ST 501 R project

For this project you will be using R to simulate random data, approximate quantities, and create graphs. The goals of this project are to

- create simulated data sets
- visual convergence concepts by graphing simulated data
- use Monte Carlo simulation to find approximate probabilities

This project will be done in groups. Only one member of your group needs to provide a submission. The R file you turn in must adhere to the R file submission guidelines. You should create a final report that includes your work and relevant R code (R markdown is great for this but no need to use it if you don't know it).

It is ok to discuss this project with students outside your group. However, you cannot share code across groups. This will be considered academic misconduct. Any instances of this or other acts of academic misconduct will be prosecuted.

Visualizing Convergence in Probability

This part will look at the behavior of the minimum order statistic from a random sample from an $exp(1)$ distribution.

1. Show that the minimum order statistic converges in probability to 0. Hint: We know the CDF of an exponential and how to find the CDF of the minimum order statistic. Start with the probability you want to show from the definition of convergence in probability to 0 ($P(|Y_{(1)} - 0| < \epsilon)$) and take the limit as n goes to infinity and show this is 1.

2. To visualize this we'll simulate data and approximate the probability statement proven in the previous part.

   (a) For a sample size of $n = 1$, generate N = 1000 data sets from an $exp(1)$ distribution.

   (b) For each data set, find the minimum value (for a sample of size 1 that will just be the value itself).

   (c) Save these minimum values for plotting.

3. Consider an ϵ of 0.05. Approximate the probability of interest ($P(|Y_{(1)} - 0| < \epsilon)$) using these 1000 simulated minimum values. (This is a Monte Carlo estimate of the probability.) Save this probability.

4. Repeat the above simulation and approximation of the probability of interest for $n = 2, 3, ..., 50$.

5. Create a plot with the sample size on the x-axis and the probability of interest on the y-axis. All created plots should have titles and appropriate axis labels.

6. In a comment explain how this plot can help someone understand convergence in probability to a constant.

7. You should also have the 1000 values of the minimum for each n saved from above. Create a plot with the sample size on the x-axis and the values of the minimum on the y-axis. (You should have 1000 values plotted above $n = 1$, 1000 plotted above $n = 2$, etc.). All created plots should have titles and appropriate axis labels.

8. In a comment explain how this plot can help someone understand convergence in probability to a constant.

   Note: The above R code to create the data sets can be done in a similar way to what was done on page 161 except here we are generating 1000 values for each sample size instead of 1. The code from page 163 is pretty similar in spirit to what you are trying to do here as well.

Visualizing Convergence in Distribution

This part will consider how well the Central Limit Theorem applies to sample means from Poisson data.

1. Consider a sample size of $n = 5$ from a $Poi(1)$ distribution.

   (a) Generate $N = 50000$ data sets of size n from the Poisson distribution.

   (b) For each data set, find the sample mean value (Hint: if you saved the above data in a large matrix the apply function or the colMeans or rowMeans functions can be handy here).

(c) Create a histogram of the sample means. Make the bins of appropriate width so that each bin only has one value of the support. For instance, the possible values for the sample mean here are 0, 0.2, 0.4, 0.6, ... Make sure that each bar only has one of these values included (so the bins would go from say -0.1 to 0.1, 0.1 to 0.3, 0.3 to 0.5, ...).

(d) The Central Limit Theorem says that $\bar{Y} \overset{\bullet}{\sim} N(\lambda, \lambda/n)$. Overlay this large-sample distribution on the histogram (hint: use freq = FALSE on you histogram and the curve function with add = TRUE to overlay the normal distribution). All plots should have appropriate titles and axis labels.

(e) Use the $N = 50000$ values to approximate the probability that $\bar{Y}$ is greater than or equal to $\lambda + 2 * \sqrt{\lambda/n}$. Also report this probability as approximated by the normal distribution (you can use a continuity correction if you'd like but that isn't required).

2. Repeat the above for $n = 10, 30$, and 100.

3. Repeat all of that for $\lambda = 5$ and $\lambda = 25$. You should have a total of 12 scenarios/plots.

4. Discuss how these plots and probabilities can help someone understand convergence in distribution.

5. Why do you think the large-sample approximation works better for larger λ values?

   Hint: Here are examples of plots that I am looking for (your plost won't look exactly like this due to random variation of course.)