Math 23C Term Project

Robi Rahman and Sharon Tai

24 March - 14 May 2021

Question

Can the price changes in certain commodities reveal the kind of recession the U.S. is in? How do qualitatively different recessions affect commodities' prices? We are specifically interested in the dotcom crash of the early 2000s, the Great Recession, and the COVID-19 pandemic.

Motivation

"The stock market is not the economy." This refrain frequently sounds. We have decided to assess whether commodity prices that could signal the health of the general public correlated with unemployment rates across two decades. The three significant recessions in this time period affected different populations differently. Could the differences in commodity prices signal which populations were significantly affected, and how affected they were? Are there differences in how each type of recession affects them?

In general, the risk-adjusted long-run expected returns to all publicly traded assets are the same throughout the economy. If one commodity had a higher expected return than another, traders would sell the less profitable asset and buy the more profitable one, until the costs of each are proportionate to their future returns. The best null hypothesis for changes in asset prices is that all of them (specifically all the ones that are equally risky) will rise or fall by the same amount, all other factors equal.

However, different goods change in value differently in relation to different world events, and some commodities are correlated with other ones. We have selected a list of commodities that we hypothesize will behave differently during qualitatively different recessions. Did some of the selected commodities respond differently to these qualitatively different recessions?

Auxiliary questions we'll be considering are 1. What the price signals about the good and its consumers 2. Whether there were supply and demand shocks that affected the prices

Commodities, Recessions, and Unemployment - Robi

Hypothesis

Commodities tend to increase in price when the economy is growing. Recessions are periods with negative GDP growth, so commodity prices will decrease, or at least increase by a lesser amount on average than during growth periods.

Unemployment generally rises during recessions and is low when the economy is strong. Therefore, we will find that unemployment rates are higher during recessions than during non-recession periods.

The 2020-21 coronavirus pandemic is an exception to these, with high unemployment but a prosperous stock market. It will appear different than the other two recessions, such that you can distinguish it from the rest of the data set based on its upward trend in commodity prices.

Analysis

We obtained monthly time series data for nine commodities (crude oil, sugar, soybeans, wheat, beef, rubber, cocoa beans, gold, ice cream)[1], the US unemployment rate[2], the US dollar to Euro exchange rate[3], and an indicator of past US economic recession dates[4]. The data were cleaned, then merged into a dataframe with one observation of each piece of information for every month from February 2001 to February 2021, giving twenty years of historical data going back from this course's spring semester. This interval includes the 2020-21 coronavirus pandemic, the 2007-2009 financial crisis, and the early-2000s dotcom stock crash. The three recessions were assigned a categorical/factor variable in a new column of the dataframe.

To test the hypothesis that commodities had better performance (in the sense of prices increasing over time) during during non-recession periods than during recessions, a new dataframe was created by dividing each commodity's price in each month by its price in the previous month to obtain a table of percentage changes. Histograms were created for each good individually and for all goods together, showing the distribution of changes in price for the whole period, recessions only, and growth periods only. The histograms were extremely similar across these different categories, somewhat contravening the hypothesis that prices would go up during non-recession periods and down during recessions.

Next, logistic regression was used in an attempt to differentiate between recession and non-recession periods. If price changes in goods are distributed differently between these two types of periods, it should be possible to construct a logistic function of a month's price changes whose output is the probability that a month was during a recession, and which is highly correlated with the historical recession indicator. This turned out not to be the case. Logistic models of some of the variables individually, all of the variables together, and some combinations of variables were all completely ineffective at predicting past recessions. Not only did the models exhibit almost no change in their values of recession probability over the range of observed price increases, the predictions created by deploying these models on the training dataset performed far worse (around 55%) than the no-information accuracy rate, a strategy of simply guessing that every month is not a recession (which is true around 85% of the time).

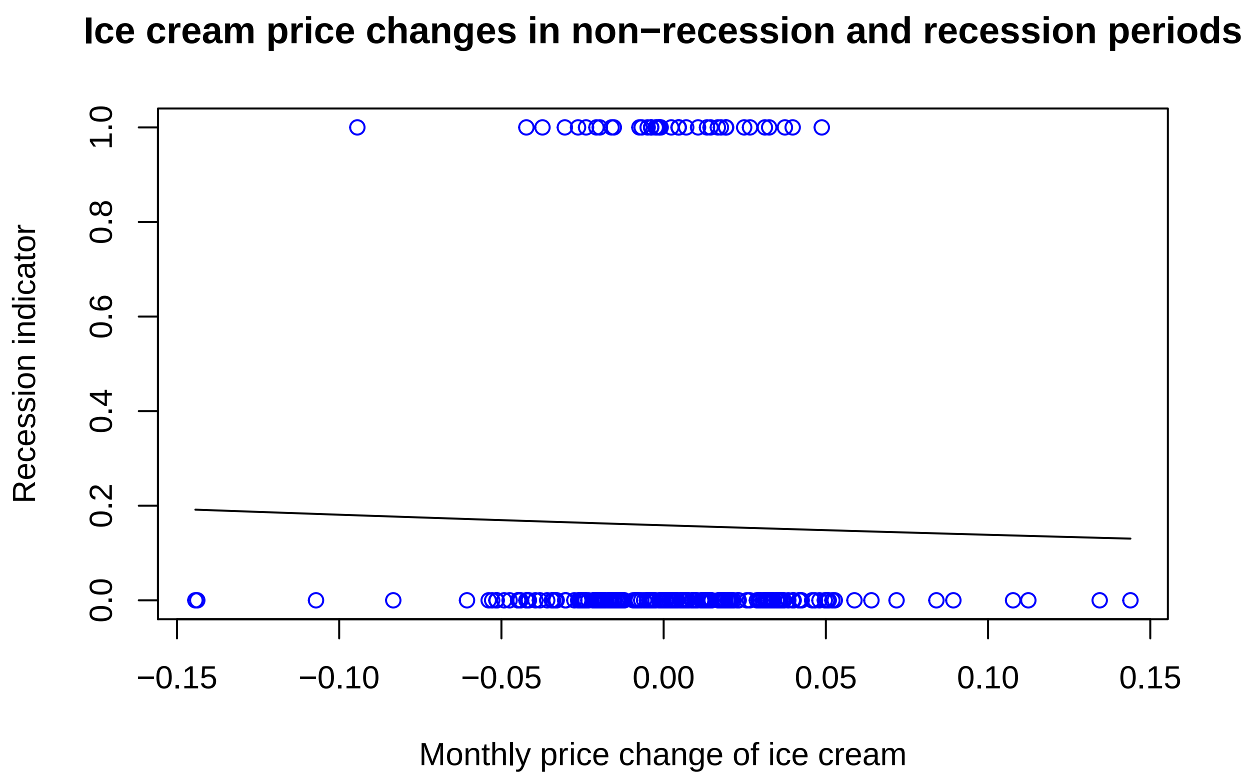
```
# Loading the cleaned data
price_changes <- read.csv("price changes.csv")[,2:15]

# Let's use ice cream for this example, since all the goods fare so poorly.
icecream_regression <- glm(recession_bool ~ Ice_cream, data = price_changes, family = 'binomial')
# Logistic model of recession probability based on price change of ice cream:
summary(icecream_regression)
```

```
##
## Call:
## glm(formula = recession_bool ~ Ice_cream, family = "binomial",
##     data = price_changes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6524  -0.5946  -0.5845  -0.5708   1.9529
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.6687     0.1769  -9.434   <2e-16 ***
## Ice_cream    -1.5900     4.9778  -0.319    0.749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[1] https://www.indexmundi.com/commodities/
[2] https://beta.bls.gov/dataViewer/view/timeseries/LNS14000000
[3] https://fred.stlouisfed.org/series/DEXUSEU
[4] https://fred.stlouisfed.org/series/JHDUSRGDPBR

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 209.71  on 239  degrees of freedom
## Residual deviance: 209.61  on 238  degrees of freedom
## AIC: 213.61
##
## Number of Fisher Scoring iterations: 3
```

```r
plot(price_changes$Ice_cream, price_changes$recession_bool, col=c("blue"),
     xlab="Monthly price change of ice cream", ylab="Recession indicator",
     main="Ice cream price changes in non-recession and recession periods")
curve(inv.logit(-1.5900*x-1.6687), add=TRUE)
```




```r
# You can see that the prediction only ranges from 20% probability that a recession
# is happening in a month when the price of ice cream crashes by 15%, compared to
# a 15% chance that there is currently a recession while the price of ice cream
# is soaring by 15%.


icecream_predictions <- as.factor(predict(icecream_regression, newdata=price_changes, type='response')
confusionMatrix(icecream_predictions, reference = as.factor(price_changes$recession_bool==1))
```

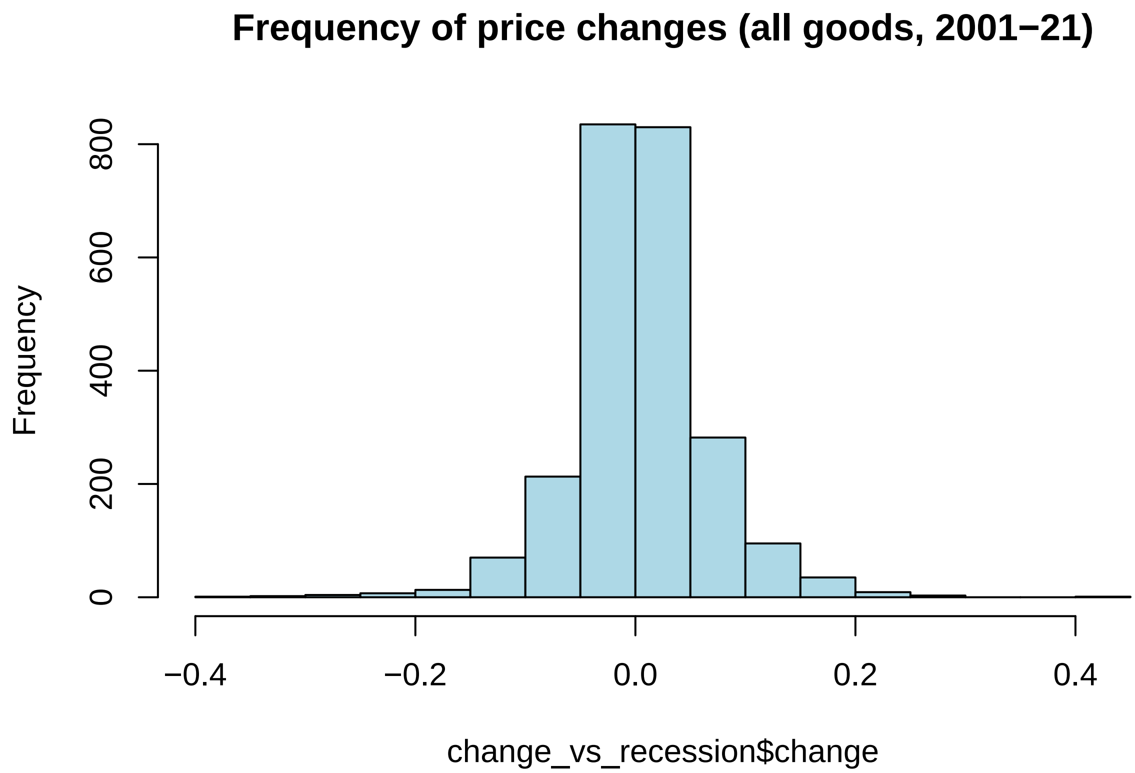
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
```

```
##      FALSE   109   18
##      TRUE     93   20
##
##                Accuracy : 0.5375
##                  95% CI : (0.4722, 0.6019)
##     No Information Rate : 0.8417
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0366
##
##  Mcnemar's Test P-Value : 2.16e-12
##
##             Sensitivity : 0.5396
##             Specificity : 0.5263
##          Pos Pred Value : 0.8583
##          Neg Pred Value : 0.1770
##              Prevalence : 0.8417
##          Detection Rate : 0.4542
##    Detection Prevalence : 0.5292
##       Balanced Accuracy : 0.5330
##
##        'Positive' Class : FALSE
##
```

As you can see from the confusion matrix, the model makes 20 correct positive predictions, 109 correct negative predictions, 13 false negative predictions, and 93 false positive predictions, for an accutacy of 54%. In contrast, simply claiming that all months are *not* recessions would produce 0 correct positive predictions, 202 correct negative predictions, 0 false positive predictions, and 38 false negative predictions, for an accuracy of 84%!

Therefore, contingency tables were used as another method to either show that price changes in recession and non-recession months could be distinguished, or confirm the model's findings that they cannot be.

```
change_vs_recession <- read.csv("cvr1.csv")
hist(change_vs_recession$change, col="lightblue", main="Frequency of price changes (all goods, 2001-21)
```




```r
table(change_vs_recession$change >= 0, change_vs_recession$recession_bool)
```

```
##
##            0    1
##   FALSE  890  168
##   TRUE  1130  212
```

```r
# During recessions: price diff >= 0 212 times; price diff < 0 168 times
# During other times: price diff >= 0 1130 times; price diff < 0 890 times

212/(212+168) # Price goes up 55.8% of the time during recessions
```

```
## [1] 0.5578947
```

```r
1130/(1130+890) # Price goes up 55.9% of the time without recession
```

```
## [1] 0.5594059
```

```r
# Repeat the above analysis, but break up the recession category into three.
change_vs_recession <- read.csv("cvr2.csv")
table(change_vs_recession$change >= 0, change_vs_recession$which_recession)
```

```
##
##            0    1    2    3
##   FALSE  890   42   79   47
##   TRUE  1130   38  101   73
```

```r
1130/(890+1130) # Prices increased 55.9% of the time outside of recessions
```

```
## [1] 0.5594059
```

```r
38/(38+42) # # Prices increased 47.5% of the time during the dotcom recession
```

```
## [1] 0.475
```

```r
101/(101+79) # Prices increased 56.1% of the time during the housing crisis
```

```
## [1] 0.5611111
```

```r
73/(73+47) # Prices increased 60.8% of the time during the COVID pandemic
```

```
## [1] 0.6083333
```

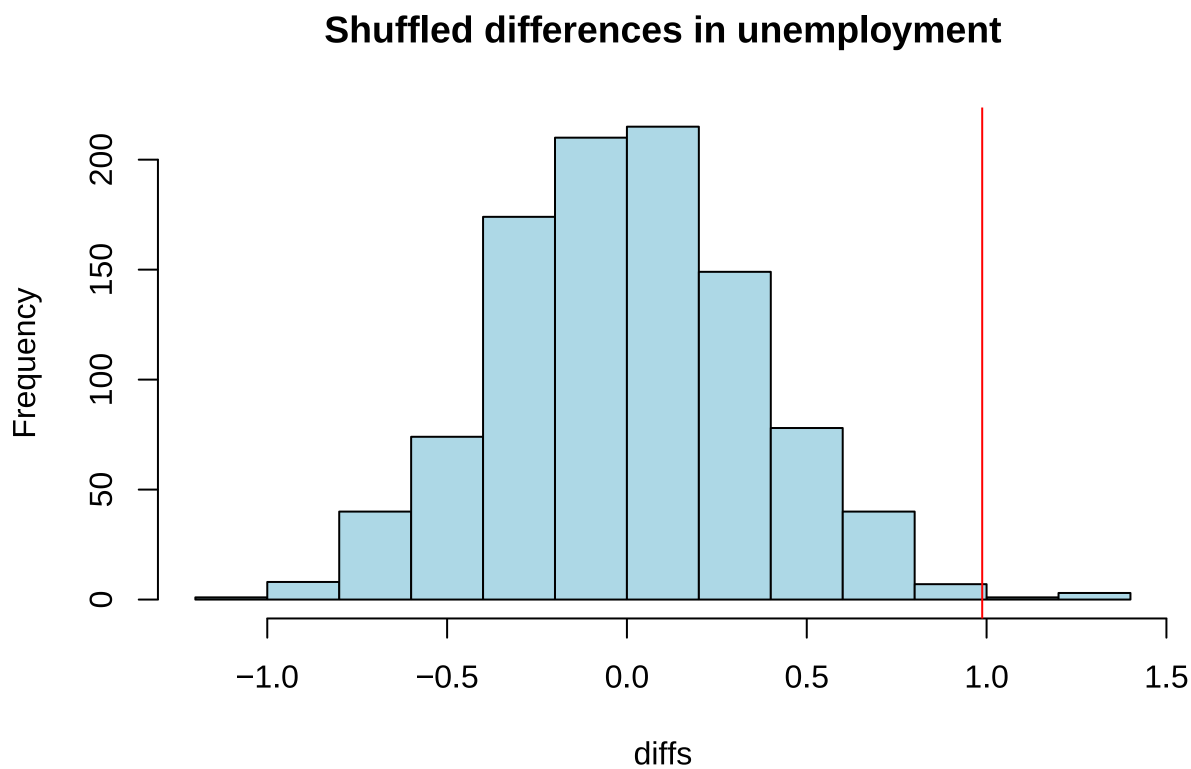
```r
df <- tibble::tribble(
  ~x, ~y, ~price,
  "non-recession", 1130,  "same/increased",
  "non-recession", -890,  "decreased",

  "recession", 212,  "same/increased",
  "recession", -168,  "decreased",

)
ggplot(data = df, aes(x, y, group = price)) +
  geom_col(aes(fill = price), position = position_stack(reverse = TRUE)) +
  geom_hline(yintercept = 0) +
  xlab("") + ylab("") + ggtitle("Likelihood of price increases vs recession")
```

```r
df <- tibble::tribble(
  ~x, ~y, ~price,

  "dotcom crash", 38,"same/increased",
  "dotcom crash", -42,  "decreased",

  "financial crisis", 101,"same/increased",
  "financial crisis", -79,  "decreased",


  "COVID pandemic", 73,"same/increased",
  "COVID pandemic", -47,  "decreased",
)
ggplot(data = df, aes(x, y, group = price)) +
  geom_col(aes(fill = price), position = position_stack(reverse = TRUE)) +
  geom_hline(yintercept = 0) +
  xlab("") + ylab("") + ggtitle("Likelihood of price increases vs recession")
```

As seen by the contingency tables, there is no difference in probability that commodity prices were flat/increasing rather than decreasing in the overall categories of recessions vs non-recession periods, though within the recession category, commodities performed worse during the dotcom crash than during the financial crisis or the coronavirus pandemic. These results all contradict the hypotheses that commodities perform better outside of recessions, and that they performed better during the coronavirus pandemic than during the other two recessions.

For a comparison of classical and simulation methods of statistical inference, a bootstrap test and a two-sample t test were conducted to compare the mean unemployment rate during recession and non-recession periods.

```
hist(diffs, col="lightblue", main="Shuffled differences in unemployment")
abline(v=observed_difference, col="red")
```

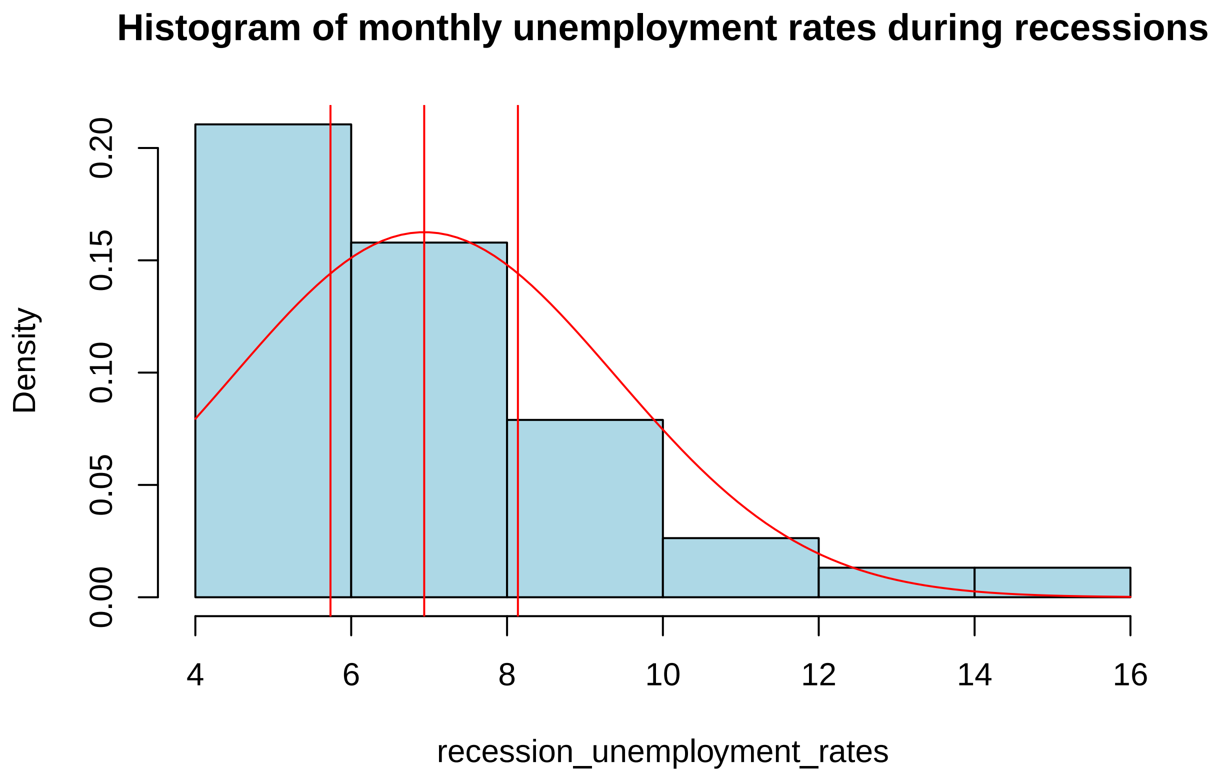


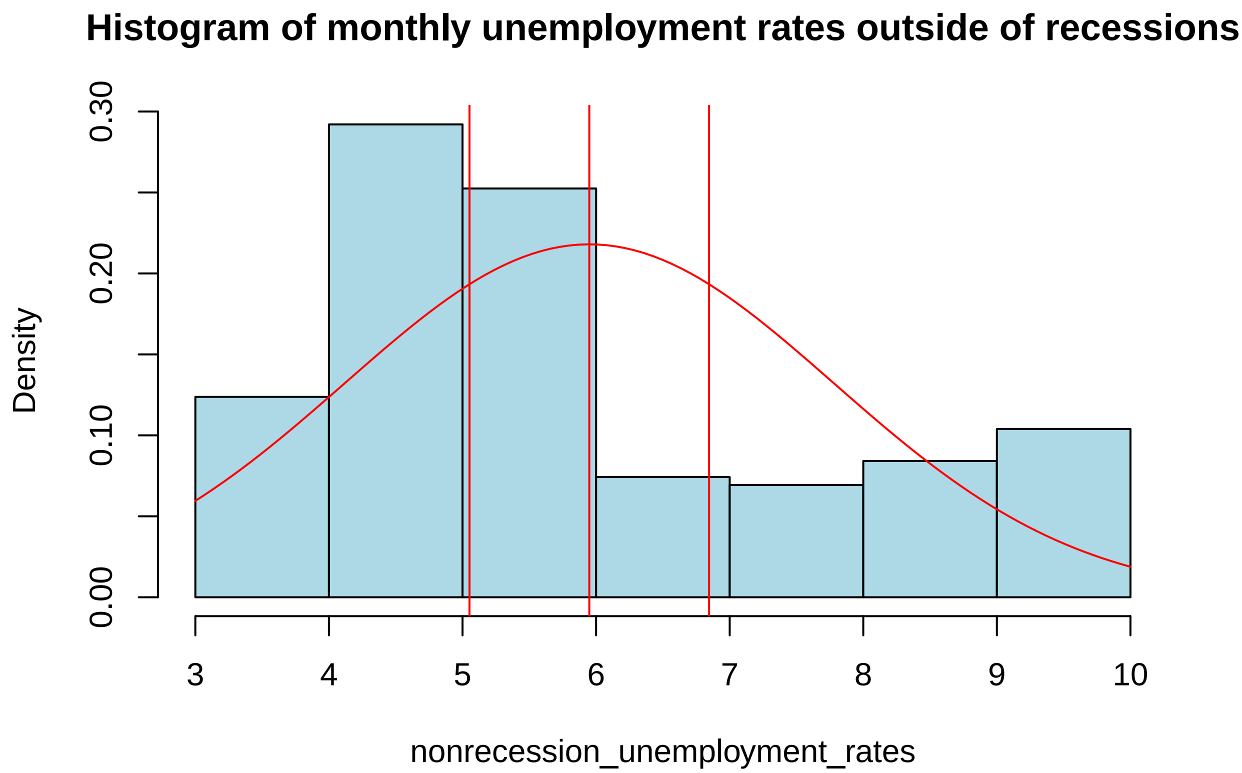

```
pvalue
```

```
## [1] 0.004
```

```
# p-value = 0.004, so there is a significant difference in unemployment during
#                    recessions and non-recession periods!
```

As shown by the observed difference overlaid on a histogram of simulated differences, and the the p-value of 0.004 for the null hypothesis that the two categories have the same mean unemployment rate, the bootstrap test suggests that unemployment is significantly higher during recessions.


Histogram of monthly unemployment rates during recessions
Density
0.00
0.05
0.10
0.15
0.20
4
6
8
10
12
14
16
recession_unemployment_rates




```r
t.test(recession_samples,nonrecession_samples,alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  recession_samples and nonrecession_samples
## t = 1.3824, df = 23.808, p-value = 0.1797
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6078307  3.0703307
## sample estimates:
## mean of x mean of y
##   6.80625   5.57500
```

For comparison, a two-sample t test does not find that there is a significant difference in the mean unemployment rate for recession and non-recession months. In this case, the bootstrap test is likely to be a more reliable method because several prerequisite assumptions for the statistical validity of the t test are not met, such as normality of the distribution of unemployment rates, or equal variance in unemployment within the recession vs non-recession periods.

Finally, some advanced regression techniques can be used for modeling applications based on this dataset. Stepwise regression is a technique wherein a model of many variables is evaluated, then based on adjusted R^2 or other criteria, variables are added or removed and the model is re-evaluated, until it has reached some optimal condition. Using the MASS and car libraries, unemployment was modeled as a linear function of all the other variables, which were then pruned if they did not contribute to the predictive capability of the model.

```
## Loading required package: carData
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4

##
## Attaching package: 'car'

## The following object is masked from 'package:psych':
##
##     logit

## The following object is masked from 'package:boot':
##
##     logit
```

```r
unemployment_reg <- lm(Unemployment ~ ., data=price_changes[,2:13])
vif(unemployment_reg)
```

```
##      Crude_oil         Sugar      Soybeans         Wheat          Beef
##       1.479551      1.233794      1.435639      1.338308      1.131973
##         Rubber   Cocoa_beans          Gold       USD_EUR     Ice_cream
##       1.328514      1.177654      1.220497      1.348971      1.064697
## recession_bool
##       1.009710
```

```r
summary(unemployment_reg)
```

```
##
## Call:
## lm(formula = Unemployment ~ ., data = price_changes[, 2:13])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1333 -1.4326 -0.4351  1.4397  8.2499
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.8967     0.1423  41.432  < 2e-16 ***
## Crude_oil        1.5760     1.6253   0.970  0.33325
## Sugar            1.0118     1.8595   0.544  0.58690
## Soybeans        -0.2237     2.8077  -0.080  0.93657
## Wheat            0.1043     2.0216   0.052  0.95888
## Beef             2.9737     3.1338   0.949  0.34367
## Rubber           0.4761     1.7901   0.266  0.79051
## Cocoa_beans     -2.8164     2.2607  -1.246  0.21411
## Gold             4.0086     3.7396   1.072  0.28489
## USD_EUR         -3.4094     6.6841  -0.510  0.61049
## Ice_cream        1.1013     3.6708   0.300  0.76444
## recession_bool   0.9982     0.3478   2.870  0.00449 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.957 on 228 degrees of freedom
## Multiple R-squared:  0.05806,    Adjusted R-squared:  0.01261
```

```
## F-statistic: 1.278 on 11 and 228 DF,  p-value: 0.2384
```

```r
stepAIC(unemployment_reg)
# Output too long for PDF - view results in R script.
```

First, a model was created using all of the variables. Then, the variables were cross-examined to find their variance inflation factors. All of the values are less than 2, indicating that the predictors are not related to each other, and this regression model does not suffer from multicollinearity. Therefore, all of these variables may be used in a multiple linear regression model of recessions.

Based on the model summary, none of the variables are significant predictors of unemployment except for the recession indicator. (If there is multicollinearity, it is possible to erroneously find that all of the independent variables are not significantly related to the dependent variable even though they may be strongly correlated to it as a whole. However, based on the VIFs, that is not occurring here.)

A stepwise regression improves this model by optimizing the Akaike information criterion to remove unnecessary variables. The stepwise regression shows that the model is optimized when all variables are eliminated except for the recession indicator and the price change of crude oil. It is impractical to predict unemployment using current commodity price fluctuations!

However, this may be more achievable using not only data from the current month, but several months of data history. Vector autoregression is a technique that allows a model to take into account a vector of lagged values of the independent variables.

```
## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: urca

## Loading required package: lmtest
```

```r
var_aic <- VAR(price_changes[,2:13], type = "none", lag.max = 5, ic = "AIC")
```

```r
summary(var_aic)
# Output too long for PDF - view results in R script.
```

It turns out that trying to predict recessions from price history of these goods is futile. None of the variables, nor their histories, are significant contributors to the likelihood of recession in the next month, except for whether there was a recession during the previous month. Of note, however, recessions and crude oil prices are significantly predictive of unemployment, and so could be used to forecast upcoming job market trends.

Distribution Analysis on Daily Data - Sharon

```r
# loading the daily data
daily_data <- read.csv("dailydata.csv", stringsAsFactors=FALSE); head(daily_data)
```

```
##   X       Date priceGOLD priceWTI priceCOCOA priceSUG priceHYS priceWEAT
## 1 1 2001-01-02   272.800    27.29      49.44    25.11 20.41789     24.57
## 2 2 2001-01-03   269.000    27.93      50.15    25.57 19.24219     24.87
```

```
## 3 3 2001-01-04   268.750    27.95      50.52    24.73 17.87054     24.23
## 4 4 2001-01-05   268.000    28.02      50.46    23.76 18.10569     23.00
## 5 5 2001-01-08   268.600    27.44      52.20    23.33 18.55637     23.56
## 6 6 2001-01-09   267.750    27.72      51.96    23.42 18.57597     23.59
##   priceSOYB rec_inds_use.USRECD rec_types_use.USRECD
## 1     24.55                   0                    0
## 2     25.08                   0                    0
## 3     24.34                   0                    0
## 4     24.00                   0                    0
## 5     23.00                   0                    0
## 6     23.46                   0                    0
```

```
dailydata_ALL <- data.frame(daily_data)
```

Please see the long R script (LongRScript_dailydata_Sharon.R) for the full daily data analyses, in which we analyze the daily price data, price changes, and differences in price changes.

Comparison Across Goods

We selected four different goods to do in-depth analysis: gold, oil, sugar, and wheat. The thinking behind this was to compare the distributions of the different price behaviors, both between recessionary and non-recessionary periods, as well as comparison across different types of recessions. Gold and oil were selected because they are goods that are the classically unusual goods. We expected gold's safe haven investment status and oil's inelastic prices to be apparent in their price changes and differences in their price changes. Sugar and wheat, as traditional commodities, were expected to behave differently from these two unusual goods.

Assumptions: The population of goods' prices has an underlying normal distribution.

Hypothesis: We hypothesized that at least either sugar or wheat prices would follow a normal distribution during non-recessionary months, and diverge during recessionary periods. We hypothesized their price changes and differences in prices changes would do the same. This was motivated by understanding that these are goods with substitutes. Oil and gold do not have historically have substitutes, but we made the same underlying assumption.

Conclusion: Instead, we found that none of the goods' daily prices, price changes, or changes in price changes followed a normal distribution. Using a Chi-square test, we rejected the null hypothesis of a normal distribution for all goods; our p-values were all close to 0. For prices, the values clustered too consistently around the mean, no matter the status of whether or not there was a recession, or the type of recession. Price changes always had long tails, but stayed clustered around zero, indicating the stability of prices. And as for changes of the price changes, these tails were very long and thin as well. Prices do change, but rarely. And when they do, they rarely change greatly.

As a follow-up, we used our own code for testing a Pareto distribution using quantiles (this used code from problem set #5). Our initial results do not indicate a Pareto distribution for these goods' prices during these twenty years. A Pareto distribution was not found for these goods' prices changes either. A follow-up study would investigate more the underlying distribution of each good's prices, focusing particularly on other stable distributions, such as a Levy distribution.

We present the results for sugar here.

Prices: The Example of Sugar

```
# Note: the values are strings, not numbers. So need to not include the
# missing values.
daily_price_SUG <- as.numeric(dailydata_ALL$priceSUG[which(dailydata_ALL$priceSUG != ".")])
```

```r
# Recession variables
sug_no_rec <- daily_price_SUG[which(dailydata_ALL$rec_inds_use.USRECD == 0)]
sug_any_rec <- daily_price_SUG[which(dailydata_ALL$rec_inds_use.USRECD == 1)]

#Types of recessions variables
sug_no_rec_type <- daily_price_SUG[which(dailydata_ALL$rec_types_use.USRECD == 0)]
sug_dotcom <- daily_price_SUG[which(dailydata_ALL$rec_types_use.USRECD == 1)]
sug_GreatRec <- daily_price_SUG[which(dailydata_ALL$rec_types_use.USRECD == 2)]
sug_COVID <- daily_price_SUG[which(dailydata_ALL$rec_types_use.USRECD == 3)]
```

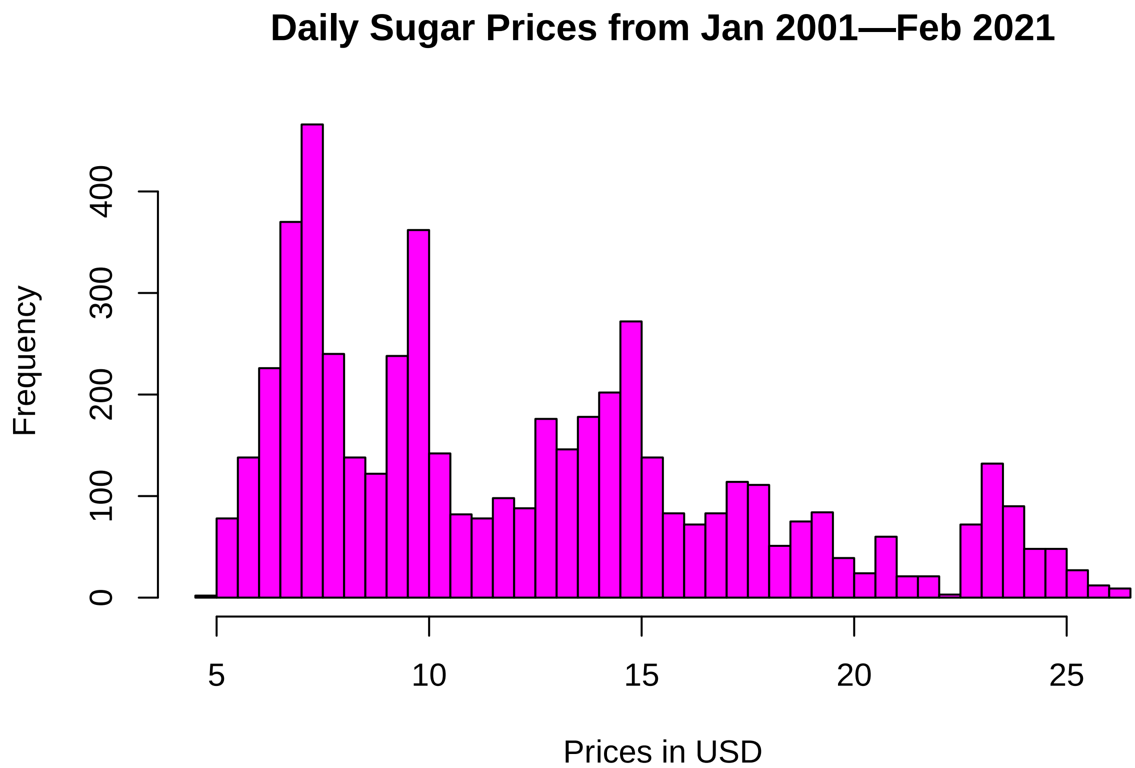
Price of the Good

Note that assessing the prices alone provides an incomplete picture, as we are only looking at twenty-year period for the prices. These prices depend on too many factors for us to meaningfully treat them as random variables with just this set of data. However, the price changes and magnitudes of changes of the price changes are more random, and treating them as random variables could provide more meaningful information about the goods' prices.

Overview of sugar's prices

We did analysis on prices for the different recessionary periods.

```r
# Histogram of prices
hist(daily_price_SUG, breaks=50,
     main="Daily Sugar Prices from Jan 2001-Feb 2021",
     xlab = "Prices in USD", ylab = "Frequency", col="magenta")
```




```r
# Possibly follows a gamma distribution, with greatest frequency between 5 and 10
summary(daily_price_SUG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.92    7.54   11.22   12.37   15.28   26.31
```

```r
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 4.92    7.54   11.22   12.37   15.28   26.31
var(daily_price_SUG)
```

```
## [1] 28.22823
```

```r
#28.22823
sd(daily_price_SUG)
```

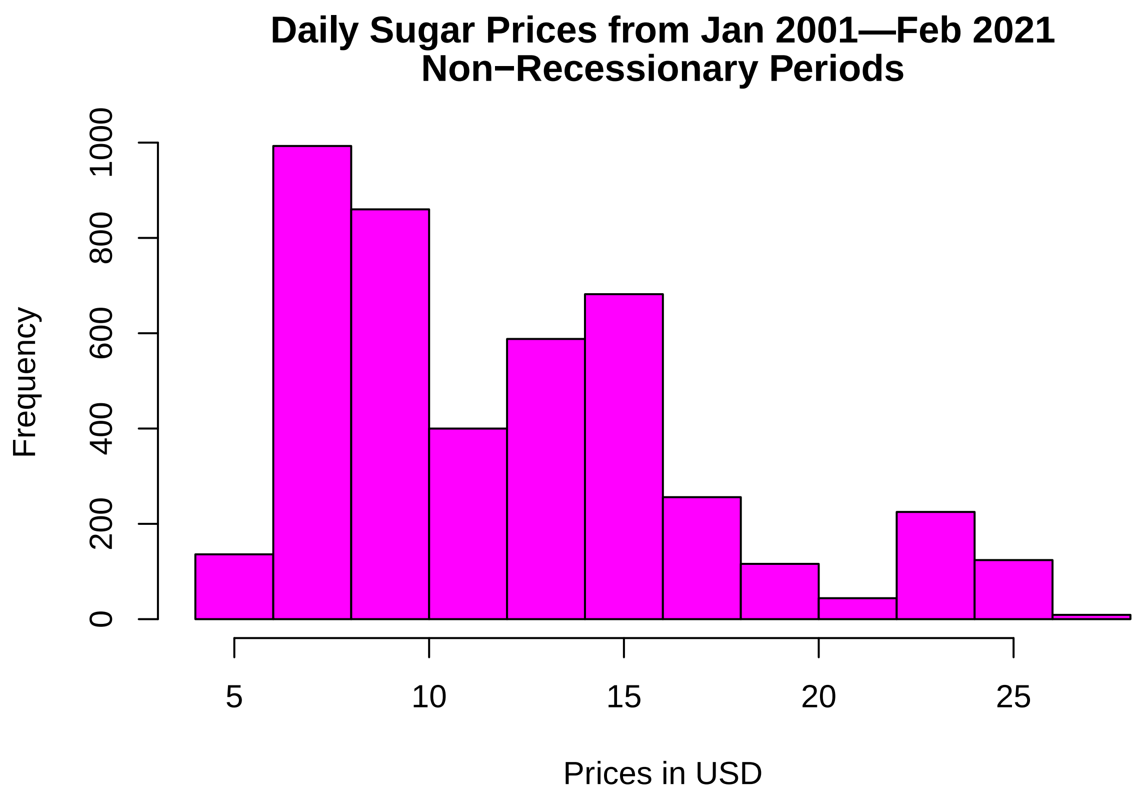
```
## [1] 5.313025
```

```r
#5.313025

# Comparing prices during recessions
summary(sug_no_rec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.92    7.93   11.02   12.17   14.81   26.31
```

```r
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 4.92    7.93   11.02   12.17   14.81   26.31
hist(sug_no_rec,
     main=c("Daily Sugar Prices from Jan 2001-Feb 2021", "Non-Recessionary Periods"),
```

```r
     xlab = "Prices in USD", ylab = "Frequency", col="magenta")
```




```r
# Definitely not normal. Looks similar to wind speeds' distribution, but sugar
# prices do not fit the usual use case for a Weibull distribution.
var(sug_no_rec)
```

```
## [1] 24.92819
```

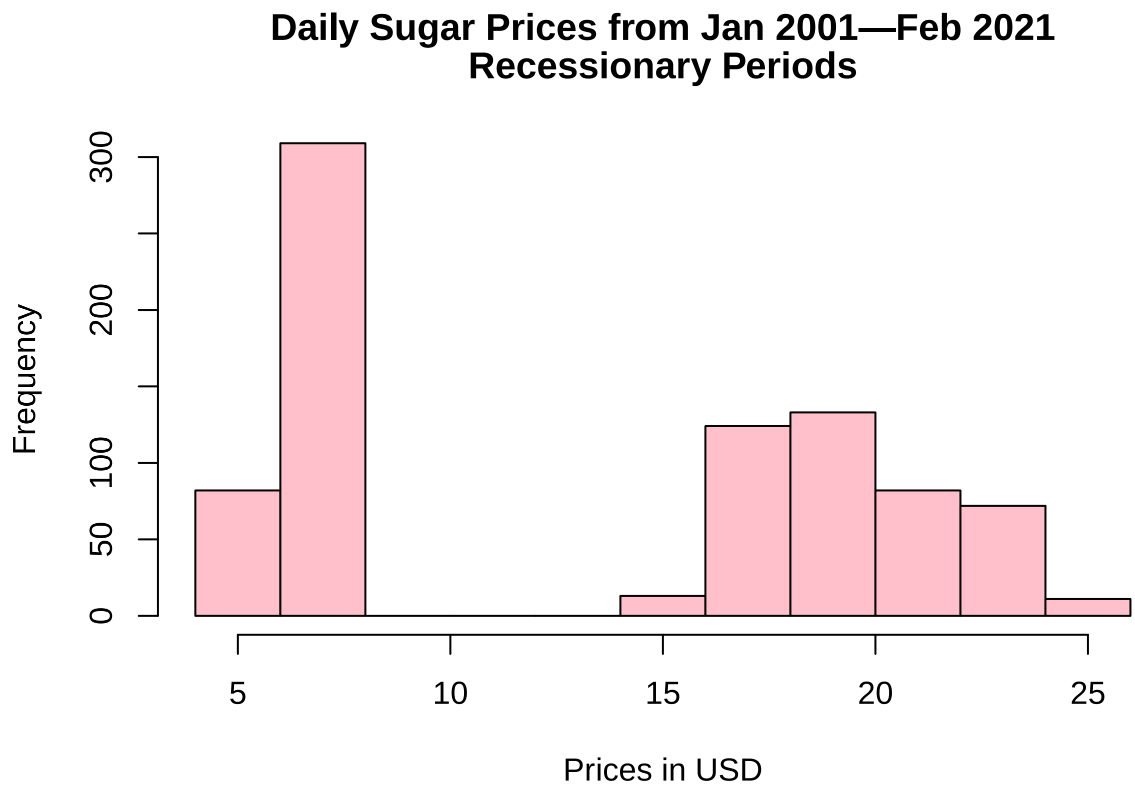
```r
# 24.92819
sd(sug_no_rec)
```

```
## [1] 4.992814
```

```r
# 4.992814

sug_any_rec <- dailydata_ALL$priceSUG[which(dailydata_ALL$rec_inds_use.USRECD == 1 & dailydata_ALL$pric
# Cast the variable to ensure usage as numbers, not strings
sug_any_rec <- as.numeric(sug_any_rec)
summary(sug_any_rec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.92    6.93   16.13   13.50   19.29   24.66
```

```r
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 4.92    6.93   16.13   13.50   19.29   24.66
hist(sug_any_rec,
     main=c("Daily Sugar Prices from Jan 2001-Feb 2021", "Recessionary Periods"),
     xlab = "Prices in USD", ylab = "Frequency", col="pink")
```




```r
# Extremely bimodal. It appears to be 50/50 split in prices < $7 and prices >$14.
var(sug_any_rec)
```

```
## [1] 44.49444
```

```r
# 44.49444
sd(sug_any_rec)
```

```
## [1] 6.670416
```

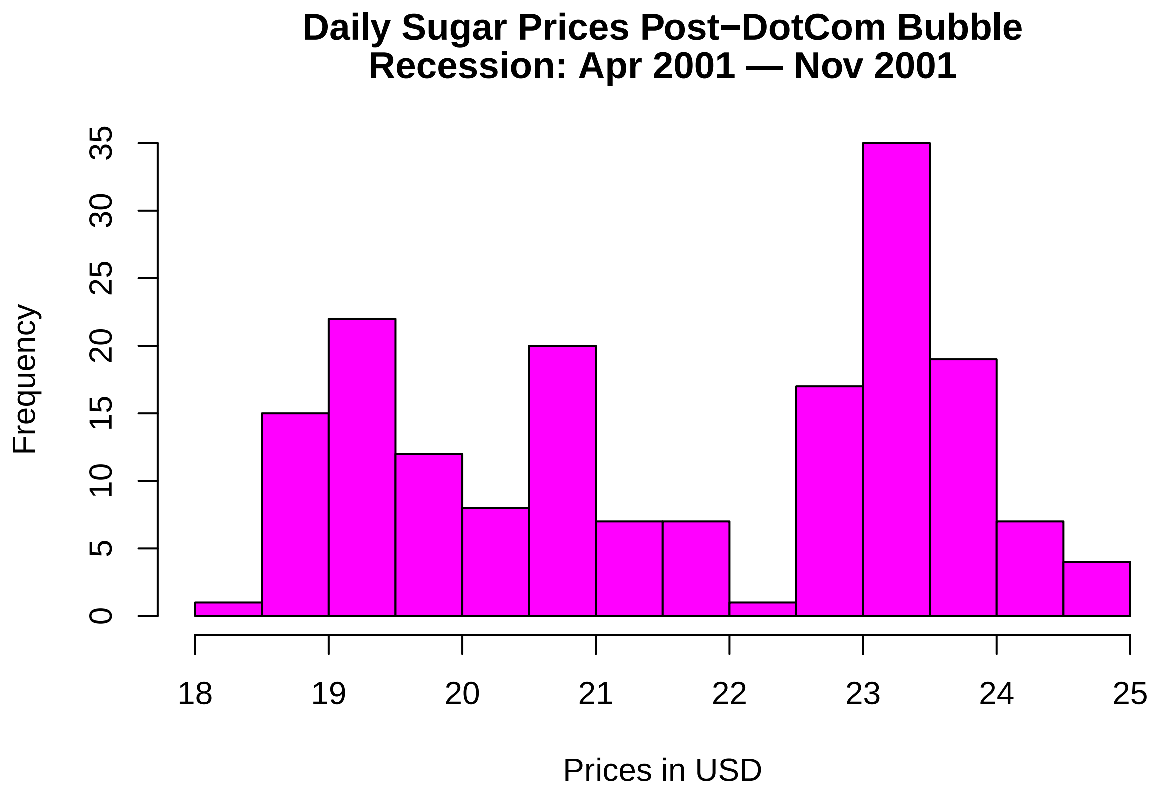
```r
# 6.670416
```

For sugar, the prices are heavily weighed on the lower half. When we look at prices during the recession, the prices are bimodal. For sugar, as for other goods, we found some difference in variance of the prices among different recessions.

```r
summary(sug_dotcom)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.37   19.80   21.73   21.62   23.35   24.66
```

```r
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 18.37   19.80   21.73   21.62   23.35   24.66
hist(sug_dotcom,
     main=c("Daily Sugar Prices Post-DotCom Bubble", "Recession: Apr 2001 - Nov 2001"),
     xlab = "Prices in USD", ylab = "Frequency", col="magenta")
```




```
# A lot of fluctuation during the dotcom recession.
# More evenly distributed prices than during no recession period.
# Comparisons to other recessions below.
var(sug_dotcom)
```

```
## [1] 3.594999
```

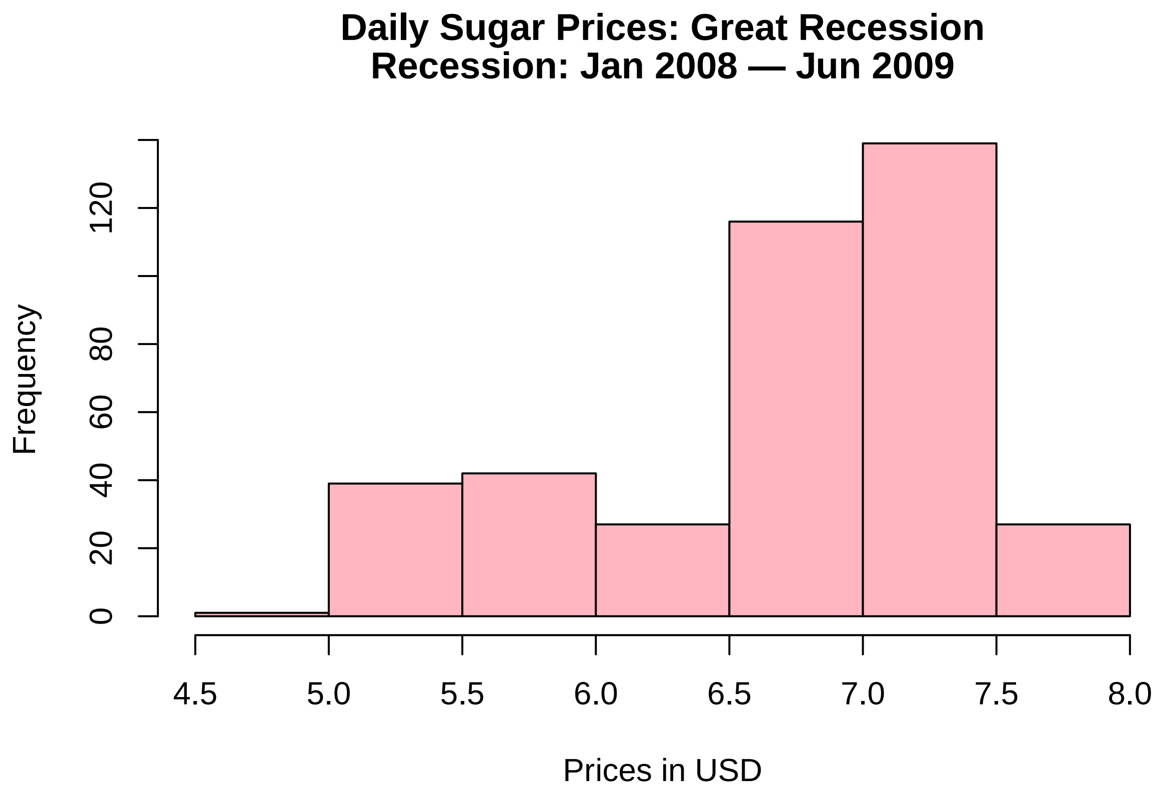
```
# 3.594999
sd(sug_dotcom)
```

```
## [1] 1.896048
```

```
# 1.896048

summary(sug_GreatRec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.920   6.445   6.900   6.724   7.270   7.810
```

```
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 4.920   6.445   6.900   6.724   7.270   7.810
hist(sug_GreatRec,
     main=c("Daily Sugar Prices: Great Recession", "Recession: Jan 2008 - Jun 2009"),
     xlab = "Prices in USD", ylab = "Frequency", col="light pink")
```




```r
# long lower tail, negative skewness.
var(sug_GreatRec)
```

```
## [1] 0.4793186
```

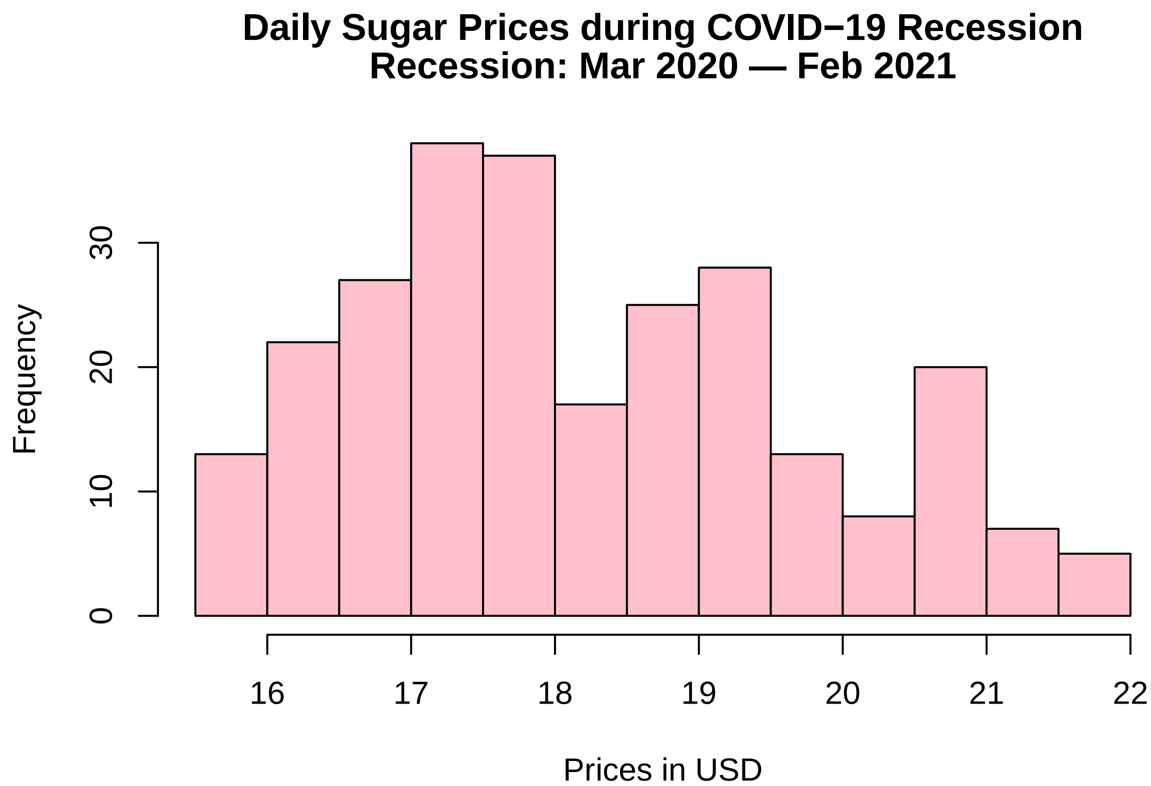
```r
# 0.4793186
sd(sug_GreatRec)
```

```
## [1] 0.6923284
```

```r
#  0.6923284
# The variance during the Great Recession is much lower than during
# the dotcom recession.

sug_COVID <- dailydata_ALL$priceSUG[which(dailydata_ALL$rec_types_use.USRECD == 3 & dailydata_ALL$priceS
sug_COVID <- as.numeric(sug_COVID)
summary(sug_COVID)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.78   17.04   17.84   18.21   19.25   21.80
```

```r
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 15.78   17.04   17.84   18.21   19.25   21.80
hist(sug_COVID,
     main=c("Daily Sugar Prices during COVID-19 Recession", "Recession: Mar 2020 - Feb 2021"),
     xlab = "Prices in USD", ylab = "Frequency", col="pink")
```




```
# Heavier upper tail than during Great Recession; has a slight positive skewness.
var(sug_COVID)
```

```
## [1] 2.385077
```

```
# 2.385077
sd(sug_COVID)
```

```
## [1] 1.544369
```

```
# 1.544369

# See the full R-script for the same analysis rescaled logarithmically.
```

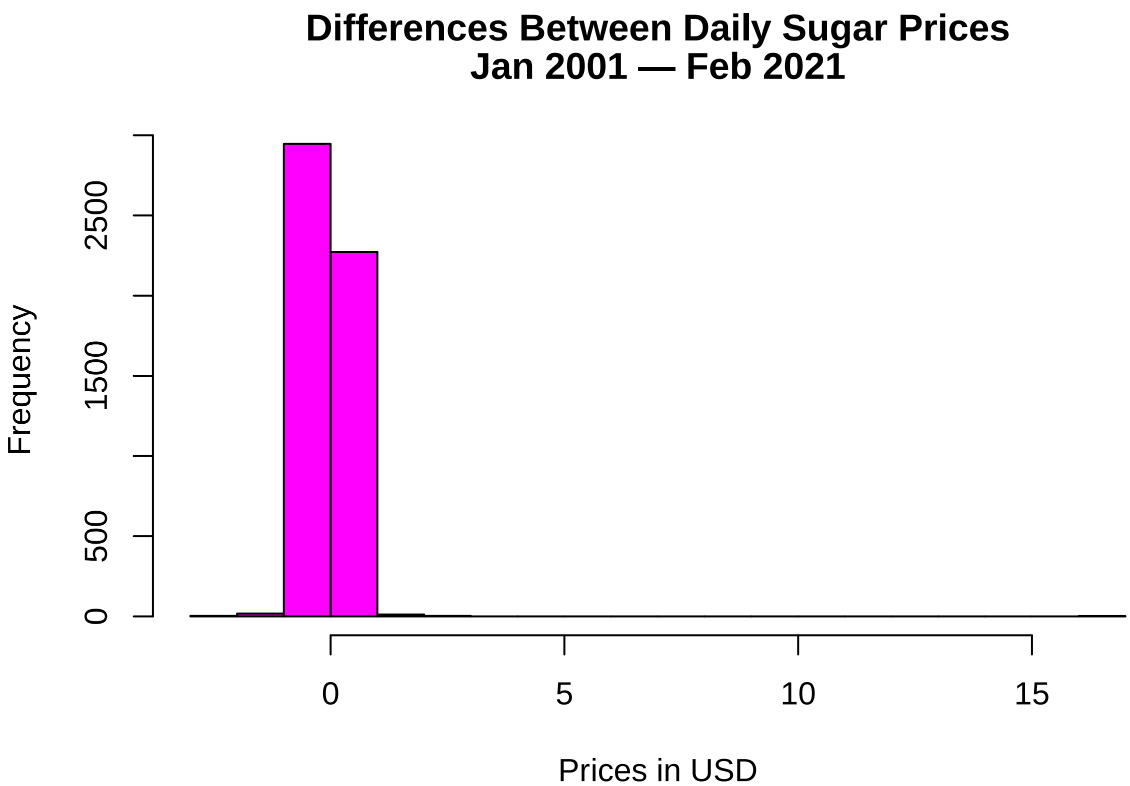
Sugar: First Order and Second Order Price Changes:

Testing Price Changes & Differences in Price Changes

We did the same analysis for price changes, and found more evidence of stable price changes clustered around 0. That is, not only are the price changes clustered around 0, but so are the changes of the price changes. Using a QQ plot and Chi-square test, we reject normality for both the first and second order price changes.

```
daily_sug_price_change <- diff(daily_price_SUG)
daily_sug_price_change_no_rec <- diff(sug_no_rec)
daily_sug_price_change_rec <- diff(sug_any_rec)
daily_sug_price_chng_dotcom <- diff(sug_dotcom)
daily_sug_price_chng_GR <- diff(sug_GreatRec)
```

```r
daily_sug_price_chng_C19 <- diff(sug_COVID)

hist(daily_sug_price_change,
     main=c("Differences Between Daily Sugar Prices", "Jan 2001 - Feb 2021"),
     xlab = "Prices in USD", ylab = "Frequency", col="magenta")
```




```r
# Very stable price changes, stable around 0
summary(daily_sug_price_change)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.219999 -0.090000  0.000000 -0.001721  0.070000 16.810001
```

```r
# Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
# -2.219999 -0.090000  0.000000 -0.001721  0.070000 16.810001
var(daily_sug_price_change)
```

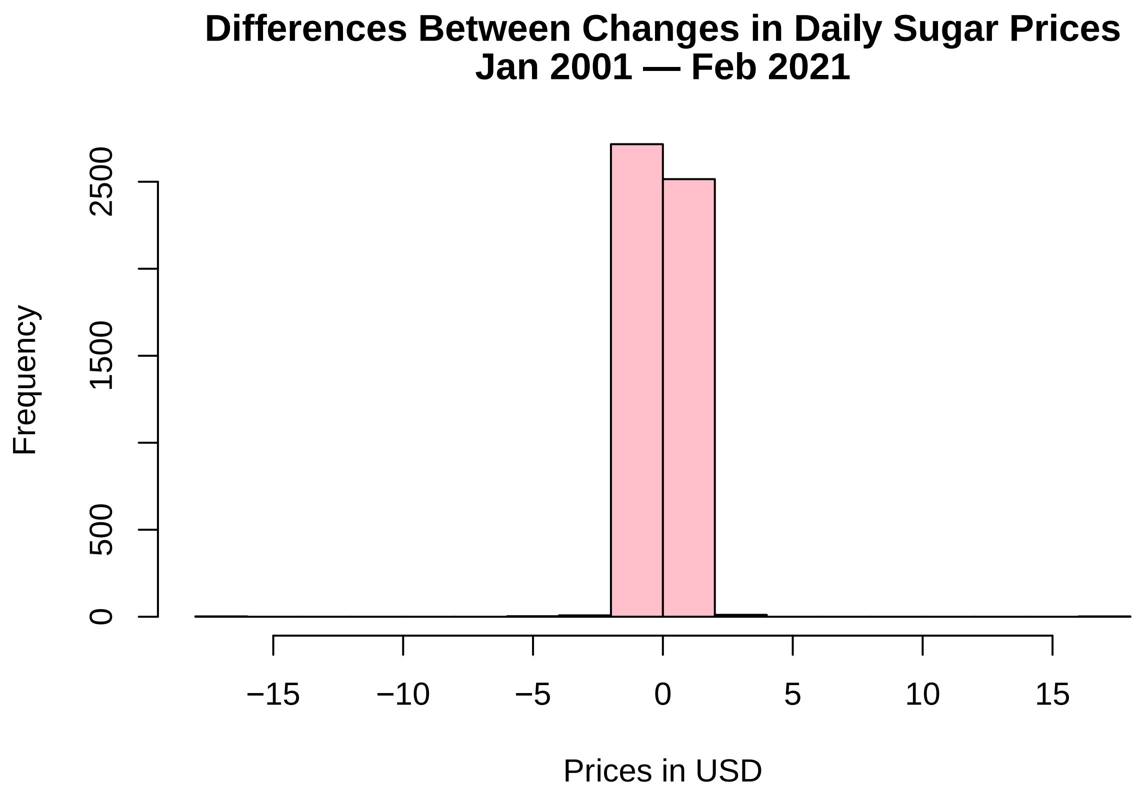
```
## [1] 0.1608409
```

```r
# 0.1608409
sd(daily_sug_price_change)
```

```
## [1] 0.4010498
```

```r
# 0.4010498

# difference in price changes
daily_sug_diff_diff <- diff(diff(daily_price_SUG))
hist(daily_sug_diff_diff,
```

```r
    main=c("Differences Between Changes in Daily Sugar Prices", "Jan 2001 - Feb 2021"),
    xlab = "Prices in USD", ylab = "Frequency", col="pink")
```




```r
# Very small differences in price changes themselves! Always clustered around 0
summary(daily_sug_diff_diff)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -16.350002  -0.140000   0.000000  -0.000074   0.130002  16.810001
```

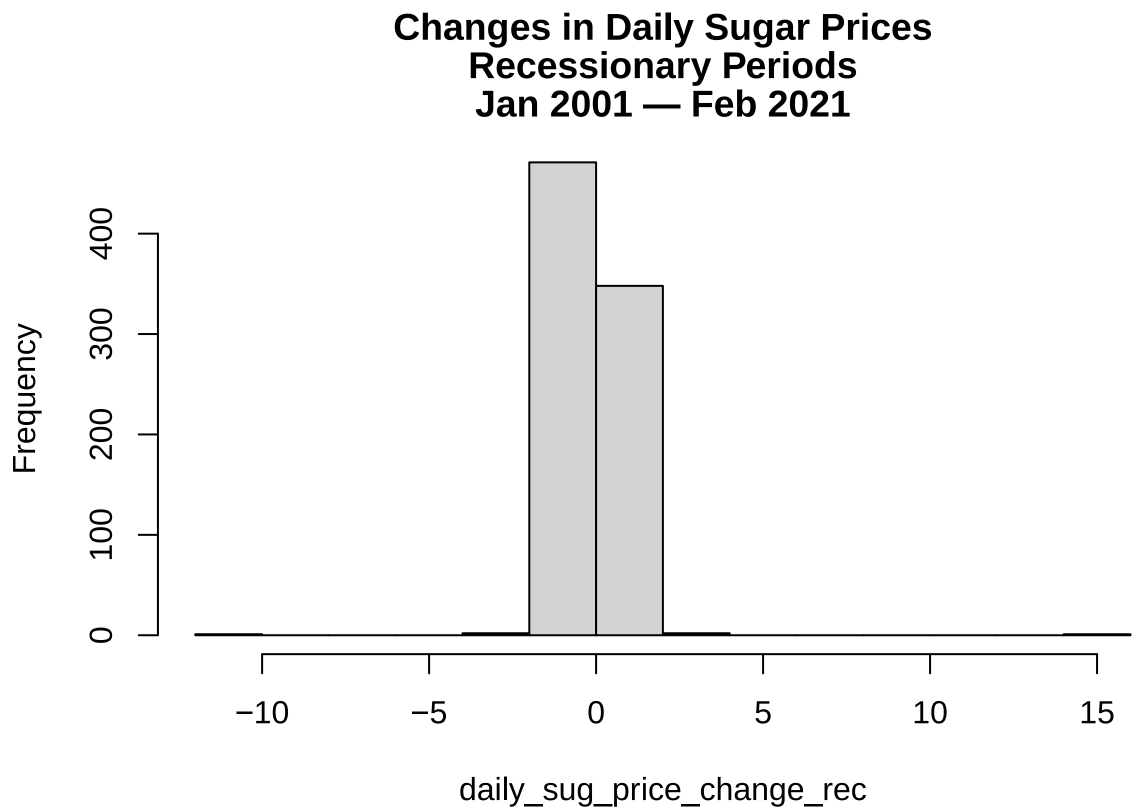
```r
#      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
# -16.350002  -0.140000   0.000000  -0.000074   0.130002  16.810001
var(daily_sug_diff_diff)
```

```
## [1] 0.3365551
```

```r
# 0.3365551
sd(daily_sug_diff_diff)
```

```
## [1] 0.5801336
```

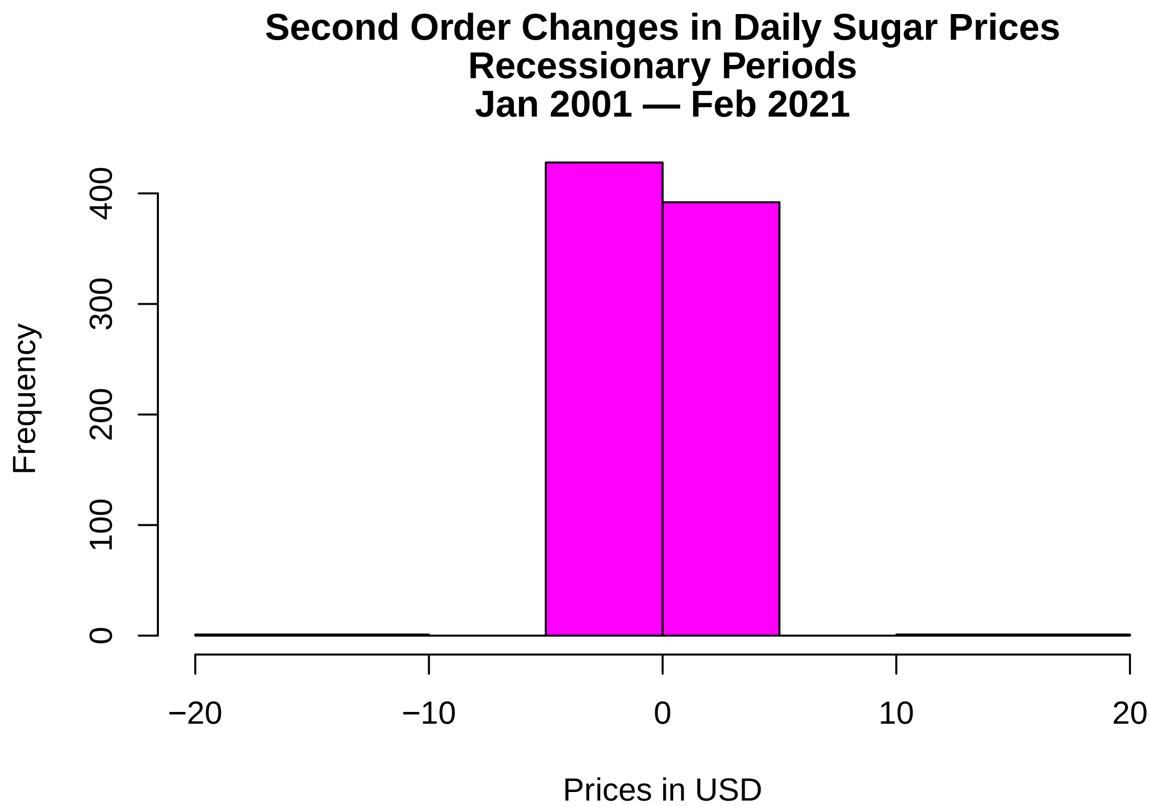
```r
# 0.5801336

hist(daily_sug_price_change_rec,
    main=c("Changes in Daily Sugar Prices", "Recessionary Periods",
           "Jan 2001 - Feb 2021"))
```




```r
summary(daily_sug_price_change_rec)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -11.430001  -0.090000   0.000000  -0.008352   0.050000  15.960000
```

```r
#      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
# -11.430001  -0.090000   0.000000  -0.008352   0.050000  15.960000

# difference in price changes
daily_sug_diff_diff_rec <- diff(daily_sug_price_change_rec)
hist(daily_sug_diff_diff_rec,
     main=c("Second Order Changes in Daily Sugar Prices", "Recessionary Periods",
            "Jan 2001 - Feb 2021"),
     xlab = "Prices in USD", ylab = "Frequency", col="magenta")
```




```r
# Changes in price changes for sugar cluster around 0, but there
# are long thin tails
summary(daily_sug_diff_diff_rec)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -16.050001  -0.130000   0.000000   0.000085   0.122500  16.030000
```

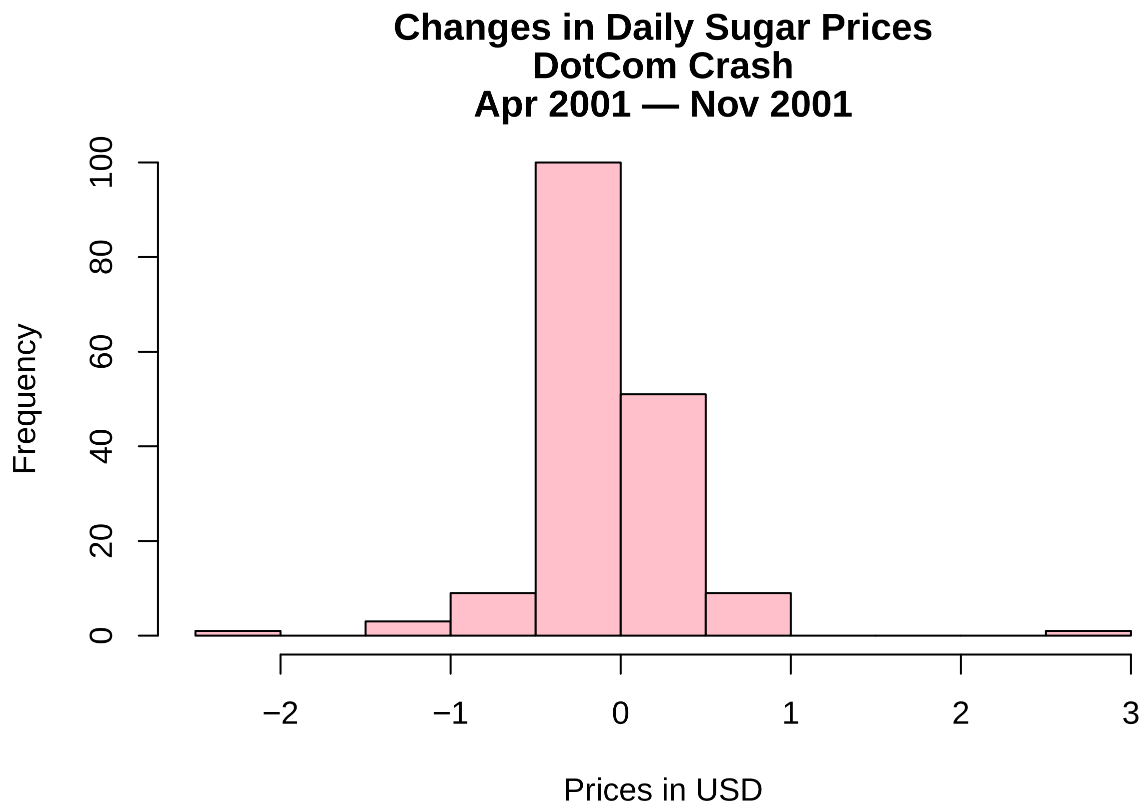
```r
#      Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
#-16.050001  -0.130000   0.000000   0.000085   0.122500  16.030000
var(daily_sug_diff_diff_rec)
```

```
## [1] 1.128211
```

```r
# 1.128211
sd(daily_sug_diff_diff_rec)
```

```
## [1] 1.062173
```

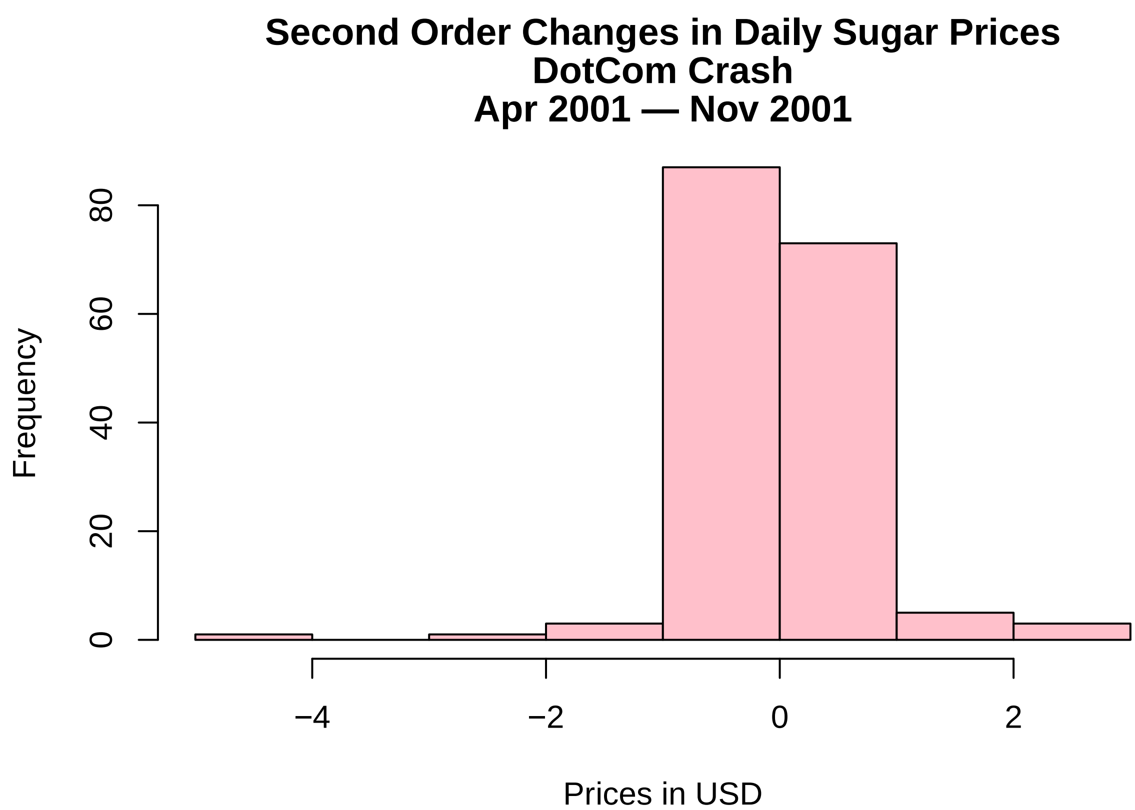
```r
# 1.062173

#****************

hist(daily_sug_price_chng_dotcom,
     main=c("Changes in Daily Sugar Prices",
            "DotCom Crash", "Apr 2001 - Nov 2001"),
            xlab = "Prices in USD", ylab = "Frequency", col="pink")
```




```
# Long thin tails with a tight curve
summary(daily_sug_price_chng_dotcom)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.22000 -0.18000  0.00000 -0.02448  0.10750  2.51000
```

```
#    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
# -2.22000 -0.18000  0.00000 -0.02448  0.10750  2.51000

# difference in price changes
daily_sug_diff_diff_dotcom <- diff(daily_sug_price_chng_dotcom)
hist(daily_sug_diff_diff_dotcom,
     main=c("Second Order Changes in Daily Sugar Prices",
            "DotCom Crash", "Apr 2001 - Nov 2001"),
     xlab = "Prices in USD", ylab = "Frequency", col="pink")
```




```
# During the dotcom crash, we see that the changes
# in price changes have very long and thin tails; there is a little volatility.
# However, most of the changes are still clustered around 0, and the variance is low.
summary(daily_sug_diff_diff_dotcom)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.73000 -0.25000  0.00000  0.00185  0.23000  2.51000
```

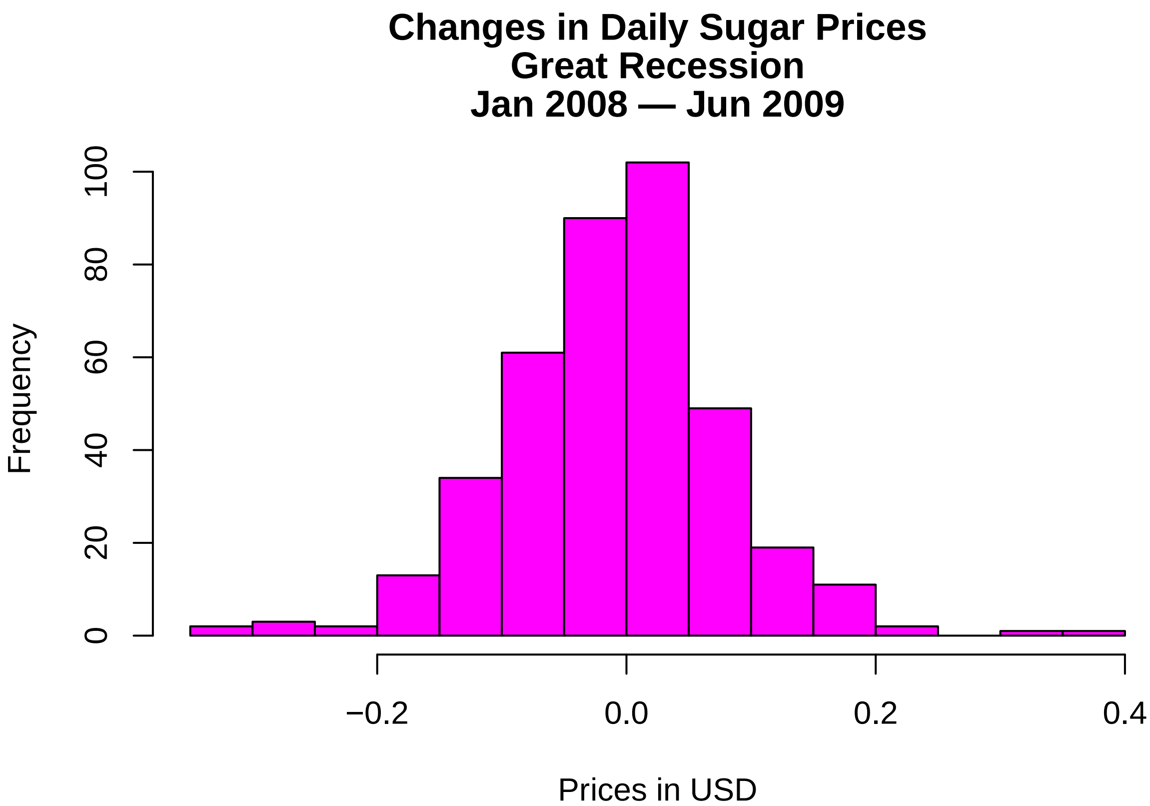
```
# Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
# -4.73000 -0.25000  0.00000  0.00185  0.23000  2.51000
var(daily_sug_diff_diff_dotcom)
```

```
## [1] 0.4446871
```

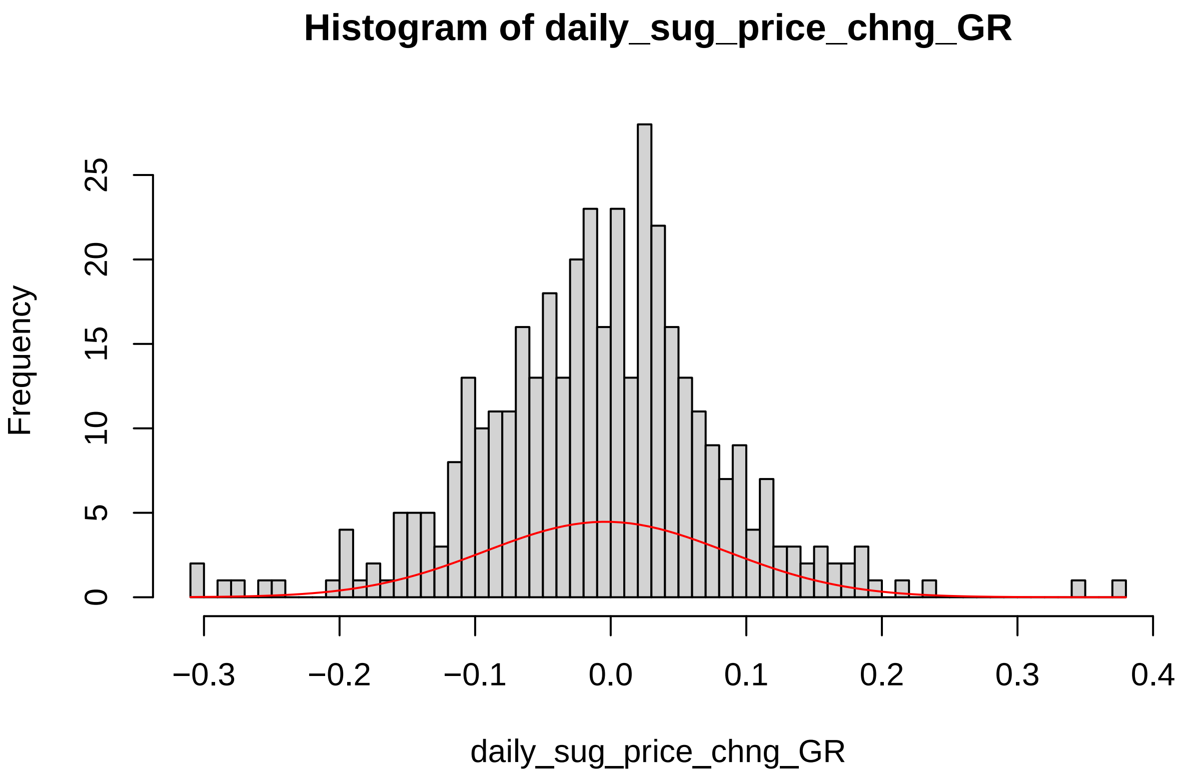
```
# 0.4446871
sd(daily_sug_diff_diff_dotcom)
```

```
## [1] 0.6668486
```

```
# 0.6668486

#****************
hist(daily_sug_price_chng_GR,
     main=c("Changes in Daily Sugar Prices",
            "Great Recession", "Jan 2008 - Jun 2009"),
     xlab = "Prices in USD", ylab = "Frequency", col="magenta")
```



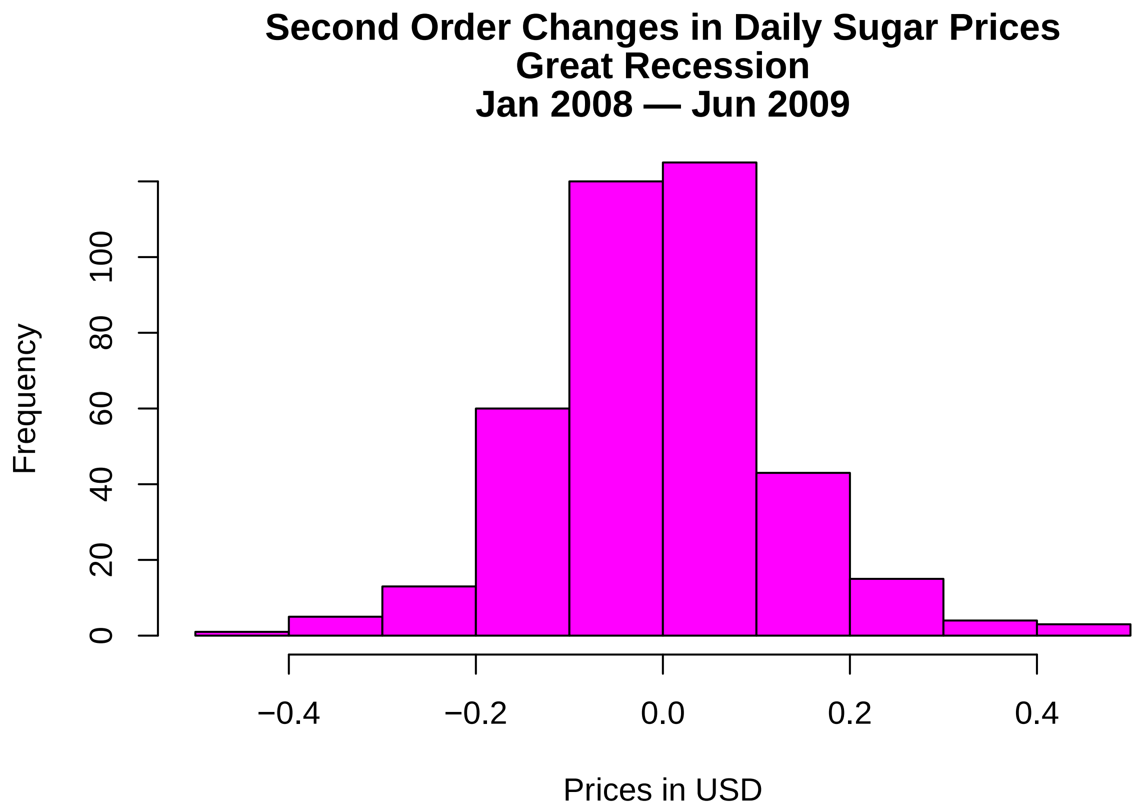

```
hist(daily_sug_price_chng_GR, breaks = 50)
# A much more normal-looking distribution than compared to the others.
curve(dnorm(x, mean(daily_sug_price_chng_GR), sd = sqrt(var(daily_sug_price_chng_GR))), add=TRUE, col =
```



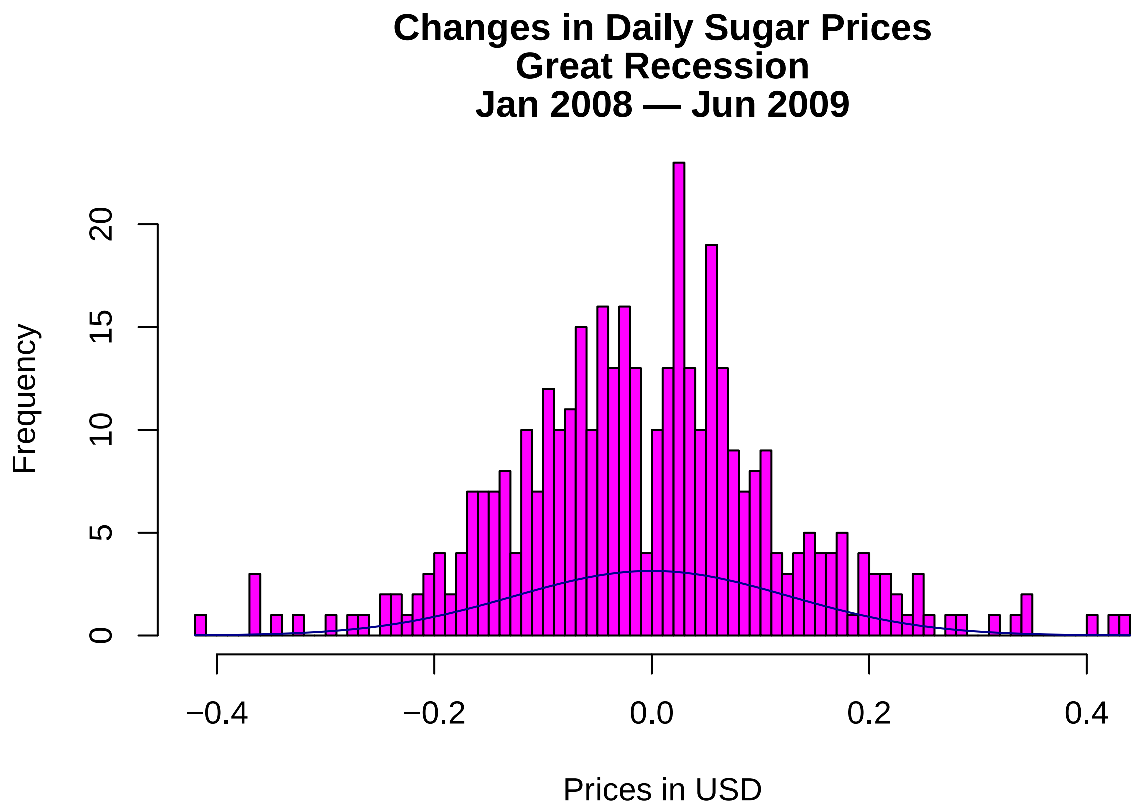

```r
# However, we see that the normal distribution for this mean and this standard deviation
# does not well-match the histogram for daily sugar price changes during the Great Recession.

summary(daily_sug_price_chng_GR)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.310000 -0.060000  0.000000 -0.003821  0.050000  0.380000
```

```r
# Min.    1st Qu.    Median      Mean   3rd Qu.      Max.
# -0.310000 -0.060000  0.000000 -0.003821  0.050000  0.380000

# difference in price changes
daily_sug_diff_diff_GR <- diff(daily_sug_price_chng_GR)
hist(daily_sug_diff_diff_GR,
     main=c("Second Order Changes in Daily Sugar Prices",
            "Great Recession", "Jan 2008 - Jun 2009"),
     xlab = "Prices in USD", ylab = "Frequency", col="magenta")
```




```r
hist(daily_sug_diff_diff_GR, breaks = 100,
     main=c("Changes in Daily Sugar Prices",
            "Great Recession", "Jan 2008 - Jun 2009"),
     xlab = "Prices in USD", ylab = "Frequency", col="magenta")
curve(dnorm(x, mean(daily_sug_diff_diff_GR), sd = sqrt(var(daily_sug_diff_diff_GR))),
      add=TRUE, col = "dark blue")
```




```r
# The curve is too tightly clustered around the mean
# for this to follow the normal distribution.
summary(daily_sug_diff_diff_GR)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.4200000 -0.0800000 -0.0100000 -0.0002571  0.0700000  0.4400000
```

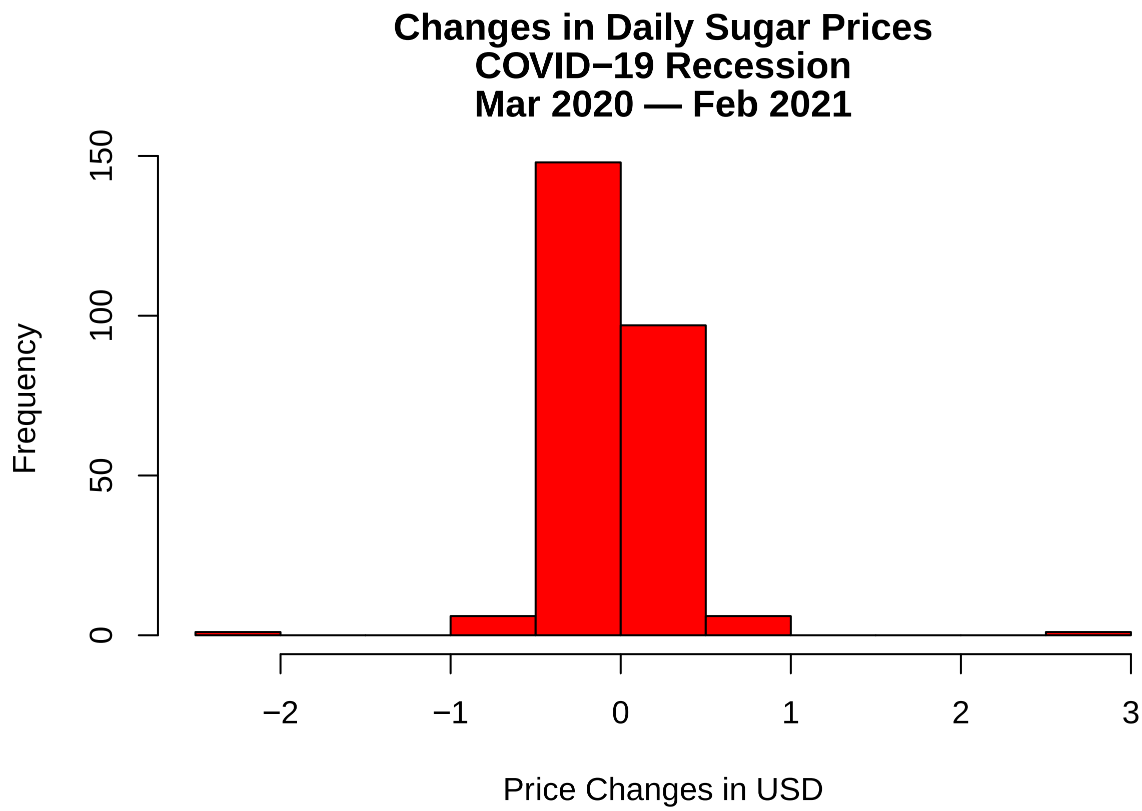
```r
#  Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
# -0.4200000 -0.0800000 -0.0100000 -0.0002571  0.0700000  0.4400000
var(daily_sug_diff_diff_GR)
```

```
## [1] 0.01612983
```

```r
# 0.01612983
sd(daily_sug_diff_diff_GR)
```

```
## [1] 0.1270033
```

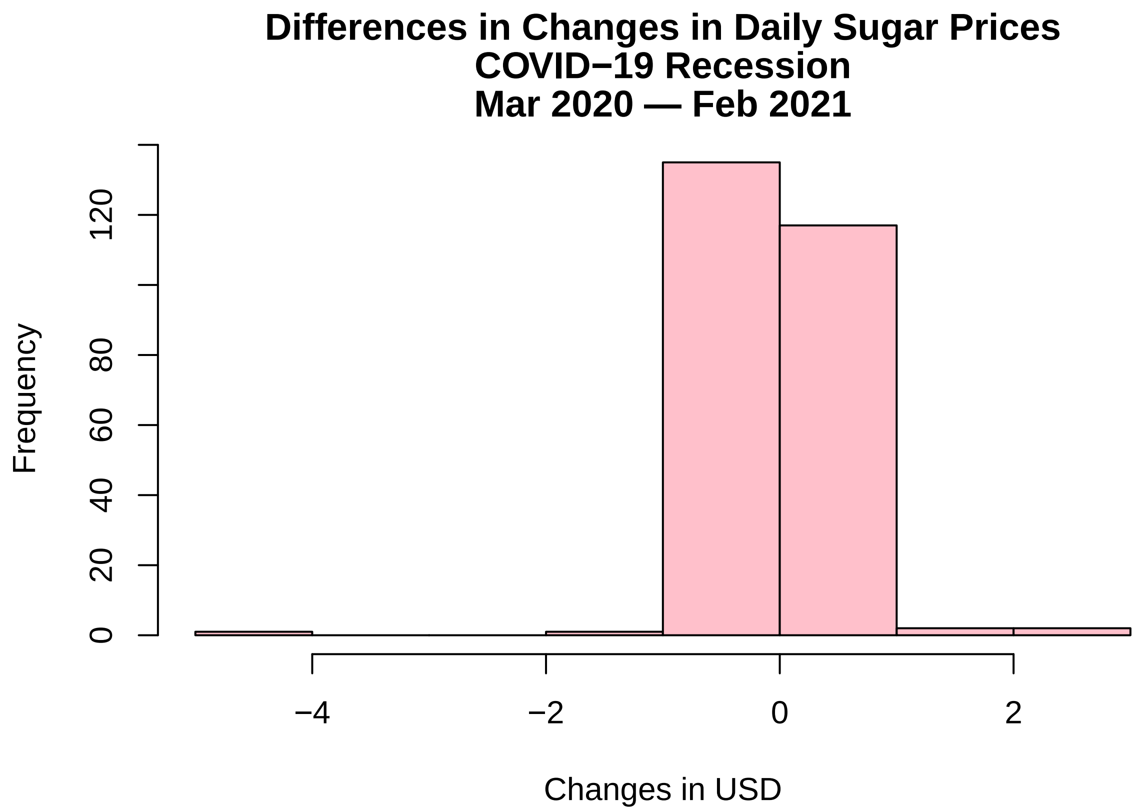
```r
# 0.1270033

#*****************

hist(daily_sug_price_chng_C19,
     main=c("Changes in Daily Sugar Prices",
            "COVID-19 Recession",
            "Mar 2020 - Feb 2021"),
     xlab="Price Changes in USD", ylab = "Frequency", col="red")
```




```r
# The price changes remain clustered around the near-0 mean with a very long and thin
# tail due to the extremes.
summary(daily_sug_price_chng_C19)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.22000 -0.15000  0.00000 -0.02189  0.08500  2.51000
```

```r
#     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
# -2.22000 -0.15000  0.00000 -0.02189  0.08500  2.51000

# difference in price changes
daily_sug_diff_diff_C19 <- diff(daily_sug_price_chng_C19)
hist(daily_sug_diff_diff_C19,
     main=c("Differences in Changes in Daily Sugar Prices",
            "COVID-19 Recession",
            "Mar 2020 - Feb 2021"),
     xlab="Changes in USD", ylab = "Frequency", col="pink")
```




```
# Continues to cluster around the near-0 mean for differences in price changes.
# Continues to have very long and thin tails for the occasional outliers.
summary(daily_sug_diff_diff_C19)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.73000 -0.20750 -0.00500  0.00062  0.20000  2.51000
```

```
#     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
# -4.73000 -0.20750 -0.00500  0.00062  0.20000  2.51000
var(daily_sug_diff_diff_C19)
```

```
## [1] 0.2405248
```

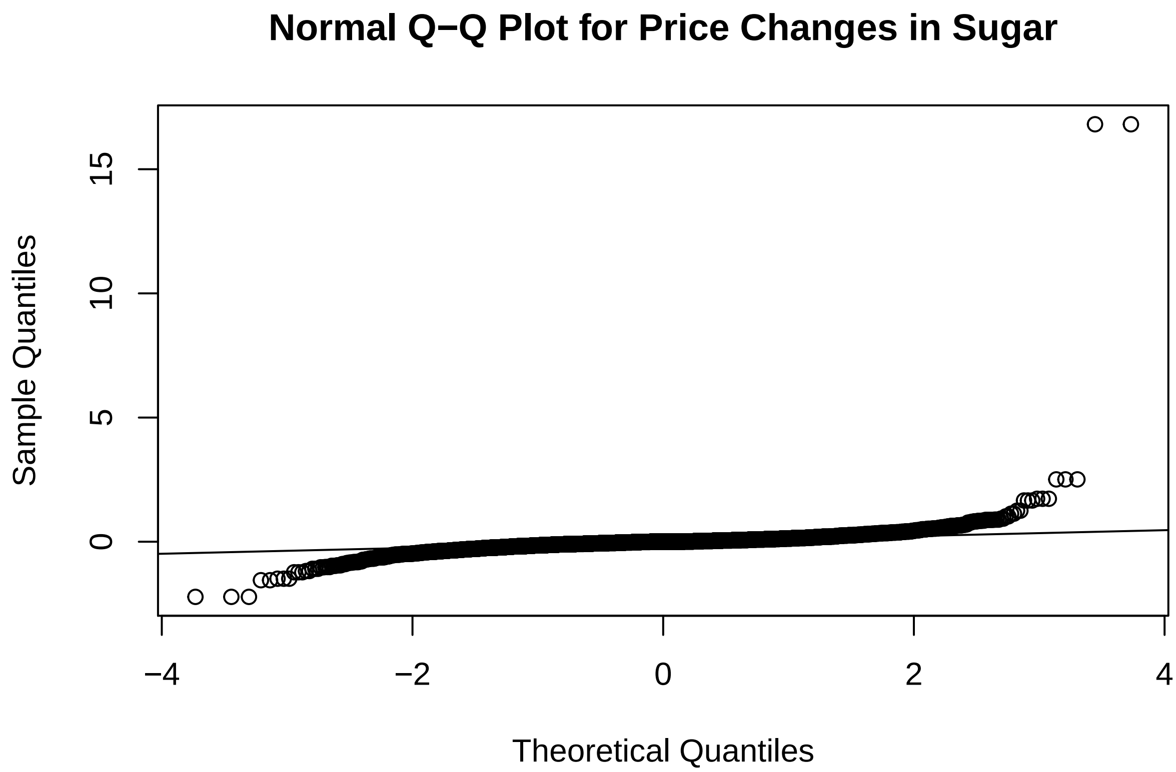
```
# 0.2405248
sd(daily_sug_diff_diff_C19)
```

```
## [1] 0.4904333
```

```
# 0.4904333
```

Testing Normality on First Order Price Changes

```
# Log Price Changes/percentage changes
qqnorm(daily_sug_price_change,
       main="Normal Q-Q Plot for Price Changes in Sugar")
qqline(daily_sug_price_change)
```




```r
# The data usually follow the normal distribution, but the outliers
# of price changes creates heavy tails. The distribution is therefore not
# normal. This matches what we have seen with gold, considered a safe haven good, and
# oil, a price-inelastic good. Sugar is traditionally not thought of as either of these
# types of goods, and yet it similarly has non-normal price changes, with heavy tails.
# This seems to be a trait native to prices themselves, regardless of type of good.
```

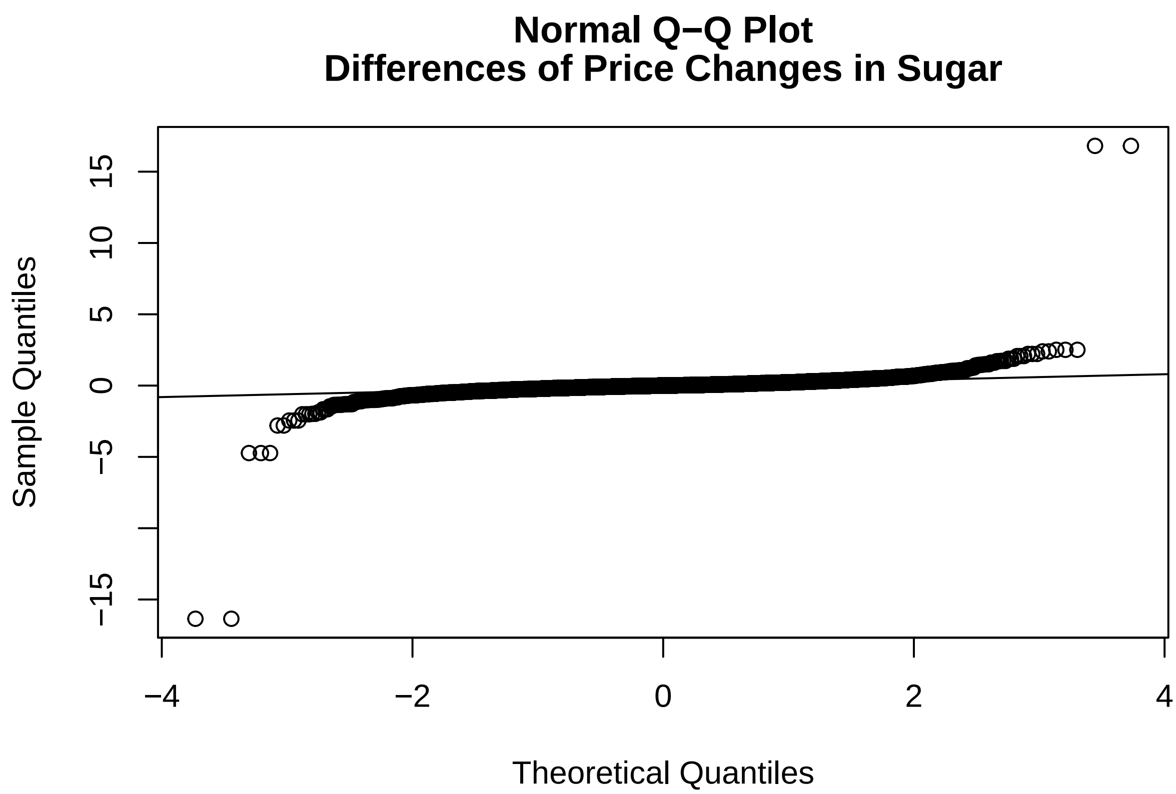
Testing Normality on Second Order Price Changes

```r
daily_pval_sug_diff_diff<- chiSqTest(daily_sug_diff_diff)
```

```
## [1] "Chi-sq test statistic:"
## [1] "9888.67053452539"
## [1] "p-value with df = {nbins - 2}:"
## [1] "0"
```

```r
# "Chi-sq test statistic:"
# "9888.67053452539"
# "p-value with df = {nbins - 2}:"
# "0"
# We definitely reject the null hypothesis that sugar's changes in price changes
# follow a normal distribution.

# Changes in price changes
qqnorm(daily_sug_diff_diff,
       main=c("Normal Q-Q Plot", "Differences of Price Changes in Sugar"))
```

```r
qqline(daily_sug_diff_diff)
```



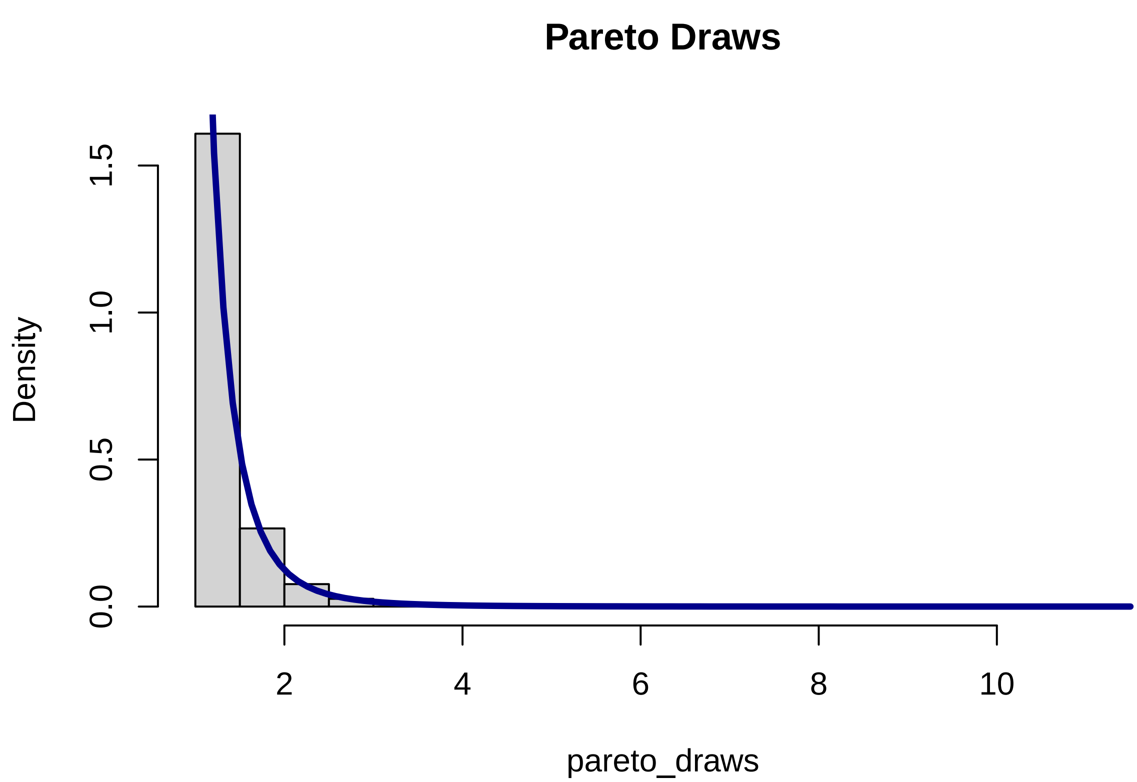

```r
# For sugar, changes in price changes have "lighter" tails, unlike for oil.
```

Testing a Different Distribution: Pareto

At this point, our hypothesis of an underlying normal price distribution for sugar has been rejected. We try instead a Pareto distribution on the prices themselves as well as on the price changes. We are unsuccessful in fitting the Pareto curve to either.

```r
#--------------------------------------
# Daily: Sugar: Pareto distribution
# Using code and notes from STai's PSet #5 R homework
#--------------------------------------
# Let's assess whether the distribution of sugar's prices
# follows a Pareto distribution instead. A Pareto distribution
# can more closely model stock prices.

# Given the density function:
paretopdf <- function(y) 4*y^(-5)

# Distribution function from integrating pdf from 1 to y:
# 1 - (1/y^4).

# Quantile
# q = 1 - (1/y^4)
```

```r
CDF <- function(y) 1 - (1/y^4)
# y = 1 / (1-q)^(1/4)
invCDF <- function(q) 1 / (1-q)^(1/4)

# Generating 10000 uniform random numbers to be the quantiles
quantiles <- runif(10000)
pareto_draws <- invCDF(quantiles)

hist(pareto_draws, prob=TRUE,
     main="Pareto Draws")
curve(paretopdf, col="darkblue", lwd=3.2, add=TRUE)
```

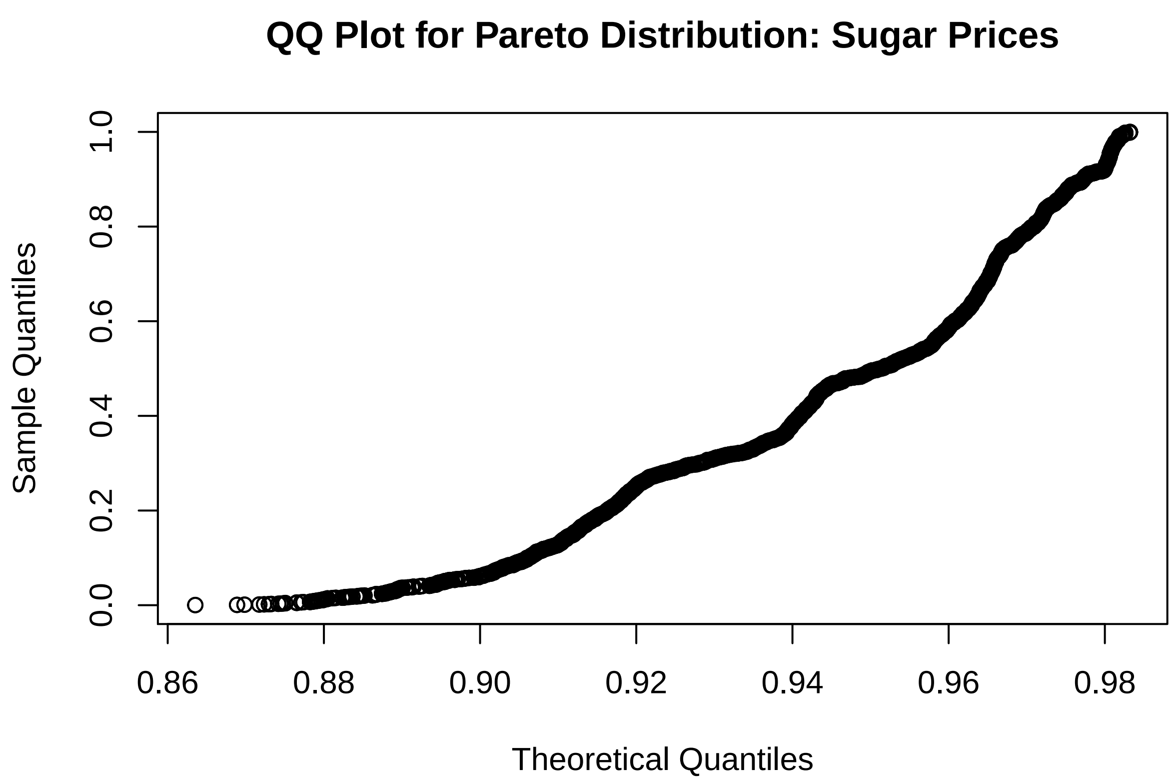



```r
# Note that the curve of the pareto's density function matches the values that were
# randomly drawn according to the distribution function's inverse.

library(fitdistrplus)
```

```
## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:caret':
##
##     cluster

## The following object is masked from 'package:boot':
##
```

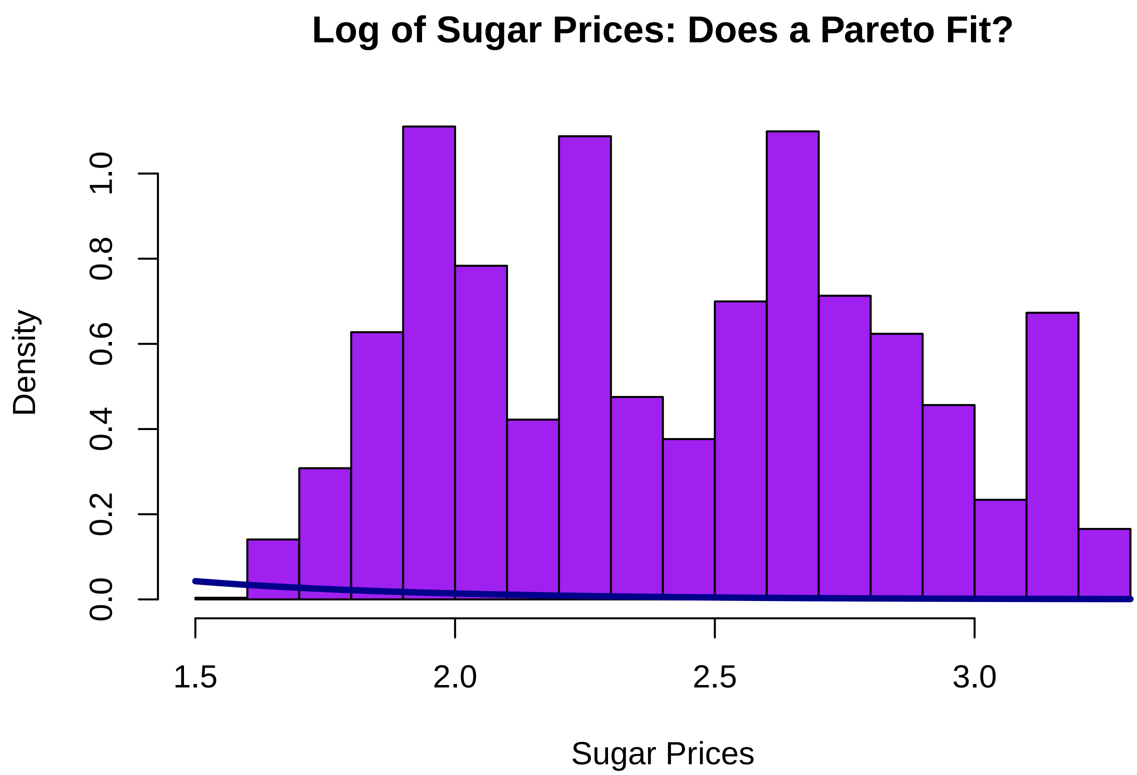
```
##      aml
```

```r
# Use a qq plot to see if the claims follow Pareto distribution with different parameters

# The 1.25 creates a straight line between the theoretical quantiles and sample ones.
CDF <- function(y) 1 - (1/y^1.25)

#-----
# generating quantiles for the number of data points in the sample
# e.g. if 100 data points, then [1/100, 2/100, 3/100, ..., 100/100]
sug_noNA <- dailydata_ALL$priceSUG[which(dailydata_ALL$priceSUG != ".")]
length_sug_noNA <- length(dailydata_ALL$priceSUG[which(dailydata_ALL$priceSUG != ".")])
sample_quantiles_sug <- (1:length_sug_noNA) / length_sug_noNA

# sorting the data set to compute each datapoint's theoretical quantile if it followed
# the given distribution function. e.g. pareto with parameter of alpha.
theoretical_quantiles_sug <- CDF(sort(as.numeric(sug_noNA)))

# This QQ plot illustrates how well the theoretical distribution matches the empirical distribution.
plot(theoretical_quantiles_sug, sample_quantiles_sug,
     main="QQ Plot for Pareto Distribution: Sugar Prices",
     xlab="Theoretical Quantiles",
     ylab="Sample Quantiles")
```

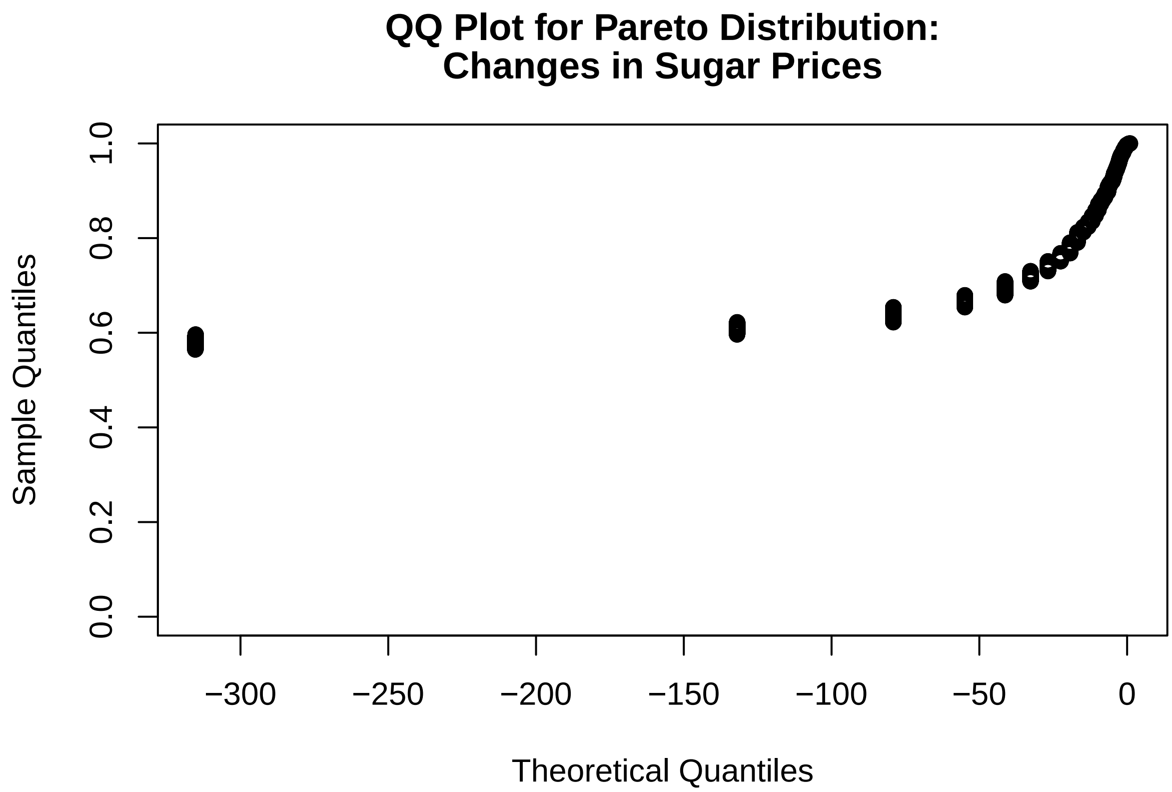



```r
# Sugar prices' Pareto theoretical vs. sample quantiles has more of a linear relationship than
# the other goods prices' Pareto theoretical vs. sample quantils' relationship!
```

```r
# Rescale using log

alpha = 1.25
pdf = function(y) alpha*exp(y)^(-alpha-1)
hist(log(as.numeric(sug_noNA)), prob=TRUE,
     main = "Log of Sugar Prices: Does a Pareto Fit?",
     xlab="Sugar Prices", col="purple")
curve(pdf, col="darkblue", lwd=3.2, add=TRUE)
```



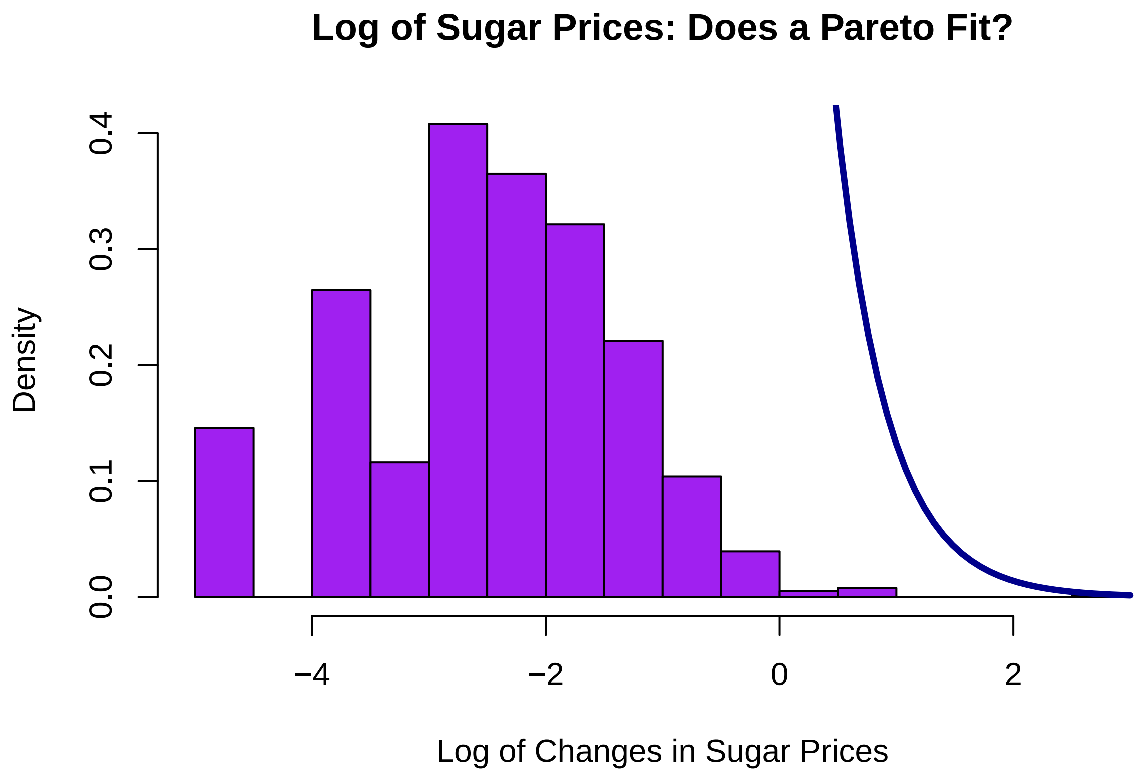

```r
# The curve does not fit the histogram.

# Quick assessment of sugar's price changes
sug_delt_noNA <- diff(as.numeric(sug_noNA))
sug_sample_quantiles_delta <- (1:length(sug_delt_noNA)) / length(sug_delt_noNA)
sug_delt_th_quant <- CDF(sort(sug_delt_noNA))
plot(sug_delt_th_quant, sug_sample_quantiles_delta,
     main=c("QQ Plot for Pareto Distribution:", "Changes in Sugar Prices"),
     xlab="Theoretical Quantiles",
     ylab="Sample Quantiles")
```




```
# Definitely does not follow a line; no Pareto distribution is established.

alpha = 1.25
pdf = function(y) alpha*exp(y)^(-alpha-1)
hist(log(sug_delt_noNA), prob=TRUE,
     main = "Log of Sugar Prices: Does a Pareto Fit?",
     xlab="Log of Changes in Sugar Prices",
     col = "purple")
```

```
## Warning in log(sug_delt_noNA): NaNs produced
```

```
curve(pdf, col="darkblue", lwd=3.2, add=TRUE)
```




```
# Rescaled logarithmically, the Pareto distribution fits the shape of the logs of sugar
# price changes' histogram. However, the Pareto curve does not overlay the
# histogram.
```

Conclusion

These findings of price stability, at both the first and second order price changes, were not unique to sugar. The prices for the other goods, including safe-haven gold and inelastic oil, followed the same behavior. While there were differences in the variance of prices, price changes, and changes of the prices changes among different recessionary periods, the values always remained stubbornly tightly clustered around 0, with either long and thing or long and heavy tails indicating the presence of sometimes unusual circumstances. From these findings, it seems prices themselves, regardless of type of good or recession, do not have a normal population distribution.