ASSIGNMENT 3- ANSWER KEY

Q2. (15 P)
The sample space of a random experiment consists of the integers 0, 1, 2, 3, 4.
The probabilities assigned to each simple event are given below.

<table>
<thead>
<tr>
<th>e_i</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(e_i)</td>
<td>0.015</td>
<td>0.235</td>
<td>0.425</td>
<td>0.245</td>
<td>0.08</td>
</tr>
</tbody>
</table>

(a) Is this an acceptable probability assignment? Why?

\[ \begin{align*}
(3) & \quad \text{yes, because } 0 \leq P(e_i) \leq 1 \quad \forall e_i \quad \text{and} \quad \sum_{i=0}^{4} P(e_i) = 1.
\end{align*} \]

(b) Define the event \( E_1 = \) "The outcome of the trial is smaller than 3". What is \( P(E_1) \)?

\[ \begin{align*}
(3) & \quad P(E_1) = 0.015 + 0.235 + 0.425 = 0.675 \\
& \quad = P(0) + P(1) + P(2)
\end{align*} \]

(c) Define \( E_2 = \) "The outcome is an odd number". Find \( P(E_1 \cup E_2) \) and \( P(E_1 \cap E_2) \).

Give a relation linking both. Use this relation to find \( P(E_2) \).

\[ \begin{align*}
(2) & \quad P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \\
(3) & \quad P(E_1 \cup E_2) = 0.015 + 0.235 + 0.425 + 0.245 = 0.92 \quad \text{(odd or } < 3) \\
(3) & \quad P(E_1 \cap E_2) = 0.235 \quad \text{(odd and } < 3 \Rightarrow P(e_i = 1)) \\
& \quad P(E_2) = 0.92 - 0.675 + 0.235 = 0.48
\end{align*} \]
Q3. 6 data points are given as follows:

<table>
<thead>
<tr>
<th>x</th>
<th>12</th>
<th>8</th>
<th>14</th>
<th>11</th>
<th>16</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>45</td>
<td>56</td>
<td>41</td>
<td>47</td>
<td>39</td>
<td>34</td>
</tr>
</tbody>
</table>

\[ \bar{x} = 13.17 \quad \bar{y} = 43.67 \quad s_x = 7.9 \quad s_y = 26.2 \]

(a) Find the equation of the least squares regression line.

\[ \beta_1 = \frac{\sum xy}{\sum x^2} = \frac{\sum xy - (\sum x)(\sum y)/n}{\sum x^2 - (\sum x)^2/n} = \frac{3345 - 3,449.67}{1,105 - 1,040.17} = -2.08 \]

\[ \beta_0 = \bar{y} - \beta_1 \bar{x} = 43.67 - (-2.08)(13.17) = 71.06 \]

\[ \hat{y} = 71.06 - 2.08x \]

(b) Find the correlation coefficient \( r \) and interpret it.

\[ r = \frac{\sum xy}{\sqrt{\sum x^2 \sqrt{\sum xy}}} = \frac{-124.67}{\sqrt{1,105 \sqrt{3345}}} = -0.987 \]

There is an almost linear correlation relationship with a negative slope.

(c) By considering the equation you found in (a), find the predicted \( y \)-value for \( x = 16 \).

\[ \hat{y} = 71.06 - 2.08 \times 16 = 37.78 \]

(d) Why is the regression line also called the least squares fitted line?

Because the sum of the squares of the distances of the data points to the regression line is the smallest value.
Q4. (15P)
An experiment consists of the following steps: Toss a coin; if the outcome is tails, roll a die, and record the number appearing on the top face; if the outcome is heads, roll two dice, and record the numbers appearing on the two top faces.

(a) Describe the sample space. How many elements does it have?

\[ S = \{ (H, 1), (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6), (H, 1, j) \mid i, j \in \{1, \ldots, 6\} \} \]

\[ |S| = 7 + 6 \times 6 = 42 \]

(b) Is the equally likely principle applicable? Explain your answer.

No.

\[ P((T, i)) \neq P((H, i, j)) \]

\[ i \in \{1, \ldots, 6\} \]

(c) Define the event E as "the sum is 3". (If one die is rolled, the outcome is a 3; if two are rolled, the sum of the appearing two numbers is 3). What is the probability of E?

\[ P(T, 3) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12} \]

\[ P(H, 1, 2) = \frac{1}{2} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \]

\[ P(H, 2, 1) = \frac{1}{2} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{92} \]

\[ \text{Therefore, } P(E) = \frac{8}{72} = \frac{1}{9} \]
Q5. (15 P)
The IQ (Intelligence Quotient) scores of seven people are 86, 90, 94, 100, 108, 115, and 177.

(a) Find the mean and median for this data set.

(2) The mean is
\[ \bar{x} = \frac{\sum x_i}{n} = \frac{770}{7} = 110 \]

For the median, the data is already sorted in increasing order: 86, 90, 94, 100, 108, 115, 177 and 100 is the middle value; or

(2) \[ np = \frac{7 	imes 100}{7} = 100 \]

so median is 100.

(b) Which is a better measure of center for this type of data? Why?

(4) Median is a better measure of center, because the person with IQ score of 177 is an outlier (it is much larger than the rest of the scores). And mean is affected by outliers, but not the median.

(c) Compare the range (sample range) and interquartile range for this data set.

(15) Sample Range = largest - smallest = 177 - 86 = 91

(15) IQR = Q_3 - Q_1, where Q_3 = 115 since np = 7 	imes (.75) = 5.25 \text{ 6th value of 7}

and Q_1 = 90, since np = 7 	imes (.25) = 1.75 \text{ 2nd value of 7}

so IQR = 115 - 90 = 25.

(4) IQR is a better measure, since it is not influenced by outliers. (Various other versions are also accepted)

(d) An IQ of 120 is said to be the 95th percentile. Explain what this means!

(3) At least 95% of the IQ scores are at or below 120 and at least 5% of the IQ scores are at or above 120.
Q-6:
\[ \hat{y} = 1.8 + 1.6x \]
\[ 1.6 = \frac{S_{xy}}{S_{xx}} \]
\[ a) S_{xx} = 30 - \frac{1}{5} \times (10^2) \]
\[ S_{xx} = 10 \]
b) \[ 1.6 = \frac{S_{xy}}{10} \] so \[ S_{xy} = 16 \]
c) \[ 1.8 = \hat{y} - 1.6 \times 2 \] so \[ \hat{y} = 5 \]

1) 1000 football fans were asked two questions: “Which team do you like?” and “Will they win the championship this year?” 400 said that they like FB, 350 like GS and 250 like BJK. Of these, 50 FB fans, 140 GS fans and 210 BJK fans don’t believe that their team will finish first.

(a) Picking a person at random, what is the (empirical) probability that he/she is a BJK fan?

(b) Given that a person picked at random is a FB fan, what is the (empirical) probability that he/she believes FB will finish first?

(c) Define the events A=’GS supporter’, and B=’Believes his/her team will be the champion this year’. Are these events independent? Show this in two different ways.

\[ \frac{210}{1000} = 0.21 \]

\[ P(CH/FB) = \frac{P(CH \cap FB)}{P(FB)} = \frac{0.350}{0.400} = 0.875 \]

\[ P(A) = 0.35 \quad P(B) = 0.6 \quad P(A \cap B) = \frac{210}{1000} = 0.21 \]

\[ P(A) \quad P(B) = (0.35)(0.6) = 0.21 = P(A \cap B) \rightarrow \text{Independent} \]

or \[ P(A/B) = \frac{210}{600} = 0.35 = P(A) \]

or \[ P(B/A) = \frac{210}{350} = 0.6 = P(B) \]

1st 5, 2 x 1 4
2) (a) Write an expression for the probability that 5 cards drawn from a standard deck of 52 will have 4 hearts and 1 club?

(b) In an experiment two fair dice are tossed and their sum is recorded. What is the probability that a 7 or 11 is the outcome?

\[
\text{(5)} \quad a) \quad \frac{\binom{13}{4} \times \binom{13}{1}}{\binom{52}{5}} \rightarrow \frac{2}{3}
\]

\[
\text{(6)} \quad b) \quad \begin{align*}
7 &= 1+6 \quad 6+1 \\
&= 2+5 \quad 5+2 \\
&= 3+4 \quad 4+3
\end{align*} \Rightarrow \text{Prob}(\text{sum} = 7) = \frac{1}{36} \quad \text{6} = \frac{1}{6}
\]

\[
\begin{align*}
11 &= 5+6 \\
&= 6+5 \\
\Rightarrow \text{Prob}(\text{sum} = 11) &= \frac{1}{36} \quad \text{2} = \frac{1}{18}
\end{align*}
\]

\[
\Rightarrow \text{Prob}(\text{sum} = 7 \text{ or } 11) = \frac{1}{6} + \frac{1}{18} = \frac{2}{9}
\]
3) The horse Fredo will jump over the fence with probability 0.8 if Jim is the rider, and with probability 0.6 if Joe is the rider. Today, either Jim or Joe will ride the horse with probabilities 0.4 and 0.6, respectively.

(a) Construct a tree diagram to obtain all elements of the sample space and assign probabilities to these simple events.

(b) What is the probability that the horse will not jump over the fence today?

(c) We have been told that the horse has jumped over the fence. What is the probability that Jim was the rider?

\[
\begin{align*}
\text{(8) } & \quad a) \quad \begin{array}{c}
\text{Jim} \\
0.4
\end{array} \\
\phantom{a) } & \quad \begin{array}{c}
\text{Jump} \\
0.8 \\
\text{No jump} \\
0.2
\end{array} \\
\phantom{a) } & \quad \begin{array}{c}
\text{P(Jim and Jump)} = (0.4)(0.8) = 0.32 \\
\text{P(Jim and No jump)} = (0.4)(0.2) = 0.08
\end{array}
\end{align*}
\]

\[
\begin{align*}
\text{(8) } & \quad a) \\
\phantom{a) } & \quad \begin{array}{c}
\text{Joe} \\
0.6
\end{array} \\
\phantom{a) } & \quad \begin{array}{c}
\text{Jump} \\
0.6 \\
\text{No jump} \\
0.4
\end{array} \\
\phantom{a) } & \quad \begin{array}{c}
\text{P(Joe and Jump)} = (0.6)(0.6) = 0.36 \\
\text{P(Joe and No jump)} = (0.6)(0.4) = 0.24
\end{array}
\end{align*}
\]

\[
\begin{align*}
\text{(4) } & \quad b) \quad \text{No } P(\text{No jump}) = P(\text{Jim and No jump}) + P(\text{Joe and No jump}) \\
& \quad = 0.08 + 0.24 = 0.32
\end{align*}
\]

\[
\begin{align*}
\text{(6) } & \quad \text{c) } P(\text{Jim | Jump}) = \frac{P(\text{Jim and Jump})}{P(\text{Jump})} = \frac{0.32}{0.68} = (1 - 0.32 \text{ from } b)
\end{align*}
\]
4) We are given the points (2,1), (3,2) and (4,2).

(a) For the line \( y = x - 1 \) compute the residual sum of squares (the sum of squares of the errors).

(b) Find the equation of the regression line (line fitted by least squares).

(c) What is the residual sum of squares (the sum of squares of the errors) for the line in part (b). Why is it called the least squares line?

\[
\begin{align*}
(a) \quad e_1 &= 1 - (2 - 1) = 0 \\
& \quad e_2 = 2 - (3 - 1) = 0 \\
& \quad e_3 = 2 - (4 - 1) = -1 \\
\text{Therefore} \quad \frac{3}{3} \sum (e_i)^2 &= 1
\end{align*}
\]

\[
\begin{align*}
(b) \quad \hat{y} &= \beta_0 + \beta_1 x \\
\beta_1 &= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\frac{\sum x_i^2}{n} - (\frac{\sum x_i}{n})^2} \\
\beta_0 &= \bar{y} - \beta_1 \bar{x}
\end{align*}
\]

\[
\begin{align*}
\beta_1 &= \frac{2 \cdot 1 + 3 \cdot 2 + 4 \cdot 2}{3} - \frac{2 \cdot (2 + 3 + 4)}{3} = 0.5 \\
\beta_0 &= \frac{2 + 3 + 4}{3} - 0.5 \cdot 3 = 0.47
\end{align*}
\]

Therefore \( \hat{y} = 0.47 + 0.5x \)

\[
\begin{align*}
(c) \quad e_1 &= 1 - (0.47 + 0.5 \cdot 2) = -0.17 \\
& \quad e_2 = 2 - (0.47 + 0.5 \cdot 3) = 0.33 \\
& \quad e_3 = 2 - (0.47 + 0.5 \cdot 4) = -0.17 \\
\frac{3}{3} \sum (e_i)^2 &= 0.0283 + 0.1083 + 0.0283 = 0.1677
\end{align*}
\]

This line is called the least squares line because the aim of the regression process is to fit a line which results in minimized error sum of squares which is achieved in part (b).
5) 20 data points, some sums and the correlation coefficient are given below.

<table>
<thead>
<tr>
<th>x</th>
<th>15</th>
<th>60</th>
<th>10</th>
<th>36</th>
<th>50</th>
<th>32</th>
<th>18</th>
<th>71</th>
<th>96</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>85</td>
<td>52</td>
<td>91</td>
<td>59</td>
<td>51</td>
<td>52</td>
<td>74</td>
<td>52</td>
<td>25</td>
<td>56</td>
</tr>
</tbody>
</table>

\[
\sum_{i=1}^{20} x_i = 956, \quad \sum_{i=1}^{20} y_i = 1102, \quad \sum_{i=1}^{20} x_i^2 = 60712, \quad \sum_{i=1}^{20} y_i^2 = 66592, \quad \sum_{i=1}^{20} x_i y_i = A, \quad r = -0.903.
\]

(a) What is the value of \( S_{xy} \)?

\[
S_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n} = S_{xy} \approx -84.38 \quad \text{[Linear Regression Calculations]}
\]

(b) What does \( r = -0.903 \) tell you? Explain.

(c) What is \( A \) (the sum of the cross products of the \( x_i \) and \( y_i \))?  

\[
A = \sum x_i y_i = 956 \times 1102 = 1050752.
\]

(d) Find the mean and standard deviation for the \( x_i \) and the mean for the \( y_i \).

\[
\bar{x} = \frac{\sum x_i}{n} = \frac{956}{20} = 47.8, \quad \bar{y} = \frac{\sum y_i}{n} = \frac{1102}{20} = 55.1
\]

\[
\text{Var}(x) = \frac{1}{n} S_{xx} = \frac{1}{20} (956 - \frac{956^2}{20}) = 758.76 \rightarrow \text{Std}(x) = 28.4
\]
6) The histogram of a set of discrete numerical data is shown below.

(a) Find the mode and the median.
(b) Find the sum of the data points. What is the sum of squares of the deviations?
(c) The 100p-th percentile is 2.5. What is p? (Give the answer in rational form)
(d) What is the interquartile range?
(e) What is the relative frequency of 2?

3 4 a) Mode = the most frequent = 4
b) Median = 50th percentile = \( \frac{n \cdot p}{100} = 45 \cdot 0.5 = 22.5 \) th = 2

5 c) Next last data point at 2 is 3.5th, \( k = 3.5 \Rightarrow np = 35 \Rightarrow p = \frac{35}{45} \)

4 d) \( \frac{45}{4} = 11.25 \rightarrow 12 \quad Q_1 = 0 \)

2 e) \( \frac{9}{45} \)
In the semifinals of a tennis tournament, Sharapova plays against Pierce and Myskina plays against Williams. The winners will play in the final, and the losers will play for third place. The probability for Sharapova beating Pierce is 0.7, for beating Myskina is 0.6, and for beating Williams is 0.4. Williams will beat Myskina with probability 0.6.

a) What is the probability that Sharapova will win the tournament?

b) Given that Sharapova has lost her semifinal match, what is the probability that she will place third in the tournament?

c) What is the probability that Sharapova has won her semifinal match, given that she won her last match?

d) The winner gets $200000, the other finalist $150000, and the third-placed player will get $100000. What is the expected prize money for Sharapova?

Hint: A tree diagram helps a lot.

\[
\begin{align*}
&0.7 \Rightarrow \text{Win} \quad 0.6 \Rightarrow \text{Win} \\
&0.4 \Rightarrow \text{Myskina} \quad 0.6 \Rightarrow \text{Win} \\
&0.3 \Rightarrow \text{Loss} \quad 0.4 \Rightarrow \text{Williams} \quad 0.6 \Rightarrow \text{Loss} \\
&0.6 \Rightarrow \text{Myskina} \quad 0.4 \Rightarrow \text{Loss} \\
&0.012 \Rightarrow \text{Loss} \quad 0.048 \Rightarrow \text{Win} \\
&0.108 \Rightarrow \text{Win} \\
&0.072 \Rightarrow \text{Loss} \\
&0.072 \Rightarrow \text{Loss} \\
\end{align*}
\]

\[a) \quad (0.7)(0.6)(0.4) + (0.7)(0.4)(0.6) = 0.336\]
\[b) \quad (0.4)(0.4) + (0.6)(0.6) = 0.52\]
\[c) \quad P(WS/wL) = \frac{P(WS \& WL)}{P(wL)} = \frac{(0.7)(0.14 + 0.24)}{0.168 + 0.168 + 0.058 + 0.108} = 0.683\]
\[d) \quad (0.336)(200000) + (0.7(0.36 + 0.16))(150000) + 0.3(0.16 + 0.36)(100000) = 134000\]