Stat 704, Homework 1

1. Let
\[ Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2), \]

independent of
\[ Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2). \]

Let \( \bar{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1i} \) and \( \bar{Y}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{2i} \) be the sample means from the two populations.

(a) Find \( E(\bar{Y}_1 - \bar{Y}_2) \).
(b) Find \( \text{var}(\bar{Y}_1 - \bar{Y}_2) \).
(c) What is the distribution of \( \bar{Y}_1 - \bar{Y}_2 \)? Hint: first find the distributions of \( \bar{Y}_1 \) and \( \bar{Y}_2 \) and argue that these two random variables are independent.

2. Let \( Y_1, Y_2, \) and \( Y_3 \) be independent random variables with means \( E(Y_i) = \mu_i \) for \( i = 1, 2, 3 \) and common variance \( \text{var}(Y_i) = \sigma^2 \). Define \( \bar{Y} = \frac{1}{3}(Y_1 + Y_2 + Y_3) \).

(a) Find \( \text{cov}(Y_1 - \bar{Y}, \bar{Y}) \).
(b) Find \( E\{(Y_1 + 2Y_2 - Y_3)^2\} \).

3. A random sample of 796 teenagers revealed that in this sample, the mean number of hours per week of TV watching was \( \bar{y} = 13.2 \), with a standard deviation of \( s = 1.6 \). Find and interpret a 95% confidence interval for the true mean weekly TV-watching time for teenagers. Why can we use a t CI procedure in this problem?

4. An engineer wants to calibrate a pH meter. She uses the meter to measure the pH in 14 neutral substances (pH = 7.0), obtaining the following data: 6.986, 7.009, 7.028, 7.037, 7.028, 7.009, 7.053, 7.028, 7.011, 7.021, 7.037, 7.070, 7.058, 7.013.

(a) Use a boxplot and Q-Q plot to determine whether the assumption of normality for these data is reasonable.
(b) Test at \( \alpha = 0.05 \) whether the true mean pH reading for neutral substances differs from 7.0. Use SAS and report the p-value of your test.
5. Suppose a sample of 10 types of compact cars reveals the following one-day rental prices (in dollars) for Hertz and Thrifty, respectively:

<table>
<thead>
<tr>
<th>Car Type</th>
<th>Renters</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hertz</td>
<td></td>
<td>37.16</td>
<td>14.36</td>
<td>17.59</td>
<td>19.73</td>
<td>30.77</td>
<td>26.29</td>
<td>30.03</td>
<td>29.02</td>
<td>22.63</td>
<td>39.21</td>
</tr>
<tr>
<td>Thrifty</td>
<td></td>
<td>29.49</td>
<td>12.19</td>
<td>15.07</td>
<td>15.17</td>
<td>24.52</td>
<td>22.32</td>
<td>25.30</td>
<td>22.74</td>
<td>19.35</td>
<td>34.44</td>
</tr>
</tbody>
</table>

(a) Explain why this is a paired-sample problem.

(b) Use a graph to determine whether the assumption of normality is reasonable.

(c) Using a p-value, test at $\alpha = 0.05$ whether Thrifty has a lower true mean rental rate than Hertz via a t-test.

6. Examine the data in Problem 16.7 on page 723 of your textbook. We will only deal with the data on the first two lines (“Low” and “Moderate”).

(a) Use a SAS procedure to prepare side-by-side box plots for the two samples. Do the spreads seem to differ across samples?

(b) Using a p-value from a t-test, test at $\alpha = 0.05$ whether the firms rated “Moderate” have a significantly higher mean productivity improvement than those rated “Low”.

(c) From SAS, obtain the p-value from the folded F-test of $H_0 : \sigma_1 = \sigma_2$. What do you conclude here?

(d) Obtain and interpret a 95% CI for the difference in mean productivity improvement between firms rated “Moderate” and those rated “Low”.

(e) Comment on the standard diagnostic plots (given by SAS graphics).