For questions 1 and 2: verbal SAT scores and math SAT scores of students in a certain county are positively correlated and the scatter diagram is football-shaped. The verbal scores have an average of 510 and a SD of 90; the math scores have an average of 480 and an SD of 120.

1. [3] One student scores 600 on both tests. Her math score (circle one)

is much less than is about equal to is much larger than cannot be compared to based on the information given the average math score of students with a 600 verbal score. Explain briefly.

(There was a correction; “Her math score” was originally “her verbal score”.)

600 verbal is \((600 - 510)/90 = 1\) standard deviation above average. The predicted math score will therefore be \(r\) standard deviations above average, where \(r\) is the correlation coefficient. Since \(r < 1\) the predicted math score will be less than \(480 + 120 = 600\); therefore her actual score is larger than the predicted score.

2. [3] Consider the following two events:

(i) a student chosen at random from \(\) all students \(\) scores above 600 on the math SAT;

(ii) a student chosen at random from those scoring 510 on the verbal SAT scores above 600 on the math SAT.

Circle one:

(i) is more likely than (ii) (ii) is more likely than (i) both are equally likely

Explain briefly.

The distribution of math scores among all students is normal with mean 480 and SD 120. The distribution of math scores among students scoring 510 on verbal is normal with mean 480 and SD \(\sqrt{1-r^2} \times 120\), which is less than 120; thus these students are less likely to score higher than 600.
For questions 3 and 4, consider the following data set:

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

3. [5] Find the coefficient of correlation between x and y. **Show your work.**

The average value of x is \((1 + 2 + 2 + 3 + 3 + 3)/6 = 14/6\);

The average value of \(x^2\) is \((1^2 + 2^2 + 2^2 + 3^2 + 3^2 + 3^2)/6 = 6\); thus \(SD(x) = \sqrt{6 - (11/6)^2} = 5/9\).

The average, average-square, and SD of y are the same as those for x, since they’re the same numbers in a different order.

The average of \(xy\) is \((1 \cdot 3 + 2 \cdot 2 + 2 \cdot 3 + 3 \cdot 1 + 3 \cdot 2 + 3 \cdot 3)/6 = 31/6\).

Therefore we have

\[
    r = \frac{ave(xy) - ave(x)ave(y)}{SD(x)SD(y)} = \frac{(31/6) - (14/6)(14/6)}{\sqrt{5/9}\sqrt{5/9}} = -\frac{1}{2}.
\]

4. [2] What is the equation of the regression line for predicting y from x, in the form \(y = mx + b\)? **Show your work.**

The regression line passes through \((14/6, 14/6)\) and has slope \(r\frac{SD(y)}{SD(x)} = -1/2\); thus its equation is

\[
    (y - 14/6) = -\frac{1}{2}(x - 14/6)
\]

or, solving for y, \(y = -x/2 + 7/2\).

Alternatively, the graph of averages is linear, and this is its equation.
For questions 5 and 6, suppose we want to assess whether students learn more in statistics classes given online or in the traditional manner. We give a standardized test of statistics to students who took Stat 21 online in the summer of 2011 and to those who took Stat 21 in the lecture format in the fall of 2011.

5. [3] The online students do better on this test than the lecture students. Can we conclude that online instruction is better than traditional instruction? Why or why not?

No, we cannot conclude that online instruction is better than traditional instruction. Many confounders are possible, including but not limited to:

- semester: the online class was in the summer and the lecture class was in the fall. Perhaps students are better able to concentrate on classes in the summer (because they have less of them). Perhaps they are less able to concentrate in the summer (because it’s summer and our culture says summer is not for learning).

- speed of the class: summer classes are shorter than fall classes, which has an effect. (In my experience students learn less in summer classes than in fall or spring classes, since the increased speed means more students fall behind and even the conscientious ones don’t have time to digest the material.)

- discipline-specificity: maybe online instruction works for statistics but not for other subjects

- cheating: a lot of people said that people might cheat on the test in the online class. I did not have this in mind when writing the problem; I was picturing a traditional test even with the online class. In fact the actual online summer Stat 21 has an in-person final.

6. [3] This is an observational study. Do you believe it would be possible to design a controlled experiment to determine if students learn more in an online version of Stat 21 at Berkeley than in a traditional version of the course? If so, explain how you would carry out such an experiment. If not, explain why you believe it is impossible.

The answer I had in mind was: no, you cannot. In order to do such a randomized study you’d have to assign students to take the online class or the traditional class at random. This would not be reasonable as Berkeley students traditionally are able to sign up for whichever classes they want to.

Some people said that it would be possible if we had the students volunteer to be part of the study, and then randomly assigned the volunteers to be in an online class or a traditional class. This would work as well. Alternatively, if you have enough control over the students you can just force everybody to “volunteer”; I recall hearing about some such experiments at West Point.
Questions 7 through 10 are true-false, for 1 point each. **Circle TRUE or FALSE; no explanation is necessary.**

7. Given a histogram for a data set, we can compute the average of the data set exactly.

8. Given a histogram for a data set, we can determine which of the histogram’s class intervals the average falls into.

9. Given a histogram for a data set, we can compute the median of the data set exactly.

10. Given a histogram for a data set, we can compute which of the histogram’s class intervals the median falls into.

7, 8, 9 are false; 10 is true. The median can’t be computed exactly. But we know the percentage of the data which falls to the left of each class boundary, and we can look to see if 50 percent is between two of those boundaries. The mean can’t be computed exactly – we’d need the sum of the data to do that. Finally, for 7, consider two data sets: 9, 9, 11 and 9, 9, 13. If we construct histograms with bins 0 – 10 and 0 – 20 these have the same histogram; but the first has mean less than 10, the second has mean greater than 10.

Questions 11 and 12 refer to the figure below. **Circle a letter for each; no explanation is necessary.**

A data set has a histogram which is sketched below.

(Lines were A, B, C, D, and E, from left to right, where C was at the peak of the distribution.)

11. [ ] Which of the lettered vertical lines is at the median of the data set?
12. [ ] Which of the lettered vertical lines is at the average of the data set?

11: D. 12: E. The distribution is right-skewed; the median is to the right of the peak of the histogram, and the average is to the right of that.
Questions 13 through 16 refer to the following data.

The average height of adult men is 70 inches, with standard deviation 2.5 inches. The average height of adult women is 64 inches, with standard deviation 2.5 inches. Both the heights of men and the heights of women are normally distributed.

13. [2: 1 + 1] There are 2.5 centimeters in an inch. I measure the heights of many men; because I don’t like writing three-digit numbers, for each man I record the number of centimeters that his height is over one meter (100 centimeters). For example a man who is 74 inches tall would have his height recorded as \((74)(2.5) - 100 = 185 - 100 = 85\).

(a) What is the average of the numbers I record? Show your work.

Applying the usual rules for change of scale, I record \((70)(2.5) - 100 = 75\).

(b) What is the standard deviation of the numbers I record? Show your work.

Again by change of scale, \((2.5)(2.5) = 6.25\).

14. [2] Which is larger: the proportion of men over six feet tall, or the proportion of women under five feet tall? (There are twelve inches in a foot.) Explain your answer.

Men over six feet tall are at least 0.8 standard deviations above average; women under five feet tall are at least 1.6 standard deviations below average. The larger deviations are less likely, so the proportion of men over six feet tall is the larger.

15. [2] What percentage of women are between 66 and 67 inches tall? Show your work.

Women between 66 and 67 inches tall are between 0.8 and 1.2 SD above the average. From the table, the proportion of them is \((0.7699 - 0.5763)/2 = 0.0968\).

16. [1] The standard deviation of the heights of all adults (men and women) taken together is (circle one, no explanation needed)

- less than 2.5 inches
- equal to 2.5 inches
- greater than 2.5 inches
Below there are four scatterplots. Each scatterplot has three lines on it, labeled with letters. For each scatterplot you should:
(a) circle the number which is closest to the correlation coefficient;
(b) circle the letter of the regression line for predicting \( y \) from \( x \).

No explanation is necessary.

Problem 17

Problem 18

Problem 19

Problem 20

Answers to correlations: 0, 0, -0.5, +0.95. To lines: A, F, H, K. Here, looking along the left end of each figure, the lines are labeled B, A, C; D, E, F; G, H, I; M, K, L. Yes, I know this labeling is a bit strange.)