STA237H1F Assignment #2 (Fall 2023) - Working with Probability Distributions in R

Lu-Wai Wong(*1008911108* & *LEC 0201)"

2023-11-13

Assignment #2 (both .Rmd & .pdf) - Due on Quercus 5 : 00*pm*, Fri Nov 24, 2023

Direct link to assignment - https://q.utoronto.ca/courses/316967/assignments/1184644

Graded out of 68 marks & worth 7.5% of your STA237H1F grade

NOTE: you must export *both* your completed R Markdown (i.e., rmd) file and your pdf file of your answers from U of T JupyterHub and save on your machine; then upload to Quercus.

NOTE - Save a copy of this rmd file as STA237A2yourname.rmd before you start editing it.

The best way to learn R is to experience coding yourself and to ask for support from the instructors or TAs in office hours if and when needed. In this assignment, you will again gain hands-on experience with RStudio and a reproducible workflow as you use R to simulate random experiments and use R functions for a variety of probability distributions that have been discussed in the course. This assignment must be completed independently so you will gain these skills and have the preparation to succeed in STA237H1F and later courses. You are strongly encouraged to start this assignment early and visit the instructor and/or TA office hours for support well before the deadline. Note that this assignment builds on Assignment 1. If you need a refresher on how to work with R Markdown and access, produce and export your assignment files from JupyterHub, please refer to https://q.utoronto.ca/courses/316967/pages/sta237h1f-assignment-number-1-introduction-to-rstudio-and-estimating-probabilities-via-simulation-due-5pm-oct-6 .

STA237H1F ASSIGNMENT 2 QUESTIONS (Fall 2023)

Answer each of the following questions with R code chunks and/or text, as appropriate.

Be sure to use `set.seed(type your student number)` ahead of every use of R functions that use a pseudo-random number generator (e.g., sample(), rbinom(), etc.) so you can answer the questions based on your results, and your simulated results will remain the same when you knit your assignment #2 rmd file to pdf.

QUESTION 1 (14 marks)

Use built-in R functions for common probability distributions to find each of the following values. Comment your code to describe your step(s).

(a) (2 marks) If a random variable Y follows a *beta distribution* with shape parameters $\alpha = 5$ and $\beta = 3$, find the median of Y.

```
set.seed(1008911108)
# 1(a) ANSWER:
# Median is the 50th percentile of the distribution
```

```r
# Calculate 50th percentile of Y
qbeta(0.5, 5, 3)
```

```
## [1] 0.6358839
```

(b) (2 marks) Simulate one observation from a *binomial distribution* with parameters 20 and 0.7.

```r
# set seed with your student number
set.seed(1008911108)

# 1(b) ANSWER:
rbinom(1, 20, 0.7)
```

```
## [1] 15
```

(c) (2 marks) If a random variable Y follows a *normal distribution* with a mean 70 and variance 64, compute $P\left(60 < Y < 75\right)$.

```r
# 1(c) ANSWER:

# Steps to Calculate P(60 < Y < 75)
# Calculate P(Y < 75) - P(Y < 60)
# Calculate P(Y < 75) using pnorm() function
less_than_75 <- pnorm(75, 70, 8)
# Calculate P(Y < 60) using pnorm() function
less_than_60 <- pnorm(60, 70, 8)

# Calculate P(60 < Y < 75)
less_than_75 - less_than_60
```

```
## [1] 0.6283647
```

(d) (2 marks) Let Y follow a *gamma distribution* with the shape and rate parameter 8 and 0.4 respectively. Find the 60^{th} percentile of Y.

```r
# 1(d) ANSWER:

# Calculate 60th percentile or, 0.6th quantile of Y
qgamma(0.6, 8, 0.4)
```

```
## [1] 20.97442
```

(e) (2 marks) Suppose a box contains 6 blue and 8 red marbles. If a child chooses three marbles together without looking, what is the probability the selected marbles contain at least two blue marbles?

```r
# 1(e) ANSWER:

phyper(1, 6, 8, 3, lower.tail = FALSE)
```

```
## [1] 0.3846154
```

(f) (2 marks) If a random variable Y follows a Pareto distribution with the $\alpha = 5$ and $\beta = 2$ respectively, calculate $P\left(Y < 6\right)$.

Recall from Week 7 tutorial that you first need the "EnvStats" R package to use the built-in Pareto R functions. The code to install this package is on line 18 of this rmd document but you will need to load the package in the R chunk below to use any of the Pareto R functions.

Note that the Pareto distribution was presented in tutorial with parameters α and β. The R function will be expecting "location" and "shape" parameters. We encourage you to access the R help documentation

for the Pareto distribution R function you are planning to use to confirm how the parameters should be entered. If they are entered in the wrong order, you will be working with a different Pareto distribution. To access the help documentation, you can type *help(R function name)* in the R console window. The relevant documentation is also available online at https://search.r-project.org/CRAN/refmans/EnvStats/html/Pareto.html.

```
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
# 1(f) ANSWER:

# Calculate cumulative probability of Y < 6
ppareto(6,2, 5)
```

```
## [1] 0.9958848
```

(g) (2 marks) If a random variable Y follows a Poisson distribution with mean 8, find the largest integer y_0 such that $P(Y \geq y_0) > 0.3$.

```
# 1(g) ANSWER:

# We are looking for the largest integer y0 such that P(Y >= y0) > 0.3
# 0.7th quantile of Y

qpois(0.7, 8)
```

```
## [1] 9
```

QUESTION 2 (14 marks)

Consider the continuous random variable Y with a pdf given by $f_Y(y) = \frac{\exp(-y)}{(1+\exp(-y))^2}$ for $-\infty < y < \infty$. Y is said to have a *standard logistic distribution.*

(a) (3 marks) Derive the cumulative distribution function (cdf) for the random variable Y. Clearly describe each of your steps with text or appropriate LaTeX code for mathematical expressions (refer to Assignment 1 for more information on LaTeX).
$2(a) ANSWER:$

The cdf of Y is given by: $F_Y(y) = \int_0^y f_Y(t)dt$

Thus, we want to find: $\int_0^y \frac{\exp(-t)}{(1+\exp(-t))^2}dt$

$\int_0^y \frac{\exp(-t)}{(1+\exp(-t))^2}dt = \frac{1}{1+e^{-y}} - \frac{1}{2}$

Thus, the CDF of $f_Y(y)$ is given by $F_Y(y) = \frac{1}{1+e^{-y}}$

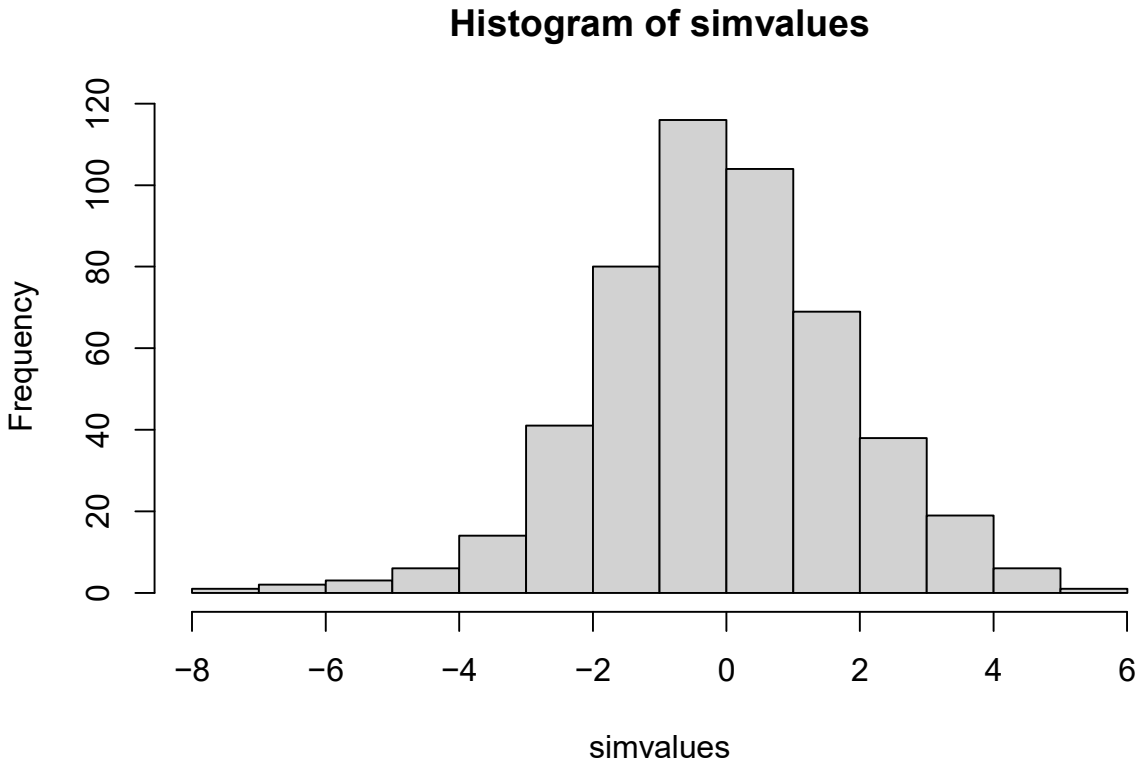
(b) (5 marks) *(i)* (2 marks) Briefly describe (in words) how to simulate an observation from a standard logistic distribution using the inverse transform method you learned in tutorial.
$2(bi) ANSWER:$

The inverse transform method is used to generate random variables from a given distribution, in this case, the logistic distribution.

First, use generate a set of numbers from a uniform distribution. Then, calculate the inverse of the CDF of the logistic distribution Finally, plug in the values from the uniform distribution into the inverse CDF Now, we have generated random variables from the logistic distribution.

(ii) (3 marks) Use the inverse transform method to generate *500* observations from the standard logistic distribution. Store these values in an R object called *simvalues* and plot your simulated values using the *hist()* R function. Comment your code to describe your step(s).

```
# 2(bii) ANSWER:
# set seed with your student number
set.seed(1008911108)

# Generate 500 random numbers from a uniform distribution
observations <- runif(500)

# Take values from uniform distribution and plug into inverse CDF of logistic distribution
simvalues <- -log((1-observations)/observations)

# Plot Histogram
hist(simvalues)
```



(c) (6 marks) *(i)* (2 marks) Compute the theoretical value of $P(Y < 1)$. Be sure to describe your steps.
$2(ci) ANSWER:$

We can find the theoretical value of $P(Y < 1)$ by plugging in 1 into the CDF of the logistic distribution.

$$P(Y < 1) = \frac{1}{1 + e^{-1}}$$
$$P(Y < 1) \approx 0.731$$

(ii) (2 marks) Use R to estimate $P(Y < 1)$ using the standard logistic distribution observations you simulated in *1(bii)*.

```r
# 2(cii) ANSWER:
# Copy standard logistic distribution code from 2(bii)
# set seed with your student number
set.seed(1008911108)

# Generate 500 random numbers from a uniform distribution
observations <- runif(500)

# Take values from uniform distribution and plug into inverse CDF of logistic distribution
simvalues <- -log(1/observations - 1)

# Calculate the proportion of values less than 1
# Add up all values less than 1 and divide by total number of observations
sum(simvalues < 1)/500
```

```
## [1] 0.734
```

(iii) (2 marks) Compare the theoretical value of $P(Y < 1)$ in *ci* to the estimated value of $P(Y < 1)$ in *cii*. If you simulated *5000* values from this distribution instead, what impact would this have on the difference between the theoretical and estimated probabilities. Justify your answer.
$2(ciii) ANSWER:$

The esitmated value was a little inaccurate in comparison to the estimated value for $P(Y < 1)$. If we simulated 5000 values from this distribution instead, the difference between the theoretical and estimated probabilities should be smaller. This is because the more values we simulate, the more accurate our estimate will be.

QUESTION 3 (14 marks)

Suppose the university is looking for *30* student representatives to serve on a variety of committees across campus. They plan to randomly select current students one at a time and send them an invitation. Suppose *75%* of students will agree to serve on a committee if invited. Let Y be the number of students the university will need to invite to recruit their target number of student representatives. *[Note: Since the number of students at the university is so large compared to the number of students they are looking for to serve on university committees, you may assume that sampling is done with replacement for this question.]*

(a) (4 marks) What probability distribution may be an appropriate model of Y? Justify your answer.
$3(a) ANSWER:$

The Negative Binomial Distribution may be an appropriate model of Y. this is because the Negative Binomial Distribution is used to model the probability of number of successes, given a number of trials of an experiment. This is the exact case presented in the question, where we want the probability of a certain number of students accepting, given a number of invitations sent out.

(b) (3 marks) Find *(i)* $E(Y)$, and *(ii)* $P(Y \leq a)$ where $a = E(Y)$. Justify your answer.

```r
# 3(b) ANSWER

# Calculate E(Y)
# E(Y) = r/p for negative binomial distribution
a <- 30/0.75
a
```

```
## [1] 40
```

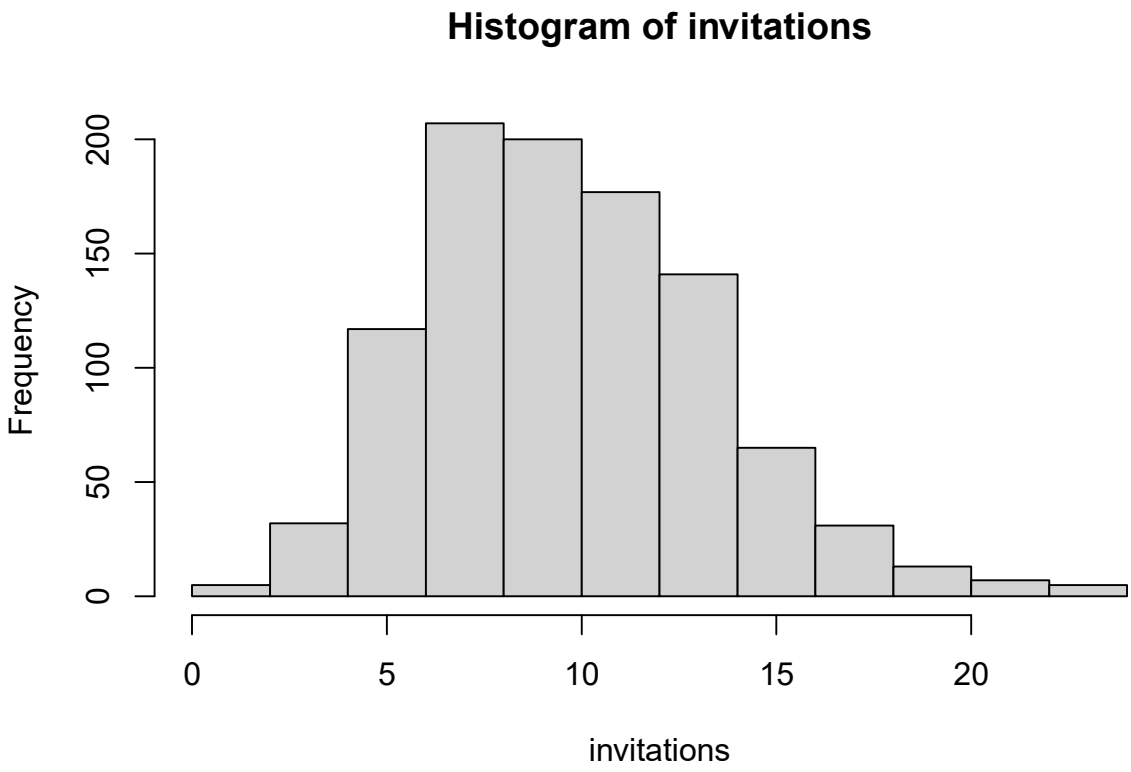
```r
# Calculate P(Y <= a)
# Want to calculate cumulative probability of Y <= a
```

```r
# Use pbinom() function
pnbinom(a-30, 30, 0.75)
```

```
## [1] 0.5839041
```

(c) (3 marks) Write R code to simulate 1000 repetitions of the random experiment described in this question. Save your simulated observations in an R vector called *invitations* and obtain of histogram of your simulated values. Comment your code to describe your step(s).

```r
# 3(c) ANSWER:
# set seed with your student number
set.seed(1008911108)

# Use rnbinom() function to simulate 1000 observations from negative binomial distribution
invitations <- rnbinom(1000, 30, 0.75)

# Plot histogram
hist(invitations)
```



(d) (4 marks) Estimate $E(Y)$ and $P(Y \leq a)$ where $a = E(Y)$ using the generated random observations in *3c*. Comment your code to describe your step(s). Are they close to the theoretical values you determined in part *3b*? Why or why not.

```r
# 3(d) ANSWER:

# Repeat code from 3(c) to simulate 1000 observations from negative binomial distribution
# set seed with your student number
set.seed(1008911108)

# Use rnbinom() function to simulate 1000 observations from negative binomial distribution
invitations <- rnbinom(1000, 30, 0.75)
```

```r
# Estimate E(Y)
mean(invitations)+30
```

```
## [1] 40.216
```

```r
# Estimate P(Y <= a)
sum (invitations <= 10)/1000
```

```
## [1] 0.561
```

The estimated values are pretty close to the theoretical values, although they are a little inaccurate. This is because we did a large number of simulations, which makes our estimate more accurate.

QUESTION 4 (14 marks)

(a) (7 marks) The time until the light in Savanna's office fails is *exponentially distributed* with mean *2 hours*.

(i) (2 marks) Find the probability that Savanna's light survives more than three hours. Show your steps.
$4a(i) ANSWER:$

We want to find $P(Y > 3)$ where Y is the time until the light in Savanna's office fails. An exponential distribution is given by $f_Y(y) = \lambda e^{-\lambda y}$ where λ is 1/mean. Thus, the probability of the time until the light in Savanna's office fails is given by:

$$\begin{aligned} P(Y > 3) &= \int_3^\infty \lambda e^{-\lambda y} dy \\ &= -e^{-\lambda y}\Big|_3^\infty \\ &= -e^{-\lambda \infty} + e^{-\lambda 3} \\ &= 0 + e^{-\frac{3}{2}} \\ &= e^{-\frac{3}{2}} \\ &\approx 0.223 \end{aligned}$$

(ii) (3 marks) Simulate 2000 observations from this exponential distribution and estimate probability that was computed in *5ai*. Comment your code to describe your steps.

```r
# 4a(ii) ANSWER:
# set seed with your student number
set.seed(1008911108)

# Using rexp() function to simulate 2000 observations from exponential distribution
observations <- rexp(2000, 1/2)

# Calculate the proportion of values greater than 3
# Add up all values greater than 3 and divide by total number of observations
sum(observations > 3)/2000
```

```
## [1] 0.227
```

(iii) (2 marks) Estimate the mean of this exponential distribution using your simulated observations. Comment your code to describe your steps.

```r
# 4a(iii) ANSWER:
set.seed(1008911108)

# Use same code from 4a(ii) to simulate 2000 observations from exponential distribution
```

```r
observations <- rexp(2000, 1/2)

# Calculate mean of observations
mean(observations)
```

```
## [1] 1.997136
```

(b) (7 marks) The time until the computer crashes in Savanna's office is *exponentially distributed* with mean *3 hours*. Suppose failure of the light and crash of the computer times in Savanna's office are independent.

(i) (3 marks) Find the probability that neither the light nor the computer fails in the next 3 hours. Show your steps.
4b(i)ANSWER :

By definition of independence,

$$P(X \geq a, Y \geq a) = P(X \geq a)P(Y \geq a)$$

Let X represent the time until Savanna's light fails, and Y represent the time until Savanna's computer crashes.

$$\begin{aligned} P(X \geq a) &\approx 0.223 \\ P(Y \geq a) &= \int_3^\infty \frac{1}{3} e^{-\frac{1}{3}y} \\ &= e^{-1} \\ &\approx 0.367 \end{aligned}$$

Thus,

$$P(X \geq a, Y \geq a) = 0.223 \times 0.367 = 0.082$$

(ii) (4 marks) Estimate the probability that was computed in part *5bi* by randomly generating 2000 light failures and computer crashes and report your estimated probability. Comment your code to describe your steps.

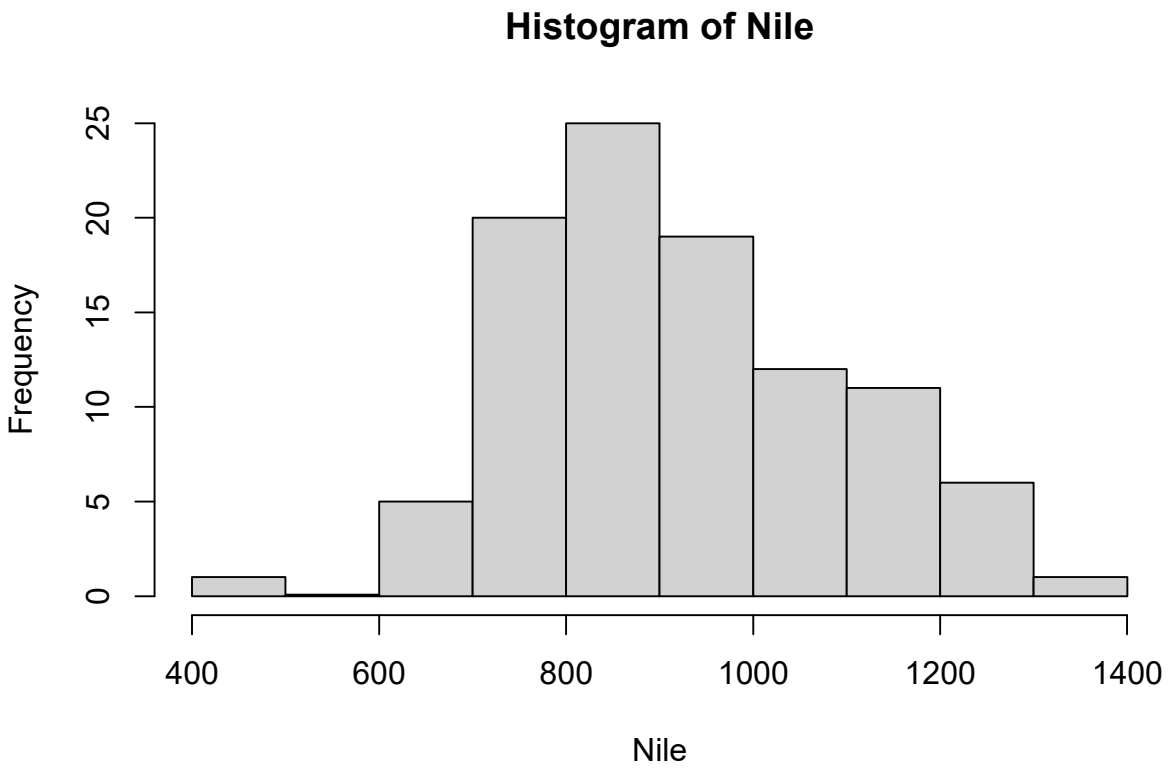
```r
# 4b(ii) ANSWER:
# set seed with your student number
set.seed(1008911108)

lamp <- rexp(2000, 1/2)
crashes<- rexp(2000, 1/3)

# Calculate the proportion of values greater than 3 for both lamp and crashes
lamp_prob <- sum(lamp > 3)/2000
crashes_prob <- sum(crashes > 3)/2000
lamp_prob * crashes_prob
```

```
## [1] 0.082401
```

QUESTION 5 (8 marks)

Consider the R data set "Nile" that contains measurements of the annual flow of the river Nile (in 10^8 m^2) at Aswan. The data set consists of 100 measurements and the following code produces a histogram of these data.

```r
hist(Nile)
```

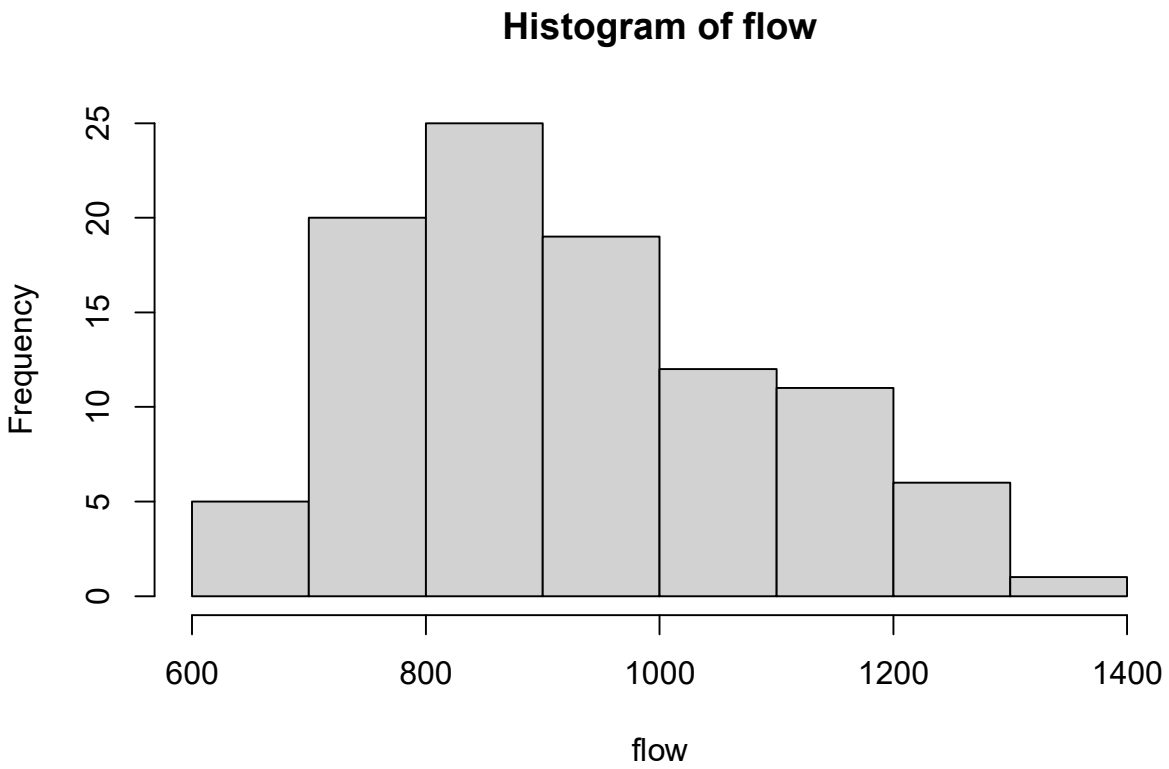


You can see that one of measurements (456 10^8 m^2) is an unusual observation in the data set. We will exclude this measurement from the data set using the R code below and save the 99 measurements we will work with in this question in the vector *flow*.

```
flow<-Nile[Nile > 600]
```

(a) (4 marks) *(i)* Construct a histogram for the *99* annual flow measurements of the Nile and comment on the shape of the distribution of flow measurements. *(ii)* Do these data appear to follow a normal distribution? Justify your answer based on appropriate plots.

```
# 5(a) ANSWER:
hist(flow)
```



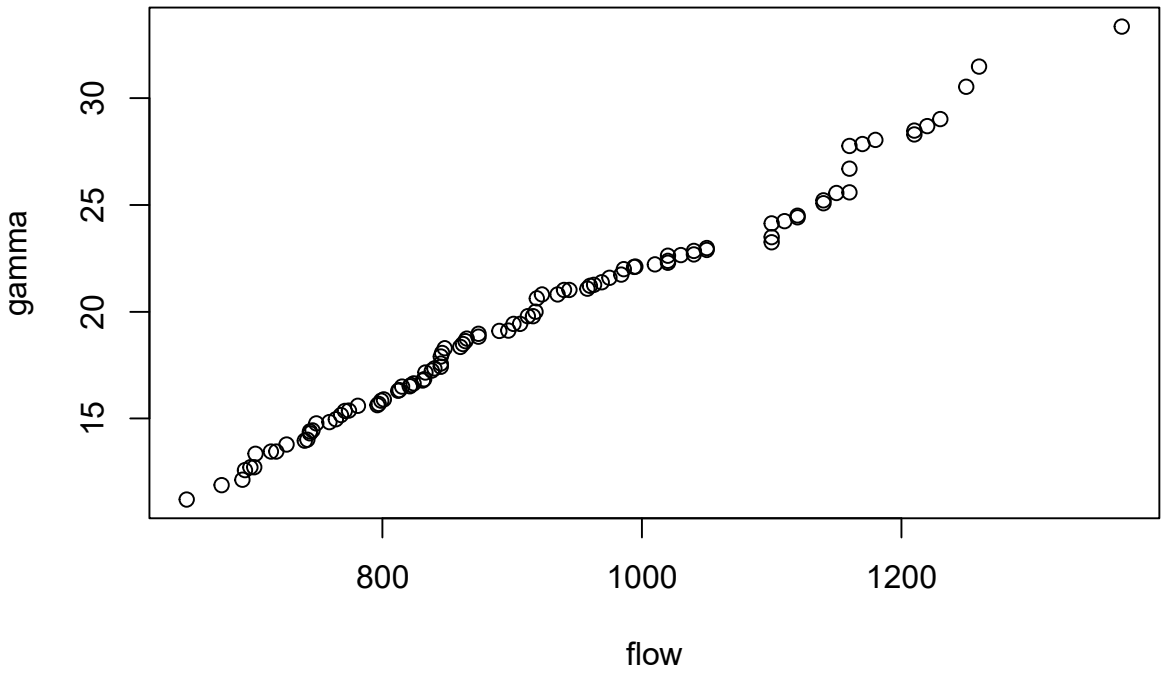
No, this data does not appear to follow a normal distribution. The data appears to be skewed to the right, which a normal distribution does not do.

(b) (2 marks) Which of the common probability distributions we've discussed in the course may be an appropriate model for these flow measurements? Briefly explain your reasoning.
$5(b) ANSWER:$

The gamma distribution may be an appropriate model for these flow measurements. A gamma distribution with the corrent shape and rate parameters could be made to resemble the shape of the flow histogram.

(c) (2 marks) Create a gamma Q-Q plot using the parameters shape parameter *20* and rate parameter *1*. Do these annual flow measurements appear to follow this gamma distribution? Justify your answer.

```
# 5(c) ANSWER:
set.seed(1008911108)
# Create gamma plot
gamma = rgamma(99, 20, 1)

# qqplot function in r package
qqplot(flow, gamma)
```



The annual flow measurements vaguely appears to follow the gamma distribution. The points do not perfectly follow a straight diagonal line, but they are close enough to resemble a gamma distribution. It deviates at some points, mostly at the end of the distribution, but it does resemble a gamma distribution.

ASSIGNMENT REPRODUCIBILITY (4 marks)

Your assignment #2 file submission in the Quercus Assignment must include *both* the rmd file with your assignment #2 answers that was compiled (or *knitted*) to produce a pdf file of your assignment #1 answers. # Rubric:

- 0/4 marks - submitted rmd file did not produce submitted pdf file
- 1/4 marks - no rmd file submitted, or rmd failed to knit.
- 1/4 marks - number other than your student number used in 'set.seed()' or seed not set in most R code chunks with randomness
- 3/4 marks - seed set in some, but not all, R code chunks with randomness.
- 4/4 marks - both your pdf file and rmd file used to produce your pdf file submitted

---

THIS IS THE END OF STA237H1F ASSIGNMENT #2

```
## [1] 0.3036931
```

```
## [1] "Fri Nov 24 21:19:02 2023"
```