1. Select the normal distribution and examine the histogram and qqplot for multiple samples by hitting the "Resample!" button a number of times and observing the variation in the data/plots.
   a) For normal data, what do you expect to see in the qqplot?
   b) Paste into your assignment the images of the histogram and qqplot for **two different** samples (ideally illustrating different behaviours). Comment on how the images compare to expected and explain any deviations.

a) For normal data, I expect to see the points lie on a straight line. Also, since the quantiles of the normal distribution change more rapidly at the tails, both ends of the line are expected to deviate more from the line.
b) The images of the relative frequency histogram and qqplot for the two different samples are in Fig.1 and Fig.2.

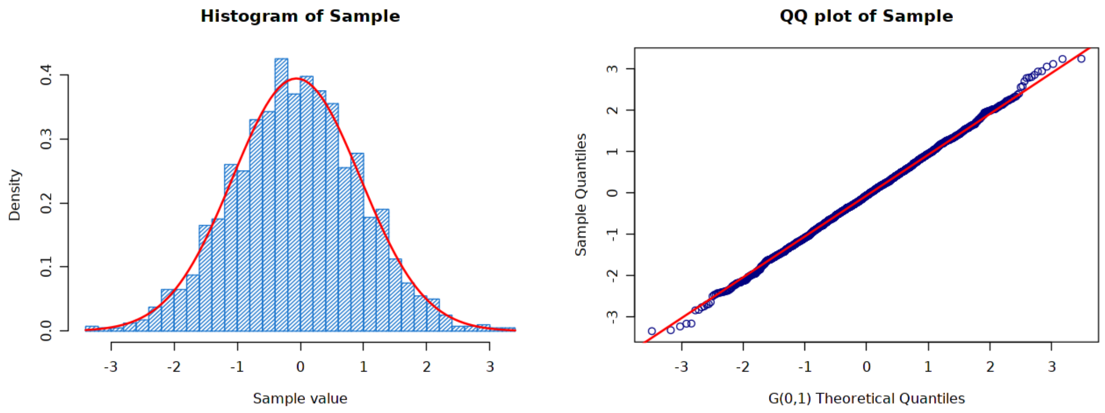


Fig.1 Relative frequency histogram and qqplot of sample from the normal distribution 1

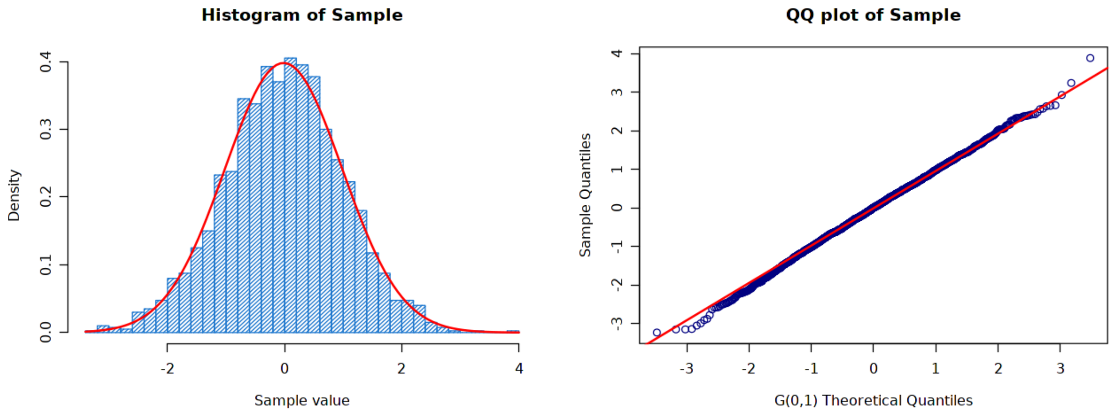


Fig.2 Relative frequency histogram and qqplot of sample from the normal distribution 2

The images are exactly as what I expected: The sample points line on a straight line, and both ends of the line deviate more from the line since the quantiles of the normal distribution change more rapidly at the tails.

2. Select the Exponential distribution and examine the histogram and qqplot for multiple samples by hitting the “Resample!” button a number of times and observing the variation in the data/plots.
   a) Paste into your assignment the images of the histogram and qqplot for **two different** samples – one that looks very **close to what you'd expect** for exponential data and one that **doesn't** quite match your expectation.
   b) For your sample in part a) that matches your expectation, describe any patterns that you observe (s-shaped, u-shaped, amount of deviation from theoretical quantiles, direction of deviation, etc.) in the qqplot. Looking at the qqplot, explain what information you can obtain from these patterns with regards to the properties of the appropriate distribution (ex. symmetry, tail weight, etc) for this data.
   c) For your sample in part a) that doesn't quite match your expectation, explain how the data is creating the deviation in the qqplot. For example, you can reference the histogram or numerical summaries to describe properties of the data.

(a) The images of the relative frequency histogram and qqplot for the two different samples are in Fig.3 and Fig.4. The QQ plot in Fig.3 matches what I expect. And the QQ plot in Fig.4 does not quite match my expectation.

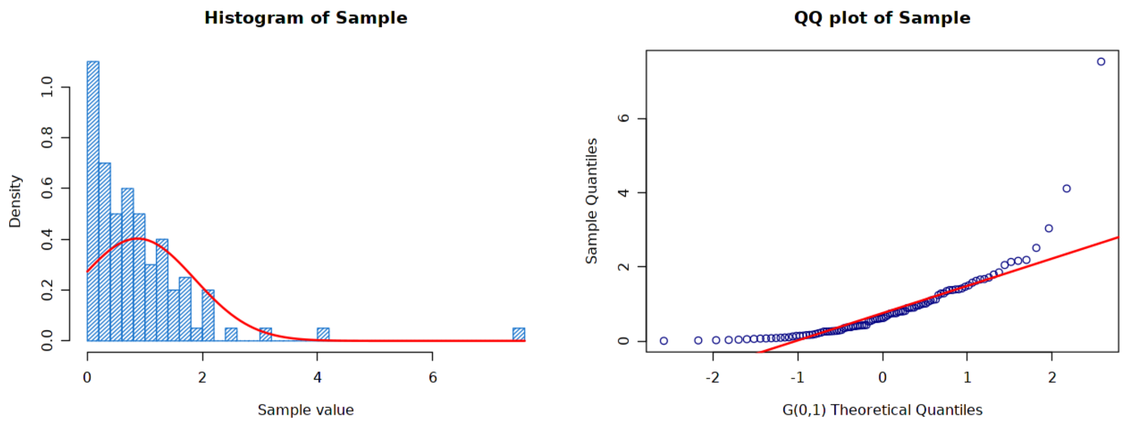


Fig.3 Relative frequency histogram and qqplot of sample from the exponential distribution 1

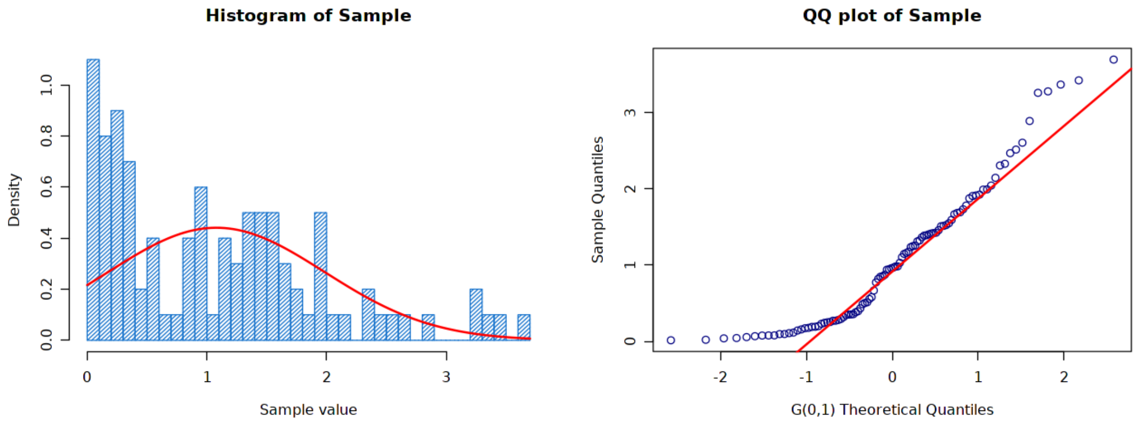
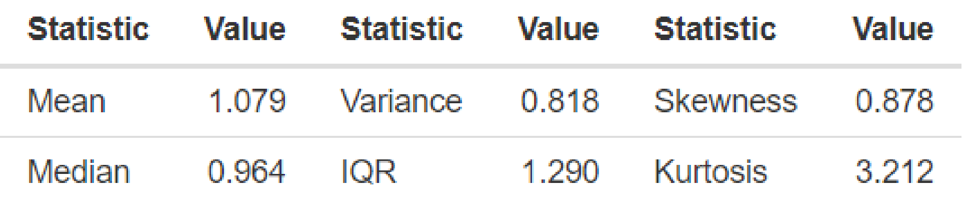


Fig.4 Relative frequency histogram and qqplot of sample from the exponential

distribution 2

(b) The sample points in Fig.3 are along a downside-up U-shape curve. Also, the sample points at the right end deviates very much from the theoretical normal distribution quantiles, which is much higher than the theoretical normal distribution quantiles. The points at the left end are very close to 0, which also deviates much from (higher than) the theoretical normal distribution quantiles. Additionally, there is a part of the points between the left and right ends that lie on the straight line, which means the part of points are very close to the theoretical normal quantiles. Overall, the points are close to the theoretical quantiles of an exponential distribution. The information we can get is: The skewness of the sample data should be positive since they are along a downside-up U-shape curve, and thus its relative frequency histogram has a longer right tail.

(c) The sample points in Fig.4 are along a rough S-shape curve rather than a typical U-shape curve. The sample points at the right end deviates from the theoretical quantiles for the normal distribution even less than some points that are not at the right end in the second half of the curve. The QQ plot means that the relative frequency histogram of the data should be more symmetric than usual exponential distributions with light tails. Note that as the numerical summery in Tab.1 shown, the kurtosis is indeed not very larger than 3, which means the data relative frequency histogram is not very peaked and has lighter tails than the usual exponential distributions; The skewness is 0.878 that is close to 0, which means the relative frequency histogram is not very positively skewed. Due to the QQ plot, graphical and numerical summery, the data distribution is not as typical exponential distributions.



Statistic	Value	Statistic	Value	Statistic	Value
Mean	1.079	Variance	0.818	Skewness	0.878
Median	0.964	IQR	1.290	Kurtosis	3.212

Tab.1 Numerical summery for the sample from the exponential distribution 2

3. Select the t distribution and examine the histogram and qqplot for multiple samples by hitting the "Resample!" button a number of times and observing the variation in the data/plots.
   a) Paste into your assignment the image of the histogram and qqplot for **one** of your samples.
   b) Describe any patterns that you observe (s-shaped, u-shaped, amount of deviation from theoretical quantiles, direction of deviation, etc.) in the qqplot. Explain what information you can obtain from these patterns with regards to the properties of the t distribution (ex. symmetry, tail weight, etc) compared to the Gaussian distribution.

(a) The image of the relative frequency histogram and qqplot for the t distribution is in Fig.5.

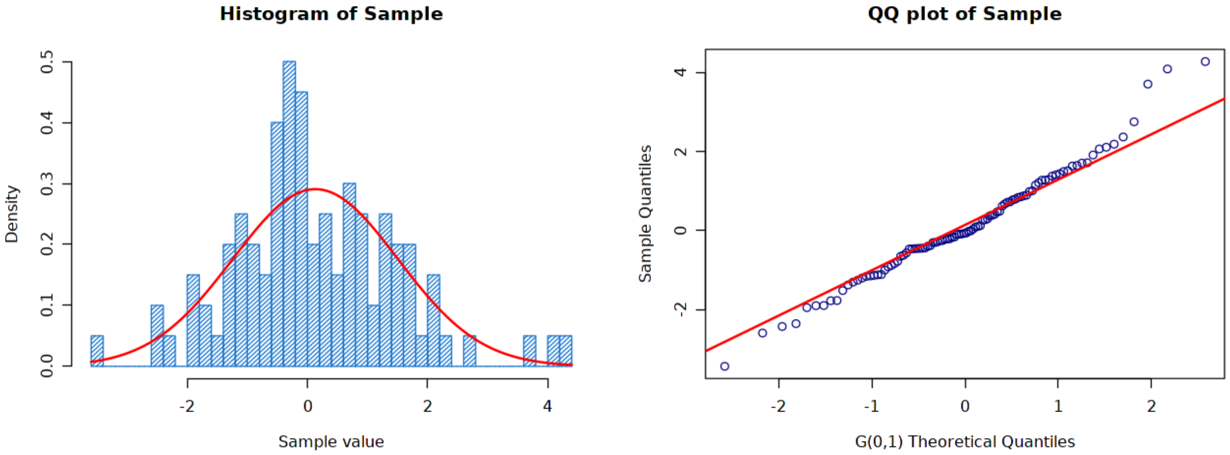
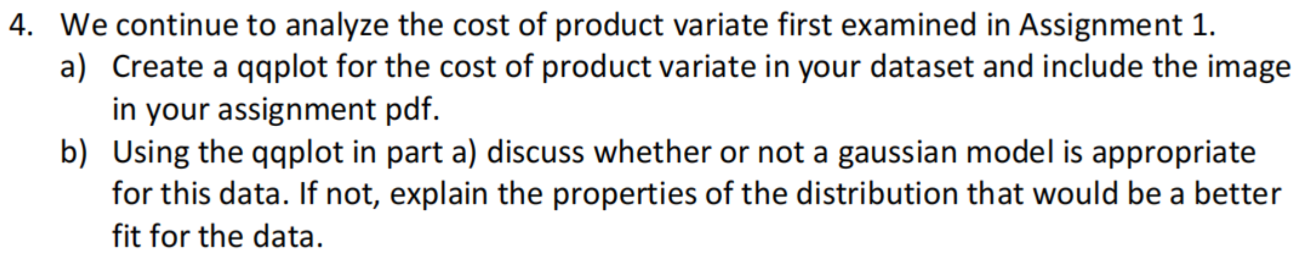


Fig.5 Relative frequency histogram and qqplot of sample from the t distribution

(b) In the qqplot, the points basically lie on an S-shape, but in the opposite direction to the S-shape for the uniform distribution. The sample points that are not at the ends basically lie along the straight line, but the points at the left and right ends deviate much from the straight line (i.e. the theoretical quantiles). To be more specific, the points at the right end deviate positively from the theoretical quantiles, and the points at the left end deviate negatively from the theoretical quantiles. We can get the following information: Most of points in the middle are roughly on a straight line while the points at the left end deviate downward and the points at the right end deviate upward relative to the theoretical normal quantiles, which means the data's relative frequency histogram is unimodal and symmetric with a skewness close to 0; The relative frequency histogram of the data should be very peaked with heavy tails since the points at the left end are lower than the theoretical normal quantiles and the points at the right end are higher than the theoretical normal quantiles, so the kurtosis should be greater than 3.



4. We continue to analyze the cost of product variate first examined in Assignment 1.
   a) Create a qqplot for the cost of product variate in your dataset and include the image in your assignment pdf.
   b) Using the qqplot in part a) discuss whether or not a gaussian model is appropriate for this data. If not, explain the properties of the distribution that would be a better fit for the data.

(a) The qqplot for the cost of product variate is in Fig.6.

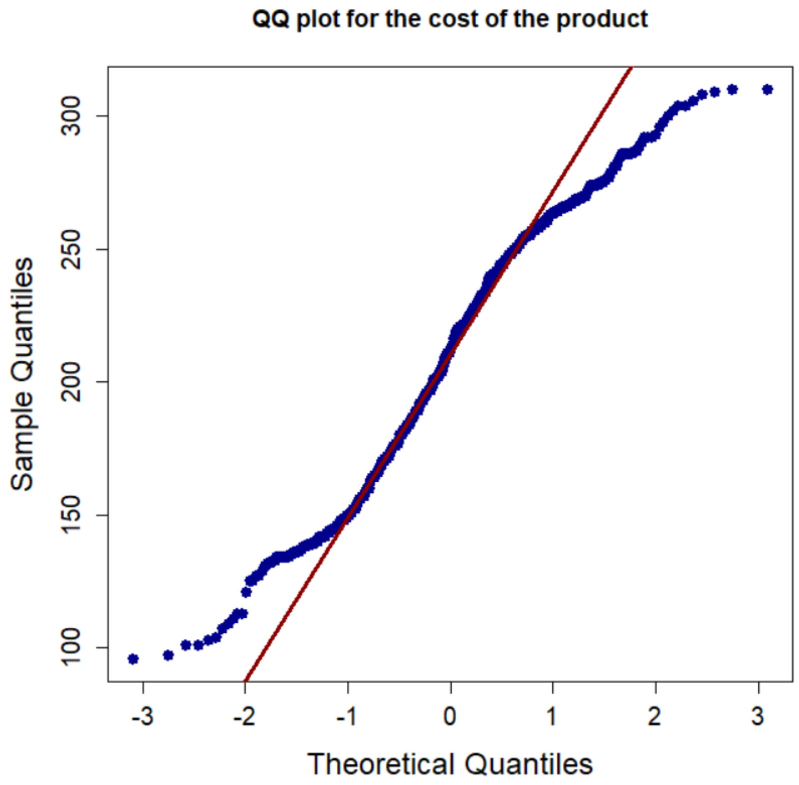


Fig.6 QQ plot for the cost of product variate

(b) Gaussian model is not appropriate. Note that the points lie along an S-shape curve, which means the relative frequency histogram should be symmetric and with short tails. Thus, maybe the uniform model can fit better than Gaussian model.

5. In this question you are going to analyze the Warehouse_block variate. Let y_1, y_2, y_3, y_4, y_5 represent the observed number of products that are retrieved from each of warehouse block A, B, C, D, and E respectively in your dataset of sample size 500. Assume a multinomial model is appropriate for modelling the warehouse block of a shipped product. Let $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ be the true proportion of products stored in warehouse blocks A, B, C, D, and E.
   a) Derive the likelihood and log likelihood function for the parameters $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$. Be sure to include the parameter space.
   b) Give a function for the maximum likelihood estimate for each of $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$. You can directly state results from the course notes. You do not have to do the maximization using Lagrange multipliers.
   c) Compute the estimates of $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ for your dataset.
   d) 10 products are to be shipped on a certain day. Find the maximum likelihood estimate of the probability that 4 products are located in Block A, 2 products are located in Block B, and the remaining products are located in one of the remaining blocks C, D, and E. What property are you using the find this maximum likelihood estimate?

Solution:

(a) According to the lecture note (2.1), the joint probability for observing $y_1, \ldots, y_5$ in a multinomial distribution with parameter θ is:

$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4, Y_5 = y_5;\ \theta) =$

$\frac{500!}{y_1!y_2!y_3!y_4!y_5!}\theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3}\theta_4^{y_4}\theta_5^{y_5}$, where $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, $0 \leqslant \theta_i \leqslant 1$ for i = 1, 2, 3, 4, 5, and $\sum_{i=1}^{5}\theta_i = 1$.

Then, we delete constant terms to get the likelihood function:

$L(\theta) = \theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3}\theta_4^{y_4}\theta_5^{y_5} = \prod_{i=1}^{5}\theta_i^{y_i}$, where $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, $0 \leqslant \theta_i \leqslant 1$ for i = 1, 2, 3, 4, 5, and $\sum_{i=1}^{5}\theta_i = 1$.

Then, we have the log likelihood function:

$G(\theta) = \log(\theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3}\theta_4^{y_4}\theta_5^{y_5}) = \sum_{i=1}^{5} y_i log\theta_i$, where $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, $0 \leqslant \theta_i \leqslant 1$ for i = 1, 2, 3, 4, 5, and $\sum_{i=1}^{5}\theta_i = 1$.

(b) By the lecture note 2.4, the maximum likelihood for the multinomial model is $\hat{\theta_i} = \frac{y_i}{500}$ for i = 1, 2, 3, 4, 5.

(c) According to Fig.7, which is the calculation result based on R and the formula $\widehat{\theta_\iota} = \frac{y_i}{500}$ for i = 1, 2, 3, 4, 5, we have the maximum likelihood estimates of θ:

$\hat{\theta} = (\widehat{\theta_1}, \widehat{\theta_2}, \widehat{\theta_3}, \widehat{\theta_4}, \widehat{\theta_5}) = (0.156, 0.172, 0.184, 0.142, 0.346)$

```
> theta_vector
[1] 0.156 0.172 0.184 0.142 0.346
```

Fig.7 The calculation result for the maximum likelihood estimates of θ

(d) By multinomial distribution, the probability is:

$Z(\theta) = P(Y_1 = 4, Y_2 = 2, Y_3 = 4, Y_4 = 0, Y_5 = 0;\ \theta) + P(Y_1 = 4, Y_2 = 2, Y_3 = 0, Y_4 = 4, Y_5 = 0;\ \theta) + P(Y_1 = 4, Y_2 = 2, Y_3 = 0, Y_4 = 0, Y_5 = 4;\ \theta) =$

$$\frac{10!}{4!2!4!0!0!}\theta_1^{4}\theta_2^{2}\theta_3^{4}\theta_4^{0}\theta_5^{0} + \frac{10!}{4!2!4!0!0!}\theta_1^{4}\theta_2^{2}\theta_3^{0}\theta_4^{4}\theta_5^{0} + \frac{10!}{4!2!4!0!0!}\theta_1^{4}\theta_2^{2}\theta_3^{0}\theta_4^{0}\theta_5^{4} = 3150\theta_1^{4}\theta_2^{2}(\theta_3^{4} + \theta_4^{4} + \theta_5^{4})$$

By the Invariance Property of the Maximum Likelihood Estimates, we know

$Z(\hat{\theta}) = 3150\widehat{\theta_1}^4\widehat{\theta_2}^2(\widehat{\theta_3}^4 + \widehat{\theta_4}^4 + \widehat{\theta_5}^4) = 3150\times0.156^4\times0.172^2\times(0.184^4 + 0.142^4 + 0.346^4) \approx 0.0008766896$

Note that this result is calculated using R, which is included in my uploaded R script.

6. a) Is this an observational study of experimental study? Justify your answer.
   b) Clearly define the Problem for this study in one or two sentences.
   c) What type of Problem (descriptive, causative, predictive) is this? Be sure to justify your answer.
   d) The target population is not explicitly listed in the article. Suggest a reasonable target population/process for this study.
   e) Based on the information provided in the article, define the study population/process for this study.
   f) Give 3 variates which were collected in this study and indicate the type of each.
   g) For each of the variates you chose in (f) give an attribute of the study population which could be estimated/determined in this study.
   h) Define study error. Give one example of study error in relation to your answers to (d), (e), and (g).
   i) Define measurement error. Give one example of measurement error for this study.
   j) Clearly state the main conclusion of this study in one or two statements.
   k) Describe an important limitation to this study.

(a) This is an observational study. This is because the target population of interest is the people recently or in the future, which is an infinite conceptual population. Also, the data are collected without any attempt to change the value of more variates for the sampled units (i.e. the interviewed 45000 people aged 15 years and over). This is not an experimental study because the experimenter does not change the values of one or more variates for the units in the sample.

(b) A UK study planned a longitudinal study to determine the relationship between vegetable/fruit consumption and mental health for people.

(c) This is descriptive, since the target population is all people recently or in the future, and the vegetable/fruit consumption and the mental health are variates of interest of the population, and the problem asks for the relationship between the variates as the attribute we want to determine.

(d) The target population should be the people recently or in the future.

(e) The study population for this study is the people aged 15 years and over in UK.

(f) 1. Mental health is a variate that is recorded in the study, whose type is continuous since the GHQ-12 scores are real numbers; 2. Age is a variate that is recorded in the study, which is discrete since ages are integer numbers; 3. Gender is a variate that is recorded in the study, which is categorical, since we only have several types of genders.

(g) 1. For mental health, we can determine the average increase in the GHQ-12 score (i.e. mental health) with an increase of one portion of vegetable/fruit intake; 2. For age, we can determine the average increase of intake of fruits and vegetables with an increase of one year of the age; 3. For gender, we can determine the sample proportion of each genders.

(h) One study error is that the average increase of intake of fruits and vegetables with an increase of one year of the age may differ in other cultures other than in UK. For example, in China, old men do not eat some vegetables and fruits such as mulberry because of cultural factors.

(i) One measurement error may be the inaccuracy of the GHQ-12 questionnaire and the response bias from some people who did the test. For example, people may lie about their feeling in some sensitive questions in GHQ-12.

(j) Mental well-being has a positive and statistically significant relationship with fruits and vegetables consumption.

(k) A limitation would be that conclusion does not give a causative relationship between eating vegetables and fruits and mental well-being, which should be valuable because people who are well-being mentally may tend to eat more fruits and vegetables.

7. 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.

In Q1-3, I become familiar with QQ plots for typical distributions (e.g. normal, exponential, and t distributions), and I trained my ability to interpret QQ plots for the real-life data in order to impose the features (e.g. symmetry, tails, etc) of the data relative to normal distributions from its QQ plot in question 4, and then decide whether normal distribution is suitable for the data, or which distribution is better to fit the data. Question 5 checked my ability to derive the likelihood and the log likelihood function for an unknown parameter for a specific model (i.e. multinomial model), and I computed the maximum estimates from the sample data, and predicted an event using the maximum estimates according to the invariance property. Finally in Q6, I tried to find the target population, study population, samples, variates, attributes, errors, problems, the conclusion, and the limitation of a real-life article using PPDAC steps, which makes me familiar with the steps better.