# TOPIC 1
# UNDERSTANDING FINANCIAL DATA

## 1. Introduction

Before building financial models, it is important to understand the empirical characteristics of financial data. Some key empirical properties investigated here are:

(i) The shape of the empirical distribution of returns
(ii) Autocorrelation structure in the mean of returns
(iii) Autocorrelation structure in the variance of returns

The return on a stock with price $P_t$ and dividend $D_t$ is computed as

$$R_t = \log(P_t + D_t) - \log(P_{t-1})$$

In the case where dividends are incorporated in prices $R_t = \log(P_t) - \log(P_{t-1})$.



## 2. Descriptive Statistics

There exists a number of descriptive measures which can be used to summarize the distribution of a financial time series $\{R_1, R_2, \ldots, R_T\}$ where $R_t$ is the return on an asset (stock) at time $t$. We assume that this distribution is stationary, so that the distribution of $R_t$ at each point in time is the same.



The descriptive statistics considered in this course are as follows:

### 2.1 Measure of Location (Mean)

The mean is computed as

$$\mu = \frac{1}{T}\sum_{t=1}^{T} R_t$$

If $R_t$ is the return on a stock, $\mu$ is the average return on the stock

### 2.2 Measures of Variation

*Standard Deviation*

The standard deviation of the return on a stock is computed as:

$$\sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(R_t - \mu)^2}$$

It can be interpreted as a measure of the stock's risk.

*Variance*

The variance of the stocks return is computed as

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(R_t - \mu)^2$$

and is also a measure of the stock's risk.

*Covariance*

Let $X_t$ be the return on stock *X* and $Y_t$ the return on stock *Y*. The covariance between the returns on the two stocks is:



$$\sigma_{XY} = \frac{1}{T}\sum_{t=1}^{T}(X_t - \mu_x)(Y_t - \mu_y)$$

The covariance measures the degree of association between $X_t$ and $Y_t$.



*Correlation*

The correlation between $X_t$ and $Y_t$ is



$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

where $-1 \le \rho_{XY} \le 1$. The correlation coefficient is a dimensionless quantity and it also measures the degree of association between $X_t$ and $Y_t$.

## 2.3 Skewness

Skewness is computed as

$$S = \frac{1}{T}\sum_{t=1}^{T}\frac{(X_t - \mu)^3}{\sigma^3}$$

For symmetric distributions, such as the normal distribution, there is no skewness. For some distributions, however, high (low) values can be more common than low (high) values. In this case the distribution is skewed to the left (right).

### 2.4 Kurtosis

Kurtosis is computed as

$$K = \frac{1}{T}\sum_{t=1}^{T}\frac{(X_t - \mu)^4}{\sigma^4}$$

An important stylised fact concerning financial data is that there are frequent extreme observations in both tails of the empirical distribution of many financial series, which is not consistent with the assumption of normality. For the *normal* distribution, K = 3. For financial data, we frequently observe K > 3. This "excess kurtosis" is caused by the "fatness" in the tails of the data distribution.

## 3. Predictability

### 3.1 Autocorrelation of Returns



Definition

Let $X_t$ be the return of an asset at time *t*. The autocorrelation between $X_t$ and $X_{t-j}$ is estimated as



$$r_j = \frac{\sum_{t=j+1}^{T}(X_t - \mu)(X_{t-j} - \mu)}{\sum_{t=1}^{T}(X_t - \mu)^2}$$



Distribution

Under the hypothesis of no autocorrelation, that is, $H_0 : \rho = 0$,

$$r_j \sim N(0,1/T),$$

where $T$ is the sample size.

Testing

A joint test of autocorrelation up to lag *m*, can be undertaken by using the Ljung-Box statistic, $Q_x(m) = T(T+2)\sum_{j=1}^{m}\frac{r_j^2}{T-j}$, which is approximately $\chi^2(m)$ under $H_0$.

Observation

Most empirical studies show that there is very little evidence of autocorrelation in returns data so that there is very little evidence of dependence in the mean.

## 3.2 Autocorrelation of Squared Returns

Testing

This can be tested by computing both $r_j$ and the Ljung-Box statistic but with $X_t$ replaced by $X_t^2$, that is, using squared returns. It is common to denote the Ljung-Box statistic when based on squared returns as $Q_{xx}(m)$.

Interpretation

Significant autocorrelation in squared returns reflects the volatility clustering characteristically observed in returns; namely, large (small) changes in returns tend to be followed by large (small) changes. As will be discussed later, significant autocorrelation in squared returns is evidence of ARCH (Autoregressive Conditional Heteroscedasticity) effects, that is, of a time-varying conditional variance in returns.

Observation

In contrast to the autocorrelation structure of returns, there is evidence of significant autocorrelation in squared returns. This implies that returns are not independent.



## 3.3 Application: Testing for Efficiency in Stock Returns



Theory

An important model used in finance to explain financial prices is based on the efficient markets hypothesis. A market is said to be *weakly* efficient if the most recent price reflects the available information. This implies that the price $P_t$, of a financial asset follows a *random walk*:



$$P_t = P_{t-1} + \varepsilon_t$$

where $\varepsilon_t$ is a disturbance term. Alternatively, the logarithmic form is

$$\log(P_t) = \log(P_{t-1}) + \mu_t$$

where $\mu_t$ is a disturbance term.

Implication

If the market is weakly efficient there should be no information contained in the disturbance term $\mu_t$ that is useful for predicting $\mu_{t+1}$.

Testing

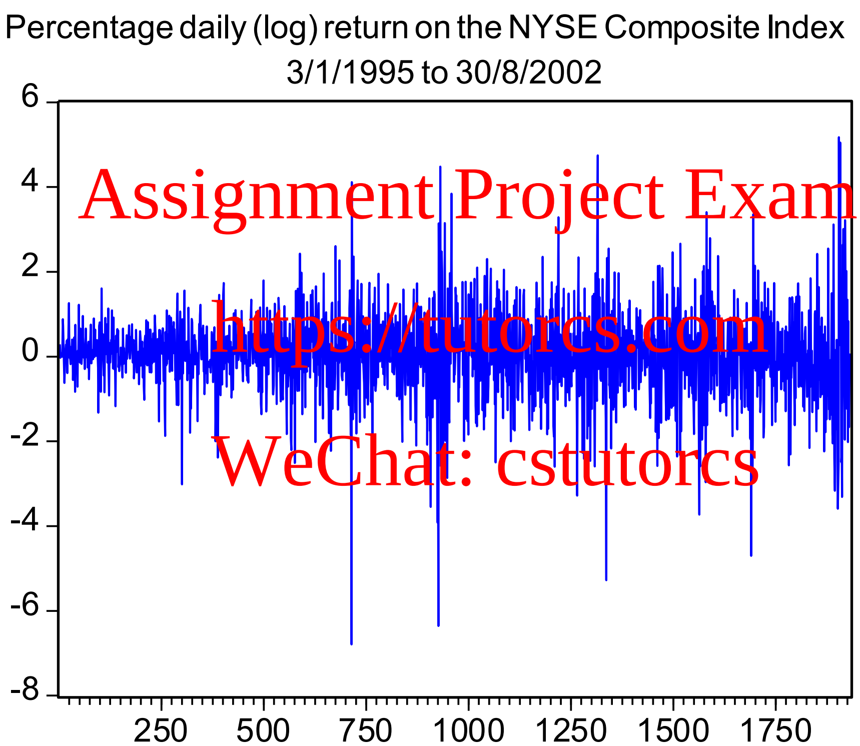
This suggests that a simple test of weak efficiency is to compute the returns

$$\mu_t = \log(P_t) - \log(P_{t-1})$$

and test for autocorrelation. If there is no significant autocorrelation, this provides support for the efficient markets hypotheis.

Application

The data are the log returns on the NYSE Composite Index, expressed as a percentage (by multiplying the daily log return by 100). The sample period is from 3/1/1995 to 30/8/2002, a total of 1931 observations. A graph of the data is shown below. The graph shows volatility clustering in the sense that large movements in returns are accompanied by further large movements in returns resulting in periods of high volatility. Similarly, tranquil periods are also evident in the graph.






Some descriptive statistics are shown in the table below. Some key points are:

1. Returns show significant autocorrelation of various orders as based on the autocorrelation coefficient ($r_j$) and the Ljung-Box statistic ($Q_x(j)$). This suggests that there is some dependency in the mean and hence the hypothesis that the stock market is efficient is rejected for this data set.

2. The autocorrelations of squared returns are much larger than those of returns. Moreover, the Ljung-Box test statistic applied to squared returns is also much larger and very significant. Both results suggest that there is considerable

dependency in the variance.

**Table 1**. Descriptive Statistics of daily (log) percentage return on the NYSE Composite Index: 3/1/95-30/8/02

| Statistic | Return ( $R_{NYSE}$ ) | p-value |
|---|---|---|
| $r_1$ | 0.068 | |
| $r_2$ | -0.046 | |
| $r_3$ | -0.031 | |
| $r_4$ | -0.001 | |
| $r_5$ | -0.052 | |
| | | |
| $Q_x(1)$ | 9.0448 | 0.003 |
| $Q_x(5)$ | 20.226 | 0.001 |
| $Q_x(10)$ | 26.723 | 0.003 |
| $Q_x(15)$ | 36.726 | 0.001 |
| $Q_x(20)$ | 39.710 | 0.005 |
| **Statistic** | **Squared Return $(R_{NYSE})^2$** | **p-value** |
| $r_1$ | 0.179 | |
| $r_2$ | 0.168 | |
| $r_3$ | 0.175 | |
| $r_4$ | 0.109 | |
| $r_5$ | 0.184 | |
| | | |
| $Q_{xx}(1)$ | 62.044 | 0.000 |
| $Q_{xx}(5)$ | 264.80 | 0.000 |
| $Q_{xx}(10)$ | 389.80 | 0.000 |
| $Q_{xx}(15)$ | 449.58 | 0.000 |
| $Q_{xx}(20)$ | 491.70 | 0.000 |

$r_j$ is the autocorrelation coefficient at lag $j$. $Q_x(j)$ and $Q_{xx}(j)$ is the Ljung-Box $Q$ statistic for the first $j$ lags of the autocorrelation function of returns and squared returns, respectively

## 4. Distribution of Returns

The above example demonstrates the autocorrelations in returns and squared returns. We now look at the shape of the empirical distribution of asset returns through the mean, variance, skewness and kurtosis. We discuss the shape of the empirical distribution of returns in relation to the normal distribution.

### 4.1 The Normal Distribution

Definition

The normal distribution of a random variable $X$ with mean $\mu$ and variance $\sigma^2$, is denoted as $N(\mu, \sigma^2)$ and is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

If a normal random variable has been standardized to have zero mean and unit variance, then the standard normal distribution is denoted as $N(0,1)$ and is given by



$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < x < \infty$$



Properties

(1) The normal distribution is bell-shaped and symmetric around the origin. Thus the normal distribution exhibits no skewness.



(2) The skewness and kurtosis coefficients for the normal distribution are respectively

$$S = 0$$

$$K = 3$$

### 4.2 Testing for Normality

Single Tests

A simple way to test for normality is to compare the computed skewness and kurtosis coefficients with the theoretical values under the assumption of normality; namely 0 and 3 respectively. Thus, the tests are

$$\text{Skewness Test: } Z_{Sk} = \frac{S}{\sqrt{6/T}}$$

$$\text{Kurtosis Test: } Z_{Kt} = \frac{K-3}{\sqrt{24/T}}$$

where $S$ and $K$ are the estimated statistics for skewness and kurtosis, respectively. Both test statistics are distributed under the null hypothesis of normality as $N(0,1)$. Thus "large" values of the test statistics, say in excess of two standard deviations (that is, greater than 2 or less than -2) constitute rejection of the hull hypothesis of normality.

Joint Test

A joint test can also be constructed as

$$JB = Z_{Sk}^2 + Z_{Kt}^2$$

which is distributed as a chi-square with two degrees of freedom (i.e. $\chi^2(2)$). The null hypothesis of normality is rejected at the 5% level when the p-value is less than $\alpha$=0.05. This is commonly referred to as the Jarque-Bera test for normality.

## 4.3 Leptokurtosis

The distribution of many asset returns series have empirical distributions which differ from normality in two respects:



(1) *Fatness in the tails*, which corresponds to points in time where large movements in returns have been excessive relative to the normal distribution.

(2) *Sharp peaks*, which corresponds to periods when there is very little movement in the return series.

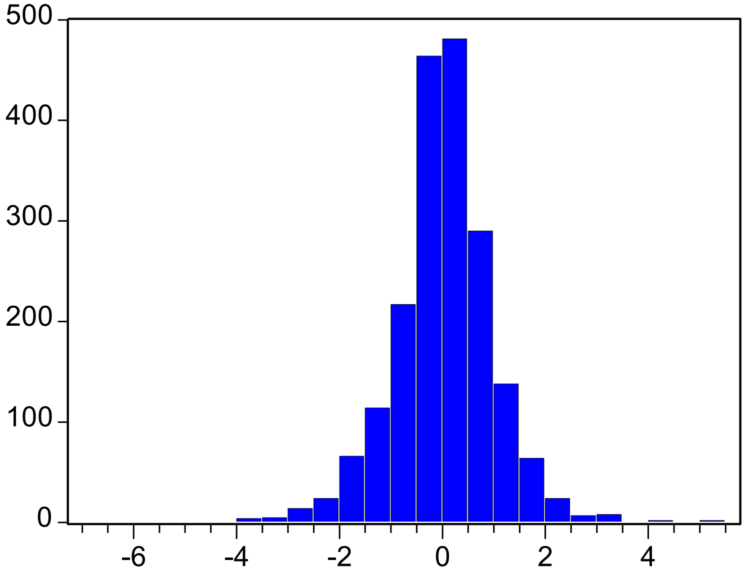


Distributions which have these two properties are known as *leptokurtic*.



## 4.4 Application: The Distribution of Stock Returns

The empirical distribution for (log) daily percentage return on the NYSE Composite Index is shown in the figure below. As before the data covers the period 3/1/95 to 30/8/02.




| | |
|---|---|
| Series: SR | |
| Sample 1 1931 | |
| Observations 1930 | |
| Mean | 0.035300 |
| Median | 0.052285 |
| Maximum | 5.178704 |
| Minimum | -6.791142 |
| Std. Dev. | 1.006207 |
| Skewness | -0.315728 |
| Kurtosis | 7.224376 |
| Jarque-Bera | 1467.129 |
| Probability | 0.000000 |

Some key features are:

1. The mean represents the average (percentage) daily return on the NYSE Composite Index. The annual average return is 0.035×250=8.75% assuming that there are 250 trading days in a year.

2. The skewness appears to be small. The test statistic $Z_{sk} = \dfrac{S}{\sqrt{6/T}} = -5.66407$ indicates that distribution of returns is negatively skewed.

3. The distribution is fat-tailed as evidenced by the high coefficient of kurtosis with the test statistic $Z_{Kt} = \dfrac{K-3}{\sqrt{24/T}} = 37.89205$.

4. This is further supported by the Jarque-Bera statistic, *JB*=1467.13 which is significant at the 1% level (p-value<0.01).

5. Conclude that returns on the NYSE Composite Index are not normally distributed.