

## Midterm

February 7

Your section: \_\_\_\_\_ Print your name: \_\_\_\_\_

Sign your name: \_\_\_\_\_

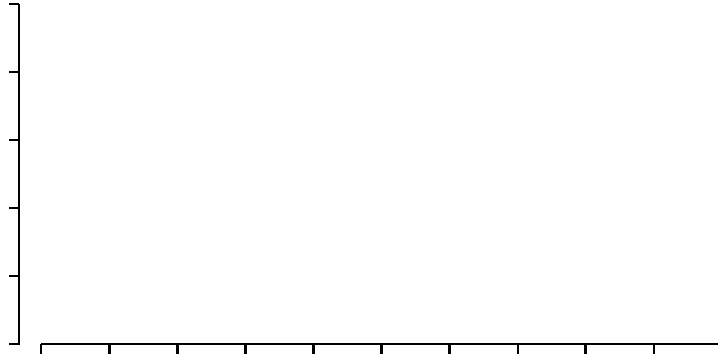
This is a closed book exam. However, you are allowed to bring two sheets (double-sided) of  $8.5 \times 11$  paper with notes. The midterm exam consists of five problems, and there is a normal table provided at the end. The exam carries 56 points but the maximum you can score is 50. Good luck!

Problem:	.. 1 ..	.. 2 ..	.. 3 ..	.. 4 ..	.. 5 ....	Sum
Points:	10	8	16	12	10	56

**Problem 1.** The cuckoo is a common European bird noted for its characteristic call and its habit of laying its eggs in the nests of other birds, which hatch and rear the young cuckoos. The following table shows the distribution of the length of 243 cuckoos' eggs measured in millimeters.

<u>Length (in millimeters)</u>	<u>Percent</u>
19-21	8.0
21-22	20.0
22-23	40.0
23-24	24.0
24-26	8.0

- Plot the histogram for this data set. Show all intermediate work. Mark the horizontal and vertical scales carefully. Label the axes.
- You find a cuckoo's egg and its length is 23.2 millimeters. Is that above or below the median? Explain your answer.



**Solution:** Millimeters ought to be plotted on the horizontal axis and percentage per millimeter on the vertical axis. The heights of the blocks over the class intervals are computed as follows:

<u>Length (in millimeters)</u>	<u>Percent/millimeter</u>
19-21	$8.0/2 = 4.0$
21-22	$20.0/1 = 20.0$
22-23	$40.0/1 = 40.0$
23-24	$24.0/1 = 24.0$
24-26	$8.0/2 = 4.0$

It is best to choose one unit of length on the vertical axis as 8 % per millimeter and 1 unit of length on the horizontal axis as 1 millimeter for the purpose of constructing the histogram.

**Problem 2.** True or false? (You need not show work here.)

	True	False
(a) If you subtract 17.5 from each entry on a list, that subtracts 17.5 from the average.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
(b) If you add 17.5 to each entry on a list, that adds 17.5 to the median.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
(c) If you add 17.5 to each entry on a list, that adds 17.5 to the SD.	<input type="checkbox"/>	<input checked="" type="checkbox"/>
(d) If you divide each entry on a list by 2, that divides the average by 2.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
(e) If you double each entry on a list, that doubles the SD.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
(f) If you change the sign of each entry on a list, that changes the sign of the average.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
(g) If you change the sign of each entry on a list, that changes the sign of the SD.	<input type="checkbox"/>	<input checked="" type="checkbox"/>
(h) If each entry on a list is between 0 and 1, then the SD is between 0 and 1 too.	<input checked="" type="checkbox"/>	<input type="checkbox"/>

### Problem 3

- (a) A list has 15 entries. Each entry can be either 2 or 3 or 5. What must the list be if its average is 5 ? Explain briefly.

**Solution:** Since each entry of the list is 2 or 3 or 5, the average can be at most 5. Since the average is given to be exactly 5, this means that each entry on the list must be exactly 5; otherwise the average would have to be strictly less than 5.

- (b) Investigators have been studying the relationship between income and education, for a very large sample of women age 25–54 who are working. Investigator A computes the correlation between individual income and individual education for all the women. Investigator B looks at each state separately, computes the average income and average education for that state – and then computes the correlation coefficient for the 50 pairs of state averages. Which investigator gets the higher correlation? Or should the correlations be about the same? Explain carefully.

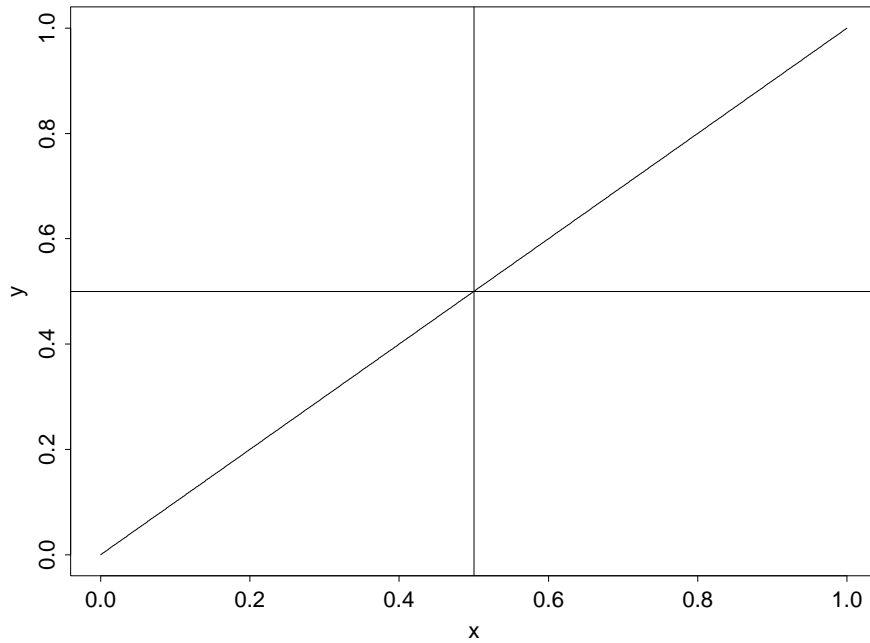
**Solution** Investigator B gets a higher correlation. This has to do with the fact that A measures the raw correlation while B looks at the ecological correlation which is formed by averaging out the income and education over each state. Ecological correlations tend to overstate the strength of an association. Within each state, there is a lot of spread in the income-education scatterplot about the averages; replacing the states by their averages eliminates the spread and gives a misleading impression of tighter clustering.

- (c) A report by the Environmental Defense Association is discussing the relationship between air pollution and annual death rates for a sample of 47 major cities in the U.S. The average death rate is reported as 9/1000 and the SD is 3/1000. The rms error of the regression line for predicting death rate from air pollution is reported as 4/1000. Is there anything wrong with the numbers ? Or do you need more information to decide ?

**Solution** There is obviously something wrong here. The rms error of the regression line for predicting death rate from air pollution has to be less than or equal to the standard deviation of the death rate. But here it is greater and that is not possible.

- (d) You have information on a pair of variables (X,Y) for 50 individuals. The variables X and Y have the same average and standard deviation and the cor-

relation between  $X$  and  $Y$  is 0. On a graph, show schematically the SD line, the regression line of  $Y$  on  $X$  and the regression line of  $X$  on  $Y$ .



The above figure shows the three lines. We take the point of averages to be  $(0.5, 0.5)$ . Because the averages and SD's of X and Y coincide, the SD line is precisely the line whose equation is  $y = x$  and this is the slant line in the figure. Because the correlation between Y and X is 0, the regression line of Y on X is the horizontal line in the figure passing through the point of averages and the regression line of X on Y is the vertical line in the figure, also passing through the point of averages.



**Problem 4.** For each of the situations described below, fill in the blank with one of the following five options:

exactly  $-1$    somewhat negative   exactly  $0$    somewhat positive   exactly  $1$

Then explain briefly.

- (a) For the students taking Statistics 220 in Winter quarter 2001, the correlation coefficient between the score on the midterm and the score on the final will be \_\_\_\_\_.

Brief explanation: somewhat positive. Students who score higher on the midterm will also in general score higher on the final.

- (b) Suppose that a class meets 30 days in a quarter. For each student in the class we record the number of days they were present and the number of days they were absent. The correlation co-efficient between the number of days present and the number of days absent is \_\_\_\_\_.

Brief explanation: exactly  $-1$ . Since no. of days present =  $30 -$  no. of days absent.

- (c) For the data set shown below, the correlation coefficient is \_\_\_\_\_.

$x$	$y$
1	10
1	-10
5	8
5	-8

Brief explanation: exactly  $0$ . Because for each  $x$  value, you get two  $y$  values which are equal in magnitude and of opposite signs, the sum of the products when the variables are expressed in standard units is exactly  $0$  (the average of the  $y$  values is  $0$ ).

- (d) The correlation between temperature readings in Fahrenheit and temperature in Centigrade on a particular day at state capitals throughout the U.S will be \_\_\_\_\_ .

Brief explanation: exactly 1. The Celsius reading  $C$  is related to the Fahrenheit reading  $F$  by the relation

$$\frac{C}{5} = \frac{F - 32}{9} .$$

**Problem 5.** In a very large class, the midterm had an average of 50 points with an SD of 20. The final scores averaged out to 60 with an SD of 15. The correlation between midterm and final scores was 0.5, and the scatter diagram was football-shaped.

- (a) Solly scores 60 on the midterm and 90 on the final. However, the instructor loses her midterm score and asks her to provide the score. Solly knows that the instructor will estimate her midterm score by the regression method if she refuses to cooperate. In order to maximize her score, should she cooperate or not? Answer with “yes” or “no”, and explain carefully. Show all your work.

**Solution:** The regression method gives Solly’s predicted score for the midterm as

$$50 + 20 \times 0.5 \times \frac{90 - 60}{15} = 50 + 20 = 70.$$

So, it does not make sense for Solly to cooperate.

- (b) Among the students who scored 60 on the midterm, about which percentage scored better than Solly, that is, higher than 90, on the final? Show all your work.

**Solution:** Consider the subpopulation of students that scored 60 on the midterm. Their average score on the final is obtained by the regression method as

$$60 + 15 \times 0.5 \times \frac{60 - 50}{20} = 63.75 = 64 \text{ (appx).}$$

The SD for this subpopulation is given by the RMS error for the regression line for final score on midterm score; this is

$$\sqrt{1 - (0.5)^2} \times 15 = 13 \text{ (appx).}$$

Converting 90 to standard units using the new average and SD we get

$$\frac{90 - 64}{13} = 2.$$

To find the desired percentage we need to find the area to the right of 2 under the normal curve; this is approximately 2.3 % using the normal table at the back of the book.