Problem Set 7

STAT-S 520

Due on February 27th, 2023

Instructions:

- Submit your answers in Canvas.
- Your answers can be typed and/or handwritten as long as your final submission is a single PDF file with answers in proper order.
- Include your R code, graphs, and output. The latter only when is relevant.
  - Check that only the relevant output is included in your submission. Pages and pages of output that is not relevant can be penalized.
- You are allowed to collaborate with your classmates as long as you write your own solutions.

Questions:

1. Consider an urn that contains 10 tickets, labelled $\{3, 3, 3, 4, 4, 7, 7, 7, 10, 10\}$. From this urn, an experiment consist on drawing $n = 60$ tickets with replacement; let Y and $\bar{X}_{60}$ the random variables that assigns the sum and sample mean of those 60 tickets, respectively; and do the following in R:
   a. Create and object called `urn` that represents the urn with the tickets shown above. Report your R code.
   b. Using R perform the following tasks, in order,
      i. Run a random seed first using `set.seed(520)`,
      ii. Obtain the sum of a random sample of 60 tickets (with replacement) from the `urn`, and
      iii. Obtain the sample mean of another random sample of 60 tickets.
   c. Obtain a big vector of 30000 sums of 60 tickets each. Call this vector `vec.sum`.
   d. Using `vec.sum`, construct a histogram, a normal probability plot, and a kernel density estimate. Does the data seem to be drawn from a normal distribution? Explain.
   e. Obtain a big vector of 50000 sample means of 60 tickets each. Call this vector `vec.mean`.
   f. Using `vec.mean`, construct a histogram, a normal probability plot, and a kernel density estimate. Does the data seem to be drawn from a normal distribution? Explain.
2. Let X_1 be a discrete random variable with probability mass function

$$f(x) = \begin{cases} 0.6 & x = 1 \\ 0.1 & x = 2 \\ 0.2 & x = 3 \\ 0.1 & x = 6 \\ 0 & \text{otherwise.} \end{cases}$$

   Let

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}$$

   i.e. they are independent from each other and each X_i has the same distribution as X_1. Let $\bar{X}_n$ be the sample mean. For each part below, show your work to receive full credit.

(a) Find $E(X_1)$.

(b) Find $\text{Var}(X_1)$.

(c) Obtain $P(1.8 < X_1 \leq 2.1)$

(d) Let $n = 80$. Find $E\bar{X}_{80}$ and $Var\bar{X}_{80}$.

(e) Let $n = 80$. Based on the CLT, approximate $P(1.8 < \bar{X}_{80} \leq 2.1)$

(f) Construct a simulation of 40000 replications, each replication results in the observed sample mean. Use your simulation to obtain the approximate probability that $P(1.8 < \bar{X}_{80} \leq 2.1)$ and compare the result to part (e).

3. ISI Section 8.4 Exercise 5, but instead use the urn

$$\{1, 1, 1, 2, 2, 5, 10, 10, 10, 10\}$$

and still use $n = 40$ tickets and approximate $P(170.5 < Y < 199.5)$ but now

a. Write in R the proposed code, evaluate `urn.model` a total of 10^5 times, share your code, and based on that answer the questions.
b. Using the plug-in principle in the urn, and properties of expected value and variance for the sum of tickets, show that the numbers given, 585.6 and 184, are no longer appropriate. Replace them with the appropriate numbers, run the modified code, and answer the questions.
c. Answer the question.

4. Assume the one can of Coke weights on average 351 grams and one can of Pepsi weights on average 350 grams and both have a standard deviation of 1 gr. If you select at random 40 cans of Coke and 42 cans of Pepsi, do the following:

(a) If you let X_i represent the weight of the i-th Coke can (randomly selected) for $i = 1, \ldots, 40$ and $\bar{X}_{40}$ the average weight of Coke cans, what is $E\bar{X}_{40}$ and $Var(\bar{X}_{40})$

(b) If you let Y_j represent the weight of the j-th Pepsi can (randomly selected) for $j = 1, \ldots, 42$ and $\bar{Y}_{42}$ the average weight of Pepsi cans, what is $E\bar{Y}_{42}$ and $Var(\bar{Y}_{42})$

(c) Can you find, approximately, $P(X_1 > 351.5)$? If yes, find it and report your value. If not, explain why not

(d) Can you find, approximately, $P(\bar{X}_{40} > 351.5)$? If yes, find it and report your value. If not, explain why not

(e) Find the probability that the average weight of 40 Coke cans is greater than the average weight of 42 Pepsi cans.

5. Recall the heuristics when applying the CLT tell us that when the sample size is $n \geq 30$ the sample mean approximately follows the normal distribution. In this question you are asked to come up with counter-examples, i.e., examples that completely violate this rule of thumb.

a. Construct a random variable (i.e., create a vector of values, as we did in class) where the population distribution is not normal, but that when $n = 5$ the sample mean is already very close to normal.
b. Construct another random variable such that when you obtain random samples of size $n = 2000$, the distribution of the sample mean, $\bar{X}_{2000}$, does not approximate at all the normal distribution.

For both parts, use the function `clt()` to obtain graphs that justify your findings.

Additional Exercises (do not turn in)

6. ISI Section 8.4 Exercise 4.

7. ISI Section 8.4 Exercise 5.

8. Use the function wlln() created in class. Come up with a vector of 10^4 values representing a random variable (similar to `x1` or `x2` created in class) such that the probability of $P(\bar{X}_{1000000} \in (\mu - \epsilon, \mu + \epsilon)) < 0.7$, that is the probability of the sample mean being in a smaller interval around μ is less than 0.7. So, in your function you need to figure out the vector argument (`x`). The other arguments should be `repl = 100` (only 100 replications for quicker computing), `size = 10^6`, and `epsilon = 0.05`. Explain why your random variable (represented by your vector) accomplishes this.

Reading assignments

- ISI Chapter 9, Sections 9.1 - 9.3