Disclaimer: It is essential to write legibly and show your work. If your work is absent or illegible, and at the same time your answer is not perfectly correct, then no partial credit can be awarded. Completely correct answers which are given without justification may receive little or no credit.

During this exam, you are not permitted to use notes, or books, nor to collaborate with others. You are allowed to use a calculator.
Problem 1 [24 points.] Answer the following as TRUE or FALSE.

(a) If the correlation between two variables is +0.07, there is a strong association between the variables.
(b) It is possible for the correlation between two variables to be -0.97.
(c) If a distribution is skewed to the right, the mean will probably be higher than the median.
(d) In regression, it is possible to have an influential point that is not a high leverage point.
(e) If $A$ and $B$ are events such that $P(A) = \frac{2}{3}$ and $P(B) = \frac{1}{3}$, then $A$ and $B$ cannot be both disjoint and independent.
(f) If $A$ and $B$ are events such that $P(A) = \frac{2}{3}$ and $P(B) = \frac{1}{3}$, then $P(A \text{ and } B)$ must equal 2/9.
(g) If $A$ and $B$ are events such that $P(A) = \frac{2}{3}$ and $P(B) = \frac{1}{3}$, then $P(A \text{ or } B)$ must equal 1.
(h) An outlier will probably have a greater effect on the median than on the mean.

Solutions: FTFT TFFF
Problem 2. Suppose we have data on the graduation year and Mathematical Reviews author ID (author ID, for short) of 25 University of Hawai‘i math PhD graduates. Assume the graduation years have a mean of 1992.1 and a standard deviation of 9.6. Next, suppose the author ID numbers have a mean of 479,867 and a standard deviation of 299,056. Suppose the correlation between the graduation year and the author ID number is .78.

a) [10 points] Find the equation of a regression line that can be used to predict the author ID from the year of graduation.

Solution: Here author ID is the y-variable and graduation year is the x-variable. Therefore, the slope of the regression line is

\[ b_1 = \frac{r_{xy}}{s_x} = \frac{(0.78)(299,056)}{9.6} \approx 24,298.3. \]

The intercept is

\[ b_0 = \bar{y} - b_1 \bar{x} = 479,867 - 24,298.3(1992.1) \approx -47,924,776.4. \]

Therefore, the regression line is given by

\[ \hat{\text{author ID}} = -47,924,776.4 + 24,298.3 \cdot \text{Graduation year}. \]

b) [6 points] Write a sentence or two explaining whether the slope of the regression line that you found in part a) seems to match the slope of the line in Figure 1.

Solution: Yes, it looks like over a decade the change in \( \hat{y} \) is about 250,000 or 10 times 25,000.
Problem 3. The Facility Condition Needs Index (FCNI) for a building is the cost to repair the building divided by the cost to replace the building.

On the University of Hawai‘i main campus, buildings that are built around 1962-1963 tend to have the highest FCNI, i.e., to be in the worst shape. In the scatterplot in Figure 2, the distance of a year \( n \) to 1962.5 is the absolute value of the difference, \(|n - 1962.5|\). The scatterplot shows FCNI values for all buildings on the campus.

(a) [8 points] The Physical Science Building was built in 1960. What is its predicted FCNI value?

Solution: \( 0.379 - 0.00505(3) = 0.36385 \approx 36\% \).

(b) [8 points] The actual FCNI value of the Physical Science Building is 0.46. Calculate its residual.

Solution: \( 0.46 - 0.36 = 0.1 = +10\% \).

(c) [8 points] In what year(s) is the FCNI predicted to be 0?

Solution: We solve the equation \( 0.379 - 0.00505x = 0 \). Then \( x = 0.379/0.00505 = 75 \). Therefore the year is 1962.5 ± 75, which is 2037.5 or 1887.5.

However, one should really do regression with the variables interchanged first. So this was not a good question.

Figure 2: Scatterplot of FCNI vs Distance to 1962.5 Regression Analysis: FCNI versus Distance to 1962.5. The regression equation is FCNI = 0.379 - 0.00505 * (Distance to 1962.5).
(d) [4 points] The $R^2$ value for this regression is 19.3%. Write a sentence about whether the regression seems to explain the data in Figure 2.

Solution: The data seem to be only partially explained by the regression. There is a lot of unexplained variation. But the general trend of the data looks like it is in agreement with the regression line.
Problem 4. [Background: The web browser Firefox is a program that can run under several operating systems (Windows, Mac OS X, etc.). But there are other browsers competing with Firefox, such as Internet Explorer and Chrome.]

During the last month, 38% of the web browsing visitors to math.hawaii.edu used the browser Firefox, whereas 58% used the operating system Windows. Moreover, 43% percent of the Windows users were using Firefox.

a) [6 points] What is the probability that a randomly chosen visitor is using Firefox running on Windows?

b) [7 points] What is the probability that a randomly chosen visitor is either using Firefox or running Windows?

c) [7 points] What is the probability that a randomly chosen visitor is a Windows user who is not using Firefox?

Solution: (a) Let $W$ be the event that Windows was used, and $F$ the event that Firefox was used. We have

$$P(W \text{ and } F) = P(W)P(F|W) = (.58)(.43) = 24.94\% \approx 25\%.$$  

(b) This is

$$P(W \text{ or } F) = P(W) + P(F) - P(W \text{ and } F) = .58 + .38 - .2494 = 71.06\% \approx 71\%.$$  

(c) This is

$$P(W \text{ and } F^c) = P(W) - P(W \text{ and } F) = .58 - .25 = 33\%.$$
Problem 5.  [12 points] Suppose that 95% percent of math PhD graduates from University of Hawai‘i have a Math Genealogy identification number (for short, “genealogy ID”). Among those who have a genealogy ID, 56 percent also have a Mathematical Reviews author ID (“author ID”). Among those who do not have a genealogy ID, 50% nevertheless have an author ID.

If a math PhD graduate from University of Hawai‘i is chosen at random, what is the probability that that person will have an author ID?

Solution: Let $A$ be the event that the person has an author ID. Let $G$ be the event that the person has a genealogy ID. We have

$$P(A) = P(G)P(A|G) + P(G^c)P(A|G^c) = (.95)(.56) + (.05)(.50) = .557 = 55.7\%.$$
Useful formulas

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]

\[
s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

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

\[
\hat{y} = b_0 + b_1 x
\]

\[
b_1 = \frac{rs_y}{s_x}
\]

\[
b_0 = \bar{y} - b_1 \bar{x}
\]

\[
P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)
\]

\[
P(A \text{ and } B) = P(A)P(B \mid A)
\]

\[
P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}
\]

\[
P(A) = P(B)P(A \mid B) + P(B^c)P(A \mid B^c)
\]

\[
P(B \mid A) = \frac{P(B)P(A \mid B)}{P(B)P(A \mid B) + P(B^c)P(A \mid B^c)}
\]

<table>
<thead>
<tr>
<th>Problem</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/24</td>
</tr>
<tr>
<td>2</td>
<td>/16</td>
</tr>
<tr>
<td>3</td>
<td>/28</td>
</tr>
<tr>
<td>4</td>
<td>/20</td>
</tr>
<tr>
<td>5</td>
<td>/12</td>
</tr>
<tr>
<td>Total</td>
<td>/100</td>
</tr>
</tbody>
</table>