



دانشکده فنی دانشگاه تهران

دانشکده برق و کامپیوتر

پروژه نهائی درس آمار و احتمال مهندسی



۱۰ خرداد ۱۳۹۸

دانشجویان عزیز، قبل از پاسخ‌گویی به سوالات به نکات زیر توجه کنید:

۱. انجام پروژه اجباری نیست و جنبه‌ی نمره‌ی امتیازی دارد. (۱ نمره)، اگرچه پروژه اجباری نیست، اما به دلیل ساده و مهم بودن مبحث، انجام آن شدیداً توصیه می‌گردد. مهلت تحویل پروژه نهایتاً تا ۱۰ تیرماه می‌باشد و به هیچ‌وجه تمدید نمی‌شود.
۲. به‌دل‌خواه خود، یکی از دو دیتاست موجود در سایت را انتخاب کنید. در صورت تمایل، می‌توانید از دیتاست دیگری نیز استفاده کنید، اما باید دیتاست را به همراه گزارش کار خود آپلود نمایید.
۳. به همراه کدهای R خود، می‌بایست گزارش کاملی از آنچه انجام دادید به همراه برداشت خود از مفاهیم جدید ارائه دهید. هفتاد درصد نمره‌ی پروژه شما مربوط به گزارش کار است و تنها ۳۰ درصد پروژه مربوط به کد می‌باشد. به گزارش‌های مشابه نمره‌ای تعلق نمی‌گیرد.
۴. تمیزی گزارش کار، واضح بودن جملات و رعایت اصول گزارش‌نویسی (مثل ارجاع مناسب به تصاویر، استفاده از پاورقی و فهرست مطالب و تصاویر و ...) بخشی از نمره‌ی گزارش هستند. خلاقیت در رسم نمودارها و تحلیل‌های جالب شما، می‌تواند نمره‌ی اضافی برای شما فراهم کند.

هدف این پروژه استفاده از مباحث و مفاهیم مطرح شده در این درس جهت تحلیل اطلاعات واقعی است.

۱. بررسی کنید که آیا در مجموعه داده‌ها متغیری وجود دارد که برای برخی از داده‌ها مقدار ندارد؟ در صورتی که جواب شما مثبت است برای هر یک از متغیرها نسبت تعداد مقادیر نامعلوم را بیابید. به نظر شما چه راهکاری برای حل مشکل این متغیرها و کارکردن با آن‌ها وجود دارد؟

۲. به دلخواه خود یک متغیر عددی را انتخاب نمایید و به سوالات زیر پاسخ دهید:

(ا) نمودار جعبه‌ای (Box Plot) را رسم کنید و مقادیر دقیق چارک‌ها (Q_1, Q_2, Q_3, Q_4) را گزارش نمایید.

(ب) هیستوگرام این متغیر را رسم نمایید.

(ج) نمودار توزیع تجمعی را برای این متغیر به دست آورید.

۳. به دلخواه خود یک متغیر عددی را انتخاب نمایید و به سوالات زیر پاسخ دهید:

(ا) از میان مقادیر مربوط به این متغیر صد نمونه به تصادف بردارید.

(ب) میانگین و واریانس و انحراف معیار این نمونه‌ها را محاسبه کنید.

(ج) یکی از ابزارهایی که برای مقایسه‌ی شهودی دو توزیع به کار می‌رود نمودار $Q-Q$ می‌باشد. توزیع نمونه‌های برداشته شده را با توزیع نرمال مقایسه کرده (با استفاده از دستور $qqnorm(x)$) و نتایج را گزارش دهید.

(د) برای میانگین به دست آمده در بند قبل بازه‌ی اطمینان ۹۵ درصد را به دست آورده و نتایج را تفسیر کنید.

(ه) برای میانگین محاسبه شده یک تست فرض دو طرفه تشکیل دهید (با استفاده از شاخص آماری t مقدار p -value را محاسبه و فرض خود را تصدیق یا رد کنید).

(و) خواسته بند قبل را با استفاده از یک تست فرض یک طرفه تکرار کنید.

۴. به دلخواه دو متغیر عددی انتخاب نمایید و به سوالات زیر پاسخ دهید:

(ا) نمودار پراکندگی نقطه‌ای (scatter plot) را برای نمایش وابستگی این دو متغیر رسم کنید و توضیح دهید که رابطه‌ی بین این دو متغیر چگونه است.

(ب) ضریب همبستگی Pearson و Spearman این دو متغیر را محاسبه کنید.

(ج) توزیع این دو متغیر را با رسم نمودار $Q-Q$ (با استفاده از دستور $qqplot(x, y)$) مقایسه کنید.

۵. یکی از توزیع‌های آماری توزیع پواسن (Poisson) می‌باشد. این توزیع یک پارامتر داشته که معمولاً با λ نشان داده می‌شود. اکنون به سوالات زیر پاسخ دهید:

(ا) نمودار این توزیع را به ازای چندین مقدار مختلف λ رسم نمایید.

(ب) از این توزیع (با پارامتر ۴) ۵ نمونه با سایز ۵۰ برداشته (با استفاده از دستور $rpois(n, \lambda)$) و میانگین هر یک از پنج نمونه را محاسبه نموده و رسم کنید. همین کار را برای ۱۰ و ۵۰۰ و ۵۰۰۰ نمونه با سایز ۵۰ نیز تکرار نمایید. نتایج را با استفاده از قضیه حد مرکزی تفسیر کنید.

۶. به دلخواه دو متغیر عددی انتخاب کنید و یکی را به عنوان متغیر وابسته و دیگری را به عنوان متغیر مستقل در نظر بگیرید. آنگاه به سوالات زیر پاسخ دهید:

(ا) دویست نمونه به تصادف بردارید.

(ب) خط رگرسیون توصیف‌کننده رابطه این دو متغیر را به دست آورید.

(ج) خط به دست آمده در قسمت قبل را رسم نمایید.