DSA3361 Inferential Data Analytics

Week 3 Workshop (Take Home Assignment 1)

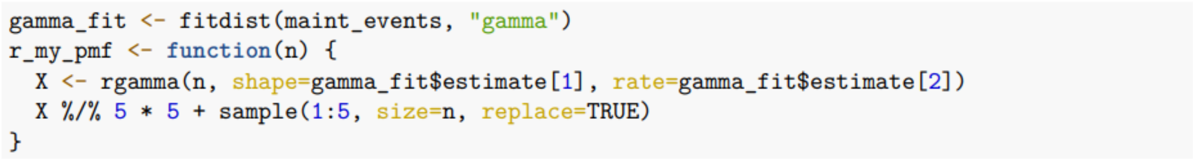
1. Recall that in the Dataset_FastFood.csv, we have the heights data and the weights data. We have done the distribution fitting to the height data in Workshop 1. In this section, we are going to apply what we have learned to the weights data.

   Firstly, we rename the weight data from the Dataset_FastFood.csv into weights. Plot a histogram (density plot) for the weights data.

   Next, we use fitdist() function to try to fit both the normal distribution and the gamma distribution to the weights data.

2. Referring to Q1, which of the following distribution do you think fits the weights better, normal distribution or gamma distribution? State your reason why.

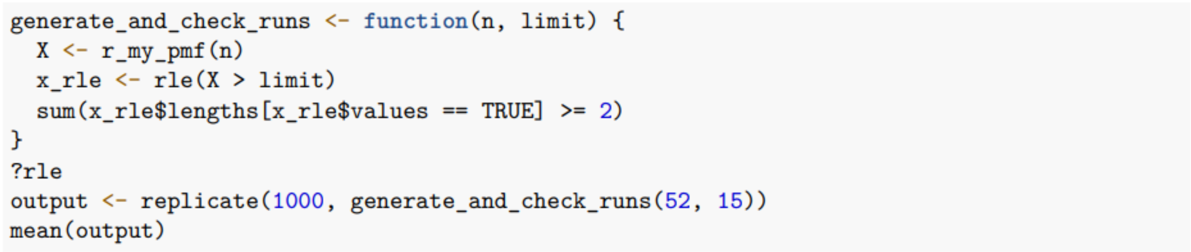
3. We tried gamma fit, a continuous distribution, to the numbers of SERIS weekly maintenance events, which is a discrete variable.



```
gamma_fit <- fitdist(maint_events, "gamma")
r_my_pmf <- function(n) {
  X <- rgamma(n, shape=gamma_fit$estimate[1], rate=gamma_fit$estimate[2])
  X %/% 5 * 5 + sample(1:5, size=n, replace=TRUE)
}
```

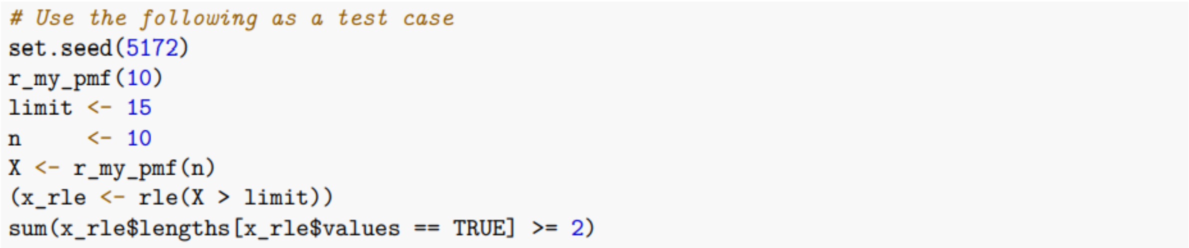
   Recall, the functions r_my_pmf roughly simulates a new sample of size n which follows a new distribution.

   Now, as a policy maker, you run the following function generate_and_check_runs and require to explain to all your fellow colleagues in a meeting. The function codes are as shown:



```
generate_and_check_runs <- function(n, limit) {
  X <- r_my_pmf(n)
  x_rle <- rle(X > limit)
  sum(x_rle$lengths[x_rle$values == TRUE] >= 2)
}
?rle
output <- replicate(1000, generate_and_check_runs(52, 15))
mean(output)
```

   Explain the mechanism behind calling this function. (Hint: you may use ?rle to find out the function.)



```
# Use the following as a test case
set.seed(5172)
r_my_pmf(10)
limit <- 15
n     <- 10
X <- r_my_pmf(n)
(x_rle <- rle(X > limit))
sum(x_rle$lengths[x_rle$values == TRUE] >= 2)
```

4. [Chihara & Hesterberg, 2019, Exercise 3.11]
   The file Phillies2009.csv contains data from 2009 season for the baseball team the Philadelphia Phillies.
   (a) Compare the empirical distribution functions of the number of strike-outs per game (StrikeOuts) for games played at home and games played away (Location).
   (b) Find the mean number of strike-outs per game for the home and the away games.
   (c) Perform a permutation test to see if the difference in means is statistically significant.

5. [Chihara & Hesterberg, 2019, Exercise 5.11]
   Import the data from Bangladesh.csv. In addition to arsenic concentration for 271 wells, the data set contains cobalt and chlorine concentrations.
   (a) Conduct Exploratory Data Analysis of the chlorine concentrations and describe the salient features.
   (b) Bootstrap the mean.
   (c) Find and interpret the 95% bootstrap percentile confidence interval.