Problem set 7

S520

Upload your answers through the Assignments tab on Canvas by 11:59 pm, Thursday 19th October. Draw graphs in R and include code.

1. (5 points.) Let X be a discrete random variable with probability mass function

$$P(X = x) = \begin{cases} 0.3 & x = -2 \\ 0.6 & x = -1 \\ 0.1 & x = 12 \\ 0 & \text{otherwise.} \end{cases}$$

   Let $X_1, \ldots, X_n$ be an iid sequence of random variables with the same distribution as X. Let $\bar{X}$ be the sample mean (of $X_1, \ldots, X_n$.)

   (a) Find EX.

   (b) Find $\text{Var}(X)$.

   (c) What is the expected value of $\bar{X}$?

   (d) What is the variance of $\bar{X}$? (Note: This will depend on n.)

   (e) Suppose $n = 100$. Use the R function `pnorm()` to find the approximate probability that $\bar{X}$ is greater than 0.5.

2. (5 points. From the Spring 2017 final.) I downloaded data on the number of citations for a random sample of 1000 journal articles published in 1981. (The data is from the ISI Citation Indexes.) I ran some analysis on the data in R, and produced the following output:

```
> citations = scan("citations.txt")
Read 1000 items
> summary(citations)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    1.00    9.06    7.25  300.00
> var(citations)
[1] 565.2476
> # Number of articles with no citations
> sum(citations == 0)
[1] 460
```

   (a) Is the distribution of the number of citations (i) exactly normal, (ii) approximately normal, or (iii) not close to normal? How do you know?

   (b) Find an approximate 95% confidence interval for the mean number of citations.

(c) Find an approximate 95% confidence interval for the *proportion* of journal articles with no citations.

3. (10 points.) As part of the 2016 American National Election Studies (ANES) pilot study conducted in January 2016, a sample of 1200 respondents were asked to rate various public figures on a "feeling thermometer" from 0 to 100, where 0 indicates "very cold" and 100 indicates "very warm" feelings toward the figure. The file `ANES2016.txt` contains the feeling thermometer scores for three Presidential candidates: Donald Trump, Hillary Clinton, and Bernie Sanders. To simplify things, assume the respondents are a simple random sample from the population of adult U.S. citizens (this is not quite true but is close enough for our purposes.)

   (a) Plot the distributions of the raw feeling thermometer scores for Trump, Clinton, and Sanders on the same scale. What's wrong with the data?

   (b) Get rid of the feeling thermometer observations that are greater than 100, e.g. by using the `subset()` function. Find the sample mean and sample standard deviation for the new variables for Trump, Clinton, and Sanders.

   (c) Find 99% confidence intervals for:

      i. Trump's mean feeling thermometer score if all adult U.S. citizens were asked in January 2016.

      ii. Clinton's mean feeling thermometer score if all adult U.S. citizens were asked in January 2016.

      iii. Sanders' mean feeling thermometer score if all adult U.S. citizens were asked in January 2016.

   (Note that since we've excluded scores over 100, the sample sizes are no longer all 1200.)

4. (5 points.) Trosset exercise 9.6 exercise 9. (Recall: The length of a confidence interval is the distance from its upper bound to its lower bound.)

5. (5 points.) In a May 2019 Gallup poll,[1] 63% of a sample of 1009 U.S. adults supported same-sex marriage.

   (a) Treating the data as a simple random sample, find a 95% confidence interval for the percentage of all U.S. adults who support same-sex marriage.

   (b) Suppose we wanted to have a 95% confidence interval for the percentage of all U.S. adults who support same-sex marriage with total length 2% (i.e. 0.02.) How large a simple random sample would we need?

[1] `https://news.gallup.com/poll/257705/support-gay-marriage-stable.aspx` . Note: The margin of error Gallup states is different from what we would calculate, because they adjust for the fact that they are not taking a true simple random sample.