PI Game

Group 17

Data Champions [*]

June 15th 2021

Abstract

Welcome to our Academic Skills (acsk) Final Project! In this project, the Probability and Inference (PI) game has been studied. We make use of the report template file for this game. Furthermore, this document is written with the help of LaTeX.

1 Introduction

The following research report reflects the knowledge and skills we have learned from the Newsvendor Problem and Python. The goal of the project is to address the 7 questions of the PI Game while working with a dataset of a European bakery chain.

We structure this report in the question-based report style in order to best divide the questions amongst the group, while also trying to convey the information and results in the clearest way possible. Each question consists of an introduction, methodology, results, and conclusion in order to achieve this goal.

As stated above, we discuss and apply our theoretical knowledge of the Newsvendor problem (NVP) from the lectures and make use of these skills we have learned. The NVP has lots of real-world applications such as in manufacturing and service industries, so it is a very useful tool to use when some resources are unknown or random. Through answering the questions, we are able to use different techniques, such as use Monte Carlo Simulations, and Parametric/Nonparametric Framework, to further our study of the NVP. The purpose of learning about the NVP is to find the most optimal way to solve real world problems that deal with unknown or random variables, and The PI Game very much exemplifies this purpose.

The variables for our report:

Order Quantity	Demand	Cost	Price	Profit	Holding Cost
Q	Y	c	p	Π	c_h

Demand Sample	Sample Size	Monte Carlo Repetitions	Significance Level
$D_n = Y_i, i = 1, ...n$	n	M	α

[*] Ronan Kenny (2691815), Guus Bouwens (2701442), Tommy Link (2691913), Jelle Khanna (2711028), Muhammet Simsek (2635634)

2 Theoretical Questions

2.1 Question 1

2.1.1 Problem Description Question 1

The first theoretical question furthers our examination of the Newsvendor problem by adjusting the profit function. Recall that the profit function is $\Pi(Q,Y;c,p) = pmin\{Q,Y\} - cQ$, where the cost c only contains the per unit purchasing cost. However, the question asks us to consider a holding cost $c_h \geq 0$ per unit, which can be charged when there are surplus goods. After adjusting the profit function, the optimal order quantity function must be adjusted accordingly, which also changes the expected profit in the process. Finally, sensitivity plots of the optimal order quantity and expected profit with respect to c_h are plotted in order to allow us to have an understanding of how c_h affects the Newsvendor problem overall.

2.1.2 Methodology Question 1

The first part of the question asks us to incorporate a given holding cost (c_h) into the profit function. c_h is only charged when there is a surplus of goods (the order quantity is greater than the demand). Therefore, we can take the maximum between Q and Y whilst then subtracting Y. c_h does not directly affect the original profit function, so if we now include c_h in a new profit function, we derive the following:

$$\Pi(Q,Y;\tilde{c},\tilde{p}) = pmin\{Q,Y\} - cQ - c_h(max\{Q,Y\} - Y) \tag{1}$$

The next part of the question asks us to calculate $Q^*(F_y;\tilde{c},\tilde{p})$. It's important that the theorems about the optimal order quantity, introduced in the first lecture, are adjusted to include the holding cost. The theorem for the optimal order quantity with continuous demand follows:

Theorem 1. Suppose Y is continuous. Given $\tilde{c}$, $\tilde{p}$, and the CDF, F_Y, and assuming a density (PDF) f_Y exists,

$$Q^*(F_Y) := Q^*(F_Y;\tilde{c},\tilde{p}) = F_Y^{-1}(\frac{p-c-c_h}{p+c_h}). \tag{2}$$

Proof. By definition, (*Note that $\tilde{c}$ = [c,c_h] and $\tilde{p}$ = p.)

$\mathbb{E}[\Pi(Q,Y;\tilde{c},\tilde{p})] = \int_{\mathbb{R}} \Pi(Q,y;\tilde{c},\tilde{p})dF_Y(y)$
$= \int_{\mathbb{R}}[\Pi(Q,y;\tilde{c},\tilde{p}) = pmin\{Q,y\} - cQ - c_h(max\{Q,y\} - y)dF_Y(y)$
$= p(\int_{\mathbb{R}} min\{Q,y\}dF_Y(y) - cQ\int_{\mathbb{R}} 1dF_Y(y) - c_h(\int_{\mathbb{R}} max\{Q,y\} - y)dF_Y(y)$
$= p(\int_{\mathbb{R}} min\{Q,y\}dF_Y(y) - cQ\int_{\mathbb{R}} 1dF_Y(y) - c_h[(\int_{\mathbb{R}} max\{Q,y\}dF_Y(y) - \int_{\mathbb{R}} ydF_Y(y))]$
$= p(\int_{\mathbb{R}} min\{Q,y\}dF_Y(y) - cQ\int_{\mathbb{R}} 1dF_Y(y) - c_h(\int_{\mathbb{R}} max\{Q,y\}dF_Y(y)) + c_h(\int_{\mathbb{R}} ydF_Y(y))$
$= p(\int_{-\infty}^{Q} ydF_Y(y) + Q\int_{Q}^{\infty} 1dF_Y(y)) - cQ - c_h[Q\int_{-\infty}^{Q} 1dF_Y(y) + \int_{Q}^{\infty} ydF_Y(y)] + c_h(\int_{-\infty}^{\infty} ydF_Y(y))$
$= p[(QF_Y(Q) - \int_{-\infty}^{Q} F_Y(y)dy) + (Q - QF_Y(Q)] - cQ - c_h[QF_Y(Q) - (QF_Y(Q) - \int_{Q}^{\infty} F_Y(y)dy] +$
$c_h(\int_{-\infty}^{\infty} YdF_Y(y))$
$= p(Q - \int_{-\infty}^{Q} F_Y(y)dy) - cQ - c_h(\int_{Q}^{\infty} F_Y(y)dy) + c_h(\int_{-\infty}^{\infty} YdF_Y(y))$
$= (p-c)Q - p(\int_{-\infty}^{Q} F_Y(y)dy - c_h(\int_{Q}^{\infty} F_Y(y)) + c_h(\int_{-\infty}^{\infty} YdF_Y(y))$

Take the first derivative with respect to Q and then set it equal to 0.

$$(p-c) - pF_Y(Q)dy - c_h(1 - F_Y(Q)) = 0$$
$$(p-c) - (p + c_h)F_Y(Q) - c_h = 0$$
$$\frac{p-c+c_h}{p+c_h} = F_Y(Q) => Q^*(F_Y; \tilde{c}, \tilde{p}) = F_Y^{-1}(\frac{p-c-c_h}{p+c_h})$$

□

In this project, we only consider the optimal order quantity with continuous demand, thus we leave out the discrete theorem and proof. The discrete case can be defined very similarly to the continuous case, but the proof uses summations instead of integrals and the generalized inverse CDF($F_Y^{\leftarrow}$) instead of the inverse CDF. We can now conclude our methodology for Question 1 and move on to the results.

2.1.3 Results Question 1

The second part of the question asks to provide the optimal order quantity $Q^*(F_Y; \tilde{c}, \tilde{p})$, given $\tilde{c}$, $\tilde{p}$, and the CDF F_Y and the third part asks us to fix values for $\tilde{c}$, $\tilde{p}$, and pick the distribution. With a normal distribution, we are able to plot sensitivity graphs of $Q^*(F_Y; \tilde{c}, \tilde{p})$ and the optimal expected profit with respect to c_h:

Variable:	Q	Y	$\tilde{p}$	$\tilde{c}$ = [c,c_h]	mu	sigma
Value:	40	50	1.5	[1,0.15]	100	10

The optimal order quantity when c_h is included, given the values above, is 89.197. For comparison, the order quantity without c_h, given the values above, is 95.693. The sensitivity plots follow:

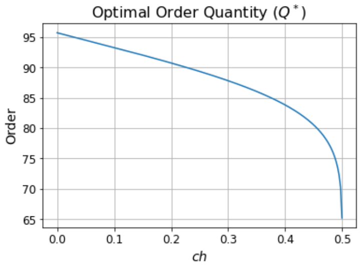


Figure 1: The Sensitivity Graph of the Optimal Order Quantity with respect to c_h

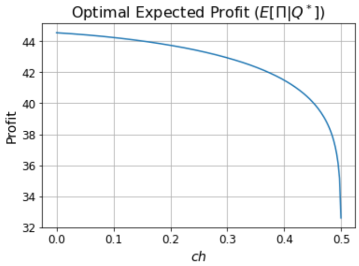


Figure 2: The Sensitivity Graph of the Optimal Expected Profit with respect to c_h

2.1.4 Conclusion Question 1

Through these results, we see the effect of adding a holding cost c_h to the profit function. For instance, the optimal order quantity without c_h (given the same values) and is 95.693, which is greater than the optimal order quantity with c_h. This implies that c_h decreases the optimal order quantity when included. Both Figure 1 and Figure 2 back this claim up. In the graphs, when c_h increases, the optimal order quantity and the optimal expected profit decrease. This intuitively makes sense, as the only difference between the original profit function and equation (1) is a decrease when the order quantity is greater than the demand. This concludes our analysis of question 1.

2.2 Question 2 and 3

2.2.1 Problem Description Question 2&3

When the distribution is known, the solution to the news vendor problem is easy to find. In practice however, this is never the case. Therefore the solution must be estimated, which can be done in different ways. In this section we compare the performance of the parametric and the non-parametric approach and illustrate strengths and weaknesses for both methods.

In the parametric approach, also known as the frequentist parametric approach(FP) it is assumed that the demand follows a parametric distribution with CDF $G(*|\theta)$ and PDF $g(*|\theta)$, where $\theta \in \Theta$. In this section we further assume that G and g are continuous. Then the solution for a given target service level τ is given by:

$$Q_n^P(G_{\widehat{\theta}_n};\tau) = G^{-1}(\tau|\widehat{\theta}_n) \tag{3}$$

Where $\widehat{\theta}$ is the maximum-likelihood estimator.

In the non-parametric approach(NP) no assumption about the underlying demand distribution is made. The solution is given by:

$$Q_n^{NP}(c,p) = \inf\{y : F_n \geq \tau\} = Y_{(\lceil \tau n \rceil)} \tag{4}$$

where F is the empirical distribution function and $Y_{(k)}$ refers to the k-smallest observation.

The obvious question is; when is which estimator preferable? As shown by Ban et al. (2020) the parametric estimator is asymptotically biased if the distribution is incorrectly specified, however Levi et al. (2015) pointed out the non-parametric estimator can be inaccurate when the target service level is large, and the sample size n is to small. We try to understand at which target service levels and sample sizes, one approach outperforms the other.

2.2.2 Methodology Question 2&3

To test when which approach is preferable, we ran Monte-Carlo simulations with M=1000 repetitions for different τ and n. The underlying distribution was either normal, log-normal, logistic or mixed normal and log-normal distribution. With the assumed distribution being either normal or log-normal. To test the performance of both methods, we focused on the profit loss ratio (PLR),but also used the empirical root mean squared error (RMSE). They are defined as the following:

$$RMSE_n^k(\tau) = \sqrt{\frac{1}{M}\sum_{j=1}^{M}[\widehat{Q}_n^{k\,j}(\tau) - Q^*(\tau)]^2}. \tag{5}$$

$$PLR_n^k(\tau) = \frac{1}{M}\sum_{j=1}^{M}\frac{|R(Q^*;\tau) - R(\widehat{Q}_n^{k\,j};\tau)|}{|R(Q^*;\tau)|}. \tag{6}$$

To compare the NP and FP approaches we looked at the ratio between the statistics for both estimators. With ratio larger than 1 indicating that the FP estimator performed better than the FP estimator.

$$\frac{RMSE_n^{NP}(\tau)}{RMSE_n^P(\tau)} \quad \text{and} \quad \frac{PLR_n^{NP}(\tau)}{PLR_n^P(\tau)} \tag{7}$$

2.2.3 Results Question 2&3

The results are very one-sided, if the distribution is correctly specified, with the FP estimator dominating the NP estimator in terms of RMSE and PLR, especially in the boundary regions of τ, as can be seen in Figure 3. This is true for all n we tried. This matches up with Theorem 7 in Levi et al. (2015), which states that, if the distribution is normal or log-normal the correctly specified parametric approach has a lower asymptotic variance than the non-parametric approach for all $\tau \in (0, 1)$. Intuitively it makes sense that the FP approach performs better compared to NP approach, when the distribution is correctly specified, because we have extra information about the underlying data. To explain the very bad result when τ is close to 0 or 1, we numerically calculate the MSE of the NP estimator, under the assumption that τn is an integer.

We need:

$$f_{Y_{(j)}}(y) = \frac{n!}{(j-1)!(n-j)!} f_y(y)[F_y(y)]^{(j-1)}[1 - F_y(y)]^{(n-j)}. \tag{8}$$

which is straightforward to show and also a Theorem in [3]. Then:

$$E[Y_j] = \int_{-\infty}^{\infty} y f_{Y_{(j)}}(y)\, dy. \tag{9}$$

$$MSE(Q_n^{NP}, Q^*) = (E[Y_j - F^{-1}(\tau))^2 + E[Y_j^2] - E[Y_j]^2. \tag{10}$$

Which we can calculate numerically. The results of this for the normal distribution are shown in Figure 3. The increasing MSE at the edges explains the bad performance of the NP estimator if τ is close to 0 or 1.

If the distribution is misspecified, the NP estimator outperforms the ML estimator most of the time. Examples can be found in Figure 4. One exception being when assuming a log-normal distribution, even though the underlying data is normally distributed Figure 4. As Ban et al. pointed out, this is because the normal distribution can be very well approximated by a log-normal. The same holds for the logistic and the normal distribution. Other insightful results arise when the data follows a mixture of a normal and log-normal distribution and we assume a normal distribution. The NP approach is better for almost all τ, except when τ is around 0.4 (Figure 5). This can be explained by Figure 5 and the following: If, $G(\tau|\theta_{ML}) \approx F(\tau)$, where F is the actual distribution, then $Q_n^P \approx Q^*$. However in practice this fact is not useful as it requires information about underlying distribution.

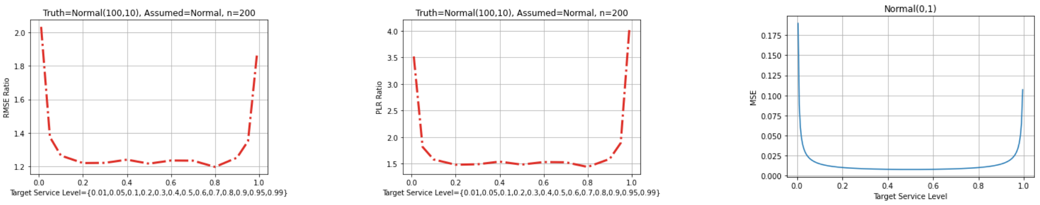


Figure 3: Performance when distribution correctly specified

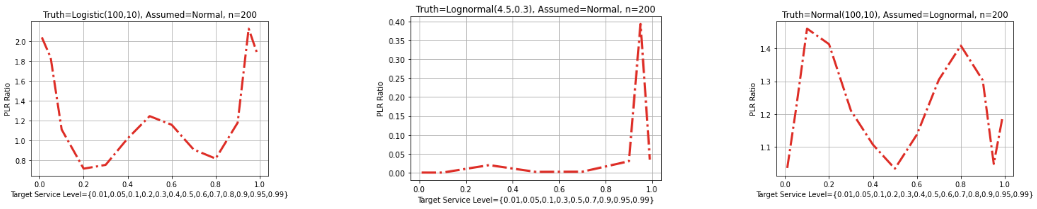


Figure 4: Performance when distribution misspecified

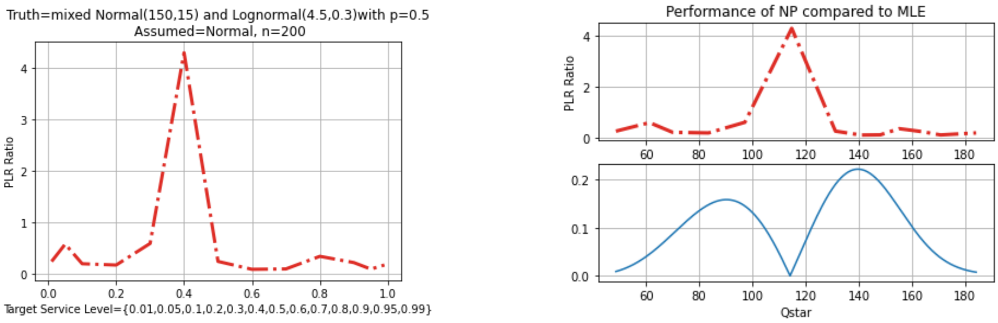


Figure 5

2.2.4 Conclusion Question 2

In practice the data is not simulated and therefore it is not possible to easily test which estimator performs better. However our results show that the parametric approach is preferable, if one can make a reliable guess about the underlying distribution. This is especially true when τ is close to 0 or 1. Should the underlying distribution be hard to determine or if no MLE can be calculated, it is probably favorable to use the NP approach.

References

[1] Model mis-specification in newsvendor decisions: Acomparison of frequentist parametric, bayesianparametric and nonparametric approaches
`https://papers.ssrn.com/sol3/papers.cfm?abstract`$_id = 3495733, 2020$

[2] Levi, R., G. Perakis, and J. Uichanco (2015). The data-driven newsvendor problem:New bounds and insights.Operations Research 63, 1294–1306
`https://doi.org/10.1287/opre.2015.1422`

[3] `http://www.math.ntu.edu.tw/ hchen/teaching/StatInference/notes/lecture37.pdf`

3 Empirical Questions

3.1 Question 4

3.1.1 Problem Description Question 4

Describe the data from an European bakery chain and do you observe any patterns in the data (especially in weekends) ? Do so by providing descriptive statistics, histograms and time series plots per store.

The dataset contains the demand data of four stores from this bakery chain. One of the shops is inside a shopping mall, the other three stores are on nearby streets. The data covers the demand of the stores individually from 26-9-2016 to 27-6-2019, which are 1005 observations in total. The costs and prices per store are as follows:

Table 1: Information on the four stores.

	distance (km)	cost	price	disposal cost
mall	0	2.95	3.27	0.07
streetA	0.8	2.53	3.27	0.02
streetB	3.2	2.21	3.27	0.02
streetC	3.3	2.19	3.27	0.02

Both the costs and prices are measured in euro per 100g of fresh pastries. The costs of storage and rent are different per store. Disposal cost is the cost for throwing the product out.

3.1.2 Methodology Question 4

We used Python to import, read and manipulate the Excel dataset about the stores.
All the statistics we found interesting are: sample mean, sum, sample mean, sample variance, minimums, median, maximums, inter quantile range, range, skewness and kurtosis. Most of these statistics will be familiar but we wanted to explain some of them:
The inter quantile range (IQR) is defined as the range of the data from the 25th until the 75th percentile. A larger IQR implies that it is more likely that the demand for that store will vary more. For the skewness a positve value means right winged and a negative value means left winged, compared to the figure of a normal distribution. For the kurtosis the normal figure (mesokurtic) would have a value of exactly 3. A lower value would imply more flat peaks (platykurtic), whereas higher values mean steeper peaks (leptokurtic).
In order to get a good observation of patterns in the data (especially in weekends), we have split up the statistics, histograms and time series plots in three parts: whole weeks, only weekdays and only weekends. We did the latter two parts by creating a mask for the dataset such that only the days monday untill friday get screened or in the case of the weekends: only saturday and sunday.

3.1.3 Results Question 4

The sample size for the whole week is 1005, for the weekdays 719 and for the weekends it is 286.

Table 2: Descriptive statistics for the store in the mall.

	whole weeks	only weekdays	only weekends
Sum	95572.1	41183.6	54388.5
Sample mean	95.1	57.3	190.2
Sample variance	4173.8	373.1	1094.4
Minimum	20.8	20.8	118
median	64	54.3	199.4
maximum	248.8	173.7	248.8
Inter Quantile Range	99.3	23.9	59.7
Range	228	152.9	130.8
Skewness	1.00	1.39	-0.31
Kurtosis	-0.54	3.96	-1.30

Table 3: Descriptive statistics for the store on streetA.

	whole weeks	only weekdays	only weekends
Sum	78488	34150	44338
Sample mean	78.1	47.5	155
Sample variance	2948	205	1572
Minimum	17.9	17.9	72.9
median	53.8	45.1	53.8
maximum	228.6	105.9	228.6
Inter Quantile Range	65.9	19.8	70.3
Range	210.7	88	155.7
Skewness	1.23	0.64	-0.22
Kurtosis	0.13	0.29	-1.34

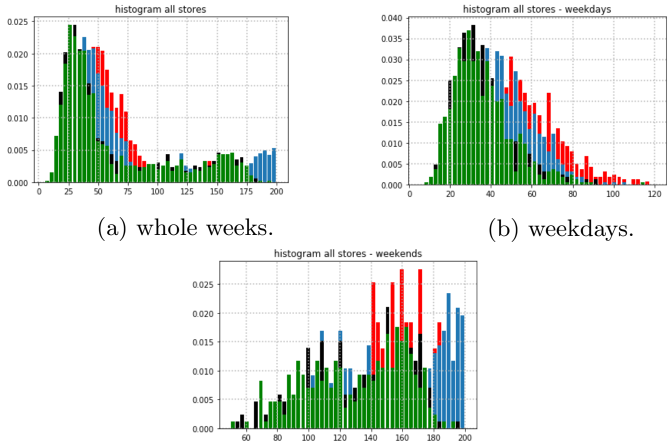
Table 4: Descriptive statistics for the store on streetB.

	whole weeks	only weekdays	only weekends
Sum	63819	26021	37798
Sample mean	63.5	36.2	132.2
Sample variance	2332	243.2	995
Minimum	11.2	11.2	55.4
median	40.5	33.2	138.1
maximum	194.2	134.2	194.2
Inter Quantile Range	64.4	16.8	50.8
Range	182.9	123	138.7
Skewness	1.09	1.84	-0.41
Kurtosis	-0.25	6.34	-0.86

Table 5: Descriptive statistics for the store on streetC.

	whole weeks	only weekdays	only weekends
Sum	62919	25599	37320
Sample mean	62.6	35.6	130.5
Sample variance	2260	205.5	985.5
Minimum	8.9	8.9	49
median	40.2	33	136.2
maximum	192.3	113.1	192.3
Inter Quantile Range	62	17.7	52.6
Range	183.4	104.2	143.3
Skewness	1.11	1.13	-0.34
Kurtosis	-0.17	1.92	-0.93

The histograms and plots for the store in the mall, on streetA, on street B and on streetC are red, blue, black and green respectively.



(a) whole weeks.

(b) weekdays.

(c) weekends.

Figure 6: Histograms for all stores during whole weeks, weekdays and weekends.

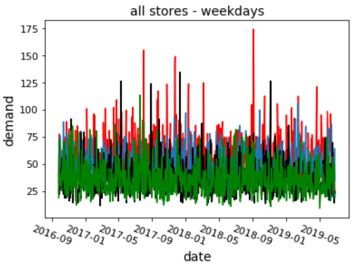


Figure 7: time series plot for all stores during weekdays.

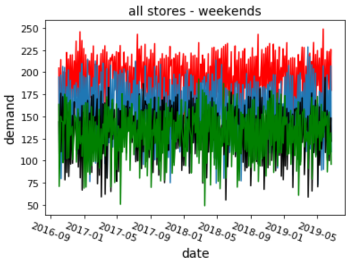


Figure 8: time series plot for all stores during weekends.

3.1.4 Conclusion Question 4

When you look at the data as whole weeks, it becomes directly visible that the demand starts high at the mall and decreases as the distance from the mall increases. So the store in the mall has the highest demand, then streetA, whereafter streetB and then lastly the store on streetC. Accordingly, the sample mean, sample variance, minimums, medians and maximums all have the same behaviour as the demand.

But when you look at the data as weekdays vs weekends it becomes quite apparent that even though the weekend has 3 days per week less in it, it still has a higher overall demand. So the mean during weekends is way higher, which brings up the overall mean by a lot. This can be seen very clearly in the time series plots. Also, the maximums are always attained in weekends whereas the minimums are always attained during the weekdays.

For the weekends all stores have negative skewness and kurtosis. Hence that data is left winged and platykurtic. Whereas during weekdays they are both always positive, however for mall and streetB the kurtosis is so high that data becomes leptokurtic. StreetA and StreetC still have platykurtic peaks. For all stores during the weekdays the positive skewness means the data is right winged.

Only the range of data for the store in the mall is higher during the weekdays. The stores on the nearby streets all have higher ranges of demand during the weekends. The range measures variability. During the weekends the IQR is way higher compared to the weekdays, this is due to the mean being higher. This implies that the variance for the weekends is also higher, this can be visualized when you compare figure 7 and 8.

The most visible trend in the histograms is the clustering of data. During the weekdays you can clearly see a log-normal pattern of the data, where the clusters are close to the mean. During the weekends you can see 2 clusters of data. We presume these clusters come close to the means of Saturday and Sunday.

3.2 Question 5 and 6

3.2.1 Problem Description Question 5

In this second empirical question, we used information in Table 1 and the parametric frequentist framework to estimate the optimal quantities for Friday, Saturday and Sunday. Were we take the constants $\tilde{c}$ and $\tilde{p}$ from Question 1 into consideration.

3.2.2 Methodology Question 5

The first store we chose was the mall because it has the most demand out of all the four stores. We also chose the store on streetC because it has the most profit per product and also it is the farthest from the mall. We specified the lognormal distrubutions because of the way the histograms take form. Also the differing skewness indicates that the data is not normally distributed (table 2 and 5). Then, we reported the point estimates and intervals of theoretically optimal order quantity. We used the direct approach of constructing the confidence interval. We did this by selecting the data from Friday, Saturday and Sunday per store, referred to as storedataperday. The τ in this question is not only decided by the costs and prices but also the disposal cost (c_h) from table 1.

This is why the τ is calculated as follows:

$$\tau = \frac{p - c - c_h}{p + c_h}. \tag{11}$$

Furthermore, the confidence interval for optimal order is calculated as follows:

$$(QL_{store}, QR_{store}) \tag{12}$$

where

$QL_{store} = \widehat{Q} - z * se$
$QR_{store} = \widehat{Q} + z * se$
$\widehat{Q} = lognorm.ppf(\tau, \widehat{\sigma}, \widehat{\mu})$ = estimated optimal order
$\widehat{\mu} = mean(\log(storedataperday))$
$\widehat{\sigma}^2 = var(\log(storedataperday))$
$z = norm.ppf(1 - 0.5 * \alpha)$ with $\alpha = 0.05$
$se = (assvar/n)$ with n = length of the data of that day
$assvar = \widehat{Q}^2 * \widehat{\sigma}^2 + 0.5 * (\log(\widehat{Q} - \widehat{\mu})^2$

Finally, the discussions have been related with our conclusions in Question 3.

3.2.3 Results Question 5

Table 6: optimal order and confidence interval for the store in the mall.

	friday	saturday	sunday
estimated optimal order	33.38	206.26	134.66
direct confidence interval	(30.80 , 35.95)	(204.18 , 208.35)	(130.96 , 138.37)

Table 7: optimal order and confidence interval for the store on streetC.

	friday	saturday	sunday
estimated optimal order	28.79	150.26	92.66
direct confidence interval	(26.87 , 30.70)	(148.08 , 152.44)	(89.31 , 96.00)

$\tau_{mall} = 0.075 \quad \tau_{streetC} = 0.32$

3.2.4 Conclusion Question 5

Parametric tests are suitable if a decently fitting distribution is specified, especially if the τ is close to 0 or 1. We are sure the data during the weekdays is log-normally distributed. For the data during the weekends we assume a mixture of normal and log-normal, but since the weekend is split up into Saturday and Sunday we assume a log-normal distribution for these days. Also if we assume log-normal when it should be normal it will still have good results (result Q2&3). The τ is close to 0 for the mall, for streetC it is not that close. So it can be concluded that the optimal order for friday for the mall is very well estimated, and it is decently estimated for the friday of streetC. The optimal order for Saturday and Sunday might be nicely estimated but perhaps they can be estimated better using NP, if not obvious: the results for the weekend of the mall will be better since it's τ is close to 0.

3.2.5 Problem Description Question 6

This question is almost identical to Question 5. The only difference is that we used the information in the nonparametric (distribution-free) framework.

3.2.6 Methodology Question 6

We chose the same two stores as in question 5 for the comparing aspect. We used the order statistics approach of constructing confidence interval. To find (12), where:

$QL_{store} = storedataperday[L]$
$QR_{store} = storedataperday[U]$
$\widehat{Q} = storedataperday[idx] =$ estimated optimal order
$idx = int(np.ceil(\tau * n) - 1)$ where τ is calculated by eq. (11)
$z = sts.norm.ppf(1 - 0.5 * \alpha)$ with $\alpha = 0.05$
$sq = np.sqrt(n * \tau * (1 - \tau))$
$U = int(np.ceil(n * \tau + z * sq))$
$L = int(np.ceil(n * \tau - z * sq))$
n = length of the data of that day

Then, we compared these results with our outcome in Question 5. Finally, the discussions have been related with our conclusions in Question 2 as well as Question 4.

3.2.7 Results Question 6

Table 8: optimal order and confidence interval for the store in the mall.

	friday	saturday	sunday
estimated optimal order	35.60	205.47	131.52
direct confidence interval	(32.15 , 36.83)	(203.54 , 207.48)	(128.64 , 140.16)

Table 9: optimal order and confidence interval for the store on streetC.

	friday	saturday	sunday
estimated optimal order	29.43	151.01	94.62
direct confidence interval	(28.08 , 32.98)	(148.68 , 154.93)	(91.02 , 98.80)

3.2.8 Conclusion Question 6

The results do not differ much from those in question 5, a variety of 0 to 4 per day. In question 4 we presumed that the clusters in the histogram of the weekends come close to the means of Saturday and Sunday. This can be verified when you look at the optimal orders and the peaks of the histograms for the concerning stores. Especially when you take into account the (disposal) costs that get applied, since the means are a bit higher.
Since the τ for streetC is the same as in Q5 (thus not close to 0 or 1) and we are not 100% sure about the weekend being log-normal, we expect the optimal order to be better estimated using NP. So the Saturday and Sunday for streetC should have the truest result here.

4 Final Word - Question 7

4.1 Conclusion

There is an effect of adding a holding cost c_h to the profit function. c_h decreases the optimal order quantity when included. This intuitively makes sense, as the only difference between the original profit function and equation (1) is a decrease when the order quantity is greater than the demand. The parametric approach is preferable, if one can make a reliable guess about the underlying distribution. This is especially true when τ is close to 0 or 1. Should the underlying distribution be hard to determine or if no MLE can be calculated, it is probably favorable to use the NP approach.
In conlusion, the best estimated values are:

Table 10: optimal order and confidence interval for the store in the mall.

	friday	saturday	sunday
estimated optimal order	33.38	206.26	134.66
direct confidence interval	(30.80 , 35.95)	(204.18 , 208.35)	(130.96 , 138.37)

Table 11: optimal order and confidence interval for the store on streetC.

	friday	saturday	sunday
estimated optimal order	28.79	151.01	94.62
direct confidence interval	(26.87 , 30.70)	(148.68 , 154.93)	(91.02 , 98.80)

Since; 1. the friday is definitely log-normally distributed for both stores. ; 2. we assumed the weekends to be log-normal and with the τ of the mall being near 0, the saturday and sunday of the mall are be better estimated using the parametric framework. ; 3. for the weekend of streetC the τ was not near 0 or 1 and thus an estimation using NP deems to be truest.

4.2 Discussion

A problem with the overall data is that the data is not specified enough, this leads to the estimations not having enough separation. In the sense that there are 1005 days where no seasons, weather conditions, quality of product etc. are specified. So whenever the store would want to use our estimations on a rainy day for example, they might have to dispose of a lot of products. Also there are a lot of outliers in the data which we have not learned about nor how to remove them, which would have made the results more correct. A problem with question 5 is that whenever you select the data from individual days, you get a relatively small sample size. This influences the correctness of using the parametric framework negatively. In the sense that the parameter in question will have a less accurate estimation, then compared to when there is a very large sample size. We assumed from the results of question 4 that the data is log-normally distributed. We were not sure whether it was normal or log-normal in the weekend, but to assume log-normal was the better option. It was the better option because if we assume log-normal when it should be normal it will still have good results, the other way around that is not the case. This was necessary to estimate the optimal orders and confidence intervals per chosen store in question 5.