1. The correlation between I.Q. scores and school G.P.A. is $r = 0.634$. The correlation between wine consumption and heart disease deaths is $r = -0.645$. Which of these two correlations indicates a stronger straight-line relationship? The correlation of -0.645 indicates a stronger relationship, as the magnitude of the correlation is what tells us about the strength.

2. The following figure is a scatterplot of school G.P.A. versus I.Q. score for all 78 seventh-grade students in a rural Midwest school. Points A, B, and C might be called outliers.

(a) Describe the overall pattern of the relationship in words. (direction, form, strength) Excluding points A, B and C and a few others close to these points, the data cloud can be encompassed by an oval or a football shape. This suggests a linear form for the relationship between GPA and IQ. By the upward slope of the cloud, we can tell that the two variables are positively associated. That is we are seeing points mainly in the top right and bottom left quadrant of the plot. Finally, the clustering about a straight line is stronger for the higher IQ scores than for the lower where we can see a loosening of the points. So, overall, there is probably a roughly moderate association.

(b) About what is the I.Q. and G.P.A. for student A? The co-ordinates for student “A” are IQ = 103 and GPA = 0.5.

(c) For point C, say how it is unusual. Point C is unusual because it does not appear to follow the trend of high IQ and high GPA. Instead this student has a high GPA but is on the lower side for IQ.

(d) Will removing the outliers increase or decrease the correlation? Visually, removing the points should tighten the clustering and therefore increase the correlation. Of course,
their actual impact will depend on their standard scores in both the X and Y direction. For example, both A and B seem to be fairly close in the horizontal direction to the average of the X values. This means that they were having relatively little impact on the correlation calculation anyway. The presence/absence of C is likely to have the most impact.

3. Drinking moderate amounts of wine may help prevent heart disease. Data on the yearly wine consumption (liters of alcohol from drinking wine per person) and yearly deaths from heart disease (deaths per 100,000 people) in 19 developed countries in 2001 show a linear association between the two variables. The least squares regression line for predicting heart disease from wine consumption is calculated as:

\[ y = 115.86 - 8.05x \]

(a) Interpret the slope and intercept in the above equation. Slope: associated with an increase of 1 liter of alcohol there is a corresponding average decrease of 8.05 deaths per 100,000 people. Intercept: this is the average number of deaths (per 100,000) for a country where the per-capita wine consumption is 0 liters.

(b) Use the equation to predict the heart disease death rate in a country where adults average 1 liter of alcohol from wine each year. What is the size of the error in your prediction? The predicted death rate for a country with 1 liter of per capita wine consumption can be obtained by plugging in \( x = 1 \) in the equation above. This is 115.86 - 8.05 or 107.81 deaths per 100,000 on average. The size of the error in prediction is given by the formula \( \sqrt{1 - r^2} \) times the S.D. of heart disease death rates.

(c) Suggest some differences between among nations that may be confounded with wine drinking habits. Diet and lifestyle could be potentially important confounding variables because they are associated with drinking habits and also impact heart disease death rates.

(d) Suppose we used the correlation for the countries to estimate the correlation for individuals. Would the result be way off or about right? Explain your answer. When we calculate correlations for countries, our \( x \) and \( y \) co-ordinates are rates. Rates are basically averages. Compared to the original raw data, rates tend to be “smoother” and don’t show the wild variations you might see at the individual level. So a correlation based on rates would be higher than a correlation calculated from the raw data. To see why, recall that the standard scores go into calculating \( r \). Standard scores have the S.D. of \( x \) and the S.D. of \( y \) in the denominator. If the S.D.’s are small (as for rates) then the standard score is high and therefore the correlation is high. If the S.D.s are big (as for individual data) then the standard scores are lower and therefore the correlation is small.

4. Probabilities work not by COMPENSATING for imbalances but by OVERWHELMING them. Suppose the first ten tosses of a coin give 10 tails, and that the tosses after that are exactly half heads and half tails. (Exact balance is unlikely, but this example illustrates how the first 10 outcomes are swamped by later outcomes.) What is the proportion of heads after
the first 10 tosses? What is the proportion of heads after the first 100 tosses if half of the last 90 produce heads (45 heads)? What is the proportion of heads after 1,000 tosses if half of the last 990 produce heads? What is the proportion of heads after 10,000 tosses if half of the first 9990 produces hits? The proportion of heads after the first 10 tosses is 0. After the first 100 tosses, the proportion of heads is 0.45. After 1000 tosses the proportion of heads is 0.495. After 10,000 tosses the proportion is 0.4995. The lesson here is what happens in any finite run of tosses has no lasting impact on the long run probability.

5. You are getting to know your new roommate, assigned to you by the college. In the course of a long conversation, you find that both of you share a birthday. Should you be surprised? Explain your answer. It all depends on how you view things. If you look at it from the perspective of this event happening to you in particular, then yes maybe you should be surprised. This is because when viewed from your individual perspective, this event does have a small probability and is therefore unlikely. However, if you think about the probability that two new roommates (somewhere) will share a birthday it is not a very unlikely event. You just got selected to experience the co-incidence.

6. The following exercise illustrates the idea of a sampling distribution from a very small population. The population is the scores of 10 students on an exam. No answer is provided since individual answers will vary

<table>
<thead>
<tr>
<th>Student</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>82</td>
<td>62</td>
<td>80</td>
<td>58</td>
<td>72</td>
<td>73</td>
<td>65</td>
<td>66</td>
<td>74</td>
<td>62</td>
</tr>
</tbody>
</table>

The parameter of interest is the mean score in this population. The sample is an S.R.S. of size $n = 4$ from the population. Because the students are labelled from 0 to 9, you can use the roll of a ten-sided die to choose a student from the sample.

(a) Find the mean of the 10 scores in the population. This is the population mean and our parameter.

(b) Use your dice to draw an S.R.S. of size 4 from this population. Write the four scores in your sample and calculate the mean of the sample scores. This statistic is an estimate of the population mean.

(c) Repeat this process 10 times using different die rolls each time. Make a histogram of the 10 values of the sample mean. You are constructing the sampling distribution of the sample mean. Is the center of your histogram close to the population mean you found in (a)?