

This exam has 3 parts:

1. Twenty multiple choice questions (2 points each, total 40 points, use the scantron to record your answers)
2. Open ended questions (40 points, write your answers in the space provided)
3. Data analysis part using the computer; you will be allowed to work with the computer AFTER you turn in the first two parts of the exam. You will type a report with your analysis here in the classroom and email it to the instructors, email addresses will be provided. (20 points)

Tables for the Normal, t-Student and Binomial distributions will be provided.

You will have 3 hours to answer this exam

You can take a 10 minute break when you finish parts 1 & 2 before working with the computer

This is a closed book exam. A formula sheet will be provided

PART I OF THE EXAM – MULTIPLE CHOICE QUESTIONS (40 points, 2 points each). You might write in the exam but please MARK YOUR ANSWERS IN THE SCANTRON and turn in the scantron before proceeding to answer PART 2

1. Which of these questions from the Fall 2015 MATH1530 class survey produced variables that are categorical?
 - i. What is your age?
 - ii. Should military personnel be armed in military buildings on United States soil (Yes or No)?
 - iii. How much did you spend on your last meal that you ate at an off campus restaurant (Round to the nearest whole dollar)?
 - iv. If you had to lose one of your five senses, which would you choose (Hearing, Seeing, Touch, Smell, Taste)?
 - v. What is your favorite number between 0 and 30??

(A) ii (B) iv (C) ii, iv (D) i, iii, v (E) ii, iii, iv
2. The Fall 2015 MATH1530 class survey asked “Approximately, how many times do you eat at an off campus restaurant during the week?” The table below represents the responses of 811 students.

Eat at an off campus restaurant	0	1	2	3	4	5	6	7	8 or more
Number of Students	44	150	175	159	91	85	13	40	54

Estimate the median number of times that the 811 students ate at an off campus restaurant during the week.

- (A) 3 (B) 4 (C) 91 (D) 159 (E) 406
3. You look at real estate ads for houses in Naples, Florida. There are many houses ranging from \$200,000 to \$500,000 in price. The few houses on the water, however, have prices up to \$15 million. The distribution of house prices will be
 - (A) roughly symmetric.
 - (B) skewed to the left.
 - (C) skewed to the right.
 - (D) roughly normal.
 - (E) correlated to land and water.
 4. What are all the values that a standard deviation s can possibly take?
 - (A) $0 \leq s \leq 1$
 - (B) $0 \leq s$
 - (C) $-1 \leq s \leq 1$
 - (D) $-\infty < s < \infty$
 - (E) $68 - 95 - 99.7$

Use the following for the next 2 questions. The table below represents the responses of 814 students to the MATH1530 survey question “If you had to lose one of your five senses, which would you choose?”

Senses	Hearing	Seeing	Smell	Taste	Touch
Count	40	36	504	132	102

5.

Which is the most appropriate graph to display the data?

- (A) Histogram (B) Scatterplot (C) Stemplot (D) Minitab (E) Bar Chart

6.

What can we conclude about this distribution?

- (A) It is somewhat symmetric with smell as the center of the distribution.
(B) It is slightly left-skewed with smell as the center and a spread of hearing to touch.
(C) It is slightly right-skewed with smell being the median.
(D) Smell is the most popular sense to lose and sight/hearing are the least popular senses to lose.
(E) There is a 1 in 5 chance that smell will be selected as the most popular sense to lose.

7.

Which of the following is least affected if an extreme high outlier is added to your data?

- (A) The mean (B) The median (C) The standard deviation (D) The z-score (E) The correlation

8.

There is a positive correlation between the size of a hospital (measured by number of beds) and the median number of days that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing to go to a small hospital?

- (A) Yes, the data show that stays are shorter in smaller hospitals.
(B) Yes, the correlation can't be just by an accident.
(C) No, the positive correlation is probably explained by the fact that seriously ill people go to large hospitals.
(D) No, a negative correlation would allow that conclusion, but this correlation is positive.
(E) No, this is reverse cause-and-effect.

9.

Researchers asked mothers how much soda (in ounces) their kids drank in a typical day. They also asked these mothers to rate how aggressive their kids were on a scale of 1 to 10, with larger values corresponding to a greater degree of aggression. The correlation between amount of soda consumed and aggression rating was found to be $r = 0.3$. If the researchers had measured amount of soda consumed in liters instead of ounces, what would be the correlation? (There are about 34 ounces in a liter.)

- (A) 0.3
(B) $0.3/34 = 0.009$
(C) $(0.3)(34) = 10.2$
(D) 0.3 ounces per point on scale.
(E) Cannot be determined without converting ounces to liters.

10.

Archaeologists plan to examine a sample of two-meter-square plots near an ancient Greek city for artifacts visible in the ground. They choose separate samples of plots from floodplain, coast, foothills, and high hills? This is an example of

- (A) a simple random sample. (C) an observational sample.
(B) a voluntary response sample. (D) a stratified random sample.
(E) an experiment.

11.

Which of the following correlation values indicates the strongest linear relationship between two quantitative variables?

- (A) $r = 0.30$ (B) $r = 0.00$ (C) $r = 0.65$ (D) $r = 0.11$ (E) $r = 0.60$

12.

The website www.twiigs.com allows you to vote on polls that interest you or to post one of your own. Once you have found a poll of interest, you just click "Vote," and your response becomes part of the sample. One of the questions in July 2010 was, "How many times have you been pulled over by the police?" Of the 780 people responding, 70% said "1–5 times." You can conclude that

- (A) about 70% of Americans have been pulled over by the police "1–5 times."
(B) the poll uses voluntary response, so there results tell us little about the population of all adults.
(C) more people still need to vote on the question, as a larger sample is required to reduce bias.
(D) the result is statistically significant since the percent of people that responded yes is greater than 0.
(E) nothing because this an observational study.

13.

Which of the following questions does a test of significance answer?

- (A) Is the observed effect important? (C) Is the null hypothesis true?
(B) Is the sample or experiment properly designed? (D) Is it based on a very small random sample?
(E) Is the observed effect due to chance?

14.

A simple random sample of size n requires that

- (A) the population be unbiased.
(B) every set of n individuals in the population has an equal chance of being selected.
(C) the distribution or original data be approximately normal.
(D) the population size is large.
(E) sample size is large.

Use the following for the next 2 questions. A large representative random sample of 6906 U.S. adults collected over 20 years showed that "parents reported higher levels of life satisfaction than non parents," with the observed difference in life satisfaction between the two groups being statistically significant.

15.

This is an example of

- (A) a randomized comparative experiment. (C) a block design.
(B) a matched pairs experiments. (D) a simple random sample.
(E) an observational study.

16.

In this study, we can conclude

- (A) parents tend to have higher satisfaction in their lives than non parents. Because this was a survey, we cannot make conclusions about cause-and-effect.
(B) having children leads to higher levels of life satisfaction. We can reach this conclusion because we have a representative sample.
(C) having children leads to higher levels of life satisfaction. We can reach this conclusion because we have both a large and a representative sample.
(D) having children leads to higher levels of life satisfaction. We can reach this conclusion because this was an experiment.
(E) that a practical significance exists also. We can reach this conclusion because the result was statistically significant.

17.

An athlete suspected of having used steroids is given two tests that operate independently of each other. Test A has probability 0.9 of being positive if steroids have been used. Test B has probability 0.8 of being positive if steroids have been used. If steroids have actually been used, the probability that Test A is positive and Test B is positive is

- (A) $0.9 + 0.8 = 1.7$
- (B) $0.1 \times 0.2 = 0.02$
- (C) $0.9 \times 0.8 = 0.72$
- (D) $0.1 + 0.2 = 0.3$
- (E) $0.9 \times 0.8 - 0.1 \times 0.2 = 0.7$

18.

A medical researcher is working on a new treatment for a certain type of cancer. The average survival time after diagnosis on the standard treatment is one year. In an early trial, she tries the new treatment on four subjects who have an average survival time after diagnosis of three years. Although the survival time has tripled, the results are not statistically significant even at the 0.10 significance level. The explanation is

- (A) the placebo effect is present, which limits statistical significance.
- (B) that although the survival time has doubled, in reality the actual increase is really two years.
- (C) the calculation was in error. The researchers forgot to include the sample size.
- (D) the sample size is small.
- (E) the results are not of practical importance.

19.

The most important condition for sound conclusions from statistical inference is usually that

- (A) the data can be thought of as a random sample from the population of interest.
- (B) the population distribution is exactly Normal.
- (C) the number of observations is large.
- (D) the t-test is used.
- (E) the p-value is small.

20.

What is the difference between the computed confidence interval and the hypothesis test?

- (A) There is no difference between the two inference procedures since both methods say that $\mu \neq 500$.
- (B) The confidence interval gives us a range of values (\$397.10 to \$423.10) for the mean book cost rather than just telling us whether or not the hypothesis test is significant.
- (C) The confidence interval is 95% where as the P is 0%.
- (D) The hypothesis test requires a different set of assumptions than the confidence interval.
- (E) The hypothesis test describes the magnitude of an effect where as the confidence interval gives you a strict comparison with a pre-specified value.

PART 2 OF THE EXAM. (40 points) Write your answers in the space provided

PLEASE SHOW YOUR WORK

1. (6 points) These are the lengths of a random sample of 20 sugar maple leaves:

7.6 5.8 7.7 9.0 8.1 8.6 6.3 7.5 8.6 8.7 7.8 6.5 6.1 8.0 7.5 8.0
7.0 6.2 7.9 8.2

a) Prepare a stem-and-leaf display of the data

b) sketch the boxplot

c) Calculate the five number summary

2. (10 points) In a large population of college-educated adults, the IQ of the individuals has an approximately normal distribution with mean IQ score is 112 with standard deviation of 25.

a) (3p) Sketch the distribution. What is the upper quartile of the IQ scores?

b) (3p) Peter got an IQ score of 140. In which percentile is he? (i.e. what percent of students score less than him?)

c) (4p) Suppose that somebody claims that the distribution of IQ is not exactly normal in that population, that it is moderately skewed with a somewhat longer tail to the right. Assume that you could take all the possible samples of size 400 from this population and for each sample you would calculate the sample mean. What would be the shape, mean and standard deviation of those sample means? Explain your answer.

3. (13 points) A random sample of 759 ETSU full time undergraduate students was selected to participate in a survey. One of the questions in the survey was: 'Approximately how much did you spend on books for school this semester?' The sample mean of the quantities reported by the students was \$410.1 and the standard deviation \$182.59,

a) (4p) Calculate and interpret a 95% confidence interval for the population mean.

b) (2p) A National research indicates that full-time undergraduate college students spend on average \$500 on textbooks per semester. We want to know if the mean amount spent at ETSU differs from \$500. Write the null and alternative hypothesis.

c) (3p) What is your conclusion about the null hypothesis in b)? Why? If you perform any calculations to answer this question, show your work.

d) (4p) A very large university with more than 50,000 students wants to do a similar survey. How many students should they include in the sample if they want their 95% confidence interval to have a margin of error of 10 dollars. Based on the information gathered at ETSU assume that the standard deviation in that larger university is \$200.

4. (11 points) The question 'Do you think the government should require mandated childhood vaccinations?' was asked to a random sample of 814 students of a large university; 656 of them answered YES.

a) (4p) Build a 95% confidence interval for the proportion of all students in that university who agree with mandated childhood vaccinations.

b) (4p) Assume that we want to conduct a similar survey at ETSU. What sample size do we need if we want the margin of error of the 95% confidence interval to be 0.05? You could use or not use the information you have for the large university; it is your choice. Show your work.

c) (3p) Lets go now to a community where we know that only half of the people agree with mandated childhood vaccinations. Assume you take a random sample of 20 individuals. What is the probability that 12 or more of them agree with mandated childhood vaccinations? Explain your answer.

TURN IN YOUR ANSWERS BEFORE WORKING ON THE NEXT SECTION WITH THE COMPUTER

PART 3 OF THE EXAM (20 points)

PLEASE, TURN ON THE COMPUTER NOW. You are expected to use statistical software (Minitab or R) to solve this part. You will prepare a Word document with your analysis inserting the computer output you get including graphics, interpretations answers to the questions etc. and email it to lewisn2@etsu.edu & seier@etsu.edu You will work with the data set 'pima2'. There are two versions one in Minitab format and one is a text file (pima2.mtw and pima2.txt), you are welcome to use any of them. The data file contains information for 724 members of the Pima community in Arizona. The variables are:

- glucose (plasma glucose concentration measured 2 hours into a procedure)
- diastolic blood pressure
- BMI (Body mass index, it is an index that compares the weight of the person with his/her height)
- Obese (yes or no ; this is a categorized version of BMI, if BMI>30, obese=YES, if BMI<=30, obese=NO)
- age
- test. A test of diabetes was administered, test=1 means the individual shows signs of diabetes, test=0 means the individual does NOT show signs of diabetes).

The purpose of this analysis is to explore what variables are associated to having or not having diabetes, the main research question is if body mass index is associated to having diabetes or not. The amount of glucose in the blood plays an important role in diabetes. First we want to have an idea of the demographic composition of the sample and explore the relationship of glucose with other variables. We will organize the analysis by asking specific questions. Please insert the plots and numerical output in your report

1. We want to know how old the people that we are studying are. Obtain an appropriate plot for the variable age and comment on the shape of the distribution. Calculate the five-number-summary for age and interpret the values of the two quartiles.
2. We want to know what variables glucose is associated with. Prepare plots to explore the possible association of glucose with each one of these variables: diastolic blood pressure, BMI and age. Calculate statistics to quantify the strength of those associations. Which is of those variables is more strongly related to glucose?
3. Is there a difference in mean BMI between those who have signs of diabetes and those who do not? Obtain a plot and perform a test of hypothesis to answer the question. Interpret the results.
4. Would you say that the variable BMI has a normal distribution in each of the groups defined by the variable 'test'? Produce plots and/or apply tests that would help you in answering this question. Based on your answer to this question, discuss if you are worried about the reliability of the analysis done in question 3 or if you are not worried. Explain why.
5. Now focus on the variables 'obese' and 'test'. Produce a two-way table for these variables. Apply a test to explore if having signs of diabetes is independent of being obese. Interpret the results.

After answering each question, write a paragraph summarizing your findings and answering the main research question.