1. Select the Uniform(a,b) distribution and change the parameters to a = 2, and b = 5.
Ensure your randomly generated sample is based on a sample size of 100. Change the number of bins to 20.
a) Calculate the theoretical mean and variance for a Uniform(2,5) random variable.
b) Provide the sample mean and sample variance from your randomly generated sample. How do they compare to the theoretical mean and variance found in a)?
c) Do you expect the sample skewness to be positive, negative, or approximately 0? Explain why. What is the sample skewness from your randomly generated sample?
d) Do you expect the sample kurtosis to be less 3, greater than 3, or approximately
equal to 3? Explain why. What is the sample kurtosis from your randomly generated sample?
e) Insert a screenshot of the plot of the histogram for your sample with the normal pdf overlaid. Thinking about your responses to c) and d), explain why the normal model is not a good fit for the data.
Solution:

a) For X such that X~Uniform(2, 5), E(X) = $\frac{2+5}{2}$ = 3.5 and

Var(X) = $\frac{(5-2)^2}{12}$ = $\frac{3}{4}$ = 0.75 according to the formula given in STAT 230 in Tab.1.

Notation and Parameters	Probability Density Function $f(x)$	Mean $E(X)$	Variance $Var(X)$	Moment Generating Function $M(t)$
Uniform(a,b) $b > a$	$\frac{1}{b-a}$ $a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$ $t \neq 0$; 1 $t = 0$
Exponential(θ) $\theta > 0$	$\frac{1}{\theta}e^{-x/\theta}$ $x \ge 0$	θ	θ^2	$\frac{1}{1-\theta t}$ $t < \frac{1}{\theta}$
$N(\mu,\sigma^2) = G(\mu,\sigma)$ $\mu \in \Re,\ \sigma^2 > 0$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$ $x \in \Re$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$ $t \in \Re$

Tab.1 The formulas for continuous distributions

Thus, the theoretical mean is 3.5 and the theoretical variance is 0.75 for a Uniform(2,5) random variable.

b) The sample mean is 3.496 and the sample variance is 0.745 according to the generated table Tab.2.

Statistic	Value	Statistic	Value	Statistic	Value
Mean	3.496	Variance	0.745	Skewness	0.016
Median	3.442	IQR	1.533	Kurtosis	1.727

Tab.2 The numerical summery of the randomly generated samples

The sample mean is close to the theoretical mean, which is just |3.496 – 3.5| = 0.004 smaller than the theoritical mean; The variance is |0.745 – 0.75| = 0.005 smaller than the theoretical variance.

c) Since the samples generated by uniform distribution with a relatively large sample size should basically follow uniform distribution, which should be roughly symmetric, the skewness of the of the sample is expected to be close to 0. The sample skewness from my randomly generated sample is 0.016.

d) Since the sample generated by uniform distribution should not be peaked and rarely has tails (i.e. the tails are not heavy), the kurtosis is expected to be less than 3. Also, since the samples generated by uniform distribution with a relatively large sample size should be close to a uniform distribution, the sample kurtosis is expected to be close to 1.2. The sample kurtosis from my randomly generated sample is 1.727 according to Tab.2.

e) The histogram corresponding to my randomly generated samples is Fig.1.

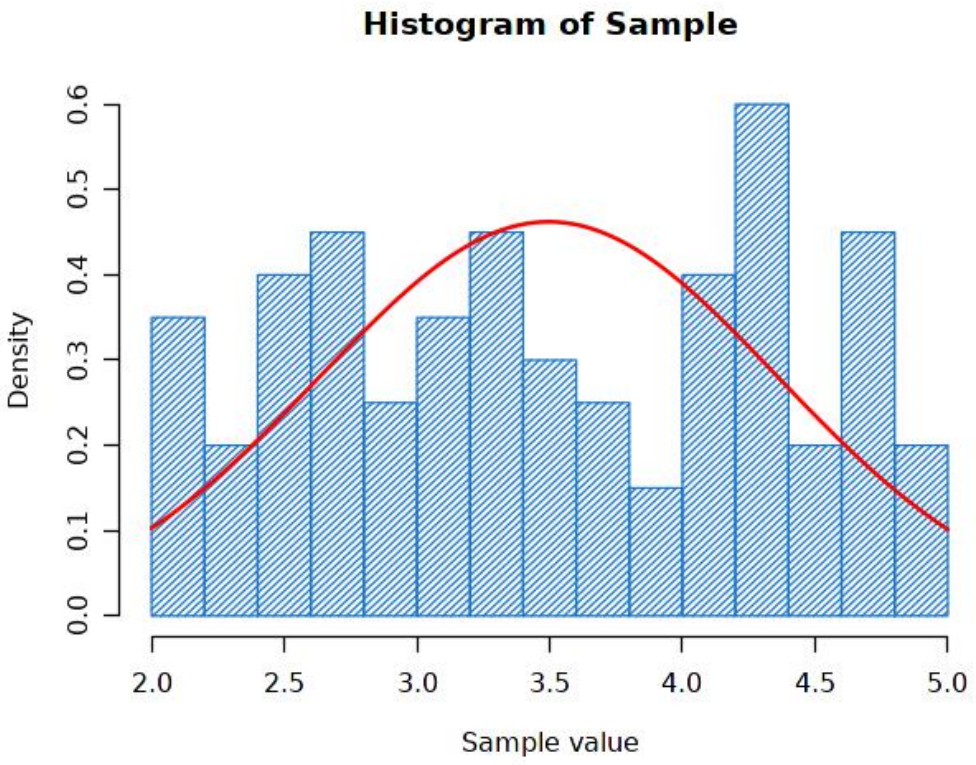

Fig.1 The histogram of randomly generated samples by uniform(2, 5)

Since the sample kurtosis is 1.727, which is much less than 3, it means the sample does not have heavy enough tails and it is not that peaked, which is against the features of the p.d.f of normal distribution.

2. Select the Exponential( ) distribution and change the parameter = (mean) to 0.5.

a) What is the theoretical mean and variance for this distribution? Provide the sample mean and sample variance from your randomly generated sample and discuss how they compare to the theoretical values.

b) Do you expect the sample skewness to be positive, negative, or approximately 0? Explain why. What is the sample skewness from your randomly generated sample?

c) Insert a screenshot of the plot of the empirical cdf for your sample with normal cdf overlaid. Discuss what the image is illustrating and explain why the normal model is not a good fit for the data.

Solution:

a) For X~Exponential(θ = 0.5), E(X) = θ = 0.5 and Var(X) = θ^2= 0.5^2 = 0.25 according to Tab.1. Therefore, the theoretical mean is 0.5 and the theoretical variance is 0.25. The sample mean is 0.472 and sample variance is 0.183 from my randomly generated sample. According to Tab.3, the sample mean is |0.472 – 0.5| = 0.028 smaller than the theoretical mean and the sample variance is |0.183 – 0.25| = 0.067 smaller than the theoretical variance. Generally speaking, the sample mean and the sample variance are close to the theory.

Statistic	Value	Statistic	Value	Statistic	Value
Mean	0.472	Variance	0.183	Skewness	1.704
Median	0.348	IQR	0.489	Kurtosis	7.491

Tab.3 The numerical summery of the randomly generated samples

b) The sample skewness is expected to be positive. Since the sample is generated by exponential distribution, the sample should not be symmetrically distributed and the right tail of the sample is expected to be very long. The sample skewness from my randomly generated sample is 1.704 according to Tab.3.

c) The histogram corresponding to my randomly generated samples is Fig.2.

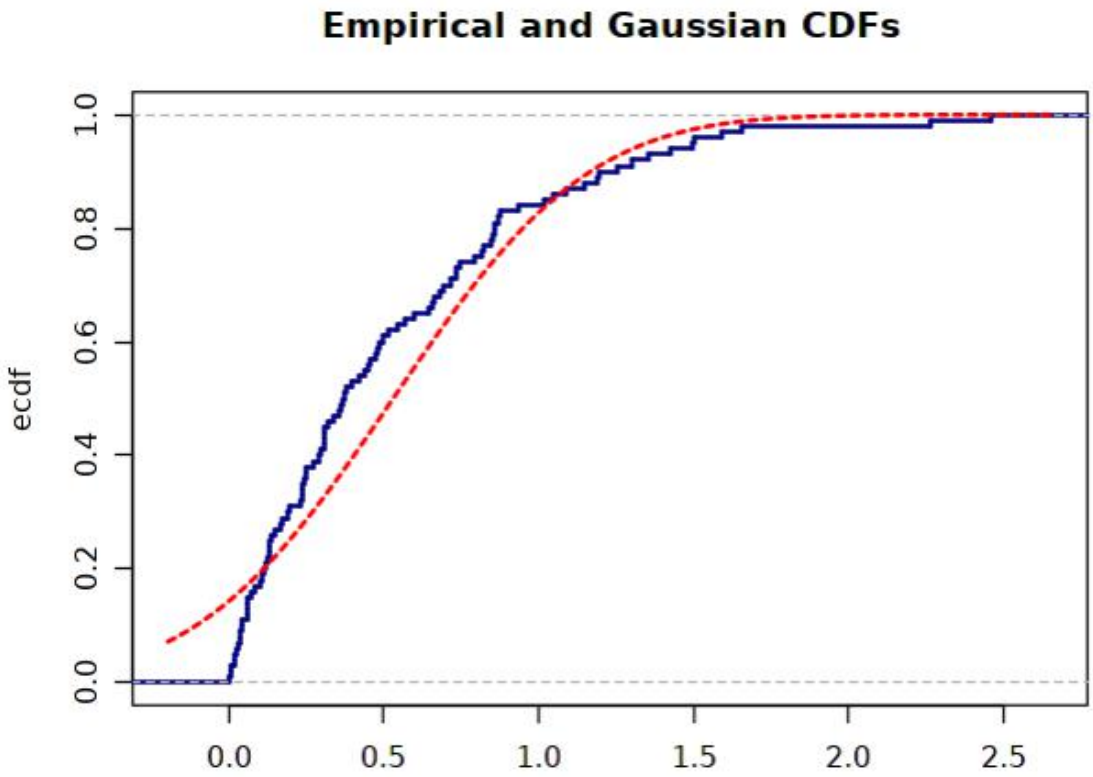

Fig.2 The empirical C.D.F for my sample with normal C.D.F overlaid

The graph shows that the exponential distribution does not fit normal distribution. First, the probability for the sample to have negative data is 0, whereas the normal CDF shows there can be a good amount of negative data in normal distribution. Also, the proportion of data smaller than 0.5 is way too higher than the expected probability for a normal distribution to have data smaller than 0.5, and the proportion of data bigger than 1.5 is way too smaller than the expected probability for a normal distribution to have data bigger than 1.5.

3. Select the G(μ, σ) distribution and ensure that μ is set to 0 and σ is set to 1.
a) Change the sample size to 50. Insert a screenshot of the corresponding histogram and provide the values for the sample mean, variance, skewness, and kurtosis.
b) Change the sample size to 100. Insert a screenshot of the corresponding histogram and provide the values for the sample mean, variance, skewness, and kurtosis.
c) Change the sample size to 500. Insert a screenshot of the corresponding histogram and provide the values for the sample mean, variance, skewness, and kurtosis.
d) Discuss how the numerical and graphical summaries for each sample size compare to their expected values. What do you notice as the sample size increases?
Solution:
a) The histogram is Fig.3.

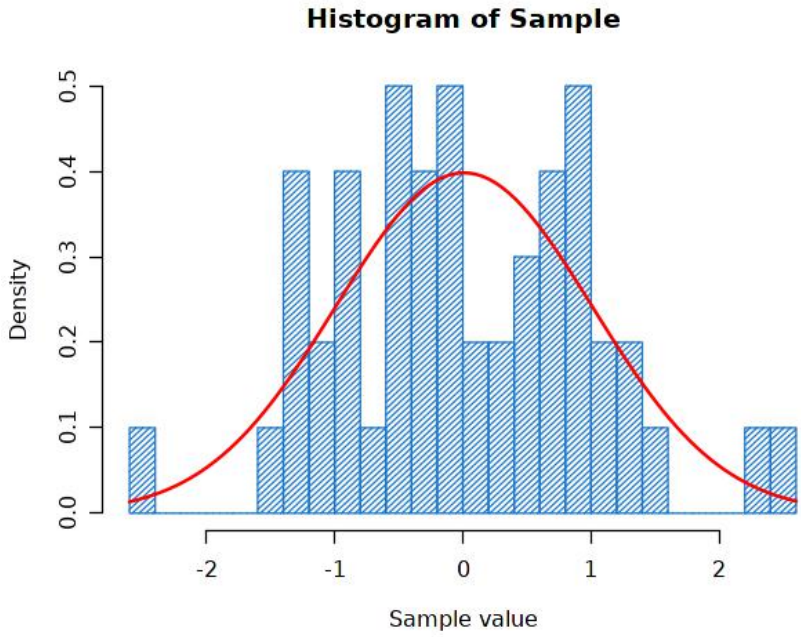

Fig.3 The histogram of randomly generated samples by G(0, 1) with sample size 50

The sample mean is 0.009, the sample variance is 1.003, the sample skewness is 0.202, and the sample kurtosis is 2.961 according to Tab.4.

Statistic	Value	Statistic	Value	Statistic	Value
Mean	0.009	Variance	1.003	Skewness	0.202
Median	-0.089	IQR	1.382	Kurtosis	2.961

Tab.4 The numerical summery of randomly generated samples by G(0, 1) with sample size 50

b) The histogram is Fig.4.

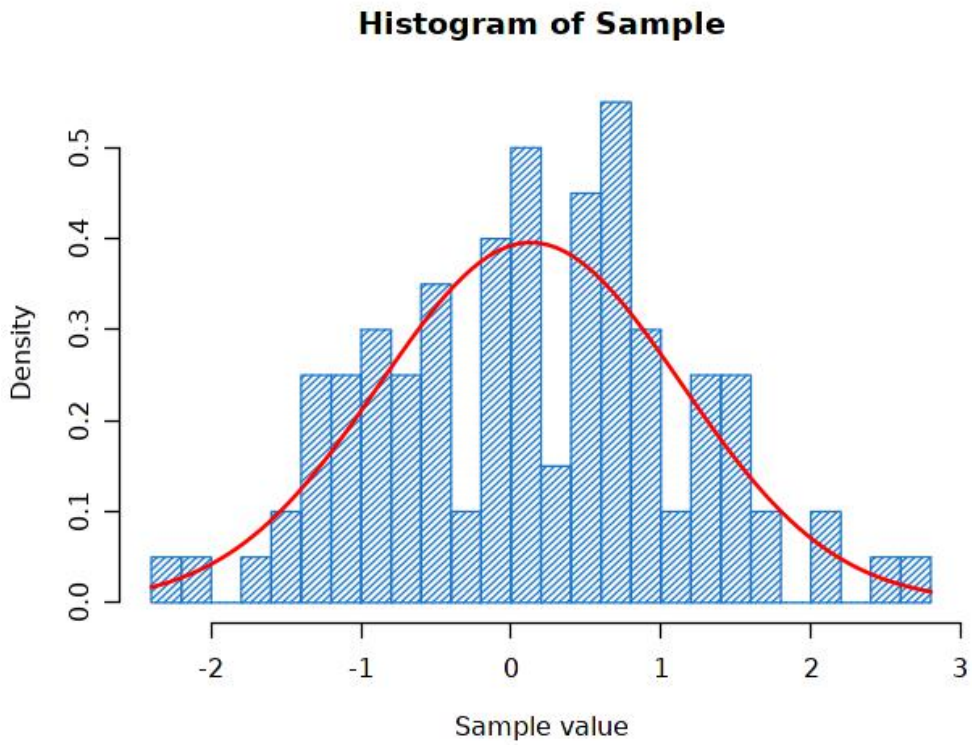

Fig.4 The histogram of randomly generated samples by G(0, 1) with sample size 100

The sample mean is 0.131, the sample variance is 1.016, the sample skewness is -0.004, and the sample kurtosis is 2.673 according to Tab.5.

Statistic	Value	Statistic	Value	Statistic	Value
Mean	0.131	Variance	1.016	Skewness	-0.004
Median	0.161	IQR	1.426	Kurtosis	2.673

Tab.5 The numerical summery of randomly generated samples by G(0, 1) with sample size 100

c) The histogram is Fig.5.

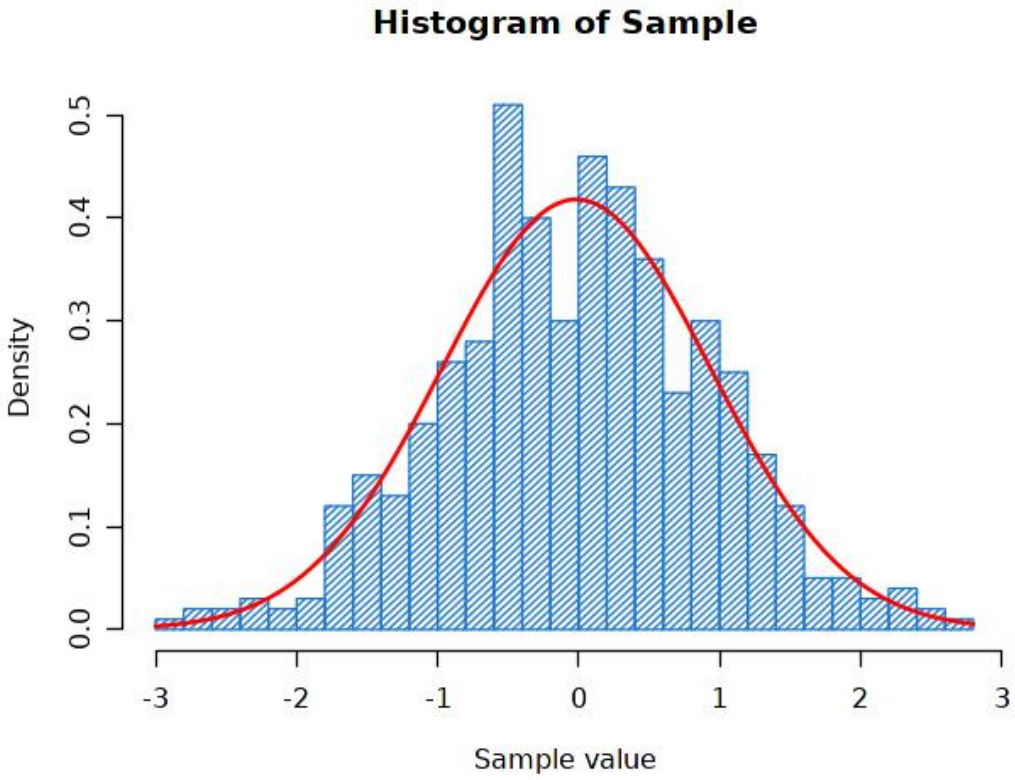

Fig.5 The histogram of randomly generated samples by G(0, 1) with sample size 500

The sample mean is -0.016, the sample variance is 0.910, the sample skewness is -0.029, and the sample kurtosis is 3.008 according to Tab.6.

Statistic	Value	Statistic	Value	Statistic	Value
Mean	-0.016	Variance	0.910	Skewness	-0.029
Median	0.012	IQR	1.225	Kurtosis	3.008

Tab.6 The numerical summery of randomly generated samples by G(0, 1) with sample size 500

d) According to the numerical summery for the different sample sizes from a) to c), the sample means are always close to 0 and the sample variances are always close to 1, which are close enough to the expected values; The skewness values for the sample size of 100 and 500 are very close to 0 that is the expected value, but the skewness value for the sample size of 50 is not very close to 0; The kurtosis for the sample size of 50 and 500 are very close to the expected value 3 for normal distribution, but the kurtosis for the sample size of 100 is far from 3. According to the graphic summery, the histogram fit better to N(0, 1) curve in terms of symmetry, centre, peak, and tails as the sample size increases. Basically, the sample data fits N(0, 1) better as the sample size increases.

4. Analyze the discount offered variate by answering the following questions.
a) What type of variate is discount offered? (ex. Discrete, continuous, categorical)
b) Provide the five number summary identifying what each number represents, followed by the range, and interquartile range for discount offered in your dataset.
c) Provide the values of the sample mean, sample standard deviation, and sample skewness. Report to 3 decimal places. How do the sample mean and sample median compare? What does this say about the tails of the distribution and how does that
compare to an exponential distribution?
d) Create and insert a plot of the relative frequency histogram with superimposed
Exponential probability density function.
e) For the Exponential(θ) distribution the mean and standard deviation are both equal to θ. Therefore, the sample mean and sample standard deviation are both estimates of θ based on the observed data. Are the sample mean and sample standard deviation close in value for your data?
f) Using both the numerical and graphical summaries, describe how well the Exponential model fits these data. You should make at least four comparisons between what you observed for your data set and what you would expect to observe if the data were generated from an Exponential model.

Solution:

a) Discount is a continuous variate.

b) According to Fig.6, the minimum value is 1, the first quantile value is 3, the median is 7, the third quantile value is 10, and the maximum value is 63.

```
> fivenum(discount)
[1]  1  3  7 10 63
```

Fig.6 The five number summery of the discount data

c) According to Fig.7, the sample mean is 12.346, the sample standard deviation is about 14.876, and sample skewness is about 1.839.

```
> mean(discount)
[1] 12.346
> sd(discount)
[1] 14.87578
> skewness(discount)
[1] 1.838955
```

Fig.7 The numerical summery of the discount data

The sample mean is |12.346 - 7| = 5.346 larger than the sample median, so the right tail of the distribution is longer than the left tail, which is similar to an exponential distribution.

d) The graph is in Fig.8.

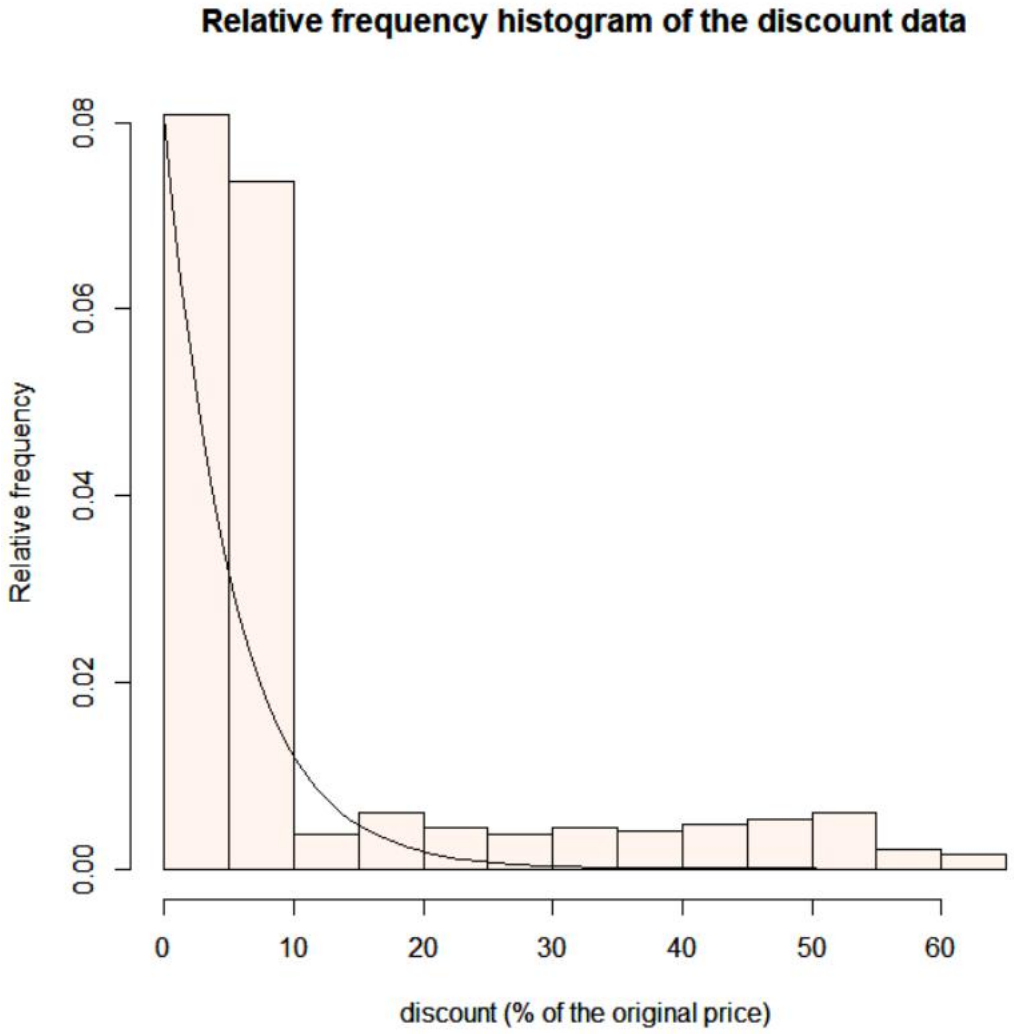

Fig.8 The relative frequency histogram of the discount data overlapped by an exponential distribution curve exp(12.346)

e) In my data, the sample mean is 12.346 and the sample standard deviation is 14.876 according to question c), so the sample mean is indeed close to the sample standard deviation.

f) 1. The expected value of the exponential distribution is θ, and θ has already been set to be 12.346, which is the mean of the data. Thus, the expected value of the exponential distribution equals the expected value of the data.
2. The standard deviation of the exponential distribution is $\theta = 12.346$, which is close to the sample standard deviation 14.876.
3. According to the graphic summery, the right tail of the sample data is longer than the left tail, and the right tail of exponential distribution is also longer than the left tail.
4. The minimum value of the sample data is 0, which equals the minimum value of exponential distributions.

5. Analyze the cost of product variate by answering the following questions.

a) What type of variate is the cost of the product?

b) Provide the sample mean, sample median, sample standard deviation, sample skewness, and sample kurtosis for the cost of the product variate in your dataset. (Round to 3 decimal places)

c) Create and insert the plot of the relative frequency histogram with superimposed Gaussian probability density function.

d) How well does the Gaussian model fit these data? Use the graphical and numerical summaries to justify your answer. You should make at least four comparisons between what you observed for your data set and what you would expect to observe if the data were generated from a Gaussian model.

Solution:

a) The cost of the product is a continuous variate. Because the unit of the variate is %, and theoretically, non-integer percent of discount is possible.

b) According to Fig.9, the sample mean is 209.484, the sample standard deviation is about 49.617, the sample skewness is about -0.132, and the sample kurtosis is 2.006.

```
> mean(cost)
[1] 209.484
> sd(cost)
[1] 49.61666
> skewness(cost)
[1] -0.1323038
> kurtosis(cost)
[1] 2.005591
```

Fig.9 The numerical summery of the cost data

c) The graph is Fig.10.

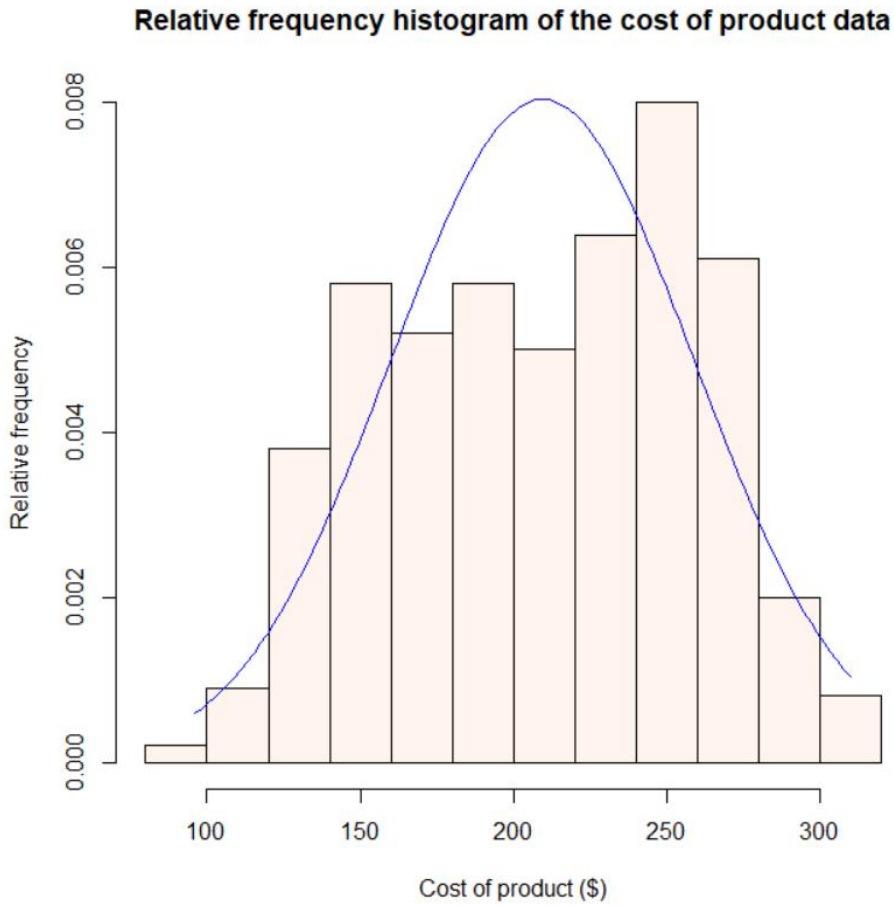

Fig.10 The relative frequency histogram of the cost of product data overlapped by a normal distribution curve norm(209.484, 49.617)

d) 1. The sample mean of the cost of product is 209.484, which equals the expected value of the normal distribution norm(209.484, 49.617) by design. Thus, they are all centered at 209.484.

2. The sample standard deviation is 49.617, which equals the standard deviation of norm(209.484, 49.617) by design. Thus, the spread of the cost of product (i.e. the variability) basically follows the normal distribution.

3. The skewness of the sample data of the cost of product is -0.132, which is close to zero. Thus, the graph is basically symmetric, which matches the feature of normal distributions.

4. According to the graphic, the relative frequency histogram of the cost of product data is not unimodal, which does not match the overlapped normal distribution norm(209.484, 49.617).

5. According to the graphic, the peak of the data is not at the sample mean 209.484, which is against the normal distribution.

6. According to the graphic, the tail of the data is too light to be normal.

7. The kurtosis of the data is about 2.006, which is much less than the expected kurtosis 3 for normal distributions.

6. a) Create and insert a scatterplot of cost of product (y) versus weight in grams (x). Use the"lm" function to add a linear line of best fit.
b) Provide the sample correlation for cost of product and weight in grams.
c) With the aid of the scatterplot, interpret the sample correlation.

Solution:

a) The graph is Fig.11.

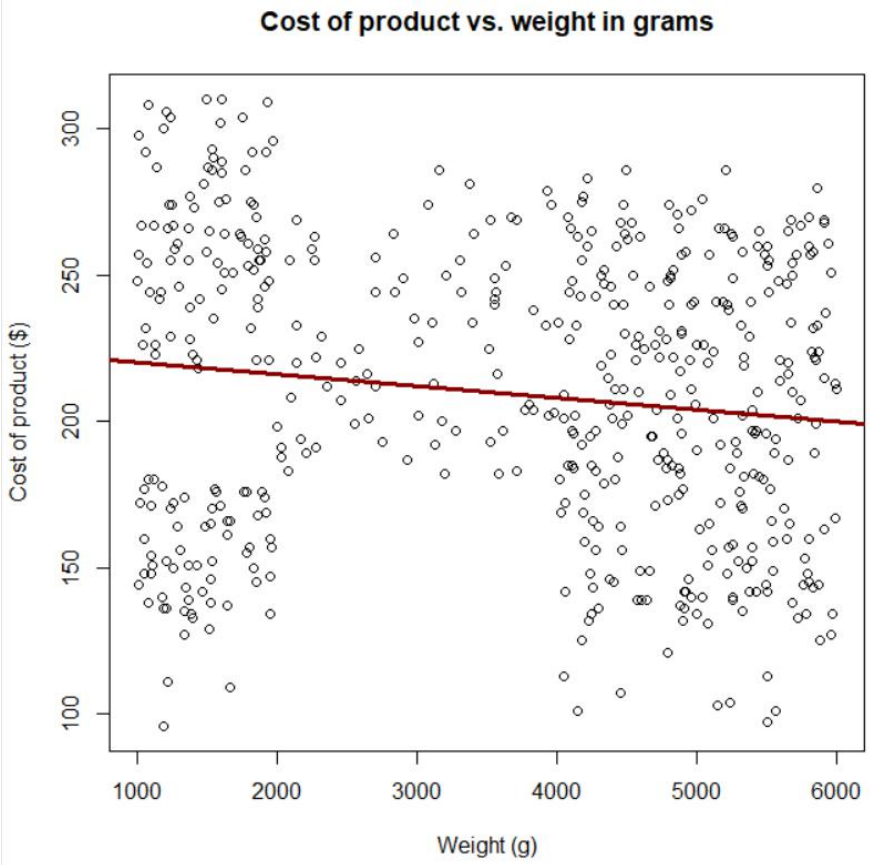

Fig.11 Cost of product vs. weight in grams scatterplot overlapped by a linear line fit to the scatterplot

b) The correlation is about -0.135, which is shown in Fig.12.

```
> cor(weight, cost)
[1] -0.1346292
```

Fig.12 The sample correlation for cost of product and weight in grams

c) According to the scatterplot, the points are not close to the linear line that best fits to the cost of product and the weight in grams, which means the two variates are not in linear relationship, so the sample correlation is close to 0; Then, since the linear line that best fits to the cost of product and the weight in grams is from top-left to bottom-right, and the height of points at the right side are more likely to be lower than the height of the points at the left side, the two variates are in negatively correlated. Thus, the sample correlation is negative.

7. a) What type of variate is the mode of shipment?
b) What proportion of purchases are shipped by Flight, Ship, and Road in your dataset?
c) Compare the weight of the purchase (in grams) between the different shipping methods by creating 3 side by side boxplots for the weight in grams for each of flight, ship, and road. Insert this plot in your pdf.
d) Suppose you were only given this plot. Describe the information about the differences and similarities between the three groups of data that you can obtain just from this plot. Things you might wish to comment on include:

- **a comparison of the symmetry of the data sets**
- **a comparison of the tail regions of the data sets**
- **a comparison of the ranges (variability) of the data sets**
- **a comparison of the medians (location) of the data sets**
- **a comparison of the number of outliers of the data sets**

Solution:

a) The variate is categorical.

b) The proportion of purchases are shipped by Flight is 0.134, by Ship is 0.698, by Road is 0.168 from the result shown in Fig.13.

```
> length(shipment_mode[which(shipment_mode == "Flight")]) / n  # the proportion of purchases are shipped by Fl
ight
[1] 0.134
> length(shipment_mode[which(shipment_mode == "Ship")]) / n  # the proportion of purchases are shipped by Ship
[1] 0.698
> length(shipment_mode[which(shipment_mode == "Road")]) / n  # the proportion of purchases are shipped by Road
[1] 0.168
```

Fig.13. The proportion of purchases are shipped by Flight, Ship, and Road

c) According to the boxplot in Fig.14, the weights of the purchase by the three modes of shipment are very close in terms of range, five number summery, and outliers. There are no outliers in the three modes, the weight of the purchase by Road has the largest maximum value, third quantile, median, and first quantile among the three modes, but has the smallest minimum value; The weight of the purchase by Flight has the smallest maximum value, third quantile, first quantile, but has the largest minimum value; The weight of the purchase by Ship has a similar median with the weight of the purchase by Flight, which are all smaller than the weight of the purchase by Road.

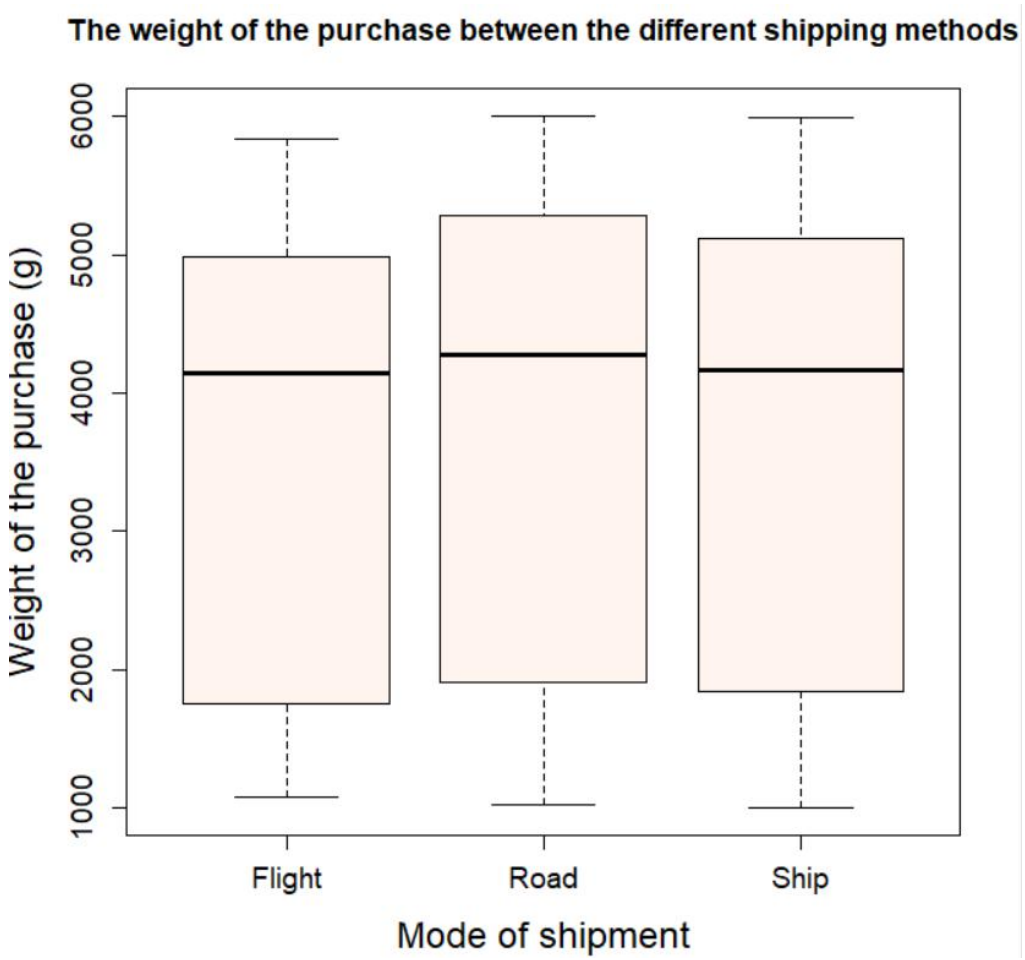

Fig.14 The weight of the purchase between the different shipping methods

d) The symmetry of the three data sets is very close, since they all have more units below the median, and less units above the median.
From the information above, they all have a longer left tail and shorter right tail since more sample data is concentrated at the region larger than median (i.e. the largest 50% of data has very narrow range).
The ranges of the data sets are very close.
The weight of the purchase by Ship has a similar median with the weight of the purchase by Flight, which are all smaller than the weight of the purchase by Road.
The data sets do not include any outliers.

8. Write an approximately 1-2 paragraph reflection (in full sentences) that summarizes how you achieved (or not if you're still not confident with them) the intended learning outcomes by completing this assignment.

By Q4-7, I understand different type of variates, observational experiments, and different types of data we collect and how we process the data we collect. Then, by most of the questions in the assignment, I repeatedly create and report the numerical summery and graphical summery of data, which includes the location, variability, shape for data sets; frequency histograms, scatterplots, and box plots. This process trained my coding ability with R and my data analysis ability, since I should pull out the information in the numerical summery including sample mean, sample variability/sample standard deviation, sample skewness, sample kurtosis, five number summery, and the graphs, to describe the location, variability, and the shape of the data sets. Also, Q1-3, Q7, and some other questions checked my ability of comparing the distribution of a data set to specific distribution, or to the distribution of other data sets, about their location, variability, and shape. Then, in Q6, I processed bivariate data using scatterplot, fit linear line, and correlation, to describe the relationship of the two variates. Also, from Q1 to Q3, I feel the relationship between the sample size and variability from expected value and how inherent variability in random samples influences the sample measures for the data compared to expected.