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

Kevin (Qifan Hu) (1008866817 LEC 0101)

2023-11-13

Assignment #2 (both .Rmd & .pdf) - Due on Quercus $5:00pm$, 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.

```r
# 1(a) ANSWER:
# median of beta distribution with shape parameters alpha = 5 and beta = 3
qbeta(0.5, shape1 = 5, shape2 = 3)
```

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

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

```
# set seed with your student number
# 1(b) ANSWER:
set.seed(1008866817) # set seed
# simulate one observation of binomial distribution with parameters 20 and 0.7
rbinom(1, size = 20, prob = 0.7)
```

```
## [1] 14
```

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

```
# 1(c) ANSWER:
standard_deviation <- sqrt(64) # sd is sqrt of var
less_than_75 <- pnorm(75, mean = 70, sd = standard_deviation) # calculate P(Y < 75)
less_than_60 <- pnorm(60, mean = 70, sd = standard_deviation) # calculate P(Y < 60)
less_than_75 - less_than_60 # calculate P(60 < Y < 75)
```

```
## [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.

```
# 1(d) ANSWER:
# 60th percentile of gamma distribution with shape 8 and rate 0.4
qgamma(0.6, shape = 8, rate = 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?

```
# 1(e) ANSWER:
# since the child chooses three marbles, there will be no replacement,
# thus we can use hypergeometric distribution to solve this problem
# calculate P(X >= 2), X is the number of blue marbles
phyper(1, m = 6, n = 8, k = 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.

```
# 1(f) ANSWER:
library(EnvStats) # load EnvStats package to ppareto()
```

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

```r
# calculate P(Y < 6), according to documentation, location is beta and shape is alpha
ppareto(6, location = 2, shape = 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 $P(Y \geq y_0) > 0.3$.

```r
# 1(g) ANSWER:
# find largest integer y0 such that P(Y >= y0) > 0.3
y0 <- 0 # initialize y0, start with 0
# iterate until P(Y >= y0) > 0.3 is false
while (ppois(y0, lambda = 8, lower.tail = FALSE) > 0.3) {
  y0 <- y0 + 1 # increment y0 by 1
}
y0
```

```
## [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:

$F_Y(y) = \int_{-\infty}^{y} f_Y(y)dy = \int_{-\infty}^{y} \frac{\exp(-y)}{(1 + \exp(-y))^2}dy = -\frac{1}{1+\exp(-y)} + c$

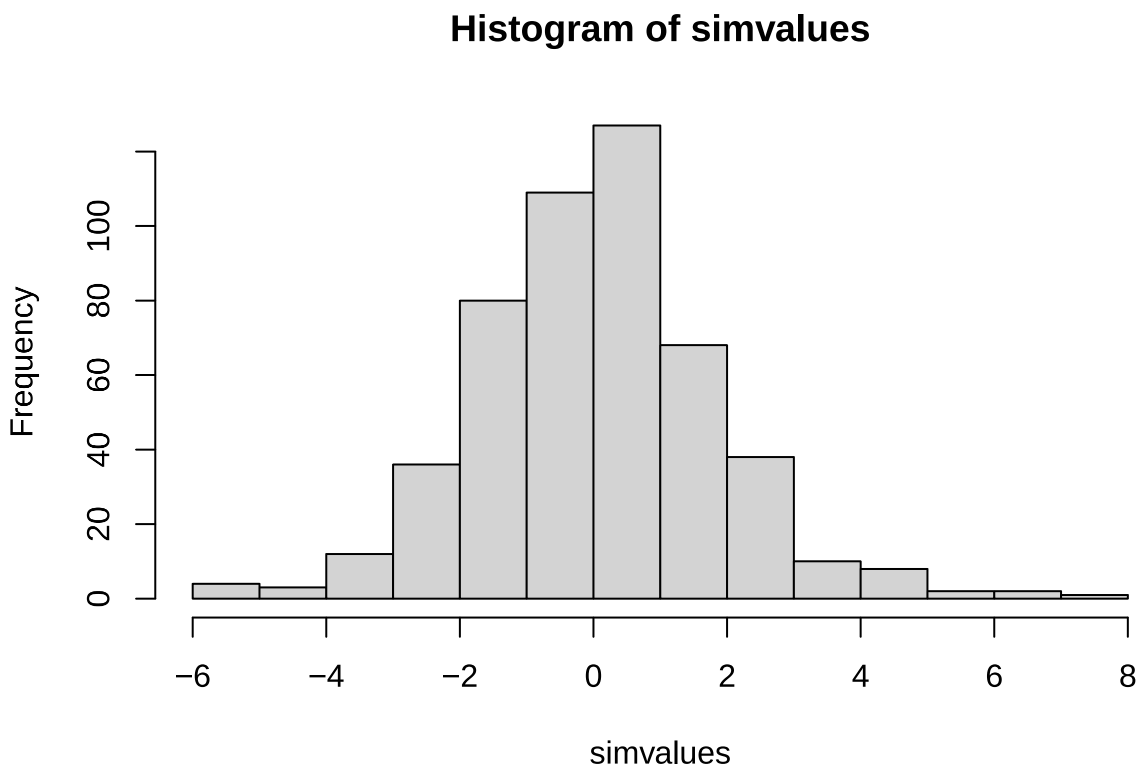
Since $F_Y(y)$ is a cdf, we know that $\lim_{y \to -\infty} F_Y(y) = 0$ and $\lim_{y \to \infty} F_Y(y) = 1$.

Thus we can solve for c and get $c = 1$. Then, the cdf $F_Y(y) = 1 - \frac{1}{1+\exp(-y)} = \frac{\exp(y)}{1+\exp(y)}$

(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: According to the inverse transform method, we know that for $U\ Unif(0, 1)$, $F^{-1}(U) = F(X)$ (The unified distribution of $F^{-1}(U)$ is equal to the distribution of X). So we can use the inverse of the cdf F to generate observations from a standard logistic distribution. In this question, we can use the inverse, $F^{-1}(U) = \log(\frac{U}{1-U})$ to generate observations from $F(X)$.

(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).

```r
# 2(bii) ANSWER:
# set seed with your student number
set.seed(1008866817)
U <- runif(500) # generate 500 random numbers from 0 to 1 in uniform distribution
simvalues <- log(U / (1 - U)) # 500 observations using inverse transform method
hist(simvalues) # plot simvalues using hist()
```



(c) (6 marks) *(i)* (2 marks) Compute the theoretical value of $P(Y < 1)$. Be sure to describe your steps. 2(ci) ANSWER: Since we know the cdf of Y is $F_Y(y)$ from 2(a), we can calculate $P(Y < 1)$ by plugging in $y = 1$ into $F_Y(y)$.

$P(Y < 1) = F_Y(1) = \frac{\exp(1)}{1+\exp(1)} = \frac{\exp(1)}{\exp(1)+\exp(0)} = \frac{\exp(1)}{\exp(1)+1} = 0.7311$

Thus, the theoretical value of $P(Y < 1)$ is 0.7311

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

```
# 2(cii) ANSWER:
sum(simvalues < 1) / 500 # estimate P(Y < 1)
```

```
## [1] 0.742
```

(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 results in ci and cii are very similar. If I were to simulate 5000 values instead, the result in cii will be closer to the theoretical value calculated in ci since we are going through 10 times more samples than what we have done.

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:We can use negative binomial distribution for this model. In this question, we are looking for the number of students needed to invite to get 30 students to serve on a committee. The negative binomial distribution looks for the number of trials needed to get a certain number of successes, which is suitable in this situation.

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

```r
# 3(b) ANSWER
# (i) E(Y)
mean <- 30 / 0.75 # mean of negative binomial distribution = r / p
mean
```

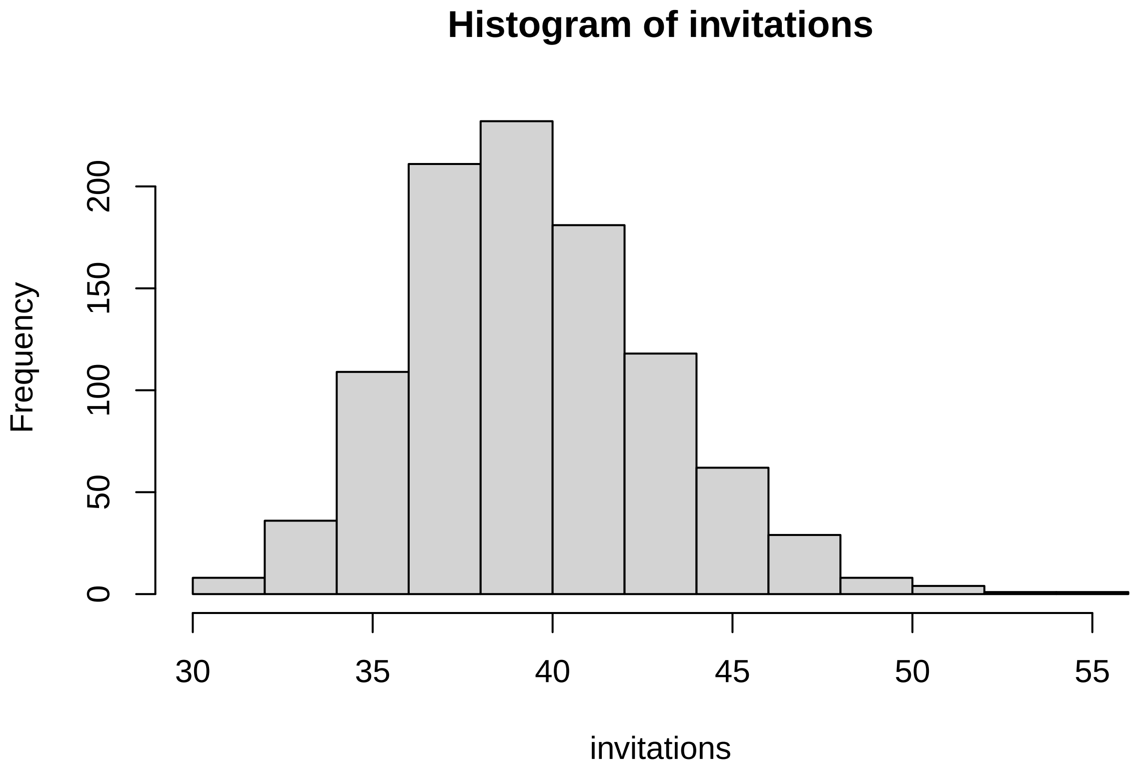
```
## [1] 40
```

```r
# (ii) P(Y <= a) = P(Y <= E(Y))
pnbinom(mean - 30, size = 30, prob = 0.75) # P(Y <= E(Y))
```

```
## [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(1008866817)
# 1000 observations from binomial distribution
# Note: we add 30 to each observation to the result because rbinom()
# generates the number of failures before the first success, not the number of trials
invitations <- rbinom(1000, size = 30, prob = 0.75) + 30
hist(invitations) # histogram of 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:
mean(invitations) # estimate E(Y)
```

```
## [1] 39.973
```

```r
sum(invitations <= mean(invitations)) / 1000 # estimate P(Y <= E(Y))
```

```
## [1] 0.473
```

3(d) ANSWER: They are close to the theoretical values we calculated in 3b because we are using 1000 observations to estimate the values, which is a large enough sample size to get close to the theoretical values. And if we increase the number of observations, the estimates will be even closer to the theoretical values.

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:

Let X be the time until the light fails.

Then $X \sim Exp(\lambda)$, where $\lambda = 1/2$ since the mean time between failure is 2 hours.

The cdf of exponentially distributed X is $F(x) = 1 - \exp(-\lambda x)$.

so $P(X > 3) = 1 - P(X <= 3) = 1 - F(3) = 1 - (1 - \exp(-3/2)) = \exp(-3/2) = 0.2231$

Thus, the probability that Savanna's light survives more than three hours is 0.2231.

(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(1008866817)
rate_value <- 1 / 2 # rate is 1 / mean
# 2000 observations from exponential distribution
light_fail <- rexp(2000, rate = rate_value)
sum(light_fail > 3) / 2000 # P(light > 3)
```

```
## [1] 0.209
```

(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:
mean(light_fail) # mean of exponential distribution simulated in 4a(ii)
```

```
## [1] 1.946743
```

(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:

Let Y be the time until the computer crashes.

Then $Y \sim Exp(\lambda)$, where $\lambda = 1/3$ since the mean time between failure is 3 hours.

The cdf of exponentially distributed Y is $F(y) = 1 - \exp(-\lambda y)$.

so $P(Y > 3) = 1 - P(Y <= 3) = 1 - F(3) = 1 - (1 - \exp(-3/3)) = \exp(-3/3) = 0.3679$

We can know that P(X > 3) = 0.2231 from 4a(i).

Since the failure of the light and crash of the computer times in Savanna's office are independent, we can multiply their probability that neither the light nor the computer fails in the next 3 hours.

Thus, the probability that neither the light nor the computer fails in the next 3 hours is 0.3679 * 0.2231 = 0.0821.

(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.

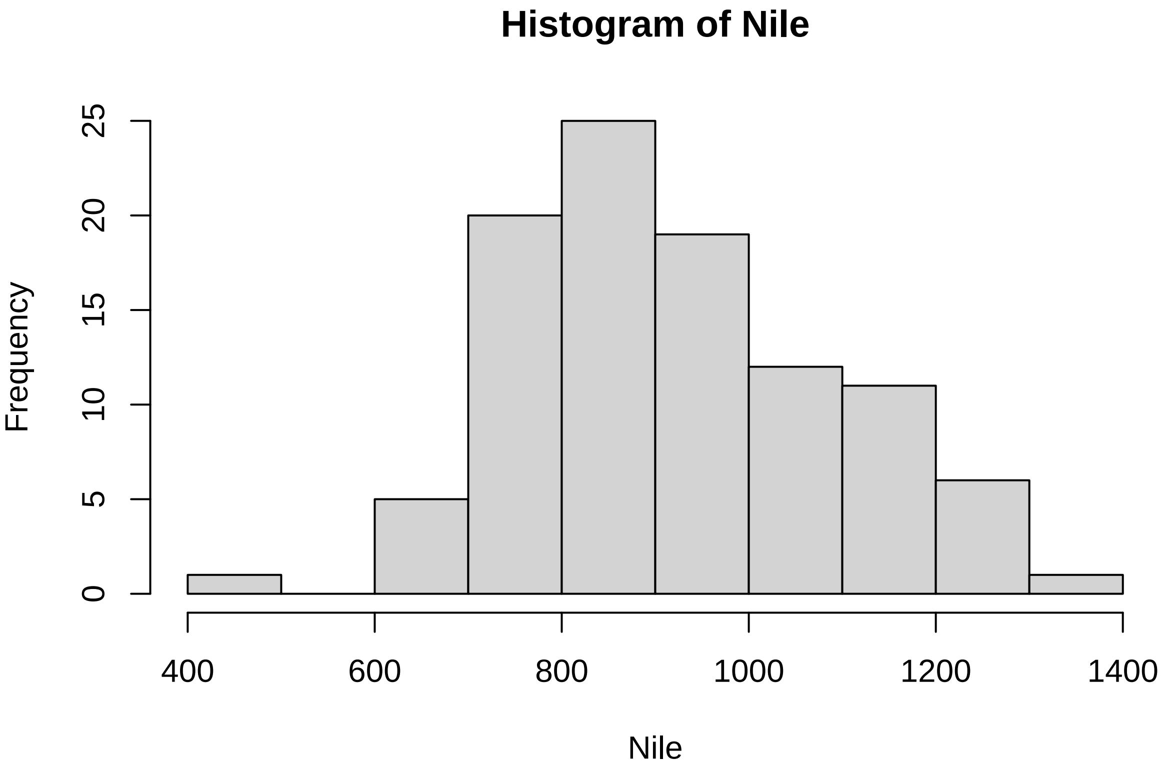
```
# 4b(ii) ANSWER:
# set seed with your student number
set.seed(1008866817)
rate_value_computer <- 1 / 3 # rate is 1 / mean
# 2000 observations from exponential distribution of light and computer
computer_crash <- rexp(2000, rate = rate_value_computer)
light_failure <- rexp(2000, rate = rate_value) # rate_value from 4a(ii)
# both computer and light survive more than 3 hours
no_fail <- computer_crash > 3 & light_failure > 3
sum(no_fail) / 2000 # P(light and computer > 3)
```

```
## [1] 0.076
```

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.

```
hist(Nile)
```

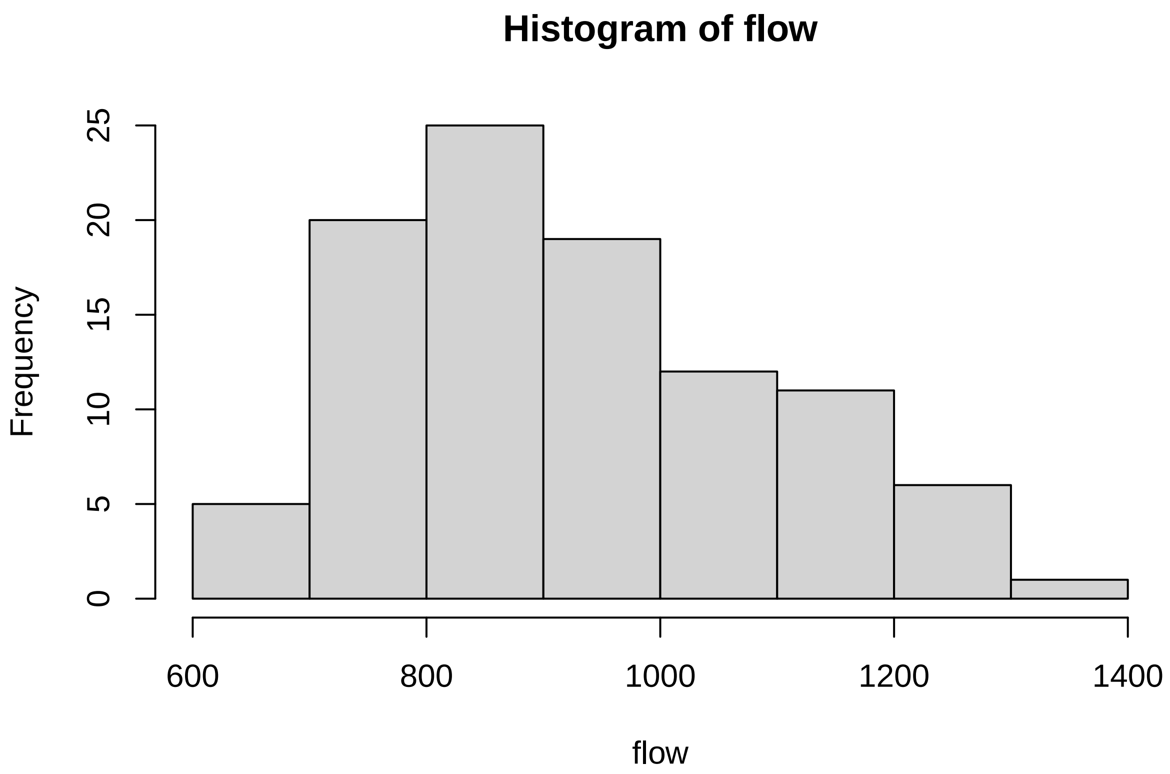


You can see that one of measurements (456 10^8 m^2) is an unusual observation in the data set. We will exclude this measurement from the data set using the R code below and save the 99 measurements we will work with in this question in the vector *flow*.

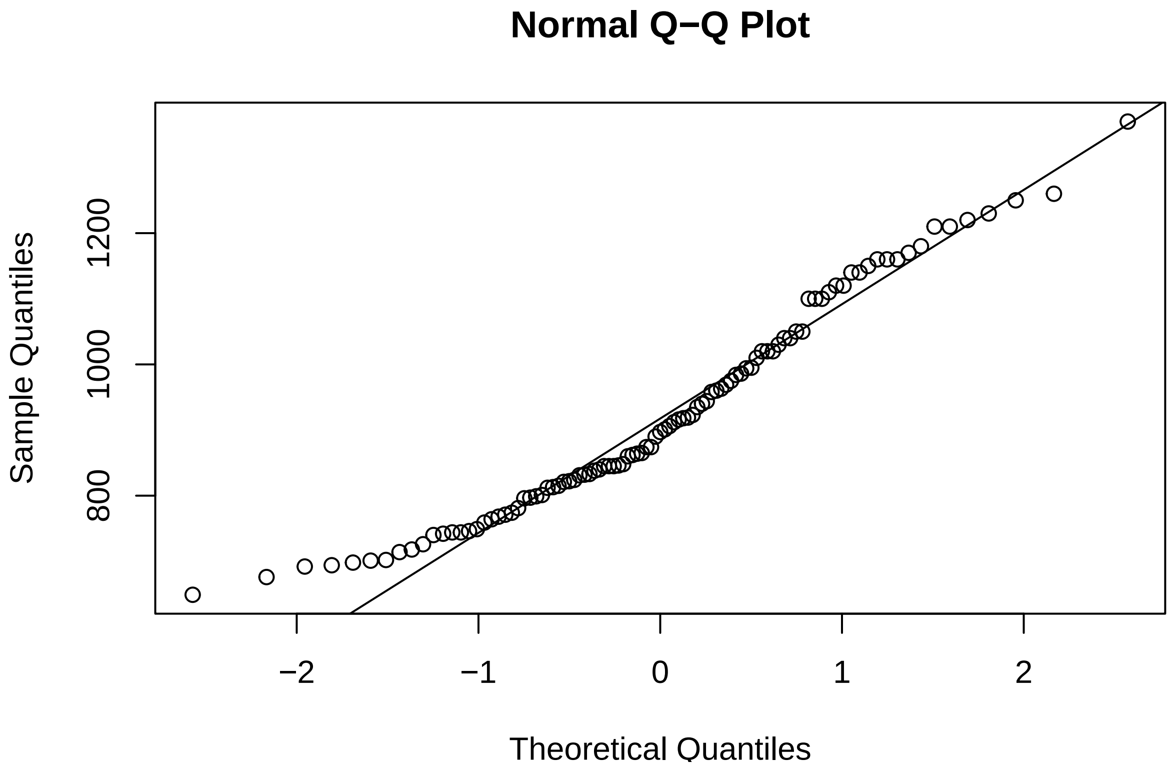
```
flow<-Nile[Nile > 600]
```

(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:
# (i)
hist(flow) # histogram of flow
```



```
# (ii)
qqnorm(flow)
qqline(flow) # a reference line
```



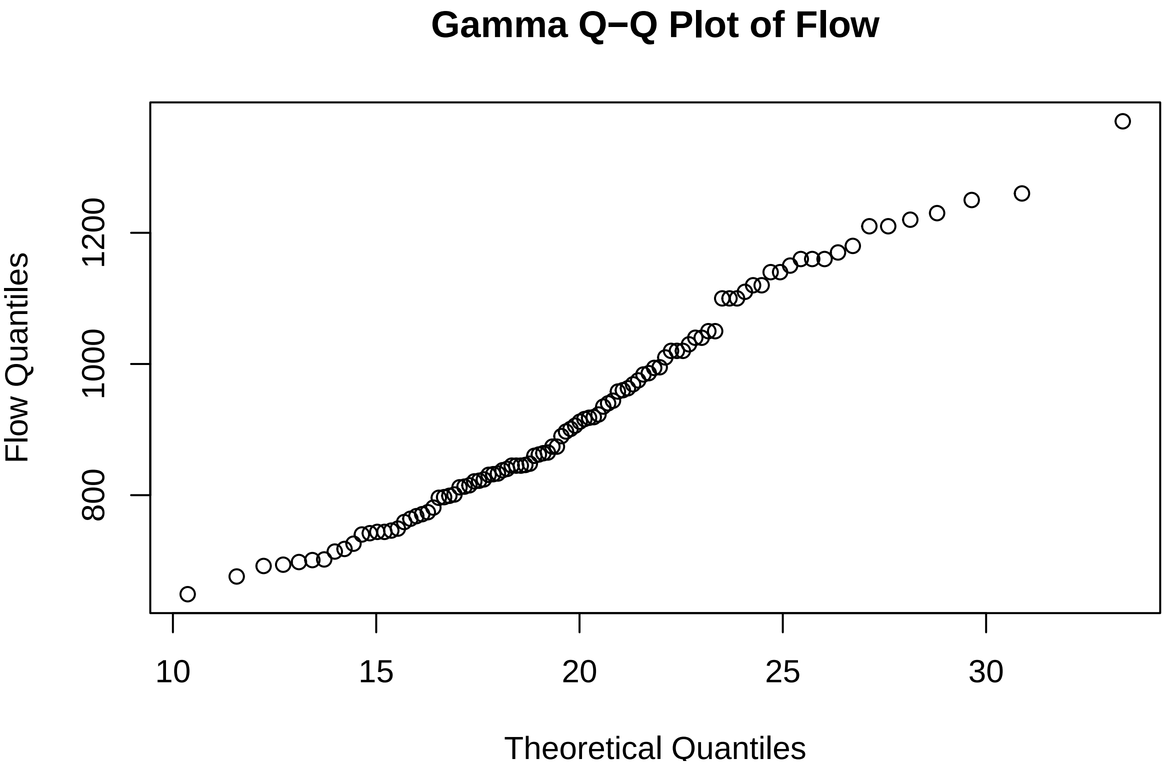
5a(ii) ANSWER: The data does not appear to follow a normal distribution. The data in the histogram does not seem symmetric, and we can also see that by looking at the plot, the data does not follow the straight normal distribution line, which does not follow the characteristic of a normal distribution.

(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: Gamma distribution may be an appropriate model for the flow measurements.

We can see that the histogram of the flow measurements is right-skewed, which is similar to the shape of a gamma distribution. Another reason for choosing gamma distribution is that the gamma distribution is often used to model waiting times, rainfall, and other physical processes. Thus, gamma distribution may be an appropriate model for these flow measurements.

(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.

```r
# 5(c) ANSWER:
# use qgamma() to generate theoretical quantiles and plot them against flow
qqplot(qgamma(ppoints(99), shape = 20, rate = 1), flow,
       xlab = "Theoretical Quantiles", ylab = "Flow Quantiles",
       main = "Gamma Q-Q Plot of Flow")
```



5(c) ANSWER: The annual flow measurements appear to mostly follow this gamma distribution. We can see that the points on the qqplot are close to a straight line, which means the flow data points are close to the theoretical quantiles of the 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] 0.1527134
```

```
## [1] "Sat Nov 25 00:44:10 2023"
```