Regression with two variables

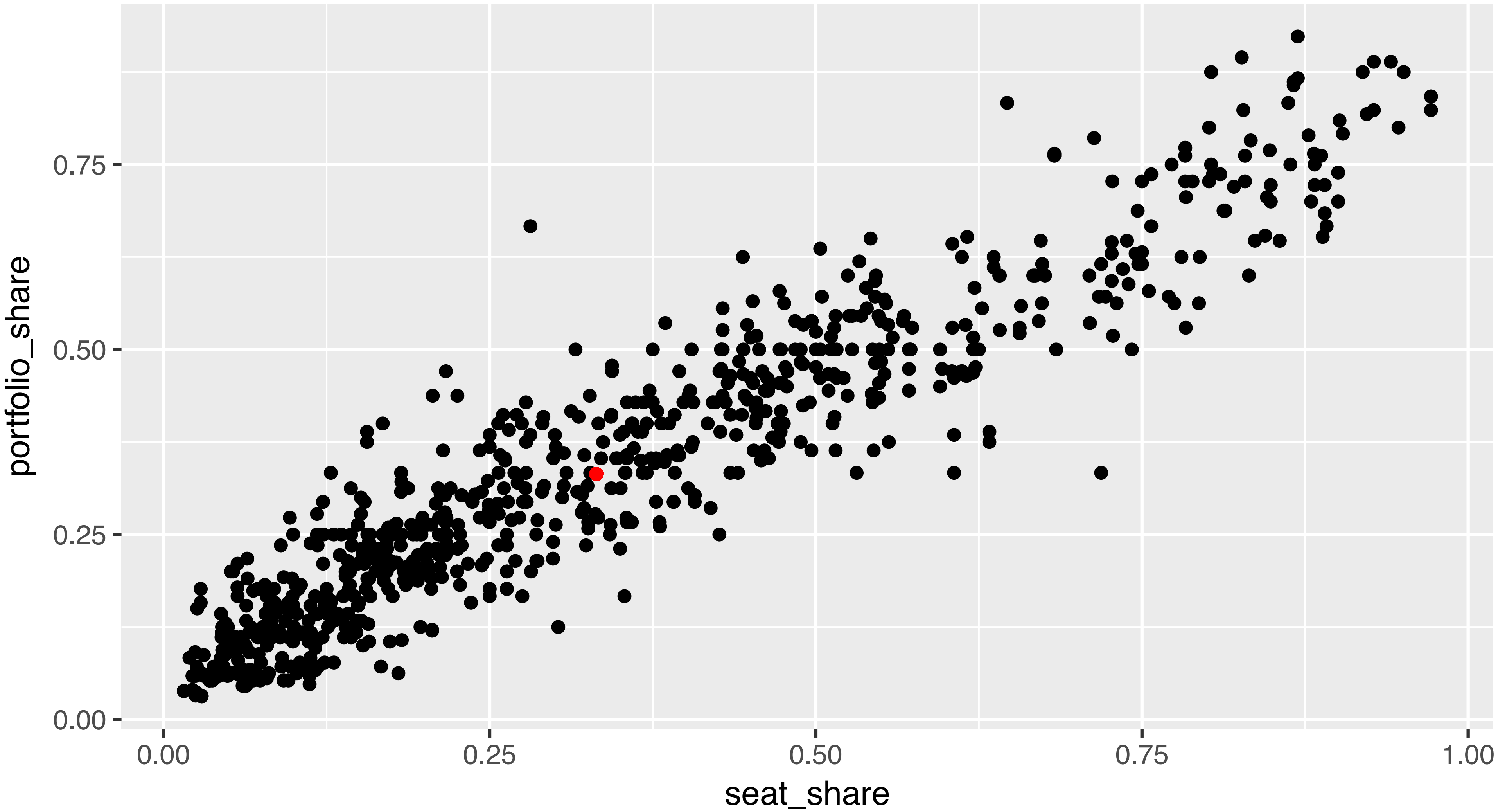
how to describe the relationship between two variables

From Wikipedia:

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables.

- We start with regression one dependent variable (Y) on one independent variable (X) - but will expand to more later

Regression means drawing a line through our scatter plot

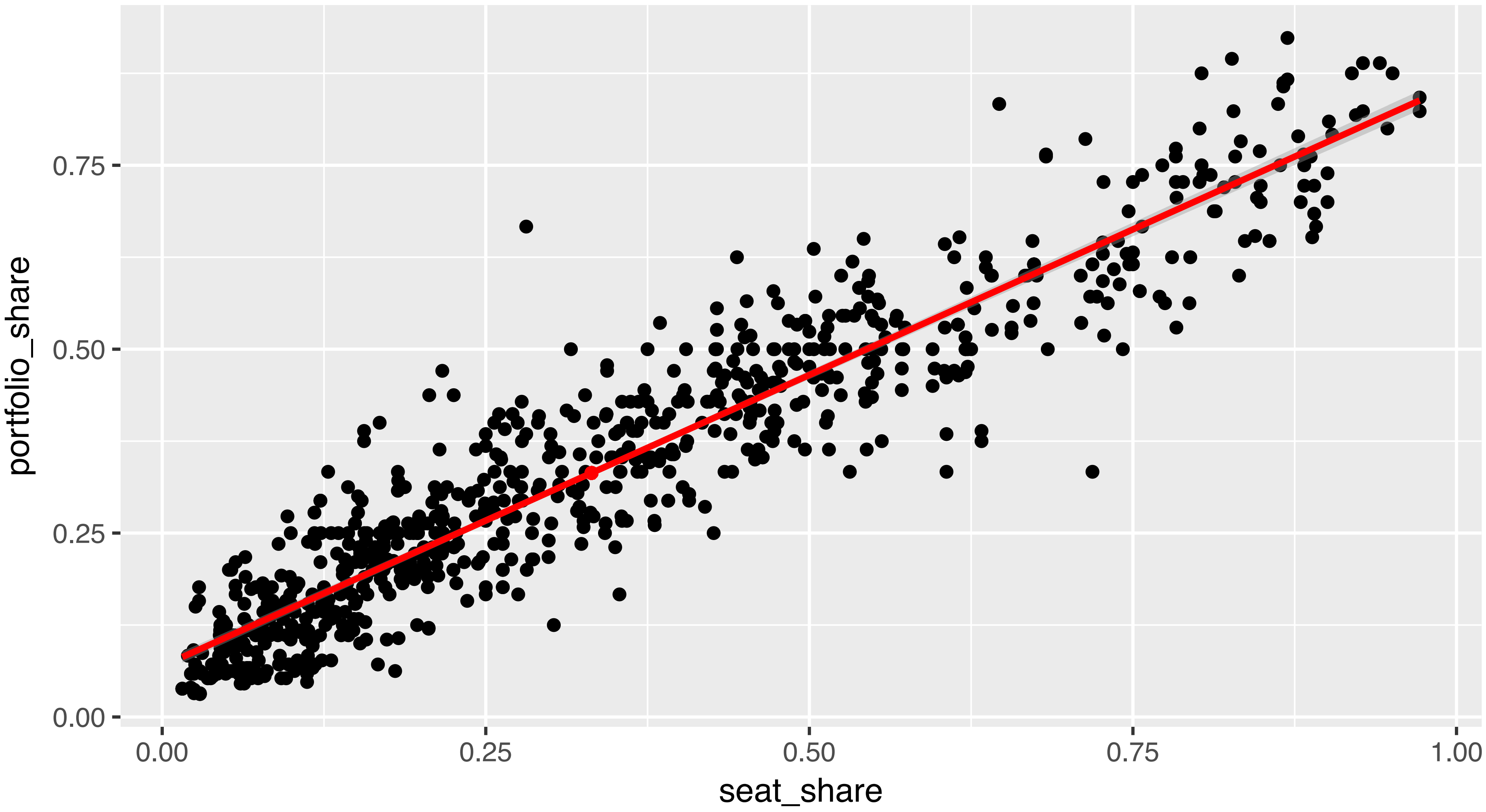


With regression the goal is to cleanly approximate the relationship between X and Y

We do so by drawing a line through these points that minimizes the distance to each point

Regression line

Slope of SD line:
0.21/0.25 = 0.84

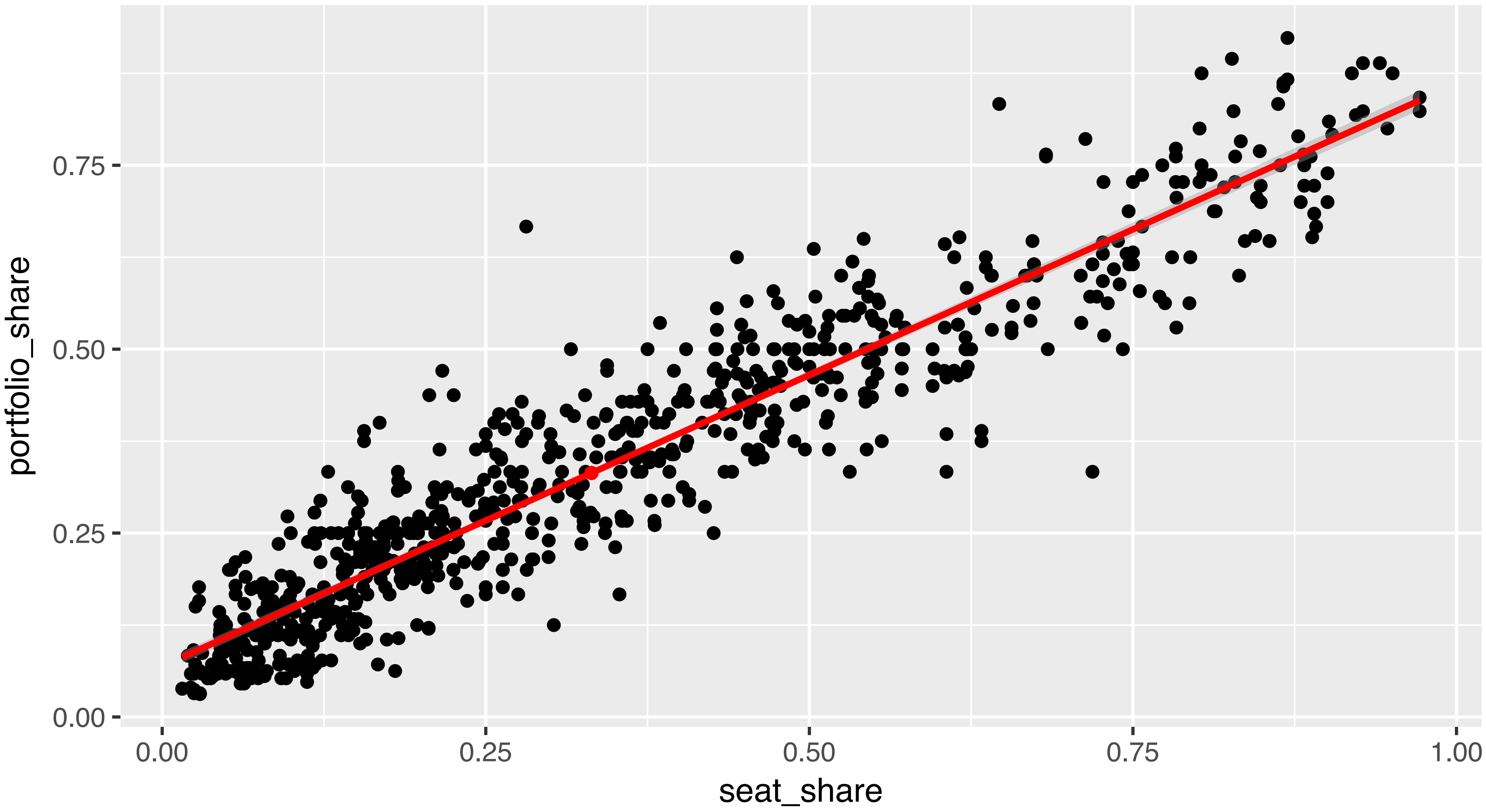


As with the sd line,
the regression line goes through the point of averages

But, the slope is not the same!

Regression line

Slope of SD line:
0.21/0.25 = 0.84



**As with the sd line,
the regression line goes through the point of averages**

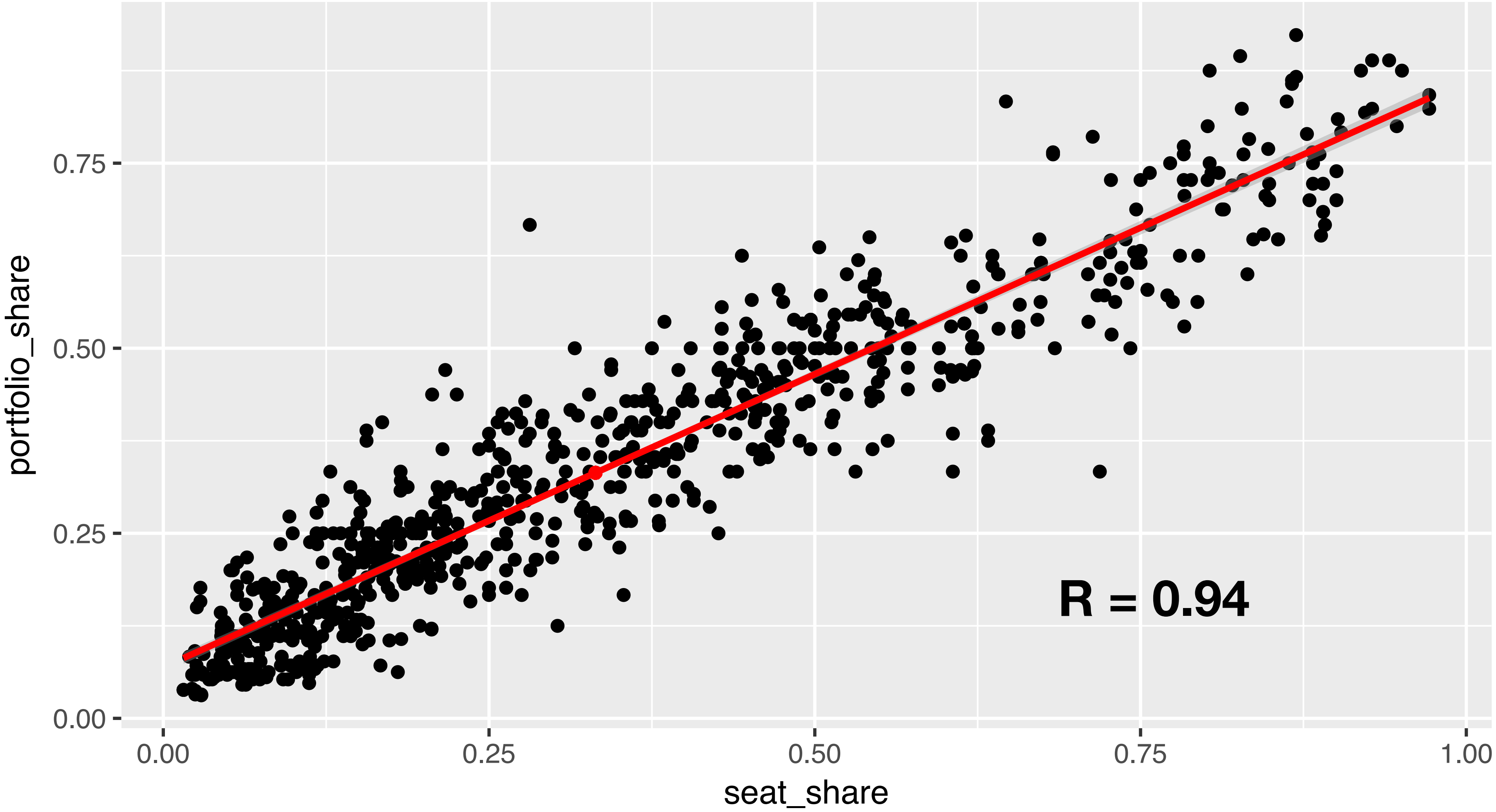
**But, the slope is not the same!
Here it is pretty close though, the slope is 0.79**

Regression line

- The regression line is an estimate of the (for now bivariate) relationship between x and y
- For each x we have a prediction of y, or what would we expect y to be on average, given the value of x
- The line goes through the point of averages
- The slope of the regression line is the slope of the SD line multiplied by the correlation (r)

Regression line

Slope of SD line:
0.21/0.25 = 0.84



As with the sd line,
the regression line goes through the point of averages

Slope of regression line = 0.84 * 0.94 = 0.79

Regression line

- So for every change in x by a standard deviation, there is The slope of the regression line is the slope of the SD line multiplied by the correlation r

- If r = 1, then all points are exactly on the SD line, i.e. slope is the same

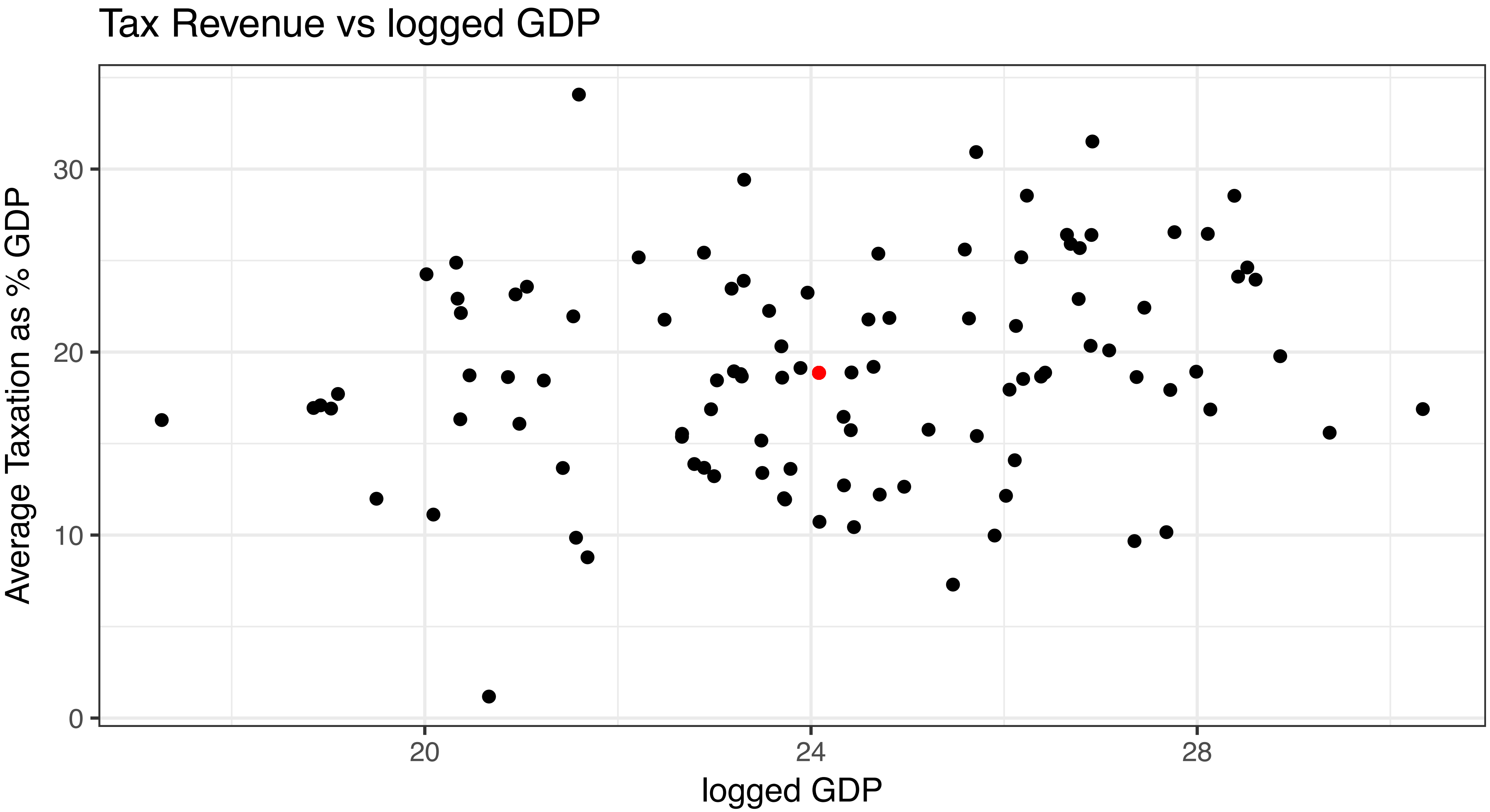
- If r = 0, then the slope is flat and no relationship exists

Regression

In a class, midterm scores average out to 60 with and SD of 15. The scores on the final have the same property. The correlation between midterm and final scores are 0.5. What are the average final scores for students with the following midterm scores:

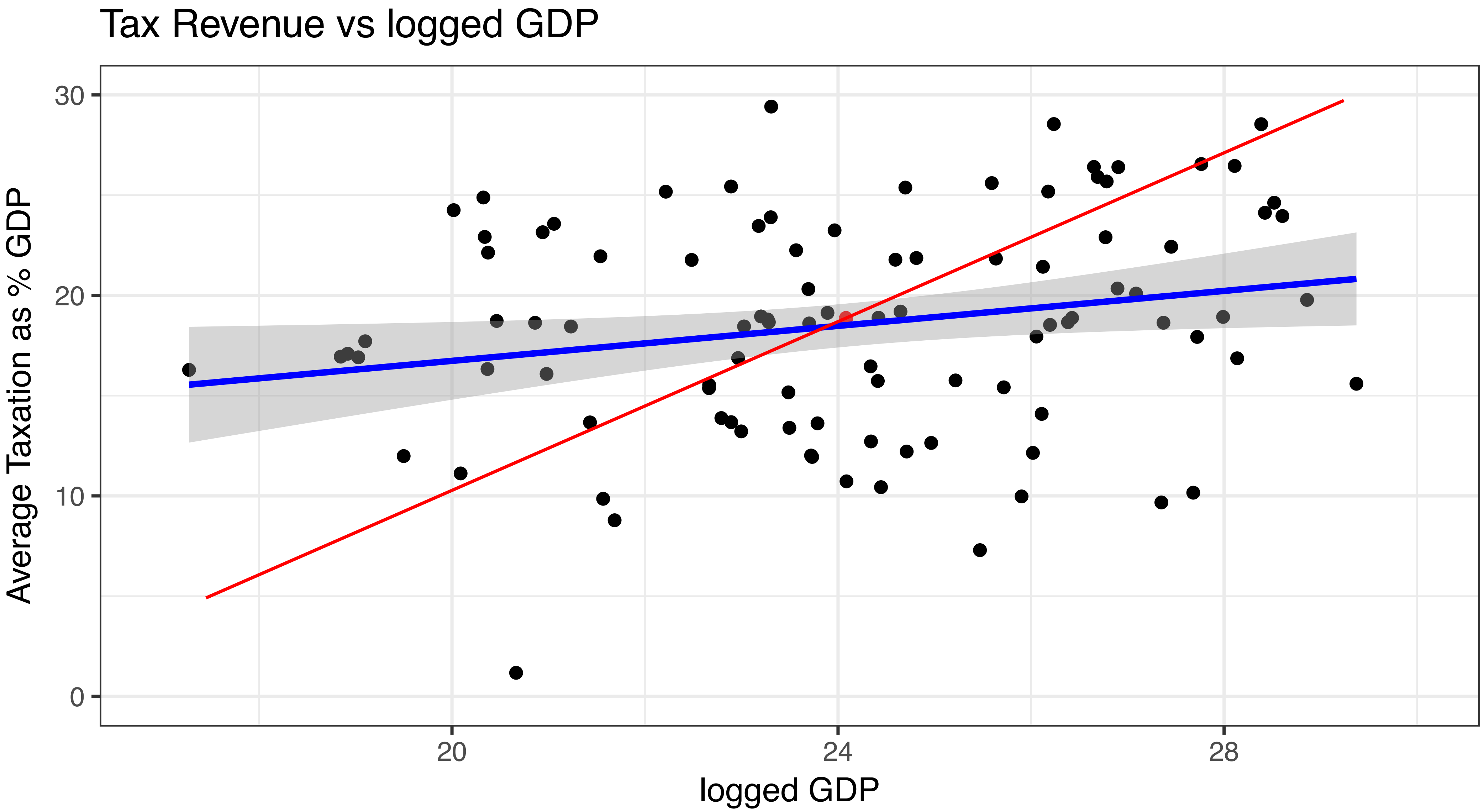
A) 75 B) 30 C) 60

Recall difference between SD line and Regression line



SD_y = 5.84
SD_x = 2.78
r = 0.2

Recall difference between SD line and Regression line



Predicting new Y

- We might use averages of smaller groups to predict new values of Y

- Example: what is the value of taxation given a certain level of GDP?

- We can group countries around our x-value of interest and take the average of those countries to predict the new value

- Turns out the point of regression line at the new value approximates it pretty well

Predicting Y — graph of averages

- The graph of average is a scatter plot of averages for small intervals of x
- Some of these averages will be too high, some too low
- The regression line smoothes out the graph of averages, some points will lie below, some above the line

Predicting new Y

- Relationship between Math SAT score and first-year GPA:
    - Average SAT 550, SD = 80
    - Average GPA 2.6, SD = 0.6, r = 0.4

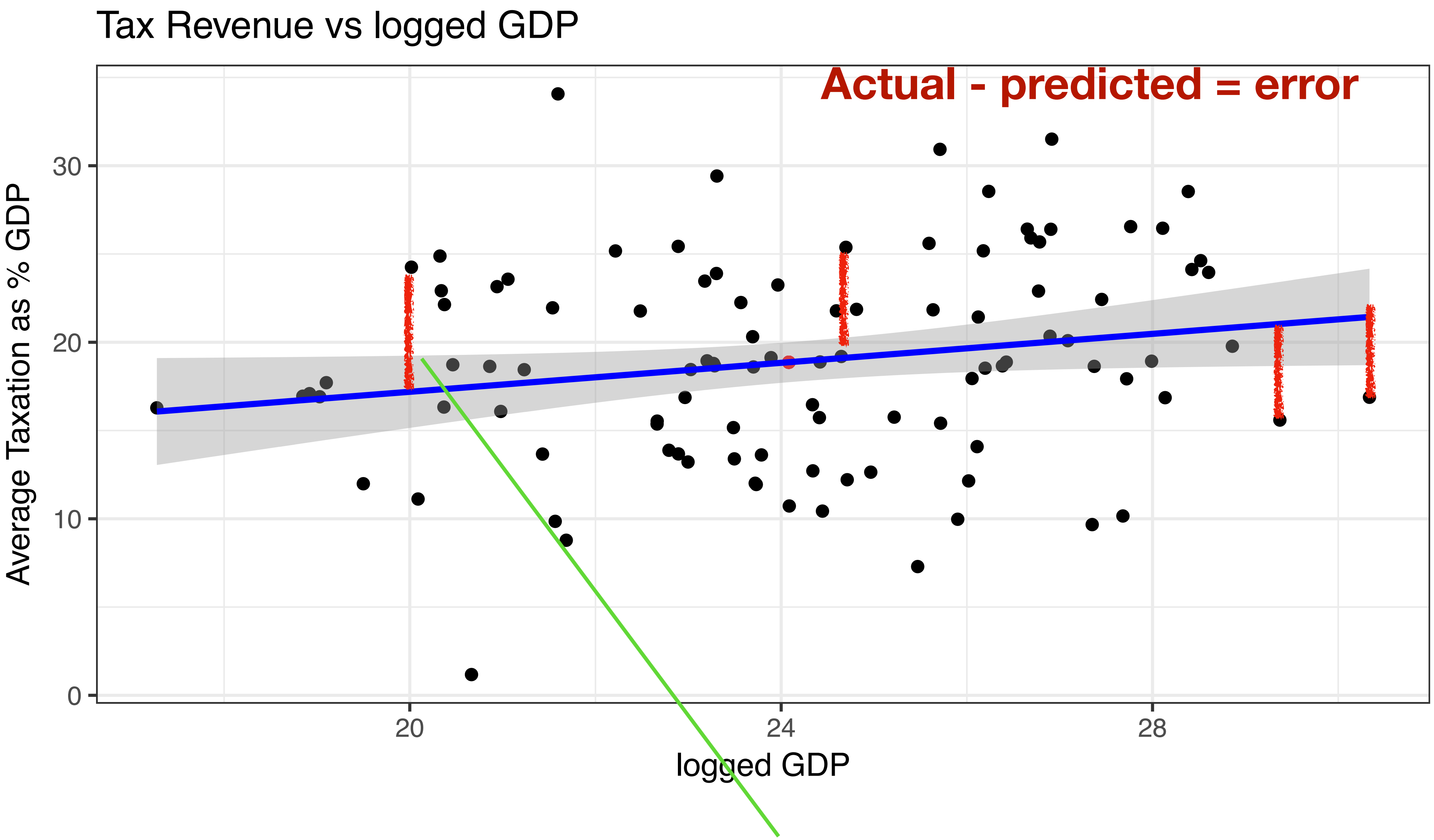
For a student with a SAT of 650, what would be the predicted GPA?

Predicting new Y

- We can do the same thing with respect to percentile ranks!
  - Average SAT 550, SD = 80
  - Average GPA 2.6, SD = 0.6, r = 0.4
  - Student in 90% percentile rank of SAT, what is the predicted rank on the GDP?

- In a study of father & son height, the sons of 72-inch fathers only averaged 71 inches in height. True of false, and explain: If you take the 71-inch sons, their fathers will average 72 inches in height.

But regression lines are not perfect



Actual - predicted = error

We always measure the error in terms of prediction error in y!
Why?

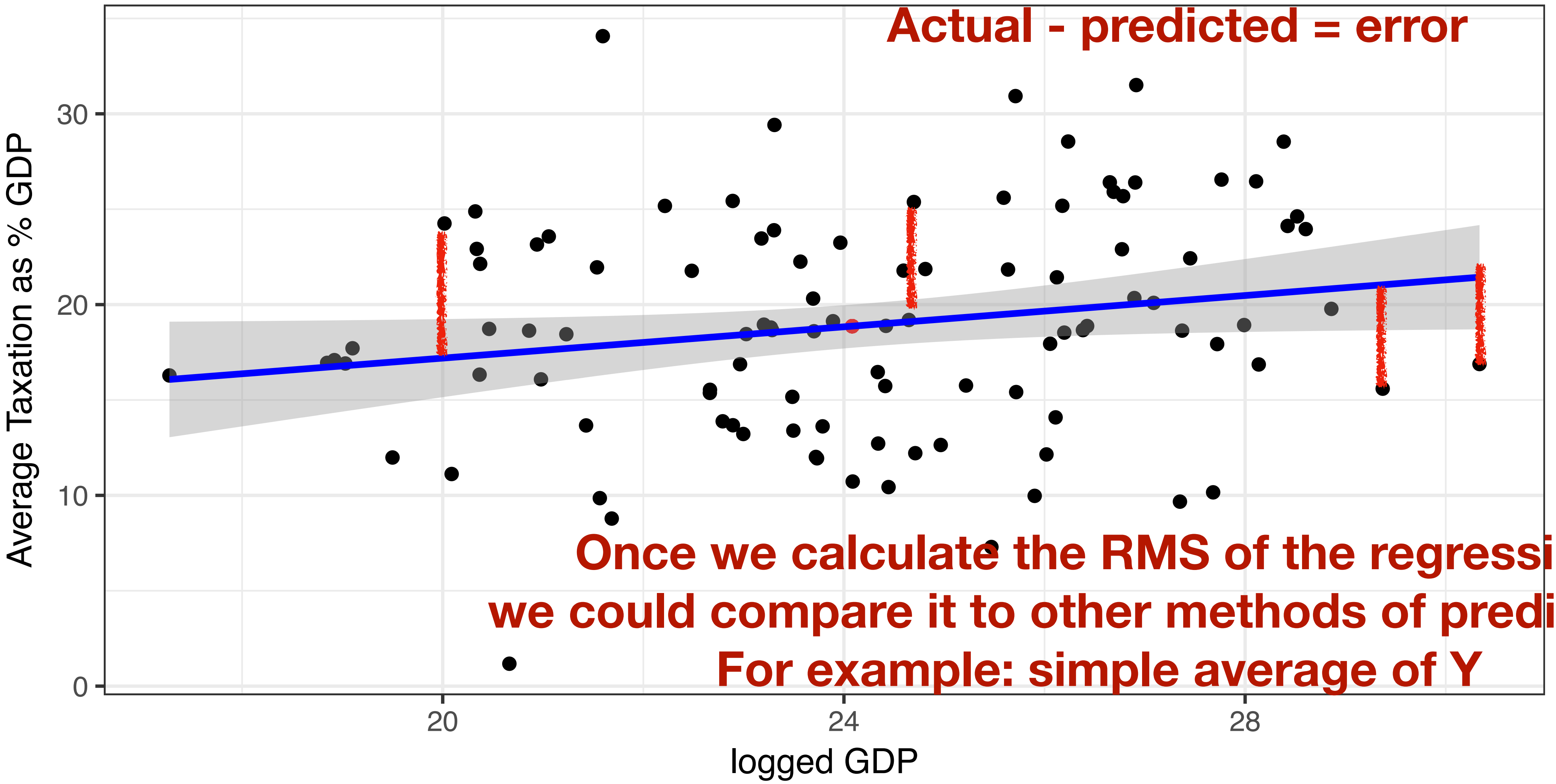
Recall the root mean squared error

- RMS = square root of the mean of the squared errors
- Approximately equal to the average of how far points are above and below the line
- RMS is always in the unit of the dependent variable (the variable to be predicted - y)
- Why can’t we just take the average of the errors?

But regression lines are not perfect

RMS = sqrt(mean ((actual-predicted)^2))

Tax Revenue vs logged GDP



Actual - predicted = error

Once we calculate the RMS of the regression, we could compare it to other methods of predicting Y. For example: simple average of Y

Recall the root mean squared error

- What is the root mean squared error of using the average of y to predict y?

Recall the root mean squared error

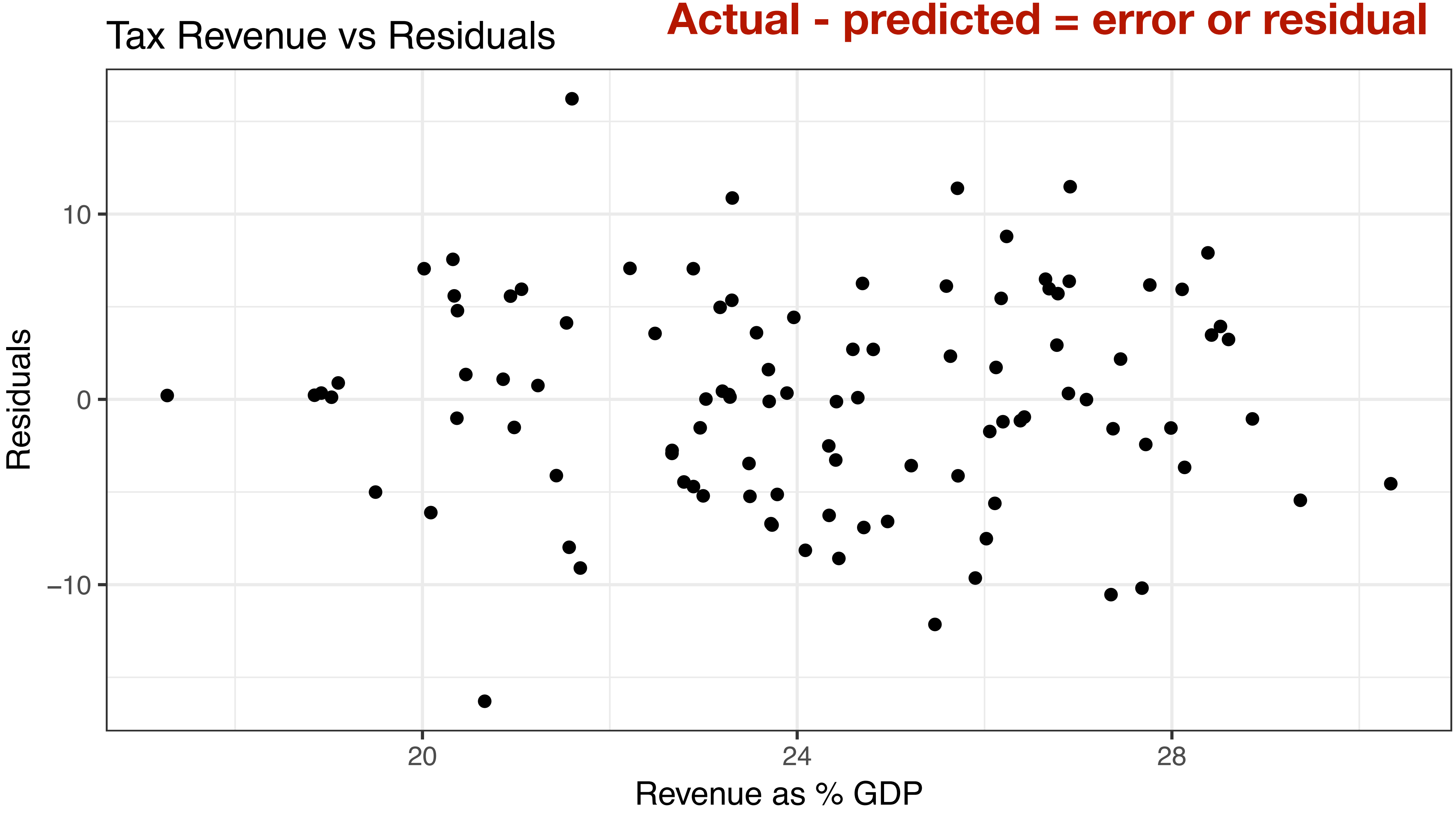
- What is the root mean squared error of using the average of y to predict y?
- The standard deviation!

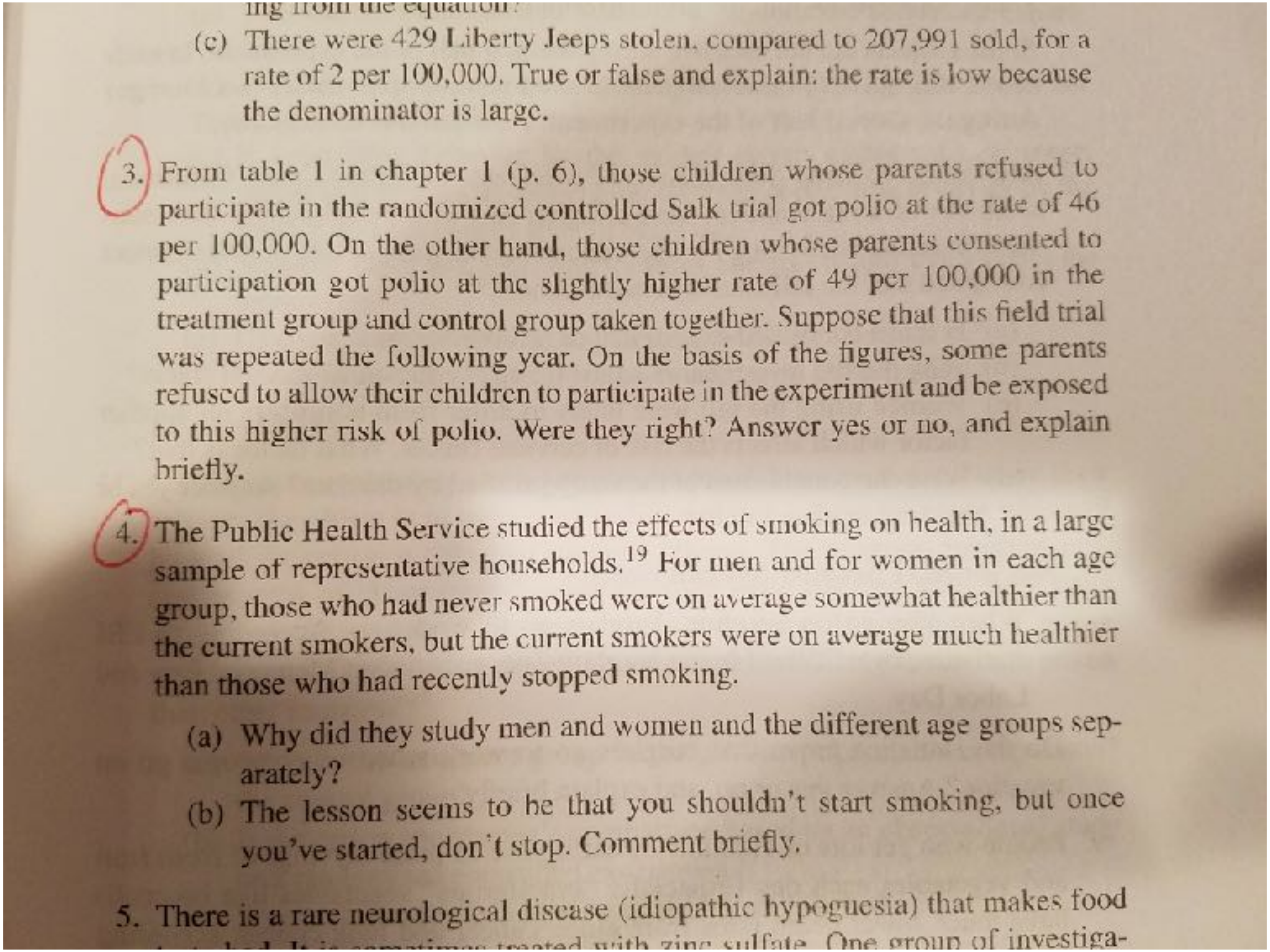
Computing the rms for the regression

- In theory, we could calculate the rms by doing the calculation for every point in our data
- Luckily, we have a formula that makes calculation much simpler: rms_regression = SD_y * sqrt(1 - r^2)
- The rms is

We can plot the error or residuals

Actual - predicted = error or residual





ing from the equation?

(c) There were 429 Liberty Jeeps stolen, compared to 207,991 sold, for a rate of 2 per 100,000. True or false and explain: the rate is low because the denominator is large.

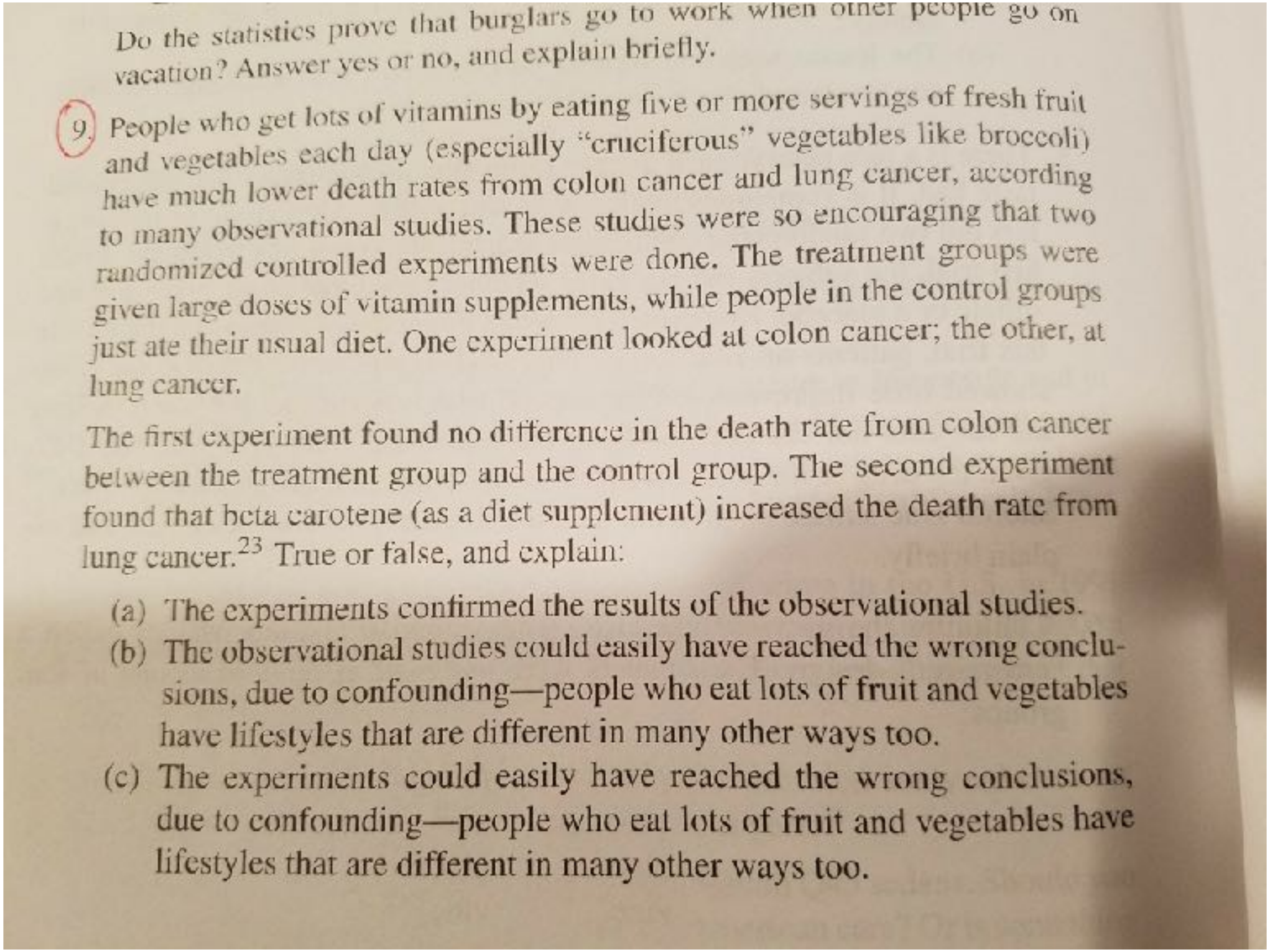
3. From table 1 in chapter 1 (p. 6), those children whose parents refused to participate in the randomized controlled Salk trial got polio at the rate of 46 per 100,000. On the other hand, those children whose parents consented to participation got polio at the slightly higher rate of 49 per 100,000 in the treatment group and control group taken together. Suppose that this field trial was repeated the following year. On the basis of the figures, some parents refused to allow their children to participate in the experiment and be exposed to this higher risk of polio. Were they right? Answer yes or no, and explain briefly.

4. The Public Health Service studied the effects of smoking on health, in a large sample of representative households.[19] For men and for women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than those who had recently stopped smoking.

(a) Why did they study men and women and the different age groups separately?

(b) The lesson seems to be that you shouldn't start smoking, but once you've started, don't stop. Comment briefly.

5. There is a rare neurological disease (idiopathic hypoguesia) that makes food taste bad. It is sometimes treated with zinc sulfate. One group of investiga-



Do the statistics prove that burglars go to work when other people go on vacation? Answer yes or no, and explain briefly.

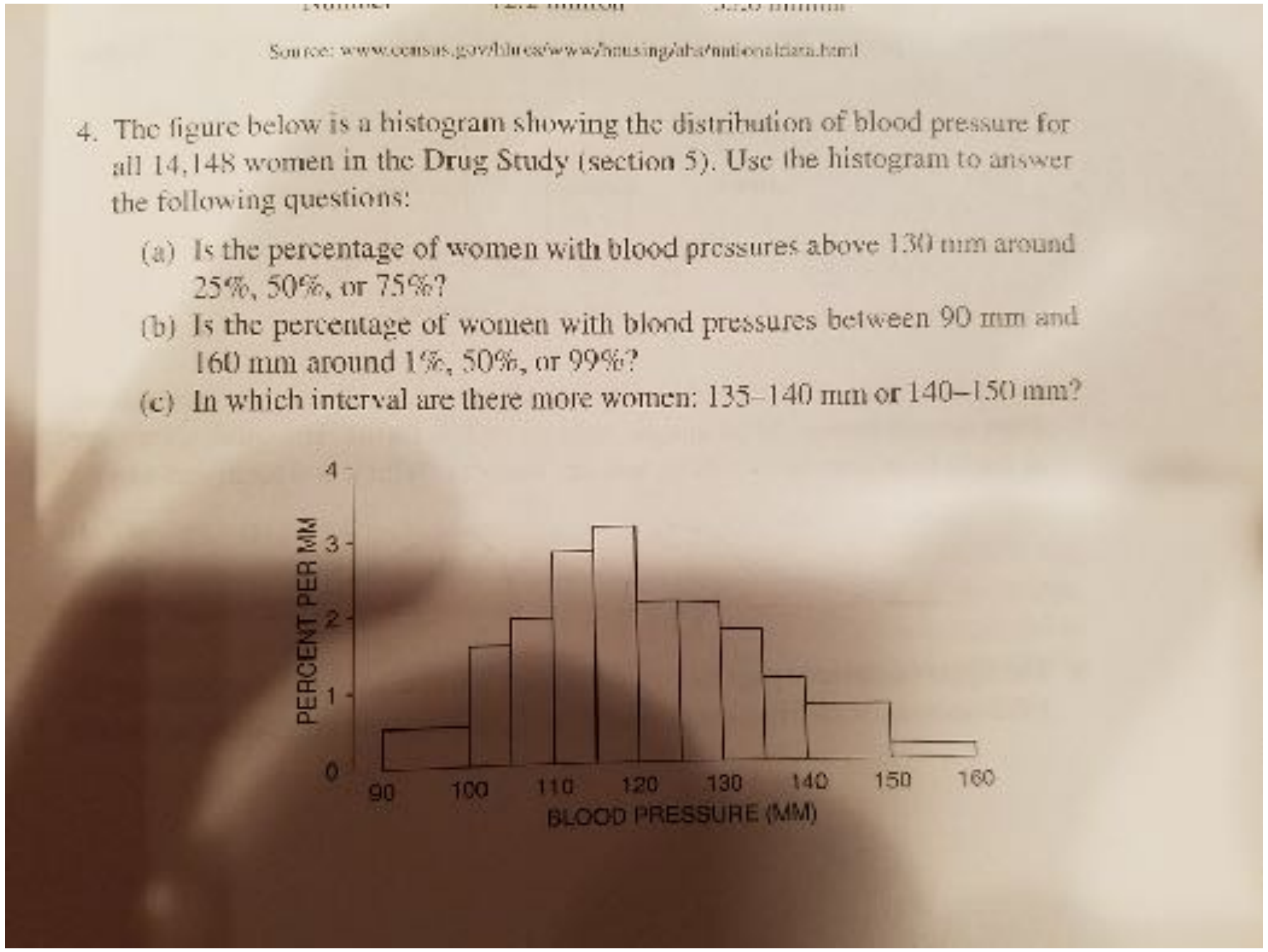
9. People who get lots of vitamins by eating five or more servings of fresh fruit and vegetables each day (especially "cruciferous" vegetables like broccoli) have much lower death rates from colon cancer and lung cancer, according to many observational studies. These studies were so encouraging that two randomized controlled experiments were done. The treatment groups were given large doses of vitamin supplements, while people in the control groups just ate their usual diet. One experiment looked at colon cancer; the other, at lung cancer.

The first experiment found no difference in the death rate from colon cancer between the treatment group and the control group. The second experiment found that beta carotene (as a diet supplement) increased the death rate from lung cancer.[23] True or false, and explain:

(a) The experiments confirmed the results of the observational studies.

(b) The observational studies could easily have reached the wrong conclusions, due to confounding—people who eat lots of fruit and vegetables have lifestyles that are different in many other ways too.

(c) The experiments could easily have reached the wrong conclusions, due to confounding—people who eat lots of fruit and vegetables have lifestyles that are different in many other ways too.

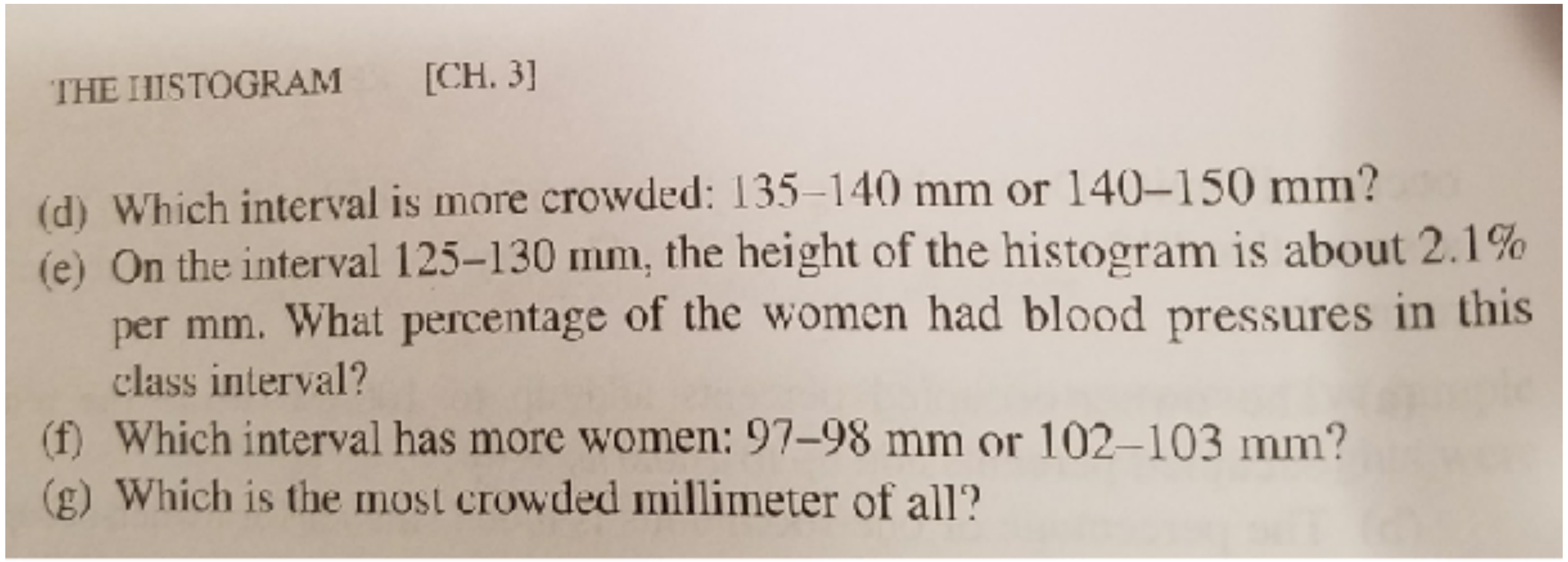


Source: www.census.gov/hhes/www/housing/ahs/nationaldata.html

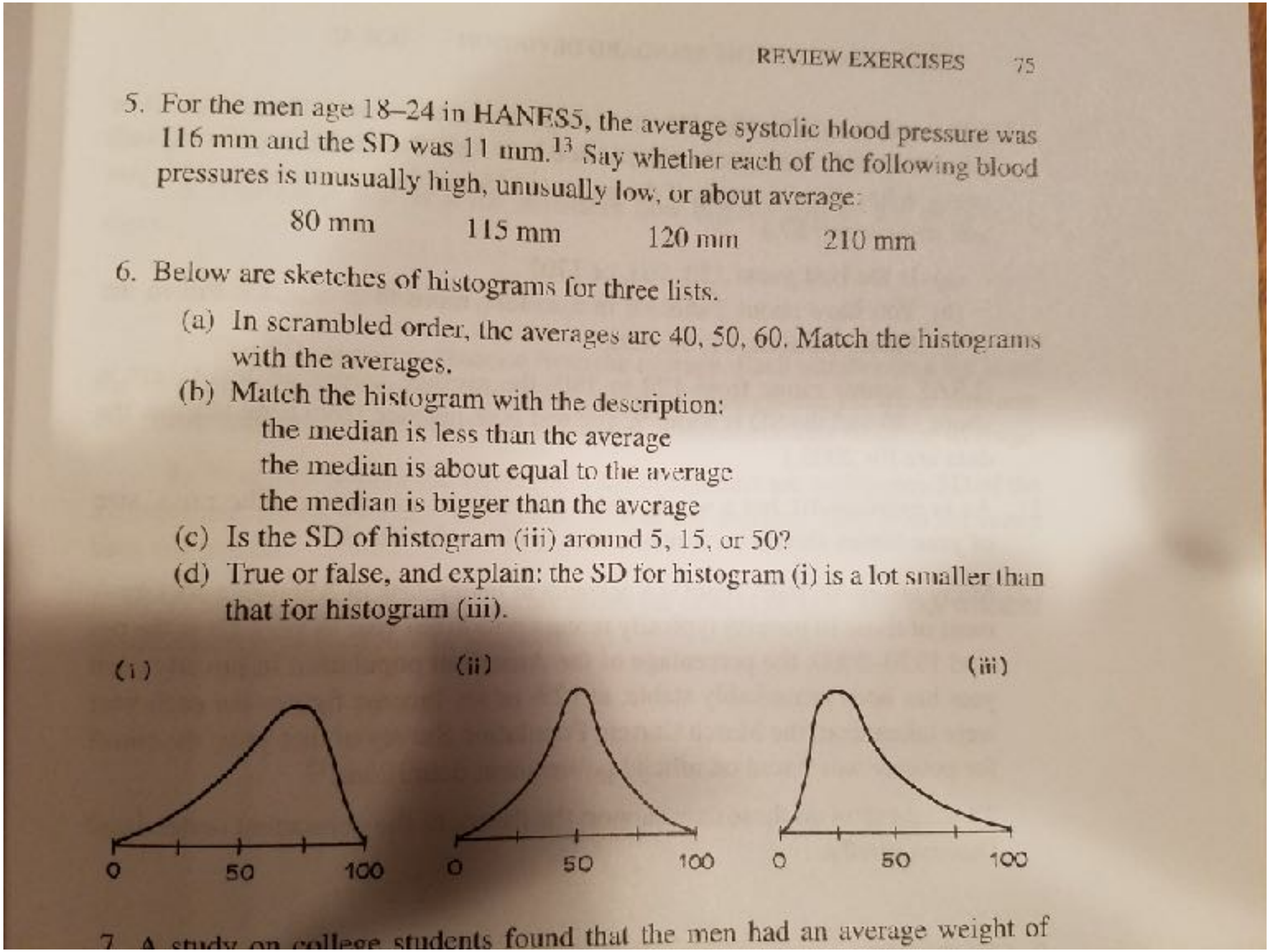
4. The figure below is a histogram showing the distribution of blood pressure for all 14,148 women in the Drug Study (section 5). Use the histogram to answer the following questions:

   (a) Is the percentage of women with blood pressures above 130 mm around 25%, 50%, or 75%?
   (b) Is the percentage of women with blood pressures between 90 mm and 160 mm around 1%, 50%, or 99%?
   (c) In which interval are there more women: 135–140 mm or 140–150 mm?

PERCENT PER MM
0
1
2
3
4
90
100
110
120
130
140
150
160
BLOOD PRESSURE (MM)



   (d) Which interval is more crowded: 135–140 mm or 140–150 mm?
   (e) On the interval 125–130 mm, the height of the histogram is about 2.1% per mm. What percentage of the women had blood pressures in this class interval?
   (f) Which interval has more women: 97–98 mm or 102–103 mm?
   (g) Which is the most crowded millimeter of all?

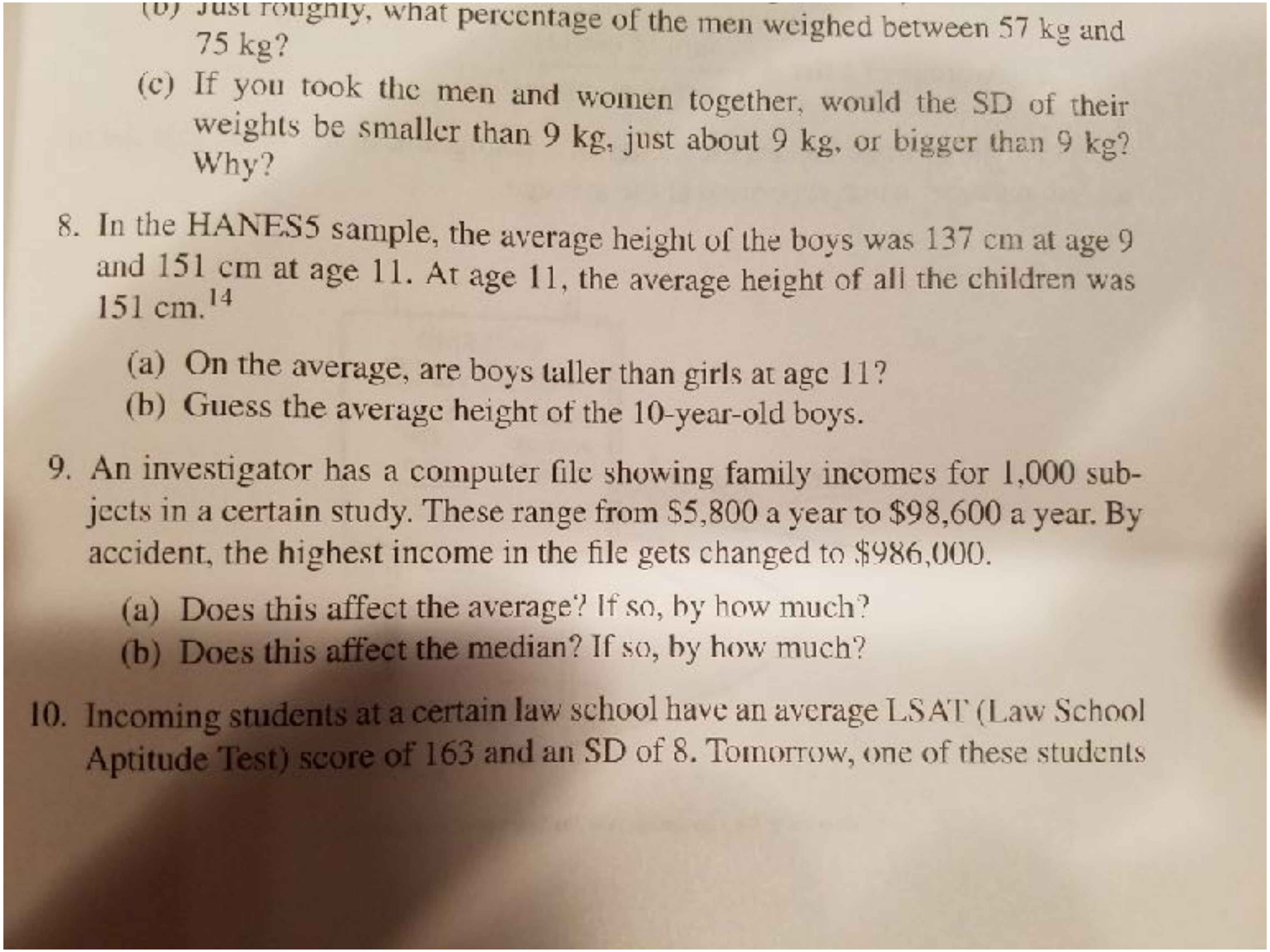


5. For the men age 18–24 in HANES5, the average systolic blood pressure was 116 mm and the SD was 11 mm.[13] Say whether each of the following blood pressures is unusually high, unusually low, or about average:

80 mm   115 mm   120 mm   210 mm

6. Below are sketches of histograms for three lists.
   (a) In scrambled order, the averages are 40, 50, 60. Match the histograms with the averages.
   (b) Match the histogram with the description:
       the median is less than the average
       the median is about equal to the average
       the median is bigger than the average
   (c) Is the SD of histogram (iii) around 5, 15, or 50?
   (d) True or false, and explain: the SD for histogram (i) is a lot smaller than that for histogram (iii).

7. A study on college students found that the men had an average weight of



(b) Just roughly, what percentage of the men weighed between 57 kg and 75 kg?

(c) If you took the men and women together, would the SD of their weights be smaller than 9 kg, just about 9 kg, or bigger than 9 kg? Why?

8. In the HANES5 sample, the average height of the boys was 137 cm at age 9 and 151 cm at age 11. At age 11, the average height of all the children was 151 cm.[14]

   (a) On the average, are boys taller than girls at age 11?

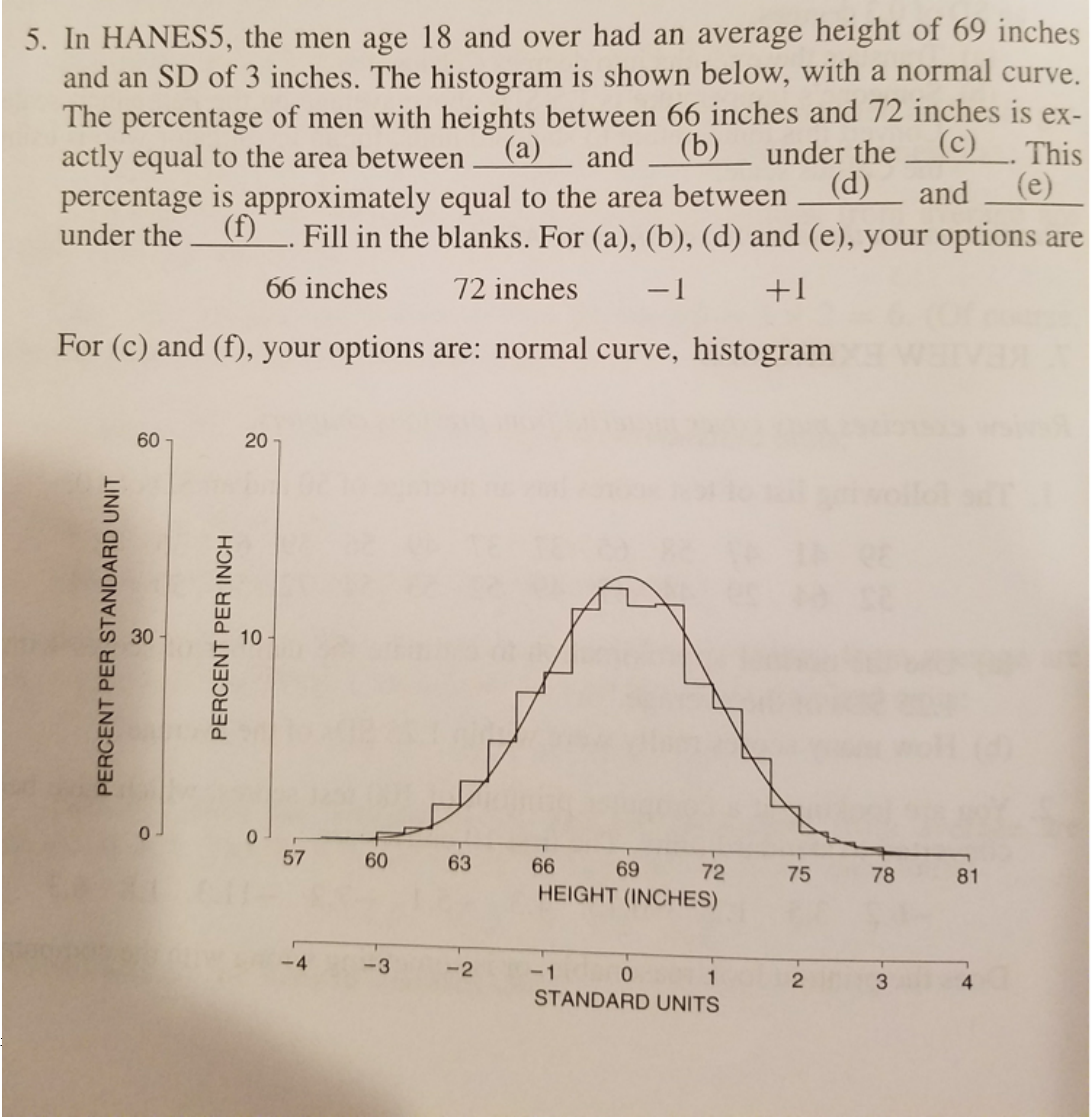
   (b) Guess the average height of the 10-year-old boys.

9. An investigator has a computer file showing family incomes for 1,000 subjects in a certain study. These range from $5,800 a year to $98,600 a year. By accident, the highest income in the file gets changed to $986,000.

   (a) Does this affect the average? If so, by how much?

   (b) Does this affect the median? If so, by how much?

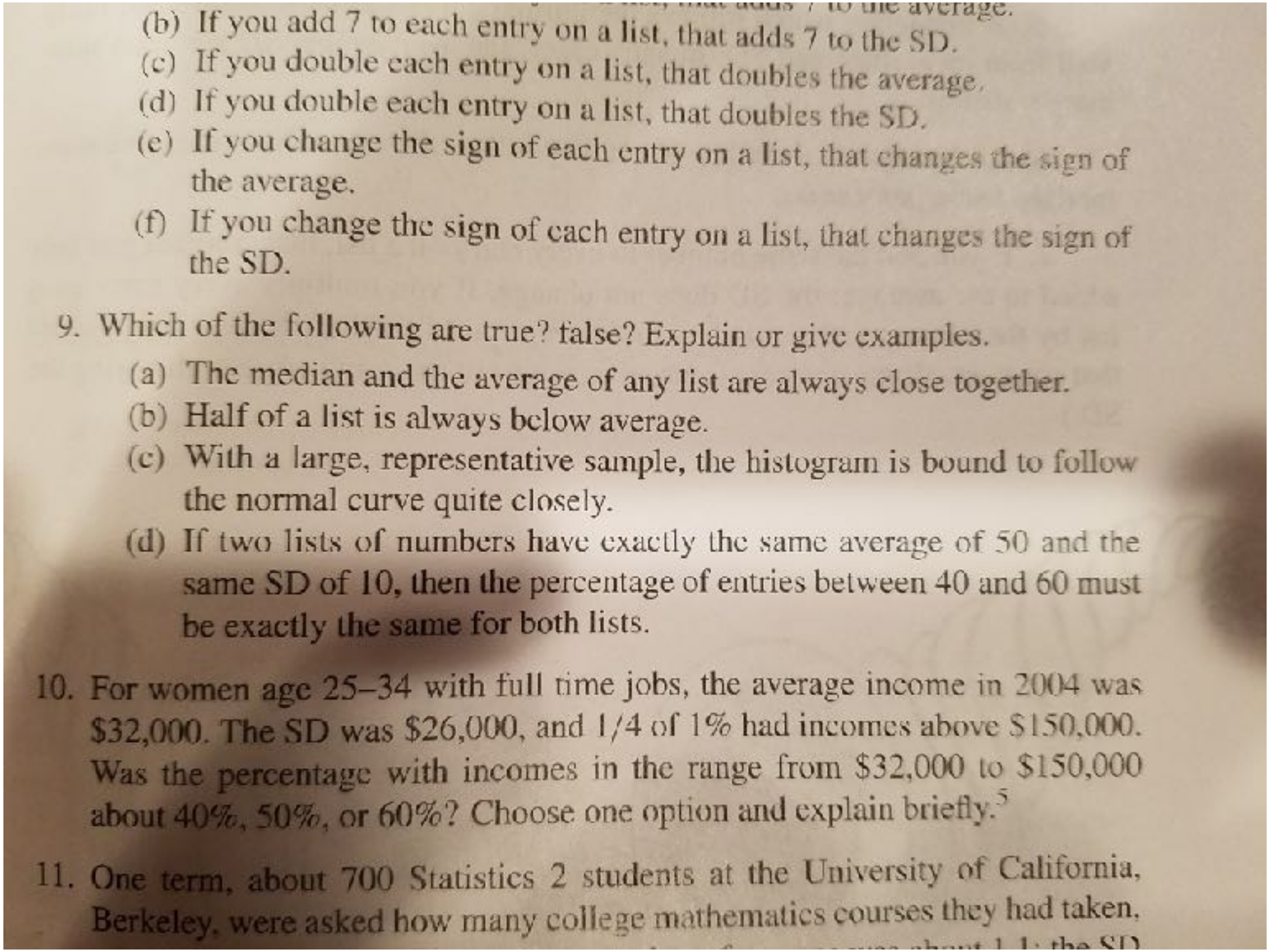
10. Incoming students at a certain law school have an average LSAT (Law School Aptitude Test) score of 163 and an SD of 8. Tomorrow, one of these students



5. In HANES5, the men age 18 and over had an average height of 69 inches and an SD of 3 inches. The histogram is shown below, with a normal curve. The percentage of men with heights between 66 inches and 72 inches is exactly equal to the area between ___(a)___ and ___(b)___ under the ___(c)___. This percentage is approximately equal to the area between ___(d)___ and ___(e)___ under the ___(f)___. Fill in the blanks. For (a), (b), (d) and (e), your options are

66 inches 72 inches −1 +1

For (c) and (f), your options are: normal curve, histogram



that adds 7 to the average.

(b) If you add 7 to each entry on a list, that adds 7 to the SD.

(c) If you double each entry on a list, that doubles the average.

(d) If you double each entry on a list, that doubles the SD.

(e) If you change the sign of each entry on a list, that changes the sign of the average.

(f) If you change the sign of each entry on a list, that changes the sign of the SD.

9. Which of the following are true? false? Explain or give examples.

(a) The median and the average of any list are always close together.

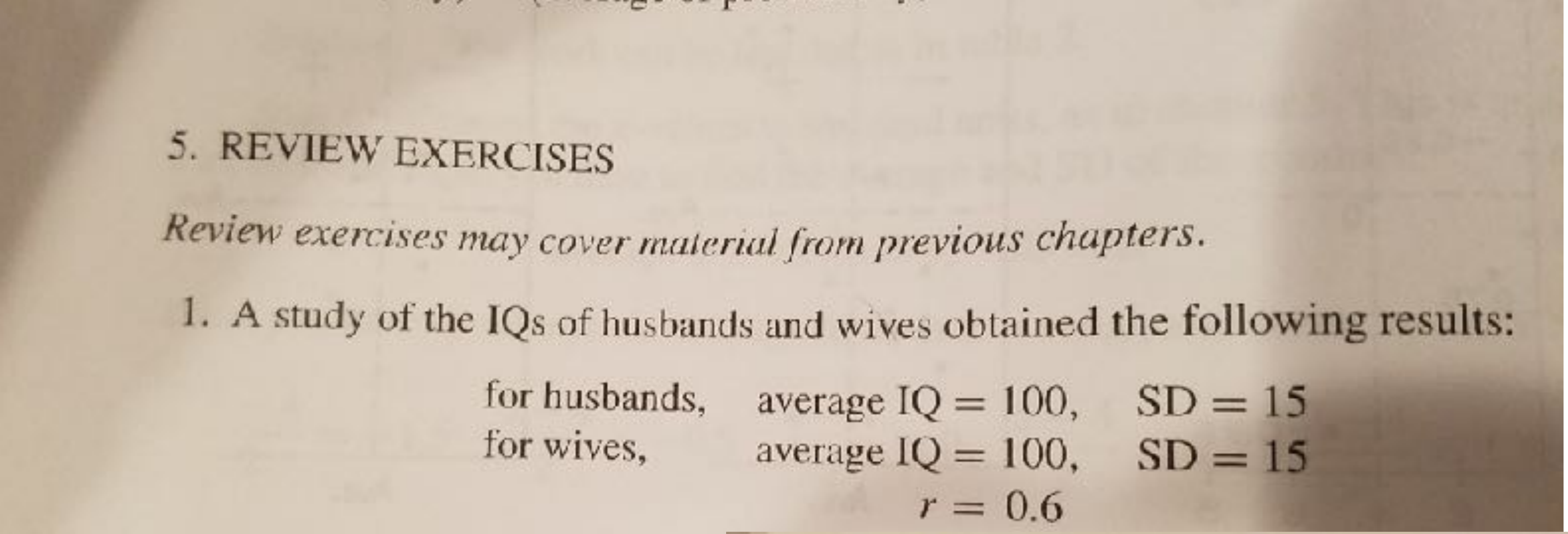
(b) Half of a list is always below average.

(c) With a large, representative sample, the histogram is bound to follow the normal curve quite closely.

(d) If two lists of numbers have exactly the same average of 50 and the same SD of 10, then the percentage of entries between 40 and 60 must be exactly the same for both lists.

10. For women age 25–34 with full time jobs, the average income in 2004 was $32,000. The SD was $26,000, and 1/4 of 1% had incomes above $150,000. Was the percentage with incomes in the range from $32,000 to $150,000 about 40%, 50%, or 60%? Choose one option and explain briefly.[5]

11. One term, about 700 Statistics 2 students at the University of California, Berkeley, were asked how many college mathematics courses they had taken, about 1.1; the SD



5. REVIEW EXERCISES

Review exercises may cover material from previous chapters.

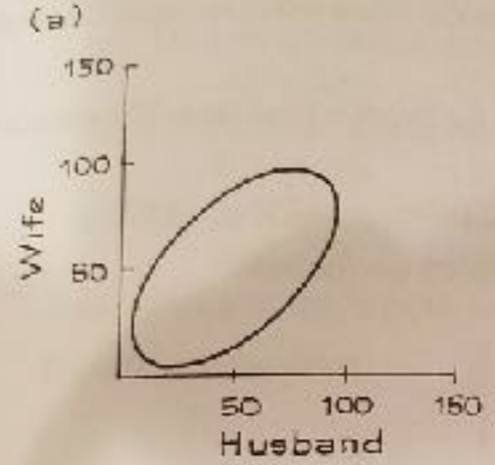
1. A study of the IQs of husbands and wives obtained the following results:

for husbands,	average IQ = 100,	SD = 15
for wives,	average IQ = 100,	SD = 15

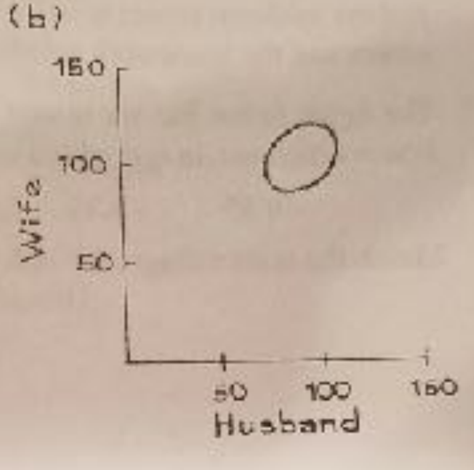
$$r = 0.6$$

One of the following is a scatter diagram for the data. Which one? Say briefly why you reject the others.

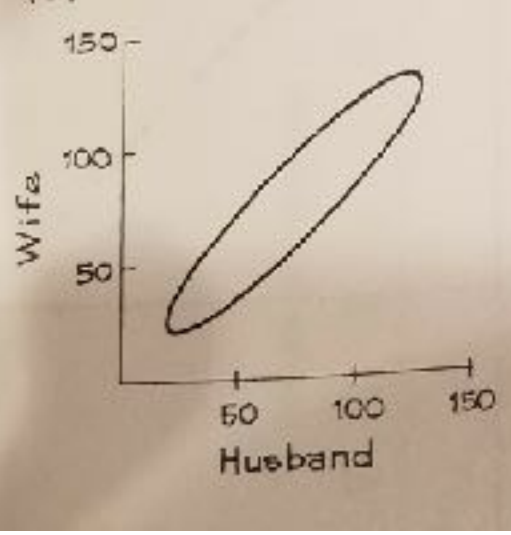
(a)



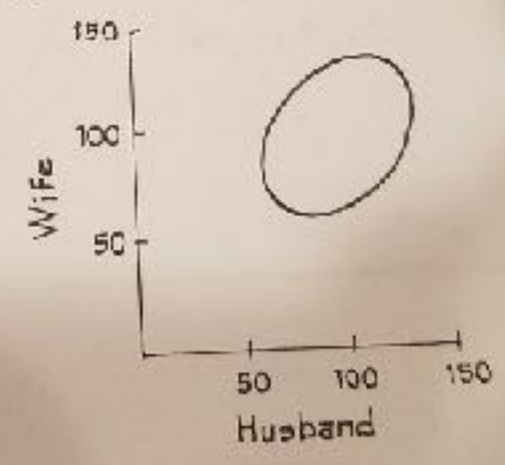
(b)



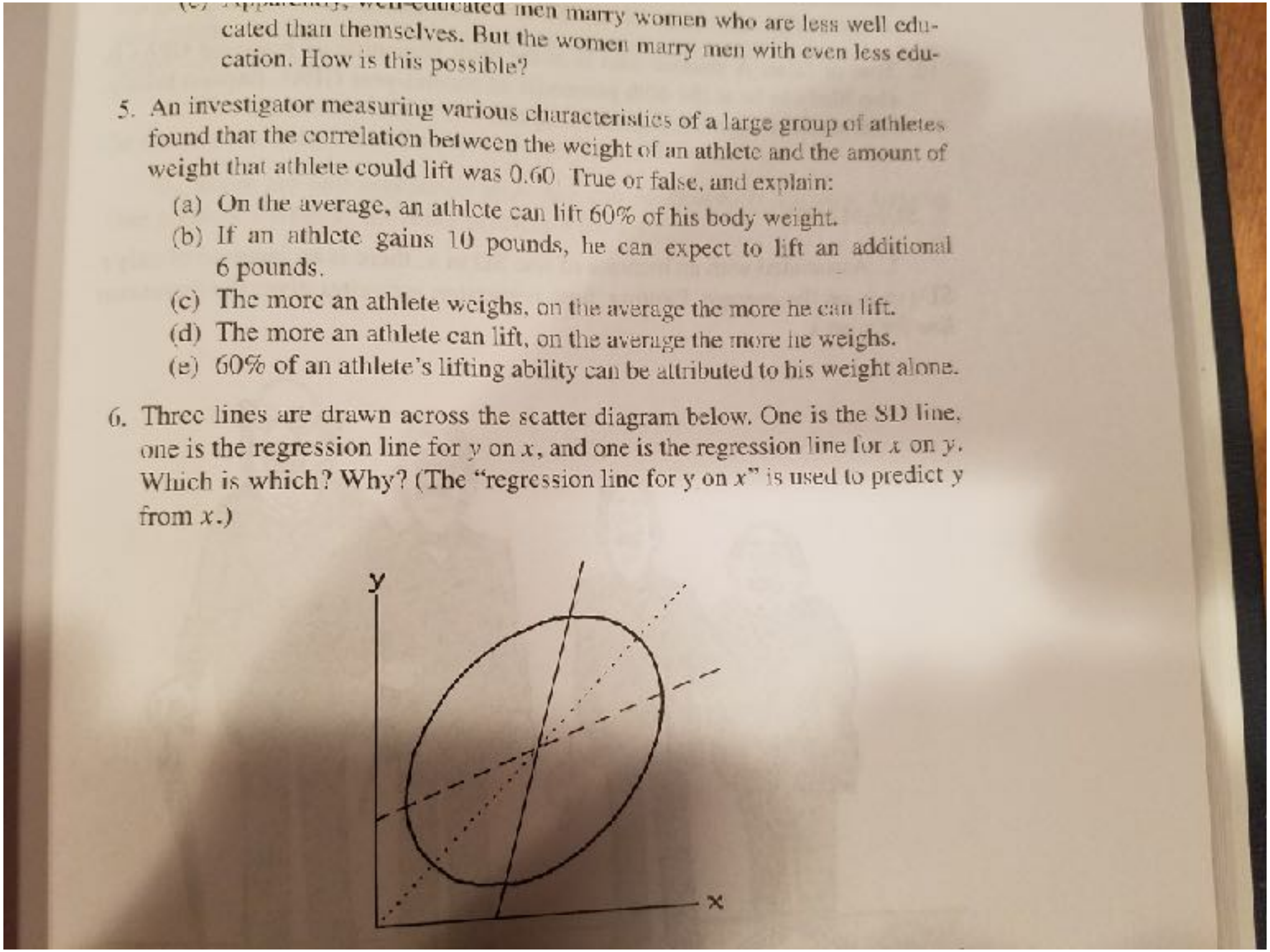
(c)



(d)



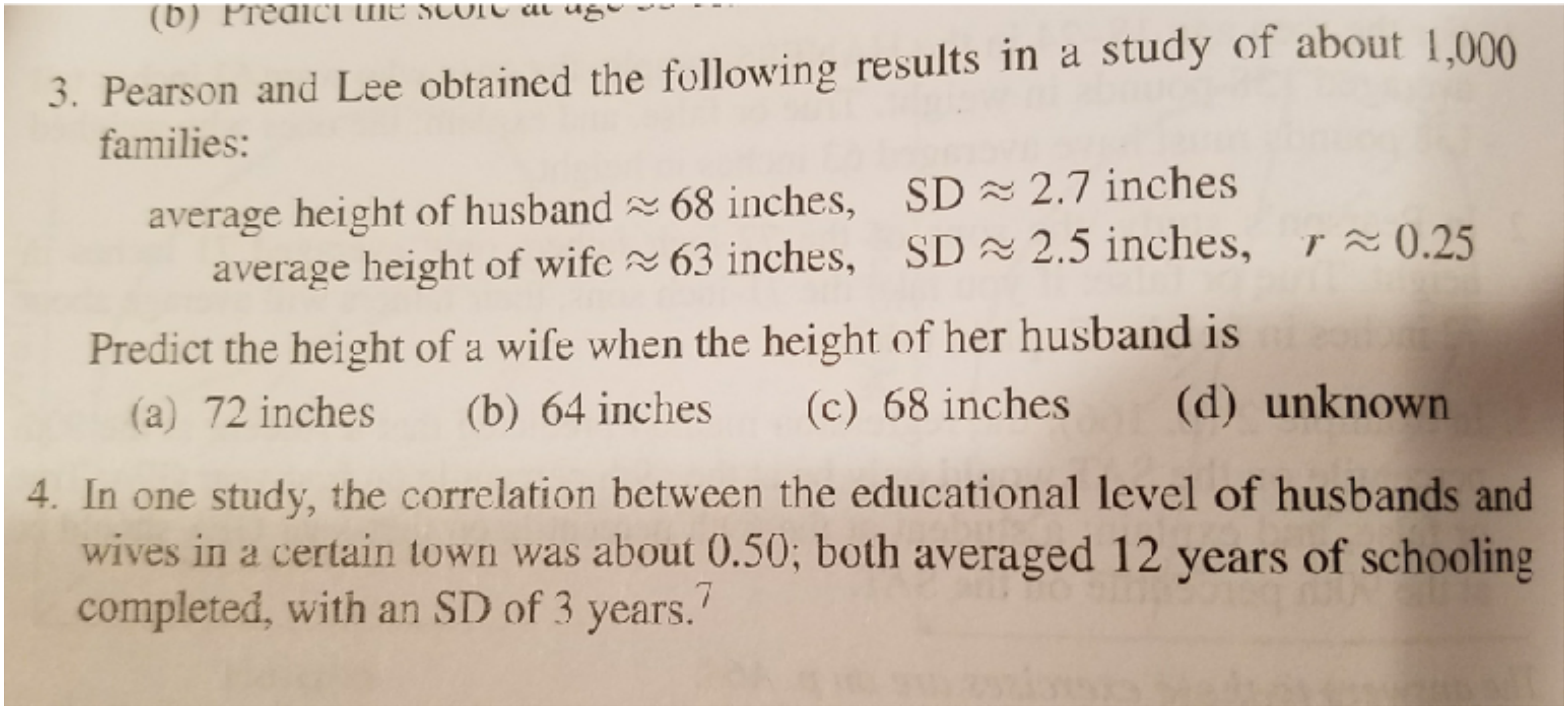
the correlation between the



(c) Apparently, well-educated men marry women who are less well educated than themselves. But the women marry men with even less education. How is this possible?

5. An investigator measuring various characteristics of a large group of athletes found that the correlation between the weight of an athlete and the amount of weight that athlete could lift was 0.60. True or false, and explain:
   (a) On the average, an athlete can lift 60% of his body weight.
   (b) If an athlete gains 10 pounds, he can expect to lift an additional 6 pounds.
   (c) The more an athlete weighs, on the average the more he can lift.
   (d) The more an athlete can lift, on the average the more he weighs.
   (e) 60% of an athlete's lifting ability can be attributed to his weight alone.

6. Three lines are drawn across the scatter diagram below. One is the SD line, one is the regression line for y on x, and one is the regression line for x on y. Which is which? Why? (The "regression line for y on x" is used to predict y from x.)



(b) Predict the score at age

3. Pearson and Lee obtained the following results in a study of about 1,000 families:

$$\text{average height of husband} \approx 68 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches}$$
$$\text{average height of wife} \approx 63 \text{ inches}, \quad \text{SD} \approx 2.5 \text{ inches}, \quad r \approx 0.25$$

Predict the height of a wife when the height of her husband is

(a) 72 inches (b) 64 inches (c) 68 inches (d) unknown

4. In one study, the correlation between the educational level of husbands and wives in a certain town was about 0.50; both averaged 12 years of schooling completed, with an SD of 3 years.[7]