

## 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" × 11" paper with notes. The midterm exam consists of five problems. The exam carries 52 points but the maximum you can score is 50. Good luck!

Problem:	.. 1 ..	.. 2 ..	.. 3 ..	.. 4 ..	.. 5 ....	Sum
Points:	10	6	12	16	8	52

**Problem 1.** The following table summarizes data on annual inches of precipitation from 1961-1990 at the Los Angeles Civic Center.

<u>Inches of Precipitation</u>	<u>Percent</u>
0-5	3.0
5-10	33.0
10-15	23.0
15-20	20.0
20-30	14.0
30-35	7

- Plot the histogram for this data set. Show all intermediate work. Mark the horizontal and vertical scales carefully. Label the axes.
- What can be said about the median of this data set ? Give as much information as possible.

**Solution:** Inches of precipitation should be on the horizontal axis; percentage per inch of precipitation should be on the vertical axis. The heights of the blocks over the given class intervals will be

Inches of Precipitation	Percent
0-5	$3.0/5 = 0.6$
5-10	$33.0/5 = 6.6$
10-15	$23.0/5 = 4.6$
15-20	$20.0/5 = 4.0$
20-30	$14.0/10 = 1.4$
30-35	$7/5 = 1.4$

For more explanation, consult your TA's.

Clearly the median is between 10 and 15, since 36 % of the observations are less than 10 and 59 % of the observations are less than 15. A good estimate of the median would be  $10 + (50-36)/4.6 = 13.04$  (approx) (the area between 10 and 13.04 under the histogram is 14 %; this added to 36 % which is the area to the left of 10 makes 50 %, so the area to the left of 13.04 is 50 %).

**Problem 2.** A few years ago, the British House of Commons discussed whether or not an hourly minimum wage should be introduced. Specifically, the Labour party suggested that all wages under 3.40 pounds sterling should be increased to 3.40 pounds sterling, which is well below the median wage.

A key question in the debate was the effect of the suggested policy on the median wage. One discussant argued that the median “would not be affected”; another one claimed that “if the lowest wage were raised to 3.40 pounds sterling an hour the median would have to rise”.

Who is right ? Would the Labour Party’s suggestion lead to an increase in the median wage in the U.K or not ? Explain carefully.

**Solution:** The Labour Party’s suggestion would not change the median wage of salaries. To see this, suppose that the median wage is 5 pounds sterling an hour. Then 50 % of labourers earn under 5 pounds sterling an hour and of course there is a certain percentage (less than 50 %) who earn under 3.40 pounds sterling an hour. If the minimum wage is raised to 3.40 pounds sterling an hour, then of course there is nobody earning below that wage, but the percentage of labourers earning below 5 pounds sterling an hour still remains the same exact 50 % (the population of labourers earning below 5 pounds sterling an hour remains unaltered), showing that the median wage is unaffected.

### Problem 3

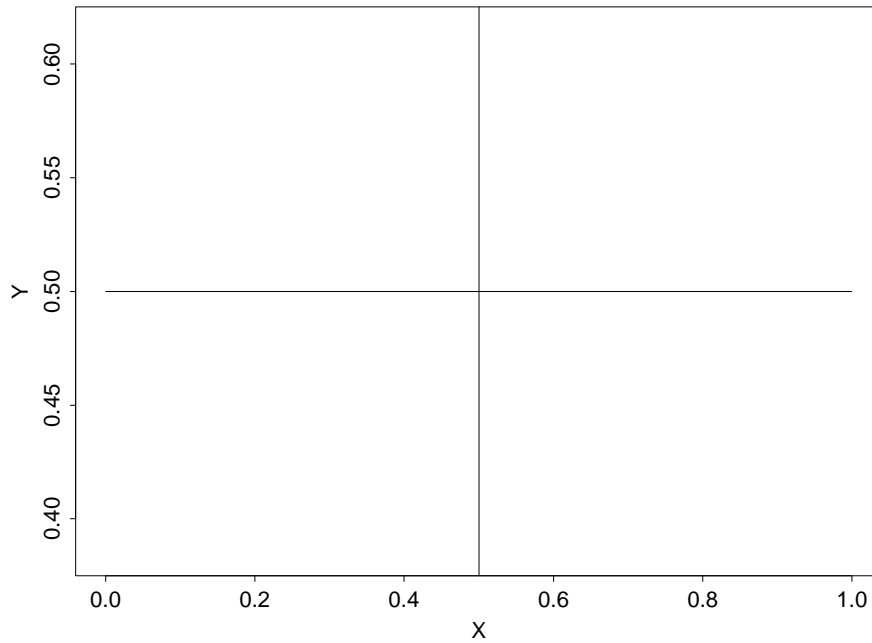
- (a) A list has 15 entries. One entry is 6 and the standard deviation of the list is 0. Can you reconstruct the list ?
- (b) An instructor standardizes her midterm and final each semester so that the class average is 50 and the SD is 10 on both tests. The correlation between tests is around 0.50. One semester, she took all the students who scored below 30 on the midterm and gave them special tutoring. They all score above 50 on the final Can this be explained by the regression effect ? Explain briefly.
- (c) When is the RMS error for the regression line used to predict Y from X, the same as the SD of Y ? In this case, graph schematically, the regression line of Y on X and that of X on Y.

**Solution**(a) If the SD is 0, then all entries in the list are the same and equal to the average value of the list. Hence, the list here consists of 15 6's.

**Solution** (b) No, this cannot be explained by the regression effect. The average score of the students, who scored say 30 on the midterm, on the final can be derived by the regression method as

$$50 + 10 \times 0.5 \times \frac{30 - 50}{10} = 50 - 10 = 40 .$$

So the regression method pulls the predicted scores on the final toward the global average but not past it. In this case, the increment in scores is too large to be explained just by the regression effect. The tutoring must have had an effect.



**Solution** (c) The RMS error for the regression line used to predict  $Y$  from  $X$  is the same as the SD of  $Y$  if and only if the correlation coefficient  $r = 0$ . In this case the regression lines of  $Y$  on  $X$  and of  $X$  on  $Y$  are shown in the above figure. The regression line of  $Y$  on  $X$  is the horizontal line and that of  $X$  on  $Y$  is the vertical line and they intersect at the point of averages which in the figure is taken to be  $(0.5, 0.5)$ .

**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) A test has 20 problems of the true-false type. The correlation between the number of right answers and the number of wrong answers for the students who took the test is \_\_\_\_\_.

Brief explanation: exactly  $-1$ . Number of wrong answers =  $20 -$  number of right answers. Hence the correlation coefficient is exactly  $-1$ .

- (b) For used cars that are less than 20 years old, the correlation between age and price is \_\_\_\_\_.

Brief explanation: somewhat negative. As age of the used car increases, price comes down because of depreciation. Antique cars don't figure, since the cars we consider, though used, are reasonably recent.

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

$x$	$y$
1	15
2	25
5	55
8	85

Brief explanation: exactly  $1$ . Because  $y = 10x + 5$  and so the scatter plot lies on a line sloping up.

- (d) For registered students at the UW, the correlation between student's age and father's year of birth is \_\_\_\_\_ .

Brief explanation: somewhat negative. Older students tend to have older fathers which means that the year of birth is lower.

**Problem 4.** We have pairs of observations on father's height and son's height for 500 pairs. The fathers are 68 inches tall on average with an SD of 2.7 inches. The sons are 69 inches tall on average with the same SD. The correlation is 0.5. For those sons that have fathers 70.7 inches tall, estimate the percentage that lie between 67 inches and 70 inches.

**Solution:** Consider the subpopulation of sons that have fathers 70.7 inches tall, this is 1 SD above the average father height. By the regression method the average height of the sons in this subpopulation is  $69 + 0.5 \times 2.7 \times 1 = 70.4$  inches (appx). The SD for the sons in this subpopulation is the RMS error for the regression line and that is  $\sqrt{1 - (0.5)^2} \times 2.7 = 2.34$  inches (appx). We can now use the normal approximation with the new average and SD to convert 67 and 70 to standard units;

$$\frac{67 - 70.4}{2.34} = -1.45 ; \quad \frac{70 - 70.4}{2.34} = -.17 .$$

To get the percentage we need to find the area between -1.45 and -.17 under the normal curve. This, using the normal table at the back of the book, is approximately,

$$\frac{85.29 - 11.92}{2} = 36.7\% .$$