

Upper St. Clair High School Summer Assignment	
Course: AP Statistics	
Assignment Title:	Naked Statistics Reading
Due Date:	First week of school
Purpose of the assignment:	☐ Review foundational material, concepts, or skills ☑ Build interest ☐ Provide a diagnostic measuring tool to help guide student differentiation & instruction ☐ Establish a supportive class culture & environment ☑ Expose students to required material, concepts, skills, and/or texts that will be referenced throughout the academic year
Resources needed to complete assignment:	☐ Textbook assigned by school ☑ Book(s) supplied by student ☐ Packet provided by teacher (please visit the Interactive Program of Studies) ☑ Other supplies: See attached assignment sheet for more details.
Estimated time for Completion:	Students may average an estimated 13 hours to complete the assignment, but we encourage students to spend time on this assignment over several weeks.
Impact on Semester 1 Grade	Graded: ☑ Yes ☐ No Teacher Note: There will be a quiz on the reading during the second week of class.

Name: _____ AP Statistics

Mr. Coffman's AP Statistics Summer Assignment

Read the book Naked Statistics: Stripping the Dread from the Data by Charles Wheelan (paperback, 304 pages, ISBN-13: 9780393347777). Amazon.com has several cheap used copies available (~$7), but you can simply borrow the book from the USC township library if you do not want to purchase a copy. **You will turn in your answers to the following questions during the first week of school.**

Chapter 1

1) Describe two specific descriptive statistics (like QB passer rating and the Gini index) that are NOT mentioned in Chapter 1. In addition, comment on the purpose of each descriptive statistic you chose.
2) Explain why one might need to use statistical inference to describe a certain characteristic of a population rather than simply calculating a descriptive statistic for the same population. You might want to come up with a specific example to aid in your explanation.
3) Explain how probability can be used to identify potential cases of cheating on standardized exams.
4) Why is it very difficult to determine cause and effect relationships for health-related phenomena in humans?

Chapter 2

1) Explain the advantages of using a single value (descriptive statistic) to summarize a complex set of variables.
2) Explain the disadvantages of using a single value (descriptive statistic) to summarize a complex set of variables.
3) Explain why it is sometimes desirable to use the median to describe the central tendency of a set of data rather than the mean.
4) Explain what the standard deviation measures.
5) You would like to describe the distribution of ages of all people in the Upper St. Clair High School building on the first day of school.
    a) Would the mean or median be a better descriptor of the central tendency of the distribution of ages of all people in the Upper St. Clair High School building on the first day of school? Explain.
    b) Rank these values from largest to smallest: 1) the standard deviation of **all teachers** in USCHS on the first day of school, 2) the standard deviation of **all students** in USCHS on the first day of school, 3) the standard deviation of **all people** in USCHS on the first day of school.
6) Explain why it is important to be careful when using percentages to describe changes in quantity.

Chapter 3

1) Explain the difference between "precision" and "accuracy."
2) Explain why it is important to compare "apples to apples" rather than "apples to oranges" by using one of the examples given in this chapter.
3) Explain one way using a descriptive statistic to evaluate or rank entities can backfire.

Chapter 4

1) Define correlation in the statistical sense.
2) What does a correlation of 1 mean? Of -1?
3) Visit https://www.tylervigen.com/spurious-correlations and explore some of the "spurious correlations" on the site. Then explain what correlation does not imply causation.

Chapter 5

1) One of the key assumptions the author made when computing the probability that a taste-tester would select Schlitz over Michelob more than 40 times during the live taste test was that the probability of the taster picking each beer was 50%. Do you think this was a reasonable assumption? Explain.
2) Define probability.

Name: _____ AP Statistics

3) Explain why the probability of a specific individual being struck by satellite debris in 2011 was 1 in 21 trillion while the probability of anyone anywhere being struck was 1 in 3,200.
4) Explain how DNA analysis relies on probabilities to make positive matches.
5) The expected value of a situation really only makes sense in terms of the "long run" or "many, many trials" of the situation. Explain why you would not necessarily use the expected value for points after a touchdown to decide whether or not to kick the extra point or go for the two-point conversion in the Super Bowl.
6) Briefly summarize the context of this statement on p. 84: "*most of the people who test positive will not have the disease.*"

Chapter 6

1) Briefly summarize each of the six problems with probability presented in this chapter.
    a) Assuming events are independent when they are not
    b) Not understanding when events ARE independent
    c) Clusters happen
    d) The prosecutor's fallacy
    e) Reversion to the mean
    f) Statistical discrimination
2) Choose **one** of the specific examples of misusing/misunderstanding probability from this chapter and describe it in your own words.

Chapter 7

1) Describe the three things we generally ask data to do.
2) Why is it important to gather a representative sample from a population for statistical analysis?
3) Briefly summarize the five types of sample bias addressed in this chapter.
    a) Selection bias
    b) Publication bias
    c) Recall bias
    d) Survivorship bias
    e) Healthy user bias

Chapter 8

1) On p. 133, Wheelan explains why household incomes in the U.S. are skewed to the right. Similarly, distributions of data can be skewed left or symmetric. Explain why the following distributions would be expected to be skewed or symmetric:
    a) The ages of all individuals in the USCHS building on the first day of school → skewed right
    b) The final grades of all students in AP Statistics class → skewed left
    c) The heights of all students in USCHS → symmetric
2) Explain why large sample sizes produce sample means with smaller standard error than small samples.

Chapter 9

1) The term "hypothesis testing" is introduced on p. 146. Briefly explain hypothesis testing in your own words.
2) Explain why it is important to not only consider the *significance* of a finding but also the effect *size* of the finding.
3) Briefly explain what Type I and Type II errors are. Be sure to also address how you know which error is more egregious.

Chapter 10

1) There are two things that go into the calculation of the margin of error – the confidence level and the standard error. After reading the chapter, answer the following questions.

   a) Which will have a larger margin of error – a poll of 1,000 individuals with confidence level 68% or the same poll of 1,000 individuals with a confidence level of 95%? Explain your answer.
   b) Which will have a larger margin of error – a poll of 500 individuals with a confidence level of 95% or the same poll, this time with 1,000 individuals, at a confidence level of 95%? Explain your answer.
2) There are three methodological questions raised in this chapter that should be considered when polling or interpreting the results of a poll. List those three questions.
3) Visit the website https://projects.fivethirtyeight.com/2016-election-forecast/national-polls/ and look at the polls that were taken in November of 2016 leading up to that year's presidential election. As you know, Donald Trump defeated Hillary Clinton in the actual election. Based on your reading of this chapter, speculate on why the polls did not accurately predict the outcome of the election.

Chapter 11

1) Explain in your own words what an "ordinary least squares" line is.
2) The author refers to the slope of the least squares regression line as the "regression coefficient." What three things are we generally interested in with regards to the regression coefficient?
3) Briefly explain what the R^2 value measures.

Chapter 12

1) Briefly summarize the seven common regression mistakes addressed in this chapter.
   a) Using regression to analyze a nonlinear relationship
   b) Correlation does not equal causation
   c) Reverse causality
   d) Omitted variable bias
   e) Highly correlated explanatory variables (multicollinearity)
   f) Extrapolating beyond data
   g) Data mining (too many variables)
2) Choose **one** of the specific examples for common regression mistakes from this chapter and describe it in your own words.

Chapter 13

1) Briefly summarize the four different approaches for isolating a treatment effect addressed in this chapter.
   a) Randomized, controlled experiments
   b) Natural experiment
   c) Nonequivalent control
   d) Difference in differences