Sampling, Experiments & EDA Review Sheet

Topics to Study for Exam

- Major Terms: Population, Sample, Census, Random, Bias, Parameter, Statistic
- Various Types of Bias
- Various ways of collecting data
- Experimental Design
- Quantitative vs Categorical
- Exploratory Data Analysis
- Letters used in statistics

1. Determine whether each of the following statements is describing a parameter (population value) or a statistic (sample value) and then give the letter that we use to represent it from the following list: $\overline{x}, \mu, \hat{p}, p, s, \sigma$

   a) The standard deviation of the heights of American men is 3.6 inches.
   b) 46% of the sample showed signs of increased rust.
   c) The average yearly salary of adults in Los Angeles is $41,000.
   d) Of the 200 dogs in the data set, 87% of them were licensed.
   e) The standard deviation for the sample data was 5.2 years.
   f) The average weight of the group in the data set was 155 pounds.

2. Jim wants to know how much money the average working COC student makes. Describe how Jim could use the following techniques to collect data and describe how well the sample data will approximate the population value.

   a) Systematic
   b) Voluntary Response
   c) Random Sample
   d) Convenience Sample
   e) Cluster Sample
   f) Stratified Sample
   g) Simple Random Sample
   h) Census
3. Define the following key terms and give an example of each.
   a) population
   b) census
   c) sample
   d) random
   e) bias
   f) parameter
   g) statistic

4. Describe and give an example of each of the following types of bias.
   a) Sampling Bias
   b) Question Bias
   c) Response Bias
   d) Deliberate Bias
   e) Non-Response Bias

5. What is the difference between a random sample and a simple random sample? Give an example of a sample that is random, but not simple random. Explain Why.

6. Rachael needs to do an experiment that will show that the nicotine patch causes a person to stop smoking. Set up the experiment for Rachael. Write a description of the experiment and include the following. What are some lurking variables that she will need to control? How can Rachael control the lurking variables? Include a description of how we will deal with the placebo effect?
7. Compare and contrast the similarities and differences between an experiment and an observational study. How can we tell if we should use an experiment or an observational study?

8. Tell if the following data is categorical or quantitative. If the data set is quantitative and we created a histogram for the data, what do you think the shape would look like? Why can’t we find the shape for categorical data?
   a) The types of cars in the different COC parking Lots.
   b) The average number of hours spent practicing ping pong.
   c) Areas in North Dakota that have wild mustangs.
   d) Each person is asked if they wear glasses, contacts, neither, or both.
   e) The average speed of the race cars at the Indianapolis 500.
   f) The test scores on a really easy test.

9. Look at the following summary statistics: max, sample size, min, mean, stand dev, median, Q1, Q3, IQR, Range, Variance, mode,
   a) Which of the statistics are measures of center (average)?
   b) Which of the statistics are measures of spread (variability)?
   c) Which of the statistics are measures of position?
   d) Are there any statistics in the list that are not a center, not a spread, nor a position?
   e) What measure of center (average) should we use when the data is bell shaped?
   f) What measure of spread (variability) should we use when the data is bell shaped?
   g) How do we find two numbers that typical values are in between when the data is bell shaped?
   h) What measure of center (average) should we use when the data is skewed or uniform?
   i) What measure of spread (variability) should we use when the data is skewed or uniform?
   j) How do we find two numbers that typical values are in between when the data is skewed or uniform?
10. The following data set describes the lengths in feet of pieces of lumber at a lumber yard. Type the data into a column of statcrunch and make a histogram, dotplot and boxplot, and find the summary statistics (max, min, mean, stand dev, median, Q1, Q3, IQR, Range, Variance, mode), best measure of center, average, best measure of spread, range for typical values, outliers.

17.4 10.7 14.4 19.7 13.5
21.6 17.8 18.2 17.3 17.2
13.2 16.3 15.7 19.1 12.7
18.6 18.2 13.6 16.7 13.1
11.8 21.3 14.8 16.4 7.6

11. The following graph was made from the final exam scores of students in a history class.

a) What is the shape?

b) Is the mean and standard deviation an accurate representative of center and spread?

c) The mean average was 77. Write a couple sentences explaining this statistic in context and what it tells us.

d) The standard deviation was 5.3 . Write a couple sentences explaining this statistic in context and what it tells us.

e) The mean average was 77 and the standard deviation was 5.3 . Use the mean and standard deviation to find two numbers that typical values are in between. Use the mean and standard deviation to find the cut off for “unusually low” test scores. Use the mean and standard deviation to find the cut off for “unusually high” test scores
Sampling / Experiments / EDA Review Sheet Answers

1.  a) Parameter, $\sigma = 3.6$

    b) Statistic, $\hat{p} = 46\%$

    c) Parameter, $\mu = $41000

    d) Statistics, $\hat{p} = 87\%$ , $n = 200$

    e) Statistic, $s = 5.2$

    f) Statistic, $\bar{x} = 155$

2.  a) Systematic : Look at a list of all COC students and pick every 20th person on the list. Would not represent the population because it is not random. If he chooses the first person randomly, then it would represent the population.

    b) Voluntary Response: Create a survey on facebook and ask COC students to respond. Will not represent the population.

    c) Random Sample: He puts the names of all COC students in a hat and shakes it up and draws out 50 names. This will represent the population since everyone had a chance of being chosen.

    d) Convenience Sample : He picks everyone that he goes to class with. Will not represent the population.
e) Cluster Sample: He randomly picks 8 classes and gets information from every individual in those classes. Since it's random, it would represent the population.

f) Stratified Sample: He separates the COC students into 1st year, 2nd year, 3rd year and then picks 50 people from each group. Would not represent the population unless he picks the 50 people randomly.

g) Simple Random Sample: He puts the names of all COC students in a hat and shakes it up and draws out 50 names. This will represent the population since everyone had a chance of being chosen.

h) Census: Attempting to get data from all COC students. The COC computer has a list of all students. He contacts and gets information from all of them.

3. a) population: The collection of all people or objects to be studied. For example, all domestic animals in Lancaster CA.

b) census: Attempting to get information (data) from everyone in a population. May or may not succeed. This is the best data and represents the population very well. For example, measuring the IQ of every employee in a software company.

c) sample: Getting information (data) from a subgroup of the population. Usually less than 10% of the population. For example, measuring the breed, age and weight of 48 dogs in Oklahoma.

d) random: When everyone in a population has an equal chance of being included in the sample. For example: Each employee of a software company has an employee ID number. Have a computer randomly choose ID numbers. Whichever employee’s number comes up, that employee will take an IQ test.

e) bias: When sample data does not represent the population. Usually specific groups have been left out and are not being represented. For example: Wanting to get information about all domestic animals in Lancaster CA and instead only getting information about dogs and cats. (Left out other domestic animals.)

f) parameter: A number that represents a population. For example a population proportion (percentage) \( p = 0.75 \)

g) statistic: A number that represents a sample. For example a sample standard deviation \( s = 10.6 \) pounds.

4. a) Sampling Bias: When the sample was too small, collected incorrectly or without randomization. For example, the sample data was collected by putting a survey up on Facebook.
b) Question Bias: When someone phrases a question in order to force people to answer the way they want. For example: In order to save children from devastating diseases, should all children have vaccinations.

c) Response Bias: When people do not feel comfortable answering truthfully. For example, we ask people if they tend to hoard possessions in their house.

d) Deliberate Bias: When the people collecting the data, deliberately leave out certain groups from the population. For example, a person wants to get data on who Americans will vote for in the next election, but does not ask any Filipino Americans.

e) Non-Response Bias: When people are asked to give data, but refuse to be part of the study. For example, a random phone number generator gave a phone number, but when the person called and tried to get data, the person said they did not want to participate.

5. In a random sample, every individual in the population has an equal chance of being chosen to be in the sample. In a simple random sample, every group of size n has a chance of being chosen. Random Cluster Sample is random sample but not simple random because not every group that could possibly be made has a chance. Only those groups designated in the cluster have a chance.

6. Rachael must randomly select two groups. One group will wear a nicotine patch and the other will wear a placebo patch. The individuals and people giving the patch must not know whether it contains nicotine or not. This will control the placebo effect. Lurking variables will be sources of nicotine, how long someone has smoked for, the number of cigarettes smoked per day. She will want her nicotine and placebo groups to be as similar as possible by picking them randomly and blocking. If the treatment group has a much higher percentage of individuals that were able to quit, then she has proven that the patch does cause people to stop smoking.

7. They are similar in that we are exploring relationships between variables. The experiment has the added condition of showing a cause and effect relationship. If we need to show blame or cause then we need to control lurking variables and therefore we need an experiment. If we are just showing a relationship, then we do not need to control lurking variables and a correlation observational study would be fine.

8. a) The types of cars in the different COC parking Lots. Categorical
b) The average number of hours spent practicing ping pong. Quantitative, Skewed Right

c) Areas in North Dakota that have wild mustangs. Categorical

d) Each person is asked if they wear glasses, contacts, neither, or both. Categorical

e) The average speed of the race cars at the Indianapolis 500. Quantitative, Bell shaped

f) The test scores on a really easy test. Quantitative, Skewed left

9. a) Mean, Median, Mode,

b) Standard Deviation, Variance, Range, IQR

c) Min, Max, Q1, Q3

d) Yes. Sample size (frequency)

e) Mean

f) Standard Deviation

g) Mean – Standard Deviation < Typical Values < Mean + Standard Deviation

h) Median

i) IQR

j) Q1 < Typical Values < Q3

10.
This data set describes the lengths in feet of pieces of lumber at a lumber yard. The data set is slightly skewed left so the median of 16.4 feet is the best measure of center. So the average length of the boards at the lumber yard is 16.4 feet. The IQR of 4.85 feet is the best measure of spread. So typical boards had lengths 4.85 feet from each other. Hence typical boards were between 13.35 feet (Q1) and 18.2 feet (Q3) in length. There were no outliers.

11. The following graph was made from the final exam scores of students in a history class.

   a) What is the shape?

   Bell Shaped

   b) Is the mean and standard deviation an accurate representative of center and spread?

   Since the data is bell shaped (normal), the mean is an accurate measure of center and the standard deviation is an accurate measure of spread.
c) The mean average was 77. Write a couple sentences explaining this statistic in context and what it tells us.

The mean is a type of average. It is also the balancing point for the data. The sum of the distances of numbers below the mean will equal the sum of the distances of numbers above the mean. It is only accurate when the data is bell shaped. In this case the average test score was approximately 77.

d) The standard deviation was 5.3. Write a couple sentences explaining this statistic in context and what it tells us.

The standard deviation is a measure of typical spread from the mean. Data with more spread tend to give less consistent values and may be difficult to predict, while data with less spread tend to give more consistent values and may be easier to predict. In this case typical scores on the history exam were within 5.3 points from the mean average of 77.

e) The mean average was 77 and the standard deviation was 5.3. Use the mean and standard deviation to find two numbers that typical values are in between. Use the mean and standard deviation to find the cut off for “unusually low” test scores. Use the mean and standard deviation to find the cut off for “unusually high” test scores.

We find typical values by adding and subtracting the mean and standard deviation. So typical final exam scores on the history final were between 77-5.3 and 77+5.3. So typical scores on the history final were between 71.7 and 82.3

The cut off for a test score being unusually low is the mean minus two standard deviations or 77 – 2(5.3). So any score lower than 66.4 was unusually low when compared to the rest of the class.

The cut off for a test score being unusually high is the mean plus two standard deviations or 77 + 2(5.3). So any score higher than 87.6 was unusually high when compared to the rest of the class.