STAT 422: Homework #3

Due: February 18, 2022, *11:59pm*

Problem 1

statement

Rusts R Us Repair Shop (part 2) The amount of time (in hours), X, needed by a local repair shop, Rusts 'R Us, to repair a randomly selected piece of equipment is assumed to be an exponential random variable with pdf

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda} I_{(0,\infty)(x)}$$

where $\lambda > 0$ and the mgf of X is given by $m_X(t) = (1 - \lambda t)^{-1}$.

a. This repair shop wants to estimate the expected amount of time needed to repair a piece of equipment. Identify the parameter Rusts 'R Us wants to estimate, and explain why the sample mean, $\overline{X}$, is a reasonable statistic for the business to use as an estimator.

b. In a random sample of 5 repair times, Rusts 'R Us observes the following values (in hours): $14, 6, 3, 4, 1.5$. Based on these data, what is the *observed value* $\overline{x}$ of the statistic (i.e., what is the *estimate*)?

c. Use MGFs to show the exact sampling distribution of $\overline{X} \sim Gamma(5, \frac{\lambda}{5})$. Be sure to provide rationale for each step!

d. **Before attempting this problem go through sampling-distributions.R very carefully.** A customer service representative for Rusts 'R Us has been telling customers to expect an average repair time of three hours, i.e., $\lambda = 3$. *Under this assumption* $(\lambda = 3hrs)$, use R to plot the following distributions:

   - The parent or population distribution
   - The exact sampling distribution of $\overline{X}$
   - An empirical sampling distribution of $\overline{X}$

Provide your code and the code used to produce the three distributions. Explain which of these three distributions are theoretical and which are approximate. *Don't forget: the* `rate` *argument in the Gamma and Exponential distribution functions in R corresponds to the reciprocal of the* β *parameter the text uses. Don't be shy about posting R questions to the discussion board!*

e. Some customers are upset, and they think the customer service representative is underestimating the average amount of time it actually takes to repair a piece of equipment. (That is, they belive the true average repair time, λ, is greater than 3 hours). Use R to find the probability of observing the given sample mean (i.e., $\overline{x}$ from part (b)) or something larger in a random sample of five repairs, if the true population mean is really 3 hours. Shade what this probability represents on the appropriate plot from part (d). Based on this probability, do you think the customers have a valid argument? Explain why or why not.

f. The customer service representative for Rusts 'R Us truly thinks the true average repair time is three hours. He claims that in this case the sample median is a better estimator than the sample mean. Use R to *estimate* the sampling distribution of the sample median of samples of size $n = 5$. (Hint: Type `? median` in R to find out more about the median function.) Based on your approximate sampling distribution, what is the estimated probability of observing the given sample median or one larger if the true population mean is really 3 hours? Based on this probability, do you think the customers have a valid argument? Explain why or why not. Provide all R code and output used.

Solution

a. The company wants to know the expected amount of time needed to repair any piece of equipment. This is the average amount of time needed to repair pieces of equipment: μ. So the sample mean $\bar{X}$ is a good statistic to approximate μ because it is the sample mean.

b. $\bar{X} = \frac{14+6+3+4+1.5}{5} = 5.7$

c. We seek $\mathbb{E}\left[e^{t/n\sum_{i=1}^{n} X_i}\right]$.

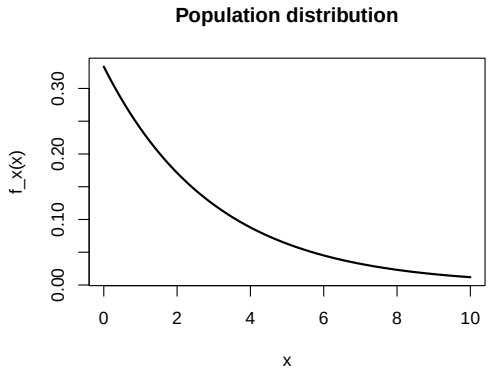
$$\begin{aligned}
\mathbb{E}\left[e^{t/n\sum_{i=1}^{n} X_i}\right] &= \mathbb{E}\left[\prod_{i=1}^{n} e^{t/nX_i}\right] \\
&\overset{iid}{=} \prod_{i=1}^{n} \mathbb{E}\left[e^{t/nX_i}\right] \\
&= \prod_{i=1}^{5} \mathbb{E}\left[e^{t/5X_i}\right] \\
&= \prod_{i=1}^{5} (1 - \lambda\frac{t}{5})^{-1} \\
&= (1 - \lambda\frac{t}{5})^{-5s}
\end{aligned}$$

Which is the mgf of $Gamma(5, \frac{\lambda}{5})$

d. Solution:

- The parent or population distribution

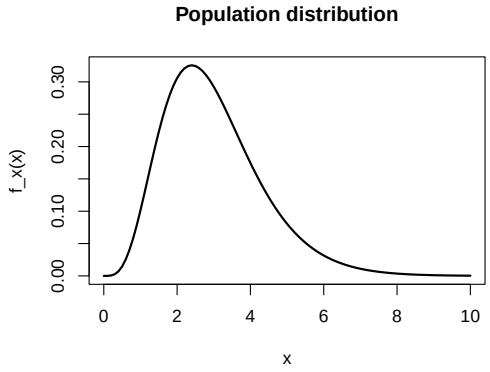
```
curve(dexp(x, 1/3), from=0, to=10, xlab = "x", ylab = "f_x(x)",
      main = "Population distribution", lwd = 2)
```




```
# we got things in weird order so rate = 1/lambda
```

- The exact sampling distribution of $\overline{X}$

```
curve(dgamma(x, 5, 5/3), from=0, to=10, xlab = "x", ylab = "f_x(x)",
      main = "Population distribution", lwd = 2)
```

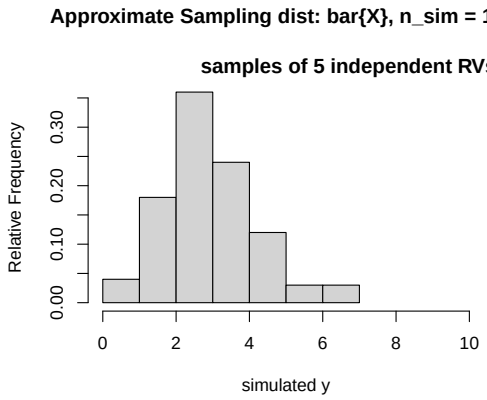


- An empirical sampling distribution of $\overline{X}$

```
y_vec <- rep(NA, 100)

for(i in 1:100){
  samp <- rexp(n = 5, rate=1/3)
  y_vec[i] <- mean(samp)
}
hist(y_vec, xlim = c(0,10),
  xlab = "simulated y",
  ylab = "Relative Frequency",
  freq = F,
```

```
main = "Approximate Sampling dist: bar{X}, n_sim = 100 \n
                    samples of 5 independent RVs")
```



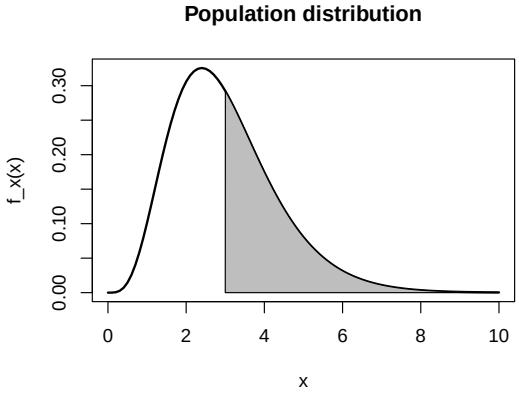
e. We want $P(\bar{X} \geq 3) = 1 - P(\bar{X} \leq 3) = 1 - F_{\bar{X}}(3)$

```
1 - pgamma(3, shape=5, rate = 5/3)
```

```
## [1] 0.4404933
```

We can visualize this by

```
x <- seq(0,10,length=1000)
y <- dgamma(x, 5, 5/3)
curve(dgamma(x, 5, 5/3), from=0, to=10, xlab = "x", ylab = "f_x(x)",
          main = "Population distribution", lwd = 2)
polygon(c(x[x>=3], max(x), 3), c(y[x>=3], 0, 0), col="grey")
```

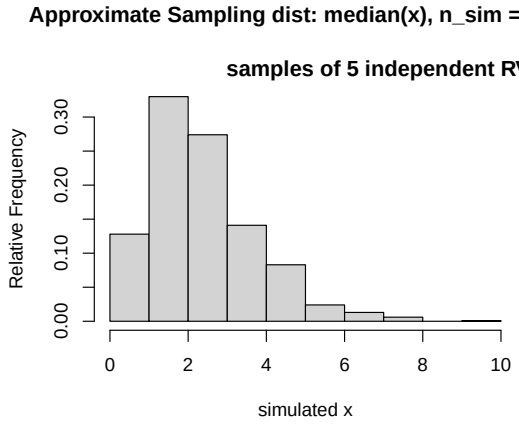


The probability is about 0.44, so it is pretty likely to see times more than 3 hours, but it is less than 50%, so I don't think that the mean is actually greater than three.

f.

```
y_vec <- rep(NA, 1000)

for(i in 1:1000){
  samp <- rexp(n = 5, rate=1/3)
  y_vec[i] <- median(samp)
}
hist(y_vec, xlim = c(0,10),
  xlab = "simulated x",
  ylab = "Relative Frequency",
  freq = F,
  main = "Approximate Sampling dist: median(x), n_sim = 1000 \n
                      samples of 5 independent RVs")
```



In this plot, we see that it is skewed even more to the left (mode around 2) which indeciates that finding the median of ≥ 3 is pretty un-likely. In fact we can compute

```
sum(y_vec >= 3) / 1000
```

```
## [1] 0.268
```

which is about 27%, which is pretty unlikely. So I think the customers are wrong.

Problem 2

Statement

We showed in class that if $Y_1, ..., Y_n$ is a random sample of size n from a χ^2_1 distribution, that $\sum_{i=1}^n Y_i \sim \chi^2_n$. Now, suppose that the Y_is are no longer a random sample, but are independent RVs with distributions $Y_i \sim \chi^2_{\nu_i}$, use mgfs to show that the sum of the Y_is, $\sum_{i=1}^n Y_i$, follows a $\chi^2_{\nu_1+\nu_2+...\nu_n}$ distribution.

Solution

$$\begin{aligned}
\mathbb{E}\left[e^{t*\sum_{i=1}^{n} Y_i}\right] &= \mathbb{E}\left[\prod_{i=1}^{n} e^{t*Y_i}\right] \\
&\overset{iid}{=} \prod_{i=1}^{n} \mathbb{E}\left[e^{t*Y_i}\right] \\
&= \prod_{i=1}^{n} (1-2t)^{-\nu_i/2} \\
&= (1-2t)^{\frac{-1}{2}\sum_{i=1}^{n}\nu_i}
\end{aligned}$$

Which this is the mgf of $\chi^2_{\nu_1+\cdots+\nu_n}$ as required.