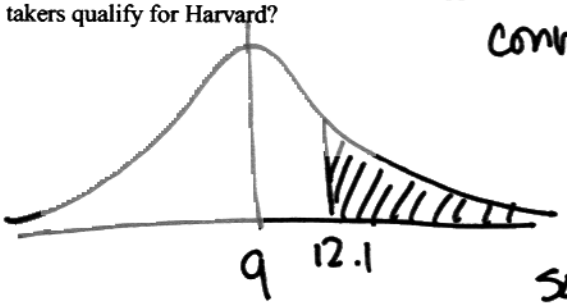
Statistics 10 Exam 1 Review Exam is 10/22/03

1. Indicate whether the following statements are true or false

	T	F	Statement
A.	X		Randomization is necessary to prevent selection bias in experiments
B.	X		Control Groups are necessary in experiments so we can compare the results from the treatment group properly
C.		X	~~Groups in experiments are always given a placebo~~
	X		Observational studies can establish association
	X		Observational studies are misleading due to confounding
		X	Selection Bias in an experiment is due to confounding

The next four questions use information from this statement, but each question is separate (i.e. you can get the first one wrong and it won't affect the others): The Medical College Admissions Test (MCAT) is constructed to be normally distributed with a mean of 9 and a standard deviation of 2. Approximately 20,000 people take the test every year. SHOW YOUR WORK FOR FULL CREDIT.

2. Harvard Medical School only considers applicants with a test score of 12.1 or greater. How many of the test takers qualify for Harvard?



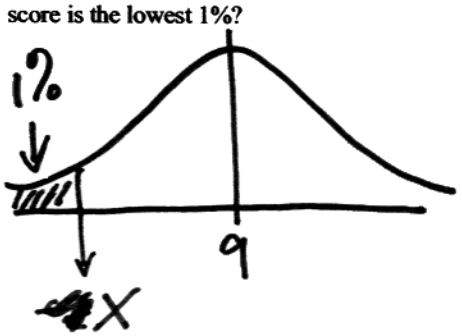
convert 12.1 to z

$z = \frac{12.1-9}{2} = +1.55$

area to left of 1.55 is .9394

so $1.0 - .9394 =$ .0606 or 6.06%

3. The lowest 1% of all test takers can enroll in Dr. Nick Riviera's School of Medicine. At and below what score is the lowest 1%?

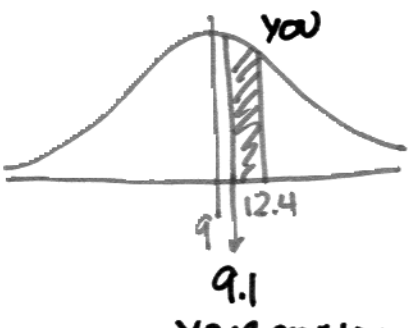


lowest 1% implies $z = -2.33$

$-2.33 = \frac{x-9}{2}$ solve for X

gives 4.34

4. You decided to take the MCAT and got an 12.4. Your cousin, who went to USC, also took the MCAT got a 9.1. What percentage of test takers have scores between yours and your cousin's?



① YOU $\frac{12.4-9}{2} = +1.70$ area is .9554

② COUSIN $\frac{9.1-9}{2} = +.05$ area is .5199

subtract areas

.9554 − .5199 = .4355

or 43.55

5. After thinking it over, you decide not to apply to Medical School, but apply to Law School instead. And to your surprise, the UCLA Law School is willing to consider applicants with a valid MCAT score – with the following condition: All applicants must add 23 to their MCAT Score first and then multiply that score by 5. So for example, you got a 12.4, your new score is 177.

If you apply UCLA rules to ALL the MCAT scores, what are the new mean, median and standard deviation?

New Mean = (23+9)*5 = 160

New Median = (23+9)*5 = 160

New SD = 2×5 = 10

mean = median for normal dist.

The next three questions refer to the list {-8, -5, -3, 0,1, 3, 0, 4}.

6. The mean is:
   (a) –1.0
   (b) 0
   (c) 0.5
   (d) 1.0
   (e) none of the above.

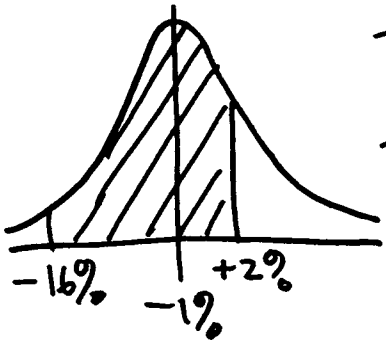
7. The median is:
   (a) –1.0
   (b) 0
   (c) 0.5
   (d) 1.0
   (e) none of the above.

8. The inter-quartile range is:
   (a) -4
   (b) 4
   (c) -6
   (d) 6
   (e) none of the above.

9. The SD can never be larger than the mean:
   (a) True
   (b) False

The next three questions refer to this statement, but each question is separate (i.e. you can get the first one wrong and that won't affect the others): Corporate securities (or publicly traded stocks) are an investment opportunity for individuals as well as institutions. The 10,000 stocks available for investment to U.S. residents are normally distributed with a mean one-year return of -1% (this means you lost 1% of the value of your investment) and a standard deviation of 12%. SHOW YOUR WORK FOR FULL CREDIT.

10. What percentage of stocks had one-year returns between -16% and +2%? (5 points)

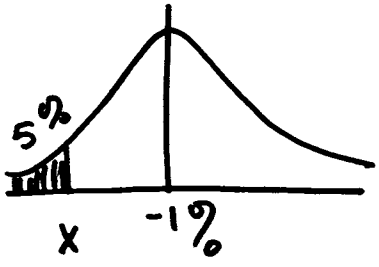


for +2 $z = \frac{2-(-1)}{12} = 0.25$ area is .5987

for -16 $z = \frac{-16-(-1)}{12} = -1.25$ area is .1056

difference is .5987 - .1056 = .4931 or 49.31%

11. A stock is at 5th percentile (i.e. 5% of the stocks have returns equal to or lower than this stock), what is its one-year return? (5 points)

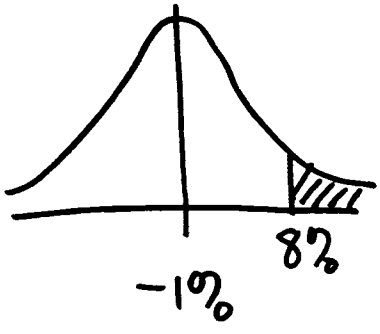


Use $z = -1.64$ or -1.65 or -1.645 and solve for X, for example

$-1.645 = \frac{X-(-1)}{12}$

X = -20.74 or -20.74% or -20.8%

12. In order to meet your retirement goals, you need to buy stocks that have a return of 8% or more. Approximately how many stocks out of the 10,000 qualify? (5 points)



$z = \frac{8-(-1)}{12} = +.75$

area to the left is .7734

so 1 - .7734 = .2266 to the right or 22.66%

22.66% of 10,000 is 2,266 stocks

13. A professor constructed a sample survey to estimate the percentage of USC undergraduates living at home. Two assistants were stationed at the "Tommy" Trojan statue (it's on the main plaza) and were instructed to interview all students who passed by at specified times. Many students would not speak with the assistants, in fact, only 369 out of 1500 approached, did. As it turned out, 39% of 369 students interviewed said they live at home with their parents the others live elsewhere. Does the investigator's procedure give an accurate estimate of the percentage of USC students who live with their parents?

First, answer yes or no and then explain your reasons for your choice. This does not need to be a long answer.

NO 1) Selection Bias - not everyone had the same opportunity to be interviewed

2) Non-Response Bias - only 369/1500 respondent for less then 50%

14. Classify the following variables as categorical, ordinal or numerical by checking the correct box, if it is a quantitative variable, further classify the variable as either discrete or continuous:

	Variable	Categorical	Ordinal	Quantitative	Discrete	Continuous
A	Hair Color	✓				
B	Frozen Food Brand	✓				
C	Number of students in a classroom			✓	✓	
D	Your age			✓		✓

15. To study the effects of exercise on the grades of college students, a researcher wishes to compare the grade point averages of students at randomly selected colleges across the United States. The researcher selects students at random and after interviewing them to find out who exercises and who does not, chose 644 students of each (exercisers and non-exercisers). The researcher made sure the two groups of 644 were similar in racial composition, gender, major, and every subject had accumulated at least 120 units towards graduation. There were a total of 1,288 students in the study from approximately 40 colleges, their overall GPA was 3.22. The average GPA for the students who exercised was 3.34 and the standard deviation was .36.

a. What is the "treatment"?

EXERCISE

b. What is the response or outcome variable?

GPA

c. Is this an observational study or an experiment?

OBSERVATIONAL

d. From this study, an example of a sample statistic is:

3.34, 3.22 (GPAs) .36 (SD)

e. What is the population of interest?

ALL COLLEGE STUDENTS

f. What is the population parameter of interest in this study?

difference in GPA between exercisers and non-exercisers

16. The next questions refer to the list {-4, -9, 0, -3}.

a. What is the mean of this list?

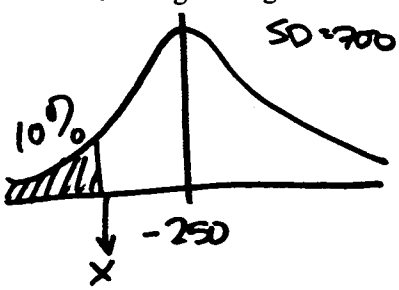
$$\frac{-4 + -9 + 0 + -3}{4} = -4$$

b. What is the standard deviation of this list?

$$s = \sqrt{\frac{(-4--4)^2 + (-9--4)^2 + (0--4)^2 + (-3--4)^2}{4}} \approx 3.24$$

The next three questions use information from this statement, but each question is separate (i.e. you can get the first one wrong and it won't affect the others): A recent study showed that the gambling income of adults age 21 and over in the United States from all forms of legalized gambling (e.g. lottery, video poker, horse racing, casinos) is normally distributed with a mean of -250 dollars (a loss) and a standard deviation of $700. SHOW YOUR WORK FOR FULL CREDIT.

17. It is believed that the gamblers with the largest losses, that is those with the lowest 10% of gambling income, should be considered gambling "addicts" and given some kind of treatment. How much money does a gambling adult need to lose to be considered an "addict"?

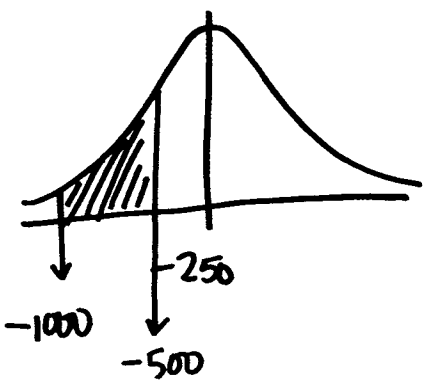


Choose $z = -1.28$ closest to 10% in left "tail"

$$-1.28 = \frac{X - (-250)}{700}$$ solve for X

$\$ -1,146$

18. What percentage of adults age 21 and over had gambling losses of at least $500 but not more than $1000?



$$z_{-500} = \frac{-500 - (-250)}{700} = -.36$$ area is .3594

$$z_{-1000} = \frac{-1000 - (-250)}{700} = -1.07$$ area is .1423

area between -500 and -1000 is $.3594 - .1423 = .2171$

or 21.71%

19. What is the median gambling income?

a loss of 250 or -250

The Super Bowl is the number one party event of the year for Americans, exceeding even New Year's Eve celebrations. Suppose it is known that the typical party has 17 partygoers on average with a standard deviation of 3.3.On a typical Sunday afternoon, the average number of calories consumed in America is 600. Please assume that calories are normally distributed

The Hardvard School of Public Health decided to study the effects of attending Super Bowl Sunday parties on the caloric consumption of Americans. 850 Americans were selected by random-digit dialing and interviewed by telephone. 490 Americans reported that they had attended a Super Bowl party, 110 did not attend a party but watched the Super Bowl on television at home. The remainder did not attend a Super Bowl party or watch the game. The calories consumed by the partygoers had a mean 1,330 with a standard deviation of 600. The calories consumed by the non-party goers had a mean of 560 with a standard deviation of 100. Among the partygoers, 77% reported getting "drunk", only 5% of the non-party goers reported getting "drunk". The average party had 19 partygoers.

20A. (2 points) The population of interest to the Hardvard School of Public Health is

(a) all Super Bowl Partygoers
(b) all Americans
(c) all Americans who watched the Super Bowl on television
(d) 850 Americans
(e) 490 Americans who reported that they had attended a Super Bowl Party
(f) 110 Americans who had not reported attending a Super Bowl Party

20B. (2 points) The parameter of greatest interest to the Hardvard School of Public Health is

(a) 17 partygoers
(b) 3.3 partygoers
(c) 600 calories
(d) The percentage who report getting "drunk"
(e) The average number of calories consumed by Americans
(f) The average number of calories consumed by party goers on Super Bowl Sunday

20C. (2 points) The sample of interest to the Hardvard School of Public Health is

(a) all Super Bowl Partygoers
(b) all Americans
(c) all Americans who watched the Super Bowl on television
(d) 850 Americans
(e) 490 Americans who reported that they had attended a Super Bowl Party
(f) 110 Americans who had not reported attending a Super Bowl Party

20D. (2 points) The statistic of greatest interest to the Hardvard School of Public Health is

(a) 19 partygoers
(b) 1330 calories
(c) 560 calories
(d) 850 Americans
(e) 77% got drunk
(f) 5% got drunk

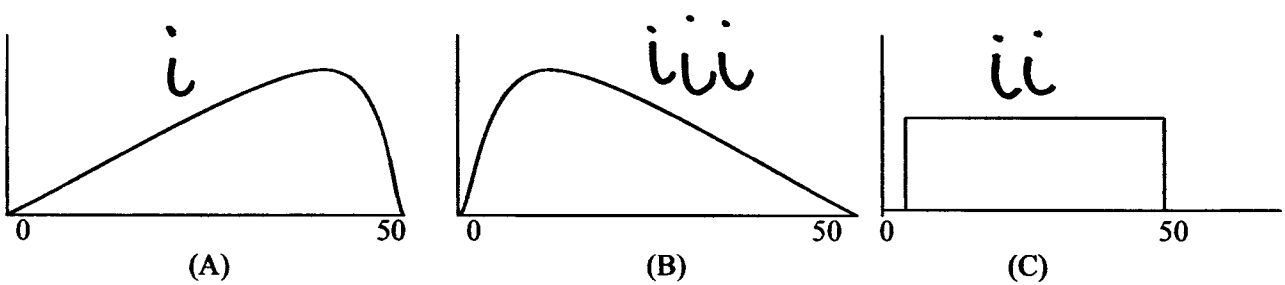
20E. (2 points) This is an example of a

(a) an Observational Study with historical controls
(b) an Observational Study utilizing multi-stage cluster sampling techniques
(c) an Observational Study that uses a random probability method for sample selection.
(d) a Randomized Experiment without Controls, but it is blind
(e) a Randomized Experiment without Controls, but it is double-blind
(f) a Randomized Controlled Experiment

21. Please indicate whether each statement is true or false **(one point each)**

	True	False	Statement
A	X		The total area of a histogram is always 100% when area is expressed as percentages
B	X		Larger samples are no better than smaller samples at preventing bias
C	X		A histogram is a graphical summary which represents percentages as areas
E	X		The area under the histogram between two values is equal to the percentage of cases in a class interval defined by those values
F	X		In a randomized controlled experiment utilizing a placebo, if the control group is comparable to the treatment group, then the difference in the responses of the two groups is likely to be a result of the treatment
G		X	Double blind experiments are better at preventing the placebo effect than single blind experiments
H		X	Confounding is not a source of bias
I		X	Random selection (or randomizing/randomization) is employed in sample designs because it is impartial however it does not minimize bias
J		X	If a large number of persons selected for a sample do not respond, problems of response bias are likely
K	X		A histogram with one peak is said to be unimodal, a histogram with two peaks is said to be bimodal

22. Here are three histograms, assume they have been correctly drawn:



Match <u>each</u> histogram above to the best choice listed below: **(2 points each, 6 points total)**

i. The average is smaller than the median.
ii. The average is equal to the median
iii. The average is larger than the median
iv. Cannot determine the average for this graphic
v. Cannot determine the median for this graphic

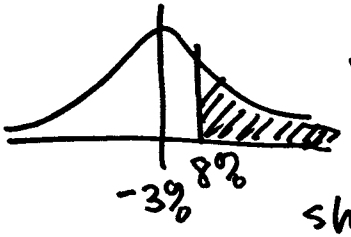
23. In an observational study (choose one) **(2 points)**

a. Investigators do not assign subjects to treatment or to control groups
b. There isn't a control group
c. Investigators can establish association but not causation
d. Confounding factors cannot be controlled
e. All of the above are true
f. Only A and C are true
g. Only D is false

24. The next 2 questions refer to this statement, but each question is separate (i.e. you can get the first one wrong and its result will not affect the others): You are on the verge of investing some of your hard-earned money in the stock market and you are examining two funds, let's call them A and B. Your investment adviser, I'll call him The Oracle, gives you some information on their performance (as measured by percentage returns over many, many days). Fund A has mean return of -3% with a standard deviation of 15%. It had a minimum of –78% and a maximum of 72%. Fund B has a mean return of 5% with a standard deviation of 2%. It had a minimum of –5% and a maximum of 15%. Assume both funds are normally distributed. SHOW YOUR WORK FOR FULL CREDIT.

A. You need to invest in a fund that spends as much time as possible giving returns in excess of 8%. Which fund is more likely to do this **(7 points)**

A => mean = -3%
SD = 15%

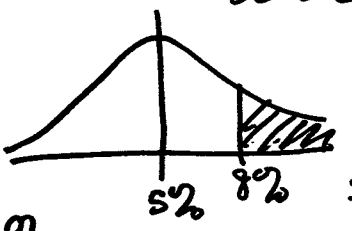


$\frac{8-(-3)}{15} = .73$

shaded area .2327 23.27%

B => mean = 5%
SD = 2%

$z = \frac{8-5}{2} = 1.5$



shaded area = .0668 or 6.68%

FUND A is more likely to be above 8% than B

B. Oracle says you know what? You need to take inflation into account in all of your calculations. So subtract 4% (professor: just subtract 4, don't worry about the percentage sign) from all of the returns and then multiply by 2. So for example, on a given day, Fund A returned –11%, so following Oracle's instructions subtracting 4 yields –15% and multiplying by 2 yields –30%. If you do this, what are the new mean, median, and standard deviations for funds A and B? **(6 points)**

FUND A → new mean = -14 new SD = 30
FUND B → " = 2 new SD = 4

work
mean A $(-3-4)*2 = -14$
SD A $(15)*2 = 30$

mean B $(5-4)*2 = 2$
SD B $(2)*2 = 4$

25. The computer output for analysis variable OHNO is being shown on a screen in a big lecture hall.

```
. summarize ohno, detail
                            ohno
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -10.65347      -15.3681
 5%    -9.295065     -15.03784
10%     -8.56562     -14.96494       Obs               1000000
25%    -7.346793     -14.91875       Sum of Wgt.       1000000

50%    -5.999892                     Mean            -5.999883
                        Largest      Std. Dev.         2.00111
75%    -4.650739      2.982908
90%    -3.436283      3.037329       Variance        4.004441
95%    -2.707139      3.061257       Skewness         .000419
99%    -1.351807      3.272228       Kurtosis        3.000024
```

Please answer the following questions based on the Stata results for variable OHNO. You may round the numbers given above to two decimal places. For example, -8.56562 can be rounded to –8.57

The professor says "The Stata results suggest that variable OHNO is normally distributed"

A) Using the Stata results above, determine how the professor came to this conclusion. Identify at least 2 conditions. (6 points)

(1) mean = median

(2) distance 50th to 25th + 75th to 50th are nearly equal ≈1.35

(3) 95%, 99.7% appear to be 2, 3 SDs away from mean

B) Does this variable have outliers (circle one): (YES) NO (1 point)

Using information from the Stata results, show how you arrived at your choice. (3 points)

$IQR = (-4.650739 - (-7.346793)) = 2.6961$

$75^{th} + (1.5 \times IQR) = -4.650739 + (1.5 \times 2.6961) = -.6067$

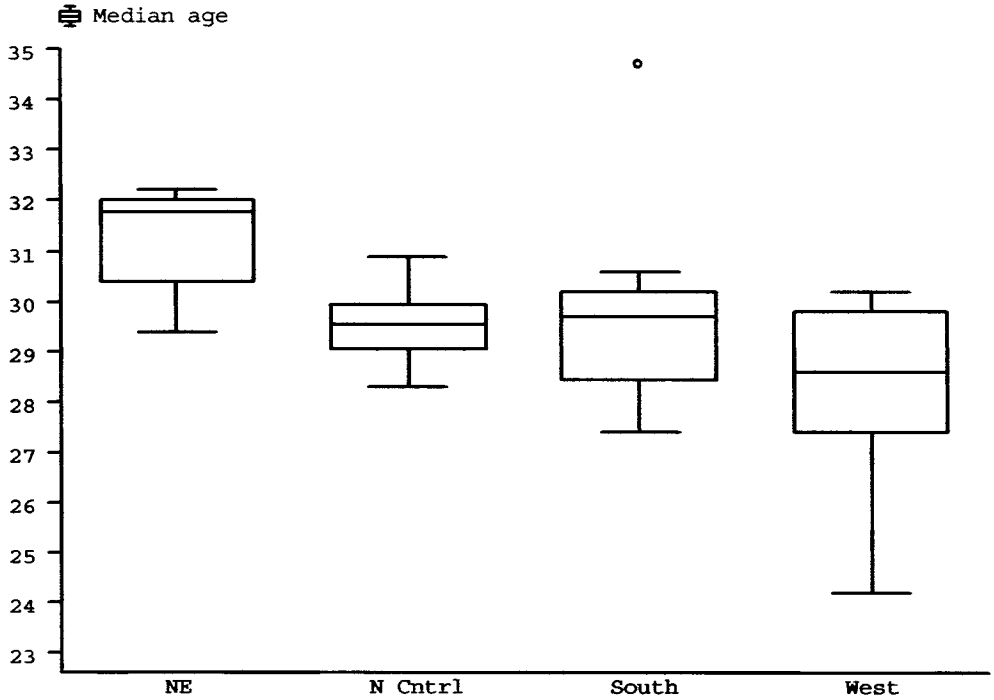
$25^{th} - (1.5 \times IQR) = -7.346793 - (1.5 \times 2.8961) = -11.3909$

there are values exceeding both of these

C) Please calculate the interquartile-range for variable OHNO (2 points)

see above

26. A pair of Geographers are studying the distribution of age for each of the 50 states in the U.S. They used a box plot to summarize the medians for each state by geographic region. Here is a box plot of the results:



Using the box plot, please answer the following questions:

a) Is there enough information present to estimate the inter-quartile range (IQR) for the North Central (N Cntrl) region? (circle one)

(YES) NO

If you answered "YES" please give an estimate of that value in the space below, if you answered "NO" please explain why it is not possible to estimate the IQR using a box plot. (3 points total)

IQR = value of 75th − value of 25th
30 − 29 = 1

b) Which region appears to be the most symmetrical in age? (circle one) (2 points)

NE (N Cntrl) South West Not enough information

c) Which region has the most skewness in age? (circle one) (2 points)

NE N Cntrl (South) West Not enough information

d) Which region has the most "left skewness" in age? (circle one) (2 points)

NE N Cntrl South (West) Not enough information

e) Is there enough information in the box plot to estimate the mean for the West region? (circle one)

YES (NO)

If you answered "YES" please give an estimate of that value in the space below, if you answered "NO" please explain why not. (3 points total)

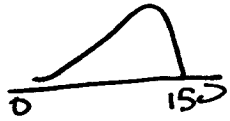
If we could establish normality the mean = median but this isn't normal so one can't accurately estimate the mean

27. In a hypothetical experiment, a new drug was compared with "standard therapy" treatment for patients suffering from inoperable cancer. These types of patients volunteered for the experiment and were randomized into treatment and control groups. The difference in survival time (in months) was selected as the response variable. Which of the following best describes the primary reason to randomize patients into treatment or control groups? (Choose the one best answer) (2pts)

(a) to prevent the bias introduced when the patients know what type of treatment they are receiving
(b) to prevent the placebo effect from confounding the results of the experiment
(c) to create "double blinding" when neither the investigators nor the patients know what type of treatment the patients are receiving
(d) to create two groups that are similar at the start of the experiment on both known and unknown factors associated with survival time.
(e) to eliminate the selection bias resulting from the fact that all of the patients had inoperable cancer
(f) all of the above

28. Suppose an examination is very easy and all but a few students in a class received very high scores (total possible points was 150). Which statement below most correctly describes the relationship between the mean and median? (2 pts)

(a) The mean is lower than the median for this examination
(b) The mean and the median are approximately equal
(c) The median will be exactly one-half the value of the mean
(d) The median is lower than mean for this examination
(e) There is not enough information to describe the relationship between the mean and median
(f) None of the above



29. Suppose an examination is very difficult and all but a few students in a class received very low scores (total possible points was 150). If these scores were graphed, the distribution of these scores would be described as: (2 pts)

(a) Normal
(b) Symmetric or symmetrical
(c) Left skewed
(d) Right skewed
(e) Uniform
(f) None of the above

30. A consulting is firm is hired to count the number of passengers on buses in the city of Los Angeles. The first bus measured had 25 passengers, the second bus had 35 and the third had 55 passengers. The difference in the number of passengers between the first and second bus is equivalent to half the difference between the second and third bus. This is an example of a: (2 pts)

(a) Quantitative variable with discrete values
(b) Quantitative variable with continuous values
(c) Quantitative variable with descriptive values
(d) Qualitative variable with descriptive values
(e) Qualitative variable with ordered values
(f) None of the above