1. I once collected data in class on how long (in seconds) it took for a chocolate chip to melt in your mouth and for a peanut butter chip to melt in your mouth. I took the differences in these times (chocolate minus peanut butter) for each person. The sorted data, and a dotplot, for the 31 differences appear below:

<table>
<thead>
<tr>
<th>-41</th>
<th>-36</th>
<th>-35</th>
<th>-33</th>
<th>-31</th>
<th>-28</th>
<th>-25</th>
<th>-25</th>
<th>-20</th>
<th>-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>-17</td>
<td>-17</td>
<td>-16</td>
<td>-14</td>
<td>-11</td>
<td>-7</td>
<td>-6</td>
<td>-5</td>
<td>-5</td>
<td>-4</td>
</tr>
<tr>
<td>-4</td>
<td>-2</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>15</td>
<td>17</td>
<td>21</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Explain what the value -41 means in terms of the student who produced that value and his/her melting times.

This value means that the chocolate chip took 41 fewer seconds to melt than did the peanut butter chip for that student.

The mean of these 31 differences is -6.65 seconds, and the standard deviation is 23.61 seconds.

c) Conduct a test of whether the sample data provide strong evidence of a difference in melting times of chocolate and peanut butter chips on average. Report the hypotheses, test statistic, and p-value as accurately as you can.

The hypotheses are:  

\[ H_0: \mu_{\text{diff}} = 0 \quad H_a: \mu_{\text{diff}} \neq 0 \]

The test statistic is:  

\[ t = \frac{\bar{x}_{\text{diff}}}{s_{\text{diff}} / \sqrt{n_{\text{d}}} = -6.65 / (23.61 / \sqrt{31}) = -1.566. \]

Using the t-table with 30 degrees of freedom reveals that the one-sided p-value would be between .05 and .10, so the two-sided p-value is between .10 and .20.

d) Determine and interpret a 95% confidence interval based on the 31 differences.

A 95% confidence interval for the population mean difference \( \mu_{\text{diff}} \) is:  

\[ \bar{x}_{\text{diff}} \pm t^* \frac{s_{\text{diff}}}{\sqrt{n_{\text{d}}} = -6.65 \pm (2.042)(6.23.61)/\sqrt{31} , \text{ which is } -6.65 \pm 8.66, \text{ which is the interval from } -15.31 \text{ to } 2.01. \]

We can be 95% confident that the mean melt time with a chocolate chip is between 15.31 seconds less than with a peanut butter chip to 2.01 seconds more than with a peanut butter chip.
e) Summarize your conclusion from this analysis.

The p-value is not small, so our class data does not provide much evidence to believe that there is a difference, on average, between the melting times of chocolate and peanut butter chips.

f) Now suppose that you were to re-do this analysis after removing the outlier value of 67. Indicate how each of the following would change. Circle your answers. Do not bother to explain or perform any calculations.

Mean: 
- Decrease 
- Increase 
- Remain the same

Standard deviation: 
- Decrease 
- Increase 
- Remain the same

Test statistic: 
- Decrease (more negative) 
- Increase (less negative) 
- Remain the same

p-value: 
- Decrease 
- Increase 
- Remain the same

(Because the outlier was in the opposite direction of the overall tendency, meaning that the outlier was a large positive value and the overall mean was negative, removing the outlier would produce stronger evidence of a difference between the groups, and therefore a smaller p-value.)

2. (10 pts) Students in an introductory statistics class were asked how many states they have visited. The following output pertains to the sample results:

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>states</td>
<td>50</td>
<td>13.36</td>
<td>1.03</td>
<td>7.27</td>
<td>2.00</td>
<td>7.00</td>
<td>12.00</td>
<td>20.00</td>
<td>31.00</td>
</tr>
</tbody>
</table>

a) (3 pts) Determine a 90% confidence interval for the population mean number of states visited among all students at this university.

The 90% confidence interval for \( \mu \) is: \( \bar{x} \pm t^* \frac{s}{\sqrt{n}} \). We need 49 degrees of freedom, but the table only gives 40 and 50 df. Using 40 df gives \( t^* = 1.684 \), so the CI is: \( 13.36 \pm 1.684(7.27)/\sqrt{50} \), which is \( 13.36 \pm 1.73 \), which is the interval (11.63, 15.09).

b) (2 pts) Check and comment on whether the technical conditions of this confidence interval are satisfied.

We do not strictly have a random sample from students at the university, because only students from one class were chosen. The sample size (50) is large enough (larger than 30) that the second condition is satisfied.
c) (1 pts) For what proportion of students in the sample is the number of states visited within the interval from a)?

Students who have been in 12, 13, 14, or 15 states fall within the interval (11.63, 15.09). The dotplot shows that this includes 8 students, which is 8/50 = .16.

d) (2 pts) Should you expect your answer to c) to be close to 90%? Explain why or why not.

No. The interval estimates the population mean, not individual values.

e) (2 pts) Based on your interval, what can you say about the p-value if you were to conduct a two-sided significance test of whether the population mean differs from 10? Explain briefly, without conducting a test or doing new calculations.

The value 10 is not within the CI, so it is a not plausible value of μ, so we would reject the null hypothesis that μ = 10, so the p-value would be less than .10.

3. In a recent study, researchers investigated possible biochemical mechanisms that could be involved in the early stages of romantic love. They measures plasma level of neurotrophins for a sample of 58 subjects who had recently fallen in love. They also asked each subject to rate his/her level of passionate love feelings on a numerical scale. Researchers calculated the correlation coefficient between the level of passionate love and plasma level of neurotrophins to be $r = 0.34$.

Conduct the appropriate test of whether this sample provides strong evidence (at the $\alpha = .05$ level) of a positive correlation between these variables in the population. Report the relevant hypotheses, test statistic, p-value, test decision, and conclusion.

The hypotheses are $H_0: \rho = 0$ vs. $H_a: \rho > 0$, where $\rho$ represents the correlation coefficient between level of passionate love and plasma level of neurotrophins in the population of all people who have recently fallen in love. (It is equivalent to state the hypotheses in terms of the population slope coefficient: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 > 0$,

The test statistic is: $t = \sqrt{\frac{n-2}{1-r^2}} \approx 2.71$.

The $p$-value is found from the right tail of a t-distribution with 56 degrees of freedom. Looking in Table T reveals that $p$-value < .005.

Because $p$-value < .05, we reject the null hypothesis at the $\alpha = .05$ significance level.

The sample data provide very strong evidence that there is a positive correlation between level of passionate love and plasma level of neurotrophins in the population of all people who have recently fallen in love.
4. Between the months of September 1990 and May 1997, a statistics teacher gathered data on the average temperature for that month (in degrees Fahrenheit) and the amount of gas usage in his home for that month (in units called therms). Summary statistics for these variables follow:

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StDev</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg temp</td>
<td>71</td>
<td>46.35</td>
<td>1.80</td>
<td>15.16</td>
<td>45.00</td>
<td>26.00</td>
</tr>
<tr>
<td>gas usage per day</td>
<td>71</td>
<td>5.311</td>
<td>0.420</td>
<td>3.538</td>
<td>5.000</td>
<td>6.600</td>
</tr>
<tr>
<td>Pearson correlation of avg temp and gas usage per day</td>
<td>-0.930</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Determine and report the equation of the least squares line for predicting a month’s gas usage per day based on its average temperature.

The slope coefficient is: \( b_1 = r \frac{s_y}{s_x} = (-0.930) \times (3.538)/(15.16) \approx -0.217 \)

The intercept coefficient is: \( b_0 = \bar{y} - b_1 \bar{x} = 5.311 - (-0.217) \times 46.35 \approx 15.369 \)

The equation of the least squares line is therefore:

predicted gas usage per day = 15.369 – 0.217 \times \text{avg temp}.

b) In the following scatterplot with the least squares line superimposed, circle the point corresponding to the month with the largest positive residual:

The largest positive residual corresponds to the point farthest above the least squares line.

c) Calculate the value of \( r^2 \), and write a sentence interpreting what this value means.

The value of \( r^2 = (-0.930)^2 \approx 0.865 \), which means that 86.5% of the variability in gas usage per day across these months is explained by knowing the average temperature for the month.

d) Predict the gas usage per day for a month in which the average temperature is 50 degrees Fahrenheit.

This prediction is: predicted gas usage per day = 15.369 – 0.217 \times 50 \approx 4.519 \text{ therms per day}.

e) A significance test or confidence interval for the slope coefficient will be based on how many degrees of freedom?

The degrees of freedom is \( n - 2 = 71 - 2 = 69 \).
f) Computer software reports the standard error of the sample slope coefficient to be 0.01036. Use this information to produce a 95% confidence interval for the population slope coefficient.

This 95% confidence interval is: \( b_1 \pm t^* \times SE(b_1) \), which is \(-0.217 \pm 2.000 \times 0.01036\), which is \(-0.217 \pm 0.021\), which is the interval \((-0.238, -0.196)\).

g) Write a sentence interpreting this interval, including what the slope coefficient means.

We are 95% confident that the predicted decrease in gas usage for each additional degree (F) of temperature is between 0.238 therms/degree and 0.196 therms/degree.

5. A recent study investigated whether seat position within a bus may be related to whether a passenger experiences motion sickness. The following table classifies each person in a random sample of bus riders by the location of his or her seat and whether nausea was reported:

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Middle</th>
<th>Rear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nausea</td>
<td>58</td>
<td>166</td>
<td>193</td>
</tr>
<tr>
<td>No nausea</td>
<td>870</td>
<td>1163</td>
<td>806</td>
</tr>
</tbody>
</table>

a) What proportion of those who rode in the front experienced nausea?

\( \frac{58}{58+870} = \frac{58}{928} \approx 0.0625 \)

b) Determine the expected count for the (middle, no nausea) cell of the table. (Show your work.)

This expected count is \( \frac{2839 \times 1329}{3256} \approx 1158.79 \).

Consider the following output (notice that one entry has been replaced by xxx):

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Middle</th>
<th>Rear</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nausea</td>
<td>58</td>
<td>166</td>
<td>193</td>
<td>417</td>
</tr>
<tr>
<td>No nausea</td>
<td>870</td>
<td>1163</td>
<td>806</td>
<td>2839</td>
</tr>
<tr>
<td>Total</td>
<td>928</td>
<td>1329</td>
<td>999</td>
<td>3256</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Middle</th>
<th>Rear</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>58</td>
<td>166</td>
<td>193</td>
</tr>
<tr>
<td></td>
<td>118.85</td>
<td>170.21</td>
<td>127.94</td>
<td>31.155</td>
</tr>
<tr>
<td></td>
<td>31.155</td>
<td>0.104</td>
<td>33.080</td>
<td>31.155</td>
</tr>
<tr>
<td>2</td>
<td>870</td>
<td>1163</td>
<td>806</td>
<td>2839</td>
</tr>
<tr>
<td></td>
<td>809.15</td>
<td>xxx</td>
<td>871.06</td>
<td>4.576</td>
</tr>
<tr>
<td></td>
<td>4.576</td>
<td>0.015</td>
<td>4.859</td>
<td>4.576</td>
</tr>
<tr>
<td>Total</td>
<td>928</td>
<td>1329</td>
<td>999</td>
<td>3256</td>
</tr>
</tbody>
</table>

\( \text{Chi-Sq} = 73.789, \ DF = 2, \ P\text{-Value} = 0.000 \)

c) What conclusion would you draw from the above chi-square output? Explain briefly.

The p-value is essentially zero, so the sample data provide overwhelming evidence that the population proportions who experience nausea are not the same in the three areas of the bus.
d) Which seating position seems to be the worst with respect to nausea? Explain how you conclusion follows from the chi-square analysis.

The largest contribution to the chi-square test statistic value is 33.080, from the (rear, nausea) cell of the table. The observed count there is much higher than expected. For the other cells in the top row (the “nausea” row), the observed count is lower than expected. So, the rear is the worst seating position with regard to nausea.

6. Researchers studied heart rates after engaging in physical exercise for adults who were also classified according to whether and how much they smoke. Data were collected to investigate whether there are differences in mean heart rates among various smoking classifications (heavy, light, moderate, non-smoker). The output below pertains to an ANOVA analysis addressing this issue:

<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>smoking status</td>
<td>3</td>
<td>1464.1</td>
<td>488.0</td>
<td>6.08</td>
<td>0.004</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>1604.8</td>
<td>80.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>3069.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) State the appropriate null hypothesis in symbols. Also explain what the symbols mean.

\[ H_0: \mu_{\text{heavy}} = \mu_{\text{moderate}} = \mu_{\text{light}} = \mu_{\text{non}} \]

(where \( \mu_i \) represents the population mean heart rate in smoking level \( i \))

b) Summarize the conclusion that you would draw from the ANOVA F-test.

Because the \( p \)-value is so small (.004), the sample data provide very strong evidence that there is a difference in population mean heart rate among these four smoking groups. At least one of the group’s mean heart rate differs significantly from another group’s.

7. Some of the statistical inference techniques we have studied include:
   A. One-sample \( z \)-procedures for a proportion
   B. Two-sample \( z \)-procedures for comparing proportions
   C. One-sample \( t \)-procedures for a mean
   D. Two-sample \( t \)-procedures for comparing means
   E. Paired-sample \( t \)-procedures
   F. Chi-square procedures for two-way tables
   G. ANOVA procedures
   H. Linear regression procedures

For each of the following research questions, indicate (by letter) the appropriate statistical inference procedure for investigating the question.

a) A researcher used data from the American Time Use Survey (ATUS) to investigate whether high school math teachers spend more time working per day than high school history teachers.

   D. Two-sample \( t \)-procedures for comparing means
b) Biologists recorded the frequency of a cricket’s chirps (in chirps per minute) and also the temperature (in degrees Fahrenheit) when the cricket measurement was recorded. They investigated whether chirp frequency is a significant predictor of temperature.

H. Linear regression procedures

c) Economists compared starting salaries of new employees across three different groups: those with graduate degrees, those with only bachelor’s degrees, and those with no higher education degrees.

G. ANOVA procedures

d) A researcher investigated whether laughter increases blood flow by having subjects watch a humorous movie and a stressful movie, randomly deciding which movie the subject would see first, measuring the blood flow through the person’s blood vessels while watching the movie.

E. Paired-sample $t$-procedures

e) Do more than two-thirds of Cal Poly students have at least one class on Fridays this quarter?

A. One-sample $z$-procedures for a proportion

f) Is there an association between a college student’s level of drinking alcohol (classified as none, some, or considerable) and her/his residence situation (classified as living on-campus, off-campus with parents, or off-campus but not with parents)?

F. Chi-square procedures for two-way tables