S&DS 220: Homework 6

Due Friday February 23

Braeden

Instructions

1. Complete the questions below. Upload your knitted PDF solutions to Gradescope by the due date.
2. Your solutions should be a combination of writing and R code. When writing, use complete sentences.
3. Previous homework assignments already had code chunks created for you. Now it is up to you to insert R code chunks within each problem as needed.
4. You should aim for clear and concise communication (in both words and R code).

Problem set questions

Problem set questions

Question 1: (Exercise 4.11) Normal distribution

Suppose that scores on an exam are normally distributed with mean 80 and standard deviation 5, and that scores are not rounded.

(a) What is the probability that a student scores higher than 85 on the exam?

Solution

This would put the student one devation above the mean. One deviation is 5 points, so the probability is $P(X > 85) = 1 - P(X \leq 85) = 1 - P(Z \leq 1) = 1 - 0.85 = 0.15$.

(b) Assume that exam scores are independent and that 10 students take the exam. What is the probability that 4 or more students score 85 or higher on the exam? Compute the exact probability, and then estimate it with simulation.

Solution Z-scores are used to find the probability that a student scores above 85. The formula for Z scores is $Z = \frac{X-\mu}{\sigma}$. The Z score for 80 is $Z = \frac{80-75}{5} = 1$. Therefore, the probability of scoring higher than 80 is the complement of the CDF for this z-score which is represented as $1 - P(Z \leq 1) = 1 - 0.80 = 0.20$.

```
n <- 10 # num students
mu <- 80 # mean score
sigma <- 5 # standard deviation
n_sims <- 10000
times_above_4 <- 0

for (i in 1:n_sims) {
  scores <- rnorm(n, mu, sigma)
  times_above_4 <- times_above_4 + (sum(scores >= 80) >= 4)
}
times_above_4 / n_sims
```

```
## [1] 0.8291
```

Question 2: (Exercise 4.18) Uniform distribution

Suppose the time (in seconds) it takes your professor to set up their computer to start class is uniformly distributed on the interval $[0, 30]$. Suppose also that it takes you 5 seconds to send your mom a nice, quick text that you are thinking of her. You only text her if you can complete it during the time your professor is setting up their computer. If you try to text your mom every day in class, what is the probability that she will get a text on 3 consecutive days?

Solution Since each day is independent, the probability of sending a text on any given day is the same. The probability of sending a text on any given day is the probability that the time it takes to set up the computer is greater than 5 seconds. The probability of sending a text on any given day is $P(X > 5) = 1 - P(X \leq 5) = 1 - \frac{5}{30} = 1 - \frac{1}{6} = \frac{5}{6}$. The probability of sending a text on 3 consecutive days is $(\frac{5}{6})^3 = \frac{125}{216}$. This computation is relatively simple because we are dealing with a uniform distribution, thereby allowing us to use the formula for the probability of a uniform distribution.

Question 3: (Exercise 4.19) Exponential distribution

Suppose the time to failure (in years) for a particular component is distributed as an exponential random variable with rate $\lambda = 1/5$. For better performance, the system has two components installed, and the system will work as long as either component is functional. Assume the time to failure for the two components is independent. What is the probability that the system will fail before 10 years have passed? Compute the exact probability, and then estimate it with simulation.

Solution.

The formula for exponential random variable with rate $= \lambda$ is $f(t) = \lambda e^{-\lambda t}$. Failure for the two components is independent, therefore, the probability is simply 1 - the sum of the probabilities that both components will last longer than 10 years. The probability that a single component will last longer than 10 years is $P(systemworkingaftertyears) = 1 - P(bothfailbytyears) = 1 - (1 - e^{-\lambda t}) * (1 - e^{-\lambda t}) = 1 - (1 - e^{-\frac{1}{5} * 10})^2 = 0.98168436111$. Simulation:

```
simulate <- (function(n_sims, lambda, t) {
  times_fail <- 0
  for (i in 1:n_sims)
  {
    times_fail <- times_fail + (min(rexp(2, lambda)) < t)
  }
  times_fail / n_sims
})
simulate(10000, 1/5, 10)
```

```
## [1] 0.9797
```

Question 4: (Exercise 4.22) Modeling with random variables

For each of the following descriptions of a random variable, indicate whether it can best be modeled by binomial, geometric, Poisson, uniform, exponential, or normal. Answer the associated questions. Note that not all of the experiments yield random variables that are *exactly* of the type listed above, but we are asking about reasonable modeling.

(a) Let Y be the random variable that counts the number of sixes which occur when a die is tossed 10 times. What type of random variable is Y? What is $P(Y = 3)$? What is the expected number of sixes? What is the Var(Y)?

Solution.

$P(Y = 3)$ is a binomial distribution becuase there are only two outcomes, therefore we can determine this value using the binomial distribution formula. The expected # of sixes, $E(Y)$, is $np = 10 * \frac{1}{6} = \frac{10}{6}$. The variance of Y, Var(Y), is $np(1-p) = 10 * \frac{1}{6} * \frac{5}{6} = \frac{50}{36}$.

Only two possible outcomes, success or failure, therefore **binomial** is a good distribution to use for modelling.

(b) Let U be the random variable which counts the number accidents which occur at an intersection in one week. What type of random variable is U? Suppose that, on average, 2 accidents occur per week. Find $P(U = 2)$, $E(U)$, and Var(U).

Solution

This problem requires that we consider time (static period of a week) and the number of accidents that occur in that period. Therefore, **Poisson** would be a fantastic model to utilize here. If on average 2 accidents occur a week, we can caclulate P(U = 2) using the Poisson distribution formula. The expected number of accidents, $E(U)$, is $\lambda = 2$. The variance of U, Var(U), is also $\lambda = 2$.

(c) Suppose a stop light has a red light that lasts for 60 seconds, a green light that lasts for 30 seconds, and a yellow light that lasts for 5 seconds. When you first observe the stop light, it is red. Let X denote the time until the light turns green. What type of rv would be used to model X? What is its mean?

Solution

This problem requires that we consider a non-static timeframe and the time it takes between events occuring. Therefore, **exponential** would be the ideal model for this case. Using exponential, the mean of X is $\frac{1}{\lambda} = \frac{1}{30}$.

(d) Customers arrive at a teller's window at a uniform rate of 5 per hour. Let X be the length in minutes of time that the teller has to wait until they see their first customer after starting their shift. What type of rv is X? What is its mean? Find the probability that the teller waits less than 10 minutes for their first customer.

Solution

Contant rate of arrival, only time of events is changing therefore **exponential** would be the ideal model for this case. The mean of X is $\frac{1}{\lambda} = \frac{1}{5}$. The probability that the teller waits less than 10 minutes for their first customer is $P(X < 10) = 1 - e^{-\frac{1}{5} * 10} = 1 - e^{-2} = 1 - 0.13533528324 = 0.86466471676$.

(e) A coin is tossed until a head is observed. Let X denote the total number of tails observed during the experiment. What type of rv is X? What is its mean? Find $P(X \leq 3)$.

Solution

One event is repeated until success is achieved. Since the number of trials is not fixed, **geometric** would be the ideal model for this case. The mean of X is $\frac{1}{p} = \frac{1}{0.5} = 2$. $P(X \leq 3)$ is $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1 + 0.5 + 0.25 + 0.125 = 1.875$.

(f) Let X be the recorded body temperature of a healthy adult in degrees Fahrenheit. What type of rv is X? Estimate its mean and standard deviation, based on your knowledge of body temperatures.

Solution

Body temperatures average at around 98.6 degrees. This is a unique case of a variable that we know little information about, and typically these cases arise in the life sciences. Therefore, **normal** would be the ideal model for this case. The mean of X is 98.6 degrees based on common knowledge. Since the vast majority of people all subsist within a degree of 98.6, I would estimate the standard deviation to be 0.75.

Question 5: Distribution of the sample mean from an exponential

For each of sample sizes 5, 20, 50, and 200, perform the following:

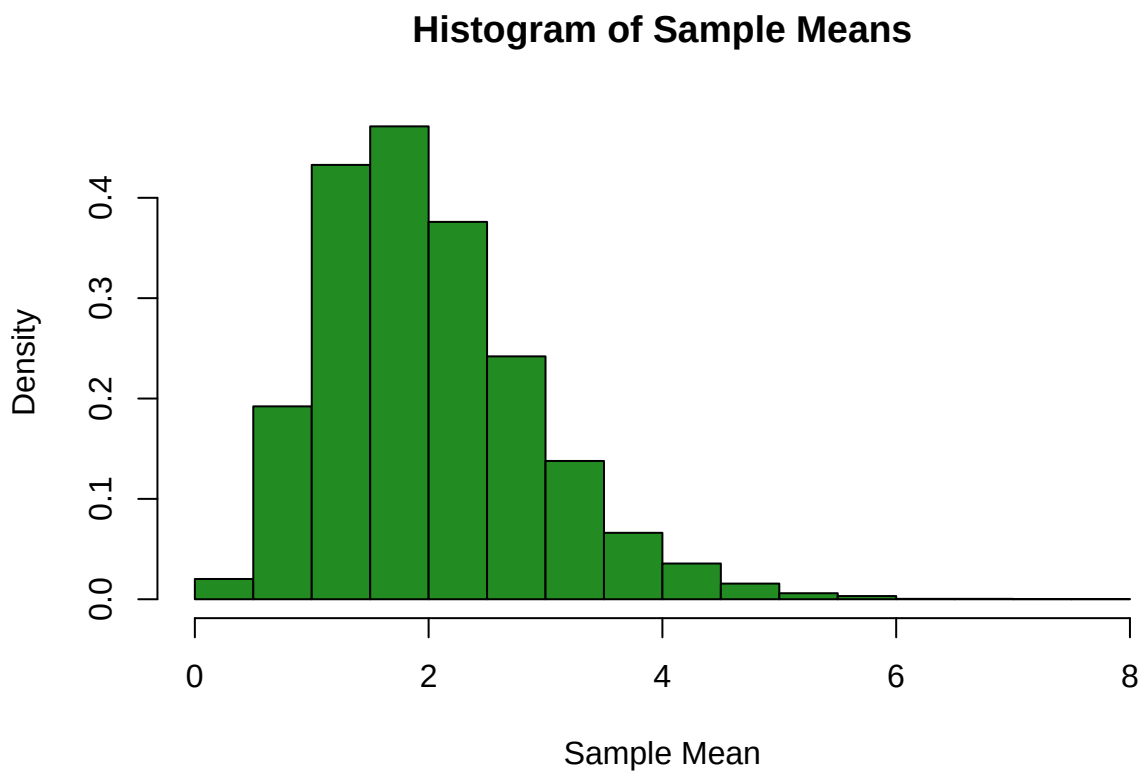
(1) Generate 10,000 sample means (using `replicate`) from a sample of size `n = size` from an exponential distribution with rate parameter `rate = 0.5` (use `rexp` to generate the samples).
(2) Compute the mean and standard deviation of the sample means generated in part (a).
(3) Plot a histogram of the sample means. Set the argument `probability = TRUE` inside the histogram. Give the histogram appropriate axes labels (which should include the sample size).

(a) sample size $n = 5$

Solution.

```
#perform steps (1), (2), (3) with
size = 5
sample_means_size <- 10000
rate <- 0.5

perform_steps <- function(sample_means_size, size, rate) {
  data <- replicate(sample_means_size, mean(rexp(size, rate)))
  mean(data)
  sd(data)
  hist(data, probability = TRUE, xlab = "Sample Mean", ylab = "Density", main = "Histogram of Sample Mea
}
perform_steps(sample_means_size, size, rate)
```

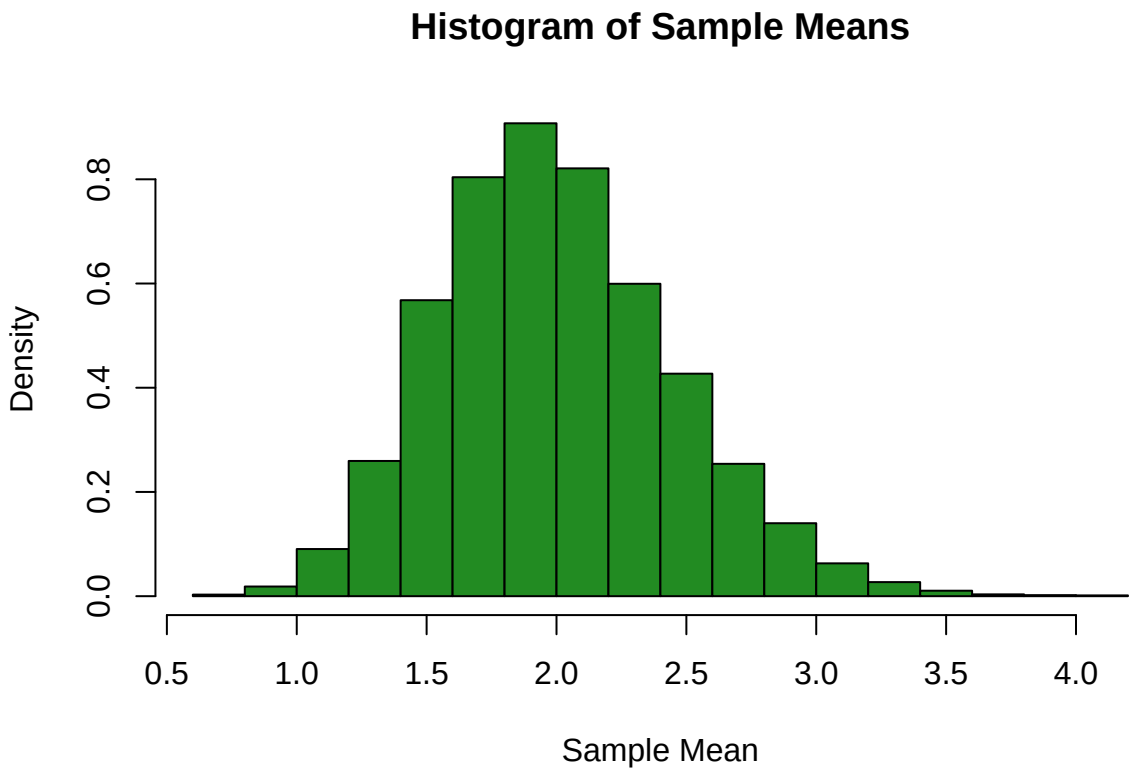


(b) sample size $n = 20$

Solution.

```
#perform steps (1), (2), (3) with
size = 20

perform_steps(sample_means_size, size, rate)
```

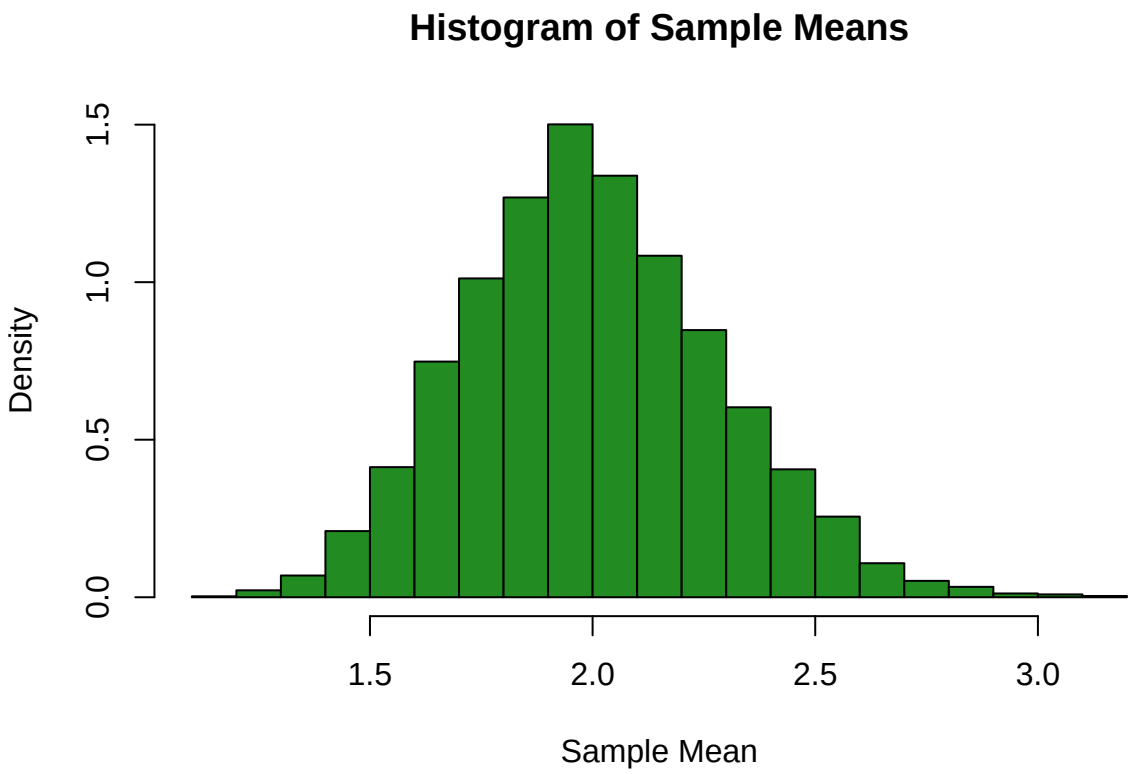


(c) sample size $n = 50$

Solution.

```
#perform steps (1), (2), (3) with
size = 50

perform_steps(sample_means_size, size, rate)
```



(d) sample size $n = 200$

Solution.

```
#perform steps (1), (2), (3) with
size = 200
```

(e) What do you notice about the mean, variance, and shape of the histogram as the sample size n increases?

Solution

As the sample size increases the mean and the variance both increase, and the shape of the histogram becomes more normal. This is consistent with the Central Limit Theorem. The Central Limit Theorum states that the distribution of a sample will approximate a normal distribution under certain conditions, as shown above by the histograms. The CLT states that the distribution of sample means will tend towards a normal distribution as the sample size increases, regardless of the original distribution of the data, provided the original distribution has a finite variance.