BIMS 8701 Assignment 1: Statistics Review

Due February 19, 2024

A. This section is to help you get familiar with Bioconductor.

A.1. What version of Bioconductor are you using? (0.5 pt)

A.2. What version of BiocManager do you have installed? (0.5 pt)

A.3. How many total packages are available to be installed by BiocManager? Answer this question using `BiocManager::available` (1 pt)

A.4. Download the Illumina TruSeq adapters bundled with the trimmomatic software from here: https://github.com/timflutre/trimmomatic/blob/master/adapters/TruSeq3-PE-2.fa -- then, use the `Biostrings` package to read the file into R, and then count the frequency of each base for each sequence. (1 pt)

A.5. What is the most frequently used nucleotide in these sequences overall, and what is its overall frequency expressed as a percentage? (1 pt)

A.6. Compute the standard deviation across sequences of the frequency of each base. Which nucleotide has the highest standard deviation and what is it? (1 pt)

B. This section is to help you better understand the central limit theorem.

B.1. Simulate a standard normal sample with sample size n = 5, plot the sample and calculate the sample mean. (0.5 pt)

B.2. Repeat the simulation for one million times, plot the distribution of the one million sample means. (0.5 pt)

B.3. Now increase the sample size to 10. Repeat the above simulation and record the distribution of the one million sample means. (0.5 pt)

B.4. Now further increase the sample size to 20, 50, 100, and 1000. Repeat the above simulation. Show how the distribution of the one million sample means changes. (1 pt)

B.5. Now let’s change the simulation to an exponential sample. Set the rate lambda = 0.1, 0.5, and 1. Repeat B.1-B.4 and show the sample mean change under each parameter setting. (1 pt)

B.6. Now let’s change to a Poisson sample. Set lambda = 1, 5, and 10. Repeat B.1-B.4 and show the sample mean change under each parameter setting. (1 pt)

B.7. What can we conclude from this simulation study? (0.5 pt)

Hint:
1) You can use these R functions for simulation:

```
rnorm()
rexp()
rpois()
```

2) Set an appropriate bin size when plotting a histogram.