Stat261 - Spring 2018
Assignment #5 - due Thursday, April 5, 2018 in class

The files: Stat261-Assign5-R-2018.pdf and StudentPerformanceData.pdf are required to complete this assignment. The Stat261-Assign5-R-2018.Rmd contains the code that generated Stat261-Assign5-R-2018.pdf. Section numbers in the assignment questions refer to the sections in Stat261-Assign5-R-2018.pdf.

We will use the variables **G3, sex,** and **Walc** in this data set. Look up the definitions of these variables in the file StudentPerformanceData.pdf.

1. Using the information (summary and graphs) in Section 1.1, comment on the variables **G3, sex,** and **Walc**.

2. Section 1.2 contains an analysis of the first 10 grade observations.

   (a) Comment on the distribution of these 10 observations given Figures 3 and 4.

   (b) What is a 95% confidence interval for the mean of the 10 observations?

   (c) Using the 10 observations, perform a test of the hypothesis that the mean grade is 10. Include a concluding sentence which could be incorporated into a report about this data to your boss. Hint: the required computed quantities are given in this section.

3. We are interested in comparing the grades for boys and girls in Section 1.3.

   (a) Comment on Figure 5.

   (b) Assuming that the variances for the boys and girls are equal, perform a test of the hypothesis that the mean grade for boys is the same as the mean grade for girls. Include a concluding sentence which could be incorporated into a report about this data to your boss.

   (c) Without assuming that the variances for the boys and girls are equal, perform a test of the hypothesis that the mean grade for boys is the same as the mean grade for girls. Include a concluding sentence which could be incorporated into a report about this data to your boss.

   (d) Were your conclusions for the two tests above the same? Why do you suppose that is?

4. In Section 1.4, we are interested in whether there is a relationship between student grades and **Walc**, weekend alcohol consumption.

   (a) Comment on Figure 6 and 7.

   (b) The results from fitting a straight line model to the grades as a function of weekend alcohol consumption are given on page 10. What is the estimated straight line model for this data?

   (c) Perform a test of the hypothesis that the slope parameter is zero. Include a concluding sentence which could be incorporated into a report about this data to your boss.

(d) Figure 8 is a qqplot of the residuals from the straight line model fit. Comment on this plot.

(e) Figure 10 contains the side-by-side boxplots of the grades by **Walc**, with the fitted line drawn on top. Comment on this graph.

BONUS QUESTIONS:

1. (BONUS) Suppose that $Y_1, Y_2, ..., Y_n$ are independent $N(\alpha, \sigma^2)$. Show that if σ is unknown, the likelihood ratio statistic for testing $H_0 : \alpha = \alpha_0$ is given by:

$$D = n \ln \left[1 + \frac{1}{n-1} T^2 \right], \text{ where}$$

$$T = \frac{\hat{\alpha} - \alpha_0}{s/\sqrt{n}}.$$

2. (BONUS) Testing equality of variances. Consider k independent normal samples of sizes $n_1, n_2, ..., n_k$. Measurements from sample i have unknown variance σ_i^2. Let $s_1^2, s_2^2, ..., s_k^2$ be the sample variances computed from the sample data which are estimates of $\sigma_1^2, \sigma_2^2, ...\sigma_k^2$. Since the measurements are normally distributed, we know that.

$$(n_i - 1)s_i^2/\sigma_i^2 \sim \chi^2_{(n_i-1)} \text{ for } i = 1, 2, ..., k.$$

Using the above distribution, the log likelihood for σ_i is therefore:

$$\ell(\sigma_i) = -(n_i - 1) \ln \sigma_i \ - (n_i - 1)s_i^2/(2\sigma_i^2).$$

(a) Find the joint log likelihood function of $\sigma_1, \sigma_2, ..., \sigma_k$ and show that it is maximized for $\hat{\sigma_i^2} = s_i^2, i = 1, 2, ..., k$.

(b) Show that if $\sigma_1 = \sigma_2 = ... = \sigma_k = \sigma$, then the MLE of σ^2 is given by,

$$s^2_{pooled} = \left(\sum_{i=1}^{k} (n_i - 1)s_i^2 \right) / \left(\sum_{i=1}^{k} (n_i - 1) \right).$$

(c) Show that the likelihood ratio statistic for testing $H_0 : \sigma_1 = \sigma_2 = ... = \sigma_k = \sigma$ is given by

$$D = \sum_{i=1}^{k} (n_i - 1) \ln(s^2_{pooled}/s_i^2)$$

.