیز صادق است و به یک تعمیم برای قضیه حد مرکزی دست یافته‌اید! اما وقتی بیشتر دقت می‌کنید حس می‌کنید که یک جای این نمودار می‌لنگد. در نتیجه نمودار قبل را در کنار نمودار توزیع نرمال استاندارد رسم می‌کنید. دو نمودار را با هم مقایسه کنید. چه تفاوتی میان آنها وجود دارد؟

۳. متاسفانه این دونمودار با وجود شباهت زیادی که به هم دارند از توزیع‌های متفاوتی می‌آیند! شما نسبت به توزیعی که به دست آوردید کنجکاوتر شده و سعی می‌کنید آن را به ازای اندازه نمونه‌های مختلف بررسی کنید. به ازای اندازه نمونه‌های $n = 2,\ n = 3,\ n = 6,\ n = 100$ مانند قبل نمونه‌گیری‌های ۱۰۰۰۰ تایی انجام داده و آن‌ها را داخل یک dataframe ذخیره کنید. همچنین یک ستون نیز برای توزیع نرمال استاندارد به دیتافریم اضافه کنید. سپس در یک نمودار، توزیع‌های مربوط به هر یک از nهای مختلف و توزیع استاندارد نرمال را رسم کرده و به هر یک از نمودارها، یک رنگ جدا اختصاص دهید به طوری که از همدیگر قابل تمیز باشند. از نموداری که رسم کردید و مقایسه توزیع‌ها چه نتیجه‌ای می‌گیرید؟

۴. روزی دوست شما، ویلیام گاست به کاشان می‌آید. شما با او در رابطه با توزیع جدیدی که کشف کردید صحبت می‌کنید. او می‌گوید که اتفاقاً درباره این توزیع مقاله‌ای چاپ کرده و توابع مربوط به این توزیع را به کتابخانه پایتون اضافه کرده است. او همچنین می‌گوید که این توزیع شامل پارامتری به نام درجه آزادی یا ν می‌باشد که در اینجا برابر با $n-1$ است.

با استفاده از دیتافریمی که ساختید توزیع مربوط به $n = 2$ و توزیع t متناظر با آن را در یک نمودار رسم کنید و آن دو را با هم مقایسه کنید. آیا اکنون به نظر شما این دو توزیع یکسانند؟

۵. فرض کنید کسی به شما ۶ عدد گل محمدی می‌دهد و ادعا می‌کند که از باغ شما چیده است. شما به او شک می‌کنید و فرآیند گلاب‌گیری را روی آن‌ها انجام داده و مقادیر گلاب به دست آمده از آن‌ها برابر ۱۳۰، ۱۲۰، ۱۳۶، ۱۴۵، ۱۵۰، ۱۴۰ میلی‌لیتر شده است. آیا به نظر شما او راست می‌گوید؟ فرض صفر و فرض جایگزین را در این مورد بیان کنید.

مقدار p-value را یکبار با استفاده از دیتافریمی که دارید و یکبار با استفاده از توابع کتابخانه‌ای پایتون به دست آورید و آن دو را با هم مقایسه کنید.

۵ داده‌بازی

۱.۵ بخش اول

۱. در این بخش قرار است تا با دیتافریم airquality که اطلاعاتی در مورد وضعیت آب‌وهوای نیویورک در یک بازه‌ی زمانی را در اختیار ما قرار می‌دهد، کار کنید.

۲. به کمک boxplotها یک نمودار مناسب ارائه دهید که وضعیت دما (temp) را بر حسب ماه‌های مختلف نشان دهد.

۳. بخش الف را برای میانگین باد و میانگین غلظت اوزون و میانگین تابش خورشید بر حسب ماه‌های مختلف تکرار کنید و نتایج خود را از این ۴ نمودار شرح دهید.

۴. به کمک نمودار scatter نمودار میانگین غلظت اوزون بر حسب دما را رسم کنید و آن را تحلیل کنید.

۵. این بار نمودار scatter را براساس ویژگی‌های جذر میانگین غلظت اوزون و ماه و باد رسم کنید. سعی کنید نمودار رسم شده اطلاعات را به خوبی نشان دهد. نتایج خود را بیان کنید. (به وابستگی میان متغیرها توجه کنید.)

۶. نمودار هیستوگرام برای باد و تابش خورشید رسم کنید، این دو نمودار را با هم و با نمودار توزیع نرمال مقایسه کنید و نتیجه‌ی خود را آن بیان کنید.

۲.۵ بخش دوم

در اتاقی یک بلندگو (فرستنده) قرار دارد که سیگنالی ارسال می‌کند و همچنین سه میکروفون (گیرنده) در فواصل و جاهای مختلف قرار داده شده‌اند که این سیگنال را دریافت می‌کنند. اطلاعات سیگنال ارسال شده و دریافتی بر حسب زمان در دیتاست speaker.csv به شما داده شده است. ابتدا داده‌ها را در یک دیتافریم لود کنید.

۱. نمودار اطلاعات فرستنده و سه گیرنده را بر حسب زمان در یک نمودار رسم کنید.

۲. داده‌های اندازه‌گیری شده به دلیل تفاوت و فاصله‌ی گیرنده‌ها و عوامل محیطی ممکن است دچار تغییراتی شده باشند، به همین منظور لازم است تا داده‌ها را نرمالایز کنیم، یعنی تبدیل خطی‌ای روی آن‌ها اعمال کنیم که توزیع نهایی دارای امید ریاضی صفر و انحراف معیار واحد باشد. ($x' = \frac{x-\bar{x}}{\sigma_x}$)

داده‌های نرمالایز شده‌ی فرستنده و سه گیرنده را به دیتافریم اضافه کنید. از این به بعد تنها با داده‌های نرمالایز شده کار خواهیم کرد. برای اطمینان میانگین و واریانس داده‌های نرمالایز شده را حساب کنید. آیا نتیجه با آنچه انتظار دارید یکسان است؟

۳. توزیع توام داده‌های نرمالایز شده‌ی سه گیرنده با فرستنده را در نمودار یا نمودارهای مناسبی نمایش دهید. (راهنمایی: جواب‌ها همگی نوعی بیضی هستند.)

۴. کوواریانس اطلاعات نرمالایز شده‌ی گیرنده‌ها با فرستنده را حساب کنید. (سه مقدار) مقادیر به دست‌آمده را با توزیع‌های توام که در قسمت قبل کشیدید مقایسه کنید. (به کوواریانس بین توزیع‌های نرمالایز شده correlation نیز می‌گویند که مقداری بین منفی یک و یک است.)

۵. همان‌طور که احتمالاً حدس زدید گیرنده‌ها علاوه بر آن که اطلاعات را با مقداری نویز و تضعیف شده دریافت می‌کنند آن را با تاخیر هم می‌بینند. یعنی نمودار داده‌های گیرنده‌ها در مقایسه با داده‌های فرستنده شیفت خورده است. به دلیل نویزی که در اطلاعات وجود دارد هیچوقت نقاط با شیفت دادن دقیقا روی هم قرار نمی‌گیرند.

یک روش خوب برای پیدا کردن تاخیر بین این دو موج شیفت دادن یک موج و محاسبه‌ی کوواریانس بین این دو است تا جایی که این کوواریانس بیشینه شود. اگر نویزی وجود نداشت این مقدار برای داده‌های نرمالایز شده، یک می‌بود چرا که کوواریانس یک متغیر نرمالایز شده با خودش برابر است.

داده‌های گیرنده‌ی اول را به ازای جابه‌جایی‌های مضرب ۵۰ از صفر تا ۱۰۰۰ میکروثانیه (نصف کل بازه‌ی زمانی) به سمت چپ شیفت دهید. سپس داده‌های ۱۰۰۰ میکروثانیه ابتدایی از فرستنده و گیرنده را که باقی می‌ماند، جدا کنید و کوواریانس این دو را محاسبه کنید. حال نمودار کوواریانس بر حسب شیفت را رسم کنید. (راهنمایی: همیشه باید ۲۰۰ داده از میان ستون receive1 نرمالایز شده بردارید و کواریانس آن را با ۲۰۰ داده‌ی ابتدایی ستون send نرمالایز شده محاسبه کنید.)

شیفتی که به ازای آن بیشترین کوواریانس بین داده‌های گیرنده‌ی شیفت خورده و فرستنده ایجاد می‌شود را به همراه کوواریانس بیشینه پیدا کنید. توزیع توام ۱۰۰۰ میکروثانیه اول گیرنده‌ی شیفت خورده با فرستنده را رسم کنید.

۶ توضیحات

دانشجویان عزیز حتما به نکات زیر توجه داشته باشند.

- پروژه به‌گونه‌ای طراحی شده که به دانش آماری فراتر از آن‌چه در این درس آموخته‌اید نیاز نداشته باشد وآن‌چه را که آموخته‌اید تثبیت و تفهیم می‌کند. به همین جهت انجام آن برای یادگیری درس اکیدا توصیه می‌شود.
- صرف‌نظر از رویکرد آموزشی این پروژه، آخرین نقطه‌ی جبران نمراتتان در این درس می‌باشد و بنا به سابقه‌ی چندساله، به اسکیل‌شدن نمرات امیدی نیست، در نتیجه از اهمیت این موقعیت غافل نشوید.
- شما می‌بایست علاوه بر کدهای پیاده‌شده، گزارشی تحلیلی از نتایج خود ارائه دهید. توجه داشته باشید که مفهوم گزارش پروژه با مفهوم توضیح کد متفاوت است در نتیجه در فایل گزارش، از درج کد جدا بپرهیزید.
- کدهای پایتون و آر خود را حتما در قالب دفترچه‌ی ژوپیتر بارگذاری کنید. دستیاران آموزشی موظف به اجرای کدهای شما نیستند.
- اسکریپت‌های خود را خوانا و تمیز بنویسید. طبیعتا این درس، درس برنامه‌نویسی نیست اما کد بسیار پیچیده و غیرقابل فهم نمره‌ی کامل را دریافت نمی‌کند. استفاده از توابع و نام‌های متغیرهای بامعنا به خوانایی کد می‌افزاید.
- گزارش‌کار، اولین و مهم‌ترین آیتم نمره‌دهی می‌باشد در نتیجه با صرف زمان مناسب، گزارشی تهیه کنید که بازتاب‌گر زحماتی باشد که برای انجام پروژه کشیده‌اید. استفاده‌ی صحیح از نیم‌فاصله، علائم نگارشی، گویا بودن جملات و پاراگراف‌بندی مناسب از جمله مواردیست که در نگاه اول جلب توجه می‌کند و نکاتی نظیر استفاده از زیرنویس برای تصاویر و بالانویس برای جداول، ارجاع‌دادن به روابط و تصاویر با شماره‌ی مربوط به هر کدام و ... از جمله خصوصیت‌های یک نوشته‌ی آکادمیک است. متن گزارش را با فونت B Nazanin و اندازه‌ی ۱۴ در قالب گزارش قرار داده شده روی سایت تایپ نمائید. از قرار دادن عکس از نوشته‌ی دست‌نویس خود در گزارش به‌شدت پرهیز کنید و روابط ریاضی را نیز تایپ کنید.
- با توجه به مفهوم امتیازی بودن پروژه، به شدت با موارد تقلب چه در کد و چه در گزارش برخورد خواهد شد.
- سعی می‌شود از برخی از دوستان از طریق تماس تصویری سوالاتی در قالب تحویل پروژه پرسیده شود. در نتیجه مشخص است که هر شخص باید به تمامی محتوایی که ارائه می‌دهد مسلط باشد.
- در نهایت یک فایل گزارش پی‌دی‌اف را در کنار دفترچه‌های ژوپیتر زیپ کرده و با نام <sid>-surname.zip در صفحه‌ی درس بارگذاری کنید.
- ابهامات خود در مورد سوالات و یا قالب گزارش در گروه تلگرامی درس مطرح کنید. در انتهای هر پیام بنده را منشن کنید. سوالات در گروه پرسیده شده و همان‌جا پاسخ داده خواهند شد تا در دسترس همه‌ی دانشجوها قرار بگیرند.