

# EXAM#1

## Part I. Multiple choice (5 pts each)

1. The proportion of observations from a standard normal distribution that satisfy  $-0.86 < Z < 0.40$  is
- A. .3211
  - B. .4605
  - C. .5395
  - D. .6789
  - E. .8300

For problems 2 and 3, suppose that SAT math scores follow a normal distribution with  $\mu = 500$  and standard deviation  $\sigma = 80$ .

2. What proportion of scores are above 540?
- A. .3085
  - B. .4013
  - C. .5000
  - D. .5987
  - E. .6915
3. What score on the SAT math would you need in order to be at the 98th percentile of the population? (a score better than the score of 98% of the population)
- A. 567
  - B. 576
  - C. 589
  - D. 631
  - E. 664

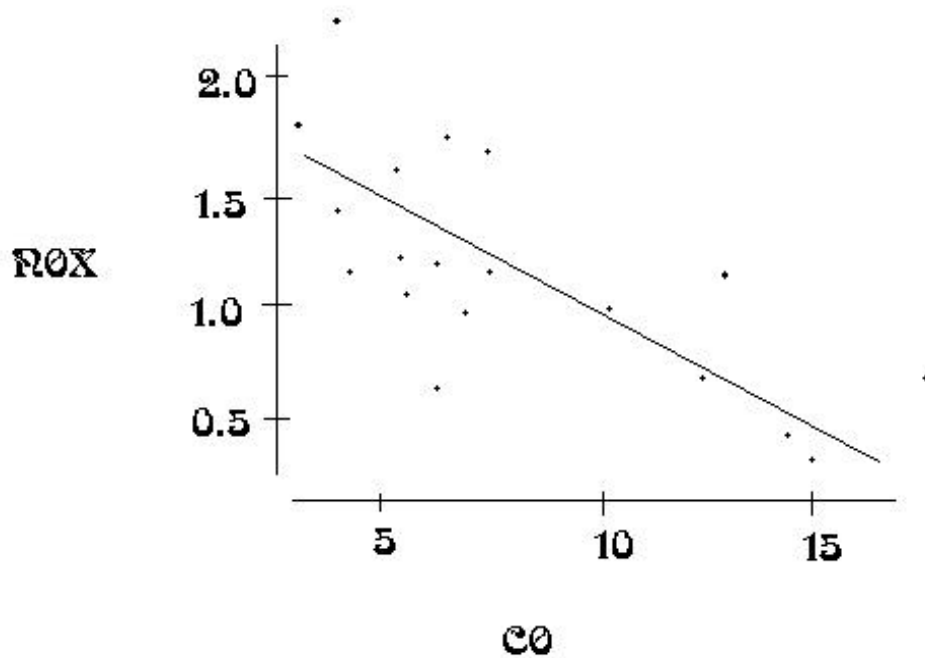
\*\*\*\*\*

4. At an elementary school there is a positive correlation between shoe size and the reading comprehension. To this correlation the variable age (or grade level) would be
- A. an explanatory variable
  - B. a response variable
  - C. a lurking variable
  - D. a coincidence variable

5. Which of the following best describes a histogram?
- A. a graph of the distribution of a quantitative variable and a categorical variable with the quantitative variable on the vertical axis and the categorical variable on the horizontal axis.
  - B. a graph of the distribution of a single quantitative variable with that variable on the vertical axis and counts on the horizontal axis.
  - C. a graph of the distribution of a single categorical variable with that variable on the vertical axis and counts on the horizontal axis.
  - D. a graph of the distribution of a single quantitative variable with that variable on the horizontal axis and counts on the vertical axis.
  - E. a graph of the distribution of a single categorical variable with that variable on the horizontal axis and counts on the vertical axis.
6. Weight and height would typically have a positive correlation. Suppose we have computed two correlations:  $r_1$  is the correlation in height and weight among 3000 individuals; and  $r_2$  is the correlation between average weight at a given height and height (e.g. we'd average the weights of all individuals 5'6" tall.)
- A.  $r_1$  will be large than  $r_2$
  - B.  $r_2$  will be large than  $r_1$
  - C.  $r_1$  and  $r_2$  will be the same
  - D.  $r_2$  must be zero
  - E.  $r_2$  can not be calculated
7. To data on the heights of 20 typical college students, we erroneously add the data on Sam, the giraffe, who is 20 feet tall. Which of the following numerical statistics will be least affected by our error?
- A. least squares regression line with height =  $x$  and weight =  $y$
  - B. mean
  - C. range
  - D. standard deviation
  - E. median
8. A study is conducted to determine if one can predict the yield of a crop based on the amount of yearly rainfall. What is the explanatory variable in this study?
- A. yield of a crop
  - B. amount of yearly rainfall
  - C. the experimenter
  - D. either bushels of inches of water

9. Which of the following statements is correct?
- A. Faculty who are good researchers tend to be poor teachers and vice-versa, so the correlation between teaching and research is 0.
  - B. Women tend to be, on average, about 3.5 inches shorter than the men they marry, so the correlation between heights of spouses must be negative.
  - C. A research finds the correlation between the shoe size of children and their scores on a reading test to be 0.22. The researcher must have made a mistake since these two variables are clearly unrelated and must have correlation 0.
  - D. If people with large heads tend to be more intelligent, then we would expect the correlation between head size and intelligence to be positive.

10. Consider the following scatterplot of amounts of CO (carbon monoxide) and NOX (nitrogen oxide) in grams per mile driven, in the exhausts of cars. The least squares regression line has been drawn in the plot.



In the above scatterplot, what is the approximate correlation between variables?

- a.  $-1.00$
- b.  $-0.73$
- c.  $0.49$
- d.  $0.00$

11. If removing an observation from data set would have a marked change on the position of the least squares regression line fit to the data, what is the point called?
- robust
  - a residual
  - influential
  - a response
12. A study of the salaries of full professors at Upper Wabash Tech showed that the median salary for female professors was considerably low than the median male salary. Further investigation showed that the median salaries for male and female full professors were about the same in every department (English, Physics, etc.) of the university. This apparent contradiction, i.e., equal salaries in every department can still result in a higher overall median salary for men, is known as what?
- the inflation rate
  - least squares regression
  - Simpson's paradox
  - the residual effect

## Part II. Problems and Discussion. Be sure to show all necessary work.

1. (16 pts) For the data: 12, 14, 24, 40, 50
- Find the mean.
  - Find the standard deviation  $s$  using the algorithm developed in class.
2. (9 pts) A study of the size of jury awards in civil cases (such as injury, product liability, and medical malpractice) in Chicago showed that the median award was about \$8,000. But the mean award was about \$69,000. Explain how this great difference between two measures of center can occur?
3. (40 pts) Below are the numbers of CD's owned by a selected group of 18 statistics students at an eastern university.

60 29 250 8 150 100 10 30 55 180 90 0 260 60 120 12 118 50

For this data, find the

- median.

- b. first quartile, third quartile, and IQR.
- c. stemplot for the data: use hundreds as the stem and use split stems.
- d. boxplot for the data.
- e. Suppose that we wanted to make a graphical comparison among student CD ownership in the above university, Bradley, Florida State, Pacific University, and Oral Roberts University. What would be an appropriate plot to use for this comparison?
- f. Would you describe this distribution as symmetric, left skewed, or right skewed? Why?

4. (44 pts) Last spring statistician S. Berry used past data and a statistical model to predict the number of home runs that would be hit by baseball's leading home run hitters during 1999. I have included the four highest model values and a random sample of other high values. Then I list the actual season number of home runs through last Sunday's games.

player	model	season
McGwire	59	60
Griffey	51	48
Sosa	50	61
Gonzalez	44	37
Belle	39	36
G. Vaughn	37	42
Ramirez	35	42
Rodriquez	34	41
Walker	33	37
Thomas	31	15

- a. Draw a scatterplot for this data by using model value as explanatory variable and season total as response variable.
- b. This data has a least squares regression line  $y = -4.0 + 1.11x$ . Draw a graph of this line with your scatterplot above.
- c. Is residual for Thomas positive or negative? Explain.
- d. Predict the season total for a hitter whose model value was 40 home runs.

- e. The correlation between  $x$  and  $y$  is  $r = 0.793$ . What fraction of the variation in actual season home runs ( $y$ ) is explained by the least squares regression line on model home runs ( $x$ )?

5. (16 pts) The US Census collects data on many variables as part of its mandate to enumerate the population. The table below gives the number of individuals (age 25 or older) in certain Illinois counties who have attained each level of education. For simplicity, the numbers are hundreds of individuals.

	County				
	Peoria	Tazewl	DuPage	Jeffer	TOTAL
0 to 8th grade	105	71	210	38	424
9 to 12 grade	151	101	364	35	651
HS graduate	364	297	1172	83	1916
College, no deg	240	173	1129	40	1582
Assoc/Bach deg	223	129	1547	35	1934
Grad/Prof deg	76	31	601	9	717
<b>TOTAL</b>	<b>1159</b>	<b>802</b>	<b>5023</b>	<b>240</b>	<b>7224</b>

- a. Give the marginal distribution (in percents) of educational attainment.
- b. Give the conditional distribution (in percents) of educational attainment for Jefferson county.
6. (15 pts) Standard scores ( $z$ -scores) can be used to compare outstanding records from different eras. Compare standard scores of the following famous home run records. Who do the scores say is the best?

HR = number of home runs by the famous player

AV = mean number of home runs per position player per full season

SD = standard deviation of number of home runs per position player per full season

	HR	AV	SD
Ruth, 1927	60	6	10
Maris, 1961	61	15	13
McGwire, 1998	70	19	12