Homework Assignment 11: Applied Probabilistic Models

Multivariate distributions and probability densities

5273

1 Introduction

For this work aspects of multivariate probability densities and distribution are discussed. To begin with, a convolution of distribution example is performed, in the area of finance, specifically in stock prices of two companies. As a second point, a Chi-Squared test is performed for two categorical variables in the analysis of the results of my thesis work [4]. Finally, a numerical covariance analysis is performed, in order to prove two properties of this statistical value, also with prices on the stock market between two other enterprises.

For the analysis, the R software is used in its version 4.0.2 [6], and the code used is available on the GitHub repository [3]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Convolution of Distributions

This section is about the convolution of probability distributions. The convolution can be considered as the operation of forming linear combinations of random variables. Then, the probability density function (PDF) of a sum of random variables is the convolution of their corresponding PDF. This sum, let be $Z = X + Y$, for example, of two random variables, can be expressed as mentioned in Equation 1 for discrete variables and Equation 2 for continuous ones.

$$P(Z = z) = \sum_{k=-\infty}^{\infty} P(X = k)P(Y = z - k) \tag{1}$$

$$h(z) = (f * g)(z) = \int_{-\infty}^{\infty} f(z - t)g(t)dt = \int_{-\infty}^{\infty} f(t)g(z - t)dt \tag{2}$$

For the convolution example data are downloaded from Yahoo! Finance [1], corresponding to the values of the stock prices of Amazon and Tesla Motors enterprises, which correspond to the timeframe of September 2019 - September 2020. The data used for the analysis is the adjusted close price of the stock, which is the closing price after adjustments for all applicable splits and dividend distributions. It includes trading days only, which means that Saturdays, Sundays, and national holidays are not quoted as the stock market is not open on those days. Figure 1 and Figure 2 show a boxplot and density plot respectively created to represent differences of the data. Then, Figure 3 shows a histogram of the convolution of the distributions.

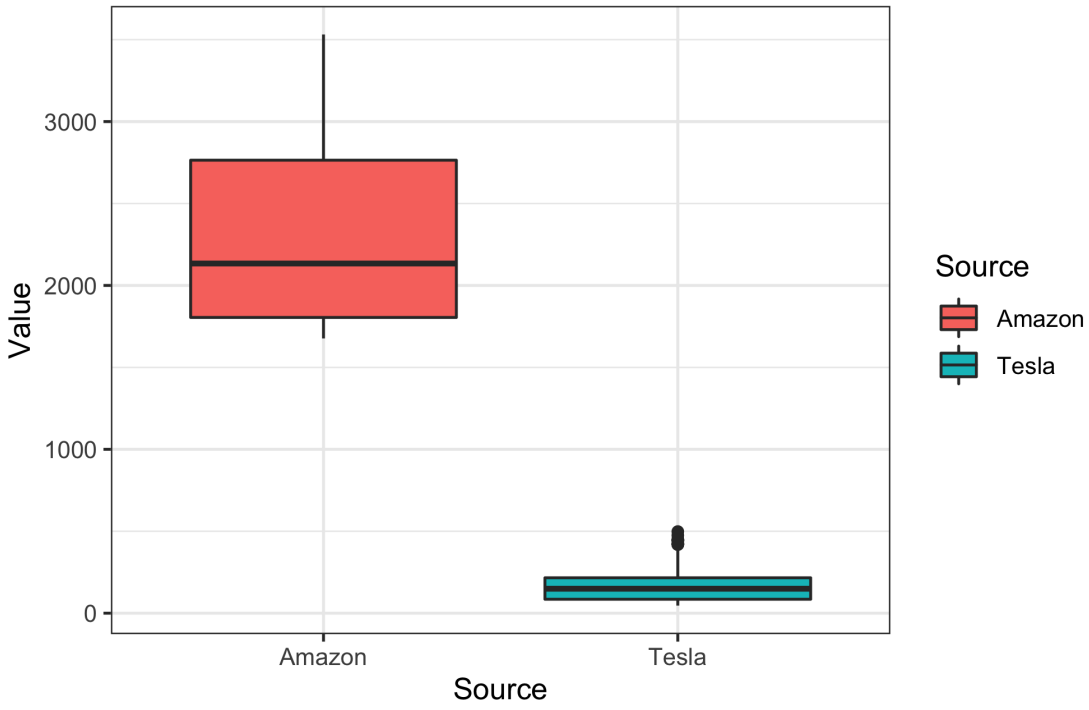


Figure 1: Boxplots of the Adjusted Close Price of Amazon and Tesla

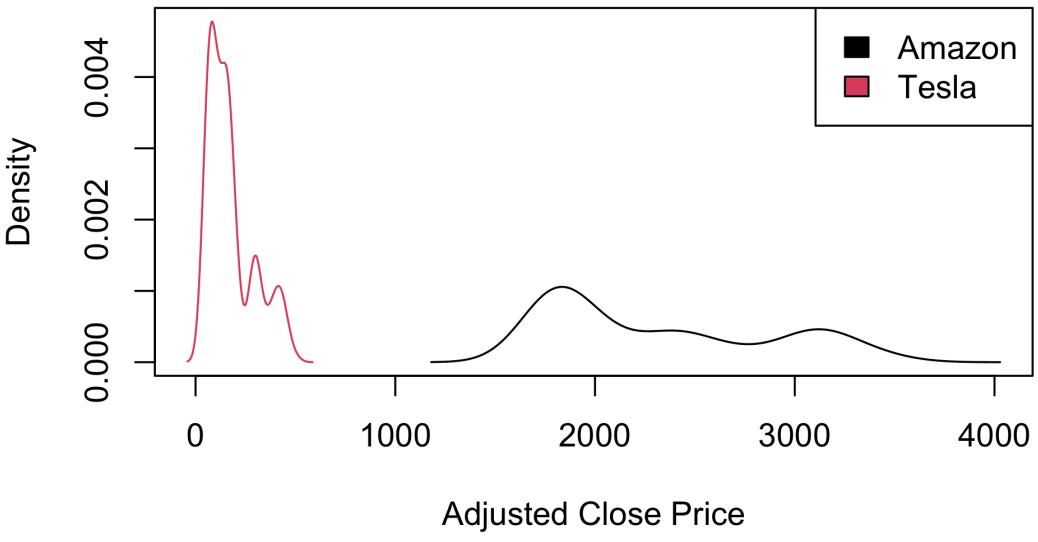


Figure 2: Densplot of the Adjusted Close Price of Amazon and Tesla

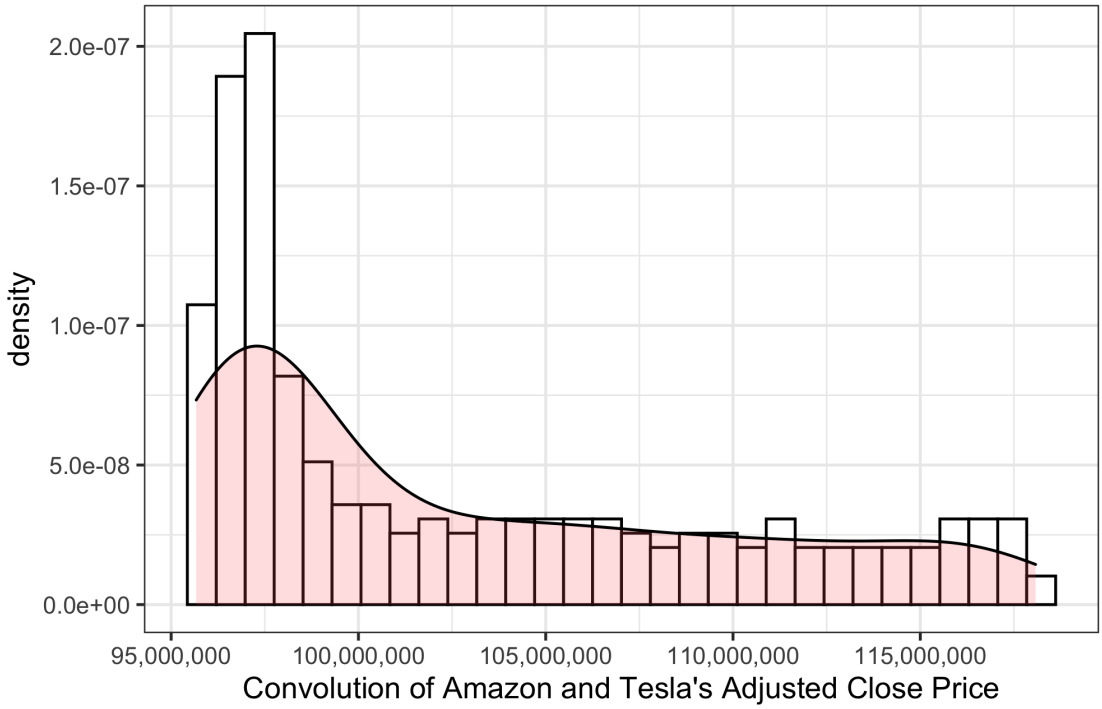


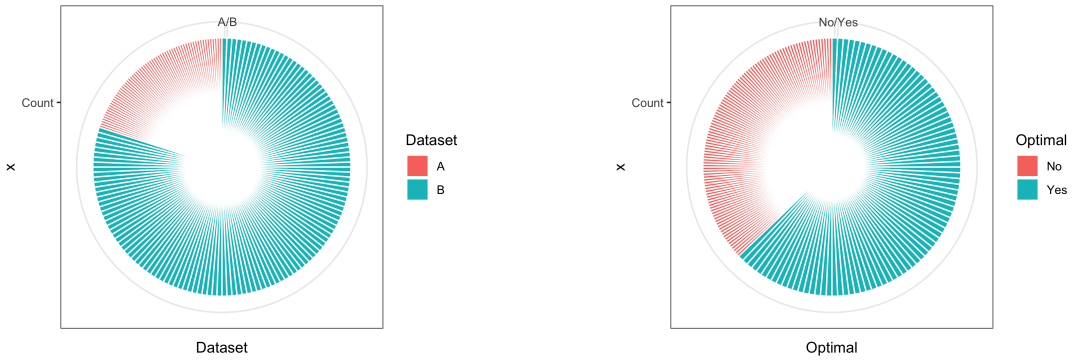
Figure 3: Histogram of the Convolution of the Adjusted Close Price of Amazon and Tesla

3 Chi-Squared Test

The Chi-Square test of independence tests whether there is a relationship between two or more categorical variables [5]. Categorical or nominal data refers that one uses labels instead of numbers; for example race and gender are categorical variables. The central tendency of categorical variables is given by its mode, since median and mean can only be computed on numerical data. Therefore, it does not follow a normal bell-curve distribution, and cannot be analyzed with tests based on a normal distribution such as the t-test or ANOVA. The hypotheses are:

- H_0: The variables are independent there is no relationship between the two categorical variables. Knowing the value of one variable does not help to predict the value of the other variable.
- H_1: The variables are dependent, there is a relationship between the two categorical variables. Knowing the value of one variable helps to predict the value of the other variable.

To perform the test, data are taken from the results of my thesis work. For this example,*Dataset* and *Optimal* are the two categorical variables which are the ones to be analyzed. Figure 4 shows how instances performed with these variables. Besides, a contingency table analysis is performed, which is shown below. The hypothesis to be tested is that *Dataset* and *Optimal* are not associated with one another. The p-value is 1.33×10^{-7}, which is less than the significance level of 0.05, so the null hypothesis is rejected, and the conclusion from this hypothesis test is that the *Dataset* and *Optimal* values are not independent, therefore are associated somehow. Finally, the last contingency table contains the expected values which would be true the null hypothesis.



(a) Groups of data of Instances

(b) Instances that reached optimal values

Figure 4: Pie Charts of Results

data.txt

```
       Contingency Table of Dataset and Optimal Solutions
         A     B
  No  0.163 0.837
  Yes 0.534 0.466
```

data.txt

```
       Pearson's Chi-squared test with Yates' continuity correction

data:  tbl
X-squared = 27.821, df = 1, p-value = 1.331e-07
```

data.txt

```
       Contingency Table of Values which H0 would be true
            A        B
  No  34.66667 69.33333
  Yes 29.33333 58.66667
```

4 Covariance

Covariance provides a measure of the strength of correlation between two variables or more sets of variables. In the covariance matrix de C_{ij} element corresponds to the covariance of x_i and x_j, whereas the element C_{ii} is the variance of x_i. The following properties can also be recognized:

- If $\text{COV}(x_i, x_j) = 0$ then variables are uncorrelated,
- If $\text{COV}(x_i, x_j) > 0$ then variables are positively correlated,
- If $\text{COV}(x_i, x_j) < 0$ then variables are negativetily correlated.

For the experiments, data of the adjusted close price of the stock market are downloaded from the Yahoo Finance website, in the timeframe from January 2007 to March 2017, the chosen companies are Procter & Gamble and its german peer Beiersdorf.

The first proof is that $\text{Cov}[aX + b, cY + d] = ac\text{Cov}[X, Y]$. For this experiment coefficients a, b, c, d of different distributions are generated 30 times each, in order to have a certain grade of diversity. In all the iterations results were the same. The used code is shown below:

```r
#1. Proof: Cov[aX+b, cY+d] = acCov[X,Y]
LHS <- numeric()
RHS <- numeric()

for (i in 1:30) {
  a<-runif(1,0,100)
  b<-rnorm(1,0,1)
  c<-rpois(1,5)
  d<-rbinom(1,5,0.4)

  lhs1 <- cov(a * df$PG + b, c * df$BEI.DE + d)
  LHS <- c(LHS,lhs1)
  rhs1 <- a * c * cov(df$PG, df$BEI.DE)
  RHS <- c(RHS,rhs1)
}
```

code/covariance.R

The second proof is that $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$. For this experiment, the mentioned variances of the adjusted close prices are calculated, and then the values of both sides of the equation are compared. The result is confirmed with a value of 968.83. The used code is shown below:

```r
#2. Proof: Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]
lhs2 <- var(df$PG + df$BEI.DE)
rhs2 <- var(df$PG) + var(df$BEI.DE) + 2 * cov(df$PG,df$BEI.DE)
```

code/covariance.R

To prove that $\text{Cov}[aX + b, cY + d] = ac\text{Cov}[X, Y]$, by definition [2] and with parameters a, b, c, and d holding constants:

$$\begin{aligned}
\text{Cov}[aX + b, cY + d] &= \mathbb{E}[(aX + b)(cY + d)] - \mathbb{E}[aX + b]\mathbb{E}[cY + d] \\
&= \mathbb{E}[acXY + adX + bcY + bd] - (a\mathbb{E}[X] + b)(c\mathbb{E}[Y] + d) \\
&= ac\mathbb{E}[XY] + ad\mathbb{E}[X] + bc\mathbb{E}[Y] + bd - ac\mathbb{E}[X]\mathbb{E}[Y] - ad\mathbb{E}[X] - bc\mathbb{E}[Y] - bd \\
&= ac(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) \\
&= ac\text{Cov}[X, Y].
\end{aligned}$$

To prove that $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$:

$$\begin{aligned}
\mathrm{Var}[X+Y] &= \mathbb{E}(X+Y)^2 - \mathbb{E}[X+Y]^2 \\
&= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2(\mathbb{E}XY - \mathbb{E}X\mathbb{E}[Y]) \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y].
\end{aligned}$$

References

[1] Yahoo! Finance. Stock Market Live, Quotes, and Business, 2020. https://finance.yahoo.com, Last accessed on 2020-11-14.

[2] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 2012.

[3] Oscar Alejandro Hernandez Lopez. Probability in R. https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R, 2020.

[4] Oscar Alejandro Hernández López. Study of Mixed Integer Programming Models for the Concrete Delivery Problem, 2020.

[5] Antoine Soetewey. Chi-square test of independence in r, 2020. https://www.statsandr.com/blog/chi-square-test-of-independence-in-r/, Last accessed on 2020-11-13.

[6] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2020.