

# Lecture 1

## Introduction and Review

CHEN Ying  
FE5209 Financial Econometrics



# Outline

- Introduction to financial econometrics
- Data features
- Review on probability and statistics
- Introduction to R

## Reading:

SDA chapter 4/5

FTS chapter 1

SFM chapter 3

<http://cran.r-project.org/doc/manuals/R-intro.pdf>

<https://cran.r-project.org/doc/contrib/usingR.pdf>



**\*\*[SDA] *Statistics and Data Analysis for Financial Engineering* (2010) by David Ruppert**

**[FTS] Tsay, R.S. (2010) *Analysis of Financial Time Series, Third Edition*, Wiley.**

**[SFM] Franke, J., Härdle, W. K., Hafner, C. M. (2015) *Statistics of Financial Markets An Introduction*. Springer.**

# Introduction

*What is ... financial econometrics?*

*Broadly speaking, it aims to study quantitative problems arising from finance. It uses statistical techniques and economic theory to address a variety of problems from finance. These include building financial models, estimation and inferences of financial models, volatility estimation, risk management, testing financial economics theory, capital asset pricing, derivative pricing, portfolio allocation, risk-adjusted returns, simulating financial systems, hedging strategies, among others (Fan, Jianqing).*

*It involves intersection of statistical techniques and finance. It seeks to test models of how financial markets operate and how financial prices are determined. Conversely, new techniques in analysing financial data can lead to empirical facts inconsistent with existing theories, begging for new models or investment strategies. One distinguishing aspect of finance is the importance of risk, both in our models and empirical implementation. (Jeffrey R. Russell)*



# Course description

This course will discuss econometrics methods and models that are appropriate for understanding, analysing and solving financial problems. Gain hands-on experience with financial data and learn about regression analysis, time series analysis, multivariate statistical methods and their applications to capital asset pricing, volatility analysis and forecast, risk management and trading strategy. Learn how to improve investment outcomes for yourself and/or clients.

Some related problems in finance:

- ☐ Are financial markets weak-form informationally efficient?
- ☐ Does CAPM represent superior model for the determination of returns on risky assets?
- ☐ How to model long-term relationships for asset returns?
- ☐ Explain the relationship among multiple financial instruments.
- ☐ Understand and measuring risks
- ☐ Determine the optimal hedge ratio in pairs trading.

*What is difference between general econometrics and financial econometrics?*

The nature of data in finance issues is very different.

# Data

Financial data are naturally time series.

Definition: A time series is a sequence of data points, measured typically at successive points in time spaced at uniform time intervals.

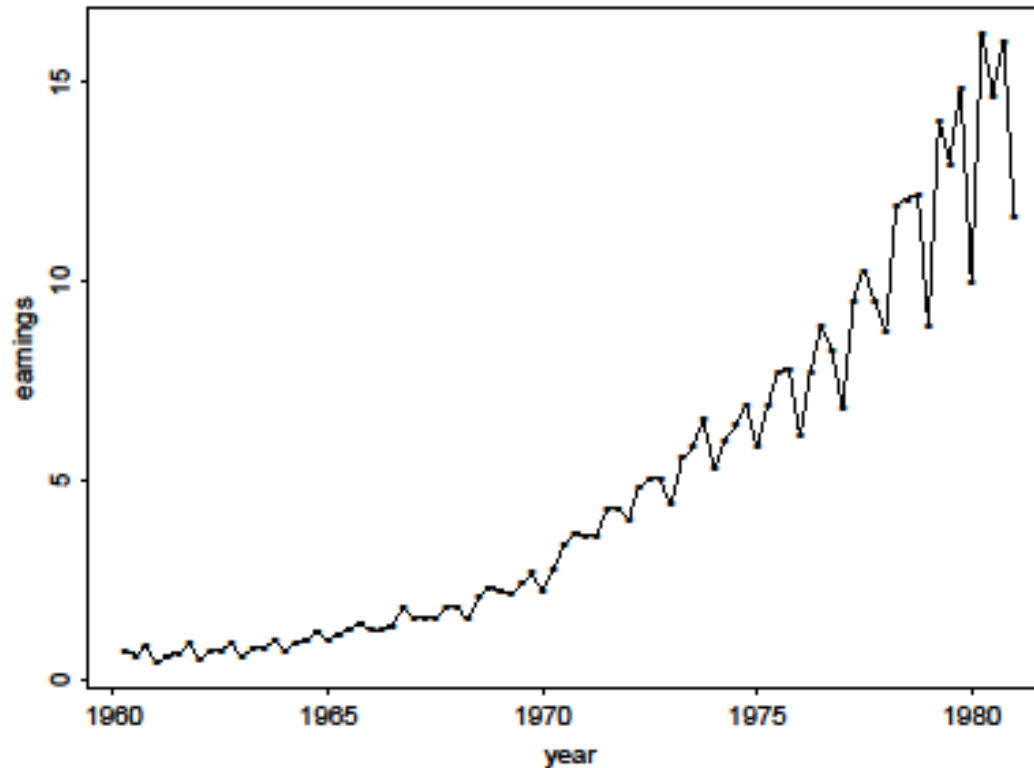
Examples of time series are the daily and the **annual** flow volume of the Nile River at Aswan.

Examples of time series in finance:

- ❑ *Quarterly* **earnings** of Johnson & Johnson.
- ❑ *Monthly* **interest rates** of Singapore.
- ❑ *Weekly* **exchange rate** between U.S. Dollar vs Singapore Dollar.
- ❑ *Daily* closing value of the Strait Times **Index** (STI).
- ❑ *Intra-daily* (tick by tick) transaction prices of **BA**.

# Quarterly earnings of Johnson & Johnson

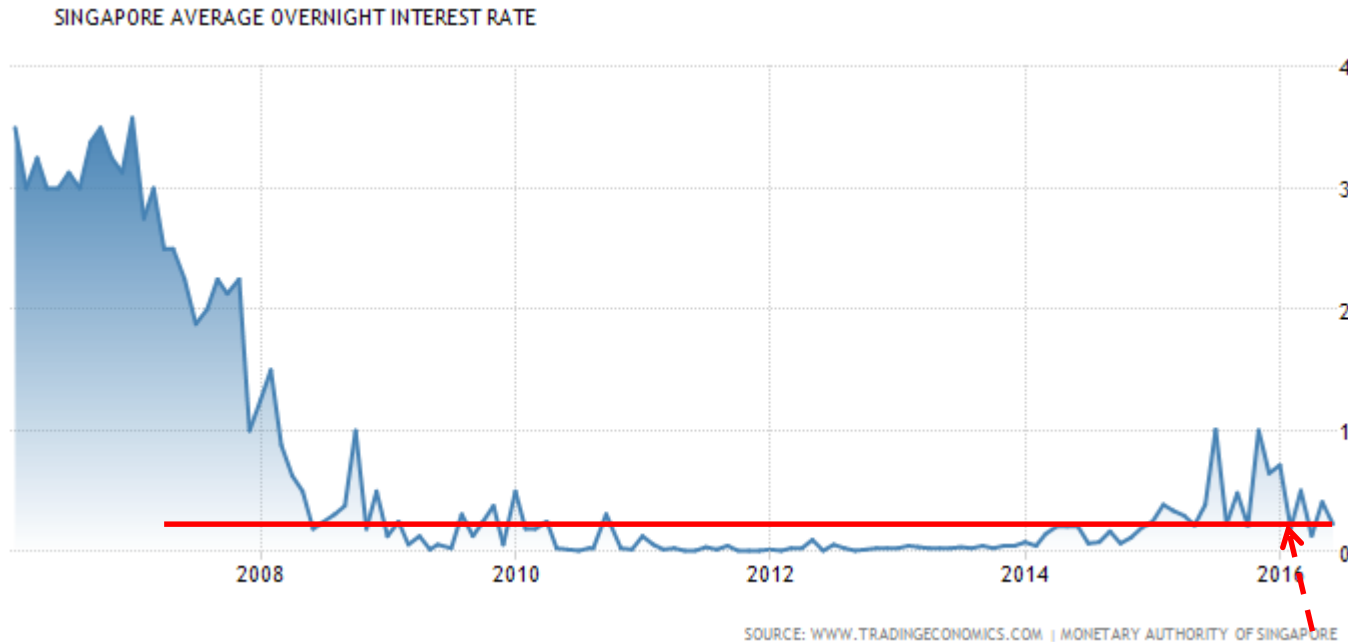
Quarterly earnings per share of Johnson & Johnson: 60-80



An exponentially increasing trend, accompanied by seasonal variations.

*Johnson & Johnson (JnJ) is an New Jersey-based multinational medical devices, pharmaceutical and consumer packaged goods manufacturer founded in 1886. Its common stock is a component of the Dow Jones Industrial Average and the company is listed among the Fortune 500.*

# Monthly interest rates of Singapore



There is no obvious pattern.

*The benchmark interest rate in Singapore was last recorded at 0.23% (June 2016). Interest Rate in Singapore is reported by the Monetary Authority of Singapore. Historically, from 1988 until 2013, Singapore Interest Rate averaged 1.69 Percent reaching an all time high of 20 Percent in January of 1990 and a record low of -0.75 Percent in October of 1993. SIBOR is a reference rate based on the interest rates at which banks offer to lend unsecured funds to each other in the Singapore interbank market.*

*U.S. interest rate (0.25-0.50%) reported by Federal reserve.*

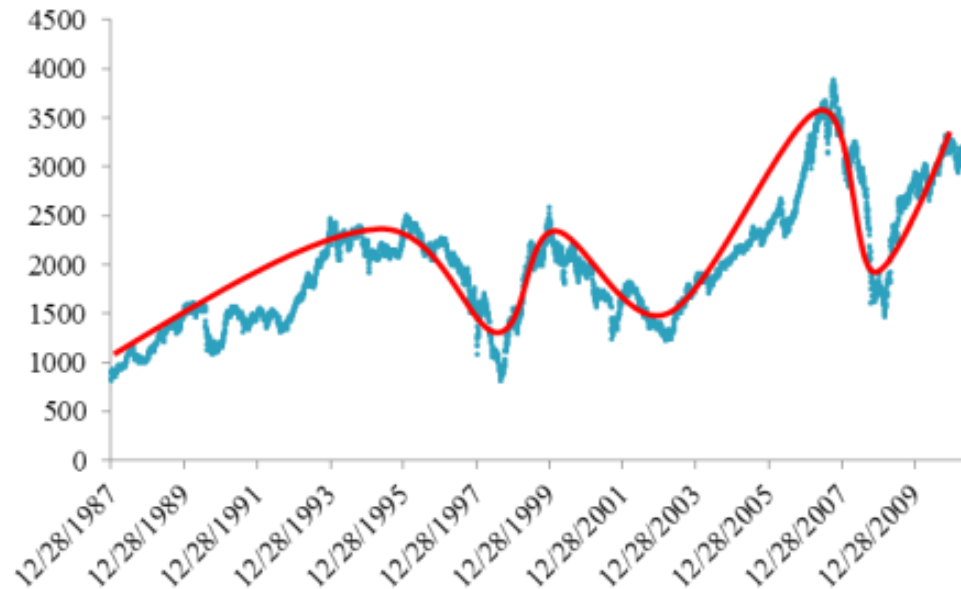
# Weekly exchange rate USD/SGD



An exponentially decreasing trend, after 2009.



# Daily closing value of the Strait Times Index



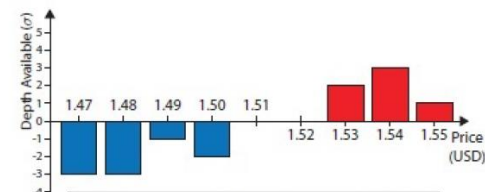
There is cyclical component.

*STI (Straits Times Index) prices exhibit a cyclical component along business cycle. The values dropped down around 1997, 2002 and 2008. The Asian financial crisis began in July 1997. During 2001–2003 there was economic recession in Singapore. During 2008 there was the global financial crisis.*

# Intra-daily (tick by tick) transaction prices of BA

symbol	date	time	price	size
BA	05DEC2005	9:31:10	69.4500	60700
BA	05DEC2005	9:31:11	69.4500	100
BA	05DEC2005	9:31:11	69.4500	200
.....				
BA	05DEC2005	9:31:11	69.4500	2500
BA	05DEC2005	9:31:11	69.4500	100
BA	05DEC2005	9:31:11	69.4500	100
BA	05DEC2005	9:31:12	69.4500	1600
BA	05DEC2005	9:31:12	69.4500	1500
BA	05DEC2005	9:31:12	69.4500	1700
BA	05DEC2005	9:31:12	69.4500	100
BA	05DEC2005	9:31:13	69.4500	100
BA	05DEC2005	9:31:15	69.4500	100
BA	05DEC2005	9:31:18	69.4700	100
.....				
BA	05DEC2005	9:31:18	69.4500	100
BA	05DEC2005	9:31:19	69.4500	100
.....				
BA	05DEC2005	9:31:19	69.4500	100
BA	05DEC2005	9:31:27	69.4500	100

## Limit order book



Arriving order $x$	Values after arrival (USD)			
	$b(t_x)$	$a(t_x)$	$m(t_x)$	$s(t_x)$
Initial Values	1.50	1.53	1.515	0.03
(\$1.48, $-3\sigma, t_x$ )	1.50	1.53	1.515	0.03
(\$1.51, $-3\sigma, t_x$ )	1.51	1.53	1.52	0.02
(\$1.55, $-3\sigma, t_x$ )	1.50	1.54	1.52	0.04
(\$1.55, $-5\sigma, t_x$ )	1.50	1.55	1.525	0.05
(\$1.54, $4\sigma, t_x$ )	1.50	1.53	1.515	0.03
(\$1.52, $4\sigma, t_x$ )	1.50	1.52	1.51	0.02
(\$1.47, $4\sigma, t_x$ )	1.48	1.53	1.505	0.05
(\$1.50, $4\sigma, t_x$ )	1.49	1.50	1.495	0.01

\*image from (Gould et al., 2013)

# Nature of financial data

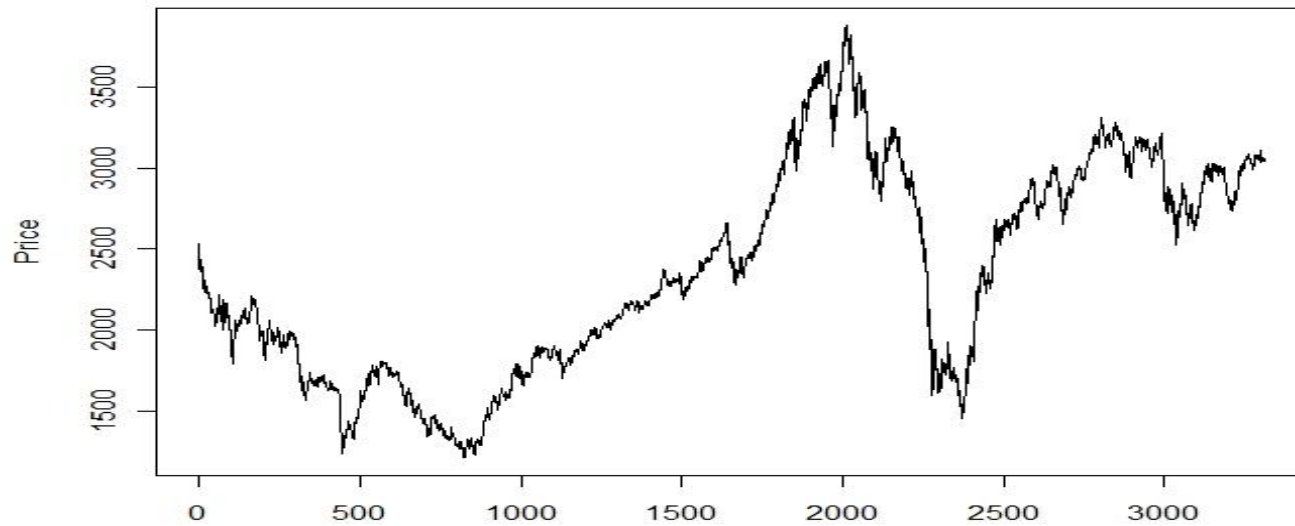
Financial data are naturally time series.

BUT financial data are observed at a much **higher frequency** than e.g. macro economic data or bio data.

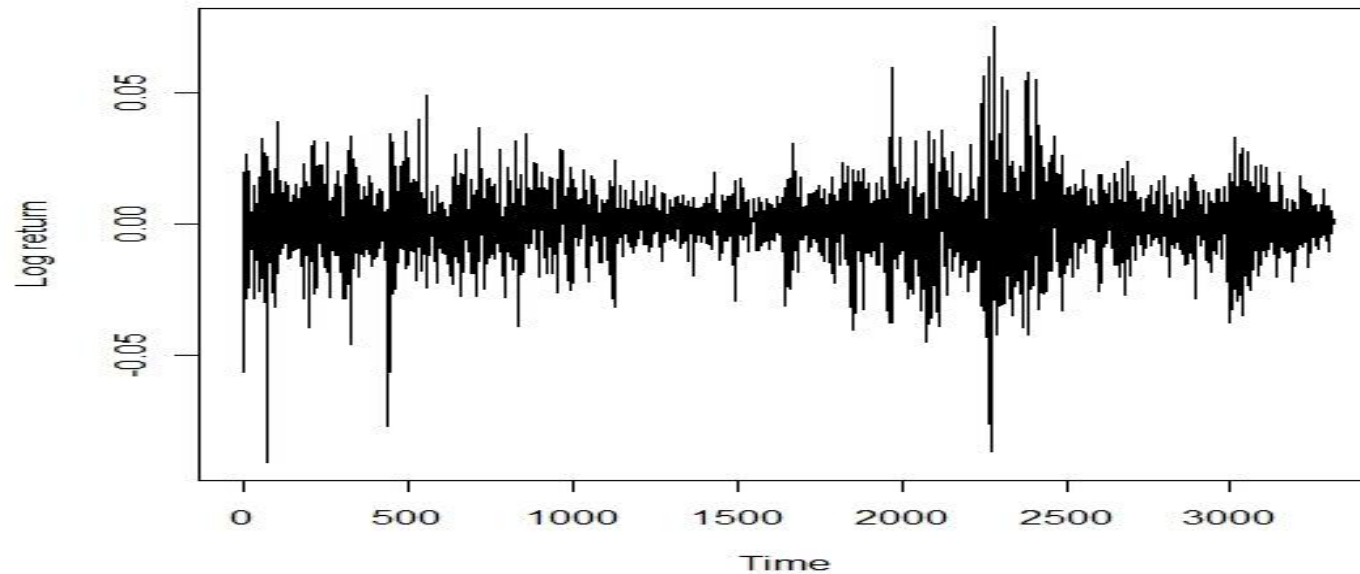
Also the **properties of financial series** differ. An important issue is whether the series has a unit root (thus its statistical property varies over time) and how to devise methods to estimate models when the variables are integrated of order one. In most cases, we observe prices most of the time, **yet deal with asset or portfolio returns**.

# Returns of STI

**Time Series of STI**



**Time Series of STI log return**



# Why returns?

Prices are generally found to be non-stationary (properties are changing over time).

Conventional statistics methods are appropriate for handling stationary data.

Returns are found to be stationary, at least relative to prices.

## Financial data analysis is easy?

The analysis of financial data brings its own challenges. As you will see, financial returns possess some common properties that need to be incorporated in econometric models.

- ❑ returns of assets such as stocks and bonds exhibit time-varying volatility.
- ❑ financial returns can exhibit asymmetry in volatility.
- ❑ financial data are not normally distributed.

We will learn

## Statistical/Econometric/IT techniques

that are appropriate for analysing, understanding and solving financial problems.



# What to do with financial data?

## Aggregation and Statistics

- ❑ Data warehouse and On-Line Analytical Processing (OLAP)

## Indexing, Searching, and Querying

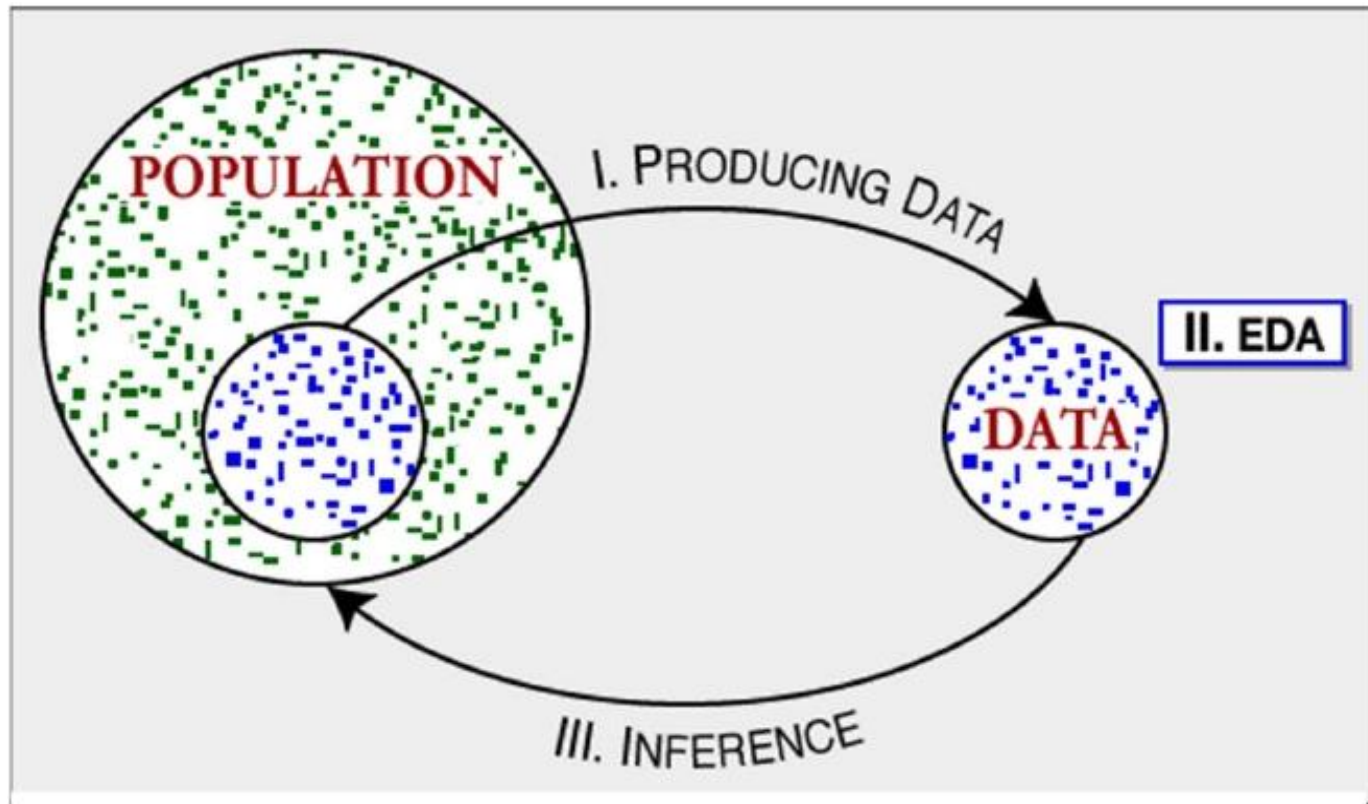
- ❑ Keyword based search
- ❑ Pattern matching

## Knowledge discovery via Statistical Modeling and Data Science:

- ❑ Discovery of useful, possibly unexpected, **patterns** in data
- ❑ Non-trivial extraction of implicit, previously unknown and potentially useful **information** from data
- ❑ Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns
- ❑ Formulation of relationships between variables in the form of **mathematical equations** to interpret the underlying mechanism and make forecast

# Sampling

Extract **information** from **data** to understand “real world”  
& enhance **decision-making**





# Literary Digest's survey in 1936

In the 1936 presidential election, Literary Digest mailed questionnaires to **10 million** people (25% of voters at the time).

Selected from telephone books, club memberships, mail order lists, automobile ownership lists

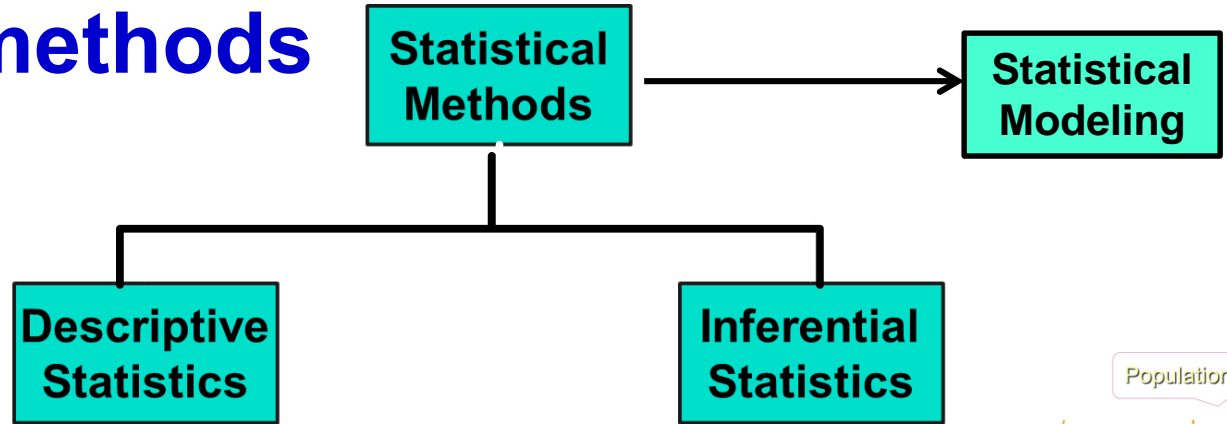
**2.4 million** people responded.

The *Literary Digest* predicted an overwhelming victory of Landon over Roosevelt: 57% to 43%.



Roosevelt won the election by a landslide – **62% to 38%**.

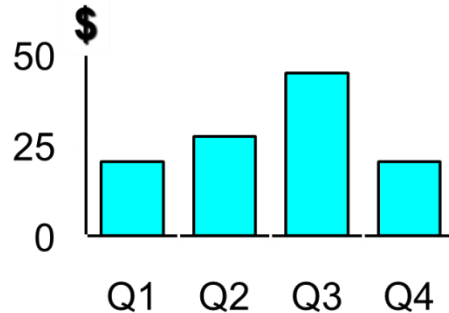
# Statistical methods



Descriptive statistics Involves

- Collecting Data
- Presenting Data
- Characterizing Data

**Purpose: Describe Data**



$$\bar{X} = 30.5 \quad S^2 = 113$$

Inferential statistics Involves

- Estimation
- Hypothesis
- Testing

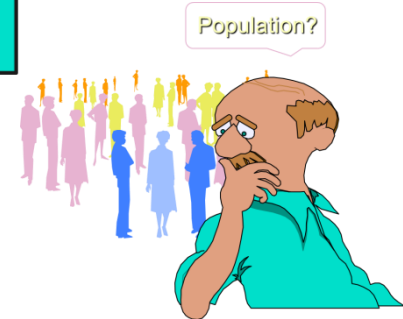
**Purpose: Make Decisions About Population Characteristics**

Population (Universe)

Sample: Portion of Population

Parameter: Summary Measure about Population

Statistic: Summary Measure about Sample

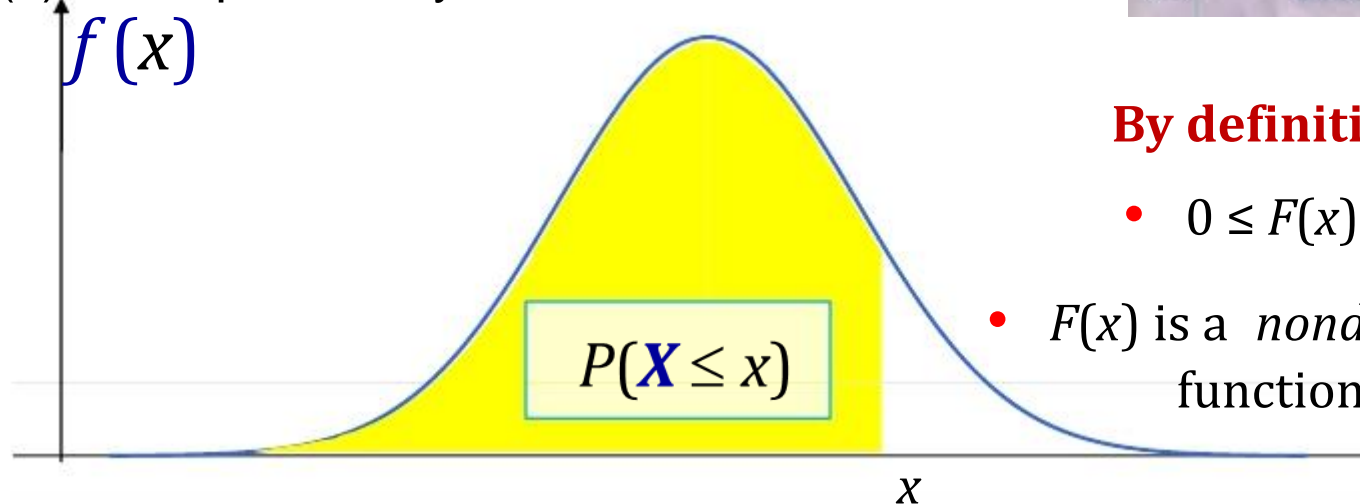


# Cumulative Distribution Function (CDF)

For a given  $x$ , the **cumulative distribution function** (**cdf**)  $F(x)$  of a continuous r.v.  $X$  is defined by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt.$$

$F(x)$  is the probability that  $X$  does not exceed  $x$ .



**By definition :**

- $0 \leq F(x) \leq 1$
- $F(x)$  is a *nondecreasing* function of  $x$

It is also easy to see that :

$$P(X > x) = 1 - F(x), P(c \leq X \leq d) = F(d) - F(c)$$

**Note:**  $P(X \leq x) = P(X < x)$

# Normal distributions & Standard Normal distribution

We can use the  $N(0,1)$  table to compute probabilities for any r.v. that obeys a Normal distribution :  $\mathbf{X} \sim N(\mu, \sigma^2)$  . **Why?**

Because there is a special relationship between  $N(0,1)$  and  $N(\mu, \sigma^2)$  .

If  $\mathbf{X} \sim N(\mu, \sigma^2)$ , then the r.v.  $\mathbf{Z}$  defined by  $N(0,1)$  obeys a standard Normal distribution.  $\mathbf{Z} = (\mathbf{X} - \mu)/\sigma$

$$\mathbf{X} \sim N(\mu, \sigma^2),$$

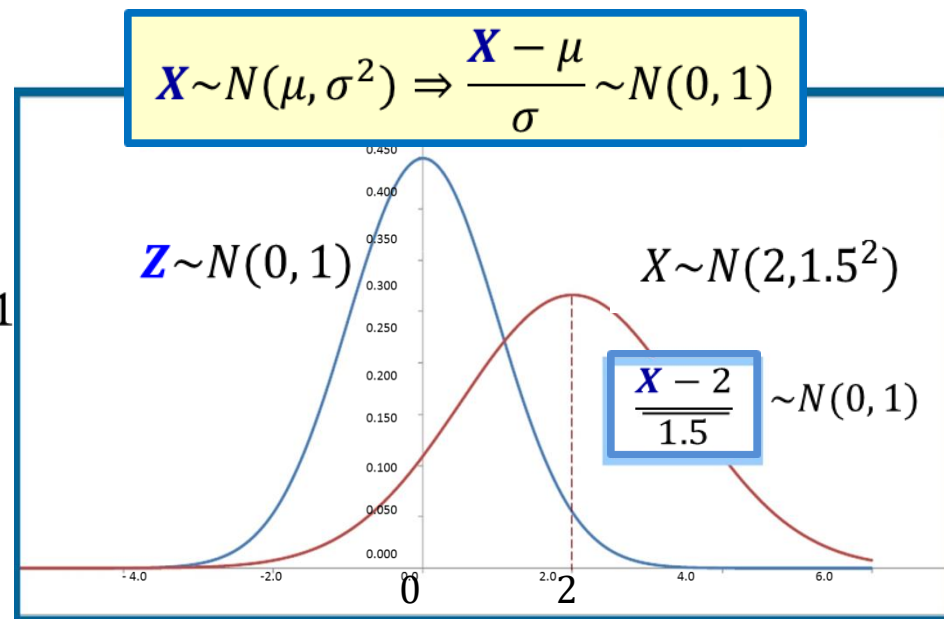
$$E(\mathbf{Z}) = E[(\mathbf{X} - \mu)/\sigma] = [E(\mathbf{X}) - \mu]/\sigma = 0$$

$$\star E(a\mathbf{X} + b) = aE(\mathbf{X}) + b$$

$$\text{var}(\mathbf{Z}) = \text{var}[(\mathbf{X} - \mu)/\sigma] = \text{var}(\mathbf{X})/\sigma^2 = 1$$

$$\star \text{var}(a\mathbf{X} + b) = a^2 \text{var}(\mathbf{X})$$

**Therefore  $\mathbf{Z} \sim N(0,1)$**



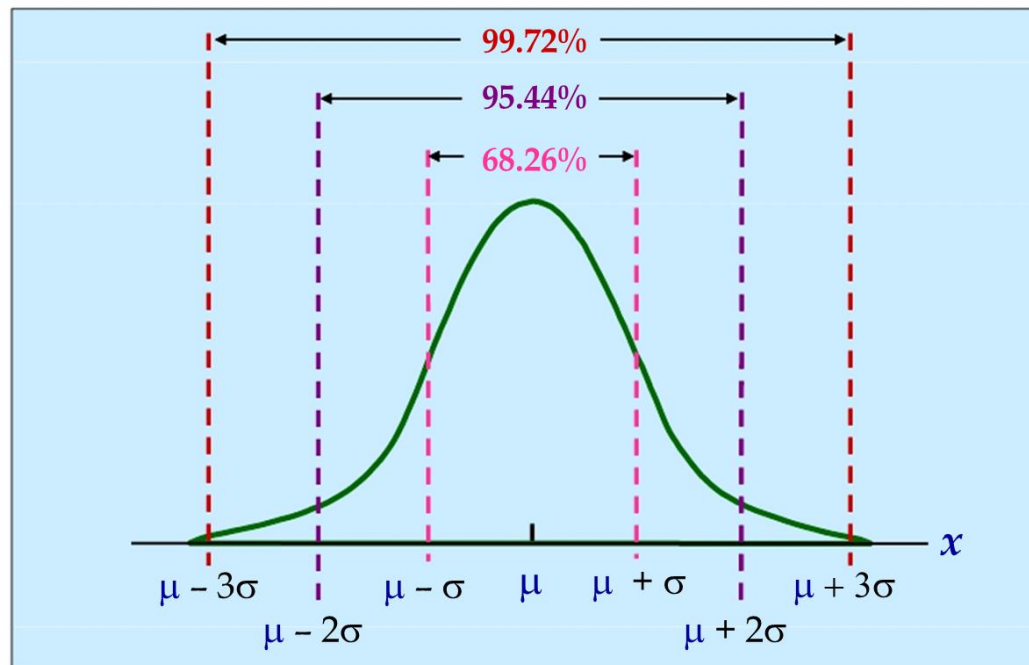
# Facts of normal distribution

$$X \sim N(\mu, \sigma^2)$$

68.26% of values of  $X$  are within 1 std deviation of its mean

95.44% of values of  $X$  are within 2 std deviations of its mean

99.72% of values of  $X$  are within 3 std deviations of its mean



# Estimator

Let  $X_1, X_2, \dots, X_n$  be a random sample. Let  $\theta$  be the **parameter of the statistical distribution describing a random variable** (e.g., the mean and variance of stock returns).

$\theta$  is estimated based on a sample.

Let  $\tilde{\theta}$  be an estimator of  $\theta$ . Two criteria of a *good estimator*

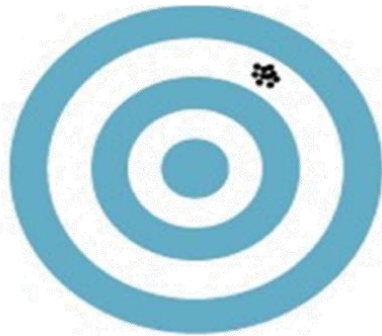
(a) **Unbiasedness**, meaning,  $E[\tilde{\theta}] = \theta$

(b) **Small MSE (mean-square error)**

MSE of  $\tilde{\theta}$  is defined by  $MSE[\tilde{\theta}] = E[(\tilde{\theta} - \theta)^2]$

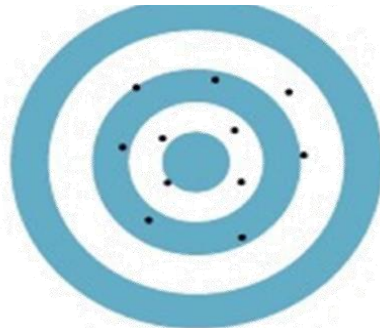
The estimator is said to be **efficient** if it is unbiased and has the minimum variance among all the unbiased estimators.

# Good estimator



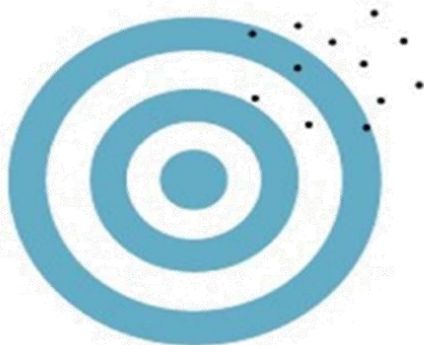
High bias, low variability

(a)



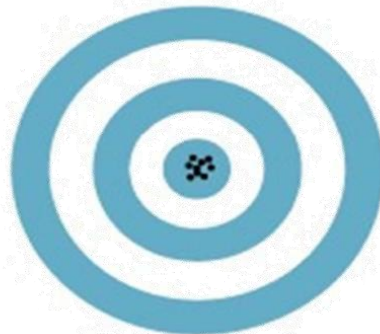
Low bias, high variability

(b)



High bias, high variability

(c)



The ideal: low bias, low variability

(d)

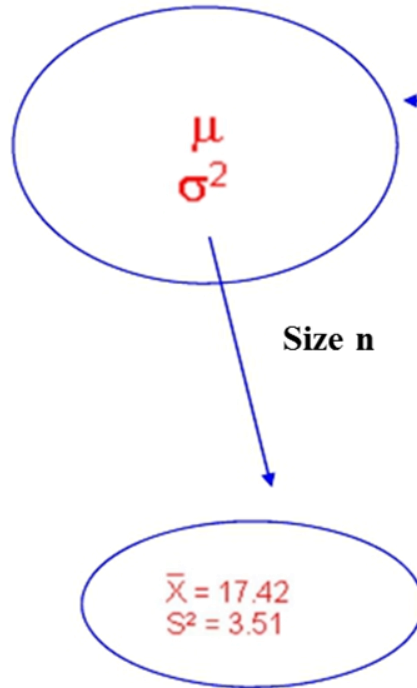
To reduce the *variability*, use a larger sample.

However *bias* can not be corrected by increasing the sample size!

To reduce *bias*, use random sampling.

# Inference

Population



Can we generalize that the population mean is near 17.42?

Can we generalize that the population variance is near 3.51?

What is the prob.  
that the population  
mean is actually more  
than  
21?

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$



# One-sided or two-sided hypothesis tests

An important application of statistics in finance is to test, based on sample data, a **hypothesis** concerning the population. Construct hypotheses for the following situations.

Return of fund manager A's portfolio is higher than that of the market

$$H_0: \mu_A = \mu_M$$

$$H_1: \mu_A > \mu_M$$

Volatility of Yen is higher than that of Euro

$$H_0: \sigma_Y^2 = \sigma_E^2$$

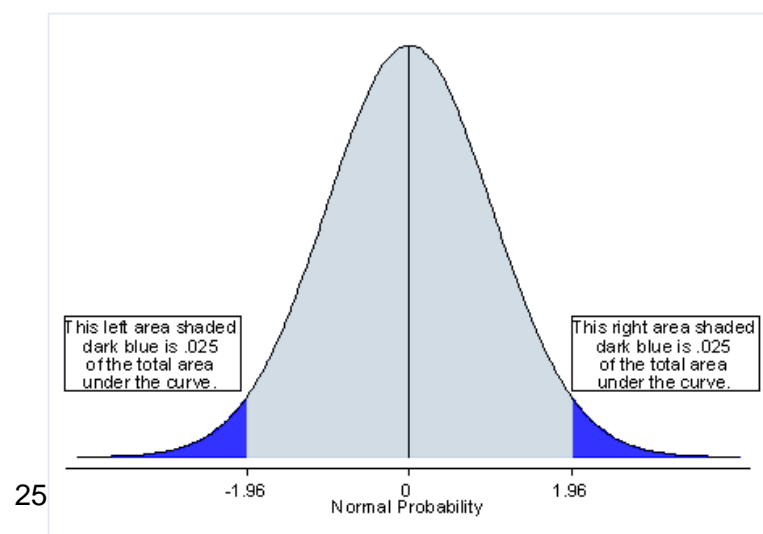
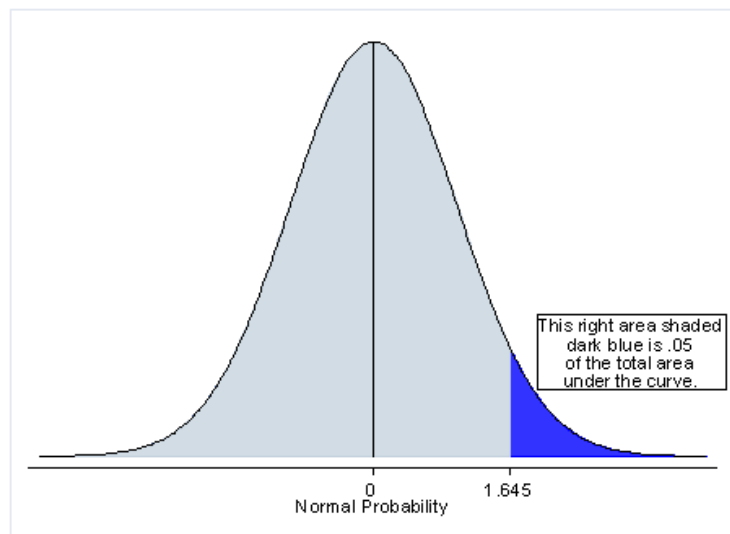
$$H_1: \sigma_Y^2 > \sigma_E^2$$

Interest rate and stock return are uncorrelated

$$H_0: \rho = 0$$

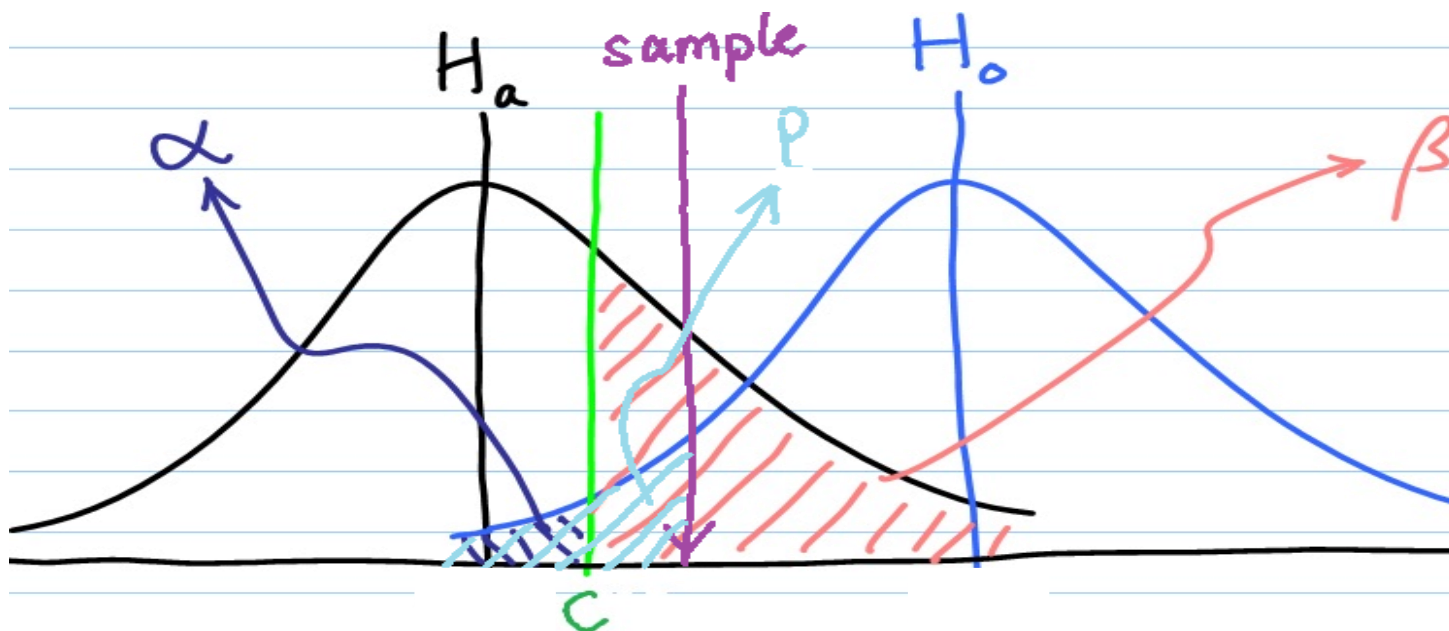
$$H_1: \rho \neq 0$$

*The first two examples are one-sided hypotheses, while the third example is two-sided hypotheses.*



# P-values and significance level

The *P-value* is the probability of seeing the observed data (or something even less likely) given the null hypothesis.  
*The risk to reject the null.*



If  $p\text{-value} < \alpha$ , we reject the null hypothesis at the level of significance  $\alpha$  (or higher). Otherwise, we don't.

# R

- ❑ The R statistical programming language is a free open source package based on the S language developed by Bell Labs.
- ❑ The language is very powerful for writing programs.
- ❑ Many statistical functions are already built in.

How to get R:

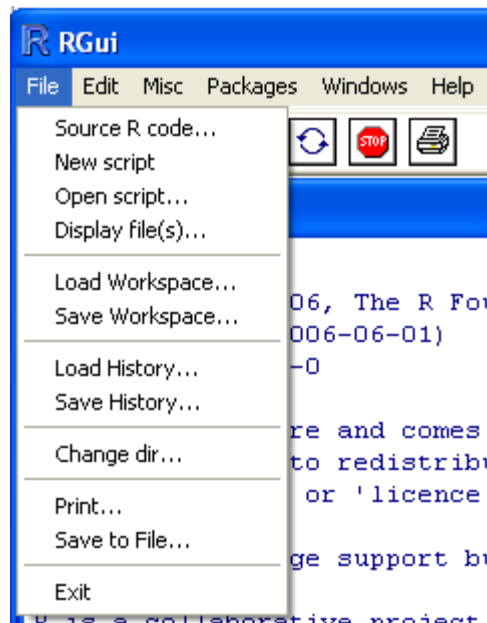
- <http://www.r-project.org/>
- Google: “R”
- Windows, Linux, Mac OS X, source

Files for introduction to R (source: Tyler K. Perrachione, Gabrieli Lab, MIT)

- [http://web.mit.edu/tkp/www/R/R\\_Tutorial\\_Data.txt](http://web.mit.edu/tkp/www/R/R_Tutorial_Data.txt)
- [http://web.mit.edu/tkp/www/R/R\\_Tutorial\\_Inputs.txt](http://web.mit.edu/tkp/www/R/R_Tutorial_Inputs.txt)

# Getting Started

- ❑ Opening a script.
- ❑ This gives you a script window.



# Getting Started

- ❑ Basic assignment and operations.
- ❑ Arithmetic Operations:
  - ❑  $+$ ,  $-$ ,  $*$ ,  $/$ ,  $^$  are the standard arithmetic operators.
- ❑ Matrix Arithmetic.
  - ❑  $*$  is element wise multiplication
  - ❑  $\%*\%$  is matrix multiplication
- ❑ Assignment
  - ❑ To assign a value to a variable use “<-”

# Math and variables

## Math:

```
> 1 + 1
```

```
[1] 2
```

```
> 1 + 1 * 7
```

```
[1] 8
```

```
> (1 + 1) * 7
```

```
[1] 14
```

## Variables:

```
> x <- 1
```

```
> x
```

```
[1] 1
```

```
> y = 2
```

```
> y
```

```
[1] 2
```

```
> 3 -> z
```

```
> z
```

```
[1] 3
```

```
> (x + y) * z
```

```
[1] 9
```

# Arrays

```
> x <- c(0,1,2,3,4)
```

```
> x
```

```
[1] 0 1 2 3 4
```

```
> y <- 1:5
```

```
> y
```

```
[1] 1 2 3 4 5
```

```
> z <- 1:50
```

```
> z
```

```
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
```

```
[16] 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
```

```
[31] 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
```

```
[46] 46 47 48 49 50
```

# Math on arrays

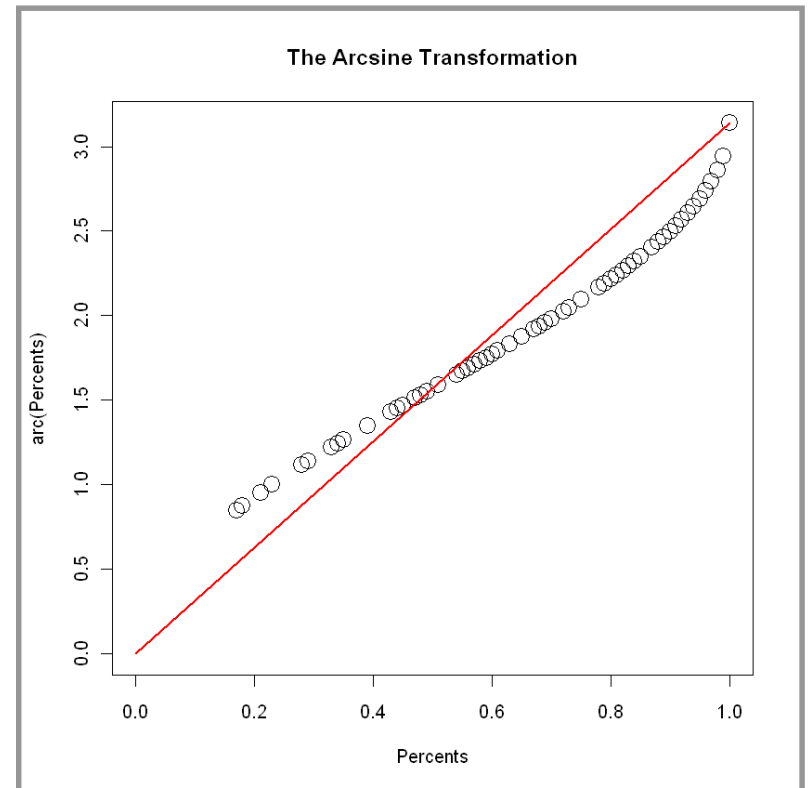
```
> x <- c(0,1,2,3,4)
> y <- 1:5
> z <- 1:50
> x + y
[1] 1 3 5 7 9
> x * y
[1] 0 2 6 12 20
> x * z
[1] 0 2 6 12 20 0 7 16 27 40 0
[12] 12 26 42 60 0 17 36 57 80 0 22
[23] 46 72 100 0 27 56 87 120 0 32 66
[34] 102 140 0 37 76 117 160 0 42 86 132
[45] 180 0 47 96 147 200
```



# Functions

```
> arc <- function(x) 2*asin(sqrt(x))
> arc(0.5)
[1] 1.570796
> x <- c(0,1,2,3,4)
> x <- x / 10
> arc(x)
[1] 0.0000000 0.6435011 0.9272952
[4] 1.1592795 1.3694384
```

```
> plot(arc(Percents)~Percents,
+ pch=21,cex=2,xlim=c(0,1),ylim=c(0,pi),
+ main="The Arcsine Transformation")
> lines(c(0,1),c(0,pi),col="red",lwd=2)
```



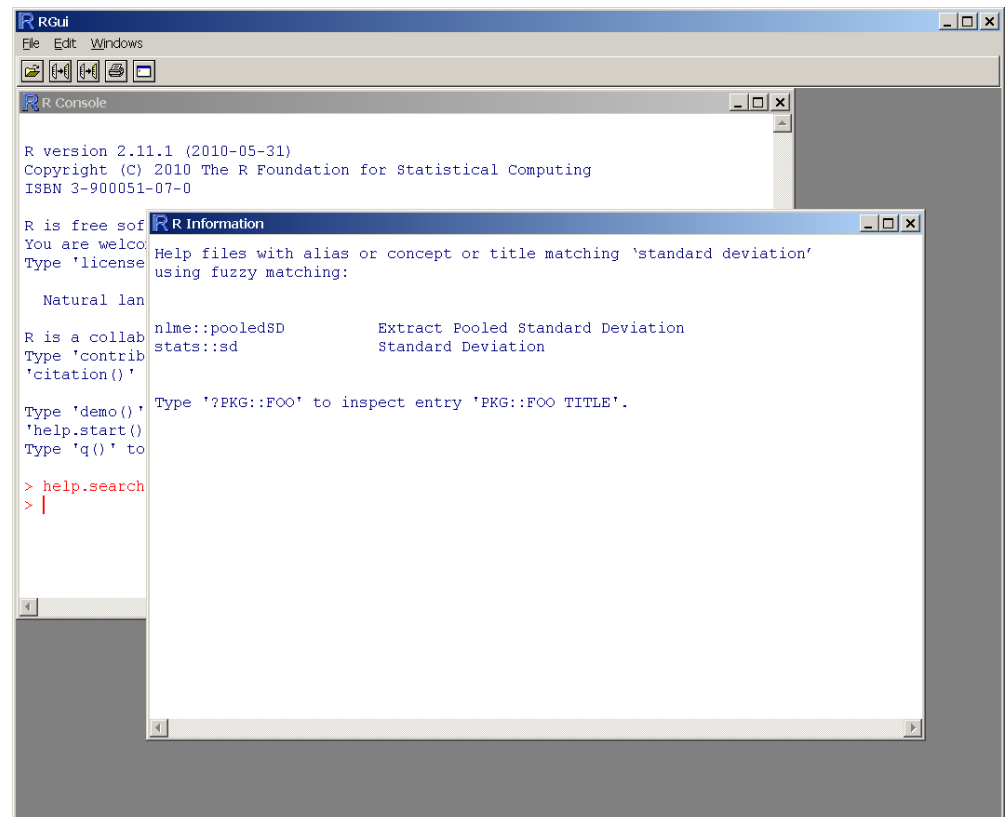
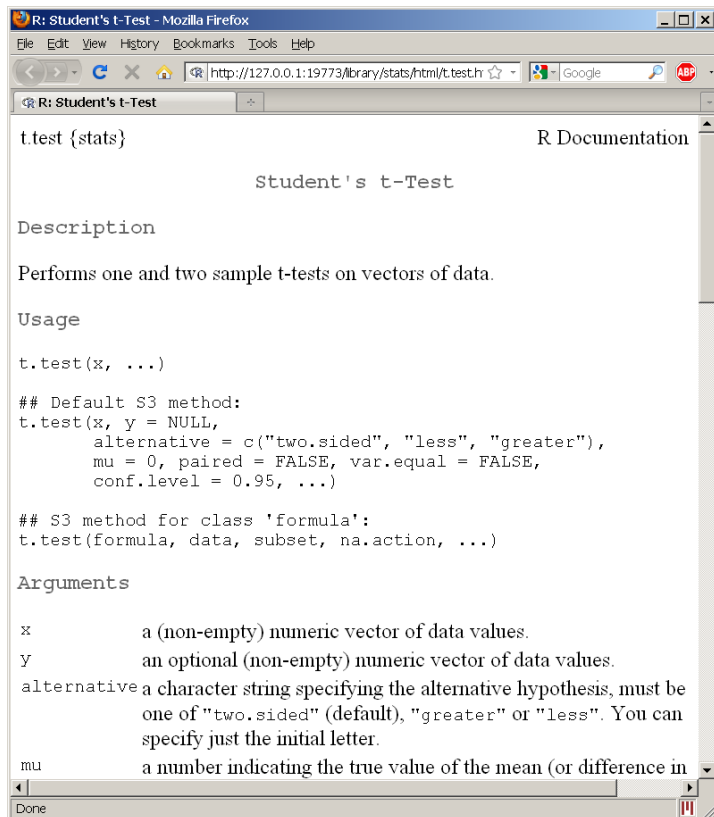
# Getting help

How to use help in R?

- ❑ R has a very good help system built in.
- ❑ If you know which function you want help with simply use `?_____` with the function in the blank.
- ❑ Ex: `?hist`.
- ❑ If you don't know which function to use, then use `help.search("_____")`.
- ❑ Ex: `help.search("histogram")`.

# Getting help

```
> help(t.test)
> help.search("standard deviation")
```



# Data analysis with R

Example experiment: Subjects learning to perform a new task:

- Two groups of subjects
  - (“A” and “B”; high and low aptitude learners)
- Two types of training paradigm
  - (“High variability” and “Low variability”)
- Four pre-training assessment tests

Example data in “R\_Tutorial\_Data.txt”

Use `setwd("c:/.../")` to set working directory



# Importing data

- ❑ How do we get data into R?
- ❑ Remember we have no point and click...
- ❑ First make sure your data is in an easy to read format such as CSV (Comma Separated Values).
- ❑ Use code:
  - ❑ `myData <- read.table("path", sep="," , header=TRUE)`

# Reading data from files

```
> myData <- read.table("R_Tutorial_Data.txt",
+ header=TRUE, sep="\t")
> myData
```

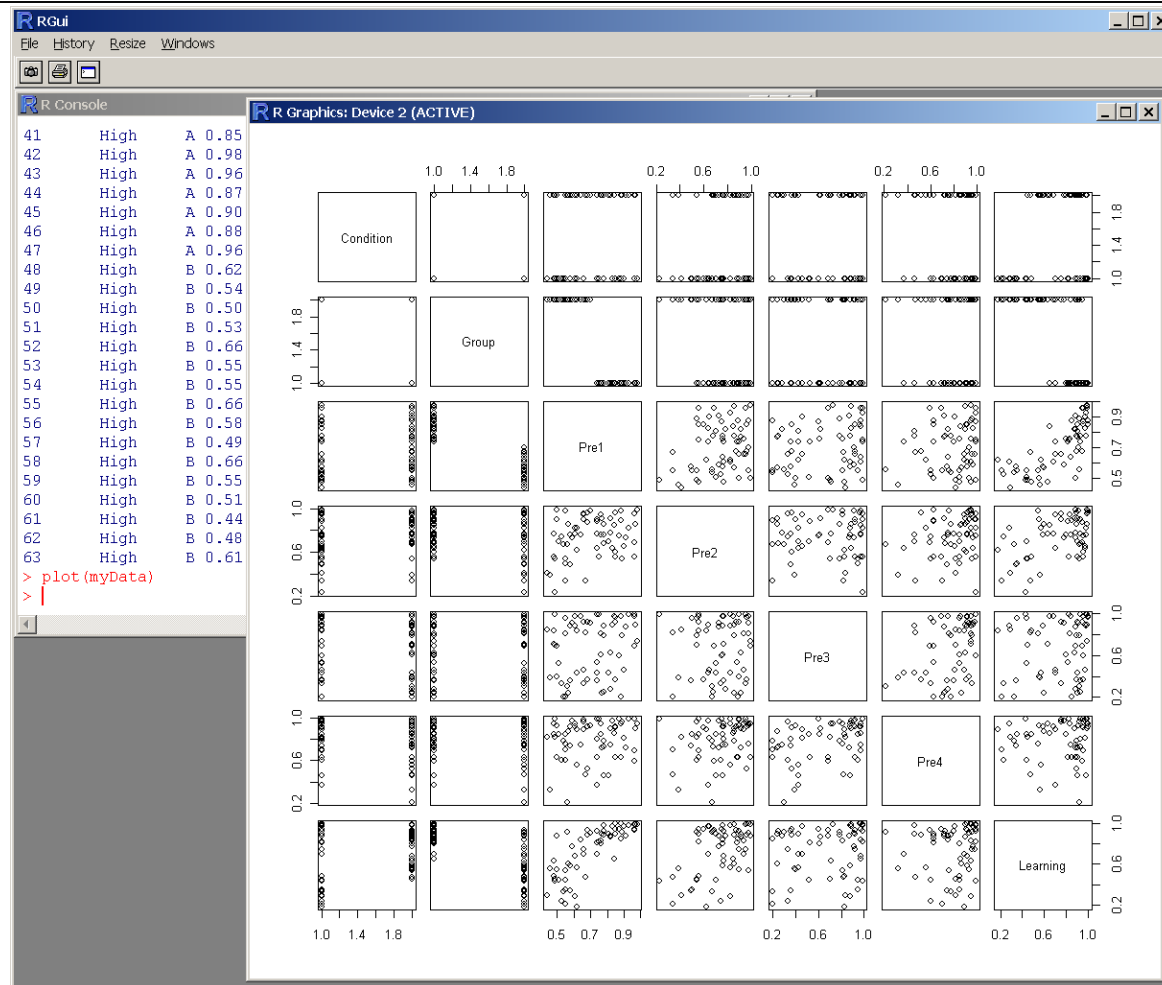
	Condition	Group	Pre1	Pre2	Pre3	Pre4	Learning
1	Low	A	0.77	0.91	0.24	0.72	0.90
2	Low	A	0.82	0.91	0.62	0.90	0.87
3	Low	A	0.81	0.70	0.43	0.46	0.90
...							
61	High	B	0.44	0.41	0.84	0.82	0.29
62	High	B	0.48	0.56	0.83	0.85	0.48
63	High	B	0.61	0.82	0.88	0.95	0.28

	A	B	C	D	E	F	G	H
1	Condition	Group	Pre1	Pre2	Pre3	Pre4	Learning	Gender
2	Low	A	0.77	0.91	0.24	0.72	0.90	M
3	Low	A	0.82	0.91	0.62	0.9	0.87	F
4	Low	A	0.81	0.7	0.43	0.46	0.9	F
5	Low	A	0.88	0.89	0.2	0.63	0.85	M
6	Low	A	0.78	0.68	0.25	0.73	0.93	F
7	Low	A	0.74	0.9	0.99	0.99	0.93	M
8	Low	A	0.78	0.86	0.79	0.78	0.89	F
9	Low	A	0.76	0.76	0.61	0.85	0.8	F
10	Low	A	0.93	0.82	0.99	0.99	0.98	M

Condition	Group	Pre1	Pre2	Pre3	Pre4	Learning
Low	A	0.77	0.91	0.24	0.72	0.9
Low	A	0.82	0.91	0.62	0.9	0.87
Low	A	0.81	0.7	0.43	0.46	0.9
Low	A	0.88	0.89	0.2	0.63	0.85
Low	A	0.78	0.68	0.25	0.73	0.93
Low	A	0.74	0.9	0.99	0.99	0.93
Low	A	0.78	0.86	0.79	0.78	0.89
Low	A	0.76	0.76	0.61	0.85	0.8
Low	A	0.93	0.82	0.99	0.99	0.98
Low	A	0.82	0.78	0.28	0.75	0.88
Low	A	0.91	0.73	0.87	0.72	0.88
Low	A	0.96	0.69	0.69	0.59	0.94
Low	A	0.97	0.86	0.89	0.9	0.99
Low	A	0.89	0.54	0.79	0.96	0.92
Low	A	0.76	0.94	0.81	0.95	0.83
Low	A	0.84	0.85	0.97	0.86	0.65
Low	B	0.62	0.82	0.43	0.56	0.57

# Visualizing datasets

```
> plot(myData)
```



# Selecting subsets of data

- ❑ Use a logical operator to do this.
  - ❑ `==`, `>`, `<`, `<=`, `>=`, `<>` are all logical operators.
  - ❑ Note that the “equals” logical operator is two `=` signs.
- ❑ Example:
  - ❑ `myData[myData$Group=="A"]`
  - ❑ This will return the rows of `myData` where `Group` is “A”.
  - ❑ Remember R is case sensitive!
  - ❑ This code does nothing to the original dataset.
  - ❑ `D.M <- myData[myData$Group=="A"]` gives a dataset with the appropriate rows.



# Selecting subsets of data

```
> myData$Learning
[1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.57 0.55
[19] 0.94 0.68 0.89 0.60 0.63 0.84 0.92 0.56 0.78
[28] 0.54 0.47 0.45 0.59 0.91 0.98 0.82 0.93 0.81
[37] 0.97 0.95 0.70 1.00 0.90 0.99 0.95 0.95 0.97
[46] 1.00 0.99 0.18 0.33 0.88 0.23 0.75 0.21 0.35
[55] 0.70 0.34 0.43 0.75 0.44 0.44 0.29 0.48 0.28
> myData$Learning[myData$Group=="A"]
[1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.98 0.82
[19] 0.93 0.81 0.97 0.95 0.70 1.00 0.90 0.99 0.95
[28] 0.95 0.97 1.00 0.99
```

Group A (high aptitude) usually performs well with high marks.

# Selecting subsets of data

```
> myData$Learning
 [1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.57 0.55
[19] 0.94 0.68 0.89 0.60 0.63 0.84 0.92 0.56 0.78
[28] 0.54 0.47 0.45 0.59 0.91 0.98 0.82 0.93 0.81
[37] 0.97 0.95 0.70 1.00 0.90 0.99 0.95 0.95 0.97
[46] 1.00 0.99 0.18 0.33 0.88 0.23 0.75 0.21 0.35
[55] 0.70 0.34 0.43 0.75 0.44 0.44 0.29 0.48 0.28

> attach(myData)
> Learning
 [1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.57 0.55
[19] 0.94 0.68 0.89 0.60 0.63 0.84 0.92 0.56 0.78
[28] 0.54 0.47 0.45 0.59 0.91 0.98 0.82 0.93 0.81
[37] 0.97 0.95 0.70 1.00 0.90 0.99 0.95 0.95 0.97
[46] 1.00 0.99 0.18 0.33 0.88 0.23 0.75 0.21 0.35
[55] 0.70 0.34 0.43 0.75 0.44 0.44 0.29 0.48 0.28
```

# Selecting subsets of data

```
> Learning[Group=="A"]  
 [1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98  
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.98 0.82  
[19] 0.93 0.81 0.97 0.95 0.70 1.00 0.90 0.99 0.95  
[28] 0.95 0.97 1.00 0.99
```

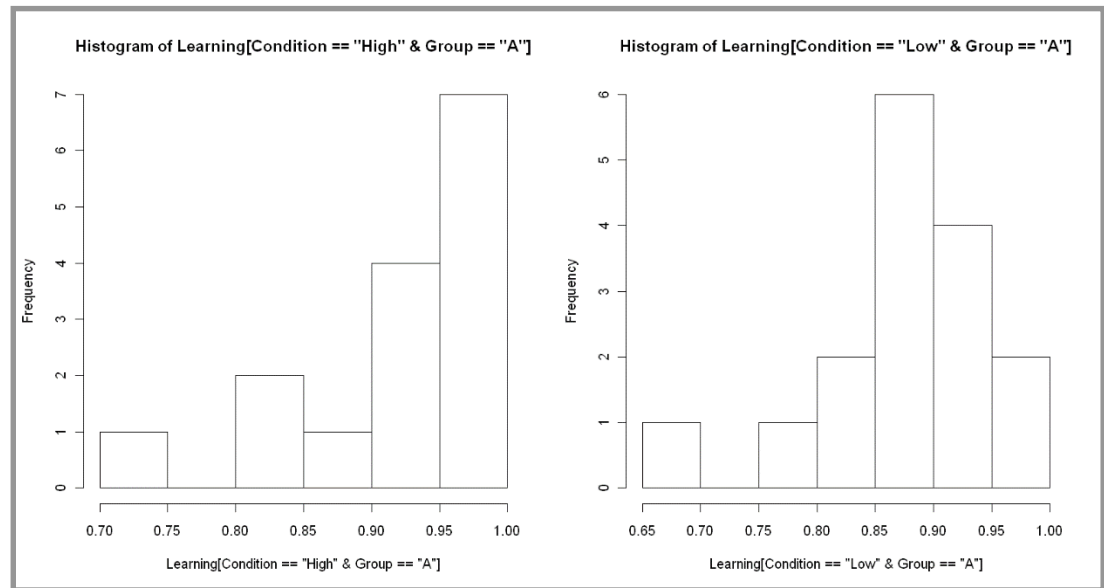
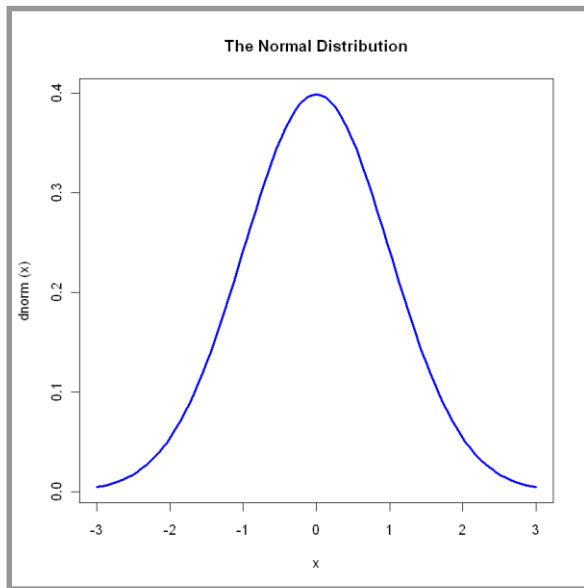
```
> Learning[Group!="A"]  
 [1] 0.57 0.55 0.94 0.68 0.89 0.60 0.63 0.84 0.92  
[10] 0.56 0.78 0.54 0.47 0.45 0.59 0.91 0.18 0.33  
[19] 0.88 0.23 0.75 0.21 0.35 0.70 0.34 0.43 0.75  
[28] 0.44 0.44 0.29 0.48 0.28
```

Is the low performance of Group B due to the teaching paradigm offered?

```
> Condition[Group=="B"&Learning<0.5]  
 [1] Low  Low  High High High High High High High  
[10] High High High High High  
Levels: High Low
```

# Are my data normally distributed?

```
> plot(dnorm, -3, 3, col="blue", lwd=3, main="The Normal Distribution")
> par(mfrow=c(1, 2))
> hist(Learning[Condition=="High"&Group=="A"])
> hist(Learning[Condition=="Low"&Group=="A"])
```



# Are my data normally distributed?

```
> shapiro.test(Learning[Condition=="High"&Group=="A"])
```

```
Shapiro-Wilk normality test
```

```
data: Learning[Condition == "High" & Group == "A"]  
W = 0.7858, p-value = 0.002431
```

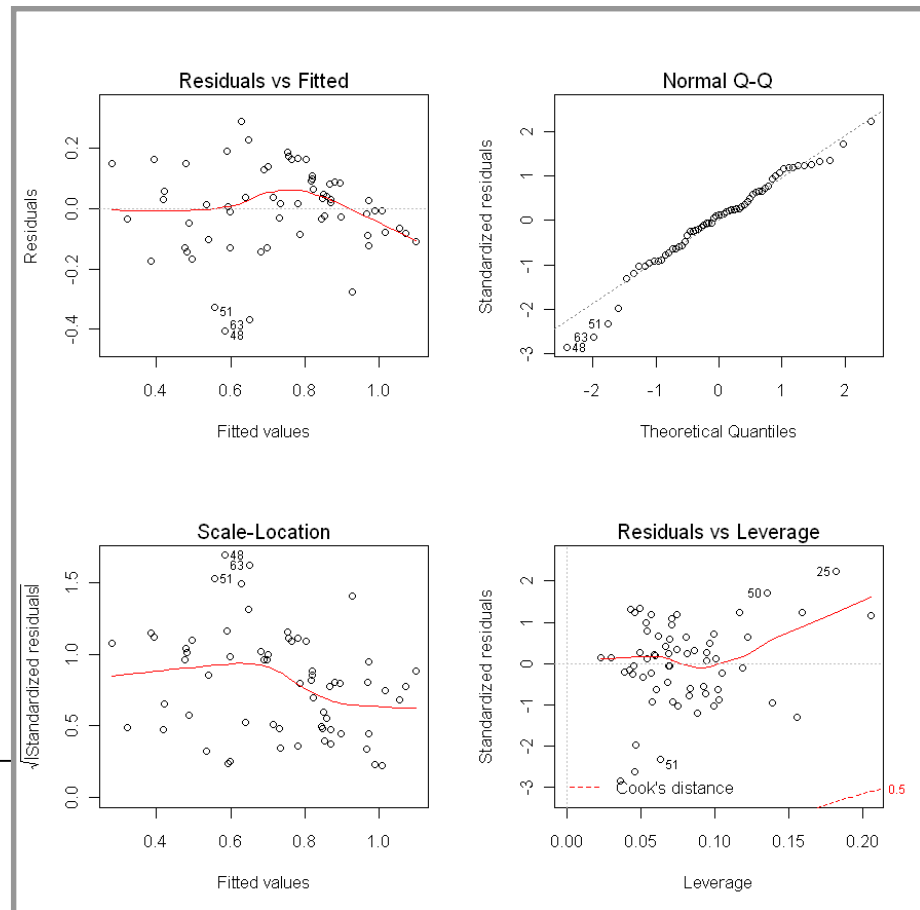
```
> shapiro.test(Learning[Condition=="Low"&Group=="A"])
```

```
Shapiro-Wilk normality test
```

```
data: Learning[Condition == "Low" & Group == "A"]  
W = 0.8689, p-value = 0.02614
```

# Linear models and ANOVA

```
> myModel <- lm(Learning ~ Pre1 + Pre2 + Pre3 + Pre4)
> par(mfrow=c(2,2))
> plot(myModel)
```



# Linear models and ANOVA

```
> summary(myModel)
```

```
Call:
```

```
lm(formula = Learning ~ Pre1 + Pre2 + Pre3 + Pre4)
```

```
Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.40518	-0.08460	0.01707	0.09170	0.29074

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-0.22037	0.11536	-1.910	0.061055	.
Pre1	1.05299	0.12636	8.333	1.70e-11	***
Pre2	0.41298	0.10926	3.780	0.000373	***
Pre3	0.07339	0.07653	0.959	0.341541	
Pre4	-0.18457	0.11318	-1.631	0.108369	

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1447 on 58 degrees of freedom
```

```
Multiple R-squared: 0.6677,    Adjusted R-squared: 0.6448
```

```
F-statistic: 29.14 on 4 and 58 DF,  p-value: 2.710e-13
```

# Linear models and ANOVA

```
> step(myModel, direction="backward")
```

```
Start:  AIC=-238.8
```

```
Learning ~ Pre1 + Pre2 + Pre3 + Pre4
```

	Df	Sum of Sq	RSS	AIC
- Pre3	1	0.01925	1.2332	-239.81
<none>			1.2140	-238.80
- Pre4	1	0.05566	1.2696	-237.98
- Pre2	1	0.29902	1.5130	-226.93
- Pre1	1	1.45347	2.6675	-191.21

```
Step:  AIC=-239.81
```

```
Learning ~ Pre1 + Pre2 + Pre4
```

	Df	Sum of Sq	RSS	AIC
- Pre4	1	0.03810	1.2713	-239.89
<none>			1.2332	-239.81
- Pre2	1	0.28225	1.5155	-228.83
- Pre1	1	1.54780	2.7810	-190.58

...

...

```
Step:  AIC=-239.89
```

```
Learning ~ Pre1 + Pre2
```

	Df	Sum of Sq	RSS	AIC
<none>			1.2713	-239.89
- Pre2	1	0.24997	1.5213	-230.59
- Pre1	1	1.52516	2.7965	-192.23

```
Call:
```

```
lm(formula = Learning ~ Pre1 + Pre2)
```

```
Coefficients:
```

(Intercept)	Pre1	Pre2
-0.2864	1.0629	0.3627



# Linear models and ANOVA

## ANOVA:

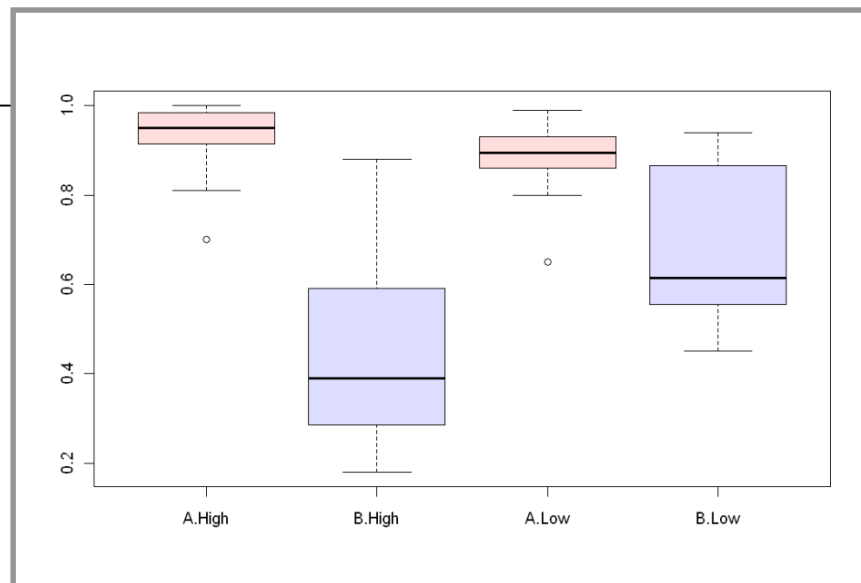
```
> myANOVA <- aov(Learning~Group*Condition)
> summary(myANOVA)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Group	1	1.8454	1.84537	81.7106	9.822e-13	***
Condition	1	0.1591	0.15910	7.0448	0.0102017	*
Group:Condition	1	0.3164	0.31640	14.0100	0.0004144	***
Residuals	59	1.3325	0.02258			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> boxplot(Learning~Group*Condition,col=c("#ffdddd","#dddddff"))
```



# Linear models and ANOVA

## ANOVA:

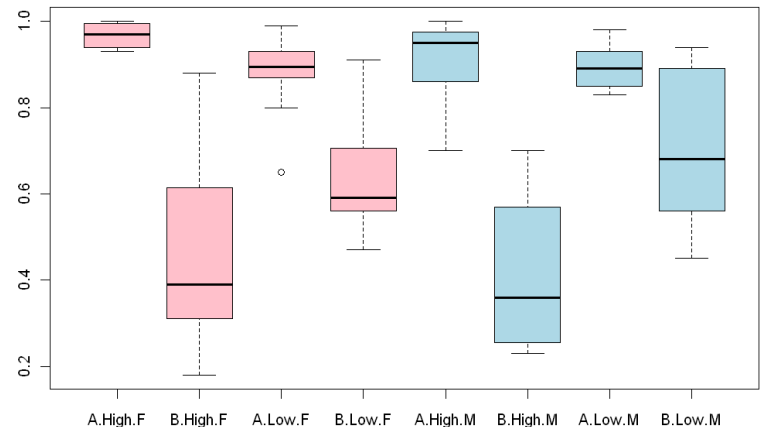
```
> myANOVA2 <- aov(Learning~Group*Condition+Gender)
> summary(myANOVA2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Group	1	1.84537	1.84537	80.3440	1.523e-12	***
Condition	1	0.15910	0.15910	6.9270	0.010861	*
Gender	1	0.04292	0.04292	1.8688	0.176886	
Group:Condition	1	0.27378	0.27378	11.9201	0.001043	**
Residuals	58	1.33216	0.02297			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> boxplot(Learning~Group*Condition+Gender,
+ col=c(rep("pink",4),rep("light blue",4)))
```



# Outline of the course

1. Introduction to R and review of Probability and Statistics.
2. Regression Analysis and Application.
3. CAPM, return based style analysis for hedge funds.
4. Modeling Univariate Time Series including stationarity and invertibility of ARMA process, identification tools.
5. Approaches to ARIMA Modeling and Forecasting
6. Autoregressive Conditional Heteroskedastic Models
7. Midterm
8. Value-at-Risk and Expected shortfall
9. Vector Autoregressive Models
10. Cointegration, Error Correction Models and Pairs Trading
11. Factor models: Principal Component Analysis and Factor Analysis with applications
12. Project presentation and review

# Books

## Textbook

1. [SDA] ***Statistics and Data Analysis for Financial Engineering*** (2010) by David Ruppert and published by Springer. While the focus of this text is on statistics and data analysis, it is an excellent text, provides a good introduction to R and also overlaps with several topics in the course. If students are going to purchase a text then I recommend this one. It may be downloaded for free from [NUS library E-Book website](#).
2. [FTS] Tsay, R.S. (2010) *Analysis of Financial Time Series*, Third Edition, Wiley. It may be downloaded for free from [NUS library E-Book website](#).
3. [SFM] Franke, J., Härdle, W. K., Hafner, C. M. (2015) *Statistics of Financial Markets An Introduction*. Springer. It may be downloaded for free from [NUS library E-Book website](#).

# Books

Other textbooks include:

3. J. Y. Campbell, A. W. Lo and A.C. MacKinlay (1996) [The Econometrics of Financial Markets](#), Princeton University Press.

4 The econometrics of financial markets by John Y. Campbell, Andrew W. Lo, A. Craig MacKinlay (1997).

5. Time Series: Theory and Methods by Peter J. Brockwell and Richard A. Davis (2006)

# Grading

1. Midterm (20% of grade)

2. Group project (40% of grade)

Students up to 5 people are working together on the project but **each student MUST specify his or her own contribution.**

***The member information should be confirmed by 4 September 2019 at GoogleDocs (link provided in IVLE announcement).***

3. Final exam (40% of grade)

Open-book, 2.5 Hours, 10 multiple choice questions and 2 essay questions

Bonus Points: You are strongly encouraged to ask/answer questions and implement R labs during lectures.



# Example of computer (R) lab and solutions

If you are unfamiliar with any of the R functions used below, then use R's help to learn about them; e.g., type `?rnorm` to learn that `rnorm` generates normally distributed random numbers.

You should study each line of code, understand what it is doing, and convince yourself that the code does what is being requested.

Note that anything that follows a pound sign is a comment and is used only to annotate the code.

# Example: R lab - Data Analysis

Obtain the data set [Stock\\_FX\\_bond.csv](#) from IVLE or the book's website and put it in your working directory. Start R and read the data with the following command:

```
dat = read.csv("Stock_FX_bond.csv",header=TRUE)
```

The data set `Stock_FX_bond.csv` contains the volumes and adjusted closing (AC) prices of stocks and the S&P 500, yields on bonds. The next lines of code print the names of the variables in the data set, attach the data, and plot the adjusted closing prices of GM and Ford.

```
names(dat)
```

```
attach(dat)
```

```
par(mfrow=c(1,2))
```

```
plot(GM_AC)
```

```
plot(F_AC)
```

Run the code below to find the sample size ( $n$ ), compute GM and Ford returns, and plot GM returns versus the Ford returns.

```
n = dim(dat)[1]
```

```
GMReturn = GM_AC[2:n]/GM_AC[1:(n-1)] - 1
```

```
FReturn = F_AC[2:n]/F_AC[1:(n-1)] - 1
```

```
par(mfrow=c(1,1))
```

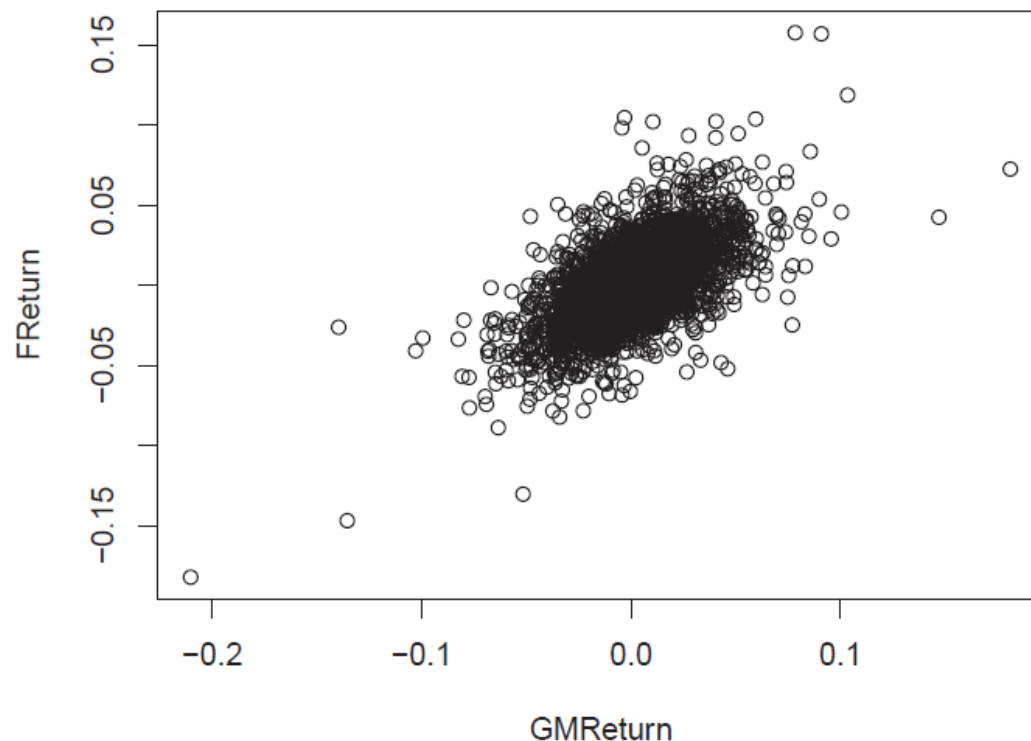
```
plot(GMReturn,FReturn)
```



# Example: R lab - Data Analysis

**Problem 1.** Do the GM and Ford returns seem positively correlated? Do you notice any outlying returns? If yes, do outlying GM returns seem to occur with outlying Ford returns?

**Answer:** The plot below shows a strong positive correlation, and the outliers in GM returns seem to occur with outlying Ford returns.



# Example: R lab - Data Analysis

**Problem 2.** Compute the log returns for GM and plot the returns versus the log returns? How highly correlated are the two types of returns? (The R function `cor` computes correlations.)

Answer: The correlation is almost 1 and the plot below shows an almost perfect linear relationship.

```
> cor(GMLogReturn,GMReturn)
[1] 0.9995408
```

When you exit R, you can Save workspace image, which will create an R workspace file in your working directory. Later, you can restart R from within Windows and load this workspace image into memory by right-clicking on the R workspace file. When R starts, your working directory will be the folder containing the R workspace that was opened.

# Excursion: Matrix

Applied Multivariate Statistical Analysis (2015) by Härdle, Wolfgang Karl; Simar, Léopold, Springer. Chapter 2. Access to [NUS library E-book](#).

A **matrix** is any doubly subscripted array of elements arranged in rows and columns.

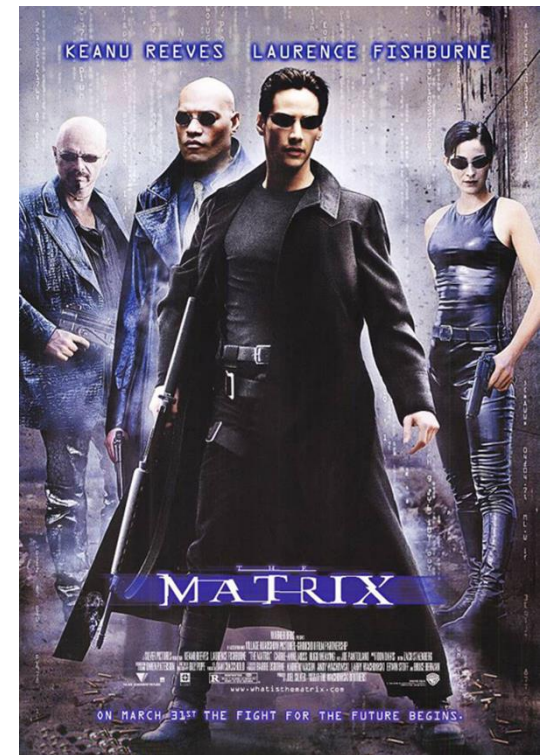
$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \{A_{ij}\}$$

$$A = [a_1, a_2, \dots, a_n] = \{a_j\}$$

**Row vector** is a  $[1 \times n]$  matrix:

**Column vector** is an  $[m \times 1]$  matrix:

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_m \end{bmatrix} = \{a_i\}$$



# Excursion: Matrix

**Square Matrix** has the same number of rows and columns.

$$B = \begin{bmatrix} 5 & 4 & 7 \\ 3 & 6 & 1 \\ 2 & 1 & 3 \end{bmatrix}$$

**Identity matrix** is square matrix with ones on the diagonal and zeros elsewhere.

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

**Transpose matrix:** Rows become columns and columns become rows

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

$$A' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

# Excursion: Matrix addition and subtraction

A new matrix  $\mathbf{C}$  may be defined as the additive combination of matrices  $\mathbf{A}$  and  $\mathbf{B}$ , where:  $\mathbf{C} = \mathbf{A} + \mathbf{B}$ .  $\{C_{ij}\} = \{A_{ij}\} + \{B_{ij}\}$

All three matrices are of the same dimension!


$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \quad \text{then } \mathbf{C} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 6 \\ 8 & 10 \end{bmatrix} = \mathbf{C}$$

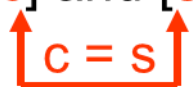
Matrix subtraction:  $\mathbf{C} = \mathbf{A} - \mathbf{B}$  is defined by  $\{C_{ij}\} = \{A_{ij}\} - \{B_{ij}\}$

# Excursion: Matrix multiplication

Matrices A and B have these dimensions:

  
 $[r \times c]$  and  $[s \times d]$

*Matrices A and B can be multiplied if  $c=s$ :*

$[r \times c]$  and  $[s \times d]$   


The resulting matrix will have the dimensions  $[r \times d]$ :

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} [2 \times 2]$$

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} [2 \times 3]$$

$$\mathbf{A} \times \mathbf{B} = \mathbf{C} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} & a_{21}b_{13} + a_{22}b_{23} \end{bmatrix} [2 \times 3]$$

# Excursion: Matrix inversion



$$\mathbf{B}^{-1}\mathbf{B} = \mathbf{B}\mathbf{B}^{-1} = \mathbf{I}$$

*Like a reciprocal  
in scalar math*

*Like the number  
one in scalar math*

Consider  $n$  equations in  $n$  variables:  $\sum_{j=1}^n a_{ij}x_j = b_i$  or  $\mathbf{Ax} = \mathbf{b}$

The unknown values of  $\mathbf{x}$  can be found using the inverse of matrix  $\mathbf{A}$  such that

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{b}$$

$$\begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 9 \end{bmatrix}$$

Inverse of  $\begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}$

$$\begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix} \times \begin{bmatrix} 6 \\ 9 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$