

## Solutions to Midterm 1

Multiple choice: 3 points/part; 27 points total. Circle the letter of the best answer.

MC1. Alice, Bob, and Carol each conduct a survey of University of Oregon students, trying to find out, for example, what they think of statistics, how often they read newspapers, how often they ride LTD busses, what they think of the football team, etc.

- Alice stands next to Deady Hall, and interviews the first 100 people who pass by.
- Bob runs an ad in the *Oregon Daily Emerald*. From the 257 people who respond, he selects a simple random sample of size 100 to interview.
- Carol obtains from the Registrar a list of all University of Oregon students, and numbers them. She interviews the the ones corresponding to the first 100 numbers in a list of random numbers in the appropriate range.

Which of the these people made an appropriate choice of sample?

- a. Only Bob and Carol.
- b. Only Alice and Carol.
- c. All of Alice, Bob, and Carol.
- d. Only Carol.
- e. Only Alice and Bob.
- f. Only Bob.
- g. None of the above is correct.

Alice chose a convenience sample. Her sample is likely to be particularly biased in its attitude toward statistics, since Deady Hall has many math classes. Bob chose a subsample of a voluntary response sample, and his sample is likely to be particularly biased about reading newspapers. Carol chose a simple random sample.

However, Carol probably got the registrar in trouble for violating privacy laws.

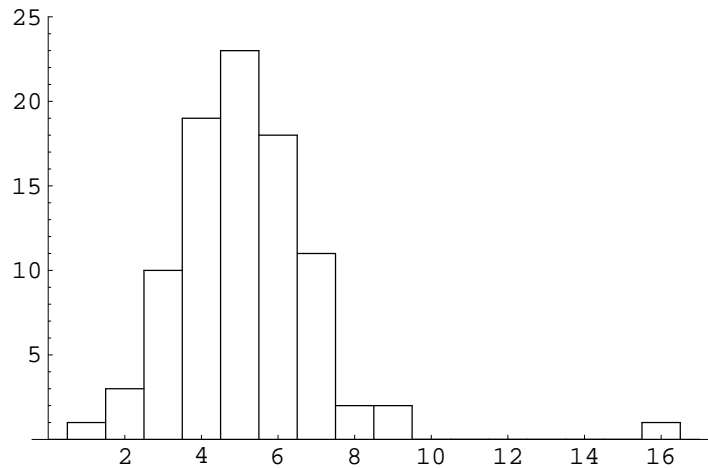
MC2. Spraying plants with a particular chemical is known to combat a particular plant disease. A research team is testing the effects of spraying differing amounts of the chemical on the recovery time of corn plants with this disease. The explanatory variable is:

- a. The chemical being tested.
- b. The amount of the chemical sprayed.
- c. The length of the recovery time.
- d. The kind of plant the tests are being done on.
- e. The disease the plants had.
- f. The research team.
- g. None of the above.

MC3. We choose a sample of 50 graduating seniors at the University of Oregon, ask for their grade point averages, and find that the mean of these 50 numbers is 2.47. Meanwhile, the registrar tells us that the mean of the grade point averages of all graduating seniors is 3.11. The number 2.47 is what:

- a. A parameter.
- b. A population.
- c. A statistic.
- d. A sample.
- e. Both (a) and (b).
- f. Both (c) and (d).
- g. None of the above.

MC4. Consider the following histogram (made from integer data):



The largest observed value is:

- a. 24.
- b. 23.
- c. 16.
- d. 9.
- e. 8.
- f. 1.
- g. None of the above.

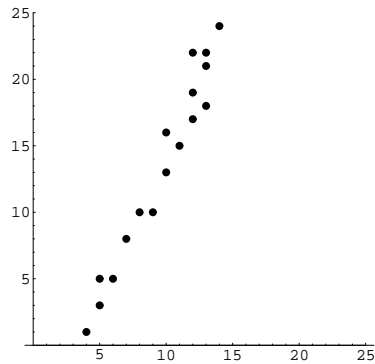
MC5. A researcher finds the age (in years) and height (in meters) of each member of a sample of Douglas fir trees, and statistically analyzes the data he gets. With age being the explanatory variable, units of the correlation of age and height are:

- a. Meters per year.
- b. Years per meter.
- c. Years.
- d. Meters.
- e. None: the correlation has no units.
- f. Douglas fir trees per year.
- g. Cannot be determined from the information given.

MC6. A college newspaper interviews a psychologist about a proposed system for rating teaching ability of faculty members. The psychologist says, “The evidence indicates that the correlation between a faculty member’s research productivity and teaching rating is close to zero.” Which of the following is the correct interpretation of this statement?

- a. Good researchers tend to be poor teachers and poor researchers tend to be good teachers.
- b. Good research and good teaching go hand in hand.
- c. Most or all of the faculty members are neither good researchers nor good teachers.
- d. Good researchers are just as likely to be good teachers as they are to be bad teachers, and likewise for poor researchers.
- e. The psychologist does not understand statistics.

MC7. Consider the following scatterplot:



The correlation is:

- a. Clearly less than  $-1$ .
- b. Close to  $-1$ .
- c. Clearly negative but not close to  $-1$ .
- d. Near zero.
- e. Clearly positive but not close to 1.
- f. Close to 1.
- g. Clearly greater than 1.
- h. Not defined.

MC8. A simple random sample of size  $n$  is drawn from a population with mean  $\mu$  and finite standard deviation  $\sigma$ . The Central Limit Theorem says that when  $n$  is sufficiently large:

- a. The distribution of the population is exactly normal.
- b. The distribution of the population is approximately normal.
- c. The distribution of the sample mean is approximately normal.
- d. The distribution of the sample mean is exactly normal.
- e. The limit of the sample size is the center of the population size.
- f. The center of the population distribution is limited.

MC9. A data set has mean 100 and standard deviation 20. The median:

- a. Must be 100.
- b. Must be 120.
- c. Must be 80.
- d. Must be 20.
- e. Cannot be determined from the information given.

Longer answer problems: follow instructions; point values as indicated.

---

1. (2 points/part; total 12 points.) (Work need not be shown.) Here is the five number summary of temperatures observed at the South Pole at 6:00 am on each day of September 1998:

-71   -44   -32   -18   -8

The questions below are about the original data, which consists of 30 numbers. For each of the following statements, circle “A” if the statement must be true, circle “S” if the statement might or might not be true, and circle “N” if the statement cannot be true. In other words, under the stated circumstances the statement is **A**lways true, **S**ometimes true, or **N**ever true.

A   S   N   (a) On at least one day, at 6:00 am the temperature was  $-8$ .

*Solution:* **A**lways true: this was the largest observed value.

A   S   N   (b) The value  $-71$  is an outlier.

*Solution:* **S**ometimes true. You can't tell if it is an outlier without looking at the original data. (The 1.5 IQR test only gives *suspected* outliers.)

A   S   N   (c) The median of the data is  $-32$ .

*Solution:* **A**lways true, by the definition of the five number summary.

A   S   N   (d) The standard deviation is  $-12$ .

*Solution:* **N**ever true: the standard deviation can never be negative.

A   S   N   (e) For about 23 days of September 1998 the temperature at 6:00 am was above  $-18$ .

*Solution:* **N**ever true:  $-18$  is the *third* quartile, so for about 23 days of September 1998 the temperature at 6:00 am was *below*  $-18$ , and for about 15 days it was even below  $-32$ .

A   S   N   (f) The mean of the data is  $-32$ .

*Solution:* **S**ometimes true. The mean might be close to the median, but might not be.

2. (4 points/part; total 12 points.) Wang's Widgets Inc. has ten employees. The mean salary is \$52,100, the standard deviation is \$24,049 (to the nearest dollar), and the five number summary is

\$25,800   \$37,000   \$43,200   \$61,000   \$102,000.

The owner has decided to give the highest paid person a \$200,000 raise. (He did something that earned the company millions in extra profits.)

In all parts of this problem, include appropriate units.

a. What is the new mean salary? Why?

Add  $\$200,000/10 = \$20,000$  (the increase in the total payroll divided by the total number of employees) to the old mean, to get \$72,100.

b. What is the new median salary? Why?

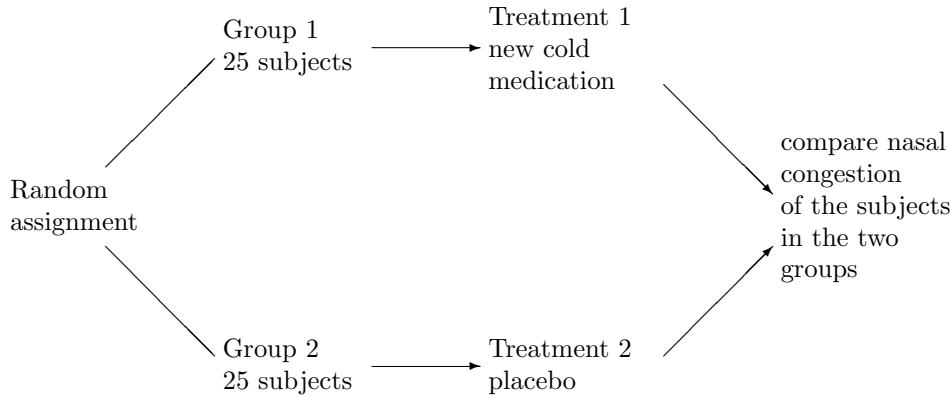
The median is unchanged at \$43,200. The median is half way between the two middle salaries. After the raise, those are still the two middle salaries, and they have not changed.

c. Suppose that instead the owner gives everyone a \$10,000 raise. What is the new standard deviation? Why?

The standard deviation is unchanged at \$24,049. The standard deviation is a measure of the spread of the distribution, and shifting it up by \$10,000 does not change the spread.

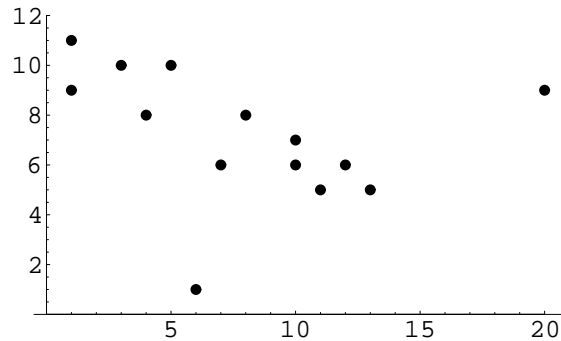
3. (15 points.) You want to test the effectiveness of a new cold medication in reducing nasal congestion. You have 50 test subjects available, each of whom is suffering from nasal congestion as a result of a cold. Use a diagram to outline in detail the design of a randomized comparative double blind experiment. Include information about the treatment groups and the response variable. Be sure that one can tell from your description that your experiment has all the characteristics expected of such experiments.

*Solution:*



The response variable is the degree of nasal congestion at the end of the experiment. The assignment to the treatment groups is made randomly. The second treatment is a placebo, something which superficially is indistinguishable from the new cold medication but which does nothing. Until after the experiment is over and the levels of nasal congestion are evaluated, neither the experimenters nor the subjects are to know who got the new cold medication and who got the placebo.

4. (4 points/part; total 8 points.) Consider the following scatterplot:

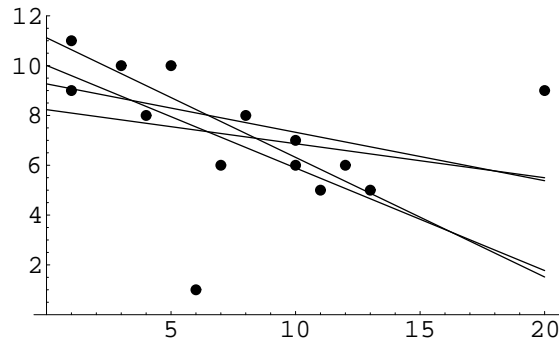


a. Explain clearly what an influential point for the regression line is.

*Solution:* A point is influential for the regression line if removing it would markedly change the regression line.

b. Identify all influential points for the regression line by giving their approximate coordinates.

*Solution:* The point at the upper right, with approximate coordinates (20, 9), is influential for the regression line. (Note that this is *not* the same as (9, 20). The answer (9, 20) will get no credit.) The other outlier, at approximately (6, 1), is not very influential for the regression line. The following graph shows four regression lines, obtained by omitting neither, one, or both the outliers.



5. (4 points.) In a study of the heights of fathers and adult sons, the correlation was found to be  $r = 0.60$ . Mention at least one lurking variable that needs to be taken into account if one is to better understand the relationship between fathers' and sons' heights.

*Solution:* The most obvious is the mother's height.

6. (4 points/part; total 12 points.) The heights of fourth graders in Seattle are approximately normally distributed with mean 46.5 inches and standard deviation 4.8 inches.

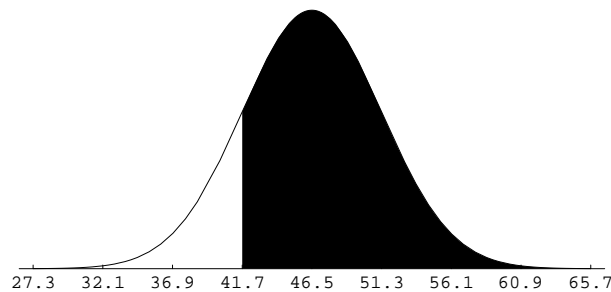
In each of the following parts, show your work, and draw a picture of the appropriate normal curve with the relevant points on the horizontal axis clearly marked and with the appropriate area shaded and clearly identified. The curve must look reasonable.

a. Approximately what percentage of fourth graders is taller than 41.7 inches?

*Solution:* The relevant area is the shaded area in the picture; the horizontal labels are at the mean plus or minus integer multiples of the standard deviation.

This one can be done by the rule of thumb: 41.7 inches is one standard deviation below the mean. About 68% of the data is within one standard deviation of the mean, so about 32% of the data is more than one standard deviation away from the mean. Of this, half will be more than one standard deviation below the mean, so that percentage is about 16%. The rest, or about 84%, will be taller than 41.7 inches.

Alternatively, looking up  $-1.0$  in Table A gives 0.1587, so the percentage is  $100\% - 15.87\% = 84.13\%$ , correct to four significant digits.



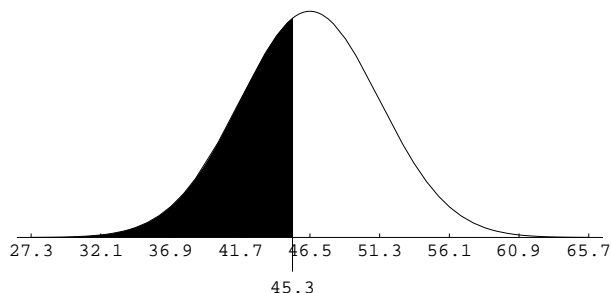
- b. Approximately 40% of the fourth graders are shorter than \_\_\_\_\_ inches.

*Solution:* This solution uses Table A. The closest number in Table A to 0.40 is 0.4013, which occurs for  $z = -0.25$ . The corresponding value of  $x$  is given by

$$x = \sigma z + \mu = (4.8)(-0.25) + 46.5 = 45.3.$$

Accordingly, 40% of the fourth graders are shorter than 45.3 inches.

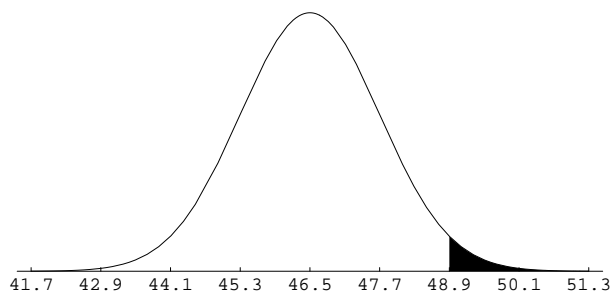
The relevant area is the shaded area in the picture; the horizontal labels are at the mean plus or minus integer multiples of the standard deviation.



- c. Approximately what percentage of simple random samples of size 16 of fourth graders has sample mean more than 48.9 inches?

*Solution:* The sample means of samples of size 16 are normally distributed with mean 46.5 inches (the same as for the original distribution) and sampling standard deviation 1.2 inches (the original standard deviation divided by  $\sqrt{16} = 4$ , the square root of the sample size).

The relevant area is the shaded area in the picture; the horizontal labels are at the mean plus or minus integer multiples of the sampling standard deviation.



This one can be done by the rule of thumb: 48.9 inches is two sampling standard deviations above the mean. About 95% of samples have sample mean within two sampling standard deviations of the mean, so about 5% have sample mean more than two sampling standard deviations away from the mean. Of this, half, or about 2.5%, will be more than two sampling standard deviations above the mean.

Alternatively, looking up 2.0 in Table A gives 0.9772, so the percentage is  $100\% - 97.72\% = 2.28\%$ , correct to three significant digits.

7. (3 points/part; total 6 points.) If you draw an M&M candy at random from a particular bag of M&M candies, the candy you draw will have one of six colors. The table below gives the probability that a randomly chosen M&M has each color.

| Color       | Brown | Yellow | Red | Green | Orange | Blue |
|-------------|-------|--------|-----|-------|--------|------|
| Probability | 0.2   | 0.3    | 0.1 | 0.1   | ?      | 0.2  |

Remember to **show your work** (even if you can do the problem in your head).

- a. Find the probability of drawing an orange M&M.

*Solution:*  $1 - (0.2 + 0.3 + 0.1 + 0.1 + 0.2) = 0.1$

- b. Find the probability of drawing a yellow or green M&M.

*Solution:* The events are disjoint, so it is  $0.3 + 0.1 = 0.4$ .

8. (4 points) John Doe's score on Midterm 1 was 96, but he missed Midterm 2 due to illness. Predict what his score would have been, given the following information on the scores of all students who took both Midterm 1 and Midterm 2.

Midterm 1: Five number summary 22 51 77 82 99; mean 69; standard deviation 23.63.

Midterm 2: Five number summary 15 42 71 84 98; mean 65; standard deviation 26.80.

Correlation  $r \approx 0.8570$ ;  $r^2 \approx 0.7345$ .

(All noninteger values given to 4 significant digits.)

*Solution:* The regression line has the formula  $\hat{y} = a + bx$  with  $b = rs_y/s_x$  and  $a = \bar{y} - b\bar{x}$ . Here  $x$  is the score on Midterm 1 and  $y$  is the score on Midterm 2. The numbers  $s_x$  and  $s_y$  are the standard deviations of these scores. Substituting the numbers, we get  $b \approx 0.97197$  and  $a \approx -2.0658$ , so the predicted score is about  $-2.0658 + (0.97197)(96) = 91.2432$ .

Here is another, more conceptual, way to do it. John Doe's  $z$ -score on Midterm 1 is  $(96 - 69)/23.63 \approx 1.1426$ . His predicted  $z$ -score on Midterm 2 is obtained by multiplying this by  $r$ , giving about 0.97922. This corresponds to an actual score of  $65 + (0.97922)(26.80) \approx 91.2431$ . (The slight discrepancy with the previous method is due to rounding errors.)