STATISTICS 401C

Examination 2 (100 points)

Name ____________________________

NOTE: Please show all work to obtain full credit.

1. Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water from various locations of a stream. Do the data suggest that the true average concentration of zinc in the bottom water exceeds that of surface water? The data are given below aggregated:

<table>
<thead>
<tr>
<th>Location</th>
<th>Bottom</th>
<th>Surface</th>
<th>Difference d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.430</td>
<td>.415</td>
<td>.015</td>
</tr>
<tr>
<td>2</td>
<td>.266</td>
<td>.238</td>
<td>.028</td>
</tr>
<tr>
<td>3</td>
<td>.567</td>
<td>.390</td>
<td>.177</td>
</tr>
<tr>
<td>4</td>
<td>.531</td>
<td>.610</td>
<td>-.079</td>
</tr>
<tr>
<td>5</td>
<td>.707</td>
<td>.605</td>
<td>.102</td>
</tr>
<tr>
<td>6</td>
<td>.716</td>
<td>.663</td>
<td>.053</td>
</tr>
<tr>
<td>7</td>
<td>.651</td>
<td>.632</td>
<td>.019</td>
</tr>
<tr>
<td>8</td>
<td>.589</td>
<td>.523</td>
<td>.066</td>
</tr>
<tr>
<td>9</td>
<td>.469</td>
<td>.411</td>
<td>.058</td>
</tr>
<tr>
<td>10</td>
<td>.723</td>
<td>.612</td>
<td>.111</td>
</tr>
</tbody>
</table>

\[ \sum d = .55 \]

\[ \sum d^2 = .072194 \]

Let \( \mu_d = \mu_{\text{bottom}} - \mu_{\text{surface}} \) be the mean difference in zinc concentration of the two populations and assume that the differences \( d_i \) have an approximately Normal distribution.

(a) (10) Compute a t-statistic to test the research hypothesis that the mean zinc concentration in the bottom water exceeds that of surface water. State the null and alternative hypothesis in terms of \( \mu_d \). Compute an approximate p-value and make the decision using \( \alpha = 0.05 \).

\[
\bar{d} = \frac{.55}{10} = .055 \quad \bar{d}^2 = \frac{.072194 - (.55)^2/10}{9} = .00466
\]

\[
H_0: \mu_d \leq 0 \quad \text{vs.} \quad H_a: \mu_d > 0
\]

\[
t = \frac{.055 - 0}{.068267/\sqrt{10}} = 2.548
\]

\[ p\text{-value} = P(T_9 > 2.548) \]

From t-table, \( .01 < p\text{-value} < .025 \)

Since the \( p\text{-value} < .05 \), ref. \( H_0 \) at \( \alpha = .05 \)
(b) (10) Construct a 90% confidence interval for \( \mu \). Use this interval to test the hypothesis in part (a), explaining how you arrived at your conclusion. What is the \( \alpha \)-level of this test?

\[
90\% \ C I : \quad \bar{d} \pm t_{.05,.9} \left( \frac{sd}{\sqrt{n}} \right) \\
\quad 0.055 \pm 1.833, \quad \frac{0.6827}{10} \\
\quad 0.055 \pm 0.0396 \\
(0.0154, 0.0946)
\]

Since \( \sigma \) is not inside this interval, we reject the null hypothesis at \( \alpha = .05 \)

2. Many people purchase sports utility vehicles (SUV's) because they think they are sturdier and hence safer than regular cars. However, data have indicated that the costs of repairs for SUVs are higher than for midsize cars when both vehicles are in an accident. To verify this observation, a random sample of 9 new SUVs and 8 midsize cars are tested for front impact resistance. The amounts of damage (in hundreds of dollars) to the vehicles when crashed at 20 mph head-on into a stationary barrier are recorded below:

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>Damage (in $100's)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midsize</td>
<td>11.95 15.42 14.27 11.42 18.12 10.36 13.65 20.25</td>
</tr>
</tbody>
</table>

Let \( \mu_1 \) and \( \mu_2 \) represent mean cost of repair for the SUVs and midsize vehicles, respectively.

(a) (10) State all evidence you can find to support or reject the assumption of equal population variances by examining the output above. Perform a statistical test using \( \alpha = .05 \) to verify this assumption. State the null and alternative hypotheses and the rejection region clearly.

1. Sample variance do not differ by a factor larger than 3
2. The box-plots have similar IQR (box-widths)
3. The normal prob plots are approximately parallel

Test \( \ H_0 : \sigma_1^2 = \sigma_2^2 \), \( \ H_1 : \sigma_1^2 \neq \sigma_2^2 \) at \( \alpha = .05 \)

\[
F = \frac{\bar{X}_1^2}{\bar{X}_2^2} = \frac{(4.86642)^2}{(3.39912)^2} = 2.0497
\]

R.R. \( F < F_{.025, 8, 7} \) or \( F > F_{.975, 8, 7} \)

\[
F_{.025, 8, 7} = 1.453 \quad \text{and} \quad F_{.975, 8, 7} = 4.90
\]

\( \Rightarrow \) R.R. \( F < 2.22 \) or \( F > 4.90 \)

Clearly, 2.0497 is not in the R.R. \( \Rightarrow \) we fail to rej. \( H_0 \) at \( \alpha = .05 \)
(b) (10) State and test the appropriate hypotheses to determine if the mean repair cost for SUVs was higher. Use $\alpha = 0.05$ to make your decision.

$$H_0 : \mu_1 - \mu_2 \leq 0 \quad H_a : \mu_1 - \mu_2 > 0$$

$$t_c = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{(n_1 - 1)A_r^2 + (n_2 - 1)A_s^2}{n_1 + n_2 - 2}}}$$

$$= \frac{23.682 - 11.554}{\sqrt{\frac{15}{1.5}}} = 4.245$$

$$R.R. \quad t > t_{0.025,15} \quad \text{i.e.} \quad t > 1.753$$

Since $t_c$ is in the R.R., we rej. $H_0$ at $\alpha = 0.05$

(c) (10) Construct a 90% confidence interval for $\mu_1 - \mu_2$. Test the research hypothesis that the repair cost for SUVs is larger by $1000$ on the average, using this confidence interval.

$$90\% \; C.I \; for \; \mu_1 - \mu_2 : (\bar{x}_1 - \bar{x}_2) \pm t_{0.05,15} \sqrt{\frac{A_r^2}{n_1} + \frac{A_s^2}{n_2}}$$

$$= 5.27 \pm 3.616$$

$$= (1.654, 8.886)$$

Need to test $H_0 : \mu_1 - \mu_2 \leq 10 \; vs. \; H_a : \mu_1 - \mu_2 > 10$

Since 10 is not inside the above interval, we rej. $H_0$ in favor of $H_a$, i.e. the cost of repair of SUV exceeds by $1000$.

3. The following data are obtained from a chemical process where the yield ($y$, gm.) of the process is thought to be related to the reaction temperature ($x$, $^\circ C$):

<table>
<thead>
<tr>
<th>x</th>
<th>50</th>
<th>53</th>
<th>54</th>
<th>55</th>
<th>56</th>
<th>59</th>
<th>62</th>
<th>65</th>
<th>71</th>
<th>72</th>
<th>74</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>122</td>
<td>118</td>
<td>128</td>
<td>122</td>
<td>125</td>
<td>136</td>
<td>144</td>
<td>142</td>
<td>149</td>
<td>161</td>
<td>167</td>
<td>169</td>
</tr>
<tr>
<td>x(cont.)</td>
<td>76</td>
<td>79</td>
<td>80</td>
<td>82</td>
<td>85</td>
<td>87</td>
<td>90</td>
<td>93</td>
<td>94</td>
<td>95</td>
<td>97</td>
<td>100</td>
</tr>
<tr>
<td>y(cont.)</td>
<td>171</td>
<td>175</td>
<td>182</td>
<td>180</td>
<td>183</td>
<td>188</td>
<td>200</td>
<td>190</td>
<td>206</td>
<td>207</td>
<td>210</td>
<td>219</td>
</tr>
</tbody>
</table>
Some summary statistics are:

\[ n = 25 \quad \Sigma x_i = 1871 \quad \Sigma y_i = 4156 \]
\[ \Sigma x_i^2 = 145705 \quad \Sigma y_i^2 = 713582 \quad \Sigma x_i y_i = 322273 \]

(a) (8) Fit a linear regression \( y = \beta_0 + \beta_1 x + \epsilon \) of the yield in this chemical process (y) on temperature (x) using the above data. (Do the intermediate computations accurately.) Report the estimates of \( \beta_0 \) and \( \beta_1 \). According to this model, what is the average increase in yield (in gms) if the temperature is increased by 10 °C?

\[
\bar{x} = \frac{1871}{25} = 74.84 \quad \bar{y} = \frac{4156}{25} = 166.24
\]
\[
S_{xx} = \frac{\Sigma x^2 - (\Sigma x)^2}{n} = 145705 - \left(\frac{1871}{25}\right)^2 = 5679.36
\]
\[
S_{xy} = \frac{\Sigma xy - \Sigma x \Sigma y}{n} = 322273 - \frac{(1871)(4156)}{25} = 11237.96
\]
\[
\hat{\beta}_1 = \frac{11237.96}{5679.36} = 1.978737
\]
\[
\hat{\beta}_0 = 166.24 + (1.978737)(74.84) = 18.15132
\]

\[
\text{Prediction equation} \quad \hat{y} = 18.15132 + 1.978737x
\]

Thus, average increase in yield for 1°C increase in temp. = 19.78737 gms

For 10°C increase in temp. = 197.874 gms

(b) (7) Compute an analysis of variance table for the regression. What is the estimate of the error variance \( \sigma^2 \)? Use the F-statistic from the analysis of variance table to test \( H_0 : \beta_1 = 0 \) vs. \( H_a : \beta_1 \neq 0 \). Use \( \alpha = .05 \) and show work.

\[
S_{xy} = \frac{713582 - (4156)^2}{25} = 22688.56 = \text{SSTot}
\]
\[
S_{SSReg} = \frac{S_{xy}}{S_{xx}} = \frac{(11237.96)^2}{5679.36} = 22236.9671
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22236.96</td>
<td>22236.96</td>
<td>1132.55</td>
</tr>
<tr>
<td>Error</td>
<td>23</td>
<td>451.59</td>
<td>19.63</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>22688.56</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
R. R.: \frac{F}{F_{0.05, 1, 23}} = 4.26 \Rightarrow \text{Reg. H is rejected at } F_{0.05, 1, 23}
\]

\[
\text{Estimate of } \sigma^2 \text{ is: } \lambda_e = MS_E = 19.63
\]
(c) Compute the estimated standard error of $\hat{\beta}_1$. Use it to compute a 95% confidence interval for $\beta_1$. Use this interval to test whether the slope is positive.

$$
\hat{\sigma} = \frac{S_{xx}}{n-2} = \frac{19.63}{15679.36} = 0.0588
$$

$$
95\% \ CI: \hat{\beta}_1 \pm t_{0.025, 25} \cdot \hat{\sigma} = 1.978737 \pm (2.069)(0.0588)
$$

$$(1.857, 2.1)$$

Need to test $H_0: \beta_1 \leq 0$ vs. $H_a: \beta_1 > 0$

Since 0 is not in the CI, we reject $H_0: \beta_1 \leq 0$, i.e. $\beta_1 > 0$

(d) Use the numbers in the analysis of variance table to calculate a statistic that measures how well your model will predict the yield based on the temperature. Explain what this statistic says to the experimenter.

$$
P^2 = \frac{SS_{Reg}}{SS_{Tot}} = \frac{222.364477}{226.8856} \approx 0.98
$$

The straight line model with temperature as the explanatory variable is able to explain 98% of the variability in the yield.

(e) Estimate the mean yield (in gms.), $\mu_{70}$, for similar chemical reactions where the temperature is maintained at 70 °C. Construct a 95% confidence interval for this mean.

$$
\mu_{70} = \hat{\beta}_0 = 18.15132 + 1.978737 \times 70 = 156.6629
$$

$$
95\% \ CI \ for \ \mu_{70}:
$$

$$
\hat{\mu}_{70} \pm t_{0.025, 25} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{70} - \bar{x})^2}{S_{xx}}}
$$

$$
154.737 \pm (2.069) \cdot \sqrt{19.63 \left( \frac{1}{25} + \frac{(70 - 24.84)^2}{5679.36} \right)}
$$

$$(154.737, 158.589)$$

(f) A new process produces a mean yield of 150 gms. at 70 °C of the same chemical. Use the above interval to perform a test of appropriate hypotheses to determine if the mean yield for reactions 70 °C using the current process, is better than the new process ($\mu_{70}$ exceeds 150 gms.), stating the $\alpha$-level of the test.

Need to test $H_0: \mu_{70} \leq 150$ vs. $H_a: \mu_{70} > 150$

Since 150 is not in the above interval, we reject $H_0$ at $\alpha = 0.025$. In fact, since 150 is entirely below the interval, the new process average yield is lower than the current process.
(g) This question concerns the 3 plots attached (see next page). Identify the plot (or plots, use letters A,B,C) you would use to answer each of the questions about model assumptions stated below. Give your answer to each question justifying your answer using the plot or plot(s). An yes/no answer alone will not earn any points.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Plot(s)</th>
<th>Your answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do the errors(e's) have a normal distribution?</td>
<td>B</td>
<td>yes; the plot does not show a specific pattern (e.g. bowl shape) indicating a significant deviation from normality.</td>
</tr>
<tr>
<td>Is the variance (σ²) constant for all x's?</td>
<td>A, C</td>
<td>yes, both plots show the same pattern of points evenly distributed around the zero line as x (or y) increase; so the spread appears to show constant variance.</td>
</tr>
<tr>
<td>Is the straight line model adequate?</td>
<td>A, C</td>
<td>yes, both plots do not show a definite pattern of curvature as the sign of the residuals change in a random fashion.</td>
</tr>
</tbody>
</table>