BME 423 Fall 2021: Homework 3

*Due **by noon on Friday, November 12th**, 2021. Please read the homework guidelines before starting your assignment. Statistical tables necessary for completing this assignment are posted on Blackboard.*

1. Confidence Intervals [20 points]

(a) Radiologists follow specific reading protocols while assessing digital mammograms. A new reading protocol called 'CARP' (Computer Assisted Reading Protocol) was developed to shorten the time taken to read a case by highlighting computationally-detected abnormal regions. A study was conducted to investigate improvements in reading time using CARP. A radiologist read 250 (n=250) mammograms using the conventional reading protocol and 250 mammograms using CARP. The mean reading time using the conventional reading protocol was 68.11 seconds, with a standard deviation of 1.77 seconds. The mean reading time using the CARP system was 63.89 seconds, with a standard deviation of 1.54 seconds. Find the 95% confidence interval of the difference in mean reading time between the conventional reading protocol and CARP. Is this difference statistically significant at a level of significance of 0.05?

This data has been adapted from Moin P, Deshpande R, Sayre J, et al., 'An Observer Study for a Computer-Aided Reading Protocol (CARP) in the Screening Environment for Digital Mammography', Acad Radiol., 18(11): 1420-9; 2011

(b) In a clinical trial, patients with hypertension were randomly assigned to either receive a placebo or a new blood-pressure medicine. At the end of the trial, the patients were reassessed and the incidence of hypertension in the two treatment groups is summarized in the table below:

Group	Hypertension	No Hypertension
Drug	52	192
Placebo	77	185

(i) What is the 95% confidence interval for the difference in the proportion of patients with hypertension for the two treatment groups?

(ii) What is the 99% confidence interval for the difference in the proportion of patients with hypertension for the two treatment groups?

(iii) Using only your calculations for parts (i) and (ii), is the difference in the proportion of patients with hypertension for the two treatment groups statistically significant with $P<0.05$? With $P<0.01$? Please state how you reached your conclusion.

2. Regression [20 points]

In a study looking at mortality due to skin cancer in relation to latitude (a stand-in for exposure to high-intensity sunlight; higher latitudes get less-intense sunlight), data were collected for mortality due to skin cancer (number of deaths per 10 million people; the dependent variable) and the latitude of the center of the associated region (degrees North; the independent variable).

The data are listed in the table below:

Latitude	Mortality	Latitude	Mortality	Latitude	Mortality
33	219	45.2	117	35.5	182
34.5	160	39	162	44	136
35	170	42.2	143	40.8	132
37.5	182	43.5	117	41.8	137
39	149	46	116	33.8	178
41.8	159	32.8	207	44.8	86
39	200	38.5	131	36	186
39	177	47	109	31.5	229
28	197	41.5	122	39.5	142
33	214	39	191	44	153
44.5	116	43.8	129	37.5	166
40	124	40.2	159	47.5	117
40.2	128	35	141	38.8	136
42.2	128	43	152	44.5	110
38.5	166	35.5	199	43	134
37.8	147	47.5	115		
31.2	190	40.2	131		

(i) Find the linear regression relationship between the latitude (independent variable) and the mortality rate due to skin cancer (dependent variable). *Note: leave the variables in the units given (°North and deaths per 10 million people)*

(ii) Calculate the standard errors of the regression coefficients a and b (in $\hat{y} = a + bx$) as well as the estimate of the variance of the line of means.

(iii) Based on your calculations for b, is the trend between mortality due to skin cancer and latitude statistically significant?

(iv) What is the 95% Confidence Interval for the mean of skin cancer mortality at 45° North?

(v) What is the 95% Confidence Interval for an observation of skin cancer mortality at 30° North?

(vi) Calculate the correlation (Pearson coefficient) between skin cancer mortality and latitude.

3. Multiple Linear Regression [20 points]

Use R to perform a multiple linear regression on the data provided in "Hwk3_Hospital_Infection_data.csv". Infection risk ("InfctRsk") is the dependent variable; length of stay ("Stay"), age of subject ("Age"), and number of X-Rays performed at the hospital ("Xray") are the independent variables, which we are interested in investigating to see if they can explain infection risk. (Note: the data file also contains other variables that will not be used in this exercise.)

a) Estimate a constant (intercept term), as well as the coefficients for all three explanatory variables noted above, and provide the final regression line equation. Show the complete results from the R **summary()** function.
b) From the results given, show how you would reject the null hypothesis of no relation between infection risk and any of the variables above.
c) Based on your results, which explanatory variables are statistically significant? Provide the null and alternate hypotheses you are testing and summarize the test using the R results for each variable.
d) Using an R function (show results), provide a 95% confidence interval for each of the three coefficients. Based on these confidence intervals, make an inference about the significance of each variable. Do these results agree with your conclusions in part c)?
e) If not all of the explanatory variables had statistically significant coefficients, re-do your analysis with only those that were statistically significant. Using the resulting regression equation (or your original equation, if all coefficients were statistically significant), estimate the mean infection risk of a 55-year-old subject who stays for 12 days at a hospital that performs 85 X-Rays.

4. Multiple Treatments [10 points]

A treatment for depression was tested using a repeated-measures study design. Before treatment, depressed subjects were evaluated using a testing instrument that evaluates overall mood (with lower scores being associated with depression). After treatment, the evaluation was performed again.

Based on the data below, is there evidence that this treatment affected the subjects' mood scores? You can assume that any quantities of interest are Normally distributed.

Patient	Mood Score Before Treatment	Mood Score After Treatment
1	66.79	64.55
2	54.81	58.69
3	58.47	57.88
4	83.55	92.07
5	63.25	67.33
6	61.56	67.68
7	86.58	91.29
8	43.52	49.85
9	77.54	74.39
10	73.03	75.93
11	53.61	56.48

5. Multiple Treatments [20 points]

A group of researchers wanted to test the effects of an erythropoietin-like drug that was intended to increase the amount of red blood cells in the blood. It was hypothesized that the drug might continue to increase the subjects' hematocrit six months or more after treatment, so the subjects were evaluated at three time points: before treatment, 30 days after treatment (a typical timeframe for the action of erythropoietin), and 180 days after treatment.

Based on the data in the table below, is there a statistically significant difference in hematocrit among these three time points? If you find that there is a difference among the three time points, identify/isolate which time points differ. You can assume the data are Normally distributed.

Subject	Hematocrit (%)		
	Before Treatment	30 Days After Treatment	180 Days After Treatment
1	34	35	38
2	32	38	39
3	30	33	37
4	31	29	36
5	32	36	39
6	31	33	38
7	34	39	38
8	33	33	35
9	35	38	36
10	30	36	35
11	32	32	34
12	31	30	35