Introduction to Data Science

Name:_____

# QUIZ 4 PRACTICE

1. Suppose you wanted to have R simulate data, specifically, the number of heads when you toss a fair coin 10 times. Give one line of R code that will print, on the console, the number of heads for 100 such trials of 10 coins. (1 pts)

rbinom(n = 100, size = 10, prob = .5)

2. An October 4, 2017, Pew Research poll published the result that **56%** of a random sample of **4135** Americans said they would **not** want to ride in a driverless car, if given the opportunity.

   a. It is possible to estimate the precision of this estimate using repeated simulated samples of 4135 people. The name for this statistical technique is called Bootstrap (1 pt)

   b. Give two lines of R code that will simulate 1000 trials of 4135 people, and calculate, from each, the **percentage** of people who say they would not want a ride in a driverless car, putting all of the percentages in a vector called **PercNoRide.** (2 pts)

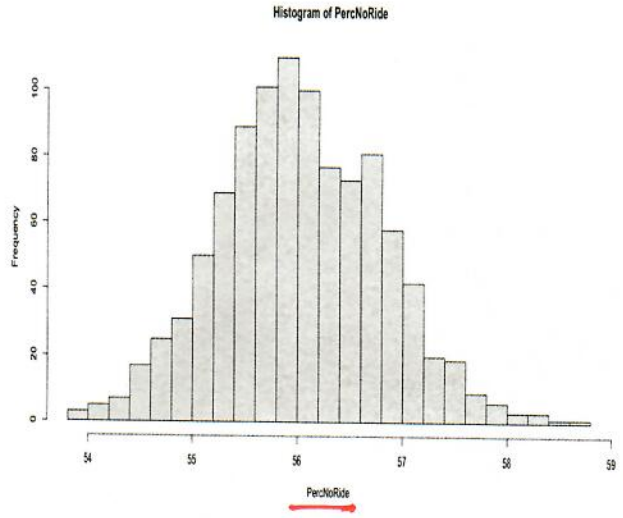
T ← rbinom(n = 1000, size = 4135, prob = .56)

PercNoRide ← 100 * T / 4135

   c. I've created the vector **PercNoRide**; below is a histogram. Its standard deviation is 0.78. Using this information, state the standard error, the margin of error, and a 95% confidence interval. (2 pts)

SE: 0.78 ME: 1.56

CI: 56 ± 1.56 Also, CI as endpoints of an interval: ( 54.44 , 57.56 )




   d. State your confidence interval in a complete sentence in terms of the problem, as we've done in class: "I'm 95% confident that…." (2 pts)

I'm 95% confident that true ~~p~~ percentage of Americans who don't want to ride in a driverless car is captured in 54.44% to 57.56%

e. Based on your interval in d, can you conclude that a **majority** of Americans would **not** want to ride in a driverless car? Explain how the confidence interval values help you reach your conclusion. (1 pts)

yes - at least 95% sure, whole interval > 50%

3. The three lines below are called a Code Chunk. What does the argument **echo=FALSE** do in this code? (2 pt)

```
```{r pressure, echo=FALSE}
plot(pressure)
```
```

don't show the code in the final document.

4. A Harvard study, published in the New England Journal of Medicine, in June, 2017, examined health data from over 60,000,000 elderly U.S. adults on Medicare. Using geographical data, they linked each adult's record of illnesses and/or death to the levels of pollution around their home. They concluded that people who live in locations that are more polluted (near highways, near coal-burning plants, near densely populated areas), are more likely to suffer from illness and death than people who do not.

a. Sampling Error: Does this study suffer from any sampling error? If so, about how much? Explain how you know. (1 pts)

yes, always Not much - huge study.

b. Sampling Bias: Does this study suffer from any sampling bias? Explain how you know that it does or does not. (1 pts)

yes! - Elderly adults do not represent all adults. Elderly adults on Medicare do not even necessarily represent all elderly adults.

c. Confounding: Is it possible that people who live in more polluted locations differ in other ways, that may affect health? If not, explain how you know there are no confounding variables. If so, suggest a plausible confounding variable, explaining how it 'works'. (2 pts)

Yes, since this study is observational, there may always be confounding variables. e.g.,
- Lower income people are more likely to live in undesirable, more polluted places. They also have poorer health, due to cost, access, etc.
- More urban locations may be more polluted. The cause of health problems could also be the social effects of urbanization.

Name:_____

6. Suppose that I asked a random sample of 163 UVM students to rate their attractiveness on a scale from 0 to 10, where 0 is least attractive, and 10 is most attractive. Here, I've summarized the vector of responses, called **Attract,** in the data frame, **surv**. Note that 6 of the 163 students left the question blank.

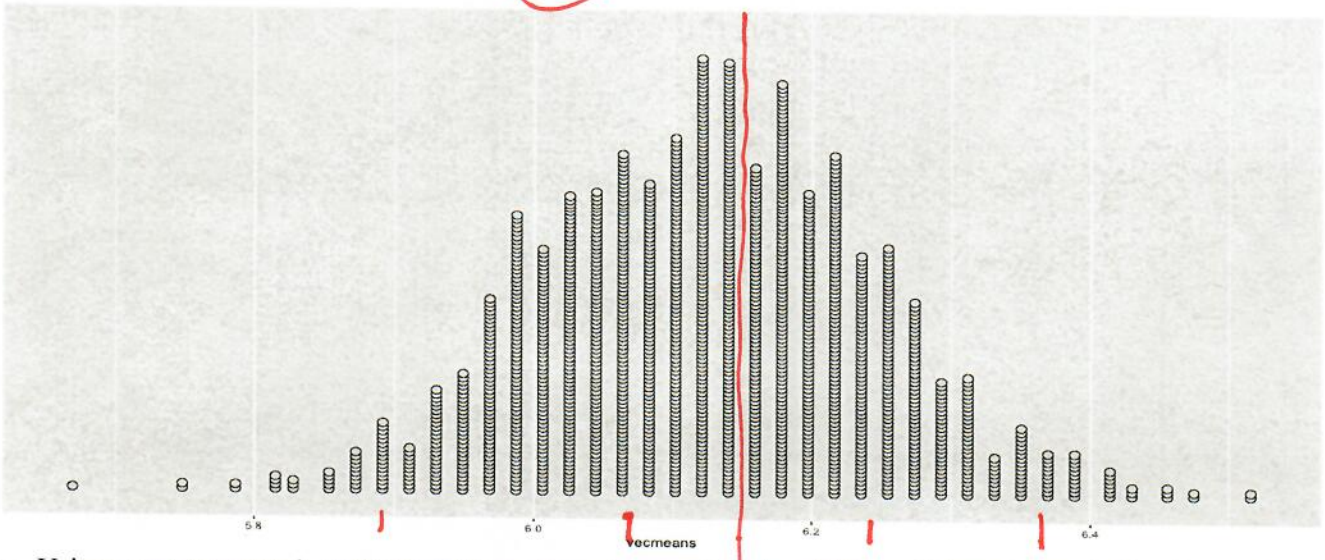
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|
| 0.0 | 5.0 | 6.0 | 6.127 | 7.0 | 10.0 | 6 |

a. Suppose I want R to calculate the distribution of 500 sample means that we would get if we were to repeat the study many times, and use it to calculate the Standard Error. Add the missing elements to the code below, to make it work for this data' (6 pts)

```
OS <- na.omit(surv$Attract)
PE <- mean(OS)
vecmeans <- c()
for (i in 1:500) {                    (157 or)
  S <- sample(x=OS, size=length(OS), replace=TRUE)
  vecmeans <- c( vecmeans, mean(S) )
}
SE = sd( vecmeans )
```

b. Suppose that I create a histogram of vecmeans, and it looks like the graph below. Remembering what you know about standard deviation, choose which of the following is closest to the standard deviation: (1 pts)

**.06** **.12** **.24** **.36** **.48**




c. Using your answer to b, and information given above, calculate a 95% confidence interval, showing work below. Suppose we consider a score of 5 to represent 'average' attractiveness. Can you conclude that UVM students consider themselves to be above average in attractiveness, based on your interval? Explain. (3 pts)

$6.127 \pm 2(.12)$

5.89 to 6.37