Lab 3 (100 pts.) - Assessing the Normality of Data
Objectives: Creating and Interpreting Normal Probability Plots (QQ plots)

In addition to the regular graphs, code, and interpretations, please submit the data for parts B, C and D ONLY (clearly labeled) **at the end of the lab report** in an Appendix to your individual submission. Make sure each section is clearly labeled. Please DO NOT include the data for parts E or F. Remember that you are only supposed to include information that is asked for. In this lab, we are not interested in the mean and standard deviations of the distributions.

A. (10 points) Online Prelab

B. (10 points) Normal random numbers (no data file required) Use software to generate 10 observations from a normal distribution with mean, μ = 9 and standard deviation, σ = 4.5.

(You must submit your data for this question. No credit will be given without data.)

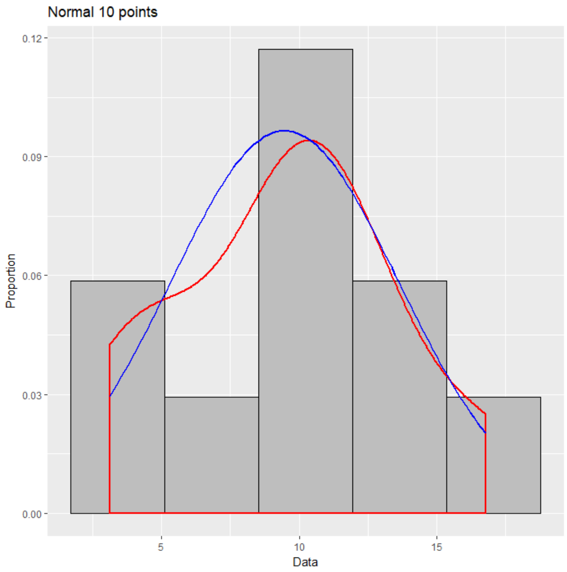
1. (3 pts.) Code:

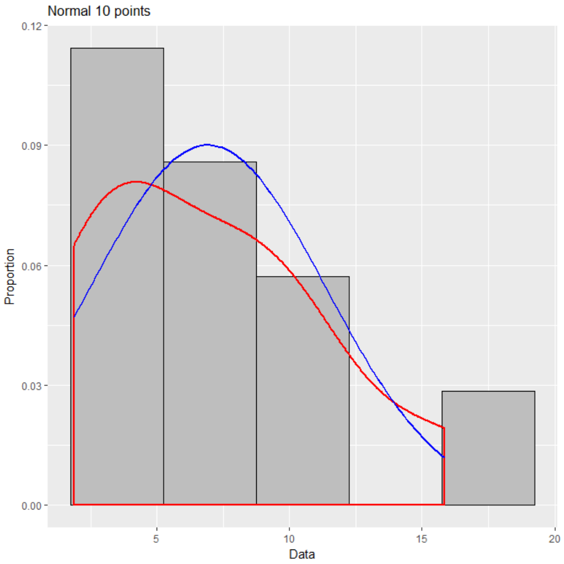
Solution:

```
#========
# Part B
#========
RandomData <- rnorm(10,mean=9,sd=4.5)
RandomData
title <- "Normal 10 points"
#---------------
# HISTOGRAM
#---------------
# Obtain xbar and s, needed for theoretical curve
xbar <- mean(RandomData)
s <- sd(RandomData)
windows()
ggplot(data.frame(RandomData=RandomData), aes(x=RandomData)) +
  geom_histogram(aes(y=..density..), bins = sqrt(length(RandomData))+2,
                 fill = "grey",col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun=dnorm, args=list(mean=xbar, sd=s), col="blue",
                lwd = 1) +
  ggtitle(title) +
  xlab("Data") +
  ylab("Proportion")
#
# QQ plot
#
windows()
ggplot(data.frame(RandomData=RandomData), aes(sample=RandomData)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle(title)
```

2. (3 pts.) Make a histogram of these observations. How does the shape of the histogram compare with a normal density curve?

Solution:





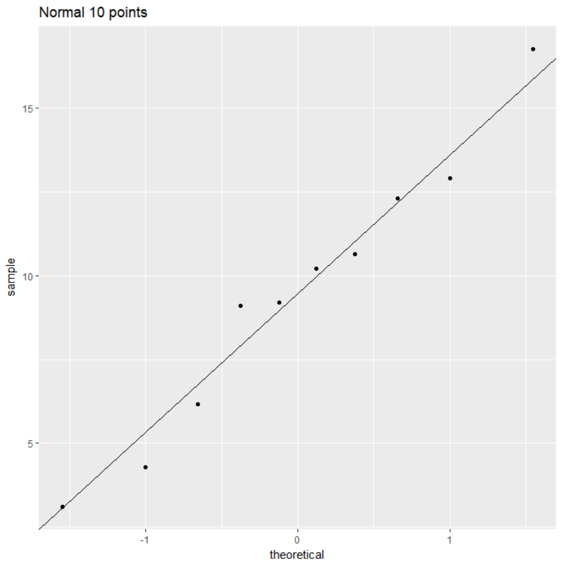
Please note that your histogram should include both the normal distribution curve and the smoothed distribution curve.

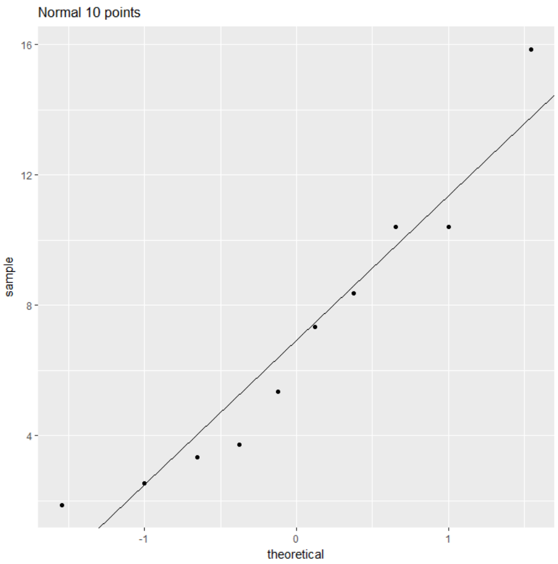
The answer to this question may vary depending on the actual shape of the histogram. I ran the code twice and generated two different histograms. The first distribution looks close to the normal distribution which can be seen via the red and blue curves and the shape although the histogram does not look exactly symmetric. The second distribution does not look normal.

You need to be aware that since the number of observations is very small, it is possible that the distribution is close to normal like the first distribution or far from normal like the second one.

3. (4 pts.) Make a normal probability plot of the data. Does the plot suggest any important deviations from normality? Please provide specifics to explain your answer.

Solution:





For this part, again, I am showing two graphs corresponding to the two histograms from 2. The first one is normal since the points are randomly above and below the line. The second one deviates from normal since the points are not randomly above and below the line.

C. (10 points) Normal random numbers (no data file required) Use software to generate 100 observations from a normal distribution with mean, μ = 9 and standard deviation, σ = 4.5.

(You must submit your data for this question. No credit will be given without data.)

1. (2 pts.) Code:

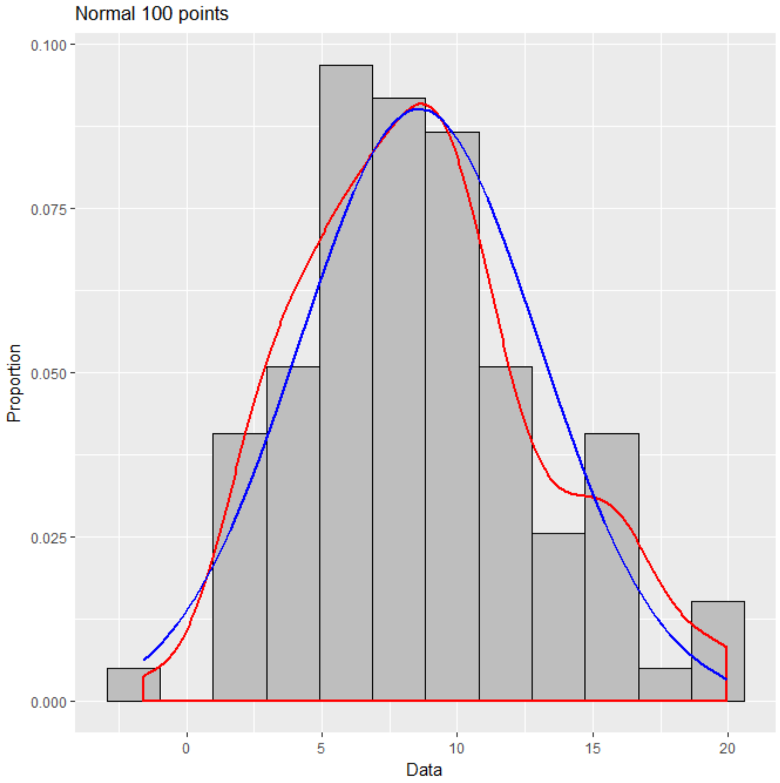
Solution:

```
#========
# Part C
#========
RandomData <- rnorm(100,mean=9,sd=4.5)
RandomData
title <- "Normal 100 points"
```

```
#---------------
# HISTOGRAM
#---------------
# Obtain xbar and s, needed for theoretical curve
xbar <- mean(RandomData)
s <- sd(RandomData)
windows()
ggplot(data.frame(RandomData=RandomData), aes(x=RandomData)) +
  geom_histogram(aes(y=..density..), bins = sqrt(length(RandomData))+2,
               fill = "grey",col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun=dnorm,  args=list(mean=xbar, sd=s), col="blue",
              lwd = 1) +
  ggtitle(title) +
  xlab("Data") +
  ylab("Proportion")
#
# QQ plot
#
windows()
ggplot(data.frame(RandomData=RandomData), aes(sample=RandomData)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle(title)
```

2. (3 pts.) Make a histogram of these observations. How does the shape of the histogram compare with a normal density curve?

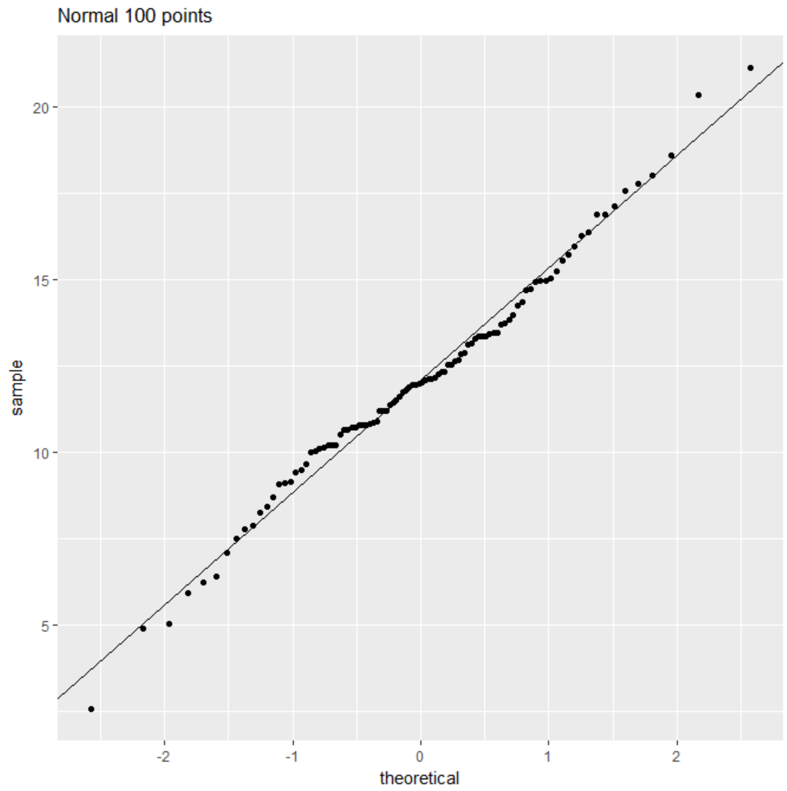
Solution:



Note that in a number of the generated distributions in R like this one, even though the shape of the distribution is close to symmetric, the two curves do not look close. Therefore, it is hard to tell if the distribution is normal or not.

3. (3 pts.) Make a normal probability plot of the data. Does the plot suggest any important deviations from normality? Please provide specifics to explain your answer.

Solution:



Most points follow the line closely and the rest are randomly above and below the line. Therefore, there is no reason to suggest that there are important deviations from normality.

Therefore, I would say that this distribution is approximately normal. Therefore, even though we plot both the QQ plot and the histogram, the QQ plot often gives a better guide as to whether the distribution is normal or not. Please keep this in mind when determining normality in the future.

4. (2 pts.) Compare and contrast the plots from part B and part C. Remember both of these parts are from the same normal distribution.

Solution:

The plots from B and C are the same in that they are produced from the same parent distribution. They will differ because they are produced from two different datasets. Most important, the large size of the second dataset provides it with a better sense of normality as visualized with the histogram and the probability plot. In fact, the plots from A don't necessarily look like they come from a normal distribution at all.

D. (40 points) Random numbers from other distributions (no data file required)

Use software to generate 100 observations from the following distributions (10 points for each of the distributions) (I) right skewed (Exponential), (II) left skewed (Beta), (III) short tailed (Uniform) and (IV) long tailed (t-distribution). **The code for how to generate the functions is in the tutorial.** You are required to use the exact code given in the tutorials including the parameters. The format of the lab report should be: a) a section title which indicates the distribution under consideration (I, II, III, or IV), b) the code for that section c) the histogram with answer to 2 below, d) the normal probability plot with answer to 3 below.

(You must submit your data for this question. No credit will be given without data.)

I: Right Skewed (Exponential)

1. (2 pts.) Code:

Solution:

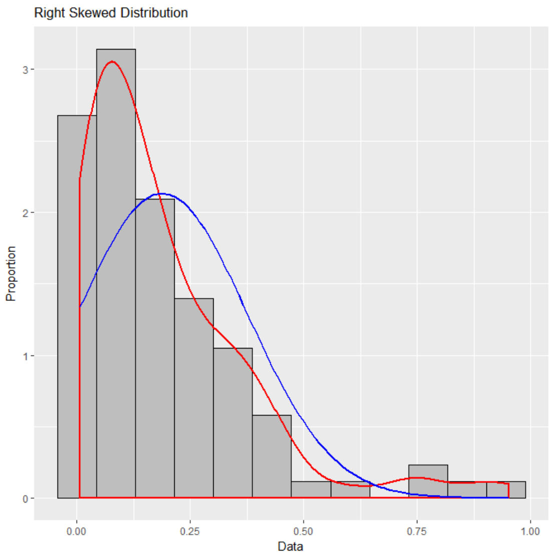
```
# right skewed: exponential distribution (lambda=5)
RandomData <- rexp(100,rate=5)
title <- "Right Skewed Distribution"
#
# HISTOGRAM
#
xbar <- mean(RandomData)
s <- sd(RandomData)
windows()
ggplot(data.frame(RandomData=RandomData), aes(x=RandomData)) +
  geom_histogram(aes(y=..density..), bins = sqrt(length(RandomData))+2,
                 fill = "grey",col="black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun=dnorm,  args=list(mean=xbar, sd=s), col="blue",
                lwd = 1) +
  ggtitle(title) +
  xlab("Data") +
  ylab("Proportion")
```

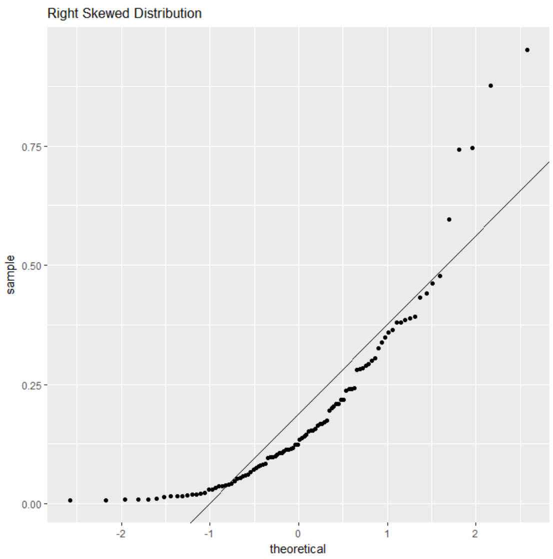
```
#
# QQ plot
#
windows()
ggplot(data.frame(RandomData=RandomData), aes(sample=RandomData)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle(title)
```

2. (4 pts.) Make a histogram of these observations. Please describe the shape of the distribution. How does the histogram compare with a normal density curve?

3. (4 pts. for each distribution) Make a normal probability plot of your data. Please describe the shape of the plot. Does the plot suggest any important deviations from normality? Please provide specifics to explain your answer You should be able to use your explanation to determine which of the four types of distributions generated the plot when you encounter plots later in the semester.

Solution:





The shape of the histogram is right skewed. The histogram deviates from the normal.	The plot is curved upward (concave up). The QQ-plot suggests important deviations from normality.

II: Left Skewed (Beta)

1. (2 pts.) Code:

Solution:

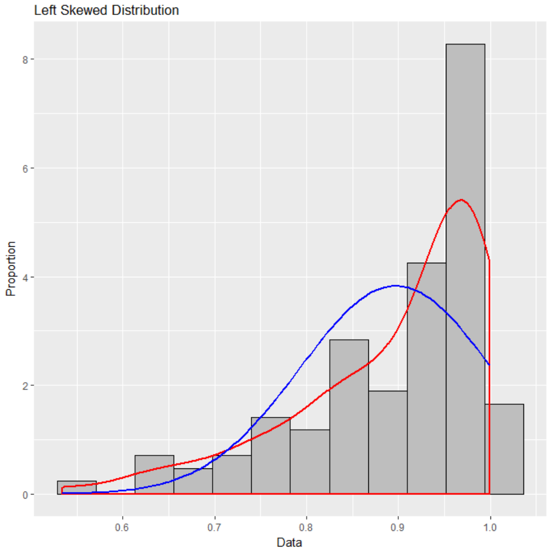
```
# left skewed: Beta distribution (on [0,1],  alpha = 7, beta = 0.8)
RandomData <- rbeta(100, 7, 0.8)
title <- "Left Skewed Distribution"
```

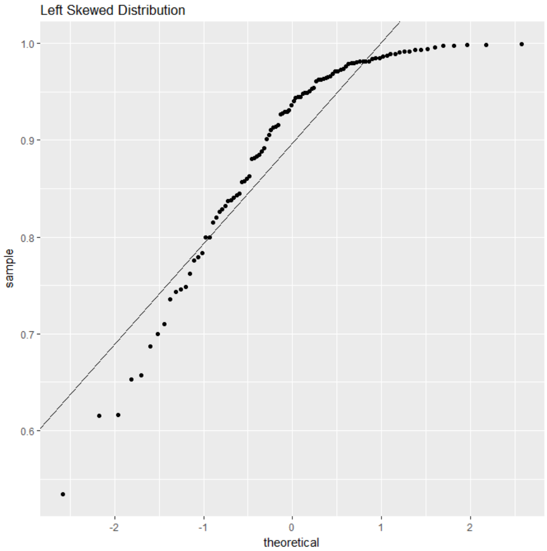
```
#
# HISTOGRAM
#
xbar <- mean(RandomData)
s <- sd(RandomData)
windows()
ggplot(data.frame(RandomData=RandomData), aes(x=RandomData)) +
  geom_histogram(aes(y=..density..), bins = sqrt(length(RandomData))+2,
                 fill = "grey",col="black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun=dnorm,  args=list(mean=xbar, sd=s), col="blue",
                lwd = 1) +
  ggtitle(title) +
  xlab("Data") +
  ylab("Proportion")
#
# QQ plot
#
windows()
ggplot(data.frame(RandomData=RandomData), aes(sample=RandomData)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle(title)
```

2. (4 pts.) Make a histogram of these observations. Please describe the shape of the distribution. How does the histogram compare with a normal density curve?

3. (4 pts. for each distribution) Make a normal probability plot of your data. Please describe the shape of the plot. Does the plot suggest any important deviations from normality? Please provide specifics to explain your answer You should be able to use your explanation to determine which of the four types of distributions generated the plot when you encounter plots later in the semester.

Solution:





It is left skewed. The histogram deviates from the normal curve.	The curve is open downward (concave down). The QQ-plot suggests important deviations from normality.

III: Short Tailed (Uniform)

1. (2 pts.) Code:

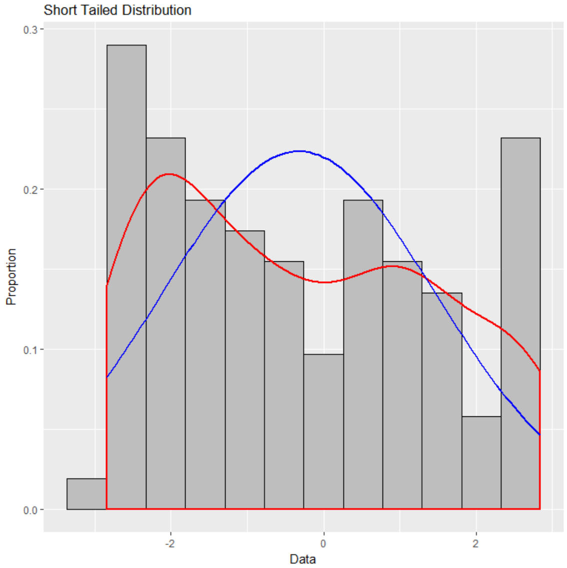
Solution:

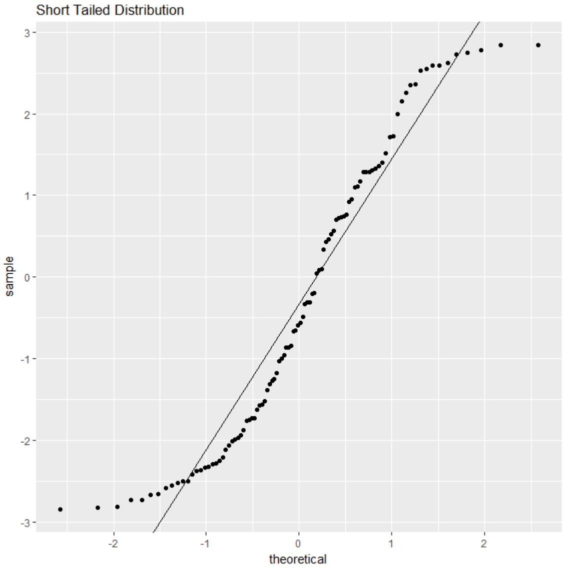
```
# short tailed: Uniform (on [-3,3])
RandomData <- runif(100,min=-3,max=3)
title <- "Short Tailed Distribution"
#
# HISTOGRAM
#
xbar <- mean(RandomData)
s <- sd(RandomData)
windows()
ggplot(data.frame(RandomData=RandomData), aes(x=RandomData)) +
  geom_histogram(aes(y=..density..), bins = sqrt(length(RandomData))+2,
                 fill = "grey",col="black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun=dnorm,  args=list(mean=xbar, sd=s), col="blue",
                lwd = 1) +
  ggtitle(title) +
  xlab("Data") +
  ylab("Proportion")
#
# QQ plot
#
windows()
ggplot(data.frame(RandomData=RandomData), aes(sample=RandomData)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle(title)
```

2. (4 pts.) Make a histogram of these observations. Please describe the shape of the distribution. How does the histogram compare with a normal density curve?

3. (4 pts. for each distribution) Make a normal probability plot of your data. Please describe the shape of the plot. Does the plot suggest any important deviations from normality? Please provide specifics to explain your answer You should be able to use your explanation to determine which of the four types of distributions generated the plot when you encounter plots later in the semester.

Solution:





The histogram does not have a clear mode or a well-defined peak like the normal density curve. Also, it fails to produce the tails of the normal density curve. It deviates from normal curve.	The points generally have high slope near the center but low slope near the end (S-shaped). The QQ-plot suggests important deviations from normality.

IV: Long Tailed (t-distribution)

1. (2 pts.) Code:

Solution:

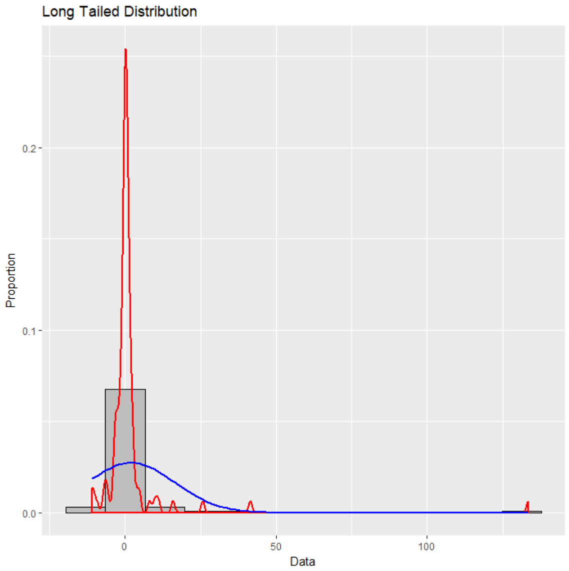
```
# long tailed: T distribution (df=1)
RandomData <- rt(100,df=1)
title <- "Long Tailed Distribution"
#
# HISTOGRAM
#
xbar <- mean(RandomData)
s <- sd(RandomData)
windows()
ggplot(data.frame(RandomData=RandomData), aes(x=RandomData)) +
  geom_histogram(aes(y=..density..), bins = sqrt(length(RandomData))+2,
                 fill = "grey",col="black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun=dnorm,  args=list(mean=xbar, sd=s), col="blue",
                lwd = 1) +
  ggtitle(title) +
  xlab("Data") +
  ylab("Proportion")
```

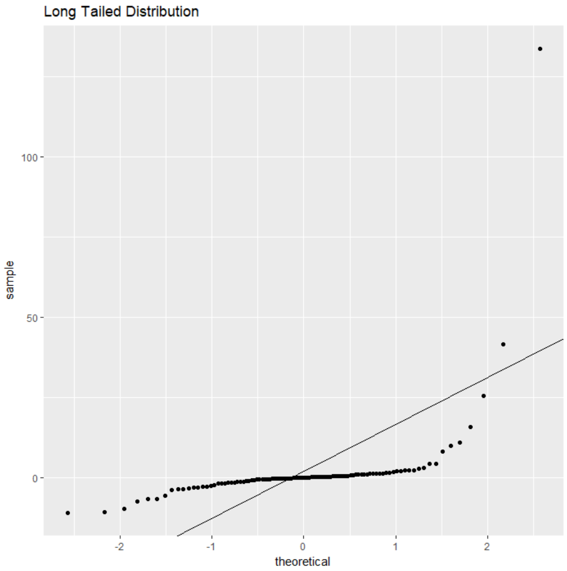
```
#
# QQ plot
#
windows()
ggplot(data.frame(RandomData=RandomData), aes(sample=RandomData)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle(title)
```

2. (4 pts.) Make a histogram of these observations. Please describe the shape of the distribution. How does the histogram compare with a normal density curve?

3. (4 pts. for each distribution) Make a normal probability plot of your data. Please describe the shape of the plot. Does the plot suggest any important deviations from normality? Please provide specifics to explain your answer You should be able to use your explanation to determine which of the four types of distributions generated the plot when you encounter plots later in the semester.

Solution:





This histogram has a sharp peak with outliers. It deviates from the normal curve.	The points generally have low slope near the center but high slope near the end (S-shaped). The QQ-plot suggests important deviations from normality.

E. (15 points) Comparison of data – GROUP (submitted separately on Blackboard)

Each group must consist of 3-4 people and will submit only one combined report. Even though it is due at the same time as the rest of the lab report, please submit this part as a separate pdf file under the “Lab 3 Group” link in Blackboard. Be sure that the names of all group members with their section time(s) are at the top of the page. If your grade is not on Blackboard (give us a week to grade the assignment), submit a regrade request with the names and sections of your group members. In addition, please indicate which person submitted the report so that we can find the grade.

Note that you still need to submit the answers to Parts B, C, D, and F as your individual report for Lab 3 in Blackboard.

1. (5 pts.) Present all of the graphs for Parts B, C and D from all of the members of the group where the graphs for the same distribution and number of data points are grouped together. Therefore, there will be 3-4 histograms and 3-4 normal probability plots for each type of random number simulation (6 sets in total: 1 from Part B, 1 from Part C, and 4 from Part D).
2. (10 pts.) For each of the six sets (histogram and normal probability) of plots (1 from Part B, 1 from Part C and 4 from Part D), please answer the following question: Do these plots look reasonably similar or do they look different? If they look different, please propose a possible explanation.

Solution:

Response for 2: I would expect all of them to look similar except for A with only 10 data points and maybe the t-distribution in D.

The problem with A is that the sample size is too small. The problem with the t-distribution is that the tails may look different.

F. (15 points) The distribution of Assaults Per Population (data file: Clean US Data)

We are interested in the normality of the Number of Assaults per 100,000 people (AssaultsPerPopulation).

1. (5 pts.) Code:

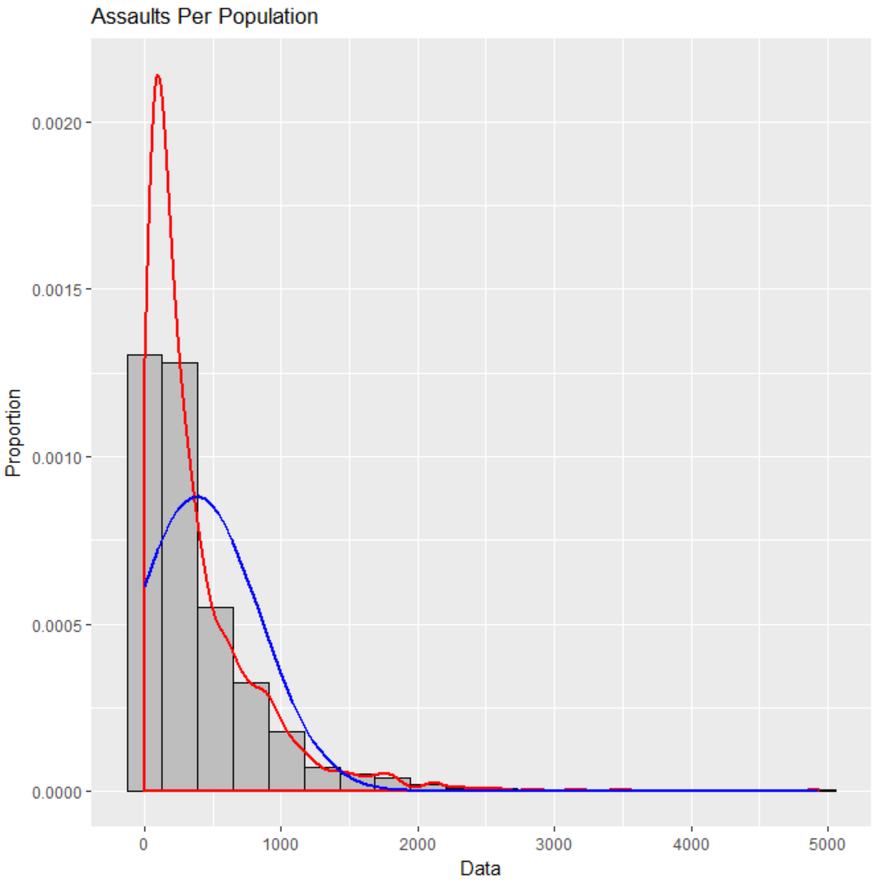
Solution:

```
# ===============================
# Part F: Distribution of Real Data
# ===============================
# Read in data
# using the interface
# Import Dataset --> From CSV --> browse to the file --> set delimiter
to tab -->
#   Change name to USData --> Import
# change variable name to make it easier.
assaults <- USData$AssaultsPerPopulation
title <- "Assaults Per Population"
```

```
# Histogram
xbar <- mean(assaults)
s <- sd(assaults)
windows()
ggplot(data.frame(assaults=assaults), aes(x=assaults)) +
  geom_histogram(aes(y=..density..),bins = 20,
                 fill="grey",col="black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun=dnorm,  args=list(mean=xbar, sd=s),
                  col="blue", lwd = 1) +
  ggtitle(title) +
  xlab("Data") +
  ylab("Proportion")
# QQ plot
windows()
ggplot(data.frame(assaults=assaults), aes(sample=assaults)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle(title)
```

2. (5 pts.) Make a histogram of these observations. Which distribution do you think this data is (normal, right skewed, left skewed, short tailed or long tailed)? Please explain your answer.

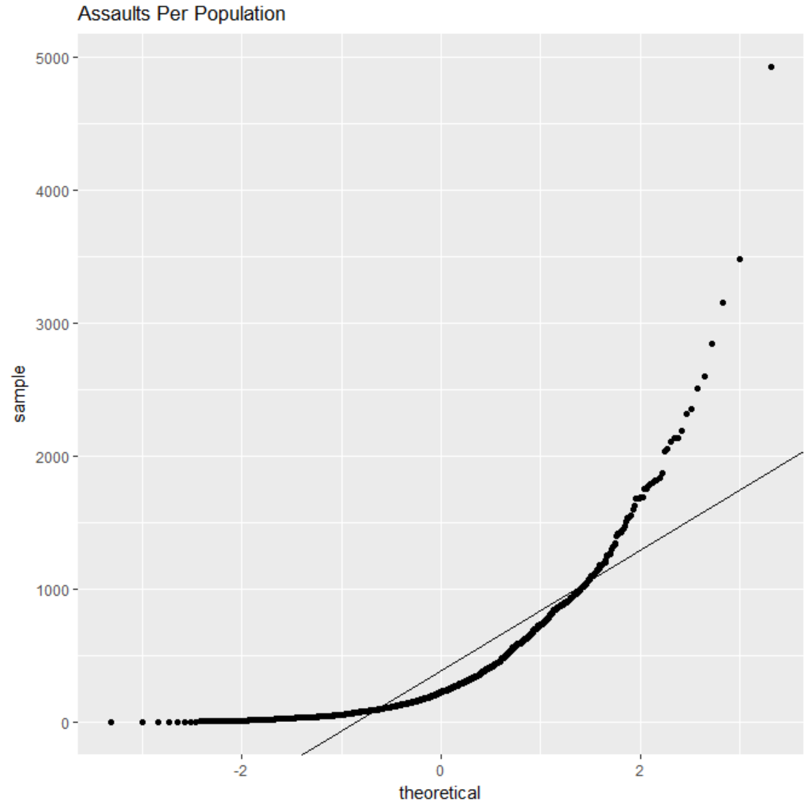
Solution:



Based on the histogram, I would say this distribution is right skewed.

3. (5 pts.) Make a normal probability plot of these data. Which distribution do you think this data is (normal, right skewed, left skewed, short tailed or long tailed)? Please explain your answer.

Solution:



This normal probability plot is very similar to the one produced for the right skewed dataset (it is concave up). This confirms my assessment of the histogram.