A Probability and Statistics Primer for Quantitative Finance

Week 4: Exploratory Data Analysis II

Jake Price
Instructor, Computational Finance and Risk Management
University of Washington
Slides originally produced by Kjell Konis

Outline

Outline

Principal Characteristics of a Distribution

Goal: numerically summarize each of the following characteristics of a distribution

i) The *central tendency*, the middle (location, position, center) of the data set

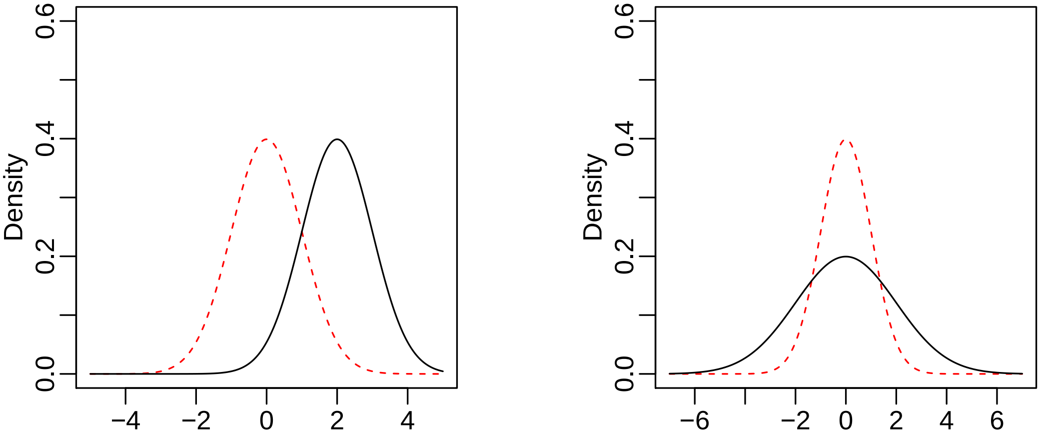
ii) The *dispersion*, the spread of the data around the center

iii) The *symmetry* (or lack or symmetry) with respect to the center

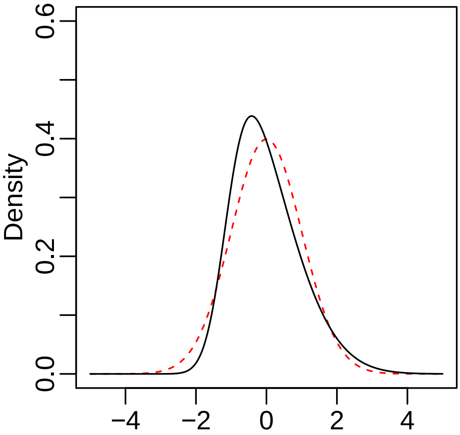
iv) The "flatness" (or *kurtosis*) of the distribution

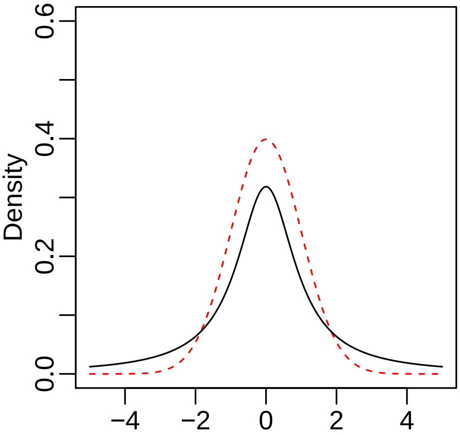
v) The number of *modes*

Principal Characteristics: Central Tendancy and Dispersion

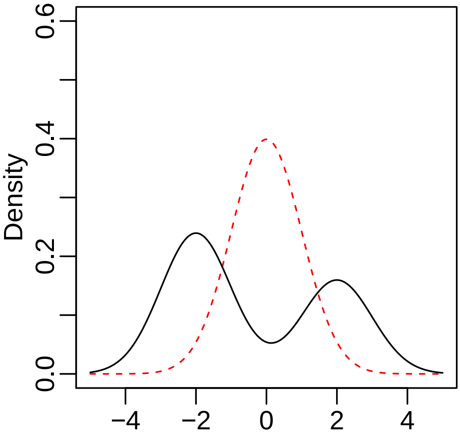


Principal Characteristics: Symmetry and Flatness





Principal Characteristics: Modality



Outline

Sample Moments and the Sample Mean

Sample Moments

The k^{th} *sample moment* of a data set $\mathbf{x} = \{x_1, \ldots, x_n\}$ is

$$\hat{\mu}'_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

Sample Mean

The *sample mean* of a data set $\mathbf{x}$ is its first sample moment

$$\text{mean}(\mathbf{x}) = \bar{x} = \hat{\mu}'_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Example: The mean 2012 Citigroup closing price is \$32.15

Sample Median

Conceptually, the *sample median* of a data set $\mathbf{x} = \{x_1, \ldots, x_n\}$ is a value m such that half of the data are greater than m and the other half less than m

When n is odd, the sample median is uniquely defined

$$\text{med}(\mathbf{x}) = \hat{q}_{0.5}(\mathbf{x}) = x_{((n+1)/2)}$$

When n is even, the sample median can be any point in the interval $(x_{(n/2)}, x_{(n/2+1)})$

For n even, define

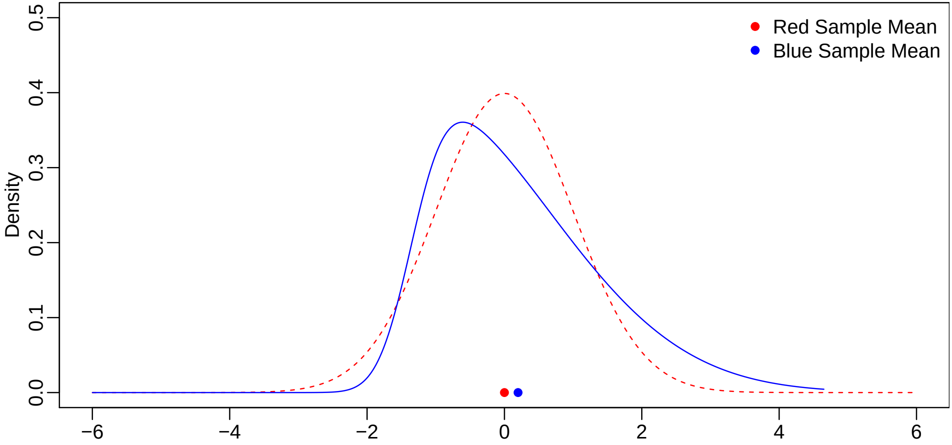
$$\text{med}(\mathbf{x}) = (x_{(n/2)} + x_{(n/2+1)})/2$$

Sample median: symmetric definition

$$\text{med}(\mathbf{x}) = \begin{cases} x_{((n+1)/2)} & n \text{ odd} \\ (x_{(n/2)} + x_{(n/2+1)})/2 & n \text{ even} \end{cases}$$

The Mean and the Median

Distribution symmetric $\implies$ sample mean $\approx$ sample median



Resistant Statistics

The mean is sensitive to extreme values

In contrast, median is *resistant*

$$\mathbf{x} = \{1, 2, 3\} \;\Rightarrow \begin{cases} \bar{x} = 2 \\ \text{med}(x) = 2 \end{cases}$$

$$\mathbf{x} = \{1, 2, 33\} \Rightarrow \begin{cases} \bar{x} = 12 \\ \text{med}(x) = 2 \end{cases}$$

Outline

Sample Central Moments and Sample Variance

Sample Central Moments

The k^{th} *sample central moment* of a data set $\mathbf{x} = \{x_1, \ldots, x_n\}$ is

$$\hat{\mu}_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^k$$

Sample Variance

The *sample variance* of a data set $\mathbf{x}$ is its second sample central moment

$$s^2 = \mathrm{var}(\mathbf{x}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The *sample standard deviation* is the square root of the sample variance

$$s = \mathrm{sd}(\mathbf{x}) = \sqrt{\mathrm{var}(\mathbf{x})} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Sample Range and Sample Innerquartile Range

Sample Range

The *sample range* of a data set $\mathbf{x} = \{x_1, \ldots, x_n\}$ is

$$\text{range}(\mathbf{x}) = \max(\mathbf{x}) - \min(\mathbf{x}) = x_{(n)} - x_{(1)}$$

Sample Innerquartile Range

The *sample innerquartile range* of a data set $\mathbf{x}$ is

$$\text{iqr}(\mathbf{x}) = \hat{q}_{0.75}(\mathbf{x}) - \hat{q}_{0.25}(\mathbf{x})$$

Comments

The range is sensitive to atypical values

Similar to the median, the innerquartile range is resistant to atypical values

Sample Skewness and Sample Kurtosis

Sample Skewness

The *sample skewness* of a data set $\mathbf{x} = \{x_1, \ldots, x_n\}$ is

$$\text{skewness}(\mathbf{x}) = \frac{\hat{\mu}_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}}$$

Sample Kurtosis

The *sample kurtosis* of a data set $\mathbf{x}$ is

$$\text{kurtosis}(\mathbf{x}) = \frac{\hat{\mu}_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}}$$

The sample *excess* kurtosis $=$ kurtosis$(\mathbf{x}) - 3$

Outline

Multivariate Data

The "experiment" we are observing is one day of the market

So far, only one measurement per experiment: Citigroup closing price

Multivariate $\implies$ more than one measurement per experiment

Example data: Citigroup closing prices during 2012

```
> library(quantmod)
> getSymbols("C", from = "2012-01-01", to = "2012-12-31")
> citi <- as.numeric(Cl(C))
```

Lets also look at the closing price of AIG during 2012

```
> getSymbols("AIG", from = "2012-01-01", to = "2012-12-31")
> aig <- as.numeric(Cl(AIG))
```

Univariate Analysis of Citigroup and AIG

Five number summeries

```
> summary(citi)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.82   28.56   32.86   32.15   35.08   40.17
```

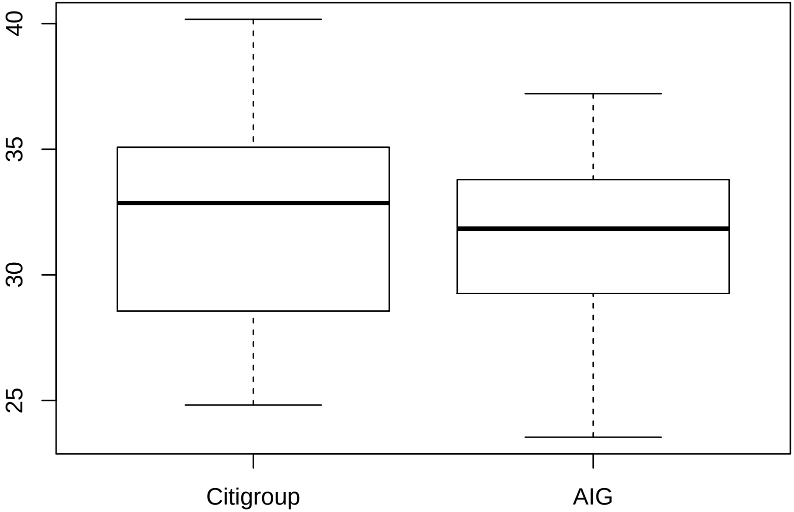
```
> summary(aig)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.54   29.26   31.84   31.31   33.79   37.21
```

Side-by-side boxplots of both variables

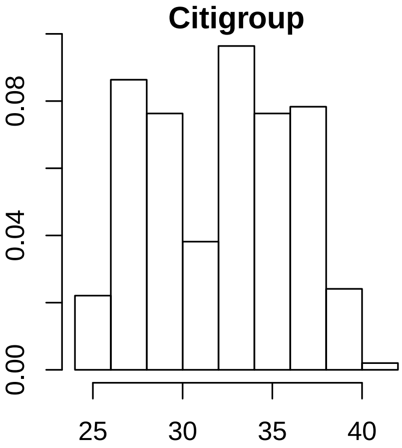
```
> boxplot(cbind(Citigroup = citi, AIG = aig))
```

Box Plots

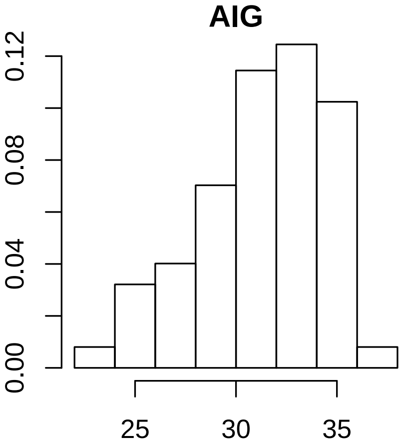


Histograms

```
> hist(citi, freq = FALSE,
+       main = "Citigroup")
```

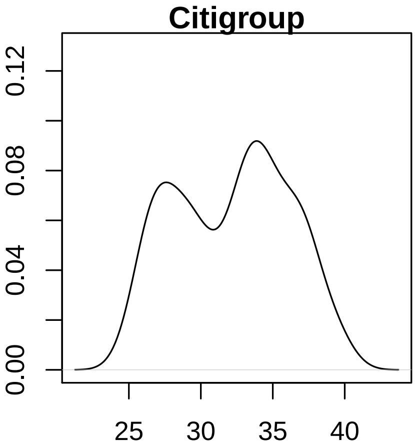



```
> hist(aig, freq = FALSE,
+       main = "AIG")
```

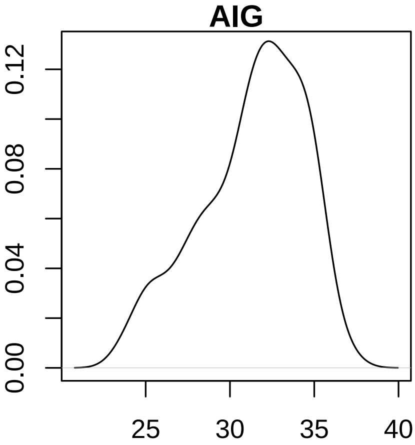


Kernel Density Estimates

```
> plot(density(citi),
+        main = "Citigroup")
```




```
> plot(density(aig),
+        main = "AIG")
```



Scatter Plots

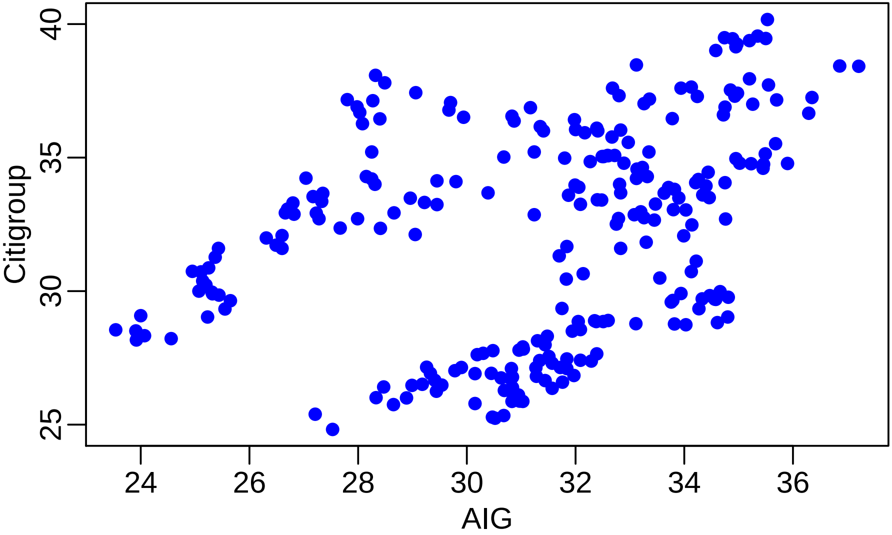
Scatter Plot: for each individual in the sample, plot one variable on the x-axis and one on the y-axis

- works very well in 2 dimensions
- works in 3 dimensions (see the `plot3d` function in the `rgl` package)
- can use a scatter plot matrix in higher dimensions

Example: plot 2012 AIG closing prices on the x-axis and 2012 Citigroup closing prices on the y-axis

```
> plot(aig, citi, xlab = "AIG", ylab = "Citigroup",
+       pch = 16, col = "blue")
```

Scatter Plots



Outline

Numerical Summaries of Multivariate Data

Consider a multivariate data set $(\mathbf{x}, \mathbf{y})$ where

$$\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$$
$$\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$$

The vector of univariate locations completely describes the multivariate location

$$\text{location}(\mathbf{x}, \mathbf{y}) = (\text{location}(\mathbf{x}), \text{location}(\mathbf{y}))$$

The vector of univariate dispersions

$$\text{dispersion}(\mathbf{x}, \mathbf{y}) \overset{☹}{=} (\text{dispersion}(\mathbf{x}), \text{dispersion}(\mathbf{y}))$$

does not tell the whole story

What's missing is a measure of the dependence of the covariation between $\mathbf{x}$ and $\mathbf{y}$

Notation: s_x is the sample standard deviation of the values in $\mathbf{x}$

Sample Covariance and Sample Correlation

Sample Covariance

The *sample covariance* measures the strength of the linear dependence between **x** and **y**

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})$$

Sample Correlation

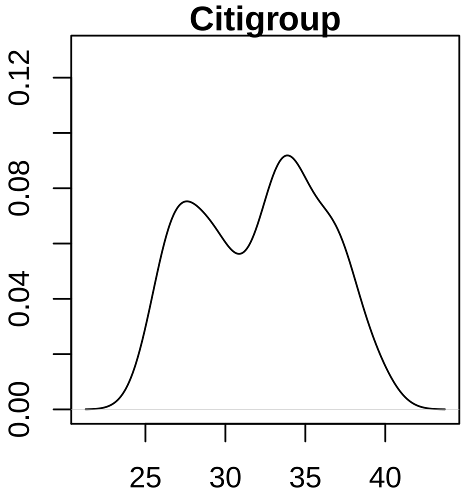
The *sample correlation* is a nondimensionalized measure of the strength of the linear dependence between **x** and **y**

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{s_x\, s_y}$$

Properties: $-1 \leq \text{corr}(\mathbf{x}, \mathbf{y}) \leq 1 \qquad \text{corr}(\mathbf{x}, \mathbf{x}) = 1 \qquad \text{corr}(-\mathbf{x}, \mathbf{x}) = -1$

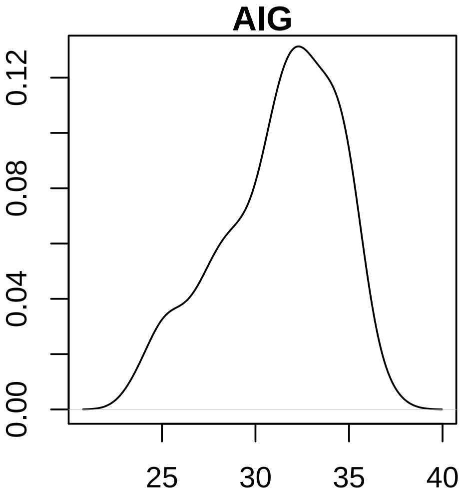
Outline

The Gaussian (Normal) Reference



Mean	32.18
Median	32.86
Standard Deviation	4.00
Innerquartile Range	6.58
Skewness	-0.04
Kurtosis	1.85

The Gaussian (Normal) Reference



Mean	31.32
Median	31.84
Standard Deviation	3.08
Innerquartile Range	4.54
Skewness	-0.58
Kurtosis	2.56

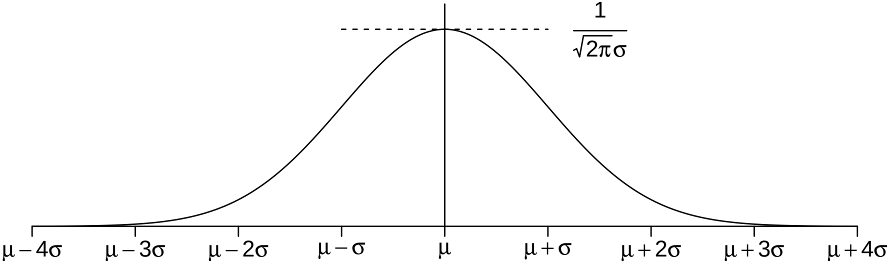
The Gaussian (Normal) Reference

The *Gaussian reference* is the well-known bell-shaped curve

$$g(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma^2 > 0$$

The location parameter μ is called the mean

The dispersion parameter σ^2 is called the variance



Simulated Data

The function `rnorm(n, mean = mu, sd = sigma)` simulates a data set such that

- The sample mean is "*close*" to `mu`
- The sample standard deviation is "*close*" to `sigma`
- The density is "*close*" to $g(x; \mathtt{mu}, \mathtt{sigma}^2) = \frac{1}{\sqrt{2\pi}\,\mathtt{sigma}} e^{-\frac{(x-\mathtt{mu})^2}{2\,\mathtt{sigma}^2}}$

Simulate a data set of 100 values with mean 3 and variance 2^2

```
> x <- rnorm(100, mean = 3, sd = 2)
```

Sample statistics:

```
> (mu.hat <- mean(x))
[1] 3.217775
> (sigma.hat <- sd(x))
[1] 1.796399
```

Simulated Data

Visually check that the estimated density is "*close*" to the Gaussian reference

Plot a kernel density estimate of the data

```
> plot(density(x), main = "", xlim = c(-4, 10))
```

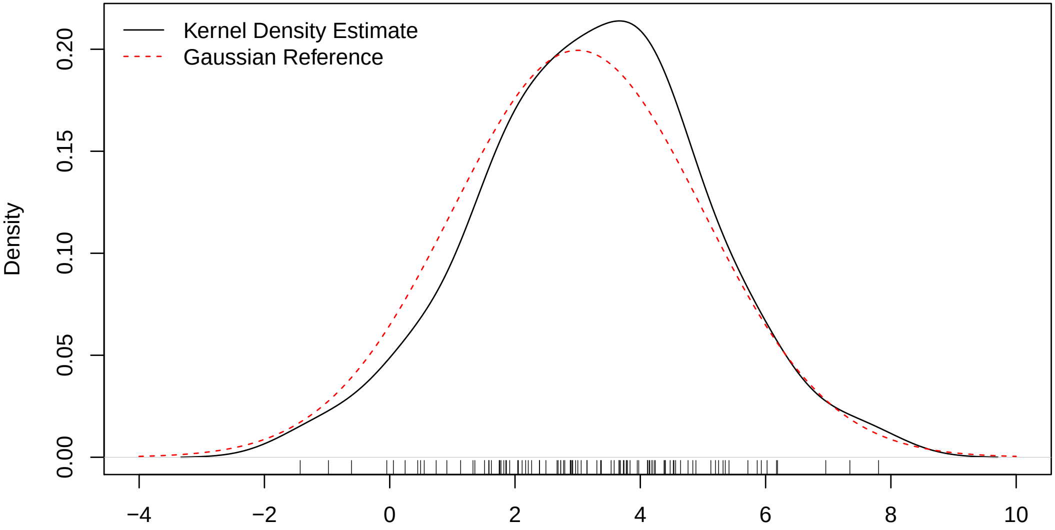
Use the `lines` function to draw the Gaussian reference on the plot

```
> u <- seq(-4, 10, 0.01)
> lines(u, dnorm(u, mean = 3, sd = 2), lty = 2, col = "red")
```

Use the `rug` function to plot the data on the x-axis

```
> rug(x)
```

The Gaussian Reference



The 68–95–99 Rule

$$\text{Gaussian Reference} \Rightarrow \begin{cases} \quad 68\% & \text{of the data in } [\mu \pm \sigma] \\ \quad 95\% & \text{of the data in } [\mu \pm 2\sigma] \\ > 99\% & \text{of the data in } [\mu \pm 3\sigma] \end{cases}$$

Let's see how we did

```
> sum( (x > 3 - 2) & (x < 3 + 2) )
[1] 74
> sum( (x > 3 - 2*2) & (x < 3 + 2*2) )
[1] 97
> sum( (x > 3 - 3*2) & (x < 3 + 3*2) )
[1] 100
```

Skewness and Kurtosis

The `PerformanceAnalytics` package provides functions for skewness and kurtosis

```
> library(PerformanceAnalytics)
```

The `skewness` function computes $\hat{\mu}_3/s^3$

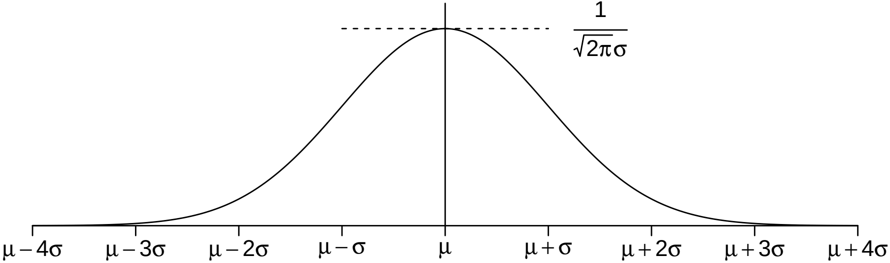
```
> skewness(x)
[1] -0.0722319
```

The `kurtosis` function returns excess kurtosis, use `method = "moment"` for $\hat{\mu}_4/s^4$

```
> kurtosis(x, method = "moment")
[1] 3.007653
```

Gaussian Reference Summary

$$g(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



mean	μ	skewness	0
variance	σ^2	kurtosis	3

Standardization

The data set $\mathbf{x}$ contains simulated Gaussian values using mean 3 and variance 2^2

The transformed values

$$\mathbf{z} = \frac{\mathbf{x} - \mu}{\sigma}$$

then appear Gaussian with mean 0 and variance 1^2

The *standard* Gaussian reference

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Standardization

Create a new vector contining the standardized values of **x**

```
> z <- (x - 3) / 2
```

The mean of `z` should be about 0 and the standard deviation should be about 1

```
> mean(z)
```

```
[1] 0.1088874
```

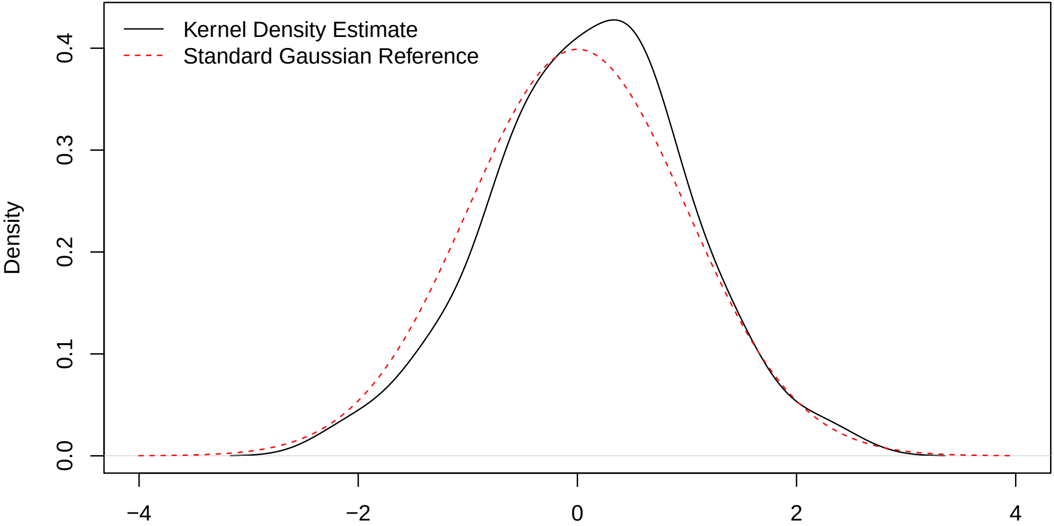
```
> sd(z)
```

```
[1] 0.8981994
```

Compare a kernel density estimate of `z` with the standard Gaussian reference

```
> plot(density(z), main = "", xlim = c(-4, 4))
> u <- seq(-4, 4, 0.01)
> lines(u, dnorm(u), lty = 2, col = "red")
```

The Standard Gaussian Reference



Outline

Quantile-Quantile Plots

A *quantile-quantile plot* compares the distributions of two data sets **x** and **y**

A quantile-quantile (qq) plot is a scatter plot of the sample quantiles of **y** vs. the sample quantiles of **x**

A linear QQ plot is evidence that the distributions are fundamentally the same

- The location and slope of the linear relationship reveal location and relative dispersion

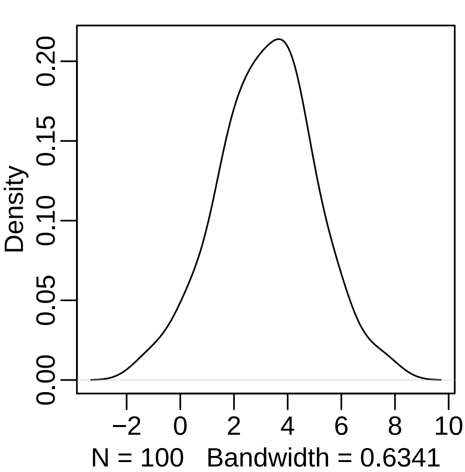
A nonlinear QQ plot is evidence that the distributions of the two data sets are fundamentally different

Example: consider a second data set simulated from the Gaussian reference

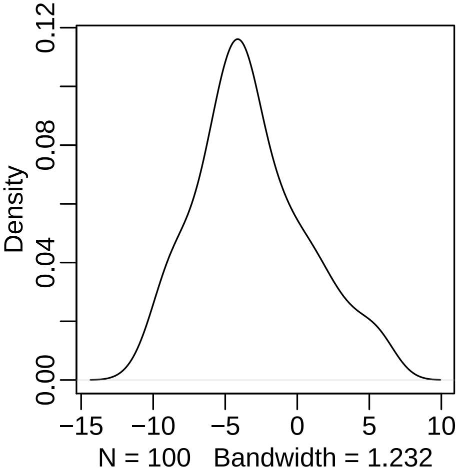
```
> y <- rnorm(100, mean = -3, sd = 4)
```

Quantile-Quantile Plots

```
> plot(density(x), main = "")
```

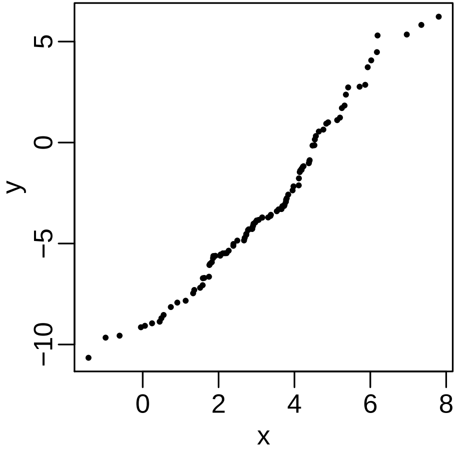



```
> plot(density(y), main = "")
```



Quantile-Quantile Plots

```
> qqplot(x, y, pch = 16, cex = 0.5)
```

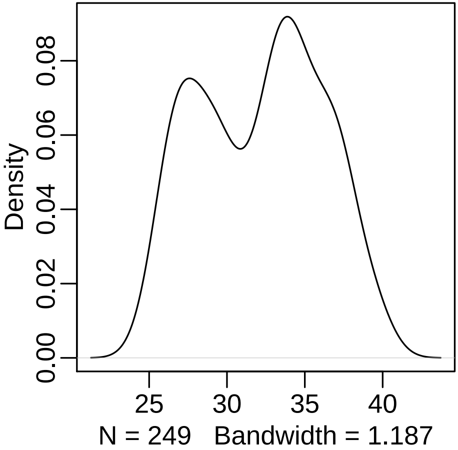


Remark: not straight but straight enough

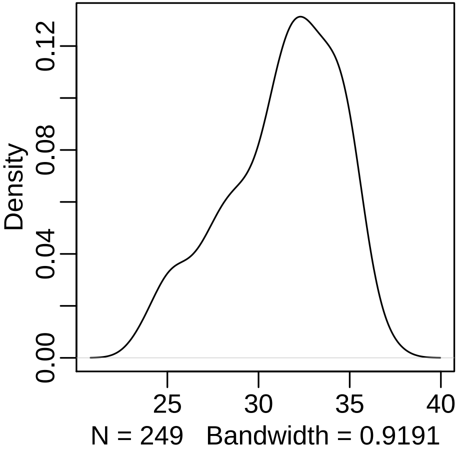
- Location: $(3, -3)$
- Slope: 2

Example: Citigroup vs. AIG

```
> plot(density(citi), main = "")
```

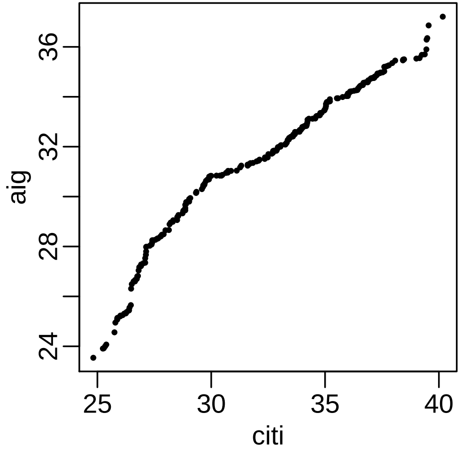



```
> plot(density(aig), main = "")
```



Example: Citigroup vs. AIG

```
> qqplot(citi, aig, pch = 16, cex = 0.5)
```



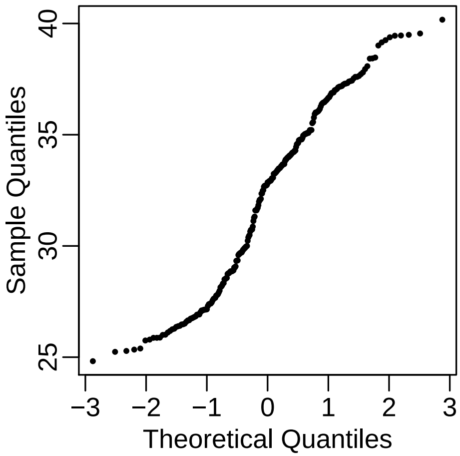
Remark: nonlinear relationship evidence that distributions fundamentally different?

Normal (Gaussian) QQ Plots

A *Normal* (Gaussian) *QQ Plot* plots the sample quantiles of a data set vs. the theoretical quantiles of the Gaussian reference

Interpretation: the more pronounced the nonlinear trend, the further from the Gaussian reference

```
> qqnorm(citi, pch = 16, cex = 0.5)
```



Outline

Transforming Data for Normality

Data analysis *"easier"* when the distribution of the data set is close to the Gaussian reference

- The center is a typical value
- Standard deviation is conceptually a good measure of dispersion
- etc. . .

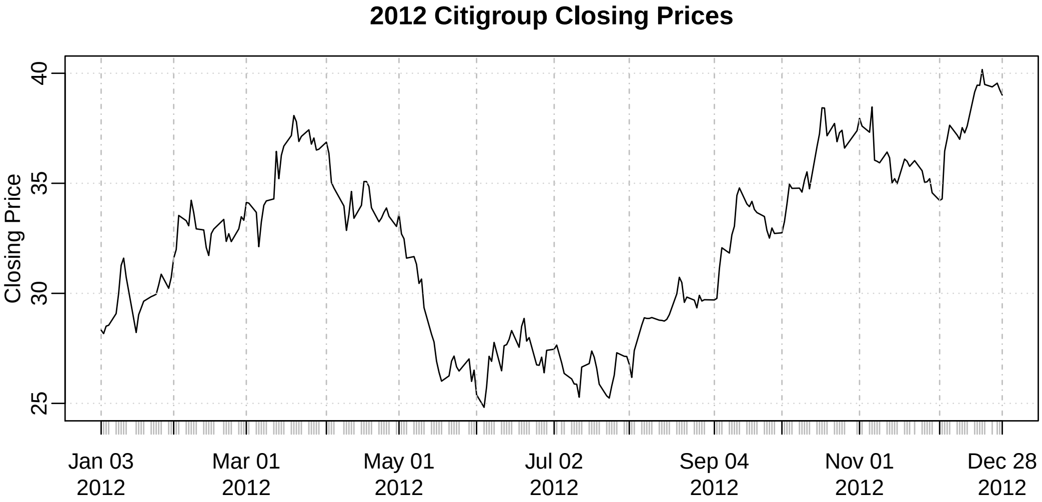
Prices are *raw data*; want to find a mathemtical function of the raw data that is closer to the Gaussian reference

Good starting point: the Box-Cox family of transformations

$$I(x) \qquad \sqrt{x} \qquad \log(x) \qquad \frac{1}{\sqrt{x}} \qquad \frac{1}{x}$$

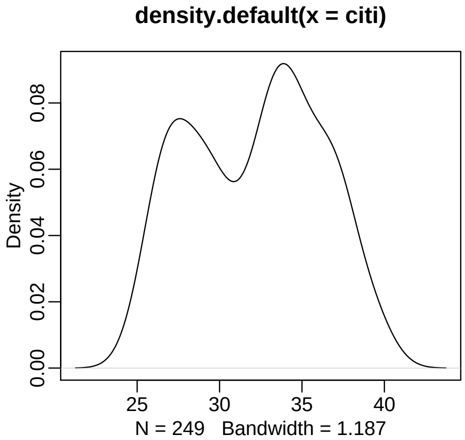
Transforming Data for Normality

```
> plot(Cl(C), ylab = "Closing Price",
+       main = "2012 Citigroup Closing Prices")
```

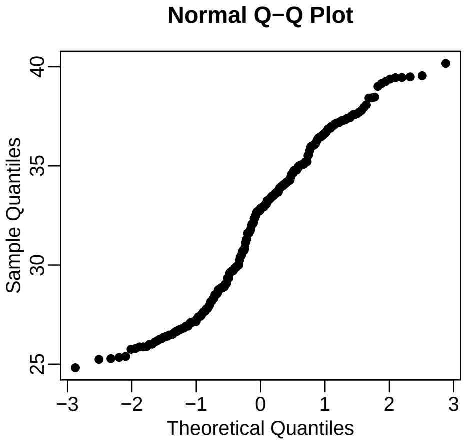


Transforming Data for Normality

```
> plot(density(citi))
```

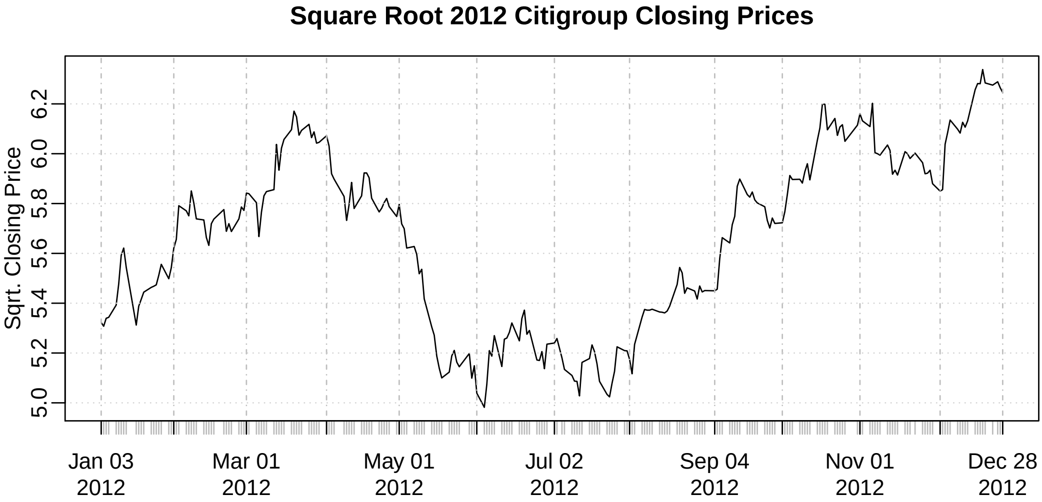



```
> qqnorm(citi, pch = 16)
```



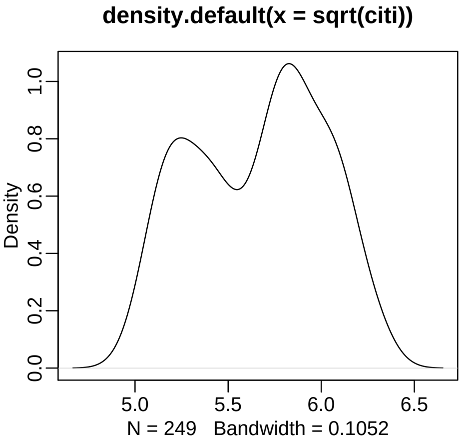
Transforming Data for Normality: Square Root

```
> plot(sqrt(Cl(C)), ylab = "Sqrt. Closing Price",
+       main = "Square Root 2012 Citigroup Closing Prices")
```

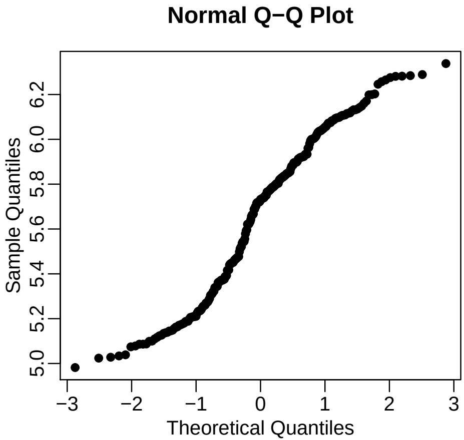


Transforming Data for Normality: Square Root

```
> plot(density(sqrt(citi)))
```

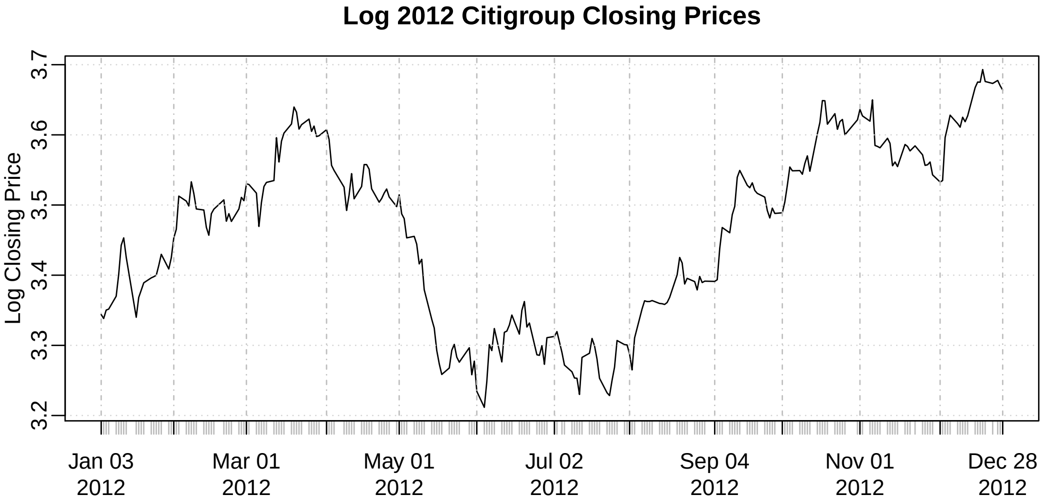



```
> qqnorm(sqrt(citi), pch = 16)
```



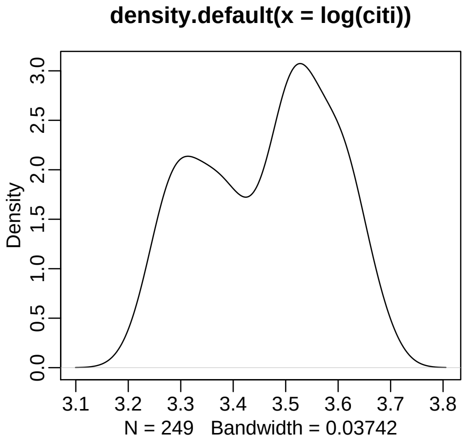
Transforming Data for Normality: Log

```
> plot(log(Cl(C)), ylab = "Log Closing Price",
+      main = "Log 2012 Citigroup Closing Prices")
```

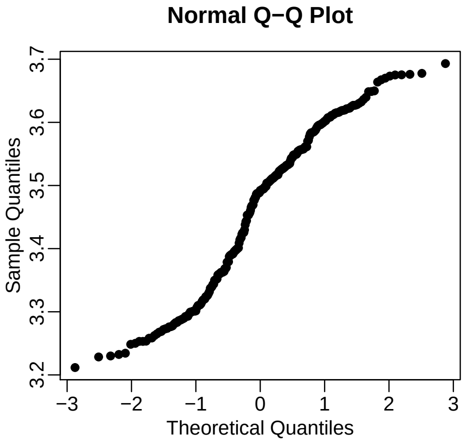


Transforming Data for Normality: Log

```
> plot(density(log(citi)))
```

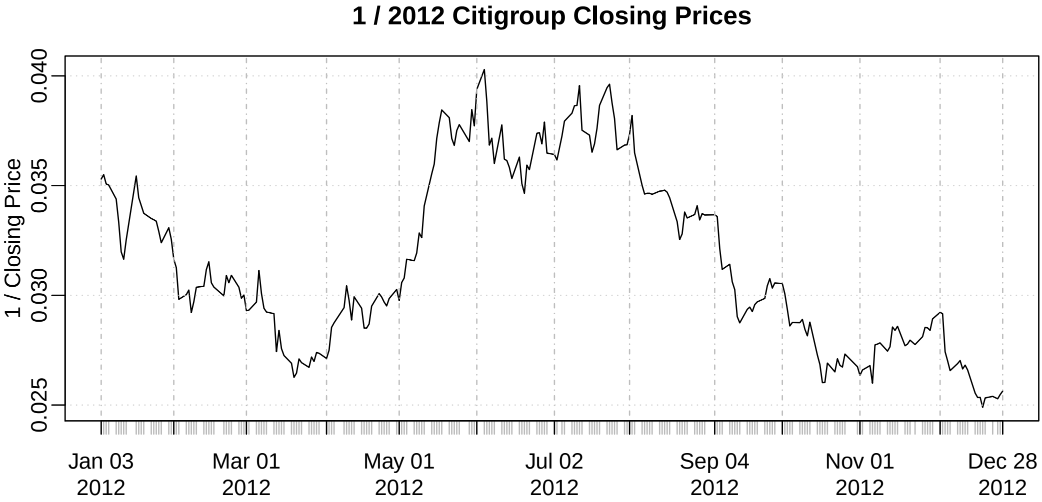



```
> qqnorm(log(citi), pch = 16)
```



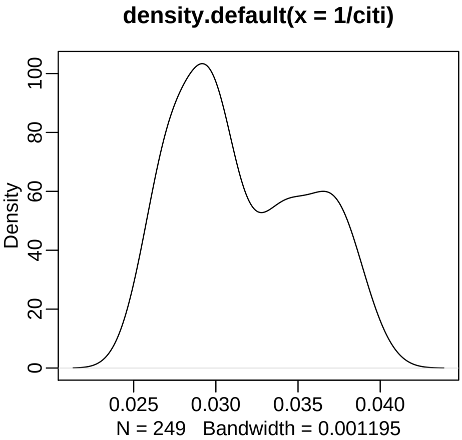
Transforming Data for Normality: Reciprocal

```
> plot(1/Cl(C), ylab = "1 / Closing Price",
+      main = "1 / 2012 Citigroup Closing Prices")
```

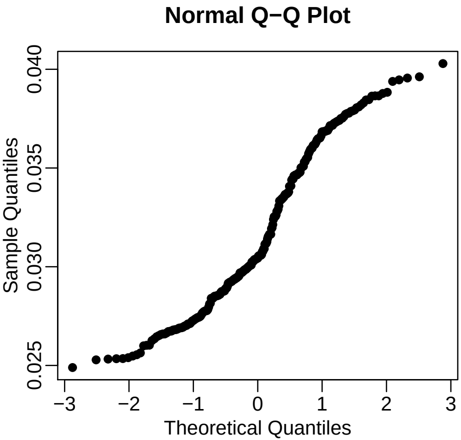


Transforming Data for Normality: Reciprocal

```
> plot(density(1/citi))
```

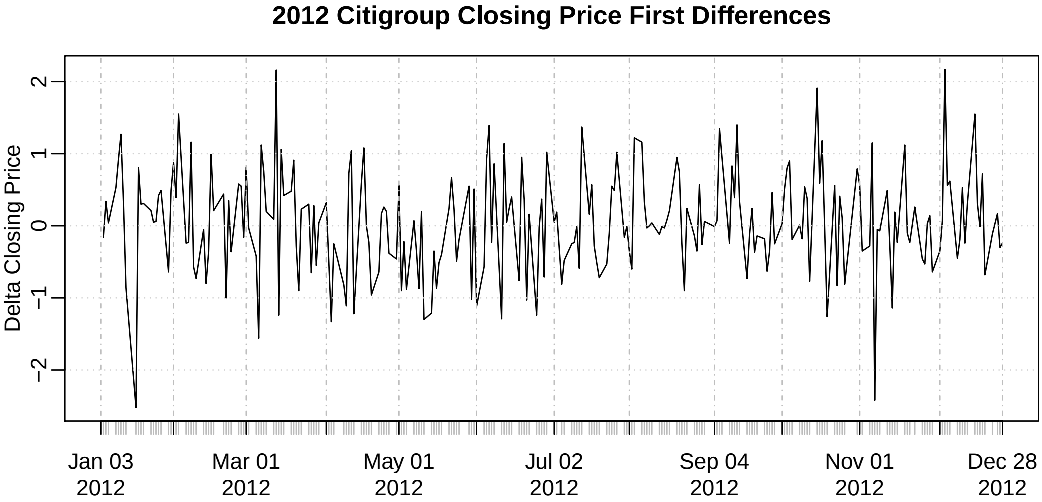



```
> qqnorm(1/citi, pch = 16)
```



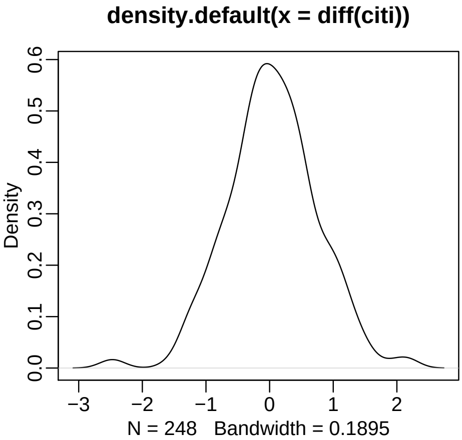
Transforming Data for Normality: First Difference

```
> plot(diff(Cl(C)), ylab = "Delta Closing Price",
+       main = "2012 Citigroup Closing Price First Differences")
```

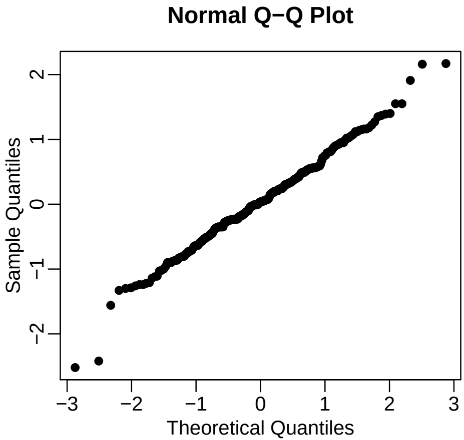


Transforming Data for Normality: First Difference

```
> plot(density(diff(citi)))
```




```
> qqnorm(diff(citi), pch = 16)
```



Outline

Returns

Use knowledge about the data to help choose a transformation

The goal of investing is to make a profit

Revenue from an investment depends on

- the change in prices of the assets held
- the amount of each asset held

Returns measure revenue relative to size of initial investment

Equivalent to measuring change in price relative to initial price

Net Returns

Let P_t be the price of an asset at time t (no dividends)

The *net return* for holding the asset from time $t-1$ to time t is

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Returns are scale-free

Returns are not unit free, their unit is time, e.g.,

$$R_t = 0.075 = 7.5\% \text{ per year}$$

for t measured in years

The *gross return* during the most recent k years is

$$\begin{aligned} 1 + R_t(k) &= \frac{P_t}{P_{t-k}} = \left(\frac{P_t}{P_{t-1}}\right)\left(\frac{P_{t-1}}{P_{t-2}}\right)\cdots\left(\frac{P_{t-k+1}}{P_{t-k}}\right) \\ &= (1+R_t)(1+R_{t-1})\dots(1+R_{t-k+1}) \end{aligned}$$

Log Returns

Log returns or *continuously compounded returns* are given by

$$r_t = \log(1 + R_t) = \log\left(\frac{P_t}{P_{t-1}}\right) = p_t - p_{t-1}$$

$p_t = \log(P_t)$ is called the *log price*

$r_t = \log(1 + R_t) \approx R_t$ for $|R_t| < 0.1 = 10\%$

A k-period log return is the sum the k single period returns

$$\begin{aligned} r_t(k) &= \log[1 + R_t(k)] \\ &= \log\big[(1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1})\big] \\ &= \log\big[(1 + R_t)\big] + \log\big[(1 + R_{t-1})\big] + \cdots + \log\big[(1 + R_{t-k+1})\big] \\ &= r_t + r_{t-1} + \cdots + r_{t-k+1} \end{aligned}$$

Getting Returns in R

```
> library(quantmod)

> getSymbols("C", from = "2012-01-01", to = "2012-12-31")

[1] "C"

> returns <- dailyReturn(Cl(C))

> head(returns, 3)

           daily.returns
2012-01-03   0.000000000
2012-01-04  -0.005647723
2012-01-05   0.012069578
```

Outline

Financial Interpretation of Sample Statistics

The focus for the remainder of the course will be the analysis of returns data

The sample mean measures the average or expected return on an asset

The sample standard deviation measures the risk of investing in an asset

The sample standard deviation of returns often called *volatility*

Example: Suppose we invest in a savings account that has a net return $R_t = \mu$ for all periods t

- The average return on the investment is $\bar{R} = \mu$ per period
- Since $R_t - \bar{R} = \mu - \mu = 0$ for all t, the sample standard deviation is zero

A savings account is a *risk free* investment

Risk free investments must be handled separate from risky investments

Risky Assets

Buying a share of Citigroup stock is a risky investment

Compute the average daily return of Citigroup

```
> mean(citi.returns.2012.daily <- dailyReturn(Cl(C)))
[1] 0.001535964
```

Compute the volatility of Citigroup returns

```
> sd(citi.returns.2012.daily)
[1] 0.02242878
```

Question: When does it make sense to invest in a risky asset rather than a risk free asset?

Rewards Relative to Variability

The expected return and the risk are characteristics of each asset in the market

Question: How to compare two risky investments?

The *Sharpe ratio* expresses expected returns in units of risk

$$S = \frac{\bar{R} - R_f}{\mathrm{sd}(R)}$$

The Sharpe ratio is commonly used by investment managers as an indicator of performance

Questions to Consider

1. Suppose an asset has produced an average yearly return of 4% during the last 5 years. Will it out perform a risk free investment this year?

2. Suppose mutual fund A claims a Sharpe ratio of 1.5 and mutual fund B 2.2. Is the manager of fund B better than the manager of fund A?

3. Suppose we make a portfolio by investing $10,000 in two risky assets. How much money would you allocate to each asset so that the Sharpe ratio of the portfolio is maximum?

COMPUTATIONAL FINANCE & RISK MANAGEMENT
UNIVERSITY *of* WASHINGTON
Department of Applied Mathematics

http://computational-finance.uw.edu