STAT502 Homework #1

due Monday, 7/3

1. (*adapted from KNNL 16.8*) The data in “paper.txt” are from an experiment investigating the effect of paper color on response rates for a questionnaire.

   (a) Create a box plot of the data separating the results by color. Does it appear there are differences in the response rates?

   (b) Calculate the mean and standard deviation of the response rate for each of the three colors. Also calculate the pooled sample standard deviation.

   (c) Denoting the population means of the blue and green groups by μ_1 and μ_2, respectively, carry out the test of $H_0 : \mu_1 - \mu_2 = 0$ versus the two-sided alternative. Report the degrees of freedom, test statistic, and conclusion with $\alpha = .05$.

   (d) Estimate $\mu_1 - \mu_2$ with 95% confidence.

   (e) What assumptions are you making for parts (c) and (d)?

2. Download the file “CH01PR19.txt” from Canvas. Put this in your STAT502 folder, and set the directory in R to this folder. The goal in this problem is to study whether a student’s Y = GPA at the end of the freshman year (the first column) can be predicted from x = ACT test score (the second column). Use the following to fit the regression model in R:

```
data=read.table("CH01PR19.txt")
y=data[,1]
x=data[,2]
fit=lm(y~x)
summary(fit)
```

   (a) Locate the quantities under “Estimate” in the output. These are the estimated intercept and slope, respectively, in the regression line. Report the estimated regression line in the form $\hat{y} = b_0 + b_1 x$.

   (b) Locate the quantity “Residual standard error” in the output. We denoted this by s in class. How is this value interpreted?

   (c) What is the estimate of the expected change in GPA when the entrance test score increases by one point? Answer this with a 90% confidence interval. Use `qt(.95,118)` to find the 90% multiplier. Where are the values .95 and 118 coming from?

   (d) What is the estimate of the expected GPA for students with ACT test score 30?

   (e) Use `plot(x,y,xlab=’ACT’,ylab=’GPA’);abline(fit)` to plot the estimated regression function and the data. You can save this plot by right-clicking on the plot window and following File → Save as

3. In this part, we continue with the data from above, but instead of working with exact ACT scores, we consider values less than or equal to 25 “low” and values greater than 25 “high”. Thus, we have two groups for comparison. To make this work with the regression model, we define the indicator variable

$$x_1 = \begin{cases} 1 & \text{if ACT} > 25 \\ 0 & \text{if ACT} \leq 25 \end{cases}$$

and fit the regression model $E(Y) = \beta_0 + \beta_1 x_1$.

(a) For students with “low” ACT scores, what is the expected (population mean) GPA? For those students with “high” ACT scores, what is the expected GPA? What is the difference between these expected values?

(b) Use the commands `x1=1*(x>25);x1` to create the indicator variable above. Why do we multiply `(x>25)` by 1?

(c) Fit the regression model with Y and x_1, and summarize the output. What is the estimated difference between the expected GPAs? What is the standard error for this estimate? What are the degrees of freedom?

(d) Use the output above to test whether the population mean GPAs differ when comparing freshmen with “low” ACT scores with freshmen with “high” ACT scores. Use $\alpha = .05$.

(e) The test above should remind you of the two-sample t-test. Is it equivalent? Use the commands `low=y[x<=25];high=y[x>25]` to create separate GPA groups, depending on ACT score. What is the bracket syntax doing here?

(f) What is the difference in means for the “low” and “high” groups? What are the degrees of freedom for the t-test? How do these values compare with those above in (c)?

(g) Use the `t.test()` function with the equal variance assumption to test for equal population mean GPAs between the “low” and “high” ACT students. Do you get the same results as in (d)? Why or why not?