Problem:

(1) In this study, the units are the purchases at the individual product level, and the target population is all purchases made by customers at GoodBuy in 2020.

(2) Variates of interests and the corresponding attributes of interests in the target population:

Customer rating: Its type is ordinal, since it has natural ordering to represent the satisfaction of customers; An attribute of interest may be the distribution of customer ratings for both genders in the target population.

Mode of Shipment: Its type is categorical, since it does not take on numerical values; An attribute of interest may be the total weight of purchases shipped by flights in grams.

Discount offered: Its type is continuous, since the unit of the variate is %, and theoretically, non-integer percent of discount is possible; An attribute of interest may be the mean discount given to customers in the target population.

Customer care calls: Its type is discrete, since it does not have non-integer values and the value must be non-negative; An attribute of interest may be the mean number of customer care calls in the target population.

Weight in grams: Its type is continuous, since things can have non-integer grams of weight; An attribute of interest may be the standard deviation for weight among items shipped by ship in the target population.

Reached on Time: Its type is binary, since it only has 2 possible values. An attribute of interest may be proportion of the target population was delivered by the expected delivery time.

(3) Motivating questions for the variates:

Motivating question for the Customer rating: Is the distribution of customer rating similar for both genders in the target population? This is a descriptive problem, since the problem is to determine if there is a significant difference between the customer rating from males and the customer rating from females, which is a function of the variate.

Motivating question for the Discount offered: What is the mean discount given to customers in the target population? This is a descriptive problem, because the problem is to determine the average discount offered, which is a function of the variate.

Motivating question for the Customer care calls: What is the mean number of customer care calls in the target population? This is a descriptive problem, because the problem is to determine the average number of customer care calls, which is a function of the variate.

Motivating question for the Weight in gms: What is the mean and standard deviation for weight among items shipped by ship in the target population? This is a descriptive problem, because the problem is to determine the mean and standard deviation for weight in gms among items shipped by ship in the target population, which is a function of the variate.

Motivating question for the Reached on Time: What proportion of the target population was delivered by the expected delivery time? This is a descriptive problem, since the problem is to determine the proportion of the target population was delivered by the expected delivery time, which is a function of the variate.

Plan:

(1) Possible sources of study error: We only study the purchases in the first 2 months in 2020. However, customers' consumption behaviour and discount offered may differ much in the following months due to the festivals, annual bonus, seasons, etc. Also, the efficacy of different or the same modes of shipment are different due to weathers, seasons, etc.

(2) Possible sources of sample error: For the variate Reached on time, many purchases that are not reached on time are likely to be returned or canceled by the customer, which means some attributes in the sample, such as the proportion of the purchases that are not reached on time, may differ from the study population. The proportion of the purchases that are not reached on time in the sample is expected to be lower than the study population.

(3) Measurement system for weight: Machines that weigh the products in grams. Possible sources of measurement error: The inaccuracy of the machine used to weigh the products. Also, since computers use floating point systems, the systems themselves have some error when storing non-integer (i.e. continuous) data; Measurement system for reached on time: Recorded and uploaded to the database by the mailer once he/she has sent the mail to the destination. Possible sources of measurement error: A mailer may have sent the product to a wrong destination within the expected time, but did not realize the destination was wrong, and recorded the purchase to be "reached on time". Measurement system for customer rating: Customers who own the purchases rate for the purchases on the website using smart devices. Possible sources of measurement error: Customers may not be familiar with the rating scheme or the correct way of rating. They also may be too lazy to think about if they are satisfied with the purchase, and then give rating that does not follow their true feelings.

Data:

1. Data may be missing because of lack of database maintenance, which influences the analysis and conclusion badly. We should backup data frequently and maintain the database carefully.
2. Mistakes in data, such as unit inconsistency, may cause great outlier, which influences the analysis and conclusion badly. We would better set a proper domain for each variate of interests to avoid the mistakes.

Analysis:

Q1:

a) The relative frequency of Customer rating is in Fig.1 below.

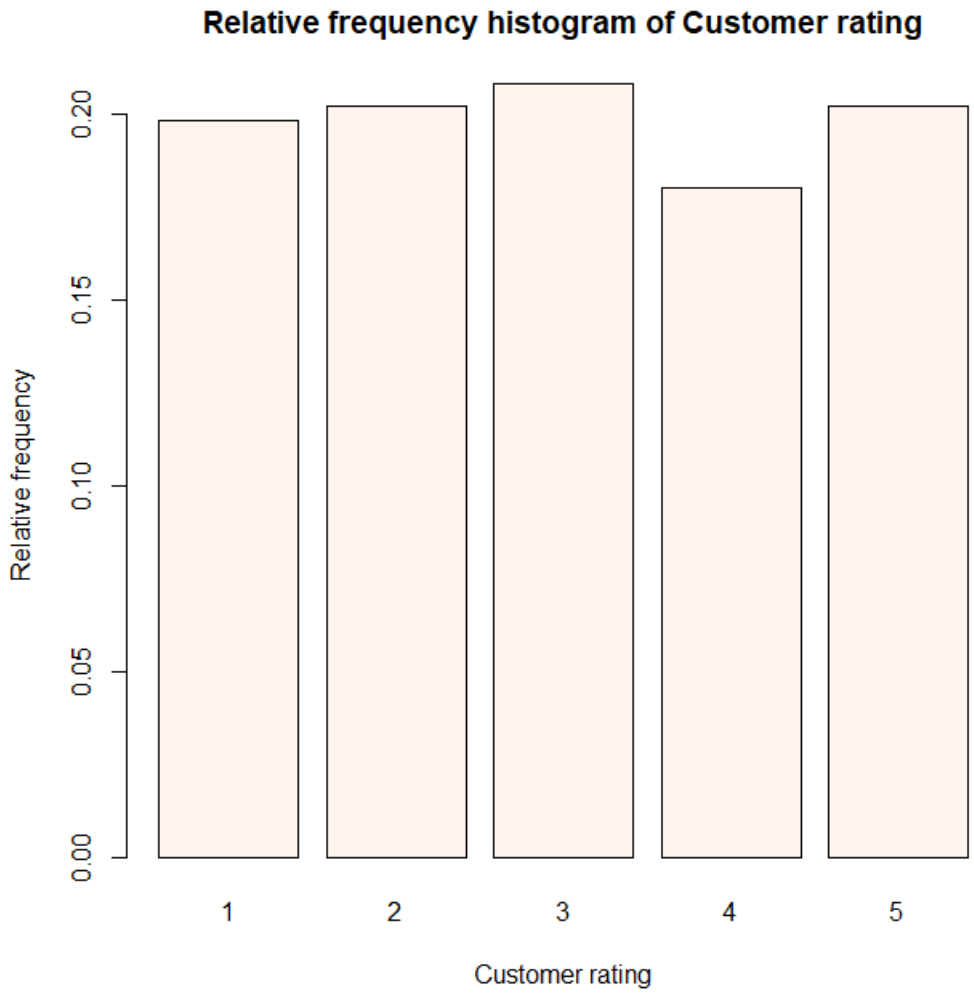

Fig.1 Relative frequency of Customer rating

b) The result using R script is:

```
   1  2  3  4  5
F 61 47 48 46 53
M 36 53 55 43 48
```

Therefore, the Frequency of each customer rating level by gender is in Tab.1 below.

		Customer Rating				
		1	2	3	4	5
Gender	M	61	47	48	46	53
	F	36	53	55	43	48

Tab.1 Frequency of each customer rating level by gender

c) The side-by-side bar graphs for Customer rating by Gender is in Fig.2.

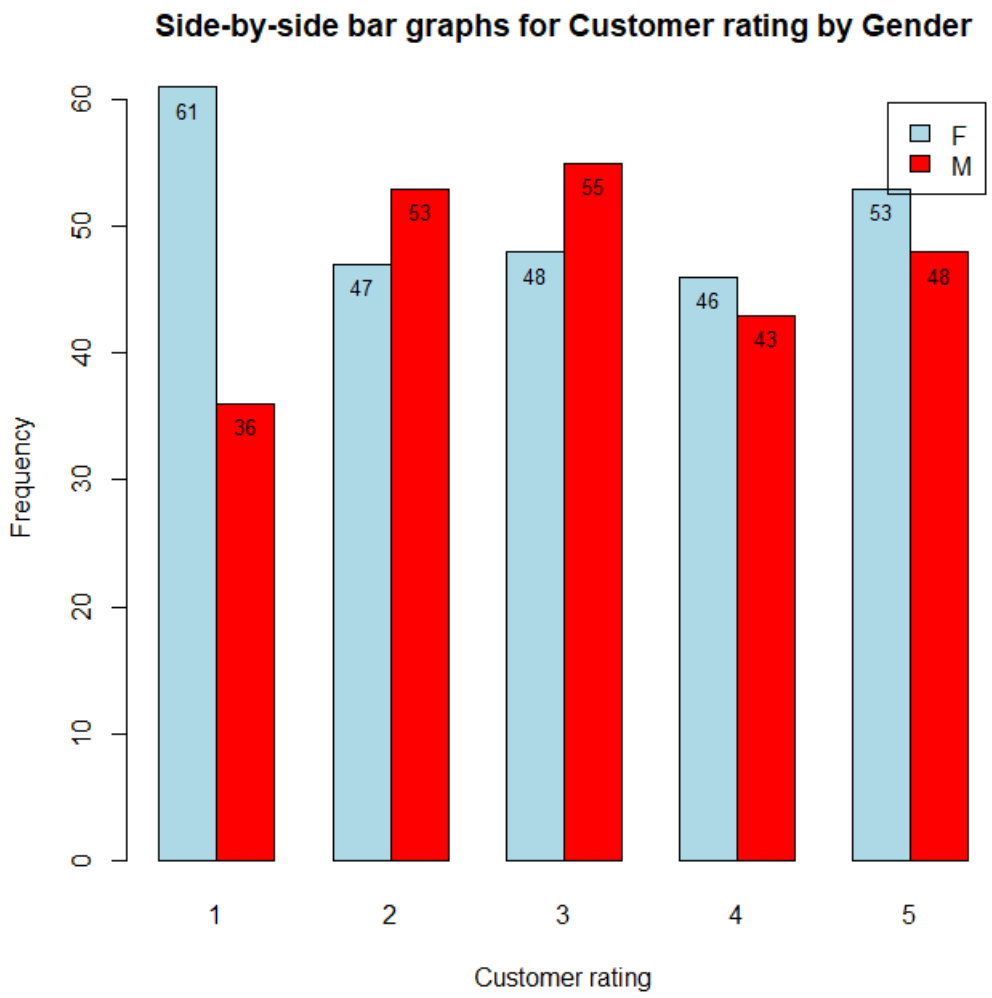

Fig.2 Side-by-side bar graphs for Customer rating by Gender

d) Since the customer rating can be regarded as randomly chosen, and it only has 2 outcomes: either “Poor rating” or “Excellent rating”, so Binomial(n, θ) is reasonable for the variate.

e) Since the output using R script is:

```
          0         1
0 0.2632653 0.1693878
1 0.3489796 0.2183673
```

We have the table in Tab.2:

		Rating	
		0	1
Reached on Time	Y = 1	0.2632653	0.1693878
	N = 0	0.3489796	0.2183673

Tab.2 relative frequency of each Rating level by whether the item was delivered on time or not

f) Since we have shown the variate satisfies binomial model, the maximum likelihood estimate is the proportion of "Excellent rating", which is 0.3877551 according to the R script result.

```
      0         1
6122449 0.3877551
```

g) The relative likelihood function plot is in Fig.3.

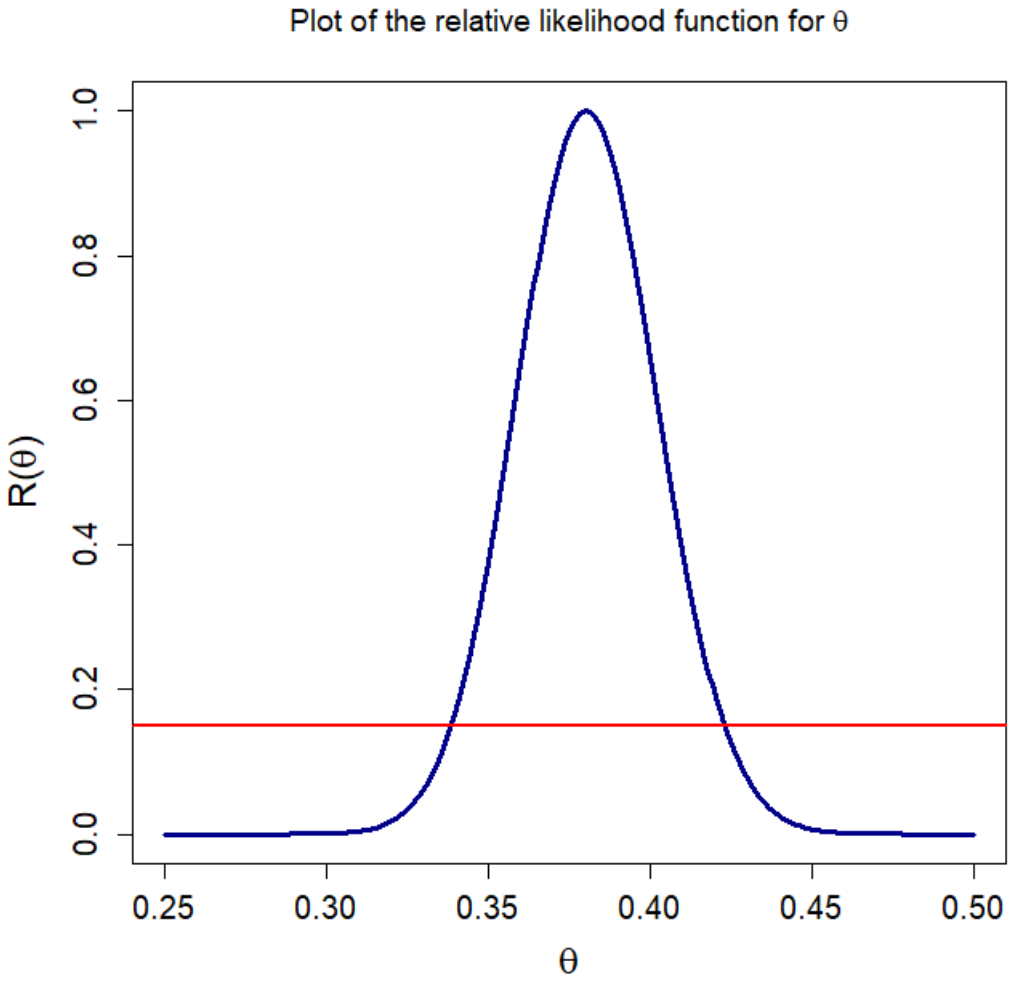

Fig.3 Relative likelihood function plot for θ

Then, according to the result of uniroot, the 15% likelihood interval is [0.3455392, 0.431122].

Q2:

a) Assuming the variate Discount offered follows Exponential(θ) model, θ corresponds to the mean discount offered in the study population.

b) The numerical summaries obtained using R are in Tab.3.

Minimum	1	IQR	7
1st Quartile	3	Range	62
Sample Median	7	Sample Mean = $\bar{y}$	12.21
3rd Quartile	10	Sample Standard Deviation = s	14.88383
Maximum	63	Sample Skewness	1.870484

Tab.3 Numerical summaries of discount offered

c) The relative likelihood function for θ is demonstrated in Fig.4.

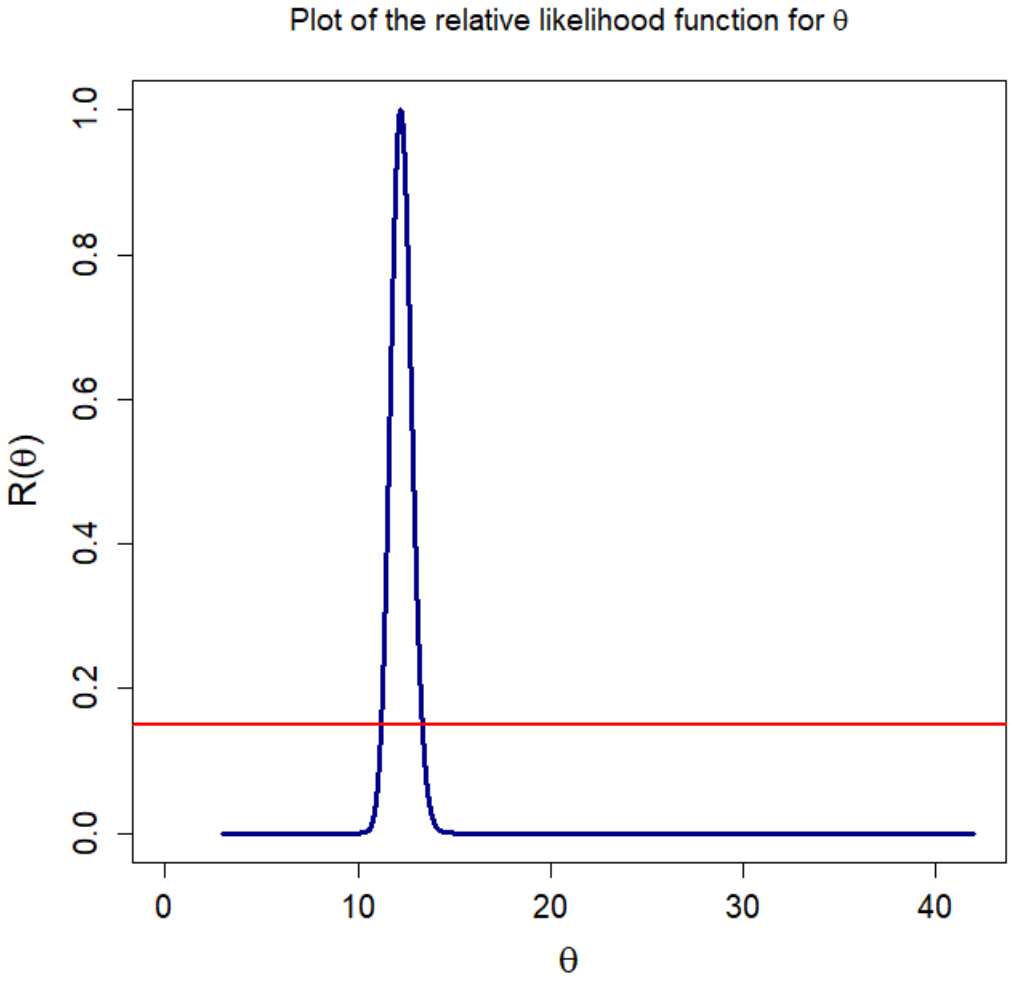

Fig.4 Plot of the relative likelihood function for θ

According to the R script result, the 15% likelihood interval for the mean discount offered in the study population is [11.19591, 13.3508].

d) Let $\tilde{\theta} = \bar{Y}$. Since the data is assumed to follow exponential model, E($\tilde{\theta}$) = θ, sd($\tilde{\theta}$) = $\frac{\theta}{\sqrt{n}}$. Let $Q_n = \frac{\tilde{\theta} - \theta}{\theta / \sqrt{n}}$, we know it approximately follows G(0, 1) when n is large.

Thus,

$$P\left(-qnorm(0.95, 0, 1) \leq \frac{\tilde{\theta} - \theta}{\frac{\theta}{\sqrt{n}}} \leq qnorm(0.95, 0, 1) \right) = 0.90$$

$$P\left(-1.644854 \leq \frac{\tilde{\theta} - \theta}{\frac{\theta}{\sqrt{n}}} \leq 1.644854 \right) = 0.90$$

$$P\left(-\tilde{\theta} - 1.644854\frac{\theta}{\sqrt{n}} \leq -\theta \leq -\tilde{\theta} + 1.644854\frac{\theta}{\sqrt{n}} \right) = 0.90$$

$$P\left(\tilde{\theta} - 1.644854\frac{\theta}{\sqrt{n}} \leq \theta \leq \tilde{\theta} + 1.644854\frac{\theta}{\sqrt{n}} \right) = 0.90$$

$$P\left(\hat{\theta} - 1.644854\frac{\hat{\theta}}{\sqrt{n}} \leq \theta \leq \hat{\theta} + 1.644854\frac{\hat{\theta}}{\sqrt{n}} \right) = 0.90$$

$$P\left(12.21 - 1.644854\frac{12.21}{\sqrt{490}} \leq \theta \leq 12.21 + 1.644854\frac{12.21}{\sqrt{490}} \right) = 0.90$$

$$P(11.30271 \leq \theta \leq 13.11729) = 0.90$$

Therefore, $[11.30271, 13.11729]$ is approximately the 0.90 confidence interval for the mean discount offered in the study population.

Q3:

a) Assuming the data follow Poisson model, the parameter θ corresponds to the mean number of customer care calls per purchase in the study population.

b) According to the result of the R script, we have the numerical summaries in Tab.4.

Minimum	0	IQR	2
1st Quartile	3	Range	7
Sample Median	4	Sample Mean = $\bar{y}$	3.978
3rd Quartile	5	Sample Standard Deviation = s	1.286037
Maximum	7	Sample Skewness	-0.3335122

Tab.4 Numerical summaries for the Customer care calls data

c) According to the result of R script, we have the table Tab.5.

Number of Customer Care calls	Observed Frequency	Expected Frequency
0	6	9.178414
1	12	36.507608
2	31	72.605437
3	118	96.263943
4	154	95.723686
≥ 5	169	179.720912
Total	490	490

Tab.5 Table of observed frequencies and expected frequencies determined using Poisson($\hat{\theta}$) model

d) According to the numerical summaries in Tab.4: The sample median is too close to the sample mean, and the sample skewness is too close to zero, which all indicate that the data is too symmetry and does not have a longer tight tail. Also, the sample variance is $1.286037^2 = 1.653891$, which is not close to the sample mean 3.978. Therefore, Poisson model is not reasonable for the data since the data do not have similar features as a Poisson distribution.

According to the frequency table in Tab.5: The expected frequency determined using Poisson($\hat{\theta}$) model is too far from the observed frequency for every number of customer care calls, which indicates that the data do not follow the Poisson distribution.

Overall, Poisson model is not reasonable for the data.

Q4:

a) Assuming the variate weight in grams can be modeled using $G(\mu_i, \sigma_i)$ for i = 1, 2, 3, μ_1, μ_2, μ_3 correspond to the mean weight in grams for purchases shipped by ship, flight , and road in the study population respectively; $\sigma_1, \sigma_2, \sigma_3$ correspond to the standard deviation for purchases shipped by ship, flight, and road in the study population respectively.

b) According to the R script result, we have the numerical summaries for the weight products shipped by Ship only Tab.6.

Minimum	1005	IQR	3278
1st Quartile	1851	Range	4990
Sample Median	4179	Sample Mean = $\bar{y}$	3663
3rd Quartile	5129	Sample Standard Deviation = s	1663.533
Maximum	5995	Sample Skewness	-0.2883923
		Sample Kurtosis	1.540062

Tab.6 Numerical summaries for the weight products shipped by Ship only

c) The qqplot is shown in Fig.5.

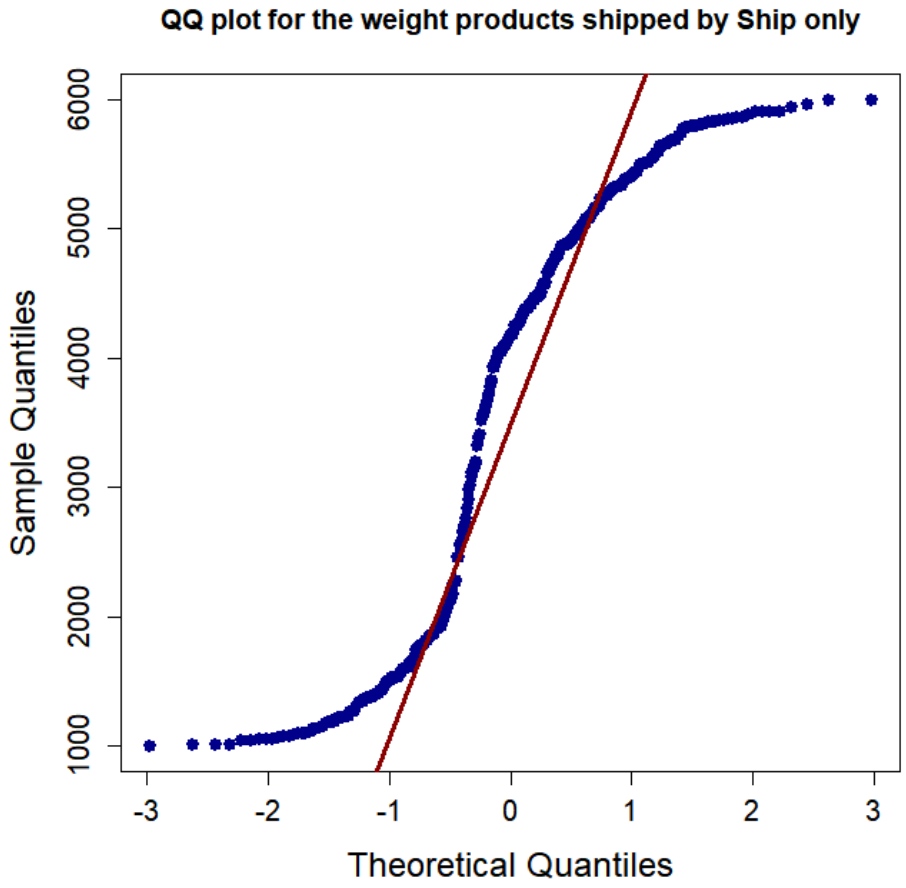

Fig.5 Qqplot of the weight of products shipped by ship only

d) Gaussian model is not suitable for these data. First, according to the qqplot, very few data points lie on the line. Then, the data points form an S-shaped curve, which is against the features of Gaussian distribution. Then, according to the numerical summaries, the sample kurtosis is too far from 3, which means the data basically has no tails. Therefore, all those features indicate that Gaussian model is not suitable for these data.

e) Let T = $\frac{\bar{Y}-\mu_1}{S/\sqrt{n}}$, then by Theorem 36, T ~ t(n - 1). Thus, by Table 4.4, we have the 95% confidence interval for μ_1:

$$\left[\bar{y}-\frac{bs}{\sqrt{n}},\bar{y}+\frac{bs}{\sqrt{n}}\right]$$

$$= \left[3663 - \frac{qt(0.975,n-1)*1663.533}{\sqrt{490}}, 3663 + \frac{qt(0.975,n-1)*1663.533}{\sqrt{n}}\right]$$

$$= \left[3663 - \frac{1.964827*1663.533}{\sqrt{490}}, 3663 + \frac{1.964827*1663.533}{\sqrt{490}}\right] = [3515.342, 3810.658]$$

The 90% confidence interval for σ_1 is:

$$\left[\sqrt{\frac{(n-1)s^2}{qchisq(0.95,n-1)}}, \sqrt{\frac{(n-1)s^2}{qchisq(0.05,n-1)}}\right]$$

$$= \left[\sqrt{\frac{(490-1)*1663.533^2}{qchisq(0.95,490-1)}}, \sqrt{\frac{(490-1)*1663.533^2}{qchisq(0.05,490-1)}}\right]$$

$$= [1580.761, 1756.269]$$

f) According to the R script, we have the numerical summaries for the weight products shipped by Road only Tab.7.

Minimum	1019	IQR	3374.75
1st Quartile	1930	Range	4981
Sample Median	4296	Sample Mean = $\bar{y}$	3850
3rd Quartile	5305	Sample Standard Deviation = s	1647.761
Maximum	6000	Sample Skewness	-0.5180879
		Sample Kurtosis	1.778084

Tab.7 Numerical summaries for the weight products shipped by Road only

g) The qqplot is shown in Fig.6.

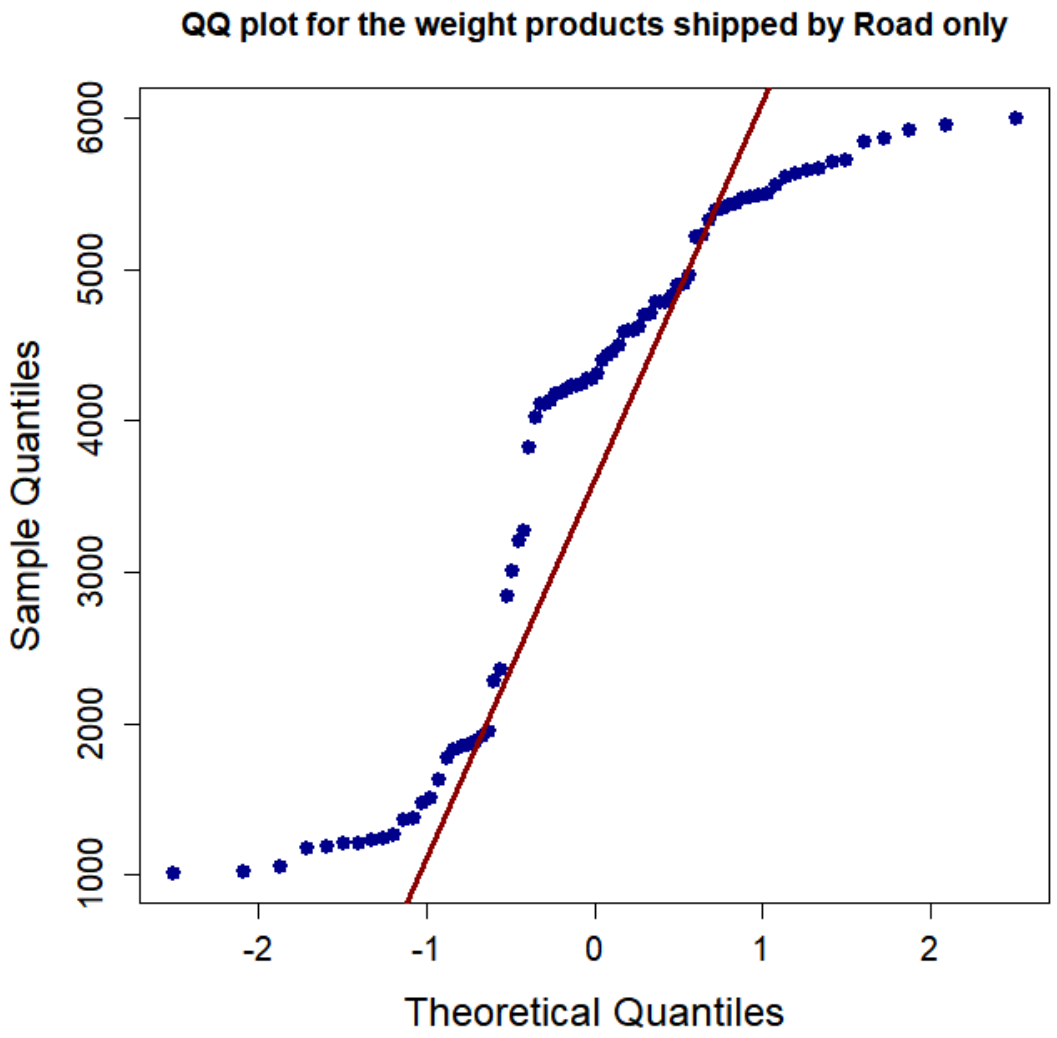

Fig.6 QQ plot for the weight products shipped by Road only

h) Gaussian model is not suitable for these data. First, according to the qqplot, very few data points lie on the line. Then, the data points form an S-shaped curve, which is against the features of Gaussian distribution. Then, according to the numerical summaries, the sample kurtosis is too far from 3, which means the data basically has no tails. Therefore, all those features indicate that Gaussian model is not suitable for these data.

i) Let $T = \frac{\bar{Y} - \mu_3}{S/\sqrt{n}}$, then by Theorem 36, T ~ t(n - 1). Thus, by Table 4.4, we have the 95% confidence interval for μ_3:

$$\left[\bar{y} - \frac{bs}{\sqrt{n}}, \bar{y} + \frac{bs}{\sqrt{n}}\right]$$

$$= \left[3850 - \frac{qt(0.975, n-1) * 1647.761}{\sqrt{490}}, 3850 + \frac{qt(0.975, n-1) * 1647.761}{\sqrt{n}}\right]$$

$$= \left[3850 - \frac{1.964827 * 1647.761}{\sqrt{490}}, 3850 + \frac{1.964827 * 1647.761}{\sqrt{490}}\right]$$

$$= [3703.742, 3996.258]$$

The 90% confidence interval for σ_3 is:

$$\left[\sqrt{\frac{(n-1)s^2}{qchisq(0.95, n-1)}}, \sqrt{\frac{(n-1)s^2}{qchisq(0.05, n-1)}}\right]$$

$$= \left[\sqrt{\frac{(490-1) * 1647.761^2}{qchisq(0.95, 490-1)}}, \sqrt{\frac{(490-1) * 1647.761^2}{qchisq(0.05, 490-1)}}\right]$$

$$= [1565.773, 1739.617]$$

j) The side-by-side boxplot is demonstrated in Fig.7.

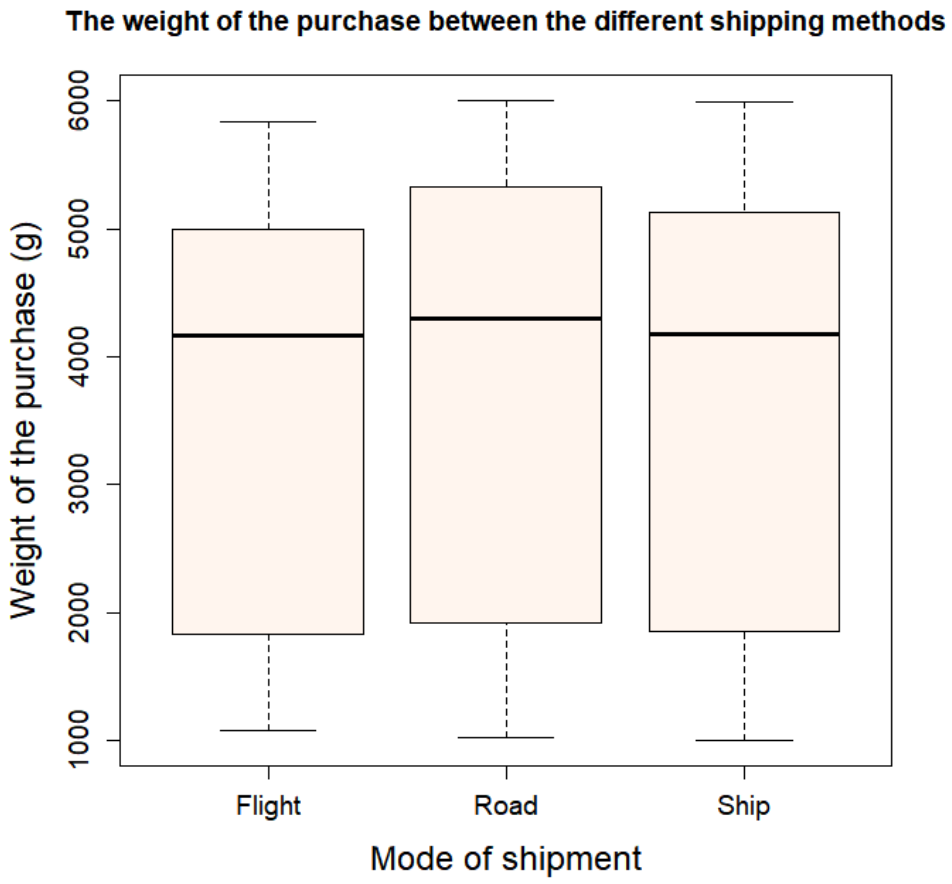

Fig.7 Side-by-side boxplots for the weight of products by mode of shipment

According to Fig.7, 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, which means they are not symmetric.
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, but the flight mode has narrower range than other 2 modes slightly.
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.

Q5:

a) Assuming the variate can be modeled by Binomial, the parameter θ corresponds to the proportion of purchases that are successfully reached on time in the study population.

b) Since the maximum likelihood estimate for the variate that is assumed to follow Binomial model should be the sample proportion of purchases that are successfully reached on time, the maximum likelihood estimate is 0.5673469.

c) According to the Table 4.4 in lecture note, we have the 95% confidence interval for θ:

$$\left[\bar{y} - qnorm(0.975) * \sqrt{\frac{\bar{y}(1-\bar{y})}{n}}, \bar{y} - qnorm(0.975) * \sqrt{\frac{\bar{y}(1-\bar{y})}{n}}\right]$$

$$= \left[0.5673469 - 1.959964 * \sqrt{\frac{0.5673469(1-0.5673469)}{490}}, 0.5673469 - 1.959964 * \sqrt{\frac{0.5673469(1-0.5673469)}{490}}\right]$$

$$= [0.5234793, 0.6112146]$$

d) By the Invariance Property of Maximum Likelihood Estimate, the maximum likelihood estimate of the probability that at least 23 are delivered on time is:

$$g(x \geq 23;\ \theta = \hat{\theta} = 0.5673469) = 1 - pbinom(22, 50, 0.5673469)$$

$$= 0.9523774$$

e) By using the R script, we see the proportion of deliveries that arrived by the expected delivery date for the high-importance products is 0.5869565:

```
high_impor_Reach
        0         1
0.4130435 0.5869565
```

f) According to the result of the R script:

```
      high       low    medium
0.4130435 0.4568966 0.4103774
0.5869565 0.5431034 0.5896226
```

We have the table Tab.8.

		Product Importance		
		Low	Medium	High
Reached on Time	Y = 1	0.5431034	0.5896226	0.5869565
	N = 0	0.4568966	0.4103774	0.4130435

Tab.8 Relative frequency of each Product importance level by whether the item was delivered on time or not

g) The side-by-side bar plot for the proportion of products delivered on time by product importance is in Fig.8.

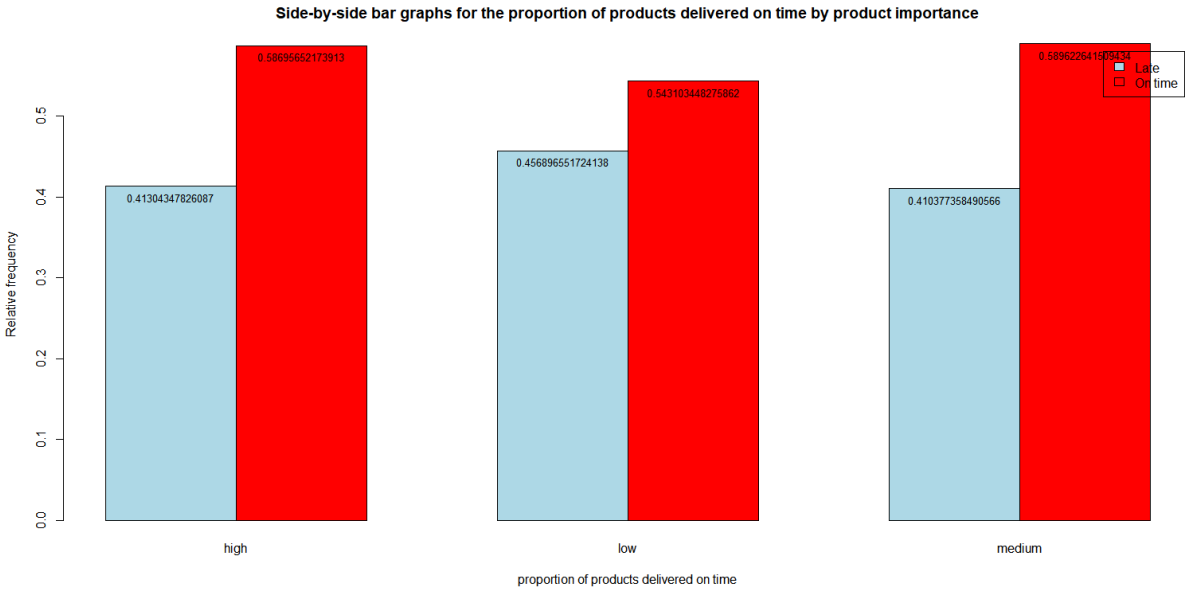

Fig.8 Side-by-side bar plot for the proportion of products delivered on time by product importance

According to Fig.8, low-importance products have the highest rate of late deliveries, whereas high-importance and medium-importance products have very close proportion of late deliveries, which are all much lower late rate than the low-importance products. Also, the proportion of late deliveries of low-importance products is the closest to the proportion of on-time deliveries, but the proportions of late deliveries of high-importance and medium-importance products are not close to the proportions of the on-time deliveries. For the similarity, all the three importance levels of products all have smaller late rates than the on-time rates.

Conclusion:

(1) In the study population, males tend to give extreme ratings, which means they are more likely to rate 1 or 5, whereas females tend to give medium ratings, which means they are more likely to rate 2 and 3. And they are approximately equally likely to rate 4. However, female customers may not be familiar with the rating scheme, rating devices, or the correct way of rating as male customers do. Thus, they may give ratings that do not match their truth feelings, which is a sources of measurement error. Also, the analysis does not give a causative relationship between gender and rating, which is a flaw.

(2) We have 15% likelihood interval that the mean discount given to customers in the study population may be within the interval [11.19591, 13.3508], and we have 90% confidence interval that the mean discount given to the customer in the study population may lie within the interval [11.30271,13.11729]. The maximum likelihood estimate of the mean discount given to customers in the study population is 12.21, which is the sample mean of the data. However, the given interval estimates and the maximum likelihood estimate do not mean the probability that the true mean discount offered in the study population lying within the intervals is certain, or the mean discount given to customers in the study population must be the maximum likelihood estimate. The conclusion can be improved by enlarging the sample size. Additionally, the conclusion can only apply to the study population, which is purchases made by customers at GoodBuy in the first 2 months in 2020. However, discount offered may differ much in the following months due to the festivals, annual bonus, seasons, etc. Thus, the conclusion includes much study error.

(3) The sample mean of the number of the customer care calls is 3.978, which is usually the maximum likelihood estimate for the mean of the number of the customer care calls in the study population. However, since we do not know which model fits the data, we can not give an interval estimate to quantify the uncertainty, which is a drawback of the study. Also, the conclusion only applies to the study population rather than the target population.

(4) The sample mean for weight among iterms shipped by ship is 3663g, and the sample standard deviation for weight among iterms shipped by ship is 1663.533, which are approximately the maximum likelihood estimates for the mean weight among iterms shipped by ship in the study population. We have 95% confidence interval that the true mean weight among iterms shipped by ship in the study population may lie within the interval [3515.342,3810.658]. And we have 90% confidence interval that the true standard deviation of weight among iterms shipped by ship in the study population may lie within the interval [1580.761,1756.269]. However, this conclusion can be inaccurate since the normality assumption is considered to be unsatisfied, but we still obtain the estimates using the assumption. Also, the conclusion only applies to study population, since the mean weight among iterms shipped by ship may change due

to the season and the weather changes, which behaves differently from the study population, which is a possible source of study error.

(5) We have a 95% confidence interval for the true proportion of the study population was delivered by the expected delivery time $[0.5234793, 0.6112146]$ based on the sample data. And the sample proportion of the purchases were delivered by the expected delivery time is 0.5673469, which is the maximum likelihood estimate for the true proportion of the study population was delivered by the expected delivery time based on the data. However, the conclusion only applies to the study population, since a possible source of study error may be that in the following months in 2020, the winter season may lead to inefficiency for shipment by ship due to the terrible voyage condition, and the summer season may lead to inefficiency for flight due to the frequent rainy days.