

STATISTICS 641 - ASSIGNMENT 1

DUE DATE: Noon (CDT), FRIDAY, January 30, 2015

Name _____

Email Address _____

Please TYPE your name and email address. Often we have difficulty in reading the handwritten names and email addresses. Make this cover sheet the first page of your Solutions.

STATISTICS 641 - ASSIGNMENT #1 - Due Noon (CDT) Friday - 01/30/2015

- Read Handouts 1 and 2
 - Read Chapters 1 and 3 in the Textbook
 - Problems to Handin:
1. (8 points) What are two major problems in the initial analysis of the O-ring failures data described in Handout 1?
 2. (8 points) What is one of the most frequent misinterpretations of statistical findings?
 3. (12 points) In each of the following studies, (i) State whether the study is experimental or observational; (ii) State whether the study is comparative or description; (iii) If the study is comparative, identify the response variable and the explanatory variable.
 - **Study 1:** To evaluate the risk of myocardial infarction and or death from coronary heart disease after discontinuation of low dose aspirin in primary care patients with a history of cardiovascular events, a group of subjects ($n = 39,513$) aged 50-84 with a first prescription for aspirin (75-300 mg/day) for secondary prevention of cardiovascular outcomes in 2000-2007 were identified, and they were followed up for a mean of 3.2 years to identify cases of non-fatal myocardial infarction or death from coronary heart disease. After the follow-up period, it was found that compared with current users, people who had recently stopped taking aspirin had a significantly increased risk of non-fatal myocardial infarction or death from coronary heart disease combined.
 - **Study 2:** In order to assess which method was more effective in soliciting contributions for their annual fund raising campaign, a non-profit organization contacted 50% of their previous contributors by telephone and 50% by personal contact from an administrator of the organization. The amount of money raised from each person contacted was used to determine the superior method of fund raising.
 - **Study 3:** An entomologist randomly selects deer to determine the proportion of deer that have ticks that have the potential to transmit Lyme disease.
 - **Study 4:** A social scientist randomly selects students from a variety of high schools to estimate the proportion of high school students who are frequent marijuana smokers. The researchers also recorded whether the students were classified as being academically successful or not.
 4. (12 points) Brazos county plans to survey 1000 out of the 60,000 registered voters in the county regarding their preference on the county paying a portion of the building of a new football stadium for Texas A&M University. A complete alphabetical list of the registered voters is available for selecting the 1000 participants. In each of the following scenarios, identify by name the type of sampling method being used.
 - a. Out of the first 60 names on the list, one name is randomly selected. That person and every 60th person on the list after that person are then included in the survey.
 - b. Each voter is randomly assigned a number between 1 and 60,000 with no repeats. The voters' names are then ordered from smallest to largest based on their assigned number. The first 1000 voters on the list are selected for the survey.
 - c. The list of 60,000 voters is divided by into 10 separate lists by voting districts within the county. The 60,000 voters are randomly assigned a number between 1 and 60,000 with no repeats. The names on each of the 10 lists are then ordered from smallest to largest by the number assigned the voter. The first 100 names on each of the ten ordered lists are selected to be in the survey.
 - d. The list of 60,000 voters is first divided into four lists consisting where the voter lived, the East, West, North, or South regions of the county. There are 120 voting precincts in the county with 30 precincts in each region. A random sample of 10 precincts is taken within each region and then a simple random sample of 25 votes is taken within each of the randomly selected precincts. The resulting 1000 voters are then interviewed in a personal survey.

5. (12 points) For each of the following studies, state what type of sampling design you would recommend and why you would select that design.
- You need to collect information on the types of electronic devices used for connecting to the internet by employees of a large corporation. There are separate lists of employees with each list classified by job categories, for example, supervisors, engineers, research scientist, clerical, etc.
 - Information is needed concerning the demographics of people arriving at an emergency room. The available resources for the project will limit obtaining information from only 10% of the arriving people.
 - The MBA program at a large university wants to evaluate the success of their program by obtaining information on the current employment status of their graduates since 2000. They plan on surveying 25% of the graduates. A complete alphabetical list of the graduates is available.
 - The Department of Education would like to estimate the average amount of educational loans incurred by undergraduate students. A national sample of undergraduate students is being planned but a comprehensive list of all college students does not exist. A list of all public and private colleges does exist and each college has a list of their students.
6. (16 points) Consider the experiment discussed in class concerning the determination of DMZ in a product. Suppose the experiment was repeated and the following data was obtained:

			Chemical Analysis				
Operator	Specimen	Run	1	2	Run Mean	Specimen Mean	Operator Mean
1	1	1	136.75	138.75	137.75	138.00	142.75
		2	137.75	139.75	138.75		
		3	136.25	138.75	137.50		
	2	4	146.75	148.75	147.75	147.50	
		5	149.25	145.25	147.25		
		6	146.75	148.25	147.50		
2	3	7	147.25	149.25	148.25	148.50	143.50
		8	150.75	148.75	149.75		
		9	146.25	148.75	147.50		
	4	10	136.65	138.85	137.75	138.50	
		11	137.55	139.55	138.55		
		12	140.45	137.95	139.20		
3	5	13	166.75	168.25	167.50	163.50	143.25
		14	157.25	161.25	159.25		
		15	162.25	165.25	163.75		
	6	16	122.75	124.75	123.75	123.00	
		17	124.25	127.25	125.75		
		18	118.75	120.25	119.50		

- Modify the R code - chemplotv1.R given in Handout 1 and in Dostat: Files/RCode, to produce a diagram similar to the diagram given in Handout 1.
 - Which of the following sources of variation in the 36 observations do you think has the largest source of variation and which source has the smallest?
 - Operator - O
 - Specimen within Operator - S(O)
 - Run within Specimen and Operator - R(S,O)
 - Analysis within Run, Specimen, and Operator - A(R,S,O)
- In Questions 7-10 on the next page, select the BEST answer.

7. (8 points) The NRC commissioned a study to evaluate the impact of nuclear power plants on the temperature of the water down stream from the discharge from the plant. The researcher was concerned about the possible effects of climate on the study so she divide the USA into 5 regions and randomly selected 10 power plants from each region. During a five month period of time, the water temperature was measured daily at each of the 50 power plants at a location above and below the point of discharge into the stream. A total of 150 measurements are taken at each of these two locations. This type of study is an example of
- A. a simple random sample.
 - B. a simple random cluster sample.
 - C. a stratified simple random sample.
 - D. a stratified cluster random sample.
 - E. a multistage cluster random sample.
8. (8 points) An advocacy group for improved health care in the US wants to estimate the cost of treating elderly patients. To obtain information about doctors across the US, the group randomly selected 100 counties from a list of the 1000 largest counties in the US. In each county, the county public health administrator is contacted and asked to randomly select 20 doctors in their county. Each selected doctor then randomly selected 25 of their patients over the age of 60 and determined the annual medical cost for each patient. This study is an example of
- A. a simple random sample.
 - B. a simple random cluster sample.
 - C. a stratified simple random sample.
 - D. a stratified cluster random sample.
 - E. a multistage cluster random sample.
9. (8 points) A study was designed to evaluate the effects of pine bark beetles on native pine trees in East Texas. The researcher divided East Texas into 35 regions. Within each of these regions, she randomly selected 15 native pine trees, each of the 15 trees was examined, and an identifier was placed on those limbs where a pine bark beetle was located. Six months later she returned and determined the amount of damage to each of the limbs where an identifier had been placed. This type of study/sampling method is an example of
- A. a stratified cluster random sample.
 - B. a stratified simple random sample.
 - C. a multistage cluster random sample.
 - D. a factorial random experiment.
 - E. an observational prospective study.
10. (8 points) FEMA wanted an assessment of the amount of damage caused by hurricane Ike to individual homes on the coast of Texas. A random sample of 50 homes was taken from each of the twenty-five coastal counties of Texas. An evaluation of the amount of damage to each of the 1250 homes was made. These 1250 measurements were then summarized into an overall average amount of damage per home. This type of study is an example of
- A. a simple random sample.
 - B. a stratified simple random sample.
 - C. a simple cluster sample.
 - D. a stratified cluster random sample.
 - E. a multistage cluster random sample.