Please complete the following problems. A calculator is allowed but no sophisticated functions will be permitted. Aside from a z-table, t-table, and $X^2$ table, and one side of an 8.5” by 11” sheet of paper, no other aids are allowed. Please show all your work and justify your answers. Round answers to 4 decimal places.
1.) The 4 scatterplots (labeled as 1,2,3,4 below) are a famous example in statistics known as Anscombe’s quartet (image source: Wikipedia).

a.) It turns out that all 4 scatterplots have the exact same coefficient of correlation. It is:
A.) 1
B.) 0.816
C.) -0.752
D.) 0.227

b.) Plot 3 does not exhibit a simple linear relationship because
A.) The corresponding residual plot would not be random scatter
B.) The slope of the regression line is not steep enough.
C.) Not all data points lie on the regression line
D.) None of the above

c.) If we were to remove the outlier in Plot 3, then the r-value of the points that remain would be
A.) 1
B.) 0
C.) -1
D.) .5
d.) Which scatterplot appears to possess a truly linear relationship?
   A.) Plot 1
   B.) Plot 2
   C.) Plot 4
   D.) None of the plots exhibit a linear relationship.

e.) Which scatterplot most exhibits a perfect nonlinear relationship?
   A.) Plot 1
   B.) Plot 2
   C.) Plot 3
   D.) Plot 4

f.) Which statement best describes coefficient of determination, $R^2$?
   A.) $R^2$ is the percentage of the variation of the response variable that can be explained by the model.
   B.) $R^2$ can detect the difference between a positive linear relationship and a negative linear relationship.
   C.) $R^2$ is the percentage of the response variable that can be explained by the model.
   D.) If the slope of the regression line is small, then $R^2$ will also be small.

g.) The point that steers the regression line in Plot 4 is not an outlier because it has a small residual.
   A.) True
   B.) False

h.) Suppose we wanted to test whether Plot 1 exhibited a linear relationship. Write down the all of the assumptions one would need to check in order to proceed with such a test, and which graphical display you would need to verify the assumption.
i.) Suppose that a two sided linear regression t-test of Plot 1 rejects the null hypothesis at a .05 level of significance. We would conclude:
   A. There is a linear relationship between the response variable and the explanatory variable, to a .05 level of significance.
   B. There is a positive linear relationship between the response variable and the explanatory variable, to a .05 level of significance.
   C. The variation of the response variable can be explained accurately by the linear model.
   D. There is evidence to suggest the observed values of the response variable are caused by the explanatory variable, in a way that is linear.

1.) More Multiple Choice:
   a.) Suppose the state decides to randomly test high school wrestlers for steroid use. There are 16 teams in the league, and each team has 20 wrestlers. State investigators plan to test 32 of these athletes by randomly choosing two wrestlers from each team. Is this a simple random sample?
   A. Yes, because the wrestlers were chosen at random.
   B. Yes, because each wrestler is equally likely to be chosen.
   C. Yes, because stratified samples are a type of simple random sample.
   D. No, because not all possible groups of 32 wrestlers could have been the sample.
   E. No, because a random sample of teams was not first chosen.

   b.) A recent survey said that 56% of college students live on campus, 62% have a campus meal program, and 42% do both. Living on campus and having a meal plan are:
   A. Independent
   B. Disjoint
   C. Both independent and disjoint
   D. Neither independent nor disjoint

   c.) Find the expected value using the chart below:

<table>
<thead>
<tr>
<th>x</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(X = x)</td>
<td>0.3</td>
<td>0.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

   A. 9
   B. 20
   C. 19
   D. 1
d.) Which of these has a Binomial model?
A. The number of people we survey until we find someone who has taken Statistics
B. The number of people in a class who have taken Statistics
C. The number of aces in a five-card Poker hand
D. The number of sodas students drink per day

e.) Which of these has a Geometric model?
A. The number of people we survey until we find someone who has taken Statistics
B. The number of people in a class who have taken Statistics
C. The number of aces in a five-card Poker hand
D. The number of sodas students drink per day

f.) A Poisson model can be used to approximate a Binomial model when
A. \( p \) is large and \( n \) is small.
B. \( p \) is equal to \( n \).
C. \( p \) is small and \( n \) is large.
D. The Poisson model cannot be used to approximate a Binomial model.

g.) Any conclusion from a statistical test, no matter how well-conceived, runs the risk of committing a Type error.
A. True
B. False

h.) Which best describes a p-value of a two tailed test?
A. The probability the null hypothesis is true.
B. The probability the null hypothesis is false.
C. The probability of observing a score as extreme as our test statistic assuming that we reject \( H_0 \).
D. The probability of observing a score as extreme as our test statistic assuming \( H_0 \) is true

i.) Which of the following is true about a 95% confidence interval constructed from a sample proportion (circle all that apply)?
A.) There is a 95% chance that the true population proportion falls within our interval
B.) There is a 95% chance that our interval captures the true proportion.
C.) If we repeat the experiment multiple times, then 95% of those 95% confidence intervals capture the true population proportion.
D.) The 98% confidence interval will be wider.

j.) We would like to conduct a test to see whether there is a significant difference in mean lifetime earnings for identical twins. We should conduct a
A.) linear regression t-test
B.) 2 proportion z-test
C.) Paired t-test
D.) 2 sample t-test
E.) ANOVA
The birthday problem of lab 5 (don’t worry if you don’t remember it) assumes that the distribution of birthdays are uniform across days of the year. To test this assumption we will use a dataset that is a random sample of 0.1% all birthdates in America in 1978. The simplified distribution of birthdays, across spans of 73 days, is listed below:

<table>
<thead>
<tr>
<th>Days</th>
<th>Births</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 73</td>
<td>596</td>
</tr>
<tr>
<td>74 to 146</td>
<td>624</td>
</tr>
<tr>
<td>147 to 219</td>
<td>675</td>
</tr>
<tr>
<td>220 to 292</td>
<td>710</td>
</tr>
<tr>
<td>293 to 365</td>
<td>677</td>
</tr>
</tbody>
</table>

a.) Is “Days” a categorical or quantitative variable?

b.) Assuming that I collected the data by selecting a year at random then taking random birthday data from that year, what type of sampling strategy did I employ?

c.) To test whether the distribution of birthdays in America is uniform, which type of test should we use?

d.) Write down the null and alternative hypotheses. Be explicit.

e.) Check all assumptions. Articulate any concerns.
f.) Conduct the test. Use .01 as your level of significance.

g.) State whether or not you reject the null hypothesis. What do you conclude?

h.) In the table, circle the 73 day span that contributed most to the test statistic. What can you conclude, in the context of this problem?
3.) The next several questions are related to the following experiment: a psychological study wants to test the effects of sleep deprivation on cognitive ability. 30 subjects with ordinary sleeping habits (8 hrs of sleep per day) are randomly assigned to 3 groups: None (no sleep overnight), Some (3 hours of sleep), and Reg (8 hours of sleep). The next morning the subjects are asked to perform 10 mental tasks, The number of tasks performed correctly for each group is listed below.

<table>
<thead>
<tr>
<th>None</th>
<th>Some</th>
<th>Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

a.) Write down the five number summary of scores for those who had no sleep.

b.) Using part a.), check numerically if there are any outliers in the “none” group.
c.) We wish to test, to a .03 level of significance, if sleep had any overall impact on cognitive ability. We will run an ANOVA to test. Write down the null and alternative hypotheses.

d.) Check all assumptions necessary to conduct the test. In particular, comment on which graphical display(s) would be necessary to check an assumption (be specific), and which assumption can be checked by the display below:

e.) Brief aside: Which boxplot(s) are skewed right?
f.) Now conduct the test, by filling in the ANOVA table below. (Take my word for it that the critical value is \(F_{0.03}=4.15\))

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment (sleep)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (within)</td>
<td></td>
<td>85.10</td>
<td>XXXX</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>142.97</td>
<td>XXXXX</td>
<td>XXX</td>
</tr>
</tbody>
</table>

g.) What do you conclude, in the context of this problem?

h.) We proceed with the Bonferroni method of multiple comparisons. What level of significance should you use when making the pairwise comparisons?
i.) Below is a partially completed table of multiple comparisons, with corresponding p values:

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>Some</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reg</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Some</td>
<td>.128</td>
<td>XXXXXXX</td>
</tr>
</tbody>
</table>

j.) To fill in the last cell of the table, which type of test should you use (be specific)?

k.) Suppose that the appropriate calculation yields a test statistic of -2.267. Was there a significant difference between cognitive ability of subjects with some sleep vs Regular amounts of sleep?

l.) Compactly summarize your findings of the ANOVA and subsequent multiple comparisons.
4.) Below is the contingency table of gender and survivors for passengers the Titanic:

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>343</td>
<td>367</td>
<td>710</td>
</tr>
<tr>
<td>Dead</td>
<td>127</td>
<td>1367</td>
<td>1491</td>
</tr>
<tr>
<td>Total</td>
<td>470</td>
<td>1731</td>
<td>2201</td>
</tr>
</tbody>
</table>

a.) What proportion of survivors are female?

b.) What proportion of females survived?

c.) Without conducting any statistical test, does there appear to be an association between gender and survival? Explain briefly.
5.) It is a commonly held belief that a greater proportion of female passengers survived the Titanic than males. So test, to a .01 level of significance, whether there is an association between gender and surviving the Titanic calamity. Be sure to state the null and alternative hypotheses, do all relevant assumption checks and what you conclude from the test.

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>343</td>
<td>367</td>
<td>710</td>
</tr>
<tr>
<td>Dead</td>
<td>127</td>
<td>1367</td>
<td>1491</td>
</tr>
<tr>
<td>Total</td>
<td>470</td>
<td>1731</td>
<td>2201</td>
</tr>
</tbody>
</table>
6.) Construct a 99% 2-proportion CI that corroborates your claim from #5.
   *You do not have to re-check any assumptions already checked in #5, but you do need to check any additional assumptions of constructing 2-proportion CI’s
   *There is more than one way to do this problem correctly. It is up to you to explain clearly what your symbols mean, what your CI tells you, and how comparing the CI with the conclusion from #5 either confirms or denies the claim in #5.
   *Only do one 2-proportion CI.