ENGS93: Statistical Methods in Engineering (Fall 2022)

Homework 6

Due at: 11:59 PM on Friday, Oct 28th 2022

<u>Note</u>: All sub-questions are worth 5 points each. In all problems, you are always allowed to do the basic arithmetic (i.e., addition, subtraction, multiplication, division, exponentiation and logarithms) using a calculator or any software (R, Excel, etc.).

For R questions, in addition to text answers, please also upload your R code files to Canvas.

1. This problem continues the scenario from R Exercise #3. Questions 3 and 4 from R Exercise #3 are examples of situations in which exactly one parameter value is specified under the alternate hypothesis (e.g. H_1: μ= 40). This is referred to as a ***simple hypothesis***. If a hypothesis specifies more than one value, then it is called a ***composite hypothesis*** (e.g. H_1: μ > 15). To get a complete picture of the error rates when the hypotheses are composite, it is often useful to plot the two error rates as *functions* of the true population parameter (e.g. μ in our example). Instead of dealing with these two functions separately, we can combine them into a single function called the ***operating characteristic (OC) function*** of the test. In general, if θ denotes the population parameter, then the OC function is the probability of failing to reject H_0 as a function of θ:

$$OC(\theta) = P\{\text{Test fails to reject } H_0 \mid \theta\}.$$

This implies that for θ values in H_0, the value of the OC function corresponds to $1 - \alpha$, or the specificity of the test. For θ values in H_1, the value of the OC function corresponds to β, or the Type II error rate. So the ***power function*** can be written as:

$$P(\theta) = P\{\text{Test rejects } H_0 \mid \theta\} = 1 - OC(\theta).$$

For our GRE example with *n*=20, the OC function can be written explicitly as a function of μ as:

$$\begin{aligned} OC(\mu) &= P\{\bar{X} \le 25 \mid \mu\} \\ &= P\left\{Z = \frac{\bar{X}-\mu}{40/\sqrt{20}} \le \frac{25-\mu}{40/\sqrt{20}} \mid \mu\right\} \\ &= \Phi\left(\frac{25-\mu}{40/\sqrt{20}}\right) \end{aligned}$$

where Φ is the standard normal cumulative distribution function (i.e. `pnorm()` in R).

   a. *As before, the organization again sets up the hypotheses as: H_0: $\mu \le 15$ (coaching program is ineffective) vs. H_1: $\mu > 15$ (coaching program is effective). Let sample size be n=20. Plot the OC curve for ETS's coaching effectiveness test using the function* `curve()`. *Plot it between* `mu0–4*se0` *and* `mu1+4*se0`. *Be sure to label your axes appropriately as* `xlab="mu"` *and* `ylab="OC(mu)"`. *Add vertical lines to your plot at μ =15 (no coaching effect), μ =25 (moderate coaching effect), and μ =40 (large coaching effect). How do you interpret the values* *of the curve at these points? In which cases do you interpret the OC curve in terms of α, and in which cases do you interpret it in terms of β?*

A good statistical test has small Type I and Type II error rates; i.e. its OC function falls steeply as the parameter values change from H_0 to H_1. A test with this property is said to strongly **discriminate** between H_0 and H_1. Due to sampling variation, no statistical test can be perfectly discriminating.

b. *How does the OC curve change with increasing sample size? To determine this, add modified OC curves to your plot using n=30, n=50, and n=70. Describe how OC curve changes with increasing sample size.*
c. *How would the OC curve change if ETS chose a rejection region of $\bar{X} > 35$? To determine this, repeat the plotting steps of Part a of this question making the appropriate substitution (keeping n=20). Compare the values of the curve at μ =15, μ =25, and μ =40 with those you obtained in Part a of this question. Explain why this happens.*

2. An airline passenger's value of time (VOT) is measured as the amount of money a passenger is willing to pay to avoid being delayed by an additional minute. A passenger is classified as either a business-passenger or as a leisure-passenger based on the purpose of travel. Economic theory suggests that the value of time should be higher for business-passengers than for the leisure-passengers, in general. It is well-known from prior studies on the variability in the passengers' socio-economic statuses that the standard deviation in VOT for business-passengers is approximately 0.25 $/min and for leisure-passengers, it is approximately 0.15 $/min. Using the results from a survey of US domestic airline passengers involving a sample of 50 business-passengers and 50 leisure-passengers, you tested the null hypothesis that the mean VOT for business-passengers (μ_b) is less than or equal to that for the leisure-passengers (μ_l), i.e., $H_0: \mu_b \leq \mu_l$ against the appropriate alternate hypothesis. Assume that, in fact, the true difference is $\mu_b - \mu_l = 0.1$, and use this as the specific alternate hypothesis.
   a. What is the power of the hypothesis test made at the 5% significance level?
   b. If we decide to increase the survey sample size for the business-passengers (owing to their greater variance in VOT) while keeping the sample size of the leisure-passengers constant, then what is the minimum sample size for the business-passengers that is needed to have a power of 95% for a 5% significance level test?
   c. For the original sample sizes (50 passengers of each type) stated in the problem, what is the significance level that needs to be used in order to get a power of 90%?
   d. For the original sample sizes (50 passengers of each type) stated in the problem, and the rejection region of $\overline{X_b} - \overline{X_l} \geq 0.075$ (where $\overline{X_b}$ and $\overline{X_l}$ are the sample means of the VOT for business- and leisure-passengers respectively), what is the power of the test?
3. All else being equal, one would expect the energy consumption to be related to the amount of CO_2 emissions. The energy consumption of buildings per unit area per unit time is measured as energy use intensity given in MJ/ft^2/year. The CO_2 emissions are measured in metric tons per capita per year (T/person/year). The data on the building energy consumption and CO_2 emissions values for a sample of 15 zip-codes is posted as an Excel file on Canvas.
   a. Find the 95% confidence interval for the correlation between emissions and energy consumption.
   b. Find the p-value for testing $H_0: \rho \leq 0.2$ versus $H_1: \rho > 0.2$.
   c. If the units for energy consumption were changed from MJ/ft^2/year to BTU/m^2/year, what is the new 95% confidence interval? Note that 1 MJ is approximately 947.8 BTU and 1 foot is approximately equal to 0.3048 m.

d. Compute the least-squares line for predicting the emissions from energy consumption. What are the units of the estimated slope? What are the units of the estimated intercept?
e. Which point has the largest magnitude of the residual?
f. Report the Total Sum of Squares (SST), error sum of squares (SSE), and regression sum of squares (SSR). What proportion of the variation in emissions is explained by energy consumption?
g. If the energy consumption increases by 1 MJ/ft^2/year, by how much would you predict the emissions to increase or decrease?
h. Compute a 95% confidence interval for the slope.
i. Compute the 95% confidence interval for the mean emissions at an energy use intensity of 145 MJ/ft^2/year.
j. Compute the 95% prediction interval for the mean emissions of a particular zip-code with an energy use intensity of 145 MJ/ft^2/year.
k. If the 95% prediction intervals are calculated at each integer value of energy use intensity (in MJ/ft^2/year) from 100 to 150, which will be the narrowest? Explain why. Also, calculate its width at the narrowest point.