Mini-Project 2
Data Collection and Cleaning
Group 5
Dobelman
STAT 486 - Market Models

Connor Barnhill, Brian Graff, Frank Portman

February 14, 2013

1 Coverage Analysis and Graphs

We used annual data[1] from January 2011 to January 2012, and calculated the market capitalization by multiplying *CSHO* (shares outstanding) by *PRCC_C* (price per share) to get *marketCap* (market capitalization) in billions of dollars. To obtain a reasonable market capitalization histogram, we had to truncate all companies with market capitalizations more than 2500 billion dollars. This way we can actually get a clear picture and comment on the distribution of the market caps. From Figure 1, we can see that the distribution is negative exponential.

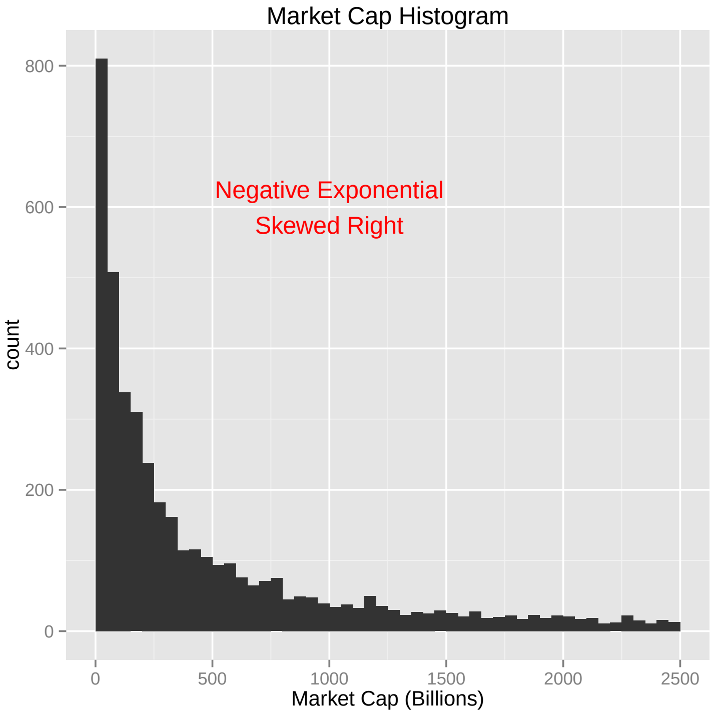

Figure 1: Above: Histogram of market capitalizations.

To calculate the proportion of stocks that pay dividends, we found the number of stocks that had the variable *DVPSP_C* (dividends per share) not equal to zero, and divided that number by the total amount of stocks in our data set. Thus, we came to the conclusion that in the 2011 calendar year approximately 48.67% of stocks paid dividends. Next, we had to add a dividend yield column to our data set. In order to calculate the dividend yield, we added a variable called *divYield* which was *DVPSP_C* (dividends per share) divided by *PRCC_C* (price per share). The resulting variable *divYield* has values that are proportions (i.e. .10 = 10%). Since we only wanted to analyze stocks that paid dividends, we then had to subset our original data set in order to removes the stocks for which *DVPSP_C* was 0. In this case, we felt that only displaying our dividend yield between 0 and .15 gave us the clearest picture of the distribution without sacrificing too much information. Looking at Figure 2, we can see that the distribution is a normal distribution that is skewed to the right.

Lastly, we decided to modify our previous plot (Figure 2) by including dividend yields for all stocks - thereby including stocks that didn't pay dividends in 2011 in our plot. As we can see from Figure 3, no amount of truncation will produce an aesthetic plot.

[1] Some observations contained NA values for some of the variables we are dealing with. For graphical and numerical summaries these observations were discarded.

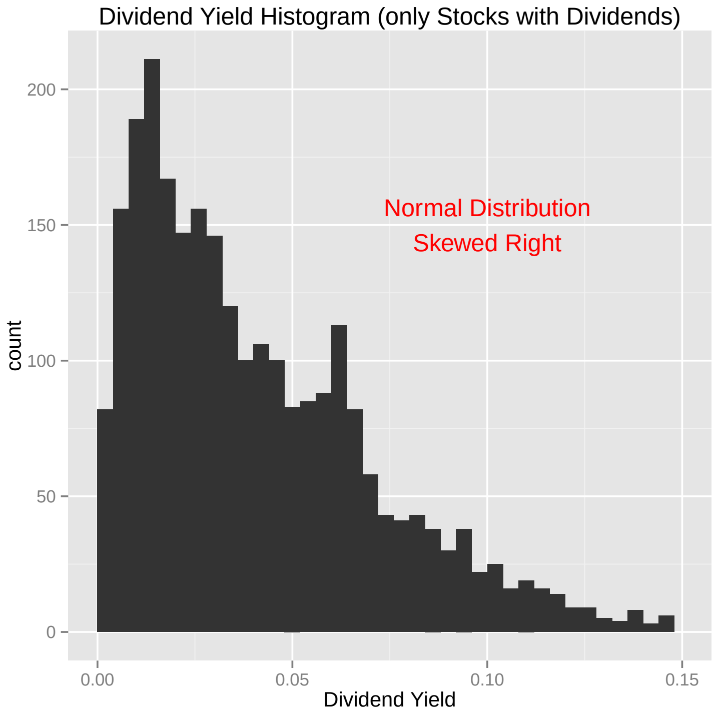

Figure 2: Above: Histogram of dividend yields for stocks that pay dividends.

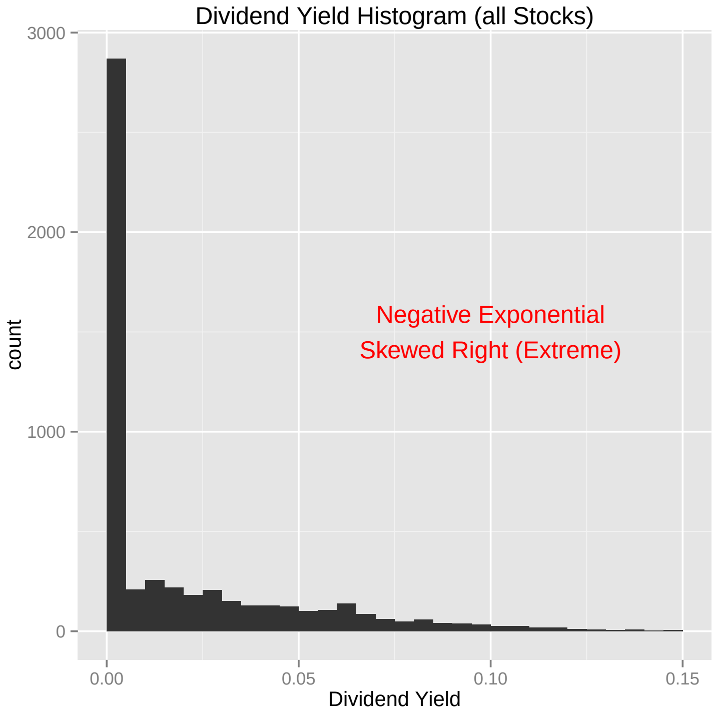

Figure 3: Above: Histogram of dividend yields for all stocks.

2 Numerical Summaries

2.1 Market Capitalization

Percentile	Market Cap
10%	32.36
20%	73.13
30%	144.07
40%	240.11
50%	414.98
60%	708.82
70%	1309.76
80%	2715.46
90%	7911.52

2.2 Dividend Yields

Percentile	Dividend Yield
10%	0.01
20%	0.01
30%	0.02
40%	0.03
50%	0.03
60%	0.04
70%	0.06
80%	0.07
90%	0.09

A Code

```
####################################################
### Mini Project 2 - Data Collection and Cleaning ##
##  Connor Barnhill, Brian Graff, Frank Portman   ##
####################################################

## Load Libraries
library(ggplot2)
library(plyr)
library(xtable)

## Load Data
crsp.data <- read.csv("data.csv", stringsAsFactors = F)

## Add Market Capitalization Column
## Call variable marketCap
crsp.data <- mutate(crsp.data, marketCap = CSHO * prcc_c)

## Plot Market Caps
## First truncate the huge companies off
## Cut off Market Caps above 2500 Billion
qplot(marketCap, binwidth = 50, data = crsp.data) + xlim(0, 2500) +
      xlab("Market Cap (Billions)") + ggtitle("Market Cap Histogram") +
      annotate("text", x = 1000, y = 600,
               label = "Negative Exponential\nSkewed Right",
               color = "red")

ggsave("marketcap.pdf", width = 6, height = 6)

# Exponential Distribution, slightly skewed right

## Calculate % of stocks of that pay dividends
stocksThatDo <- which(crsp.data$dvpsp_c != 0)

length(stocksThatDo) / length(crsp.data$dvpsp_c)
# [1] 0.4866617

## Add Dividend Yield Column
## Call variable divYield
crsp.data <- mutate(crsp.data, divYield = dvpsp_c / prcc_c)

## Make a subset of stocks that pay dividends
crsp.data.div <- subset(crsp.data, dvpsp_c != 0)

## Plot Dividend Yields
## First we look at only stocks that pay dividends
## Truncate to only 15% yields
qplot(divYield, binwidth = .004, data = crsp.data.div) + xlim(0, .15) +
      xlab("Dividend Yield") +
      ggtitle("Dividend Yield Histogram (only Stocks with Dividends)") +
      annotate("text", x = .10, y = 150,
               label = "Normal Distribution\nSkewed Right",
               col = "red")

ggsave("divyield.pdf", width = 6, height = 6)

## Now Plot Dividend Yields for all stocks
## See how the zeros affect the distribution
qplot(divYield, binwidth = .005, data = crsp.data) + xlim(0, .15) +
      xlab("Dividend Yield") +
      ggtitle("Dividend Yield Histogram (all Stocks)") +
      annotate("text", x = .10, y = 1500,
               label = "Negative Exponential\nSkewed Right (Extreme)",
               col = "red")

ggsave("divyieldall.pdf", width = 6, height = 6)


#########################################
####        Numerical Summaries      ####
#########################################

## Market Capitalization
mt <- quantile(crsp.data$marketCap,
               c(.1, .2, .3, .4, .5, .6, .7, .8, .9), na.rm = T)

mt <- as.table(mt)
xtable(mt)

## Dividend Yield
dt <- quantile(crsp.data.div$divYield,
               c(.1, .2, .3, .4, .5, .6, .7, .8, .9), na.rm = T)

dt <- as.table(dt)
xtable(dt)
```