**Week #4: Review of Chapters 1, 2, and 3**

**Problem 1.** A study of fathers’ involvement in their children’s education interviews a random sample of fathers of school-aged children. One question concerns attendance at scheduled parent-teacher conferences. The table below shows the results:

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Some</th>
<th>None</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-parent families</td>
<td>109</td>
<td>132</td>
<td>203</td>
<td>444</td>
</tr>
<tr>
<td>Single-parent families</td>
<td>15</td>
<td>10</td>
<td>13</td>
<td>49</td>
</tr>
<tr>
<td>Non-resident fathers</td>
<td>11</td>
<td>25</td>
<td>82</td>
<td>118</td>
</tr>
</tbody>
</table>

a. Create a contingency table that shows the distribution of attendance for each level of family structure.

b. Does it appear as though attendance and family structure are dependent or independent? Why?

---

A. Family Structure

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Some</th>
<th>None</th>
<th>P(all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-parent</td>
<td>109/444 = 0.245</td>
<td>132/444 = 0.297</td>
<td>203/444 = 0.457</td>
<td></td>
</tr>
<tr>
<td>1-parent</td>
<td>15/38 = 0.395</td>
<td>10/38 = 0.263</td>
<td>13/38 = 0.342</td>
<td></td>
</tr>
<tr>
<td>Non-resident</td>
<td>11/118 = 0.093</td>
<td>25/118 = 0.212</td>
<td>82/118 = 0.695</td>
<td></td>
</tr>
</tbody>
</table>

B. If they were independent → P(all) the same for each group

Since props of those who attend all vs. some vs. none varies by family structure, they are dependent.
Problem 2. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased life span. Each patient entering the program was designated an official heart transplant candidate, meaning they were gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable \textit{transplant} indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Another variable called \textit{survived} was used to indicate whether or not the patient was alive at the end of the study. Of the 34 patients in the control group, 30 died. Of the 69 patients in the treatment group, 45 died. Researchers also measured a third variable for each patient, \textit{survival time}, which recorded the length of time each patient survived for.

a. What proportion of patients in the treatment group died?

b. What proportion of patients in the control group died?

c. Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

d. What do the box plots suggest about the efficacy (effectiveness) of the heart transplant treatment?

\begin{align*}
\text{a. } P(\text{died } | \text{trt}) &= \frac{45}{69} = 0.652 \\
\text{b. } P(\text{died } | \text{control}) &= \frac{30}{34} = 0.882
\end{align*}
Problem 2. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased life span. Each patient entering the program was designated an official heart transplant candidate, meaning they were gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable \textit{transplant} indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Another variable called \textit{survived} was used to indicate whether or not the patient was alive at the end of the study. Of the 34 patients in the control group, 30 died. Of the 69 patients in the treatment group, 45 died. Researchers also measured a third variable for each patient, \textit{survival time}, which recorded the length of time each patient survived for.

![Mosaic plot and box plots showing survival and survival time](image)

a. What proportion of patients in the treatment group died?
b. What proportion of patients in the control group died?
c. Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.
d. What do the box plots suggest about the efficacy (effectiveness) of the heart transplant treatment?

C. NO, survival rates are different for \textit{trt} vs. control, they are dependent.

d. While a large proportion of those in \textit{trt} group died before end of the study, on avg., their survival time was longer than those in the control group.
Problem 3. The following set of boxplots shows the 5 number summary for the ages of all Oscar Winning Actors from 1975 to 2004, split by gender. Based on these boxplots, does it appear that there is evidence that the average age for men is different from the average age for females?

Maybe - the median ages for these two samples is different, however their Interquartile Ranges (IQR) overlap quite a bit (possibly very similar).
Problem 4. The following histogram shows the relationship between the weight of a car and the mileage it gets. How would you describe the association? If you were given a choice between -1.0, -0.79, -0.42, 0, 0.42, 0.79, and 1.0, which would you say is the best estimate of the correlation between these two variables?

- **Form:** Linear
- **Direction:** Negative
- **Strength:** $-0.79$, $-0.42$, $0$, $0.42$, $0.79$, $1.0$
- **Outliers:** No apparent outliers

![Scatterplot with outlier](image)
Problem 5. Describe the distribution in the histograms below and match them to the box plots.

Histogram 1: Uniform, Symmetric, No Outliers

Histogram 2: Unimodal, Skewed Right, No Outliers

Histogram 3: Unimodal, Symmetric, Potential Outlier

Boxplot A

Boxplot B

Boxplot C
Problem 6. Calculate the correlation between the two variables shown below:

\[ r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( y_i - \bar{y} \right) \]

Important values:

\[ n = 3 \]
\[ \bar{x} = \frac{1 + 2 + 3}{3} = \frac{6}{3} = 2 \]
\[ \bar{y} = \frac{2 + 4 + 6}{3} = \frac{12}{3} = 4 \]

Calculate \( r \):

\[ r = \frac{1}{3-1} \left( \frac{1 - 2}{s_x} \cdot \frac{2 - 4}{s_y} + \frac{2 - 2}{s_x} \cdot \frac{4 - 4}{s_y} + \frac{3 - 2}{s_x} \cdot \frac{6 - 4}{s_y} \right) \]

\[ r = \frac{1}{2} \left( \frac{-1}{s_x} + \frac{0}{s_y} + \frac{1}{s_x} \right) = \frac{1}{2} \left( \frac{-1}{1} + \frac{0}{0} + \frac{1}{1} \right) = 0 \]
Problem 6. Calculate the correlation between the two variables shown below:

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

\[
\text{Check}
\]

\[\checkmark\]
Problem 7. The smallpox data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston. Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death. Each case represents one person with two variables: inoculated and result. The variable inoculated takes two levels: yes or no, indicating whether the person was inoculated or not. The variable result has two outcomes: lived or died, indicated whether the person survived or not. This data set is summarized below.

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lived</td>
<td>238</td>
<td>5136</td>
<td>5374</td>
</tr>
<tr>
<td>Died</td>
<td>6</td>
<td>44</td>
<td>850</td>
</tr>
<tr>
<td>Total</td>
<td>244</td>
<td>5980</td>
<td>6224</td>
</tr>
</tbody>
</table>

a. What is the sample space?
b. What is the probability that a randomly selected individual survived?
c. What is the probability that a randomly selected individual who was inoculated survived?
d. What is the probability that a randomly selected individual who was not inoculated survived?
e. Does it seem like inoculation and survival are independent?

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five Guys Burgers</td>
<td>5</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>In-N-Out Burger</td>
<td>162</td>
<td>181</td>
<td>343</td>
</tr>
<tr>
<td>Fat Burger</td>
<td>10</td>
<td>12</td>
<td>22</td>
</tr>
<tr>
<td>Tommy’s Hamburgers</td>
<td>27</td>
<td>27</td>
<td>54</td>
</tr>
<tr>
<td>Umami Burger</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>26</td>
<td>20</td>
<td>46</td>
</tr>
<tr>
<td>Not Sure</td>
<td>13</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>248</td>
<td>252</td>
<td>500</td>
</tr>
</tbody>
</table>

a. Are being female and liking In-N-Out Burger best mutually exclusive?

b. What is the probability that a randomly selected males likes In-N-Out the best?

c. What is the probability that a randomly selected female likes In-N-Out the best?

d. What is the probability that a man and a woman who are dating both like In-N-Out the best?

Note any assumptions you make and evaluated whether you think they are reasonable.

e. What is the probability that a randomly selected person like In-N-Out best or that person is female?

\[ \text{A. No, 181 females picked In-N-Out} \]

\[ \text{b. } P(\text{IN0} | \text{male}) = \frac{\text{# male} \text{ # IN0}}{\text{# male}} = \frac{162}{248} = 0.65 \]

\[ \text{c. } P(\text{IN0} | \text{female}) = \frac{\text{# female} \text{ # IN0}}{\text{# female}} = \frac{181}{252} = 0.72 \]

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five Guys Burgers</td>
<td>5</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>In-N-Out Burger</td>
<td>162</td>
<td>181</td>
<td>343</td>
</tr>
<tr>
<td>Fat Burger</td>
<td>10</td>
<td>12</td>
<td>22</td>
</tr>
<tr>
<td>Tommy’s Hamburgers</td>
<td>27</td>
<td>27</td>
<td>54</td>
</tr>
<tr>
<td>Umami Burger</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>26</td>
<td>20</td>
<td>46</td>
</tr>
<tr>
<td>Not Sure</td>
<td>13</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>248</td>
<td>252</td>
<td>500</td>
</tr>
</tbody>
</table>

a. Are being female and liking In-N-Out Burger best mutually exclusive?

b. What is the probability that a randomly selected males likes In-N-Out the best?

c. What is the probability that a randomly selected female likes In-N-Out the best?

d. What is the probability that a man and a woman who are dating both like In-N-Out the best?

Note any assumptions you make and evaluate whether you think they are reasonable.

e. What is the probability that a randomly selected person like In-N-Out best or that person is female?

\[
\begin{align*}
\text{d. } P(\text{male } \text{INO} \land \text{female } \text{INO}) &= P(\text{male } \text{INO}) \times P(\text{female } \text{INO}) \\
&= (0.65)(0.72) \\
&= 0.468
\end{align*}
\]

\[
\begin{align*}
\text{e. } P(\text{INO } \lor \text{ female}) &= \frac{\text{#INO } + \text{ #female } - (\text{#INO } \land \text{female})}{\text{total } \#} \\
&= \frac{343 + 252 - 181}{500} = \frac{414}{500} = 0.828
\end{align*}
\]
Problem 9. Many times when we are discussing probabilities about diseases, we talk about the risk and the odds. The risk is the same as the probability, while the odds is the probability divided by one minus the probability. One of the most common genetic disorders in the United States is Down Syndrome. If an individual has Down Syndrome, there is an increased chance that they will have a heart defect. Approximately 47% if infants born with Down Syndrome also have a heart defect?

a. If a child has Down Syndrome, what is the risk of them having a heart defect?

b. If a child has Down Syndrome, what is the odds of them having a heart defect? What does this mean?

\[ \text{risk} = P \]
\[ \text{odds} = \frac{P}{1-P} \]

\[ a. \ P(\text{Heart Defect | Down Syndrome}) = 0.47 \]

\[ b. \ \frac{P}{1-P} = \frac{0.47}{0.53} = 0.887 \]

Children with Down Syndrome are more likely to not have a heart defect.
Problem 10. A genetic test is used to determine if people have a predisposition for thrombosis, which is the formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people actually have this predisposition. The genetic test is 99% accurate if a person actually has the predisposition, meaning that the probability of a positive test result when a person actually has the predisposition is 0.99. The test is 98% accurate if a person does not have the predisposition. What is the probability that a randomly selected person who tests positive for the predisposition actually has the predisposition?

\[ P(D^+ | T^+) = \frac{P(D^+ \cap T^+)}{P(T^+)} \]

\[ P(D^+) = 0.03 \quad P(T^+ | D^+) = 0.99 \quad P(T^- | D^-) = 0.98 \]

\[ P(D^+ \cap T^+) = (0.03)(0.99) \]
\[ = 0.0297 \]

\[ P(D^- \cap T^+) = (0.97)(0.02) \]
\[ = 0.0194 \]

\[ P(T^+) = P(D^+ \cap T^+) + P(D^- \cap T^+) \]
\[ = 0.0297 + 0.0194 \]
\[ = 0.0491 \]

\[ P(D^+ | T^+) = \frac{0.0297}{0.0491} \]
\[ = \frac{0.0297}{0.0491} \]
\[ = \boxed{0.6049} \]
Problem 11. Based on past drug testing of air traffic controllers, the FAA reports that the probability of drug use at any given time is approximately 0.007. The FAA uses a particular test to determine if the air traffic controllers are currently using drugs is 96% sensitive and 93% specific.

a. What is the probability of a positive test? $p(T^+)$

b. If a test is positive, what is the probability that the individual is actually using drugs?
Problem 11. Based on past drug testing of air traffic controllers, the FAA reports that the probability of drug use at any given time is approximately 0.007. The FAA uses a particular test to determine if the air traffic controllers are currently using drugs is 96% sensitive and 93% specific.

a. What is the probability of a positive test?

b. If a test is positive, what is the probability that the individual is actually using drugs?

\[ P(T^+) = P(T^+ \cap D^+) + P(T^+ \cap D^-) \]

\[ = 0.00672 + 0.06951 \]

\[ = 0.07623 \]

\[ P(D^+ | T^+) = \frac{P(D^+ \cap T^+)}{P(T^+)} \]

\[ = \frac{0.00672}{0.07623} \]

\[ = 0.0882 \]