

Sample Final Exam

March 13

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. This final exam consists of 12 problems; you have to answer one out of Problems 7 and 8 and any eight problems out of the remaining 10. The maximum you can score is 100.

Please select one of the following options for returning your graded exam.

- I am providing a self-addressed stamped envelope (55c postage).
Mail the exam.
- Deposit the exam in the box in front of room B-301 at Padelford Hall.
I will pick it there.
- Hold the exam. It will be available during the instructor's office hours
in Spring quarter, 2001.

Good luck!

Problem:	.. 1 2 3 4 5 6 ..
Points:	11	11	11	11	11	11
Credit:						

Problem:	.. 7 8 9 10 11 12 ..	Sum
Points:	12	12	11	11	11	11	100
Credit:							

Problem 1. (i) Some studies find an association between liver cancer and smoking. However alcohol consumption is a confounding variable. This means

(i) Alcohol causes liver cancer.

(ii) Drinking is associated with smoking, and alcohol causes liver cancer.

(ii) Sketch schematically (just show the shape by means of a curve) a histogram for income distribution in the U.S. What is the relation between mean income and median income ? (4+7=11 points)

Solution: (i) Option (ii) is correct. In this case the “treatment group” is smokers and the “control group” is non-smokers and they differ systematically with respect to drinking, drinking being associated more with smokers than non-smokers. Drinking however increase the risk of liver cancer which is the “response” being measured and hence becomes a confounder.

(ii) The histogram for income distribution is right skewed; so it has a long right tail. Look at page 34, problem 7 of the text book for a schematic shape of the income distribution. Recall that the mean is affected by large values; for the income distribution the mean is larger than the median. It is pulled to the right of the median by the large values in the right hand tail.

Problem 2. (i) Construct a list of 5 numbers for which the mean and the median are the same.

(ii) Here is a list of numbers: 0.7, 1.6, 9.8, 3.2, 5.4, 0.8, 7.7, 6.3, 2.2, 4.1, 8.1, 6.5, 3.7, 0.6, 6.9, 9.9, 8.8, 3.1, 5.7, 9.1

(a) Without doing any arithmetic guess whether the average is around 1, 5 or 10. Justify briefly.

(b) Without doing any arithmetic guess whether the SD is around 1, 3 or 6. Justify briefly.

(3+8 = 11 points)

Solution (i) Any list that is symmetric about the median will serve the purpose. Take for example the numbers 1,2,3,4,5. Both the mean and the median are 3.

(ii) (a) All numbers in the list are less than 10, so 10 cannot be the average. There are quite a few large values like 9.1, 8.1 7.7 etc. so the average must be larger than 1. That leaves 5 as the only choice for an approximate value for the average.

(b) All numbers in the list deviate from the average by less than 6, so the SD has to be less than 6. Most of them deviate by more than 1, so that the SD has to be more than 1. That leaves 3 as the only option.

Problem 3. Use the techniques developed in class to find the correlation coefficient for the following data set. (Do not use the correlation function on a statistical calculator.) Show all your work. (11 points)

x	y
3	11
5	6
7	4
5	9
8	1
2	3

Solution: The exact computations will not be shown. Follow the procedure shown on pages 132-133 of the text book. The correlation between the variables x and y is -0.464 .

Problem 4. (i) A law school finds the following relationship between LSAT scores and first-year scores (for students who finish the first year):

average LSAT score = 162, SD = 6

average 1st year score = 68, SD = 10, $r = 0.60$

Of the students who scored 165 on the LSAT, about what percentage had first-year scores of over 75 ?

(ii) Find the correlation between X and Y where

x	y
3	7
5	13
7	19
9	25
11	31

(7 + 4 = 11 points)

Solution: (i) The solution to this problem is available on pages 195-196 of the text book. We use the normal curve inside a vertical strip to figure out the required chance. Work through the steps carefully.

(ii) It is not difficult to see that $y = 3 * x - 2$; hence x and y lie on a straight line with positive slope and the correlation is 1.

Problem 5. (i) When is the R.M.S error for the regression line for Y on X, the same as the SD of Y ?

(ii) A statistician computes the slope of the regression line of Y on X as 2.5 and the slope of the regression line of X on Y as -.3. Do you think there is something wrong with his computations ? Justify your answer.

(iii) Another statistician now recomputes the slope of the regression line of Y on X as 2.5 and the slope of the regression line of X on Y as 0.5. Do you think he is computing the slopes correctly, or is he wrong ? (3 + 4 + 4 = 11 points)

Solution: (i) The R.M.S error for the regression line is given by

$$R.M.S. \text{ error} = \sqrt{1 - r^2} * S.D.(Y) .$$

Hence the R.M.S. error and the S.D. coincide if and only if $r = 0$. Here r is the correlation coefficient.

(ii) There is indeed something wrong. The slope of the regression line of Y on X and that of X on Y must have the same sign, the sign of the correlation coefficient.

(iii) There is something wrong here as well. The slope of the regression line of Y on X is

$$\text{slope}_{YX} = r * \frac{S.D.(Y)}{S.D.(X)} .$$

Consider now the regression line of X on Y. If Y is still plotted on the vertical axis and X on the horizontal axis the slope of this line is

$$\text{slope}_{XY} = (1/r) * \frac{S.D.(Y)}{S.D.(X)} .$$

Since r in absolute value is at most 1 this means that the absolute value of the slope of Y on X is less than the absolute value of the slope of X on Y. But this is clearly not the case here.

Problem 6. A die is rolled four times. What is the chance that

- (i) not all the rolls show 3 or more spots ?
- (ii) none of the rolls show 3 or more spots ?
- (iii) the sum of the numbers on the rolls is 24 given that the first roll is a 5 ?

(4 + 4 + 3 = 11 points)

Solution:(i) The total number of outcomes on 6 rolls is 6^4 . The chance that not all the rolls show 3 or more spots is 1 minus the chance that all the rolls show three or more spots. The total number of outcomes for which all four rolls show 3 or more spots is 4^4 (since the first roll can show 3 or 4 or 5 or 6 and so on for the second, third and fourth, showing that there are 4 possible outcomes on each of the four rolls). Thus the chance that all rolls show at least 3 spots is $4^4/6^4 = (2/3)^4$. Therefore the chance that not all the rolls show 3 or more spots is $1 - (2/3)^4$.

(ii) Each roll then shows either 1 or 2. This leaves 2^4 or 16 possibilities and the corresponding chance is $2^4/6^4 = 1/3^4$.

(iii) Given that the first roll is a 5 there is no way that the sum of the spots on the rolls can be 24; you need four consecutive sixes for this to happen. Thus the conditional chance is 0.

Problem 7. A fair coin is tossed 400 times. You get to pick 21 numbers. If the number of tails turns out to be equal to one of your 21 numbers, you win 100 dollars. Which 21 numbers should you pick and what is approximately your chance of winning 100 dollars ? (12 points)

Solution: The experiment is like drawing 400 times at random with replacement from a box that has one 0 and one 1, where 1 stands for tails and 0 for heads. The number of tails is the sum of the numbers in 400 draws. The average and the S.D. of the box are both equal to $1/2$. The expected value for the sum of the draws is $1/2 \times 400 = 200$ and the S.E. for the sum of the draws is $\sqrt{400} \times S.D. = 20 \times 1/2 = 10$. Thus the sum of the 400 numbers will be around 200 give or take 10 or so and the chance that it is between 190 ($200-10$) to 210 ($200 + 10$) is approximately 68 % by the normal approximation (190 is -1 in standard units and 210 is + 1). There are 21 numbers between 190 and 210 (inclusive) and it is these 21 numbers that you must choose to maximize your chance of winning 100 dollars. The corresponding chance is 68 % as pointed out before.

Problem 8. Sam and Russ each toss a die. If Sam gets 3 or more spots she records a 3, otherwise she records a 1. If Russ gets 3 or more spots he records a 1, otherwise he records a 3. They then multiply the numbers that they have recorded. They repeat the experiment a 100 times. The average of the 100 products will then be around (a) give or take (b) or so. Fill in (a) and (b) with appropriate numbers. (12 points)

Solution: The key step in the problem is to set up a box model for the product of the numbers that Russ and Sam get. Since Russ records 1 or 3 and so does Sam, the product of Russ's number and Sam's number can take 3 values, 1, 3 or 9. Thus we have the numbers on the tickets that should be in the box. To find how many of each kind we have to find the chance that the product takes on a specific value. Now

Chance that Sam records 1 = Chance that Sam gets 1 or 2 on the roll = $2/6 = 1/3$.

Chance that Sam records 3 = $1 - 1/3 = 2/3$.

Chance that Russ records 1 = Chance that Russ gets 3 or more spots on the roll = $4/6 = 2/3$.

Chance that Russ records 3 = $1 - 2/3 = 1/3$.

Chance of product being 1 = Chance that Sam records 1 \times Chance that Russ records 1 = $1/3 \times 2/3 = 2/9$.

Chance of product being 3 = Chance that Sam records 1 \times Chance that Russ records 3 + Chance that Sam records 3 \times Chance that Russ records 1
= $1/3 \times 1/3 + 2/3 \times 2/3 = 5/9$.

Chance of product being 9 = Chance that Sam records 3 \times Chance that Russ records 3 = $2/3 \times 1/3 = 2/9$.

Thus we now have the composition of the box as well, to match up with the chances of getting various values of the product; there should be 2 tickets marked 1, 5 tickets marked 3 and 2 tickets marked 9.

The expected value for the average of 100 draws from this box is equal to the average of the box. The average of the box is

$$\frac{2 \times 1 + 5 \times 3 + 2 \times 9}{9} = 35/9 = 3.89 \text{ (} \textit{appx.} \text{)} .$$

The S.D. of the box is 2.8 (appx). The S.E. for the sum of draws is

$$\sqrt{\textit{number of draws}} \times 2.8 = 10 \times 2.8 = 28 .$$

The S.E. for the average is

$$28/100 = 0.28 .$$

Thus (a) gets filled with the average of the box which is 3.89 and (b) gets filled with the S.E. for average of 100 draws which is 0.28.

Problem 9. A multiple-choice quiz has 50 questions. Each question has four possible answers, one of which is correct. Eight points are given for each correct answer, but 2 points are taken off for a wrong answer.

A student answers all 50 questions at random (by guessing).

- (a) The student's score is like the _____ of _____ draws made at random _____ replacement from the box



(specify an appropriate box)

For the first space, your options are *sum* and *average*. For the second space, select a number, and for the third space, choose from *with* and *without*.

- (b) If the passing score is 100, what is the student's chance of passing? Show all your work.

(11 points)

Solution: The blanks get filled with *sum*, 50 and *with* respectively. The box has 1 ticket marked 8 and 3 tickets marked -2.

Average of box = $(8 - 2 - 2 - 2)/4 = 1/2$.

S.D. of box = 4.33. The expected value for the sum of 50 draws is then given by $1/2 \times 50 = 25$. The S.E. for the sum of 50 draws is $4.33 \times \sqrt{50} = 30.6$. Now the sum of 50 draws approximately follows the normal curve. In standard units 100 corresponds to

$$\frac{100 - 25}{30.6} = 2.45.$$

The student's chance of passing, that is scoring over 100 is then given approximately by the area under the normal curve to the right of 2.45 and this using the normal table is approximately 0.7 %.

Problem 10. (a) If a sampling procedure is biased, can I rectify it by taking a larger sample ? Explain briefly.

(b) Is it always true that the absolute size of the sample and not its size relative to the population determines the accuracy of the sample percentage for the population percentage ? Discuss briefly. (4+7 = 11 points)

Solution: (a) No, you cannot. Look at page 335 of the text book for a discussion. With a biased selection procedure you are systematically excluding a certain part of the population and taking a larger sample will still exclude that part. The flaw lies in the method that use to pick an individual in the sample.

(b) No, this is not the case. The size of the sample relative to the population needs to be accounted for, if the sample is a substantial fraction of the population size. The accuracy is measured by the S.E. and the correction factor needs to be used if the sample is not very small compared to the population; the correction factor takes the relative size into account. See the discussion on page 368 of the text book.

Problem 11. You want to estimate the proportion of Democrats in a town with 100000 eligible voters. You also want an accuracy measure for your estimate for the proportion. Cost constraints allow you to interview only 500 voters. Describe clearly how you would proceed to do so. (11 points)

Solution: You first need to get a simple random sample of 500 voters. This can be done in principle by taking a huge box with 100000 tickets bearing the voter identities and then selecting 500 at random without replacement. In practice this is done by different methods.

The estimate of the population proportion(percentage) of Democrats is precisely the proportion(percentage) of Democrats in the sample (since the expected value for the sample percentage equals the population percentage).

If we knew the proportion(percentage) of Democrats in the population the accuracy of the sample percentage could be computed as:

$$\frac{1}{500} \times \sqrt{500} \times \sqrt{f \times (1 - f)} \times 100\%$$

where f is the proportion or fraction of Democrats in the population. Since we do not know f , we shall estimate it by the fraction of Democrats in the sample of size 500 and plug that estimate in the above expression. This is the method of bootstrap. The estimate of f is precisely the sample fraction(proportion) of Democrats.

Problem 12. One ticket is drawn at random from two boxes A and B below.

Box A: 1,2,3,4,5.

Box B:1,2,3,4,5,6.

Find the chance that the sum of the numbers is 7. Also find the chance that one of the numbers is strictly bigger than twice the other. (4+7=11 points)

Solution: The sum of the numbers is 7 if and only if we get the pairs

$(3, 4), (4, 3), (2, 5), (5, 2), (1, 6),$

where the first component in each pair is what we get from Box A and the second is what we get from Box B. Each pair has chance $(1/5) \times (1/6) = 1/30$ to crop up. There are 5 such pairs. Hence the required chance is $5 \times 1/30 = 1/6$.

For one of the numbers to be strictly bigger than twice the other we have the following pairs:

$(1,3), (1,4), (1,5), (1,6), (2,5), (2,6), (3,1), (4,1), (5,1), (5,2).$

Thus we have 10 pairs each having chance $1/30$; hence the required probability is $10/30 = 1/3$.

