

## Midterm

MAY 2

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 three problems, and there is a normal table provided at the end. The exam carries 60 points but the maximum you can score is 50. Good luck!

Problem:	.. 1 ..	.. 2 ..	.. 3 ....	Sum
Points:	26	14	20	60

**Problem 1.** In a test given to a group of 200 students with a total score of 100 points the following distribution table was obtained.

Score on the test	Number of students
0-10	30
10-40	50
40-60	60
60-90	40
90-100	20

- (a) Plot the histogram for this data set. Show all intermediate work. Mark the horizontal and vertical scales carefully. Label the axes.

You need to convert to percentages from the raw numbers for each class interval and then compute percentages per unit score as in the following table.

<i>C.I.</i>	raw no.	%	% per unit
0-10	30	15	3/2
10-40	50	25	5/6
40-60	60	30	3/2
60-90	40	20	2/3
90-100	20	10	1

The histogram can now be constructed as usual. Note that on the x-axis you should have score and on the y-axis percentage per unit score.

- (b) Give an estimate the median of the above histogram.

Since 40 % of the students scored below 40 and 70 % of the students scored below 60, the median is between 40 and 60. Now there are 30 % of students who scored between 40 and 60, so 30 % of the area of the histogram is taken up by the block between 40 and 60, which is 20 units long. Now  $20/3 = 6\frac{2}{3}$  and so 10 % of the area lies under the block between 40 and  $46\frac{2}{3}$ ; thus 50 % of the area lies to the left of  $46\frac{2}{3}$  showing that this is the median.

- (c) Assume now that in each of the class intervals above, all students had scores exactly at the mid-point of the class intervals. Thus, all students between 0 and 10 scored exactly 5, all students between 10 and 40 scored exactly 25 and so on. Find the average score of the students on the test.

$$\text{Average} = \frac{5 \times 30 + 25 \times 50 + 50 \times 60 + 75 \times 40 + 95 \times 20}{200} = 46.5.$$

- (d) Make the best guess as to the S.D. of the scores, among the three given options : 15, 25, 35.

The best guess is 25. 15 is clearly too small because you get like more than 30 % of observations outside 2 standard deviations from the mean; 35 is too large since almost all scores are then within 1.5 standard deviations of the mean. 25 seems about the right spread.

(10 + 5 + 5 + 6 = 26 points)

## Problem 2

- (a) Two different investigators are working on a growth study. The first measures the heights of 100 children in inches. The second prefers the metric system and changes the result to centimeters (using the conversion factor, 1 inch = 2.54 centimeters). A scatter diagram is plotted, showing for each child, its height in inches on the horizontal axis and the corresponding height in centimeters on the vertical axis.

- (i) If no mistakes are made in the conversion, what is the correlation ?

**Solution:** Height (converted) in centimeters = 2.54 times height in inches; so the scatter plot is a perfect straight line with positive slope and the correlation is 1.

- (ii) What happens to  $r$ , the correlation if there is a mistake in the arithmetic ?

**Solution:** If there is a mistake in multiplication the scatter plot will typically cease to be a straight line and the correlation will go down from 1. If a wrong conversion factor is used but the multiplication is still correct the correlation will still be 1.

- (iii) What typically happens to  $r$ , if the second investigator goes out and measures the same children again, using metric equipment ? (3 + 2 + 3 = 8 points)

**Solution:** Here, we are plotting the first investigator's measurement converted to centimeters against the second investigator's measurement (in centimeters). If they both measured the students accurately, their measures would coincide and you would get a correlation of 1. But typically there will be little variations in measurement (measurement error) and the correlation will go down from 1, though it will still have a high positive value.

- (b) Scatter diagrams are plotted for two different data sets and eye-estimation shows that the scatter diagram for the first data set is more clustered around the SD line as compared to the scatter diagram for the second data set. What can we conclude about the correlation between X and Y in the first data set, compared to the correlation between X and Y in the second data set ? (6 points)

**Solution:** Recall that correlation measures scatter relative to the standard deviation. We do not know about the standard deviations for the variables

in the different data sets, hence cannot judge whether a higher scatter in one data set is actually a reflection of a higher correlation or a reflection of a higher standard deviation. Recall the example that I showed in class.

**Problem 3.** Scores on midterm and final for 200 students are provided and the following information is obtained. The average midterm score is 70 with an SD of 10 and the average final score is 55 with an SD of 20. The correlation between the midterm score and final score is 0.6.

(i) Find the regression equation for predicting the final score from the midterm score.

**Solution:** The slope of the regression line is

$$\text{slope} = r \times \frac{SD(\text{final})}{SD(\text{midterm})} = 0.6 \times \frac{20}{10} = 1.2.$$

The intercept is

$$\text{avg}(\text{final}) - \text{slope} \times \text{avg}(\text{midterm}) = 55 - 1.2 \times 70 = -29.$$

Thus the equation to the regression line is

$$y = -29 + 1.2x.$$

(ii) Predict what a student who gets 85 on the midterm would score on the final.

**Solution:** We plug in 85 for  $x$  in the above equation to get

$$\text{Predicted final score} = -29 + 1.2 \times 85 = -29 + 102 = 73.$$

(iii) How would you interpret the slope of the regression line in this example?

**Solution:** The slope of the line gives the difference in the average final score between the subpopulation of students who get a certain fixed score on the midterm and the subpopulation who score one point more; thus it is the change in the average final score per unit increase in the midterm score.

(iv) Suppose that the histogram for the scores on the final follows the normal curve. Estimate the number of students that scored between 25 and 85 on the final.

**Solution:** 25 is 1.5 S.D.'s below the avg. final score and 85 is 1.5 S.D's above. Thus we need to find the area under the normal curve between -1.5 and 1.5 to get the percentage that scored between 25 and 85 on the final. This is approximately 87%, which translates to around 174 students.

( 6 + 3 + 4 + 7 = 20 points)