FINAL EXAM

Please show your work in all the problems to receive full credit. Please note that this exam is double-sided.

**Problem 1:** Biologists, researching the effects of adding limestone sand as buffer for acid rain effects in streams, monitored the pH levels of two streams each month for 36 months. The first stream had a mean pH level of 6.8 with a standard deviation of 2.3. The control stream had a mean pH level of 9.2 with a standard deviation of 1.5. Assume a .05 significance level for testing the claim that the mean pH of the first stream was less (more acidic) than the mean pH of the control stream. Also, assume the two samples are independent simple random samples selected from normally distributed populations.

1. [2 points] Identify the Null Hypothesis and the Alternative hypothesis and use the proper notation to write them in a symbolic form.
   
   
   \[ H_0: \mu_1 = \mu_2 \quad \text{(Claim)} \quad H_1: \mu_1 < \mu_2 \]  
   
   (Population mean pH 1st stream)

2. [5 points] Use a 5% significance level to test the given claim, assuming that the population standard deviations are not the same.

   \[ t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{6.8 - 9.2}{\sqrt{\frac{2.3^2}{36} + \frac{1.5^2}{36}}} \]

   Critical value: \( t \) with 35 df \( t_{0.05} = -1.69 \)

   Since \( t \leq -1.69 \)
   
   Reject \( H_0 \)  
   Support claim

3. [5 points] Build a 90% confidence interval for the difference between the two means.

   \[ E = t_{0.05} x \sqrt{\frac{s_1^2 + s_2^2}{n_1 + n_2}} = 1.69 x 0.457251 = 0.77343 \]

   \[ 90\% \ CI \quad \bar{x}_1 - \bar{x}_2 - E < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + E \]

   \[ -3.47343 < \mu_1 - \mu_2 < -1.6286 \]

4. [3 points] What does the confidence interval suggest about the difference between the two means?

   The CI has negative limits. We are 90% confident that:

   \[ \mu_1 - \mu_2 < 0 \Rightarrow \mu_1 < \mu_2 \]

   mean pH 1st stream < mean pH 2nd stream.
**Problem 2:** Researchers studying sleep loss followed the length of sleep, in hours, of 8 individuals with insomnia before and after cognitive behavioral therapy (CBT). Assume a .05 significance level to test the claim that there is a difference between the length of sleep of individuals before and after CBT. Also, assume the data consist of matched pairs, the samples are simple random samples, and the pairs of values are from a population having a distribution that is approximately normal.

<table>
<thead>
<tr>
<th>Individual</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before CBT</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>After CBT</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>$\bar{d}$</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

(1) [2 points] Find $\bar{d}$ of the length of sleep.

$$\bar{d} = 2.25$$

(2) [3 points] Find $s_d$ of the length of sleep.

$$s_d = 1.035098$$

(3) [2 points] Identify the Null hypothesis and the Alternative hypothesis.

$H_0$: $Md=0$ (There is no sig. dif.)

$H_1$: $Md \neq 0$ (There is sig. dif.)

(4) [3 points] Find the t test statistic.

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = 6.14842$$

(5) [2 points] Find the critical values.

$$t_{0.025, df} = 2.365$$

(6) [3 points] What do you conclude from this hypothesis test?

Since $t > 2.365$, Reject $H_0$

There is enough evidence in the sample to support the claim that there is a difference in the length of sleep before and after CBT.
Problem 3: The following scores represent a nurse’s assessment ($X$) and a physician’s assessment ($Y$) of the condition of 8 patients at time of admission to a trauma center.

<table>
<thead>
<tr>
<th>X</th>
<th>18</th>
<th>13</th>
<th>18</th>
<th>15</th>
<th>10</th>
<th>12</th>
<th>8</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>23</td>
<td>20</td>
<td>18</td>
<td>16</td>
<td>14</td>
<td>11</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

You can use the following results: $\sum X = 98$, $\sum Y = 119$, $\sum X^2 = 1366$, $\sum Y^2 = 1975$, $\sum XY = 1618$

(1) [2 points] Construct a Scatter plot for this data.

(2) [3 points] State the hypotheses and make the test for a significant linear relationship between $X$ and $Y$. Be sure to define your notation.

$H_0: \rho = 0$  \hspace{1cm} $H_1: \rho \neq 0$

$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}} = 0.870274$

$t = \frac{r}{\sqrt{1 - r^2}} = 4.3277$

Critical Value: $t_{0.025 (6 df)} = 2.447$. Reject $H_0$ since $t > 2.447$

(3) [2 points] What do you conclude from this hypothesis test?

There is a significant linear relationship between $X$ and $Y$.

(4) [4 points] Find the equation of the regression line.

$b_1 = 0.9683$  \hspace{1cm} $\hat{Y} = 30.136 + 0.9683X$

$b_0 = 30.136$

(5) [2 points] Interpret the value of the fitted slope.

For each point in the nurse assessment score, there is a 0.9683 increase in the physician assessment score.

(6) [2 points] What is the best predicted value of $Y$ for a value of $X = 17$?

$\hat{Y} = 30.136 + 0.9683 \times 17 = 19.4747$

(7) [3 points] Find the coefficient of determination. How good is this fit?

$R^2 = 0.7573718$

75.7% of the variation in $Y$ is explained by the regression line.
(8) [2 points] Find the residual values $y - \hat{y}$ for the values of $X = 4$ and $X = 10$.

\[ X = 4 \quad y - \hat{y} = 7 - 6.8867 = 0.1132931 \]
\[ X = 10 \quad y - \hat{y} = 14 - 12.6963 = 1.303625 \]

Problem 4: The flu season in Southern Nevada for year 2005 – 2006 ran from December to April, the coldest month of the year. Use a 5% significance level to test the claim that the number of flu cases are equally distributed among the five flu season months.

<table>
<thead>
<tr>
<th>Month</th>
<th>Dec 05</th>
<th>Jan 06</th>
<th>Feb 06</th>
<th>Mar 06</th>
<th>Apr 06</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Cases</td>
<td>62</td>
<td>84</td>
<td>17</td>
<td>16</td>
<td>21</td>
<td>200</td>
</tr>
</tbody>
</table>

(1) [4 points] Identify the Null and Alternative hypothesis?

\[ H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.2 \quad \text{(claim)} \]

\[ H_a : \text{At least one proportion is different from 0.2} \]

(2) [6 points] What is the expected frequency of flu cases for each of the five months?

\[ E_1 = E_2 = E_3 = E_4 = E_5 = \frac{200 \times 0.2}{5} = 40 \text{ Cases} \]

(3) [5 points] What is the value of the test statistic?

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]
\[ = \frac{(62 - 40)^2}{40} + \frac{(84 - 40)^2}{40} + \frac{(17 - 40)^2}{40} + \frac{(16 - 40)^2}{40} + \frac{(21 - 40)^2}{40} = 97.15 \]

(4) [3 points] What is the critical value?

\[ \chi^2_{0.05} (4 \text{ d.f.}) = 9.488 \]

(5) [2 points] What do you conclude about the given claim?

We reject the null hypothesis (20 pts)

Data provide enough evidence to reject the claim that the number of flu cases are equally distributed among the five flu seasons.
**Problem 5:** Biologists studying the re-forestation of two species of tropical trees in Brazil’s Atlantic forest examined the number of trees that survived based on whether they were planted at a mine site or at a railroad embankment. Using a .05 significance level, test the claim that whether a tree survives at a mine site or at a railroad embankment is independent of the species of the tree.

<table>
<thead>
<tr>
<th></th>
<th>Planted at Mine Site</th>
<th>Planted at Railroad embankment</th>
</tr>
</thead>
<tbody>
<tr>
<td># species 1 survived</td>
<td>64 (71.6%)</td>
<td>87 (99.4%)</td>
</tr>
<tr>
<td># species 2 survived</td>
<td>45 (57.4%)</td>
<td>34 (40.0%)</td>
</tr>
<tr>
<td></td>
<td>109</td>
<td>121</td>
</tr>
</tbody>
</table>

(1) [3 points] Identify the Null and Alternative hypothesis.

\( H_0: \text{Tree survival site is independent of species} \) (claim)

\( H_1: \text{Tree survival site is not independent of the} \) species

(2) [6 points] Determine the value of the test statistic.

\[
\chi^2 = \frac{(64 - 71.6)^2}{71.6} + \frac{(45 - 37.4)^2}{37.4} \\
+ \frac{(87 - 79.4)^2}{79.4} + \frac{(34 - 40.0)^2}{40.0}
\]

\[= 4.467\]

(3) [3 points] What is the critical value?

\[
\chi^2_{0.05} \left( \frac{(2-1)\times(2-1)}{1} \right) = 3.841
\]

(4) [3 points] What do you conclude about the given claim?

Since \( \chi^2 = 4.467 > 3.841 \), we reject the null hypothesis. Data provides enough evidence to conclude that tree survival site is not independent of the species.
**Problem 6:** In the last page you have a JMP output of the multiple regression to predict the capacity to Direct Attention (CDA) in elderly subjects, from age and education level. From the JMP output identify the following:

1. [3 points] The regression equation.
   \[ y = 1.71576 - 0.15972 \times \text{Age} + 0.6735 \times \text{Ed-Level}. \]

2. [3 points] The P-value corresponding to the overall significance of the regression.
   \[ 0.0325 \]

3. [3 points] The value of the multiple coefficient of determination \( R^2 \) and the adjusted value of \( R^2 \).
   \[ R^2 = 0.364973 \quad \text{Adj-R}^2 = 0.2803 \]

4. [3 points] The explained variation and the unexplained variation.
   \[ \text{Exp.} : \sum (\hat{y} - \bar{y})^2 = 66.025 \]
   \[ \text{Un. Exp.} : \sum (y - \hat{y})^2 = 114.879 \]

5. [3 points] Is the regression equation a good equation for predicting the CDA in elderly subjects? Justify your answer.

   - The \( R^2 \) is rather low. Only 36.4973% of the variation in CDA is explained by the model.
   - The residuals have an unstructured pattern. This is good (p-value < 0.05).
   - Although the model is significant, maybe more variables need to be included to improve the model.
Response CDA

Whole Model

Summary of Fit

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
<td>0.364973</td>
</tr>
<tr>
<td>RSquare Adj</td>
<td>0.280303</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>2.767413</td>
</tr>
<tr>
<td>Mean of Response</td>
<td>-1.52444</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
<td>18</td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>66.02501</td>
<td>33.0125</td>
<td>4.3105</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>114.87864</td>
<td>7.6586</td>
<td></td>
<td>0.0332*</td>
</tr>
<tr>
<td>C. Total</td>
<td>17</td>
<td>180.90364</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimates

| Term        | Estimate  | Std Error | t Ratio | Prob>|t|  |
|-------------|-----------|-----------|---------|------|-----|
| Intercept   | 1.7157564 | 7.334951  | 0.23    | 0.8182 |
| Age         | -0.159746 | 0.084497  | -1.89   | 0.0782 |
| Ed-level    | 0.6735303 | 0.285947  | 2.36    | 0.0325*|

Residual by Predicted Plot