Part I

Give a short, but precise and clear, answer to each question. Show sketches & details of calculations where needed. You may neglect assumptions for inference problems.

I.1. We want to know the true proportion of EC students who love mathematics! We surveyed a group of students and found a 95% confidence interval estimated this proportion to lie in the range [0.75, 0.93].

(a) What is the margin of error according to this study?
   \[ 0.93 - 0.75 = 0.18 \text{ Margin of error} = 0.18/2 = 0.09 \]

(b) Would the interval get narrower, or wider, if we use 80% confidence?
   \[ ME = Z^* SE. \text{ With } 80\% \text{ confidence, } Z^* \text{ would decrease; } SE \text{ would be the same. Therefore, interval would get narrower/smaller.} \]

(c) What must we do if we want high confidence level and low margin of error at the same time?
   \[ ME = Z^* SE. \text{ For high confidence, } Z^* \text{ will be large. The only way to make } ME \text{ small would be to decrease } SE. \text{ The only way to decrease } SE \text{ is to increase sample size.} \]

Answer: Increase sample size.

I.2. [4 pts] A sociologist wants to know the opinion of employed adult women regarding proposed new regulations on day-care centers. She has conducted telephone interviews with a random sample of 102 employed women and found 60% of them oppose the new regulations. Identify the following as precisely as possible: the population, the sample, the parameter, and the statistic.

(1) The population = All employed adult women, presumably in some geographic region, e.g., U.S.

(2) The sample = The 102 women who were contacted for telephone interviews.

(3) The parameter = The true proportion of all adult women who support (or oppose) proposed new regulations on day-care centers.

(4) The statistic = The 60% of the surveyed sample who oppose the new regulations.
I.3. [4 pts.] A study of average January temperatures for 50 cities in the U.S. gave the correlations shown in the table with 3 different variables – longitude, altitude and distance from the coast.

<table>
<thead>
<tr>
<th>JanTemp</th>
<th>Long</th>
<th>Altitude</th>
<th>Coast</th>
</tr>
</thead>
<tbody>
<tr>
<td>JanTemp</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long</td>
<td>0.029</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Altitude</td>
<td>-0.369</td>
<td>0.494</td>
<td>1.000</td>
</tr>
<tr>
<td>Coast</td>
<td>-0.065</td>
<td>0.201</td>
<td>0.310</td>
</tr>
</tbody>
</table>

(a) Which of the 3 variables seems to have the strongest effect on January temperature? Give reason.

Altitude seems to have the strongest effect on January temp. This is because its correlation has the largest magnitude (-0.369).

(b) Which 2 variables are the weakest in their correlation with each other?

January temperature and longitude seem to have the weakest correlation, as the magnitude of 0.029 is the smallest.

(c) State the key assumption made when interpreting correlations in this way.

Linear association between each pair of variables is assumed when interpreting correlations in this way.

I.4. [4 pts.] Given the information in the previous question, suppose we want to build a linear regression model to predict average January temperatures from altitude. The following additional summary statistics are available for the 50 cities in our sample

<table>
<thead>
<tr>
<th>January temp (°F)</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altitude (1000 feet)</td>
<td>43</td>
<td>12.3</td>
</tr>
<tr>
<td>Altitude (1000 feet)</td>
<td>1.8</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Find the slope and intercept of the regression line, with correct units.

\[ x = \text{Explanatory Variable = Altitude (in 1000 feet)} \]
\[ y = \text{Response Variable = January temp (in °F)} \]

Slope of regression line: \[ m = r \frac{s_y}{s_x} = -0.369 \frac{12.3}{0.62} = -7.32 \text{ (°F per 1000 ft)} \]

m = -7.32 °F per thousand feet

Intercept calculation: \[ \hat{y} = -7.32x + b \]

Plug in mean values for \( (x, y) \): 43 = -7.32(1.8) + b

\[ b = 43 + 7.32(1.8) = 56.176 \text{ °F} \]

Answers: slope = -7.32 °F per 1000 ft.

Intercept = 56.176 °F
1.5. [4 pts.] The distribution of textbook prices at a large college bookstore is known to be strongly skewed right, with true mean and standard deviation of $120 and $43.0 respectively. Describe, with sketch, the sampling distribution of mean prices for random samples of 200 textbooks (must specify shape, mean, SD).

- **Population**: Prices of all textbooks at the bookstore, given it is strongly skewed right. \( \mu = 120 \), \( \sigma = 43 \)
- **Sample**: Randomly selected 200 textbooks.
  - Sample size = \( n = 200 \)

According to the Central Limit Theorem, since \( n \) is large and the samples are random, the sampling distribution follows the normal model \( N(120, \frac{43}{\sqrt{200}}) \)

\[ N(120, 3.0406) \]

\[ \bar{y} = \text{mean price of random sample of 200 textbooks} \]

1.6. [4 pts.] Given the probabilities \( P(C) = 0.6 \), \( P(D) = 0.4 \), \( P(D|C) = 0.5 \), find the probability \( P(C \text{ or } D) \). Show steps.

\[ P(C \text{ or } D) = P(C) + P(D) - P(C \text{ and } D) \]

We are given \( P(C) \), \( P(D) \), and can find \( P(C \text{ and } D) \) from \( P(D|C) \).

\[ P(D|C) = \frac{P(D \text{ and } C)}{P(C)} \implies 0.5 = \frac{P(D \text{ and } C)}{0.6} \]

\[ \therefore P(D \text{ and } C) = 0.5 \times 0.6 = 0.3 \]

\[ P(C \text{ or } D) = 0.6 + 0.4 - 0.3 = 0.7 \]

**Method 2**

\[ \frac{x}{0.6} = 0.5 \]

\[ x = 0.5 \times 0.6 \]

\[ x = 0.3 \]

From this, we get the rest of the numbers in the Venn diagram: \( P(C) - x = 0.3 \); \( P(D) - x = 0.1 \)

**Answer:** \( P(C \text{ or } D) = 0.7 \)
I.7. [5 pts.] The sketches below show rough (but qualitatively correct) histogram profiles for two variables $X$, $Y$. For each sketch, answer the following questions:
(a) Is the mean less than, equal to, or greater than, the median?
(b) What summary statistics would best describe the center and spread?

For graph of $X$:
(a) Because of the right skew, the mean will be larger than the median.
(b) Since the distribution is skewed, the median and IQR would be best for summarizing the center and spread.

For graph of $Y$:
(a) This distribution looks symmetric. So the mean and median should be about equal.
(b) For symmetric distribution, either mean/standard deviation or median/IQR would work well for summarizing center and spread.

Grade note:
Variable $X = 2.5$ points
Variable $Y = 2.5$ points
For each variable: (a) = 1.5 points
(b) = 1 point
Part II

Give complete and detailed solutions to each question. [7 points each]

II.1 An auto rental company’s fleet consists of 70% U.S. brands, and the rest are foreign brands. The company notes that manufacturers’ recalls seem to affect about 1% of the U.S. brands, and about 0.5% of the foreign brands.

(a) Find the probability that a randomly selected vehicle in their fleet is recalled.
(b) Find the probability that a recalled vehicle is a foreign brand.
(c) What is the probability that a vehicle is recalled and it is a foreign brand.

[Show details of work & reasoning.]

A tree diagram would work well here for organizing the information.

\[
\begin{align*}
\text{Recall} & \quad 0.3 \quad \text{Recall} \\
\text{Not Recall} & \quad 0.99 \quad \text{Not Recall}
\end{align*}
\]

\[
\begin{align*}
\Pr(\text{Recall}) & = \Pr(\text{U.S. and Recall}) + \Pr(\text{Not U.S. and Recall}) \\
& = 0.7 \times 0.01 + 0.3 \times 0.005 = 0.0085
\end{align*}
\]

Answer: \( \Pr(\text{Recall}) = 0.0085 \) = 0.85%.

(b) This conditional probability is asking for \( \Pr(\text{Not U.S.} \mid \text{Recall}) \).
\[
\Pr(\text{Not U.S.} \mid \text{Recall}) = \frac{\Pr(\text{Not U.S. and Recall})}{\Pr(\text{Recall})} = \frac{0.3 \times 0.005}{0.0085}
\]

Answer: \( \Pr(\text{Not U.S.} \mid \text{Recall}) = 0.1765 \) = 17.65%.

(c) This is asking for \( \Pr(\text{Not U.S. and Recall}) \).
We have already computed that in (b): \( 0.3 \times 0.005 \)

Answer: \( \Pr(\text{Not U.S. and Recall}) = 0.0015 \)

Grade notes:
1 pt for correct tree diagram (or other correct organizing scheme)
2 pt for (a); 2 pt for (b); 2 pt for (c).
For each case, 1 pt = correct interpretation of probability,
1 pt = correct computation.
II.2 The following table shows the incarceration rate of White and Black prisoners in the State of Indiana for the years 1980-2000. Incarceration rates are shown in number of prisoners per 100,000 of population.

<table>
<thead>
<tr>
<th>Year</th>
<th>1980</th>
<th>'82</th>
<th>'84</th>
<th>'86</th>
<th>'88</th>
<th>'90</th>
<th>'92</th>
<th>'94</th>
<th>'96</th>
<th>'98</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>4.2</td>
<td>5.1</td>
<td>5.9</td>
<td>6.3</td>
<td>7.0</td>
<td>7.4</td>
<td>8.1</td>
<td>8.3</td>
<td>7.9</td>
<td>7.8</td>
<td>7.7</td>
</tr>
<tr>
<td>Black</td>
<td>381</td>
<td>502</td>
<td>620</td>
<td>752</td>
<td>873</td>
<td>1022</td>
<td>1327</td>
<td>1425</td>
<td>1500</td>
<td>1630</td>
<td>1728</td>
</tr>
</tbody>
</table>

(a) For each race, calculate appropriate summary statistics of the incarceration rates. No need to show steps.

(b) Suppose we want to estimate the true mean difference in incarceration rates using a confidence interval based on these data. Check whether the conditions for inference are met. Part of your check should include a neat, hand-drawn histogram, with clear labels, for at least one of the variables.

(c) Construct a 90% confidence interval for the true mean difference in incarceration rates, and interpret what it says.

(a) From my calculator, for Whites: \( \bar{Y}_{\text{White}} = 6.882 \), \( SD = 1.3415 \)

for Blacks: \( \bar{Y}_{\text{Black}} = 1069.091 \), \( SD = 475.984 \)

\[ \begin{align*}
\text{Min} &= 4.2 \\
Q_1 &= 5.7 \\
Q_3 &= 7.2 \\
Med &= 7.7 \\
\text{Max} &= 8.3
\end{align*} \]

The conditions that must be satisfied for comparing 2 independent means using inference techniques are:

1. Each sample is independent
2. Samples are indep. of each other
3. Each sample is approximately normal.

For (1), since we are looking at 20 consecutive years of data, it is unlikely that either sample is independent.

For (2), it is probably reasonable to assume independence of each other, though a random sample of years would be more compelling.

For (3), Whites:

<table>
<thead>
<tr>
<th>Range</th>
<th># of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-5</td>
<td>1</td>
</tr>
<tr>
<td>5-6</td>
<td>2</td>
</tr>
<tr>
<td>6-7</td>
<td>1</td>
</tr>
<tr>
<td>7-8</td>
<td>5</td>
</tr>
<tr>
<td>8-9</td>
<td>2</td>
</tr>
</tbody>
</table>

I looked at the histogram for Blacks on my TI-84 calculator.

The assumptions/conditions for inference satisfied by these samples.

(c) Sampled stats are:

\[ \bar{Y}_{\text{Black}} = 1069.091, \quad \bar{Y}_{\text{White}} = 6.882, \quad SD_{\text{Black}} = 475.984, \quad SD_{\text{White}} = 1.3415 \]

\[ ME = \bar{Y}_{\text{Black}} - \bar{Y}_{\text{White}} \]

\[ ME = 802.16, 1322.36 \]

The confidence interval for the true mean incarceration rate for Blacks is between 802.16 and 1322.36 higher than the true mean incarceration rate for Whites.

\[ SE = 143.515, \quad t_{10} = 1.812 \]

From calculator, \( df = 10 \)

\[ ME = 260.049 \]
II.3 A team of scientists, researching the consequences of vitamin B₁₂ deficiency, tracked a group of 36 adults with B₁₂ deficiency for 7 years. At the end of this period they found 9 people in their sample exhibited symptoms of major depression.

(a) What kind of study was this? (E.g., survey? observation? experiment?)

(b) What is the nature and scope of conclusions this study can reach regarding B₁₂ deficiency and depression? Explain.

(c) Suppose the true rate of depression for all adults is 15%. Test an appropriate hypothesis to determine whether the rate of depression for adults with B₁₂ deficiency is significantly different. Be sure to include all steps.

(a) This was an observational study. It was prospective.

(b) Because it was not an experiment, the most the study could conclude is there is some association between B₁₂ deficiency and symptoms of major depression among adults. However, even this claim would be difficult to make without knowing more about details of the study.

(c) Let \( p = \text{true rate of depression for adults with B₁₂ deficiency} \)

\[ H₀: p = 0.15 \]
\[ Hₐ: p ≠ 0.15 \] (two tailed, since we want to know if there is a difference from the rate of 15%)

* Sampling distribution model: If the conditions are met, then sample proportions should follow \( \hat{p} \) ~ N(0.15, \( \sqrt{\frac{0.15 \times 0.85}{36}} \))

\[ \mu = 0.15, \sigma = 0.0595 \]

Conditions check:

1. Is the sample independent?
   Not clarified whether any random selection of sample was involved

2. Sufficiently large? \( n \cdot p = 36 \times 0.15 = 5.4 \)
   \( n \cdot (1-p) = 30.6 \)

Fails to satisfy sample-size requirement.

I will proceed with inference, but note that conditions are not met.

* Compute p-value: From the given data, \( \hat{p} = \frac{9}{36} = 0.25 \)

\[ z = \frac{0.25 - 0.15}{0.0595} = 1.68 \]

Area to the right of \( z = 1.68 \) is 0.0464

\[ P-value = 2 \times 0.0464 = 0.0928 \]

Assume significance level is 10%.

Conclusion: Since the p-value is below our significance level, I reject the null hypothesis and conclude the rate of depression is different for adults with B₁₂ deficiency. The caveat is, of course, the conditions for inference were not met by this sample.