Note: You may use a spreadsheet, but it’s not necessary to complete the exam. Only #1 requires calculations (and those calculations can also be done using a calculator.)

1. (20 points) Suppose a class of 15 students takes a quiz. The teacher records the students’ quiz scores as follows:
   \[ \{8, 3, 2, 7, 10, 9, 10, 0, 9, 8, 3, 2, 5, 4, 10\} \]

   (a) Create a frequency table for this data set:

   \[
   \begin{array}{|c|c|c|}
   \hline
   \text{Class} & \text{Frequency, } f & \text{Relative frequency} \\
   \hline
   0-1 & 1 & 1/15 \approx 0.067 \\
   2-3 & 4 & 4/15 \approx 0.267 \\
   4-5 & 2 & 2/15 \approx 0.133 \\
   6-7 & 1 & 1/15 \approx 0.067 \\
   8-9 & 4 & 4/15 \approx 0.267 \\
   10 & 3 & 3/15 \approx 0.2 \\
   \hline
   \end{array}
   \]

   (b) Use your frequency table to sketch a frequency histogram. Label both axes:

   ![Frequency histogram]

   (c) Calculate the following:

   \[
   \sum_{i=1}^{15} x_i = 90
   \]

   mean:

   \[
   \bar{x} = \frac{\sum_{i=1}^{15} x_i}{n} = \frac{90}{15} = 6
   \]
(d) Listed in the table below are the quiz scores (first column, under \( x \)). Fill in the rest of the table as follows:

- use your result for the mean \( \bar{x} \) from (c) to calculate each of the deviations and list them in the 2nd column
- then calculate each of the squared deviations (third column)
- finally calculate the sum of squared deviations \( SS_x \)

(Hint: you can enter the data set in a spreadsheet and do the calculations there.)

\[
\begin{array}{c|c|c}
 x & \text{Deviations: } x - \bar{x} & \text{Squared deviations: } (x - \bar{x})^2 \\
8 & 8 - 6 = 2 & 2^2 = 4 \\
3 & 3 - 6 = -3 & (-3)^2 = 9 \\
2 & 2 - 6 = -4 & (-4)^2 = 16 \\
7 & 7 - 6 = 1 & 1^2 = 1 \\
10 & 10 - 6 = 4 & 4^2 = 16 \\
9 & 9 - 6 = 3 & 3^2 = 9 \\
10 & 10 - 6 = 4 & 4^2 = 16 \\
0 & 0 - 6 = -6 & (-6)^2 = 36 \\
9 & 9 - 6 = 3 & 3^2 = 9 \\
8 & 8 - 6 = 2 & 2^2 = 4 \\
3 & 3 - 6 = -3 & (-3)^2 = 9 \\
2 & 2 - 6 = -4 & (-4)^2 = 16 \\
5 & 5 - 6 = -1 & (-1)^2 = 1 \\
4 & 4 - 6 = -2 & (-2)^2 = 4 \\
10 & 10 - 6 = 4 & 4^2 = 16 \\
\hline 
\text{SS}_x = \Sigma (x - \bar{x})^2 = 166 \\
\end{array}
\]

(c) Use your result from (d) to calculate the (sample) variance and standard deviation of the quiz scores (first express the sample variance as a ratio, and then divide to express it as a single number):

\[
\text{Solution:} \\
\frac{SS_x}{n-1} = \frac{166}{14} \approx 11.86 \\
s = \sqrt{s^2} = \sqrt{11.86} \approx 3.44
\]

(f) Now suppose the class takes another quiz, and every student gets a score of 6. What is the mean score on this quiz? What is the variance and the standard deviation? (You should be able to write these down immediately without doing any calculations!)

\[
\text{Solution:} \text{ If every student gets the same score, then clearly that score is the mean. And in that case, each of the deviations } x - \bar{x} = 0, \text{ so each of the squared deviations is also 0. So the sum of squared deviations } SS_x = 0, \text{ and thus the variance is 0, and thus the standard deviation is also 0.}
\]
2. (5 points) The graph below shows frequency histograms for two different data sets, labelled A & B. As indicated on the graph, the two data sets have the same mean.

Which data set has the higher standard deviation? Briefly (in 2-3 complete sentences) explain why.

(Hint: Recall that we described the standard deviation as a measure of \textit{variability} or \textit{dispersion}, and that it is defined in terms of the sum of squared deviations, i.e.,

$$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Think about which histogram exhibits greater variability, or equivalently, which data set will have a larger sum of squared deviations.)

![Histograms A and B with average 100]

\textbf{Solution:} The data set B clearly has the higher standard deviation, since the histograms show that the values of B are much more dispersed than those of A, i.e., on average the values of B are much further away from the mean than values of A, which are concentrated relatively close to the mean.
3. (5 points) The graph below (taken from a blog called “Stats in the Wild”) shows boxplots for the ages of Olympic athletes in 2000-2008 for each sport (sorted from left to right by median age):

(a) Which sport has the highest median age? Which sport has the lowest median age? What are the median ages in those sports (approximately)?

Solution:
- The sport with the highest median age is equestrianism (the rightmost sport/boxplot), with a median age of approximately 36.
- The sport with the lowest median age is women’s gymnastics (the leftmost sport/boxplot), with a median age of approximately 18.

(b) Recall that the interquartile range (IQR) is defined as

\[ IQR = Q_3 - Q_1 \]

Which 3 sports appear to have the largest interquartile ranges? Which 3 sports appear to have the smallest IQRs?

Solution:
- The 3 sports with the largest IQRs seem to equestrianism, men’s shooting, and women’s shooting (men’s archery and women’s archery also have large IQRs).
- The 3 sports with the smallest IQRs seems to be women’s gymnastics, women’s rhythmic gymnastics, and men’s football (i.e., soccer).

The blog includes the following description:

“Here is a graph with a linear fit depicting the relationship between 175 countries’ life expectancies and corresponding mean years of schooling in each country. Clearly, a direct relationship exists.”

(a) What is the nature of the “direct relationship” between life expectancy and average years of schooling? Are they positively or negatively correlated?

**Solution:** Strong positive correlation: life expectancy tends to increase as average years of schooling increases.

(b) Which of the following is closest to the correlation coefficient for this paired data set?

(I) 0.7  (II) 0.1  (III) −0.1  (IV) −0.7

**Solution:** Since the scatterplot shows a strong positive correlation, the correlation coefficient must be positive and relatively close to 1, so the only reasonable choice is (I) = 0.7.

(c) The graph includes the linear regression line and its equation: “Life Exp = 50.47 + 2.44*Yrs Sch”. What do the numbers 50.47 and 2.44 represent in terms of the line and its equation? What do the numbers represent in terms of the relationship between the variables, i.e., in terms of life expectancy and average years of schooling (according to the linear regression model)?

**Solution:** 50.47 is the y-intercept of the regression line, which represents the linear regression model’s predicted life expectancy in a country with 0 average years of schooling. 2.44 is the slope of the regression line, which represents the gain in a country’s life expectancy if its average years of schooling increases by 1.

(d) Use the equation of the regression line to predict the life expectancy in countries with the following mean years of schooling (i.e., plug in the given values of ”Yrs Sch” into the given equation of the linear regression line to get the corresponding (predicted) value for “Life Exp”):

(i) 1 year:

**Solution:** Life Exp = 50.47 + 2.44*1 = 52.91

(ii) 5 years:

**Solution:** Life Exp = 50.47 + 2.44*5 = 50.47 + 12.2 = 62.67
The blog post also has the following scatterplot, for Child Mortality Rate vs. Mean Years of Schooling:

(e) What is the relationship between child mortality rate and average years of schooling? Are they positively or negatively correlated?

**Solution:** Strong negative correlation: a country’s child mortality rate tends to decrease as its average years of schooling increases.

(f) Extra credit: Use the graph of the linear regression line to estimate the linear regression parameters.

(Hint: Extend the line to the left to the $y$-axis in order to estimate the $y$-intercept, and then use that value together with the fact that the line appears to cross the $x$-axis at $x = 11$ in order to calculate the slope.)

The regression line is: “Mortality Rate = $\alpha + \beta \cdot \text{Yrs. Sch}$” where

\[
\alpha \approx \quad \beta \approx
\]

**Solution:**

The $y$-intercept $\alpha \approx 140$, since that’s the approximate $y$-value where the regression line intersects the $y$-axis.

To calculate the slope, we need the coordinates of two points on the line, so that we can calculate (rise/run). From our estimate of the $y$-intercept, we know the point $(0, 140)$ is on the line, and since the regression line appears to cross the $x$-axis at $x = 11$, the point $(11, 0)$ is also on the line. Hence we can compute the slope as

\[
\beta \approx \frac{0 - 140}{11 - 0} \approx -12.7
\]

We can interpret the regression parameters in the usual way: the $y$-intercept $\alpha \approx 140$ represents the regression model’s prediction of child mortality rate in a country with 0 years average schooling, and the slope $\beta \approx -12.7$ represents that a country’s child mortality rate will decrease by 12.7 years for each additional year of average schooling.
5. (5 points) Consider the probability experiment consisting of flipping a coin three times in a row.

(a) What is the sample space of the experiment?

Hint: There are $2^3 = 8$ elements in the sample space; one of them is the coin coming up heads (H) all three times, which we can represent as “HHH”. List the other 7 elements of the sample space:

Solution:

\[ \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \]

(b) Now consider the events

- $A$ = “getting heads three times in a row”
- $B$ = “getting heads twice and tails once” (i.e., in any order)
- $C$ = “getting heads once and tails twice” (in any order)
- $D$ = “getting tails three times in a row”

For each of these events, list the outcomes that make up the event (as already written below for $B$):

Solution:

\[ A = \{HHH\} \]

\[ B = \{HHT, HTH, THH\} \]

\[ C = \{HTT, THT, TTH\} \]

\[ D = \{HHH\} \]

Note that therefore, according to the definition of “theoretical” probability:

\[ P(A) = \frac{1}{8} = 0.125 \]

\[ P(B) = \frac{3}{8} = 0.375 \]

\[ P(C) = \frac{3}{8} = 0.375 \]

\[ P(D) = \frac{1}{8} = 0.125 \]