Problem 1. [7 points] Some of the logistic regression output is given below.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>StDev</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.3049</td>
<td>0.2382</td>
<td>-1.28</td>
<td>0.200</td>
</tr>
<tr>
<td>Dioxin</td>
<td>-0.029710</td>
<td>0.005485</td>
<td>-5.42</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Test that all slopes are zero: $G = 33.277$, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

<table>
<thead>
<tr>
<th>Method</th>
<th>Chi-Square</th>
<th>DF</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>9.234</td>
<td>15</td>
<td>0.865</td>
</tr>
</tbody>
</table>

(a) [1] Explain how you can tell from the above output that increase**d** dioxin level appears to be associated with a decrease in the probability of fish eggs hatching.

The coefficient of dioxin, $-0.0297$, is negative.

(b) [2] If $\pi$ denotes the probability that a fish egg hatches in pond water with a dioxin level of 50 parts per billion, what is the estimated value of $\pi$ according to the logistic regression model? Show work.

First, find the logit or log-odds: $-0.3049 - 0.0297(50) = -1.79$

Next, exponentiate and convert from odds to a probability:

$$\frac{\exp(-1.79)}{1 + \exp(-1.79)} = \frac{.167}{1.167} = .143$$

(c) [1] The output says that $G = 33.277$. What does this statistic (and its associated p-value) tell us?

The small p-value gives evidence that dioxin is a meaningful predictor of the probability that a fish egg hatched.

(d) [1] The output says that Chi-square $= 9.234$. What does this statistic (and its associated p-value) tell us?

The large p-value indicates no evidence that this model is deficient in explaining the probability that a fish egg hatched.

(e) [2] Can you infer, based on what you know about this experiment, that increased dioxin levels cause a smaller proportion of fish to hatch? Explain.

No. This is an observational study, so causal inference is not possible. The different levels of dioxin pollution were not assigned randomly to ponds.
Problem 3. [9 points]

(a) [2] The plot allows us to assess visually two of the assumptions made by a 2-sample t test. Name these two assumptions. For each assumption, tell what this plot indicates about the validity of that assumption for this dataset.

The normality and equal variance assumptions both appear to be justified in the stem and leaf plot for this dataset.

(b) [2] State, using appropriate mathematical symbols, the correct null and alternative hypotheses for this experiment. Then express these hypotheses in words.

\[ H_0 : \mu_U = \mu_R \quad \text{and} \quad H_a : \mu_U > \mu_R \]

The null hypothesis says that the mean cholesterol for the urban population is equal to the mean cholesterol for the rural population. The alternative says that the urban mean is greater than the rural mean.

(c) [1] Below is some Minitab output. Find the missing T statistic. Show work.

Two sample T for choles

<table>
<thead>
<tr>
<th>code</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>216.9</td>
<td>39.9</td>
<td>6.0</td>
</tr>
<tr>
<td>2</td>
<td>49</td>
<td>157.0</td>
<td>31.8</td>
<td>4.5</td>
</tr>
</tbody>
</table>

95% CI for \( \mu_1 - \mu_2 \): (45.1, 74.6)

T-Test \( \mu_1 = \mu_2 \) (vs >): \( T = ?? \) \( P = 0.0000 \) DF = 92

Both use Pooled StDev = 35.9

\[ T = \frac{\bar{x}_U - \bar{x}_R}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{216.9 - 157.0}{35.9 \sqrt{\frac{1}{45} + \frac{1}{49}}} = 8.081 \]

(d) [2] It is stated that the samples are not random samples. What does this mean about the types of inferences we can or cannot make from the results of the study?

We cannot infer the results of this study to the populations in question. In other words, we cannot assume that \( \mu_U \) is truly larger than \( \mu_R \).

(e) [2] A nonparametric test (a test that doesn’t make the assumptions in part (a)) is run on this dataset in Minitab. The test statistic found by Minitab is \( W = 2983.0 \).

Give the name of this test (either the name given in the textbook or the name used by Minitab). Without actually attempting to do it, explain how to go about finding the W statistic for this dataset.

This is the rank-sum, or Mann-Whitney, test. The W statistic is found by ranking all 94 observations, then adding the ranks found in one of the groups.
Problem 4. [5 points]

(a) [2] Fill in the blanks in the ANOVA table below.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>2</td>
<td>694</td>
<td>347</td>
<td>0.58</td>
<td>0.567</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>11900</td>
<td>595</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>12594</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) [1] Find the pooled estimate of variance based on the ANOVA table above.

The pooled sample variance is the same as the mean square error (595 in this case).

(c) [2] Does the above plot indicate anything about the appropriateness of the model assumed here? Explain.

The assumption of equality of variances among the three groups is clearly inappropriate since the rightmost group has visibly a much smaller variance than the other two.
Problem 5. [5 points]

(a) [2] If \( \omega \) denotes the odds that an individual will be involved in a violent crime, give a 95% confidence interval for the odds ratio \( \frac{\omega_{\text{Abused}}}{\omega_{\text{Control}}} \). Show all work.

The estimated odds ratio \( \hat{\omega} \) is \( \frac{104 \times 608}{52 \times 801} = 1.518 \). We use the fact that its log is approximately normally distributed with mean \( \log(\omega_{A}/\omega_{C}) \) and standard error

\[
\sqrt{\frac{1}{n_A \hat{\pi}_A (1 - \hat{\pi}_A)} + \frac{1}{n_C \hat{\pi}_C (1 - \hat{\pi}_C)}} = \sqrt{\frac{905}{104 \times 801} + \frac{660}{52 \times 608}} = .178
\]

Thus, the interval for \( \log(\omega_{A}/\omega_{C}) \) is \( \log(1.518) \pm 1.96(.178) = (.0685, .7663) \). Exponentiating gives the confidence interval for \( \omega_{A}/\omega_{C} \):

\[
(1.071, 2.152)
\]

(b) [2] This is an observational study, not a randomized experiment. Explain what this means about the types of inferences we can make, and then explain why a randomized experiment would not be feasible in this case.

We may not make causal inferences. In other words, we may not claim that abuse as children CAUSES them to become involved in violent crime later.

A randomized experiment would involve randomly assigning certain children to abusive homes and others to non-abusive homes, which is clearly unethical.

(c) [1] The Minitab output below begins to compute a statistic, X. Fill in all four blanks below to tell what the statistic is called, finish computing it, and give its associated degrees of freedom.

X is a chi-squared statistic on 1 df, calculated by adding for each of the 4 cells (observed – expected)^2/expected. The missing summand, corresponding to the upper left cell, is \( (104 - 90.21)^2/90.21 = 2.108 \).

Degrees of freedom equal \( (\text{rows} - 1) \times (\text{columns} - 1) \).

\[
X = 2.108 + 0.233 + \quad \text{++++}
\]

\[
2.890 + 0.320 = 5.551 \quad \text{++++}
\]

DF = 1, P-Value = 0.018

++++

The statistic above is called chi-squared statistic

__________________________