DATA 605: Computational Mathematics

Gabriel Campos

Last edited December 06, 2024

Contents

Library

```
library(dplyr)
library(e1071)
library(ggplot2)
library(MASS)
library(readr)
```

Final Examination: Business Analytics and Data Science

Instructions:

You are required to complete this take-home final examination by the end of the last week of class. Your solutions should be uploaded in **pdf** format as a knitted document (with graphs, content, commentary, etc. in the pdf). This project will showcase your ability to apply the concepts learned throughout the course.

The dataset you will use for this examination is provided as retail data.csv, which contains the following variables:

- Product_ID: Unique identifier for each product.
- Sales: Simulated sales numbers (in dollars).
- Inventory_Levels: Inventory levels for each product.
- Lead_Time_Days: The lead time in days for each product.
- Price: The price of each product.
- Seasonality_Index: An index representing seasonality.

Problem 1:

Business Risk and Revenue Modeling

Context: You are a data scientist working for a retail chain that models sales, inventory levels, and the impact of pricing and seasonality on revenue. Your task is to analyze various distributions that can describe sales variability and forecast potential revenue.

Data Load

```
retail_df <- read_csv("synthetic_retail_data.csv")
```

Part 1:

Empirical and Theoretical Analysis of Distributions (5 Points)

Task:

1.

Generate and Analyze Distributions:

- **X ~ Sales:** Consider the Sales variable from the dataset. Assume it follows a Gamma distribution and estimate its shape and scale parameters using the fitdistr function from the MASS package.
- **Y ~ Inventory Levels:** Assume that the sum of inventory levels across similar products follows a Lognormal distribution. Estimate the parameters for this distribution.
- **Z ~ Lead Time:** Assume that Lead_Time_Days follows a Normal distribution. Estimate the mean and standard deviation. Calculate Empirical Expected Value and Variance:

Calculate the empirical mean and variance for all three variables. Compare these empirical values with the theoretical values derived from the estimated distribution parameters.

```
head(retail_df)
```

```
## # A tibble: 6 x 6
##   Product_ID Sales Inventory_Levels Lead_Time_Days Price Seasonality_Index
##        <dbl> <dbl>            <dbl>          <dbl> <dbl>             <dbl>
## 1          1  158.             367.           6.31  18.8             1.18
## 2          2  279.             427.           5.80  26.1             0.857
## 3          3  699.             408.           3.07  22.4             0.699
## 4          4 1832.             392.           3.53  27.1             0.698
## 5          5  460.             448.          10.8   18.3             0.841
## 6          6 1693.             547.          10.1   23.5             1.13
```

```
glimpse(retail_df)
```

```
## Rows: 200
## Columns: 6
## $ Product_ID        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ Sales             <dbl> 158.43952, 278.99020, 698.85868, 1832.39467, 459.703~
## $ Inventory_Levels  <dbl> 367.4421, 426.6512, 407.6394, 392.3912, 448.3120, 54~
## $ Lead_Time_Days    <dbl> 6.314587, 5.800673, 3.071936, 3.534253, 10.802241, 1~
## $ Price             <dbl> 18.795197, 26.089636, 22.399985, 27.092013, 18.30782~
## $ Seasonality_Index <dbl> 1.1839497, 0.8573051, 0.6986774, 0.6975404, 0.840725~
```

```
summary(retail_df)
```

```
##    Product_ID         Sales         Inventory_Levels Lead_Time_Days
## Min.   :  1.00   Min.   :  25.57   Min.   : 67.35   Min.   : 0.491
## 1st Qu.: 50.75   1st Qu.: 284.42   1st Qu.:376.51   1st Qu.: 5.291
## Median :100.50   Median : 533.54   Median :483.72   Median : 6.765
## Mean   :100.50   Mean   : 636.92   Mean   :488.55   Mean   : 6.834
## 3rd Qu.:150.25   3rd Qu.: 867.58   3rd Qu.:600.42   3rd Qu.: 8.212
## Max.   :200.00   Max.   :2447.49   Max.   :858.79   Max.   :12.722
##     Price       Seasonality_Index
## Min.   : 5.053   Min.   :0.3305
## 1st Qu.:16.554   1st Qu.:0.8475
## Median :19.977   Median :0.9762
## Mean   :19.560   Mean   :0.9829
## 3rd Qu.:22.924   3rd Qu.:1.1205
## Max.   :29.404   Max.   :1.5958
```

```
# Isolate Sales data
sales_retail_df <- retail_df$Sales
summary(sales_retail_df)
```

X ~ Sales

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.57  284.42  533.54  636.92  867.58 2447.49
```

```
sum(sales_retail_df<0)
```

```
## [1] 0
```
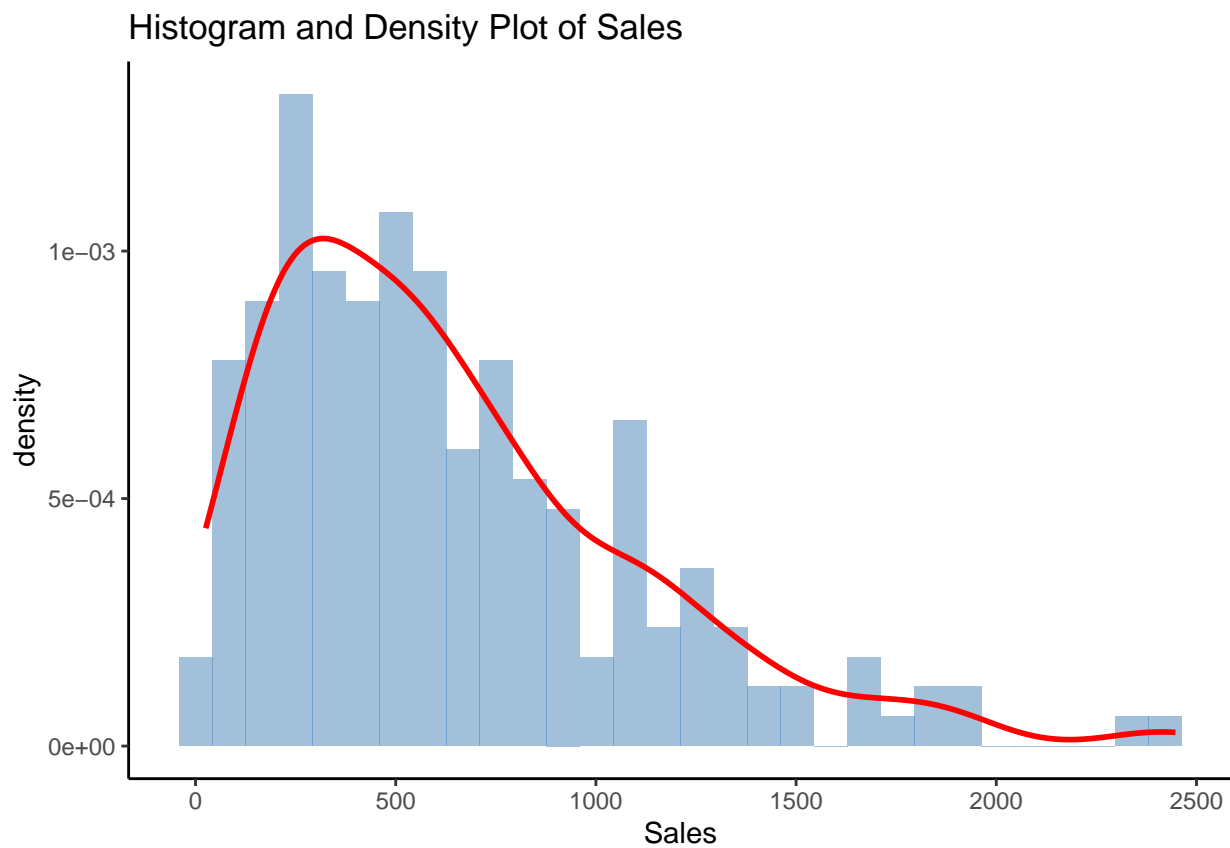
```
sum(is.na(sales_retail_df))
```

```
## [1] 0
```

```
shapiro.test(sales_retail_df)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sales_retail_df
## W = 0.90377, p-value = 4.397e-10
```

```
ggplot(retail_df, aes(x = Sales)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30, fill = "steelblue",
                 alpha = 0.5) +
  geom_density(color = "red", size = 1) +
  ggtitle("Histogram and Density Plot of Sales") +
  theme_classic()
```
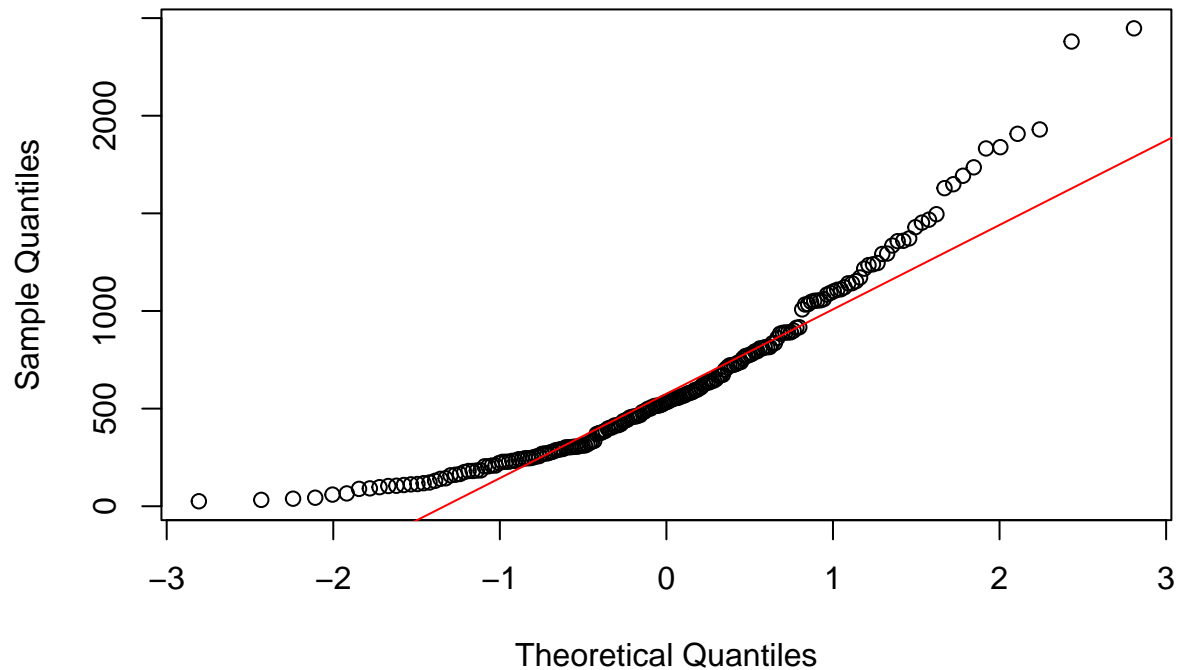
Histogram and Density Plot of Sales


```
qqnorm(sales_retail_df,
       main = "Q-Q Plot of Sales")
qqline(sales_retail_df,
       col = "red")
```
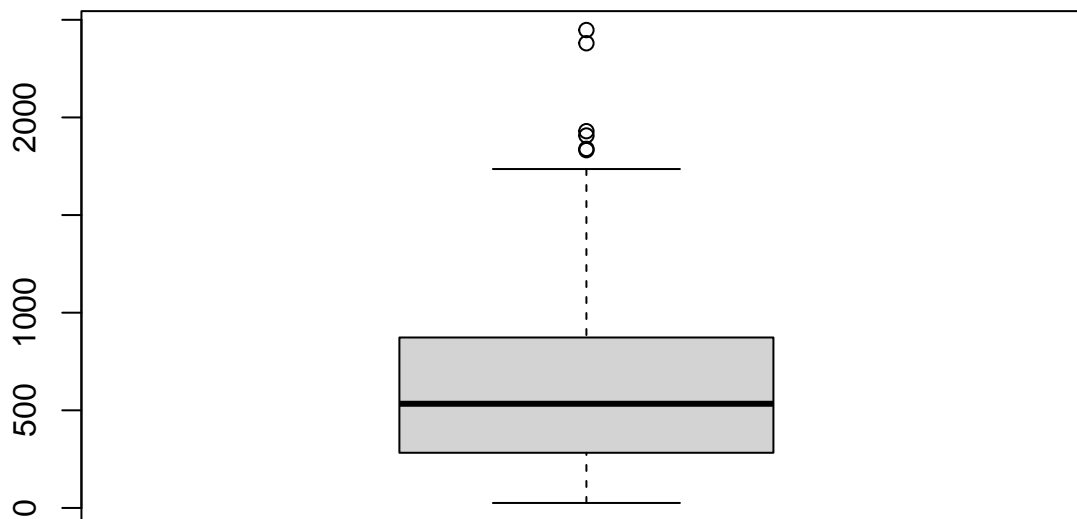
Q–Q Plot of Sales


```r
boxplot(sales_retail_df,
        main = "Boxplot of Sales Data")
```

Boxplot of Sales Data

Initial analysis

- For our **Sales** data our $Mean > Median$ ($636.92 > 533.54$) which indicates that our data is right skewed and not normalized. This is supported by our Histogram, our Q-Q plot and the Shapiro test's $p-value$ of less than 0.05.
- No NAs are noted with the **Sales** data
- Our range for the values within **Sales** is 25.57 to 2447.49, encompassing a wide range.

- Our Box plot indicates that there are outliers, primarily for values > 1000

fitdstr Sales Solution Assume $X \sim Gamma(\alpha, \beta)$ the parameter *"gamma"* will be used with *fitdistr()*.

```
sales_gamma_fit <- fitdistr(sales_retail_df, "gamma")
```

```
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced
```

```
print(sales_gamma_fit)
```

```
##        shape          rate
##    1.8349640762    0.0028810166
##   (0.1511756159) (0.0002556985)
```

Considering no NAs or negative values were noted in our original data set, the *NaNs produced* warning, is likely a result of the right-skewed data or from our outliers. I will remove the outliers to see if it removes the warning. Regardless, dealing with these outliers should improve precision.

The below steps should remove values above our 99% quantile or below the 1%

```
# compute quantiles at 1% and and 99%
sales_retail_quantiles <-
  quantile(sales_retail_df, probs =c(0.01,0.99))
# remove outliers below the 1% and above 99%
sales_retail_df_clean<- sales_retail_df[
  sales_retail_df >= sales_retail_quantiles[1] &
    sales_retail_df <= sales_retail_quantiles[2]
]
sales_gamma_fit_clean <- fitdistr(sales_retail_df_clean, "gamma")
```

```
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced
```

```
print(sales_gamma_fit_clean)
```

```
##        shape          rate
##    2.0323543224    0.0032518379
##   (0.1724396139) (0.0002975715)
```

The cleaned model still creates an error therefore I would like to see visually how well the values fit.
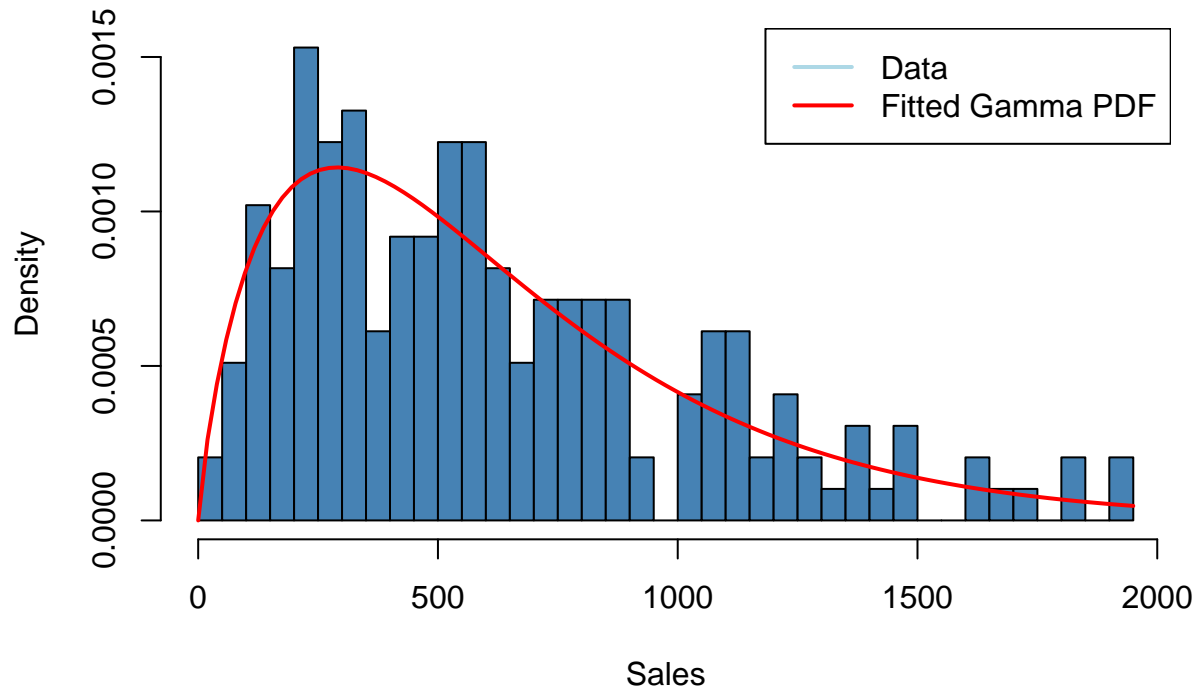
```
hist(sales_retail_df_clean,
     breaks = 30,
     probability = TRUE,
     main = "Fitted Gamma Distribution",
     xlab = "Sales",
     col = "steelblue")

curve(dgamma(x,
             shape = 1.8349640762,
             rate = 0.0028810166),
      col = "red",
      lwd = 2,
      add = TRUE)

legend("topright",
       legend = c("Data",
                  "Fitted Gamma PDF"),
```

```
col = c("lightblue",
        "red"), lwd = 2)
```

Fitted Gamma Distribution
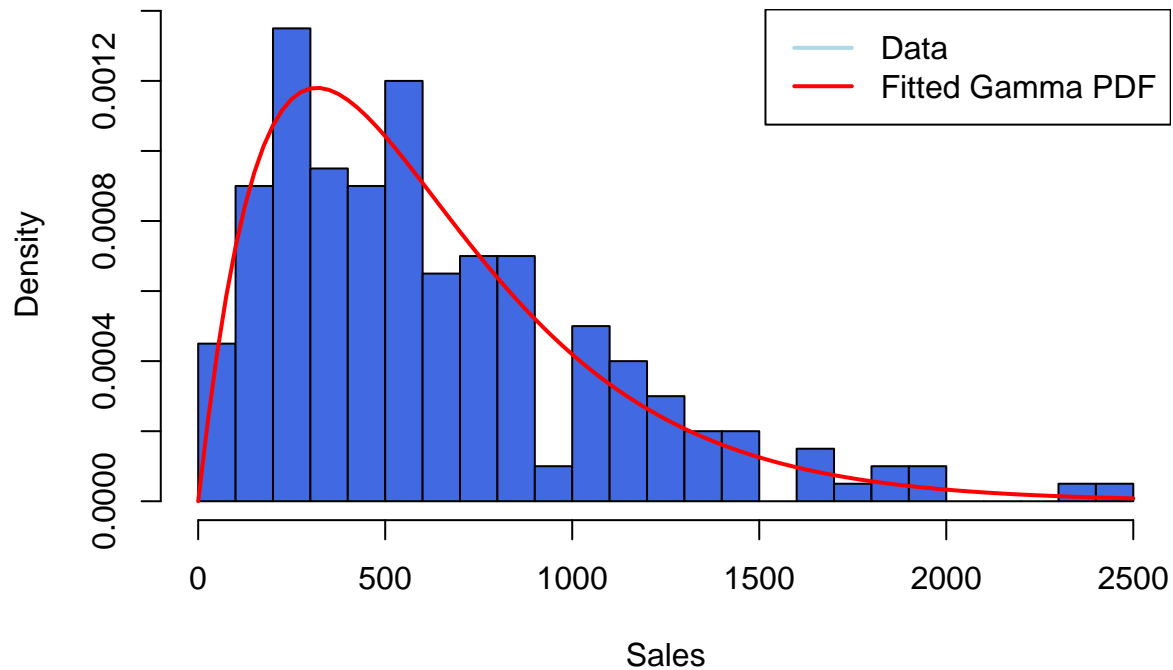


```r
hist(sales_retail_df,
     breaks = 30,
     probability = TRUE,
     main = "Fitted Gamma Distribution",
     xlab = "Sales",
     col = "royalblue")

curve(dgamma(x,
             shape = 2.0323543224,
             rate = 0.0032518379 ),
      col = "red",
      lwd = 2,
      add = TRUE)

legend("topright",
       legend = c("Data",
                  "Fitted Gamma PDF"),
       col = c("lightblue",
               "red"), lwd = 2)
```

Fitted Gamma Distribution

Visually, both values seem to match the *Sales* behavior. By calculating the absolute difference in mean, skewness and variance, I might get a better indication on which gamma distribution and gamma parameters, better emulates the datas behavior.

```r
cmp_metrics <- c("Mean", "Variance", "Skewness")

cmp_data_values <- c(617.595, 165599.6, 0.8891198)

# Original Gamma differences
cmp_shape1 <- 2.0323543224

cmp_rate1 <- 0.0032518379

cmp_original_gamma_values <-
  c(cmp_shape1 / cmp_rate1,
    cmp_shape1 / (cmp_rate1^2),
    2 / sqrt(cmp_shape1))

cmp_original_differences <-
  abs(cmp_original_gamma_values - cmp_data_values)

# New Gamma differences
cmp_shape2 <- 1.8349640762
cmp_rate2 <- 0.0028810166

cmp_new_gamma_values <-
  c(cmp_shape2 / cmp_rate2,
    cmp_shape2 / (cmp_rate2^2),
    2 / sqrt(cmp_shape2))
cmp_new_differences <-
```

```r
    abs(cmp_new_gamma_values - cmp_data_values)

# Prepare data for ggplot
plot_data <- data.frame(
  Metric = rep(cmp_metrics,
               times = 2),
  Difference =
    c(cmp_original_differences,
              cmp_new_differences),
  Gamma = rep(c("Original Gamma",
               "New Gamma"),
             each = length(cmp_metrics))
)

# Create the clustered bar plot
ggplot(plot_data, aes(x = Metric, y = Difference, fill = Gamma)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9)) +
  geom_text(aes(label = round(Difference, 2)),
           position = position_dodge(width = 0.9),
           vjust = -0.5, size = 3.5) +
  labs(
    title = "Comparison of Differences from Data Metrics",
    x = "Metric",
    y = "Absolute Difference"
  ) +
  scale_fill_manual(values = c("Original Gamma" = "steelblue", "New Gamma" = "royalblue")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.title = element_blank()
  )
```

Comparison of Differences from Data Metrics
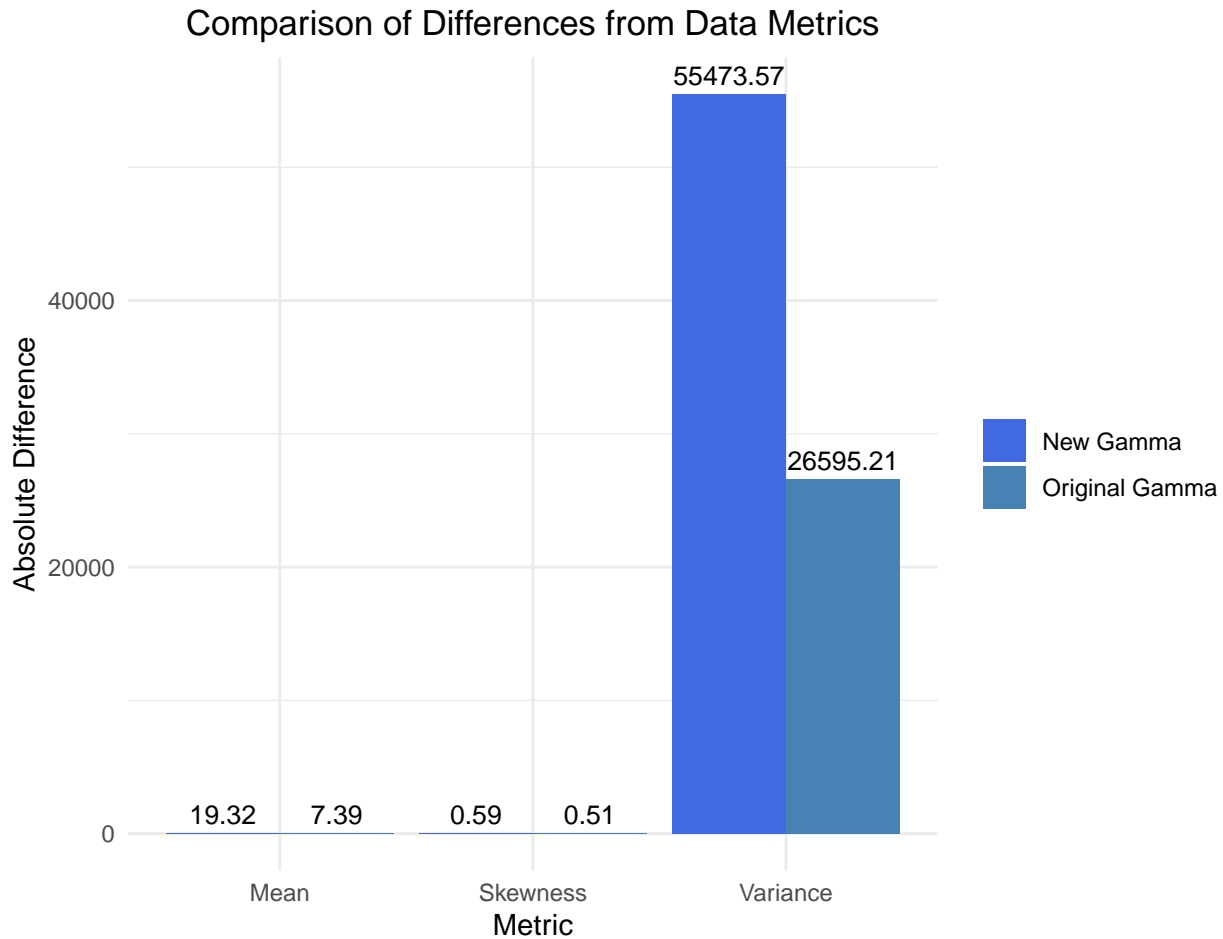

```
rm(list = ls(pattern = "^cmp_|plot_data|_clean$|_quantiles$"))
```

Mean and Variance Solution Our Empirical mean and variance is just

```
sales_emp_mean<-mean(sales_retail_df)
sales_emp_var<-var(sales_retail_df)
```

and our theoretical mean and variance is calculate as

$\mu_{gamma} = \frac{\alpha}{\beta}$

$\sigma^2_{gamma} = \frac{\alpha}{\beta^2}$ or $\frac{shape}{rate^2}$

```
sales_gamma_shape <- sales_gamma_fit$estimate["shape"]
sales_gamma_rate <- sales_gamma_fit$estimate["rate"]

sales_theo_gamma_mean <- sales_gamma_shape / sales_gamma_rate
sales_theo_gamma_var <- sales_gamma_shape/ (sales_gamma_rate^2)
```

```
comparison <- data.frame(
  Metric = c("Mean", "Variance"),
  Empirical = c(sales_emp_mean, sales_emp_var),
  Theoretical = c(sales_theo_gamma_mean, sales_theo_gamma_var)
)
```

and our rate

Answer the shape=1.8349640762 and rate=0.0028810166 are best for this gamma distribution.

```
print(comparison)
```

```
##      Metric   Empirical Theoretical
## 1     Mean     636.9162    636.9155
## 2 Variance 214831.7509 221073.1799
```

```
retail_inv_df <- retail_df$Inventory_Levels
summary(retail_inv_df)
```

Y ~ Inventory Levels

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    67.35  376.51  483.72  488.55  600.42  858.79
```

```
sum(retail_inv_df<0)
```

```
## [1] 0
```

```
sum(is.na(retail_inv_df))
```

```
## [1] 0
```

```
shapiro.test(retail_inv_df)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  retail_inv_df
## W = 0.99303, p-value = 0.4646
```

```
ggplot(retail_df, aes(x = Inventory_Levels )) +
  geom_histogram(aes(y = ..density..),
                 bins = 30, fill = "steelblue",
                 alpha = 0.5) +
  geom_density(color = "red", size = 1) +
  ggtitle("Histogram and Density Plot of Inventory Levels") +
  theme_classic()
```
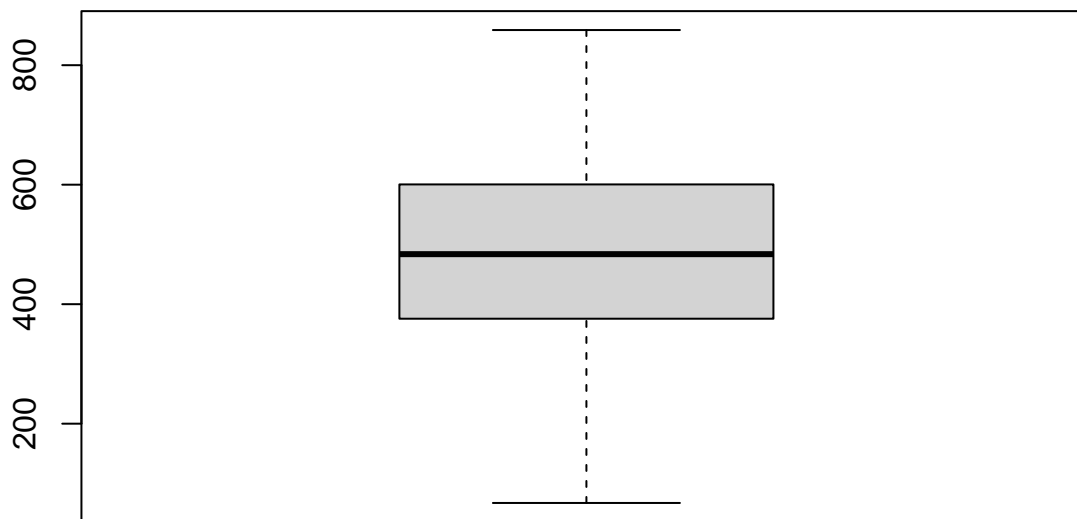
Histogram and Density Plot of Inventory Levels
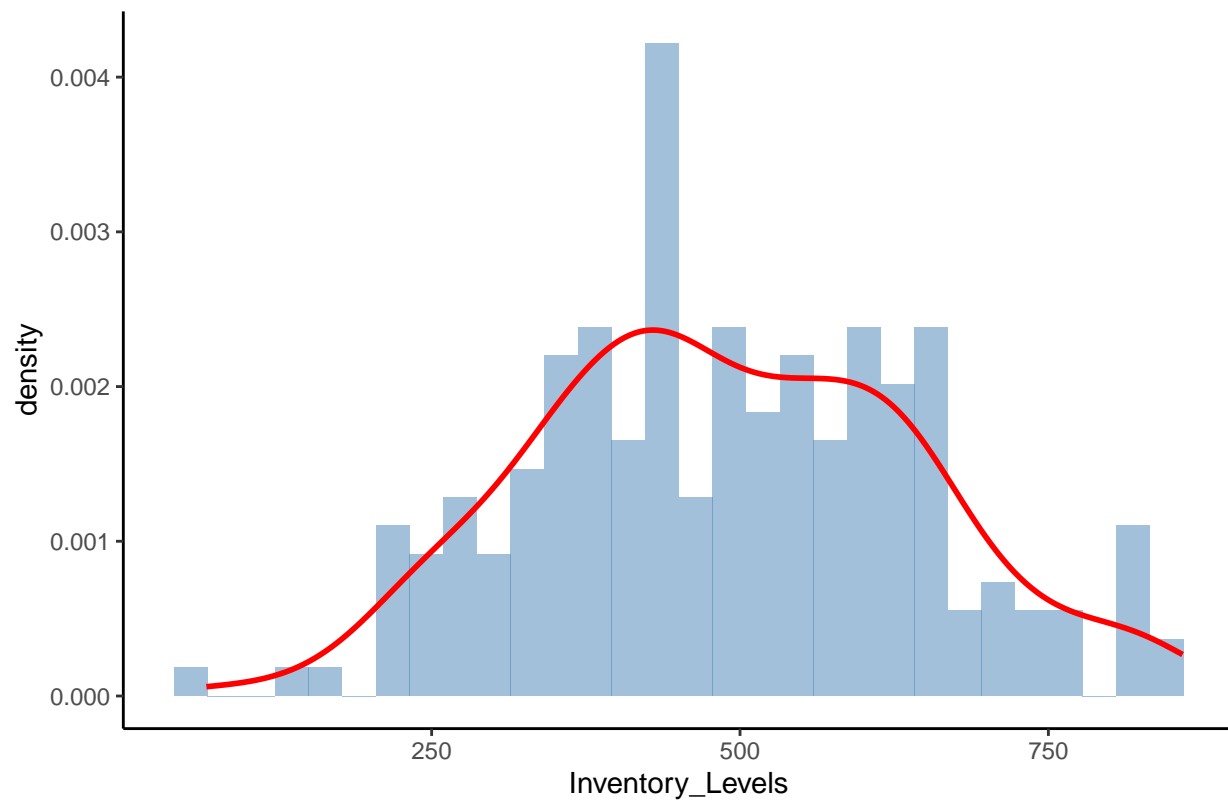
```r
qqnorm(retail_inv_df,
       main = "Q-Q Plot of Sales")
qqline(retail_inv_df,
       col = "red")
```

Q–Q Plot of Sales

(Sample Quantiles vs Theoretical Quantiles)
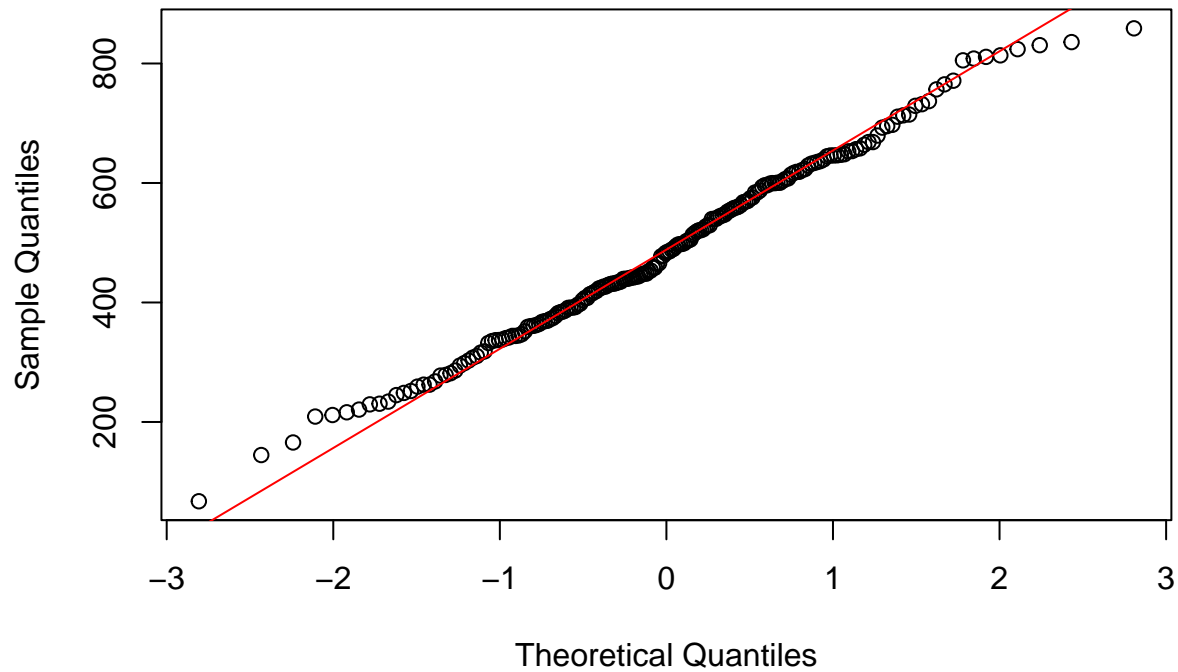
```
boxplot(retail_inv_df,
        main = "Boxplot of Sales Data")
```

Boxplot of Sales Data

Initial analysis

- No NAs found in *Inventory_Levels*
- No values below 0 for the *Inventory_Levels* values.
- 488.55 (*Mean*) > 483.72 (*Median*) suggests the data may be right skewed.
- Shapiro test had a $p-value = 0.4646$ which although below 0.5, suggesting it is not normalized, it is relatively close to being normal.

- Q-Q plot and, Histogram and Density plot, show the data as near normal.

fitdistr Inventory Levels   Since we are assuming the Inventory Levels across products follow a Lognormal distribution, and *fitdistr()* does not directly support *lognormal* distributions, we will have to first transform the data.

The parameters of this distribution will ultimately be $\mu_l og$ and $\sigma_l og$

```
inv_log <-
  fitdistr(log(retail_inv_df),"normal")
```

2.

Part 2:

Probability Analysis and Independence Testing (5 Points)

Task:

1.

Empirical Probabilities: For the Lead_Time_Days variable (assumed to be normally distributed), calculate the following empirical probabilities:

- $P(Z > \mu | Z > \mu - \sigma)$
- $P(Z > \mu + \sigma > Z > \mu)$
- $P(Z > \mu + 2\sigma > Z > \mu)$

2.

Correlation and Independence:

- Investigate the correlation between Sales and Price. Create a contingency table using quartiles of Sales and Price, and then evaluate the marginal and joint probabilities.
- Use Fisher's Exact Test and the Chi-Square Test to check for independence between Sales and Price. Discuss which test is most appropriate and why.

Problem 2

Advanced Forecasting and Optimization (Calculus) in Retail

Context: You are working for a large retail chain that wants to optimize pricing, inventory management, and sales forecasting using data-driven strategies. Your task is to use regression, statistical modeling, and calculus-based methods to make informed decisions.

Part 1

Descriptive and Inferential Statistics for Inventory Data (5 Points)

Task:

1.

Inventory Data Analysis:

- Generate univariate descriptive statistics for the Inventory_Levels and Sales variables.

- Create appropriate visualizations such as histograms and scatterplots for Inventory_Levels, Sales, and Price.
- Compute a correlation matrix for Sales, Price, and Inventory_Levels.
- Test the hypotheses that the correlations between the variables are zero and provide a 95% confidence interval.

2.

Discussion:

- Explain the meaning of your findings and discuss the implications of the correlations for inventory management. Would you be concerned about multicollinearity in a potential regression model? Why or why not?

Part 2

Linear Algebra and Pricing Strategy (5 Points)

Task:

1.

Price Elasticity of Demand:

- Use linear regression to model the relationship between Sales and Price (assuming Sales as the dependent variable).
- Invert the correlation matrix from your model, and calculate the precision matrix.
- Discuss the implications of the diagonal elements of the precision matrix (which are variance inflation factors).
- Perform LU decomposition on the correlation matrix and interpret the results in the context of price elasticity.

Part 3:

Calculus-Based Probability & Statistics for Sales Forecasting (5 Points)

Task:

1.

Sales Forecasting Using Exponential Distribution:

- Identify a variable in the dataset that is skewed to the right (e.g., Sales or Price) and fit an exponential distribution to this data using the fitdistr function.
- Generate 1,000 samples from the fitted exponential distribution and compare a histogram of these samples with the original data's histogram.
- Calculate the 5th and 95th percentiles using the cumulative distribution function (CDF) of the exponential distribution.
- Compute a 95% confidence interval for the original data assuming normality and compare it with the empirical percentiles.

2.

Discussion:

- Discuss how well the exponential distribution models the data and what this implies for forecasting future sales or pricing. Consider whether a different distribution might be more appropriate.

Part 4

Regression Modeling for Inventory Optimization (10 Points)

Task:

1.

Multiple Regression Model:

- Build a multiple regression model to predict Inventory_Levels based on Sales, Lead_Time_Days, and Price.
- Provide a full summary of your model, including coefficients, R-squared value, and residual analysis.

2.

Optimization:

- Use your model to optimize inventory levels for a peak sales season, balancing minimizing stockouts with minimizing overstock.

References

Statology fitdistr-r

rdocumentation MASS package fitdistr

Statology - fit gamma distribution to dataset in r

Wiki Gamma Distribution

rdocumentation qualtile