Statistics 21

Problems from past midterms: midterm 1

1. (5 points) The quotations below are taken from an article in the San Francisco Chronicle of May 31, 1989. The article begins:

> "In recent years, statistics has shown that drinking a little alcohol is good for your health."

(a) Has this been shown on the basis of controlled experiments?

The article continues:

> Some studies even suggest that moderate drinkers live longer than people who don't drink any alcohol.
>
> Researchers refer to this phenomenon as the U-shaped alcohol-mortality curve. Non-drinkers and heavy drinkers are on opposite sides of the curve with the highest death rates. Moderate drinkers fall in the middle, with fewer early deaths from disease.
>
> A new study of 7,700 British men, however, suggests that there's more to this curve than meets the eye."

The article goes on to describe how the investigators in the new study analyzed the non-drinking group. These were the men who, at the time of the study, said they were not drinking. As the investigators expected from the earlier studies, the non-drinkers, as a group, turned out to be not as healthy as the moderate drinkers. But the investigators went further. They found that a large number of the men in the non-drinking group had something in common. The investigators then removed the men with this common characteristic from the non-drinking group. They found that, overall, the remaining non-drinkers were just as healthy as the moderate drinkers.

(b) As a group, were the removed men more, or less, healthy than the moderate drinkers?

(c) What was the common characteristic?

2. (10 points) One number is missing from the data set below:

x	y
1	2
1	2
3	3
3	

If possible, fill in the blank to make the correlation coefficient equal to 0. If it is not possible, say why not.

3. (10 points) An instructor gives two quizzes to the ten people in her course. On the first quiz, five of the people were above average; but on the second quiz, these people all scored below average. The other five people moved in the opposite direction. They were all below average on the first quiz, and above average on the second one.

   (a) True or False and explain: This is an example of the regression effect.

   (b) The correlation coefficient between the scores on the two quizzes:

   _____ must be zero.
   _____ must be a positive number.
   _____ must be a negative number.
   _____ could be any of the above, depending on whether or not there are outliers in the data.

   Check (√) one option. Explain your choice.

4. (10 points) Here are the summary statistics for a large group of male students at a certain university.

   Height: Average = 70.0 inches, SD = 3.0 inches
   Weight: Average = 162 pounds, SD = 30 pounds
   correlation coefficient = 0.50

   The scatter diagram is football shaped.

   About what percent of the men 74 inches tall weigh less than the average weight of the men 66 inches tall?

5. (10 points) In the mid-1980's, the Educational Testing Service compared the SAT scores of college-bound seniors with those obtained by a large representative sample of high school juniors. On the verbal SAT, the 40th percentile of the scores for the college-bound seniors happened to be equal to the 60th percentile of the scores for the sample of juniors. For both groups, the SD was 100 points. The two histograms followed the normal curve closely.

   Find, approximately, the number of points separating the average of the two groups.

6. (10 points) Here are the summary statistics for a very large class:

   midterm score: average = 57 points, SD = 18 points
   final score: average = 52 points, SD = 25 points
   correlation coefficient = 0.80

   The scatter diagram is football shaped.

   (a) One person in the class scored 66 points on the midterm and 59 points on the final. What would be the regression prediction of his final score from his midterm score?

   (b) Out of all those who got the same score on the final as he did, what percent scored below him on the midterm?

7. (5 points) Here is a passage from a column in the San Francisco Sunday Examiner and Chronicle of September 27, 1987.

   "Scholastic Aptitude Test scores for 1987 were released last week. SATs

are administered to American high school seniors by the College Entrance Examinations Board. State Superintendent of Public Instruction Bill Honig called it a "noteworthy accomplishment" that California's seniors kept pace with the national average of 430 in verbal skills and 476 in math. (A perfect score is 800.) California kids scored 424 verbal, 482 math.

Honig said he was proud of the state's seniors for being average while 'going to school in the nation's second most crowded classrooms.'

Who is he kidding? Since when is it permissible much less a 'noteworthy accomplishment' for California to be merely average? High school seniors in such reputedly benighted states as Alabama, Arkansas, Georgia, Louisiana, Kentucky, Nevada, Nebraska, and Tennessee, among many others, did better on their SATs than our kids. That's a disgrace for a state that calls itself golden."

The columnist was referring to the following facts. Take Alabama for example; that year, the two averages for the Alabama high school seniors taking the SAT were 478 on the verbal part and 515 on the math part; the two averages for California were 424 and 482, respectively.

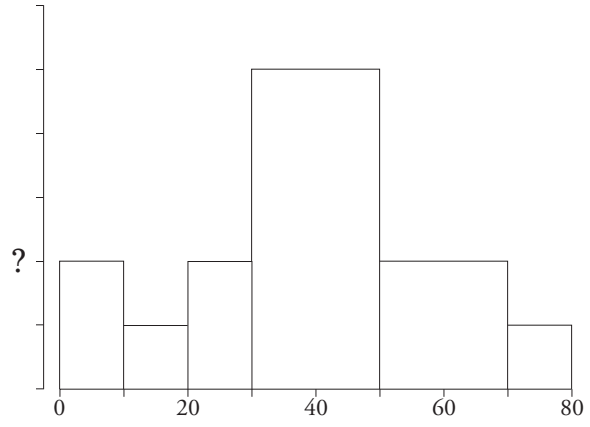
True or false, and explain: The math averages for Alabama and California show that on the whole, the schools in California are doing a poorer job than the schools in Alabama at teaching the mathematics required for the SAT."

8. (5 points) An instructor in a class of 300 students enters all test results in a computer file. A program calculates the following summary statistics:

midterm score:	average = 57 points,	SD = 18 points
final score:	average = 52 points,	SD = 25 points
	correlation = 0.60	

The program also runs through the 300 students and calculates for each, the regression prediction of final score from midterm score. For some of the students, the regression prediction was off by more than 20 points. About how many?

9. (10 points) The histogram below shows the distribution of family income in a small town. The data is hypothetical.



(a) If the density scale is used on the vertical axis, what number belongs at the question mark.

(b) What are the units for the answer to (a)?

(c) What percent of the families earned between $10,000 and $20,000?

(d) Is the average income under $40,000, over $40,000 equal to $40,000? Circle your choice and explain your reasoning. (Note: Assume that family income is spread evenly within each of the six class intervals used in the histogram: 0–10, 10–20, 20–30, 30–50, 50–70 and 70–80.)

10. (10 points) (a) What is the correlation coefficient for the data set below?

x	y
15	3
15	17
14	19
12	20
14	20
16	21
19	40

Note: To save you a little time, the sum of the x-column is 105, and the sum of the y-column is 140.

(b) What is the correlation coefficient between x and y in the data set below?

x	y
15	5
15	5
14	6
12	8
14	6
16	4
19	1

11. (10 points) An aerobics study involved 645 men. The average weight of the men was 166.5 pounds. The histogram of weight followed the normal curve closely. Out of the 645 men, there were 200 men who weighed under 150 pounds.

(a) How many weighed over 183 pounds?

(b) How many weighed under 140 pounds?

12. (5 points) For the 1,000 men in a medical study, the relationship between height and weight can be summarized as follows:

height: average = 70 inches, SD = 3 inches
weight: average = 160 pounds SD = 30 pounds
correlation coefficient = 0.60

The data is plotted on a scatter diagram, with height (x) on the horizontal axis and

weight (y) on the vertical axis. The 1,000 points form a football shaped cloud. Two lines

are also plotted on the diagram. Their equations are:

$$y = 6x - 236$$
$$y = 6x - 284$$

The units on the 6 are pounds per inch; the units on the 236 and 284 are pounds; and the units for x are inches.

About what percent of the 1,000 points fall between the two lines? Explain your reasoning.

13. (5 points) Two meteorologists are comparing noon temperature in Boston and Washington, D.C. using data from 2005. The first meteorologist computes the correlation between daily noon temperatures in Boston and Washington. The second meteorologist computes the correlation between the monthly average noon temperatures in those two cities.

    Check (√) one option below and explain your choice.

    _____ The first meteorologist gets a smaller correlation.

    _____ The two correlations are equal.

    _____ The first meteorologist gets a bigger correlation.

14. (10 points) For the first-year students at a certain university, the correlation coefficient between SAT scores and first year GPA was 0.60.

    (a) Predict the percentile rank on the first-year GPA for a student who ranked at the 90th percentile on the SAT.

    (b) There are two blanks in the paragraph below. Please fill in both blanks with the answer you got in part (a). Then answer the question.

    The regression method predicts that a student in the 90th percentile on the SAT would be in the _____th percentile on first-year GPA. True or False, and explain: "A student in the _____th percentile on first-year GPA should be in the 90th percentile on the SAT."

15. (10 points) The distribution table below is taken from a Public Health Service study. The table gives the distribution of subjects in the study by the number of cigarettes smoked daily. The class intervals include the right endpoint, but not the left. For example, the second line of the table says that 35% of the subjects in the study smoked more than 10, but not more than 20 cigarettes per day.

Number Per Day	Percent of Subjects
0 – 10	15
10 – 20	35
20 – 40	30
40 – 80	20

    (a) Plot the histogram. Mark the horizontal and vertical scales carefully. Label the axes.

    (b) Is the average around 15, 20 or 25? Circle your choice and explain your reasoning.

16. (5 points) A New York city resident is wondering whether to get a car alarm. A friend advises him not to do it. "Your car will be even more likely to be stolen. The New York Times had an article about it. Listen to this." The friend reads the following from The New York Times (of February 19, 1991):

> "One trend that does seem clear in the data is that cars equipped with anti-thefts devices, which are most commonly alarms, are more likely than other cars to be stolen or broken into.
>
> In New York, cars with these devices have a rate of claims that's three times higher than that of other cars,' said Dale Nelson, an actuary with State Farm, the nation's largest car insurer."

Is the friend's claim justified by the evidence cited by the actuary? Answer yes or no and explain your answer.

17. (5 points) The men enrolled in a large sports medicine course had an average weight of 160 pounds and an SD of 30 pounds. Their weights followed the normal curve closely.

Consider the men in the course who weighed somewhere between 180 and 200 pounds. The average weight of these men would be:

_____ equal to 190 pounds.
_____ bigger than 190 pounds.
_____ smaller than 190 pounds.
_____ can't tell without more information.

Check (√) one of the above options. Please explain your choice.

18. (5 points) There are 1600 first-year students at a certain university. Their scores on the verbal SAT followed the normal curve closely, and the average score was 550 points. Around 360 students had scores in the range from 550 to 600 points. How many had scores in the range from 600 to 650 points?

19. (10 points) For women age 25-34 in the HANES5 sample, the relationship between height and income can be summarized as follows:

height:	average ≈ 64 inches,	SD ≈ 2.5 inches
income:	average ≈ $21,000,	SD ≈ $20,200
	correlation ≈ 0.2	

(a) Find the regression equation for predicting income from height.

(b) Use the equation to predict the income of a woman 66 inches tall.

(c) Someone decides to predict income by using the equation:

predicted income = $21,000

What is the r.m.s. error for this method?

20. (10 points) (a) What is the correlation coefficient for the data set below.

x	y
5	7
17	9
19	11
23	13
36	10

(b) One number is missing in the data set below:

x	y
8	1
8	3
8	5
9	2
9	4
10	

Here are three numbers: 1 3 5.

One of them, when placed in the blank, will make the correlation coefficient between x and y close to 0.4. Which one? Please circle your choice and give your reasoning and/or calculation.

21. (10 points) An investigation in exercise physiology involves 1100 men students at a large university. For these men, the correlation between height and weight was 0.40, and the scatter diagram for the two variables was football shaped. The average weight was 158 pounds and the SD, 30 pounds.

One man in the study happened to be at the 40th percentile amongst the 1100 men for both height and weight. What percent of the men of his height weighed more than him?

22. (5 points) The summary statistics below come from a study of the employees at a large engineering firm.

education: average = 13.5 years SD = 3.1 years
income: average = $52,900 SD = $8,580

Someone uses the following equation to predict income from years of education for these employees:

predicted income = ($2860 per year) x (education) + $14,290

Please check (√) one option below and then explain your choice.

_____ The equation is the regression equation for predicting income from education for these employees.

_____ The equation is not the regression equation.

_____ There is not enough information to decide.

23. (10 points) For the men in a large medical study the average systolic blood pressure was 126 mm and the SD was 16 mm. The histogram followed the normal curve closely.

(a) Estimate the percentage of men in the study with blood pressures between 110 mm and 114 mm.

(b) Take the men in the study with blood pressures over 130 mm. The median of their blood pressures is _____.

24. (5 points) (a) Fill in the four blanks below:

x	y	x in standard units	y in standard units
1	5	____	–1.5
___	17	____	–0.3
4	19	0.0	–0.1
5	23	0.5	+0.3
7	36	____	+1.6

Remember to show work and/or give reasons.

(b) Calculate the correlation coefficient. (If you cannot fill in enough blanks in part (a) to do the calculation, just describe the arithmetic you would do if all the blanks were filled in.)

25. (5 points) A large class has two midterms. On the first midterm, the average score was 52 points out of a total of 100; on the second, it was 62 points. So the average increased by 10 points. The SDs for the two tests were about the same, and the scatter diagram was football shaped.

One person in the course scored 69 on the first midterm and 79 on the second one. True or false, and explain. Compared to everyone else who got 69 on the first midterm, this person was about average on the second midterm

26. (10 points) The table below shows the distribution of age in a California town of 25,000 residents:

Age range	Percent in range
0–4	8%
5–18	21%
19–29	22%
30–54	40%
55–84	9%

(a) Use the information in the table to draw a histogram for the age distribution in the town. Please mark the horizontal and vertical axes carefully. Label the axes. (Remember to show all calculations.)

(b) True or false, and explain: In this town, the 20th percentile of age is higher than 20 years.

27. (5 points) An investigator has the heights and weights of a sample of 1000 fit men in their early twenties. The summary statistics are:

Height: Average = 70.0 inches, SD = 3.0 inches
Weight: Average = 162 pounds, SD = 30 pounds

The scatter diagram is football shaped.

The investigator plots the graph of average weight against height—the graph of averages. (The graph shows height on the horizontal axis and average weight on the vertical axis.) The investigator also plots the regression line on the graph:

$$y = 4.7\,x - 167$$

(The units on the 4.7 are pounds per inch and units on the 167 are pounds.)

(a) Find the r.m.s. error of regression line. Copy your answer into the blank in part (b) below.

(b) True or false, and explain: A typical point on the graph of averages will be off the regression line by vertical distance of around _____ pounds.

28. (10 points) A study is made of the Math and Verbal SAT scores for the entering class at a certain college. The summary statistics are as follows:

M-SAT: Average = 560, SD = 120
V-SAT: Average = 520, SD = 110

correlation coefficient = 0.64

The scatter diagram is football-shaped.

(a) Some one who scored 500 on the M-SAT would have a percentile rank (within the entering class) on the M-SAT of _____.

(b) Of all those who scored 500 on the M-SAT, about _____ percent had a higher percentile rank (within the entering class) on the V-SAT than on the M-SAT.
