A recent study allocated 100 patients to a study of two competing migraine headache remedies, Axert and Topamax. Each patient tried the drugs in a random order, taking each drug for 4 weeks, with a 2-week “washout” period in between. While the effectiveness of the drugs is of course of interest, the investigators were also concerned about the drugs’ negative side effects, the most common of which is nausea. The results were as follows:

<table>
<thead>
<tr>
<th></th>
<th>Nausea</th>
<th>No Nausea</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axert</td>
<td>20</td>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>Topamax</td>
<td>15</td>
<td>85</td>
<td>100</td>
</tr>
<tr>
<td>totals</td>
<td>35</td>
<td>165</td>
<td>200</td>
</tr>
</tbody>
</table>

After the trial is completed, one of the trial’s statisticians goes back to the database and creates the following alternate tabular summary:

<table>
<thead>
<tr>
<th></th>
<th>Nauseated by Axert</th>
<th>Not Nauseated by Axert</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nauseated by Topamax</td>
<td>10</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Not Nauseated by Topamax</td>
<td>10</td>
<td>75</td>
<td>85</td>
</tr>
<tr>
<td>totals</td>
<td>20</td>
<td>80</td>
<td>100</td>
</tr>
</tbody>
</table>

(6 pts) 1. Statistically speaking, which of these tables contains more useful information? Why?

The second one does, because the first contains only the marginal totals from the second! The first is deceptive, since it indicates 200 patients are present, when really there are only 100.

(5 pts) 2. Use the data in one of these tables to compute the odds ratio in favor of being nauseated by Topamax relative to Axert.

\[
\text{OR} = \frac{\frac{b}{c}}{\frac{d}{e}} = \left(\frac{5}{10}\right) = \left(\frac{1}{2}\right)
\]
3. A student health service offers a new “quick” test for a sexually transmitted disease for which the standard test, while very accurate, is also invasive, expensive and time-consuming. In a group of 50 students known to have the disease, the test correctly identifies 48; however in a group of 500 students known to be disease-free, the test flags 126 as diseased. This “quick” test has

(a) good sensitivity but poor specificity
(b) good specificity but poor sensitivity
(c) good positive predictive value but poor negative predictive value
(d) good negative predictive value but poor positive predictive value
(e) high correlation but a low \( \kappa \) statistic

4. Which of the following is a correct statement regarding simple linear regression?

(a) If the fitted slope \( \hat{b} > 0 \), then \( X \) and \( Y \) have a positive linear relationship.
(b) Least squares methods enable computationally convenient intercept and slope estimates \( \hat{a} \) and \( \hat{b} \) that are resistant to the effects of outliers. Hardly!
(c) \( \hat{Y} = \hat{a} + \hat{b}x \) offers a sensible estimate of both the true mean level and a future observation \( Y \) taken at this \( x \).
(d) We expect about 95% of the observed data points to lie within the 95% prediction bands estimated from these data. No, that would only be true for a second, independent set of data not used to obtain \( \hat{a} \) and \( \hat{b} \).
(e) none of these

5. An investigator wishing to fit a simple linear regression model to two variables \( X \) and \( Y \) notices an annoying number of high outliers in the \( Y \) variable. As a result, he transforms \( Y \) to the square root scale, and fits a linear regression relating \( \sqrt{Y} \) and \( X \). Back on the original scale, his model relating \( Y \) and \( X \) is

(a) linear
(b) quadratic
(c) exponential
(d) multiplicative
(e) none of these

6. Consider a multiple linear regression relating blood pressure (Y) to gender (\( X_1 \), coded as 0 for men, 1 for women) and smoking status (\( X_2 \), coded as 0 for never smokers, 1 for ever smokers). If the mean structure of the model is given as

\[
\mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 ,
\]

what is the average increase in blood pressure for female ever smokers relative to male ever smokers?

(a) \( \beta_2 \)
(b) \( \beta_3 \)
(c) \( \beta_2 + \beta_3 \)
(d) \( \beta_1 + \beta_3 \)
(e) none of these

\[
\begin{align*}
\text{male, ever} &= \beta_0 + \beta_2 \\
\text{female, ever} &= \beta_0 + \beta_1 + \beta_2 + \beta_3 \\
\text{difference} &= \beta_1 + \beta_3
\end{align*}
\]
7. An investigator wishes to examine the "broken heart effect," or the apparent tendency of many long-term marriage partners to die within one year of their spouses' deaths. The investigator suspects that surviving spouses who have a "supportive community" (children, grandchildren, friends, church, etc.) may live longer than those who do not. To check this, 90 recently widowed partners were classified as having a supportive community or not, and by whether they survived for one year. The data are:

<table>
<thead>
<tr>
<th>status after one year</th>
<th>supportive community</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
</tr>
<tr>
<td>alive</td>
<td>28</td>
</tr>
<tr>
<td>dead</td>
<td>11</td>
</tr>
</tbody>
</table>

\[
\begin{array}{c|c|c}
\text{status after one year} & \text{supportive community} & \text{Total} \\
\hline
\text{alive} & 28 & 48 \\
\text{dead}  & 11 & 3 \\
\hline
\text{Total} & 39 & 51 \\
\hline
\end{array}
\]

Given the null hypothesis \( H_0 \): survival is independent of supportive community, find the \( X^2 \) statistic, its degrees of freedom, and an approximate \( p \)-value.

\[
E_{ij} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}
\]

\[
\begin{array}{c|c|c|c}
\text{Survival} & \text{Support} & \text{Total} \\
\hline
\text{Alive}     & 32.93 & 48.07 & 81 \\
\text{Dead}      & 6.07  & 7.93  & 14 \\
\hline
\text{Total}    & 39    & 51    & 90 \\
\hline
\end{array}
\]

\[
X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

\[
= \frac{(28 - 32.93)^2}{32.93} + \frac{(48 - 48.07)^2}{48.07} + \frac{(11 - 6.07)^2}{6.07} + \frac{(3 - 7.93)^2}{7.93}
\]

\[
= 8.37
\]

\[
\text{df} = (r-1)(c-1) = 1 \text{ df}
\]

From Table F, \( .0025 < p \)-value < .005

8. Now suppose we obtain the following breakdown by gender:

<table>
<thead>
<tr>
<th>status after one year</th>
<th>men (widowers)</th>
<th>women (widows)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>supportive community</td>
<td>supportive community</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>alive</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>dead</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Is this an example of Simpson's Paradox? Why or why not?

\[
\text{OR}_{\text{total}} = \frac{28(3)}{11(48)} = 1.59
\]

\[
\text{OR}_{\text{men}} = \frac{9(1)}{3(2)} = 1.57
\]

\[
\text{OR}_{\text{women}} = \frac{19(2)}{8(46)} = .103
\]

So NOT Simpson's Paradox, since that occurs when **both** conditional OR's are in the opposite direction (relative to 1.0) from the total OR.
(9 pts) 9. A recent social epi study related $X$, a city’s 2002 spending on drug treatment programs (in millions of dollars), and $Y$, the number of drug-related homicides in the city in 2002. The published paper on the study (authored by Prof. May Eichal-Joakes) does not include the original data, but only the following data summaries:

$$
\bar{x} = 4.0, \quad \bar{y} = 3.0, \quad s_x^2 = s_y^2 = 8.0, \quad \text{and} \quad s_{xy}^2 = 4.0.
$$

Find the Pearson sample correlation $r$, and the regression intercept and slope $\hat{a}$ and $\hat{b}$.

$$
\begin{align*}
    r &= \frac{s_{xy}}{s_x s_y} = \frac{4.0}{\sqrt{8} \sqrt{8}} = \frac{4}{8} = \frac{1}{2} \\
    \hat{b} &= r \frac{s_y}{s_x} = \frac{1}{2} \frac{\sqrt{8}}{\sqrt{8}} = \frac{1}{2} \quad \hat{a} = \bar{y} - \hat{b} \bar{x} = 3 - \frac{1}{2} (4) = 1
\end{align*}
$$

(8 pts) 10. Now suppose Prof. Eichal-Joakes responds positively to an email request from you for the original data; they turn out to be:

<table>
<thead>
<tr>
<th>$X$</th>
<th>1 2 3 4 5 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>4 3 2 1 0 8</td>
</tr>
</tbody>
</table>

Construct a scatterplot of these data. Is the correlation you obtained in the previous problem misleading? Why or why not?

**YES** the correlation is misleading: it indicates a linear, positive relationship, but in fact without the outlier $(9, 8)$, the relationship is perfectly linear with a negative slope!

(12 pts) 11. Construct a residual plot for these data; that is, a plot of the (unstandardized) residuals $r_i$ versus $X_i$. Discuss any problems with the assumptions underlying regression analysis suggested by your plot.

$$
\hat{y}_i = \hat{a} + \hat{b} x_i = 1 + \frac{1}{2} x_i, \quad \text{and} \quad r_i = y_i - \hat{y}_i
$$

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$\hat{y}_i$</th>
<th>$r_i$</th>
<th>$r_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>1  1/2</td>
<td>2  1/2</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2  2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2  2  1/2</td>
<td>-1/2</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3  3</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>3  3  1/2</td>
<td>-3 1/2</td>
<td>10/4</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>3  3  1/2</td>
<td>-3 1/2</td>
<td>10/4</td>
</tr>
</tbody>
</table>

![Residual Plot](residual_plot.png)
(9 pts) 12. You have fit a multiple regression relating $Y$, a high school junior's grade point average (GPA), to three potential predictor variables: $X_1$, the adjusted gross income of the student's household, $X_2$, the number of unexcused absenses for the student, and $X_3$, an indicator for the number of parents or guardians living in the home (1=both parents living at home, 0=one parent living at home). The partial ANOVA table is as follows:

<table>
<thead>
<tr>
<th>source</th>
<th>df</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>3</td>
<td>240.859</td>
<td>80.286</td>
<td></td>
<td>0.0001814</td>
</tr>
<tr>
<td>error</td>
<td>6</td>
<td>11.047</td>
<td>MSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>9</td>
<td>251.906</td>
<td>27.989</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Find the values of the MSE and the $F$ statistic that are missing from the table. Also, compute $R^2$, the goodness-of-fit statistic.

$$\text{MSE} = \frac{\text{SSE}}{dFE} = \frac{11.047}{6} = 1.841$$

$$F = \frac{\text{MSM}}{\text{MSE}} = \frac{80.286}{1.841} = 43.61$$

$$R^2 = \frac{\text{SSM}}{\text{SST}} = \frac{240.859}{251.906} = 0.9561$$

So model explains 95.61% of the variability in the data.

(6 pts) 13. Now here is the coefficients table:

|          | Estimate | Std. Error | $t$ value | $Pr(T > |t|)$ |
|----------|----------|------------|-----------|-------------|
| Intercept| 5.28839  | 1.49060    | 3.548     | 0.01210     |
| $X_1$    | 0.84062  | 0.90166    | 0.932     | 0.38716     |
| $X_2$    | -3.68858 | 0.83845    | -4.399    | 0.00457     |
| $X_3$    | 0.14203  | 0.05187    | 2.738     | 0.03380     |

Find an approximate 95% CI for $\beta_1$, the coefficient of $X_1$.

$$CI = \hat{\beta}_1 \pm t^* \text{SE}_{\hat{\beta}_1}, \text{ where } t^* = \frac{t_{n-p-1, \alpha/2}}{\sqrt{\text{MSE}}} = \frac{t_{16}}{\sqrt{1.841}} = 2.447$$

$$= 0.84062 \pm 2.447(1.90166)$$

$$= \left(-1.366, 3.047\right)$$

(7 pts) 14. Suppose we wish to compare GPA among the kids in the two $X_3$ groups (with and without two parents living at home). Give the precise name of a nonparametric test procedure we could use to compare these two groups, and a few advantages and disadvantages of such a procedure relative to a more standard parametric approach.

**Precise name:** Wilcoxon Rank Sum Test (or Mann-Whitney test)

**Advantages:**
- More robust to departures from normality
- Inferences will be closer to nominal level (accurate CI's, p-values)

**Disadvantages:**
- May be a bit less powerful (wider CI's, bigger p-values)
- Can also be computationally difficult