R project for ST 501/601

Objective For this project you will be using **R** to simulate random data, approximate quantities,and create graphs. More specifically, 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 probabilitiesT

Submission Please submit your report as a single pdf file that include both the output and the relevant R code. The use of **Rmarkdown** is strongly encouraged (but it is not necessary). **Rmarkdown** is quite easy to use and will help you to generate nice reports that include both code, and prose, and output (for this project and for your future work). For more on **Rmarkdown**, see here.

Note It is ok to discuss this project with other students and/or the TA and course instructor. Just please do not share code between yourselves. This will be considered academic misconduct.

Part 1: Visualizing Convergence in Probability

This part will look at the behavior of the minimum order statistic from a random sample from an exponential distribution with rate parameter $\lambda = 1$. Proceed as follows.

- 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 $Y_{(1)}$. Start with the probability you want to show from the definition of convergence in probability to 0, i.e., $\mathbb{P}(|Y_{(1)} - 0| < \epsilon)$ and take the limit as n goes to infinity and show that this probability converges to 1.
- To visualize this we'll simulate data and approximate the probability statement proven in the previous part.
  - For a sample size of $n = 1$, generate $N = 1000$ data sets from an exp(1) distribution
  - For each data set, find the minimum value (for a sample of size 1 that will justbe the value itself)
  - Save these minimum values for plotting
- Now set $\epsilon = 0.05$. Next approximate the probability of interest $\mathbb{P}(|Y_{(1)} - 0| \leq \epsilon)$ using the $N = 1000$ simulated minimum values. (This is a Monte Carlo estimate of the probability.) Save this probability
- Repeat the above simulation and approximation of the probability of interest for $n = 2, 3, ..., 50$.
- Now create a plot with the sample size on the x-axis and the probability of interest on the y-axis. The plot should have an appropriate title and appropriate axis labels. In a comment explain how this plot can help someone understand convergence in probability to a constant.
- Now for each value of $n \in \{1, 5, 10, 25, 50\}$, draw one histogram of the minimum values for a sample of size n. You will thus have 5 histogram plots and, for example, the histogram plot for $n = 10$ will be a histogram plot of the $N = 1000$ minimum values for the $N = 1000$ samples of size $n = 10$. In a comment, explain how these histogram plots (for n changing), can help someone understand convergence in probability to a constant.

Note See also page 161 and page 163 of your guided notes for relevant codes that you can adapt to this problem.

Part 2: Visualizing Convergence in Distribution

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

- Consider a sample size of $n = 5$ from a Poisson distribution with rate parameter $\lambda = 1$.
  - Generate $N = 50000$ data sets of size n from the Poisson distribution.
  - For each data set, find the sample mean value. **Hint** If you saved the above data in a large matrix the `colMeans` or `rowMeans` functions can be handy here), e.g.,

```
n <- 5
N <- 50000
## A matrix of N = 50000 rows, each rows containing
## n = 5 sample from a Poisson distribution with rate 1
X <- matrix(rpois(n*N, lambda = 1), nrow = N)
rowMeans(X) ## This will give you a vector of N = 50000 sample means.
```

  - 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 $0, 0.2, 0.4, 0.6, \ldots$. 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. The use of the `breaks` argument for the histogram function `hist` will be helpful here.
  - The central limit theorem says that $\bar{X} \sim \mathcal{N}(\lambda, \frac{1}{n}\lambda)$ when $\bar{X}$ is the sample mean of n i.i.d Pois(λ) random variables. Overlay this large-sample distribution on the histogram (**Hint**: use `freq` = FALSE in your histogram and the `curve` function with `add` = TRUE to overlay the normal distribution). All plots should have appropriate titles and axis labels.
  - Use the $N = 50000$ values to approximate the probability that $\bar{X}$ is greater than or equal to $\lambda + \frac{2}{\sqrt{n}}\lambda$. Also report this probability as approximated by the normal distribution.
- Repeat the above for $n = 10$, $n = 30$, and $n = 100$.
- Repeat all of the above for $\lambda = 5$ and $\lambda = 25$. You should have a total of 12 scenarios/plots
- Discuss how these plots and probabilities can help someone understand convergence in distribution.
- Why do you think the large-sample approximation works better for larger λ values?

Note The following are examples of plots that we are looking for (your plost won't look exactly like this due to random variation)