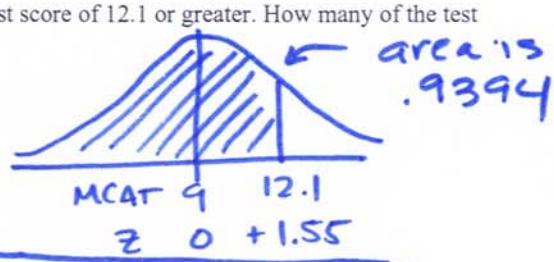
Statistics 10 Exam 1 Review Exam is 10/25/04

1. 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.

(a) 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 a Z score

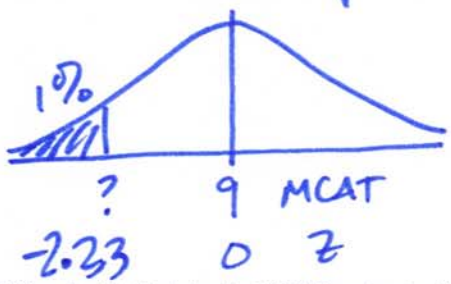
$Z = \frac{12.1 - 9}{2} = +1.55$



so 1.0 − .9394 = .0606 or 6.1%

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

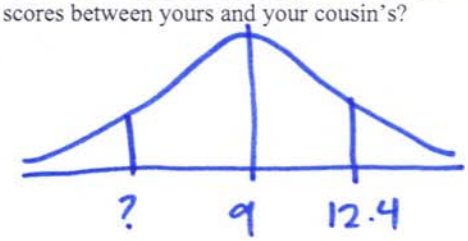
lowest 1% implies $z = -2.33$



$-2.33 = \frac{Y - 9}{2}$

$Y \approx +4.34$

(c) You decided to take the MCAT and got a 12.4. Your cousin, who went to USC, wouldn't tell you his score, but also took the MCAT and told you he is score was 2.2 Z scores below yours. What percentage of test takers have scores between yours and your cousin's?



Your $z = \frac{12.4 - 9}{2} = +1.7$

$1.7 - 2.2 = -0.5$ = His z

so $-0.5 = \frac{Y - 9}{2}$ $Y = 8.0$

OR 12.4 − (2.2 × 2) = 8.0 faster.

(d) 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 median, standard deviation and 75th percentile?

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

so mean = median = 160 for normal dist.

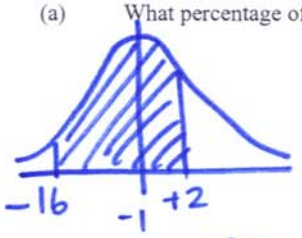
New SD = 2 * 5 = 10 SD not affect by addition

New 75th? OLD is $z = +.67 = \frac{Y - 9}{2}$ $Y = 10.34$ old 75th

so new is (23 + 10.34) * 5 = 166.7

2. The next four 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.

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



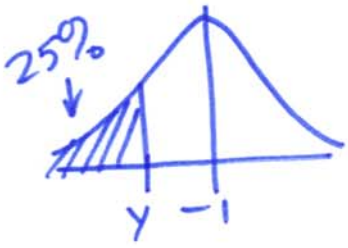
for +2 $z = \frac{2-(-1)}{12} = 0.25$ area is .5987

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

difference is .5987 − .1056 = .4931 or 49.3%

(b) A stock is at 25th percentile (i.e. 25% of the stocks have returns equal to or lower than this stock), what is its one-year return? (5 points)

Use $z = -.67$ and solve

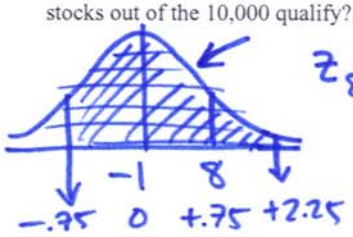


$-.67 = \frac{y-(-1)}{12}$

$y = -9.04$

9% is

(c) Financial advisors search for stocks within ±1.5 Z scores of a return of 8%. Approximately how many stocks out of the 10,000 qualify? (5 points)



$z_8 = \frac{8-(-1)}{12} = \frac{9}{12} = +.75$ so w/i ±1.5z

& .75 is $z = -.75$ to $z = 2.25$

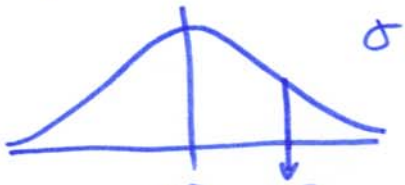
area is .2266 area .9878

So .9878 − .2266 = .7612 or 76.1% or 7,612 stocks of 10,000

(d) Your financial advisor is able to predict the future with some accuracy and says that next year, stocks will be normally distributed and the median return will be +2% and a stock at the 90th percentile will have a return of 11%. Given this information, is it possible to calculate the standard deviation for next year? Please answer YES or NO. If you answer YES, please calculate the standard deviation. If you answer NO, please explain why this is not possible to calculate. (5 points)

YES median = mean = 2%

$\sigma = ?$ recall $z = \frac{y-\mu}{\sigma}$



+1.65 or +1.64

so $+1.65 = \frac{11-2}{\sigma}$ solve

$\sigma = 5.4545$

3. 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 approached for an interview

2) Non-response bias - the majority who were approached would not speak w/ the assistants

4. 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	X				
B	Frozen Food Brand	X				
C	Number of students in a classroom			X	X	
D	Your age			X		X

5. 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?

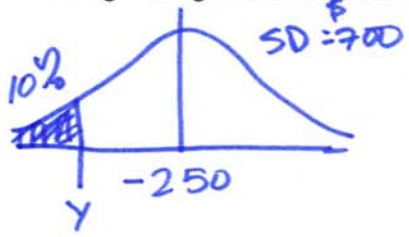
ALL COLLEGE STUDENTS

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

specifically - GPA difference between exercisers and non-exercisers

6. 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): 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.

(a) 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"?

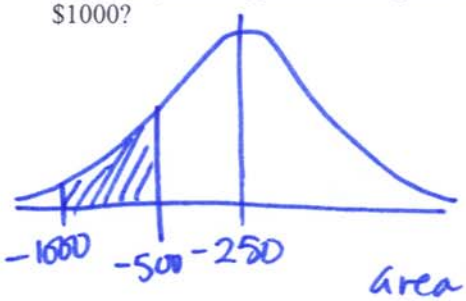


choose $z = -1.28$ b/c closest to 10% in the left tail

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

$\$-1,146$

(b) 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

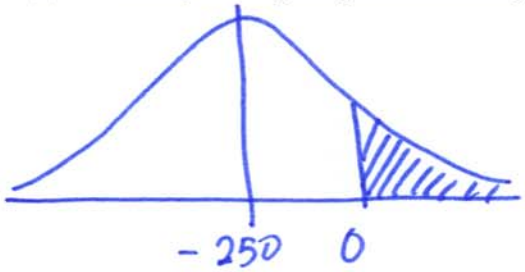
area between is $.3594 - .1423 = .2171$ or 21.7%

(c) What is the median gambling income? What is the variance for gambling income?

Median = -250 dollars

VARIANCE = $SD^2 = (700)^2 = 490{,}000$ dollars squared

(d) What percentage of gamblers actually break even or make money?



= 0 positive $$

$$z = \frac{0-(-250)}{700} = +.357 \approx +.36$$

area to the left is .6406

so shaded area is

$1.0 - .6406 = .3594$

about 36%

7. 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.

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

(a) all Super Bowl Partygoers
(b) all Americans] did not all watch the S.B.
(c) all Americans who watched the Super Bowl on television] did not go to a party necessarily
(d) 850 Americans
(e) 490 Americans who reported that they had attended a Super Bowl Party] not populations
(f) 110 Americans who had not reported attending a Super Bowl Party

B. (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] the population parameter of greatest interest (not necessarily an exact value)

C. (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 → this will generate the statistics of interest
(f) 110 Americans who had not reported attending a Super Bowl Party

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

(a) 19 partygoers
(b) 1330 calories → may be generalizable to the population
(c) 560 calories
(d) 850 Americans
(e) 77% got drunk
(f) 5% got drunk

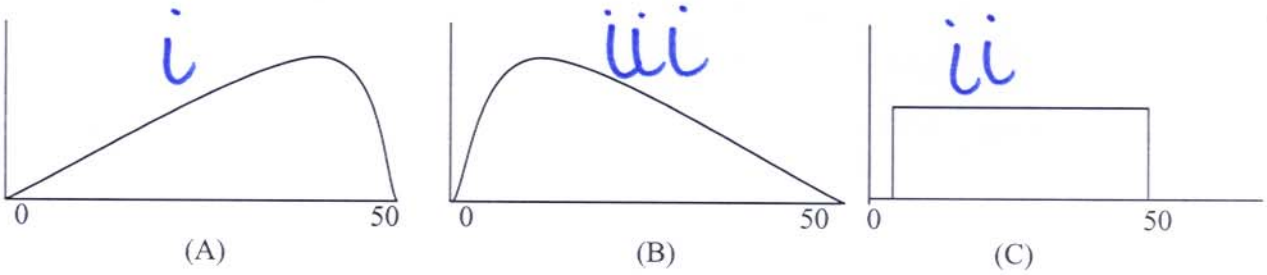
E. (2 points) This is an example of a

(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

8. 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			Confounding is not a source of bias toss this
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 non-response bias
K	X		A histogram with one peak is said to be unimodal, a histogram with two peaks is said to be bimodal

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



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

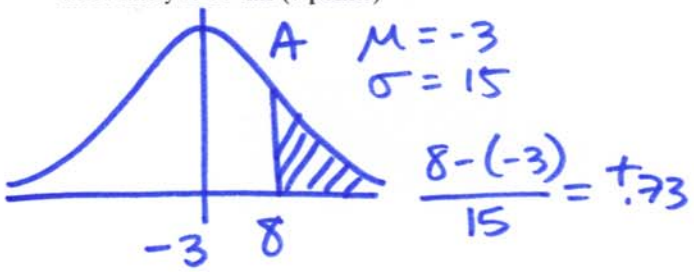
- **i.** The average is smaller than the median.
- **ii.** The average is equal to the median
- **iii.** The average is larger than the median
- **iv.** Cannot determine the average for this graphic
- **v.** Cannot determine the median for this graphic

10. 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 it's implied e.g. non-smokers, but not assigned
c. Investigators can establish association but not causation
d. Confounding factors cannot be controlled can control some
e. All of the above are true
f. Only A and C are true
g. Only D is false

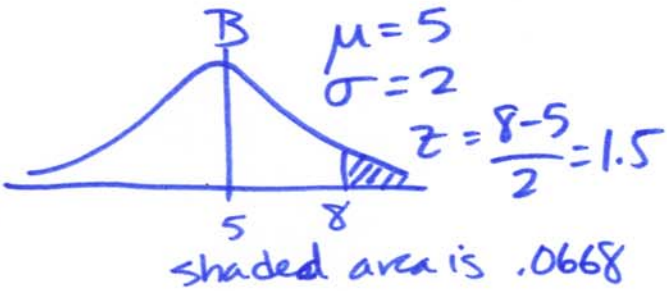
11. The next 2 questions refer to this statement, but each question is separate (i.e. you can get the first one wrong and its result will not affect the others): You are on the verge of investing some of your hard-earned money in the stock market and you are examining two funds, let's call them A and 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)**



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

shaded area is .2327



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

shaded area is .0668

.2327 > .0668

so A is more likely

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 median

FUND B → " = 2 = new median

A → new SD = 30

B → new SD = 4

Work

mean A = (−3 − 4) * 2 = −14

mean B = (5 − 4) * 2 = 2

SD A = 15 x 2 = 30

SD B = 2 x 2 = 4

12. Please indicate whether the statements below are true or false (1 point each)

	True	False	Statement
A	X		Variables whose values are categorical but not quite quantitative are often called ordinal variables
B	X		A categorical variable is a variable that names categories with text or numbers p.8 p.9
C	X		A relative frequency table differs from a frequency table by giving percentages rather than counts of the values in each category of a categorical variable p.16
D		X	A representative sample is a sample whose statistics reflect the corresponding sample parameters accurately ↑ population
E	X		Retrospective and Prospective studies are types of observational studies p.241-242

13. Some computer output from a database of 907 movies produced between 1996-2000. The variable totalgrossreceipts is the total amount of money earned (in millions) domestically.

totalgrossreceipts

Percentiles		Smallest		
1%	.437118	.235049		
5%	1.184727	.242093		
10%	2.019237	.258212	Obs	907
25%	5.672903	.297527	Sum of Wgt.	907
50%	17.0678		Mean	33.85867
		Largest	Std. Dev.	48.02333
75%	42.42211	293.5017		
90%	90.4436	306.1693	Variance	2306.241
95%	125.6034	431.0883	Skewness	4.075231
99%	215.3973	600.787	Kurtosis	32.94395

A. This variable most probably came from a(n): (2 points)

(a) Observational Study with no treatments or factors
(b) Study confounded by voluntary response bias
(c) Study which utilizes convenience sampling
(d) Experiment which utilizes simple random sampling
(e) Prospective Study with totalgrossreceipts as its primary response variable
(f) None of the above

B. (1 point) Are there outliers present in this variable? (circle one) (YES) NO

C. (3 points) Justify your answer in the space below (an answer utilizing numbers is required for full credit)

IQR is 42.42211 − 5.672903 = 36.7492

1.5 × IQR = 55.1238

so 42.42211 + 55.1238 = 97.5459 there are values greater than this
75th (1.5×IQR)

there are no outliers at the low end

D. Please calculate the range (not interquartile range) and list the values of the quartiles (i.e. Q1, Q2, Q3) for this variable in the space below. (4 points)

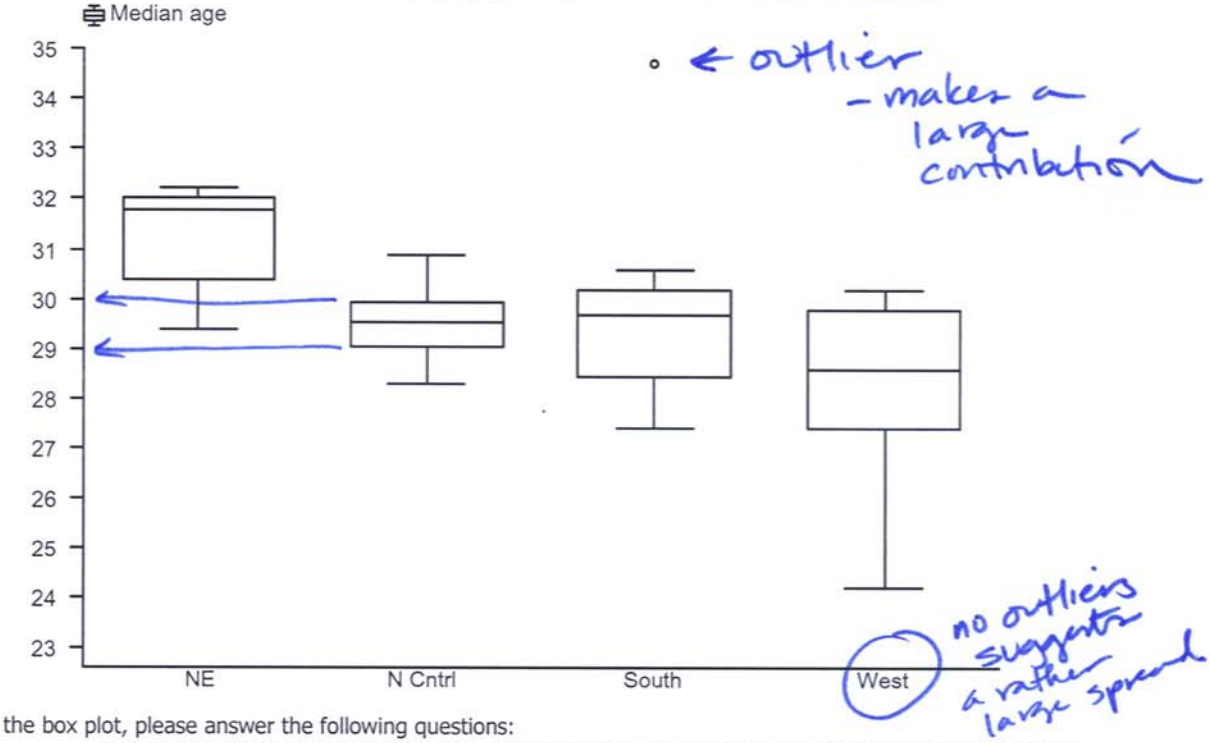
range = 600.787 − .235049 = 600.55195

Q_1 = 5.672903

Q_2 = 17.0678

Q_3 = 42.42211 there is no Q4

14. 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:



o ← outlier - makes a large contribution

no outliers suggests a rather large spread

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 probably has the most skewness in age? (circle one) (2 points)

NE N Cntrl (South) West (Not enough information)

accept either

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

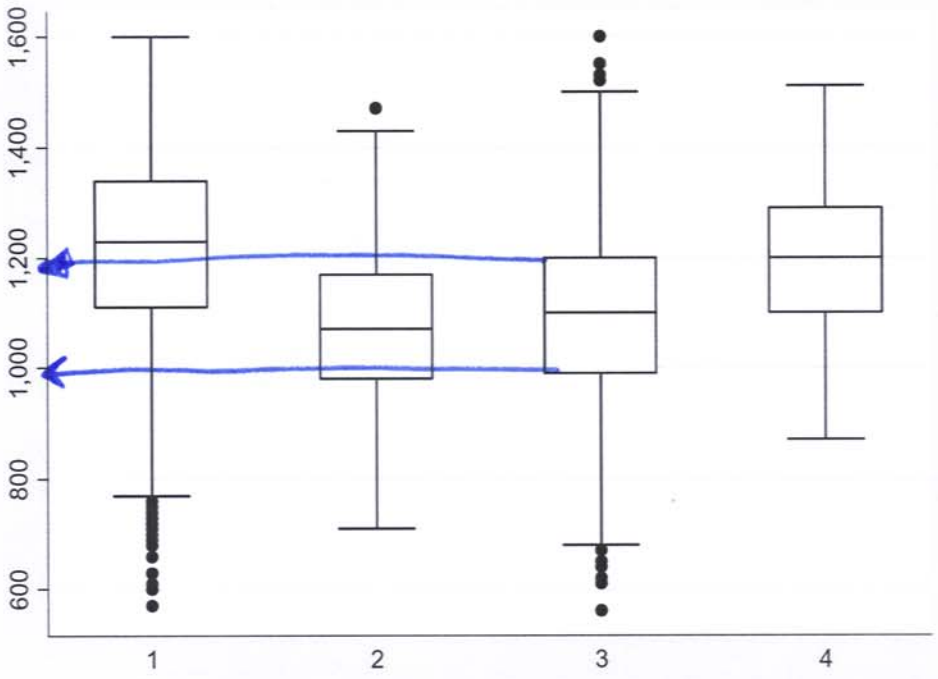
NE N Cntrl South (West) Not enough information

e) Is there enough information in the box plot to accurately 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 or symmetry we might say mean = median, but we can't so the safe answer is NO.



15. The horizontal axis should be labeled "GROUP" and the vertical axis should be labeled "POINTS". The dark dots should actually look like open circles or asterisks. Using the box plot shown above, please answer the following questions:

a) Is there enough information present to estimate the inter-quartile range (IQR) for group 3? (circle one)

(YES) NO

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

IQR = value 75th − value 25th

1200 − 1000

≈ 200

b) Which group appears to be the most symmetrical of the four groups ? (circle one) (2 points)

1 2 (3) OR (4) accepted Not enough information

c) Which group appears to be the most left skewed? (circle one) (2 points)

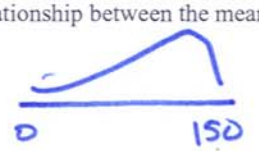
(1) 2 3 4 Not enough information

↑ 3 has outliers at BOTH ends

16. 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

17. 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

18. 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

19. 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

20. There are approximately 10000 publicly traded mutual funds. Morningstar.com gives a 0 to 5 star rating for each of the funds, but a new rival investment guide has come out with a 0 to 3 "thumbs up" rating for each of the funds. Apparently 1500 of the 10000 funds received a "3", 1500 of the 10000 received a "2", 3000 of the 10000 funds received a "1" and the rest received a "0" thumbs up. I have some software that produces a relative frequency table, but I have messed it up during the writing of this exam:

thumbs up	Freq.	Percent	Cum.
0	4,000	40.00	40.00
1	3,000	30.00	70.00
2	1,500	15.00	85.00
3	1,500	15.00	100.00
Total	10,000	100.00	

less than 2 thumbs up .70
at least 2 thumbs up .30
PART B

Using the information above, please assume that the ratings do not change and these are all of the funds in existence. Please answer the following questions:

a. What is the mean number of "thumbs up" and what is the median number of "thumbs up" for this population of funds?

$$\text{MEAN} = \frac{(4000\times 0)+(3000\times 1)+(1500\times 2)+(1500\times 3)}{10{,}000}$$

$$= 1.05$$

MEDIAN = 1 (that's where 50% is)

b. You select 3 funds at random with replacement from the 10000. What is the probability (or chance) that at least 2 of the 3 funds you select will have at least a 2 "thumbs up" rating (i.e. score 2 or more "thumbs up")? Don't care if its a 2 or 3, they are effectively the same rating

fast way $P(2 \text{ or } 3) = .30$ so $.30^3$ = all 3 at least 2 or 3

then $(.70*.30*.30)+(.30*.70*.30)+(.30*.30*.70) = .1890$

c. You select 3 funds at random again, what is the chance that all 3 will receive "0" thumbs up?

Assume the probabilities do not change or you sampled w/ replacement

$.40 * .40 \times .40$

$= .064$ or 6.4%

$\rightarrow .1890 + .027 = \boxed{.216}$

or $(.30)^3 + [(.70*.30*.30)\times 3] = \boxed{.216}$

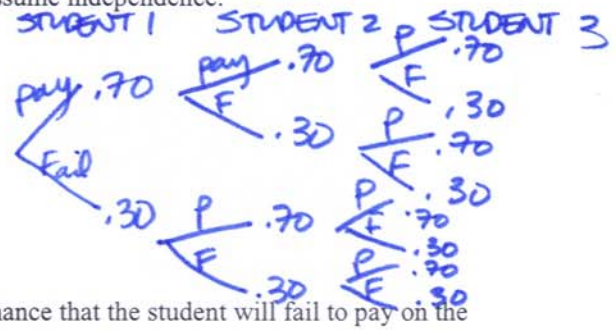
3 arrangements

21. You work for a credit card issuer and it is your job to issue cards to new customers. Since you also go to school, you decide to randomly issue cards to college students. Suppose it is known that 30% of all college students will eventually fail to pay their credit card debt within the first year of possessing a credit card.

A. You issue credit cards to 3 students selected at random, what is the chance that at least one of them will fail to pay their credit card debt within the first year? Assume independence.

(a) 10% or .10
(b) 30% or .30
(c) 34% or .34
(d) 66% or .66
(e) 70% or .70
(f) 90% or .90
(g) less than 10% (less than .10)
(h) greater than 90% (greater than .90)

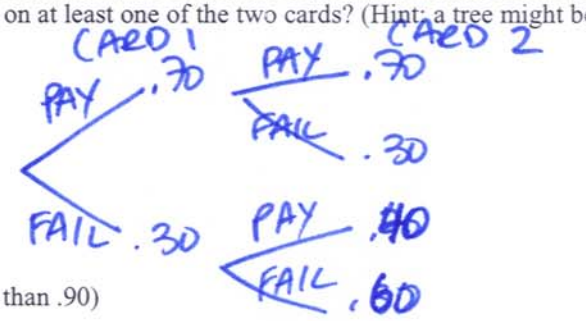
$1.0-(.7\times.7\times.7)$ — all 3 pay



B. Suppose once a student fails to pay on a credit card, the chance that the student will fail to pay on the next credit card that is issued rises to 60%. If a student has not failed to pay on a credit card, the chance still remains 30% for the next card that is issued. A randomly chosen student has been issued two cards. What is the chance that the student will fail to pay on at least one of the two cards? (Hint: a tree might be helpful here):

(a) 60% or .60
(b) 30% or .30
(c) 36% or .36
(d) 49% or .49
(e) 51% or. 51
(f) 64% or .64
(g) 90% or .90
(h) greater than 90% (greater than .90)

$1.0-(.7\times.7)$ — paid both



C. You issue credit cards to 5 students selected at random, what is the chance that all 5 will be able to pay their credit card within the first year? Please show your work for full credit.

$(.7)^5 = .1681 \approx .17$ or 17%

D. You issue credit cards to 2 students selected at random, please fill in the missing information in this table:

# of student failures	probability
0	$.7\times.7 = .49$
1	
2	$.3\times.3 = .09$
Total	1.0

→ so $1.0 - .49 - .09 = .42$

OR

$(.7\times.3) + (.3\times.7) = .21 + .21 = .42$

2 ways to pick 2 students