# STATISTICS 641 - ASSIGNMENT 1

## DUE DATE: Noon (CDT), WEDNESDAY, September 8, 2021

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) WEDNESDAY, September 8, 2021**

- Read Handouts 1 and 2
- Problems to Submit for Grading:

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:** A study monitors the occurrence of heart disease over a 5-year period in men randomized to each high fiber or low fiber diets.
   - **Study 2:** An industrial pump manufacturer monitors warranty claims and surveys customers to assess the failure distribution of its pumps.
   - **Study 3:** A biologist randomly selects fish in a river to determine the proportion of fish which show signs of health problems due to pollutants that were poured in the river upstream by a chemical plant.
   - **Study 4:** A study from hospital records found that women who had low weight gain during their pregnancy were more likely to have low birth weight babies than women who had high weight gains during their pregnancy. The researchers also recorded the age and ethnicity of the women.

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,

   i. state whether the study is a survey, a prospective study, or a retrospective study;

   ii. state whether the study is comparative or descriptive;

   iii. if the study is comparative, identify the response and explanatory variable(s).

   a. A sample of members listed in the directory from a professional organization is used to estimate the proportion of female members.

   b. To assess the effect of smoking during pregnancy on premature delivery, mothers of preterm infants are matched by age and number of previous pregnancies to mothers of full term infants and then both are asked about their smoking habits during pregnancy.

   c. A sociologist interviews juvenile offenders to find out what proportion live in foster care.

   d. A marketing study uses the registration card mailed back to the company following the purchase of a video game to gauge the percentage of purchasers who learned about the purchased game from advertising on facebook, television, or by word of mouth.

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 |  |  |

   a. Modify the R code - DMZplot.R in Canvas: R Files, to produce a diagram similar to the diagram given in Handout 1.

   b. 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 cluster random sample.

   D. a stratified simple 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 multistage cluster random sample.

   E. a stratified 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 identifer 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 identifer had been placed. This type of study/sampling method is an example of

   A. a stratified simple random sample.

   B. a stratified cluster 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 simple cluster sample.

   C. a stratified simple random sample.

   D. a stratified cluster random sample.

   E. a multistage cluster random sample.

# STAT 641 Fall 2021

## Solutions for Assignment # 1

- Problem 1 - ( 8 Points) Two major problems are
  - Ignoring the launches in which there were no O-ring failures and
  - Extrapolating the data from previous launches in which the temperature was above 50 degrees to a launch in which the temperature would be in the low 30's.

- Problem 2 - ( 8 Points) One of the most frequent misinterpretations of statistical findings is attributing a "causal" relationship between two events when only a strong correlation exists between the events.

- Problem 3 - ( 12 Points)

  **Study 1:** (i) Experimental. (ii) Comparative (iii) Response: Occurence of heart disease.
  Explanatory: Amount of fiber in diet.

  **Study 2:** (i) Observational. (ii) Descriptive.
  The manufacturer is recording why pumps fail.

  **Study 3:** (i) Observational. (ii) Descriptive.
  The biologist is acquiring health data on the collected fish

  **Study 4:** (i) Observational. (ii)Comparative.(iii) Response: Baby's birthweight.
  Explanatory: Mother's weight gain during pregnancy.

- Problem 4 - ( 12 Points)

  (a.) Systematic sampling.

  (b.) Simple random sampling.

  (c.) Stratified random sampling with the strata being the voting districts.

  (d.) Stratified Multi-Stage Cluster sampling. The voting precincts are stratified into one of four regions, and then 10 voting precincts are randomly selected from each of the four regions. The selected precincts consist of clusters of voters and a random sample of 25 voters is selected from each of the selected precincts.

- Problem 5 - ( 12 Points)

  (a) (i) Survey. (ii) Descriptive

  (b) (i) Retrospective Study. (ii) Comparative. (iii) Response: Term of pregnancy. Explanatory: Mother's smoking habits.

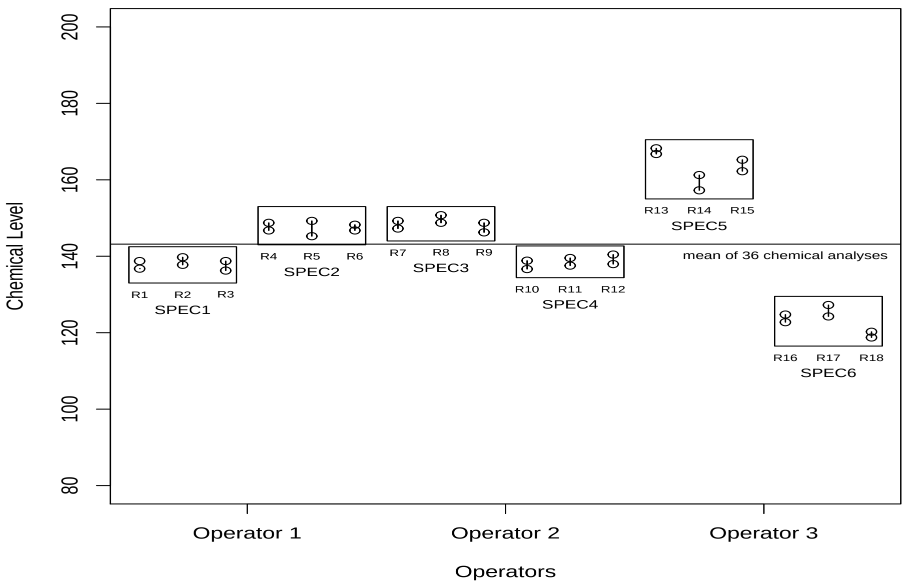
  (c) (i) Survey. (ii) Descriptive

  (d) (i) Survey. (ii) Descriptive

- Problem 6 - ( 16 Points)

  (a.) See the following plot. The code is at the end of this document.

**Figure 3: Chemical Variation Plot**



  (b.) S(O) means show the greatest amount of variability in comparison to the other three sources and the O means have the least amount of variability. In fact, 97% of the overall variability in the 36 response is attributable to S(O); 2.2% of the variability is attributable to R(S,O); 0.5% of the variability is attributable to A(R,S,O); 0.0% of the variability is attributable to O, and .3% of the variability is attributable to all other sources. We will demonstrate how to obtain these percentages in STAT 642.

7. - ( 8 points) **C**. The strata are the 5 Regions with a SRS of 10 Power Plants selected in each Region. The Power Plants are clusters with the daily measurements being the units within the clusters.

8. - ( 8 Points) **D**. The First Stage Clusters are Counties which are clusters of doctors. The Second Stage Clusters are Doctors with patients being the units within the clusters.

9. - ( 8 Points) **B**. The strata are the 35 Regions with a SRS of 15 Pine Trees selected in each Region. The Pine Trees are clusters with the limbs having pine bark beetles being the units within the clusters.

10. - ( 8 Points) **C**. The strata are the 25 Costal Counties with a SRS of 50 homes selected in each County.

```
run = c(1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,13,13,14,14,
15,15,16,16,17,17,18,18)
Res = c(
 136.75, 138.75,
 137.75, 139.75,
 136.25, 138.75,
 146.75, 148.75,
 149.25, 145.25,
 146.75, 148.25,
 147.25, 149.25,
 150.75, 148.75,
 146.25, 148.75,
 136.65, 138.85,
 137.55, 139.55,
 140.45, 137.95,
 166.75, 168.25,
 157.25, 161.25,
 162.25, 165.25,
 122.75, 124.75,
 124.25, 127.25,
 118.75, 120.25)
spec = seq(1,6)


plot(run,Res,type="p",xlab="Operators",ylab="Chemical Level",
        main="Figure 3: Chemical Variation Plot ",cex=.99,
        ylim=c(80,200),xaxt="n")
rect(0.75,133,3.25,142.5)
segments(1,136.75,1,136.75)
segments(2,137.75,2,139.75)
segments(3,136.25,3,138.75)
text(1,130,"R1",cex=.55)
text(2,130,"R2",cex=.55)
text(3,130,"R3",cex=.55)
text(2,126,"SPEC1",cex=.75)

rect(3.75,143,6.25,153)
segments(4,146.75,4,148.75)
segments(5,149.25,5,145.25)
segments(6,146.75,6,148.25)
text(4,140,"R4",cex=.55)
text(5,140,"R5",cex=.55)
text(6,140,"R6",cex=.55)
text(5,136,"SPEC2",cex=.75)

rect(6.75,144,9.25,153)
segments(7,147.24,7,149.25)
segments(8,150.75,8,148.75)
segments(9,146.25,9,148.75)
text(7,141,"R7",cex=.55)
text(8,141,"R8",cex=.55)
text(9,141,"R9",cex=.55)
text(8,137,"SPEC3",cex=.75)

rect(9.75,134.4,12.25,142.7)
segments(10,136.65,10,138.85)
segments(11,137.55,11,139.55)
segments(12,140.45,12,137.95)
text(10,131.4,"R10",cex=.55)
text(11,131.4,"R11",cex=.55)
text(12,131.4,"R12",cex=.55)
text(11,127.4,"SPEC4",cex=.75)

rect(12.75,155,15.25,170.5)
segments(13,166.75,13,168.25)
segments(14,157.25,14,161.25)
segments(15,162.25,15,165.25)
text(13,152,"R13",cex=.55)
text(14,152,"R14",cex=.55)
text(15,152,"R15",cex=.55)
text(14,148,"SPEC5",cex=.75)

rect(15.75,116.5,18.25,129.5)
segments(16,122.75,16,124.75)
segments(17,124.25,17,127.25)
segments(18,118.75,18,120.25)
text(16,113.5,"R16",cex=.55)
text(17,113.5,"R17",cex=.55)
text(18,113.5,"R18",cex=.55)
text(17,109.5,"SPEC6",cex=.75)

axis(side=1,at=c(3.5,9.5,15.5),
labels=c("Operator 1","Operator 2","Operator 3"))
abline(143.1667, 0)
text(16,140,"mean of 36 chemical analyses", cex = 0.7)
```