

۱ مسابقه‌ی تلویزیونی

مسابقه‌ی تلویزیونی‌ای در شبکه‌ی دانشگاه تهران در حال برگزاری است. روال مسابقه به این شکل است که مجری برنامه ۳ در بسته روبه‌روی شما قرار می‌دهد و شما می‌بایست یکی از این درها را انتخاب کنید. پشت یکی از این درها خودرویی قرار دارد و پشت دو در دیگر بزر است! هر کدام از این درها را که انتخاب کنید، شی پشت آن نصیب شما می‌شود. هنگامی که شما یکی از درها را انتخاب کنید، مجری می‌تواند از بین دو در باقی‌مانده یکی را برای شما پوچ کند (می‌دانیم که مجری صادق است و قصد گول و حتماً دری را باز می‌کند که پشت آن بزر است). بعد از این کار مجری از شما سوال و سوسه‌برانگیزی می‌کند، «آیا مایل هستید که انتخاب خود را عوض کرده و در دیگری که پشت آن مشخص نیست را بردارید؟» آیا به نفع شماست که انتخاب خود را تغییر دهید؟ سه پاسخ وجود دارد:

۱. بله، ارزش دارد که انتخاب اولیه‌ی خودم رو عوض کنم.

۲. خیر، انتخاب اولیه‌ام بهتر است و اون رو حفظ می‌کنم.

۳. انتخاب هر کدام از این درها هیچ تفاوتی در نتیجه ندارد و یکسان است.

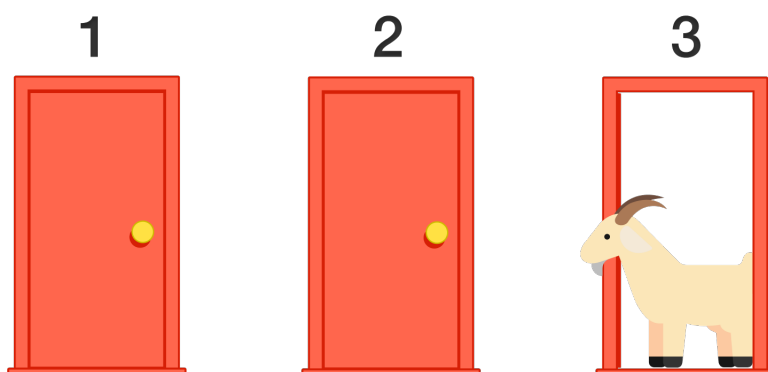
فرض کنید $3n$ دانشجو از دانشکده‌های مختلف دانشگاه در این مسابقه شرکت کرده‌اند. این $3n$ دانشجو را به سه گروه n تایی تقسیم می‌شوند که هر گروه بعد از اینکه مجری یکی از درها را حذف کرد (مجری حتماً یکی از درهایی که پشت آن بزر است را حذف می‌کند) یکی از تصمیم‌های زیر را اتخاذ می‌کنند:

۱. دسته‌ی اول انتخاب اولیه‌ی خود را عوض نمی‌کند.

۲. دسته‌ی دوم انتخاب اولیه‌ی خود را عوض کرده و در باقی‌مانده را انتخاب می‌کند.

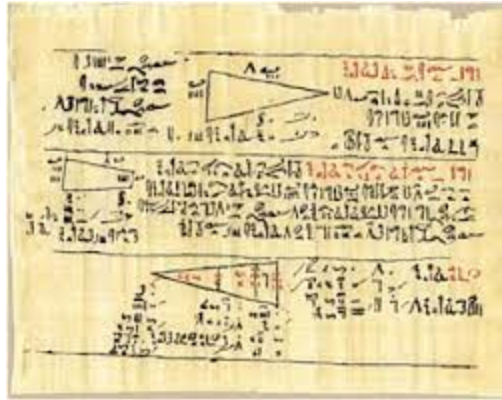
۳. دسته‌ی سوم به احتمال ۰.۵ انتخاب خود را عوض می‌کند.

با فرار دادن n با مقادیر ۱۰، ۱۰۰، ۱۰۰۰، ۱۰۰۰۰ احتمال برنده شدن هر دسته را بررسی کنید و در یک نمودار نشان دهید. سعی کنید پاسخ بدست آمده را توجیه کنید.



۲ تخمین عدد پی

حتما با عدد π آشنا هستید. تلاش برای تخمین این عدد در زمان بابل کهن و سپس در مصر باستان آغاز شد اما نام خاصی برای آن تعیین نشده بود و در آن زمان نمی‌دانستند که عدد π گنگ است که با ارائه نظریه پاپیروس توسط احمس (ریاضی‌دان مصر باستان) گنگ بودن آن مشخص شد. پس از



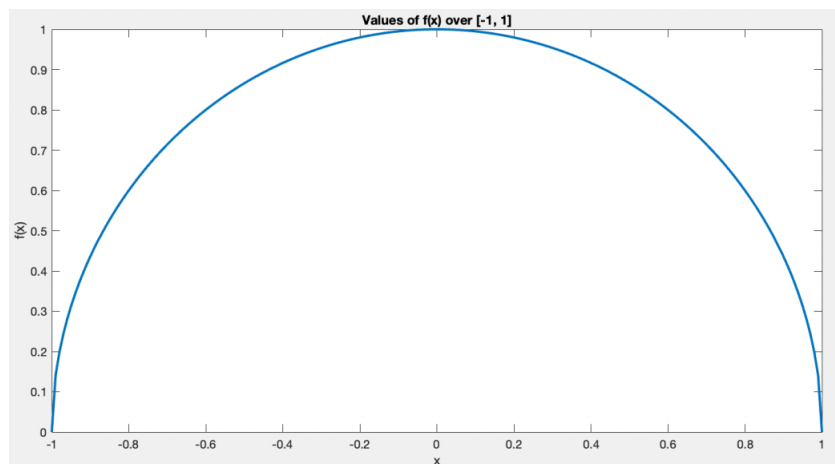
آن، اولین نظریه در مورد مقدار تقریبی عدد π با تقریب زدن مساحت دایره به وسیله یک شش ضلعی منتظم محیطی و یک شش ضلعی منتظم محاطی توسط ارشمیدس بیان شد. از جمله دانشمندان اروپایی که به تقریب عدد π نزدیک شدند، جیمز گریگوری بود که برای پیدا کردن مقدار تقریبی عدد π از فرمول زیر استفاده کرد:

$$\pi = 4\left(\frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \frac{1}{13} - \frac{1}{15} + \dots\right)$$

یکی از مشکلاتی که در این روش وجود دارد این است که برای پیدا کردن مقدار عدد π تا ۶ رقم اعشار باید پنج میلیون جمله از سری فوق را با هم جمع کنیم که نشان می‌دهد این الگوریتم رشد نمایی شدیدی دارد و زمان زیادی را برای محاسبه ارقام صرف می‌کند.

در این تمرین قصد داریم به روش مونته کارلو، مقدار این ثابت را تخمین بزنیم. روش‌های مونته کارلو دسته وسیعی از الگوریتم‌های محاسباتی هستند که بر پایه نمونه‌گیری تصادفی از محیط بنا شده‌اند. در حقیقت در این روش‌ها سعی می‌کنیم از تصادفی بودن نمونه‌ها استفاده کنیم تا مقداری را حساب کنیم که ذاتا غیر تصادفیست. نخستین پژوهش‌ها روی این متد با حل مساله سوزن بوفون آغاز شد و پس از آن این متد در شبیه‌سازی‌های حوزه‌های بیولوژی، فیزیک، مهندسی و ... استفاده شد. در طی شرح مساله بهتر با این مفهوم آشنا خواهید شد.

برای تخمین عدد π منحنی $\sqrt{1-x^2}$ را در نظر بگیرید.



۱. مساحت زیر منحنی در بازه $[0, 1]$ را بر حسب عدد π محاسبه کنید.

۲. با استفاده از فرمول زیر و نتیجه محاسبه بخش قبل، میانگین مقدار این منحنی در بازه $[0, 1]$ را به دست آورید:

$$f_{avg} = \frac{1}{b-a} \int_a^b f(x) dx$$

۳. حال فرض کنید مقدار میانگین این منحنی را در بازه $[0, 1]$ از قبل می‌دانیم. با توجه به رابطه به دست آمده در بخش قبل، مقدار عدد π را بر حسب f_{avg} به دست آورید.

۴. حال اگر X را یک متغیر تصادفی با توزیع یکنواخت در بازه $[0, 1]$ و $Y(f(X))$ متغیر تصادفی‌ای با میانگین μ باشد، n نمونه از توزیع یکنواخت X استخراج می‌کنیم. در این صورت با فرض $n = 100$ و محاسبه μ تخمینی از عدد پی ($\hat{\pi}$) ارائه دهید.

عددی که با اولین دفعه اجرای برنامه به دست آمد را ثبت کنید. سپس برنامه را دوباره اجرا کنید و عدد دوم را نیز ثبت کنید. آیا دو عدد مساوی هستند؟ برای پاسخ خود چه توجیهی دارید؟

۵. این بار با فرض $n = 1000$ و $n = 10000$ تخمینی از عدد پی (π) به دست آورید.

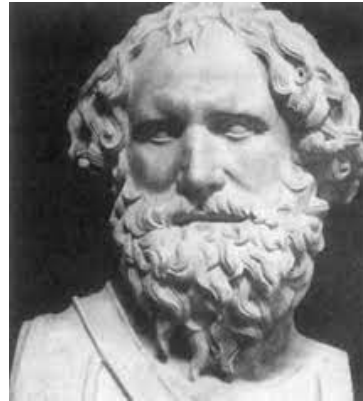
۶. تخمین‌های به دست آمده در قسمت (د) خودشان نیز متغیرهای تصادفی هستند. با استفاده از تخمین بیشینه درست‌نمایی تخمینی از مقدار میانگین و واریانس را برای $n = 100$ ، $n = 1000$ ، $n = 10000$ جداگانه محاسبه و در یک نمودار، مقادیر میانگین و واریانس نمونه‌ها را نمایش دهید.

۷. نمودار به دست آمده در قسمت (و) را توجیه کنید.

۸. واریانس‌های به دست آمده در قسمت (و) نشان‌دهنده کدام ویژگی از تخمین هستند؟



(ب) جیمز گریگوری



(آ) ارشمیدس

۳ پارادوکس سنت پترزبورگ

پارادوکس سنت پترزبورگ، بر اساس یک بازی ساده با پرتاب یک سکه سالم شکل می‌گیرد. این معمای فلسفی و اقتصادی توسط نیکولاس برنولی، در سال ۱۷۱۳ مطرح شد و دانیل برنولی اولین کسی بود که در سال ۱۷۳۸ این پارادوکس را مورد بحث و بررسی قرار داد. این بازی قانون ساده‌ای دارد. در ابتدا از شما مبلغ اولیه‌ای برای شروع بازی دریافت می‌شود. سپس از شما خواسته می‌شود تا یک سکه را پرتاب کنید. اگر اولین شیر در پرتاب اول حاصل شد، شما ۲ دلار برنده می‌شوید و با خط آوردن از بازی خارج می‌شوید. اگر در پرتاب دوم هم حاصل شد، ۴ دلار برنده می‌شوید و به همین ترتیب هربار میزان برد ۲ برابر می‌شود و با خط آوردن از بازی خارج می‌شوید. در واقع اگر n بار سکه را پرتاب کنید تا در پرتاب n م خط بیاید، 2^n دلار برنده می‌شوید. حال سوال این است شما حداکثر چه مبلغی را برای شروع بازی پرداخت کنید تا بازی برای شما منصفانه باشد؟ برنامه‌ای بنویسید که این عملیات تاس انداختن را شبیه‌سازی کند و به سوالات زیر پاسخ دهید:

۱. امید ریاضی مبلغ برد را به صورت دستی به ازای n پرتاب محاسبه کنید. در صورت میل کردن n به بی‌نهایت این مقدار چقدر است؟
۲. بوفون برای این پارادوکس راه حلی داد. او این بازی را ۲۰۴۸ بار تکرار کرد و با محاسبه امید ریاضی مبلغ برد، مبلغ منطقی را برای پرداخت اولیه مشخص کرد. حال شما آزمایش بوفون را شبیه‌سازی کنید و مبلغی را برای پرداخت اولیه پیشنهاد دهید.



(ب) نیکولاس برنولی



(آ) دانیل برنولی

۱. در این پروژه، ما قصد داریم مجموعه‌ای از داده‌های واقعی را با آن‌چه در این درس آموخته‌اید بررسی و تحلیل کنیم. برای شروع تجزیه و تحلیل یک مجموعه داده، اولین قدم آشنایی با آن است. در اولین قدم می‌توان با مشاهده مواردی مثل ویژگی‌های مجموعه داده و توزیع مقادیر و تجسم داده‌ها برای حدس زدن اولیه در مورد آن آشنایی را انجام داد. در مرحله‌ی بعدی با انجام آزمایشات آماری، اطمینان حاصل می‌کنیم که حدس‌هایمان درست است و ادعاهای خود را با اطمینان بیان می‌کنیم. برای این سوال از دیتاست StudentPerformance استفاده کنید.

(آ) به طور خلاصه دیتاست داده‌شده را توصیف کنید و تعداد ویژگی‌های آن را بیان کنید.

(ب) با این توصیف، به نظر شما کدام ویژگی اطلاعات مهم‌تری دارد و چرا؟

یک ویژگی عددی از دیتاست انتخاب کنید و به سوالات زیر پاسخ دهید.

(آ) یک هیستوگرام با bin size مناسب بکشید و ویژگی‌های بارز توزیع آن را بیان کنید. از یک نمودار مناسب برای مقایسه کردن این توزیع با توزیع نرمال استفاده کنید.

(ب) چولگی یک متغیر تصادفی توصیف مناسبی از رابطه‌ی میانگین و میانه‌ی آن است. رابطه‌ی چولگی یک متغیر تصادفی را یافته و سعی کنید آن را توصیف کنید. حال چولگی را برای این متغیر انتخاب‌شده محاسبه کرده و نتیجه را توصیف کنید.

(ج) مقادیر میانگین، میانه، واریانس و انحراف معیار را محاسبه کرده و آن‌ها را توصیف کنید.

(د) نمودار density-plot را برای این متغیر بکشید و خطوط مربوط به میانگین و میانه را به آن اضافه کنید، حال ارتباط این خطوط را با نمودار رسم‌شده بیان کنید.

(ه) بازه اطمینان ۹۵ درصدی را برای میانگین این متغیر محاسبه کنید.

(و) هیستوگرام این متغیر را کشیده و میانگین هر نمونه را به صورت خطی عمودی در بالای هیستوگرام نمایش دهید. هم‌چنین بازه اطمینان را به صورت دو خط عمودی نشان دهید.

حال دو ویژگی عددی متفاوت از این دیتاست را انتخاب کرده و به سوالات زیر پاسخ دهید.

(آ) ابتدا بدون شبیه‌سازی و بر اساس فرضیات خود، ارتباط این دو ویژگی را از روی توضیح آن‌ها حدس زده و حدس خود را بیان کنید.

(ب) نمودار پراکندگی را برای این دو متغیر رسم کرده و ارتباط آن‌ها را با هم بیان کنید.

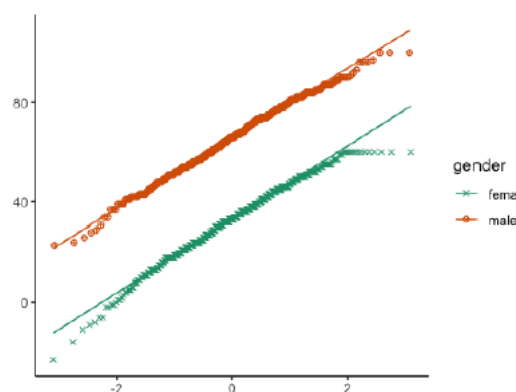
(ج) ضریب هم‌بستگی بین این دو متغیر را محاسبه کنید.

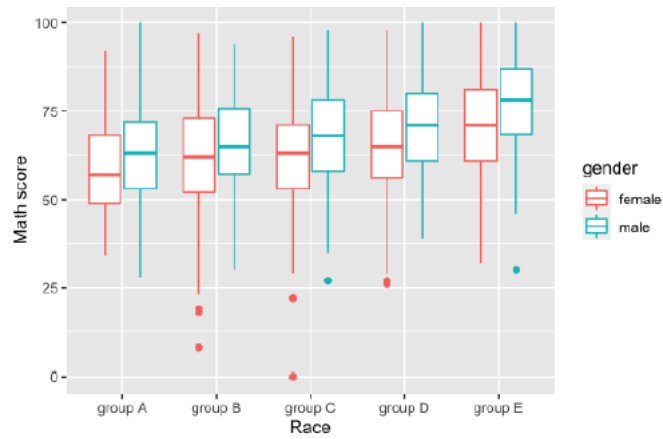
(د) حال حدسی که در قسمت آ زده بودید را با ضریب هم‌بستگی و ارتباط متغیرها مقایسه کنید.

۲. در این سوال قصد داریم با انواع نمودارهای پرکاربرد آشنا شویم. از دیتاست grades برای این کار استفاده می‌کنیم. این مجموعه، شامل نمرات دروس ریاضی، writing و reading برای ۱۰۰۰ دانش‌آموز است. توجه داشته باشید که برای کشیدن نمودارها می‌بایست از کتابخانه‌ی ggplot استفاده نمایید. هم‌چنین نمودارهای شما باید همگی دارای برچسب‌های عنوان و محورهای مربوطه باشد.

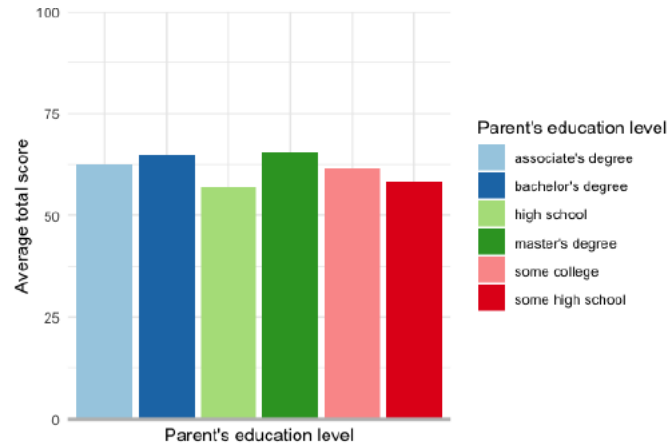
(آ) مجموعه داده را در ابتدا بخوانید، حال برای هر جنسیت، متغیر writing score را در قالب هیستوگرام نمایش دهید.

(ب) توزیع هر سه متغیر reading score، writing score و math score را با کشیدن QQ-Plot آن‌ها بررسی کنید. آیا می‌توان این برداشت را کرد که این متغیرها توزیعی مشابه با توزیع گوسی دارند؟ توضیح دهید. شکل شما باید شبیه شکل زیر باشد.





- (ج) متغیر math score را در نظر بگیرید. حال برای همه‌ی دانش‌آموزان و جنسیت‌ها و گروه‌های مختلف، به طور مجزا boxplot های آن‌ها را بکشید.
 شکل شما باید شبیه شکل زیر باشد.
- (د) نمره‌ی نهایی (میانگین تمامی نمره‌ها) را برای هر دانش‌آموز محاسبه کرده و در یک ستون ذخیره کنید و آن ستون را به مجموعه‌ی داده اضافه کنید. حال یک bar plot رسم کرده به طوری که هر مستطیل در آن نشان‌دهنده‌ی میانگین نمره دانش‌آموزانی باشد که والدین آن‌ها دارای مدرک تحصیلی یکسانی هستند.
 شکل شما باید شبیه شکل زیر باشد.



۵ توضیحات

دانشجویان عزیز حتما به نکات زیر توجه داشته باشند.

- پروژه به گونه‌ای طراحی شده که به دانش آماری فراتر از آن چه در این درس آموخته‌اید نیاز نداشته باشد و آن چه را که آموخته‌اید تثبیت و تفهیم می‌کند. به همین جهت انجام آن برای یادگیری درس اکیدا توصیه می‌شود.
- صرف نظر از رویکرد آموزشی این پروژه، آخرین نقطه‌ی جبران نمراتتان در این درس می‌باشد و بنا به سابقه‌ی چندساله، به اسکیل شدن 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 با معنا 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، ارجاع دادن 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 نمایید. از قرار دادن عکس از نوشته‌ی دست‌نویس خود در گزارش به شدت 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 طراحان را منشن کنید. سوالات در گروه پرسیده شده و همان‌جا پاسخ داده خواهند شد تا در دسترس همه‌ی دانشجویها قرار بگیرند.