Econometrics & Financial Markets

Financial data and descriptive statistics

Toulouse Business School
MSc BIF

Anna CALAMIA
a.calamia@tbs-education.fr

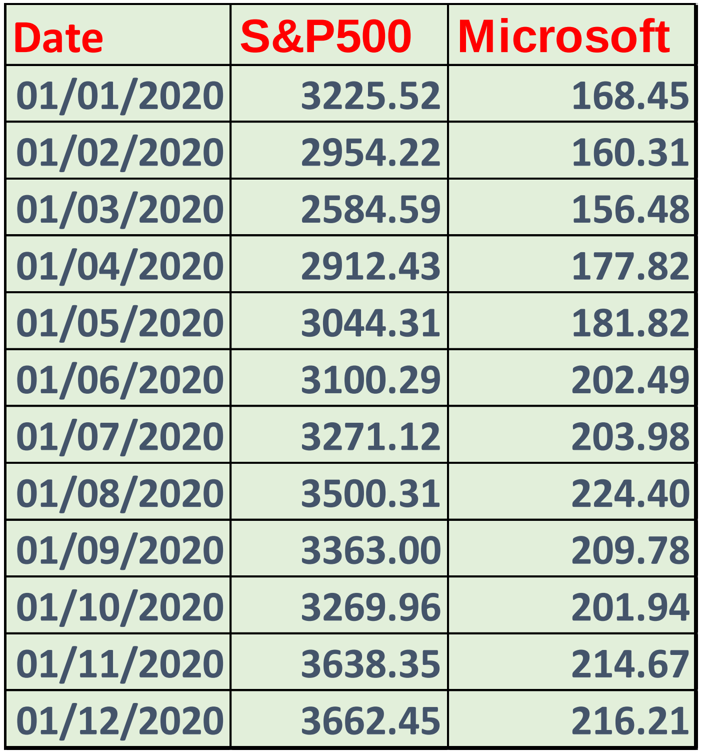
Outline

- Financial data
- Descriptive statistics: mean, variance, skewness kurtosis
- Brief overview of Normal distribution, tools to assess normality and some other standard distributions (Khi2, Student, Fisher).

Financial Data

There are 3 types of data which econometricians might use for analysis:

1. **Times series data**
   - How the value of a country’s stock index has varied with that country’s macroeconomic fundamentals.
   - How the value of a company’s stock price has varied with that country’s market index.



Date	S&P500	Microsoft
01/01/2020	3225.52	168.45
01/02/2020	2954.22	160.31
01/03/2020	2584.59	156.48
01/04/2020	2912.43	177.82
01/05/2020	3044.31	181.82
01/06/2020	3100.29	202.49
01/07/2020	3271.12	203.98
01/08/2020	3500.31	224.40
01/09/2020	3363.00	209.78
01/10/2020	3269.96	201.94
01/11/2020	3638.35	214.67
01/12/2020	3662.45	216.21

Financial Data

2. Cross-Sectional data

- Data collected at a single point in time
- Relationship between board size (board independence) and size (age) for quoted firms
- Relationship between gross investment and firm’s value and capital

Firm	Investment	Value	Capital
1	1486.7	5593.6	2226.3
2	459.3	2115.5	669.7
3	189.6	2759.9	888.9
4	172.49	703.2	414.9
5	81.43	365.7	804.9
6	135.72	927.3	238.7
7	89.51	192.7	511.3
8	68.6	1188.9	213.5
9	49.34	474.5	468
10	5.12	58.12	14.33

Financial and Business Data

3. Panel Data

- Relationship between returns and earnings for several stocks over time (dimensions of both time series and cross-sections)

Quarter	Stock	EBIT	Return
2017-Q1	AIRBUS	533	-0.00433
2017-Q2	AIRBUS	529	-0.00305
2017-Q3	AIRBUS	414	0.00513
2017-Q4	AIRBUS	878	-0.01073
2018-Q1	AIRBUS	168	-0.00192
2018-Q2	AIRBUS	871	0.02181
2018-Q3	AIRBUS	1524	-0.00661
2018-Q4	AIRBUS	2155	0.00239
2017-Q1	CARREFOUR	1385	0.005
2017-Q2	CARREFOUR	546	0.00317
2017-Q3	CARREFOUR	546	0.00826
2017-Q4	CARREFOUR	427	0.00194
2018-Q1	CARREFOUR	427	0.00328
2018-Q2	CARREFOUR	-169	-0.00431
2018-Q3	CARREFOUR	-169	-0.00151
2018-Q4	CARREFOUR	931	0.00573

Financial data

- Data may be **quantitative** (e.g. exchange rates, stock prices, number of shares outstanding).
  - ➔ Continuous data can take on any value and are not confined to take specific numbers.
  - ➔ Discrete data can only take on certain values, which may be integers (e.g. the number of shares traded during a day).
- or **qualitative** (e.g. day of the week)
  - ➔ Ordinal : a figure of 12 may be viewed as `better' than a figure of 6, but could not be considered twice as good
  - ➔ Nominal there is no natural ordering of the values at all.

Data Cleaning

- Data cleaning is an essential part of statistical analysis.
- Often time-consuming.

- Raw data: the data as it comes in (may lack headers, contain wrong data types, wrong category labels, unknown or unexpected character…)
- To get technically correct data, you have to organise data (columns) and assign variable names and types (text variables, number variable...).
- Technically correct data may still have missing values, outliers or (obvious) errors (e.g. negative age or bid-ask spread). These inconsistencies should either be removed, corrected or imputed.
- You may also choose to apply other possible filters before you implement specific methodologies (e.g. the first observations of a new stock, the 1% highest/lowest data…)

Econometric model - Steps

- General statement of the problem and formulation of an estimable model (from a theoretical model or intuition that 2 or more variables should be related)
- Collection and cleaning of data relevant to the model
- Choice of relevant estimation method and model estimation
- Statistical evaluation of the model
  - ○ yes
    - ➢ interpret (and evaluate according to theory)
    - ➢ use for analysis
  - ○ no => reformulate

Descriptive Statistics

Returns

P_t is the price of a stock or a portfolio evaluation at time t.

The return of the stock or portfolio between time t and time t-1 is :

$$\boxed{R_t = \frac{P_t - P_{t-1}}{P_{t-1}}}$$ or $$\boxed{R_t = \frac{P_t - P_{t-1} + D_t}{P_{t-1}}}$$ if dividend

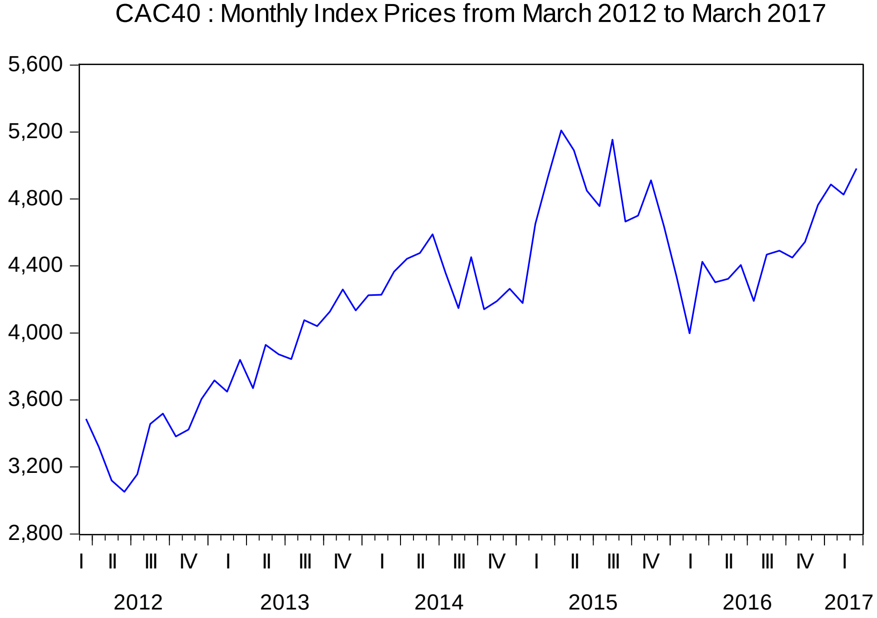
Log returns:

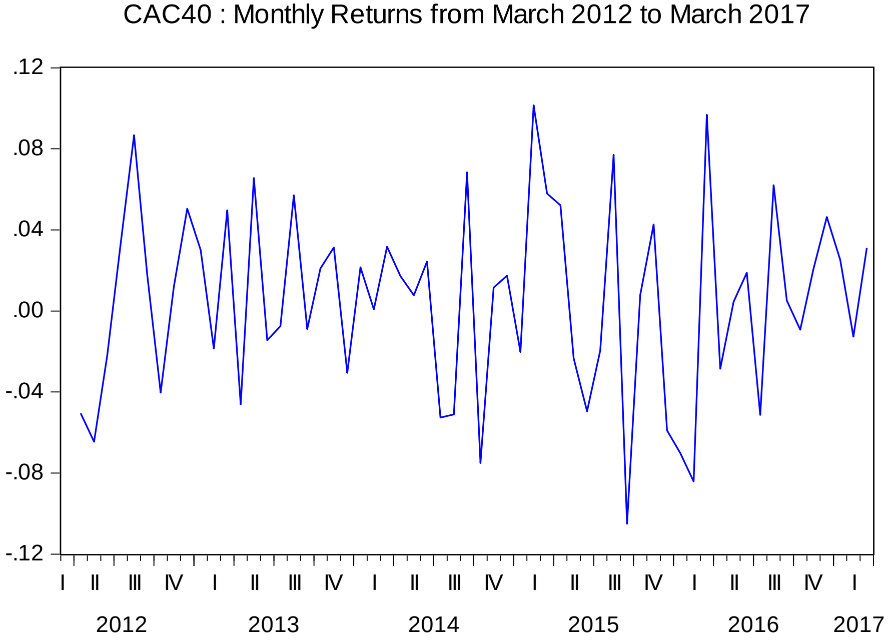
We can write $$R_t = \frac{P_t \; - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1 \Leftrightarrow R_t + 1 = \frac{P_t}{P_{t-1}}$$

Log’s property : when R_t is « small », $\ln(R_t + 1) = R_t$ then $\ln(\frac{P_t}{P_{t-1}}) = \ln(R_t + 1) = R_t$

$$\boxed{R_t = \ln(\frac{P_t}{P_{t-1}}) = \ln(P_t) - \ln(P_{t-1})}$$

Returns


CAC40 : Monthly Index Prices from March 2012 to March 2017

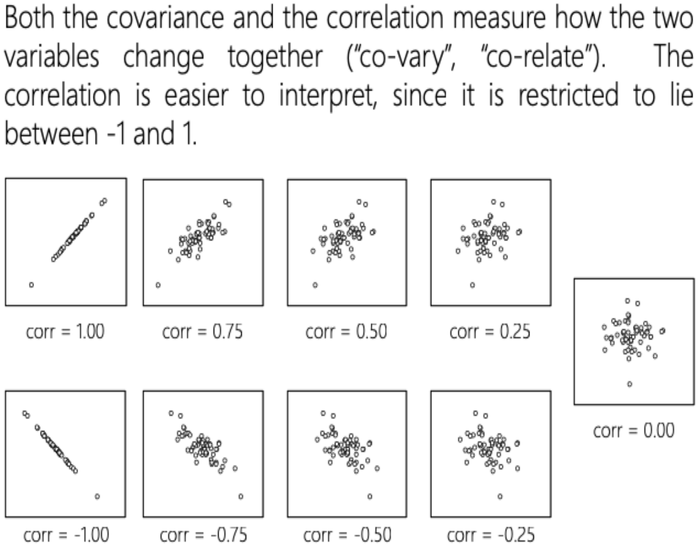

CAC40 : Monthly Returns from March 2012 to March 2017

Descriptive Statistics: Mean and variance

- Mean (expected value): $\bar{R} = \frac{1}{T}\sum_{t=1}^{T} R_t$

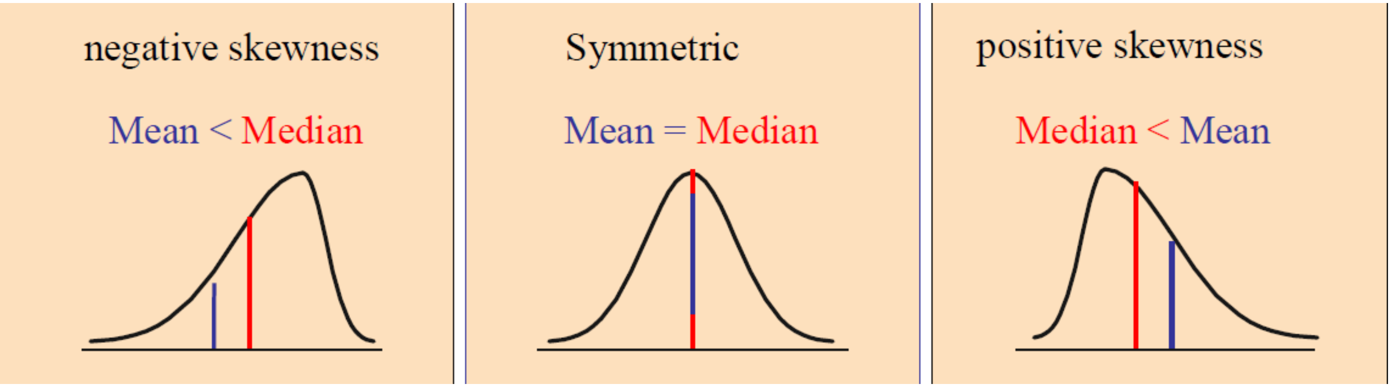
- Volatility measures:
  - → Variance : $Var(R) = \frac{1}{T-1}\sum_{t=1}^{T}(R - \bar{R})^2$
  - → Standard deviation: $\sigma(R) = \sqrt{V(R)}$
    - *T* is the number of observations
    - *Rt* is return between the dates *t-1* and *t*
    - $\bar{R}$ is the mean of the returns

- Covariance and correlation (btw 2 variables, R_1 and R_2)
  - → $Cov(R_1,R_2) = \frac{1}{T-1}\sum_{t=1}^{T} (R_1 - \bar{R}_1)(R_2 - \bar{R}_2)$
  - → $Corr(R_1,R_2) = Cov(R_1,R_2) / \sigma_1\sigma_2$



Both the covariance and the correlation measure how the two variables change together ("co-vary", "co-relate"). The correlation is easier to interpret, since it is restricted to lie between -1 and 1.

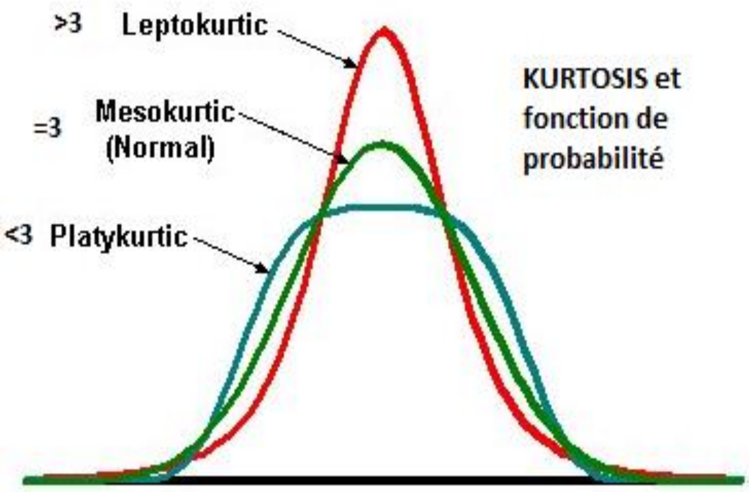
Descriptive Statistics: Skewness and Kurtosis



=> If the distribution is symmetric around the mean, S=0

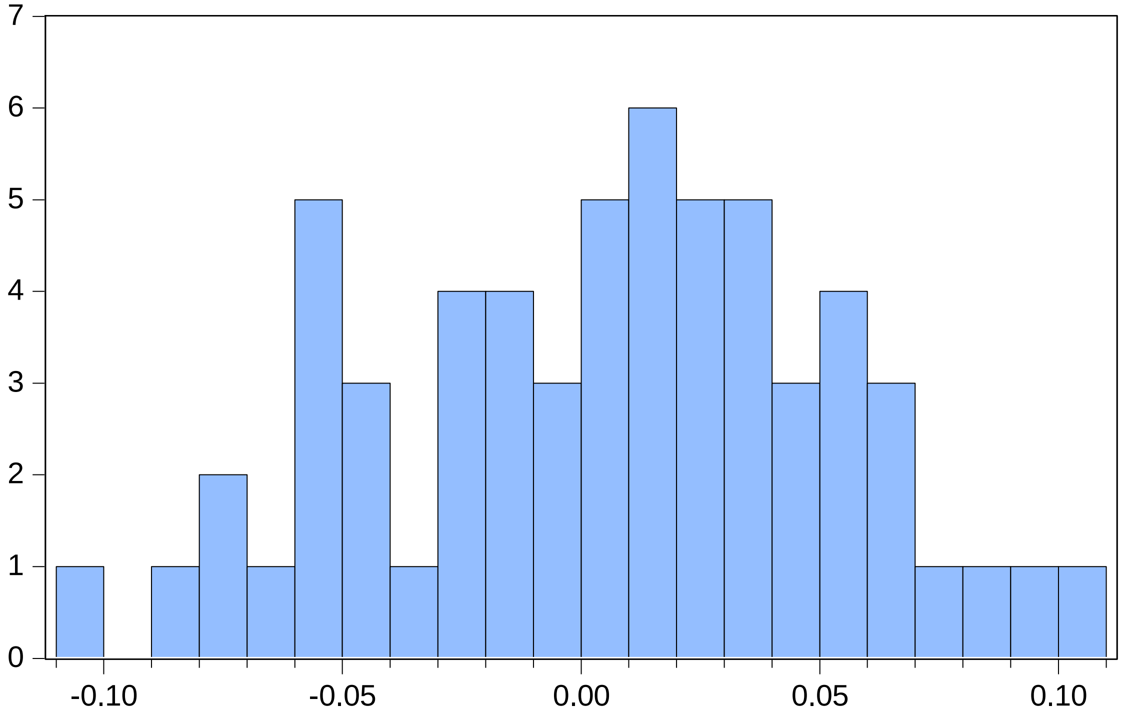
Kurtosis allows to detect extreme values

- **K-3>0**: *leptokurtic* distribution (Heavier tails and a higher peak)
  => presence of extreme values
- **K-3<0**: *platikurtic* distribution (Lighter tails and a lower peak)
  => very few extreme values
- **K-3= 0**: Normal distribution
- Kurtosis vs excess kurtosis (K-3)



Descriptive Statistics

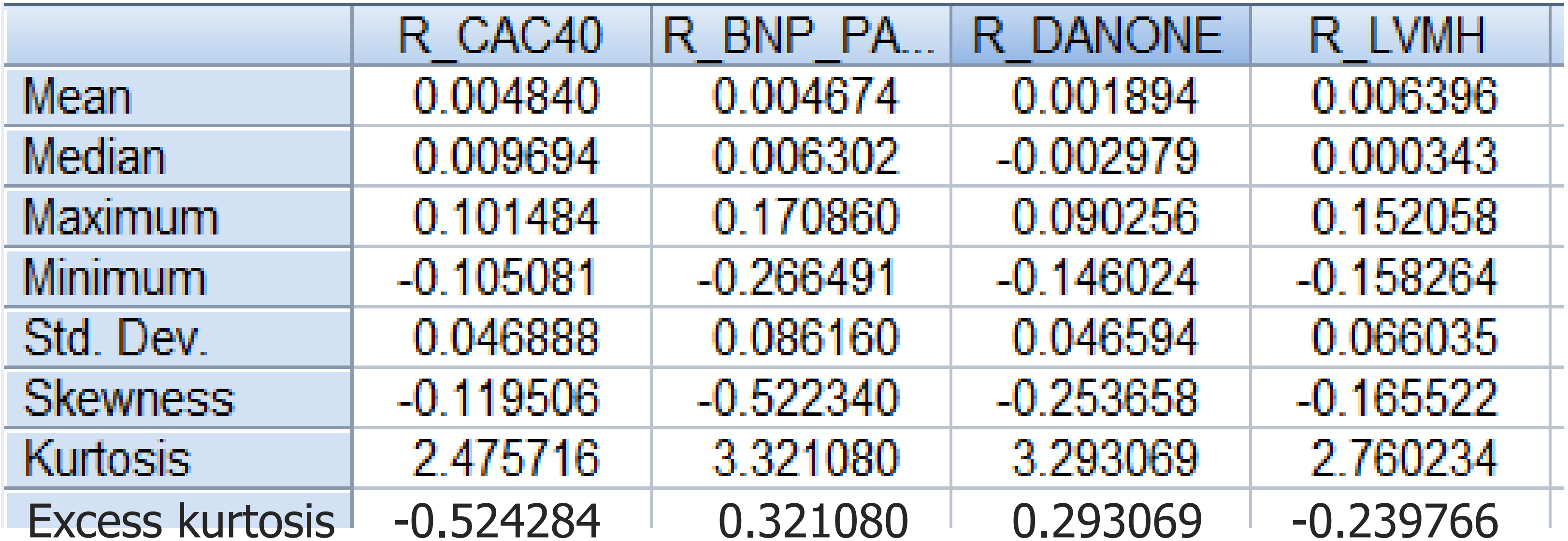
Monthly returns of the CAC40 from March 2012 to March 2017



```
Series: R_CAC40
Sample 2012M03 2017M03
Observations 60

Mean              0.004840
Median            0.009694
Maximum           0.101484
Minimum          -0.105081
Std. Dev.         0.046888
Skewness         -0.119506
Kurtosis          2.475716
Excess kurtosis  -0.524284
```

Descriptive Statistics



	R_CAC40	R_BNP_PA...	R_DANONE	R_LVMH
Mean	0.004840	0.004674	0.001894	0.006396
Median	0.009694	0.006302	-0.002979	0.000343
Maximum	0.101484	0.170860	0.090256	0.152058
Minimum	-0.105081	-0.266491	-0.146024	-0.158264
Std. Dev.	0.046888	0.086160	0.046594	0.066035
Skewness	-0.119506	-0.522340	-0.253658	-0.165522
Kurtosis	2.475716	3.321080	3.293069	2.760234
Excess kurtosis	-0.524284	0.321080	0.293069	-0.239766

Question 1 -The riskiest equity is :

A-R_CAC40

B-R_BNP_PA

C-R_DANONE

D-R_LVMH

Question 2

A-50% of the CAC_40 returns are lower than 0,48%

B-50% of the CAC_40 returns are lower than 0,96%

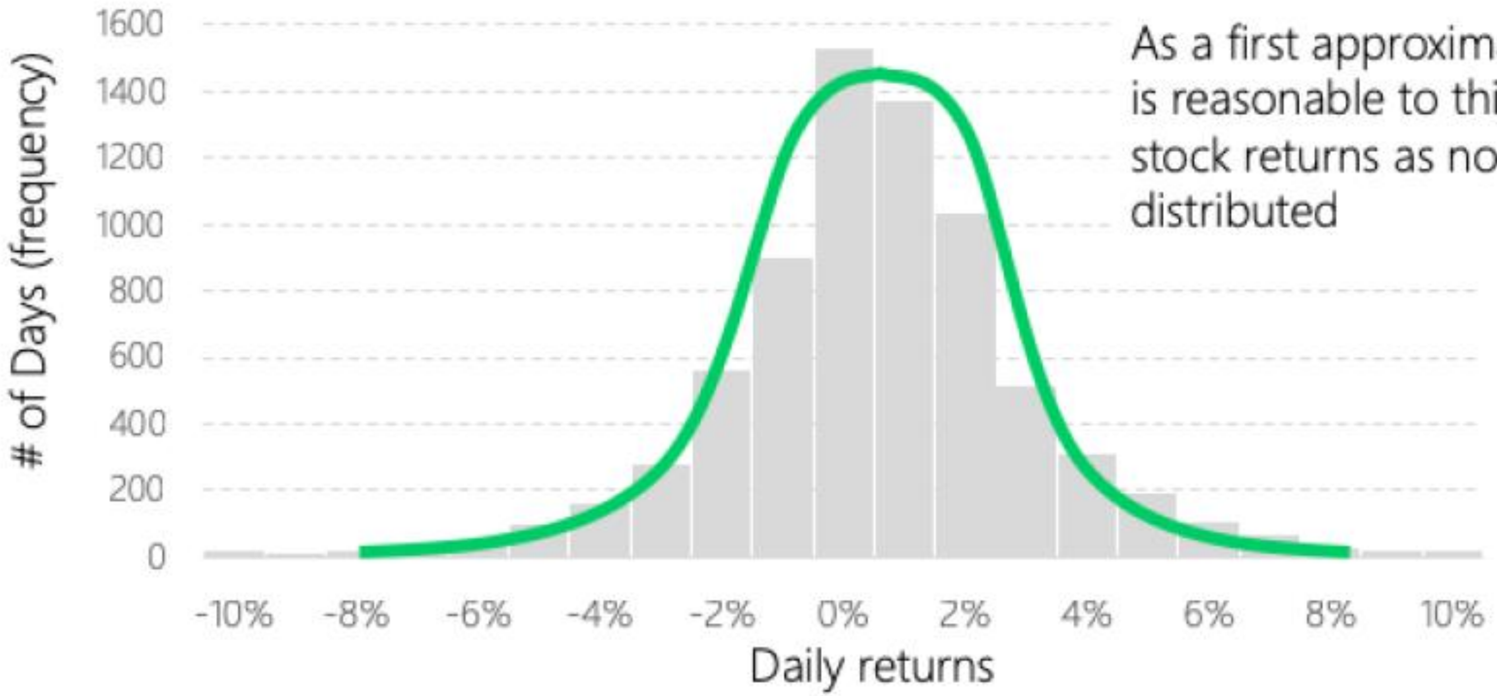
C-R_CAC40 is right skewed

D-R_CAC-40 is leptokurtic

Normality

Returns and the normal distribution

Returns and the normal distribution (= loi normale): **Apple Inc** – daily returns 1990-2019



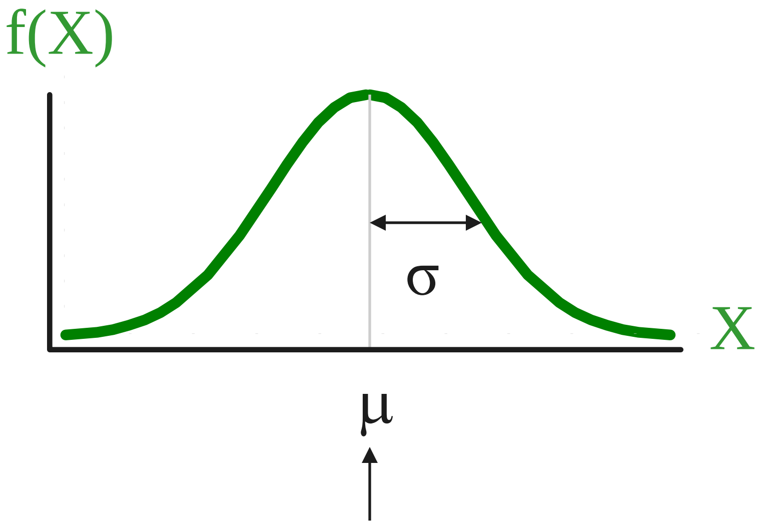
As a first approximation, it is reasonable to think of stock returns as normally distributed

The Normal Distribution

Bell Shaped

- Symmetrical (Skewness=0)
- Kurtosis =3
- Mean, Median and Mode are Equal

- Location is determined by the mean, μ (changing μ shifts the distributions left or right)
- Spread is determined by the standard deviation, σ
- The random variable has an infinite theoretical range: – ∞ to + ∞



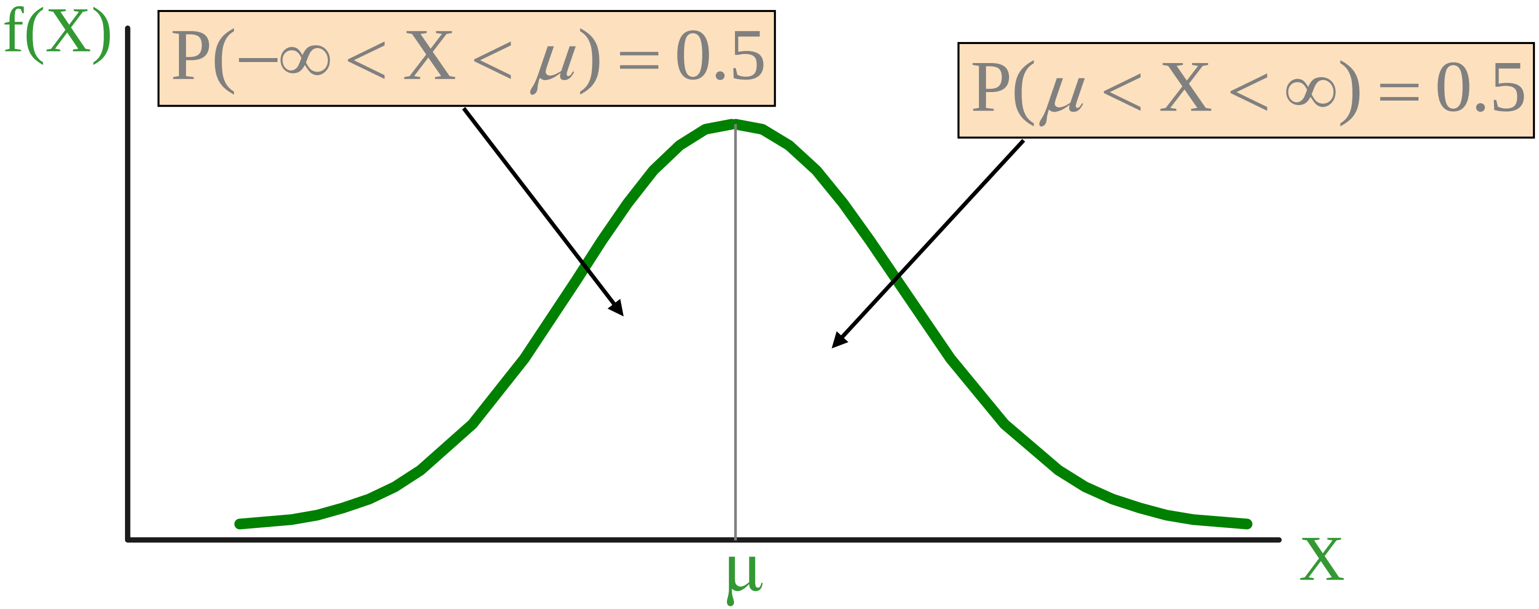
Mean
= Median
= Mode

Definition: random variable = a variable that can take on any value from a given set. Most commonly used distribution to characterize a random variable is a normal distribution.

Normal probabilities

Probability is measured by the area under the curve.

The total area under the curve is 1.0, and the curve is symmetric, so half is above the mean, half is below



$P(-\infty < X < \mu) = 0.5$

$P(\mu < X < \infty) = 0.5$

$P(-\infty < X < \infty) = 1.0$

Why normality is important?

- The probability law is known ➔ calculus of probability

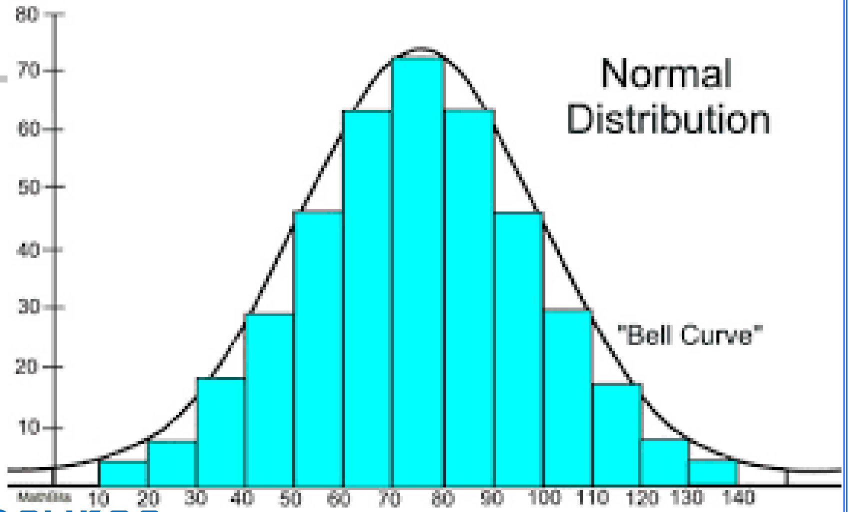
What is the probability that the returns be positive?

- No outliers ➔ Easier to model
- Mean-variance analysis and CAPM
- Specific calculus (VaR)

BUT… Not all continuous random variables are normally distributed

➔ It is important to evaluate how well the data set is approximated by a normal distribution

Assessing Normality



Construct charts or graphs

- ➔ Does the histogram appear bell-shaped?
- ➔ Is the normal probability plot approximately linear with positive slope?

Compute descriptive summary measures

- ➔ Do the mean, median and mode have similar values?
- ➔ Is the Skewness close to 0? Is the Kurtosis close to 3?

Jarque Bera Normality Test:

Based on values of skewness and excess kurtosis

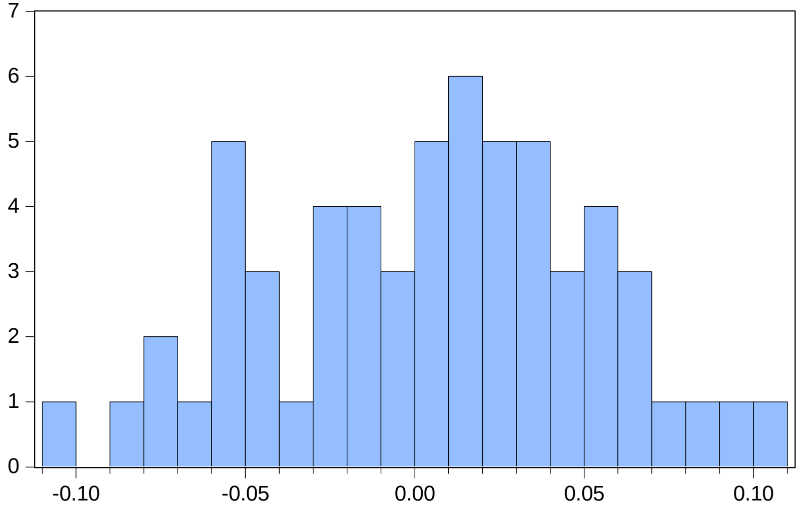
H0 : the series is normally distributed (S and EK jointly not different from 0)

H1 : the series is not normally distributed

JB ~ χ^2(2 dof)

- ➔ Reject H0 if JB > $\chi^2_{2;\alpha}$ or if pvalue < α (the test is statistically significant)

Assessing Normality



```
Series: R_CAC40
Sample 2012M03 2017M03
Observations 60

Mean            0.004840
Median          0.009694
Maximum         0.101484
Minimum        -0.105081
Std. Dev.       0.046888
Skewness       -0.119506
Kurtosis        2.475716
Excess kurtosis -0.524284
Jarque-Bera     0.830002
Probability     0.660340
```

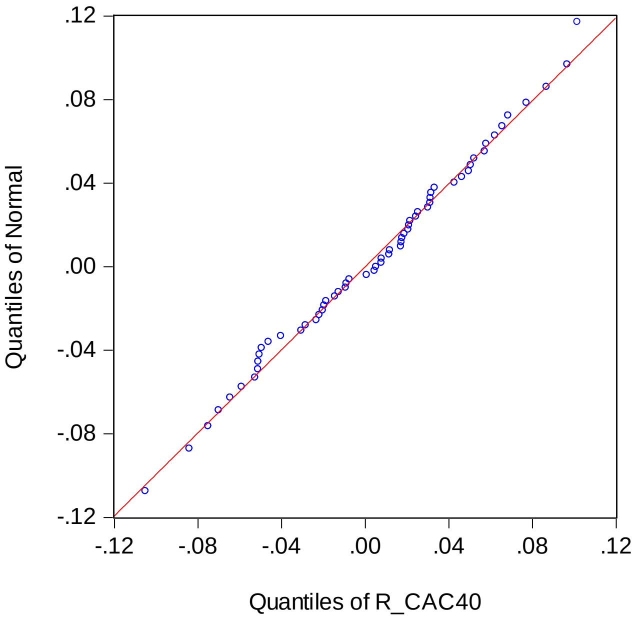
Question 3 –Are the CAC_40 returns normally distributed?

A-No, because the series is left skewed

B-Yes, because the Jarque Bera test is not significant

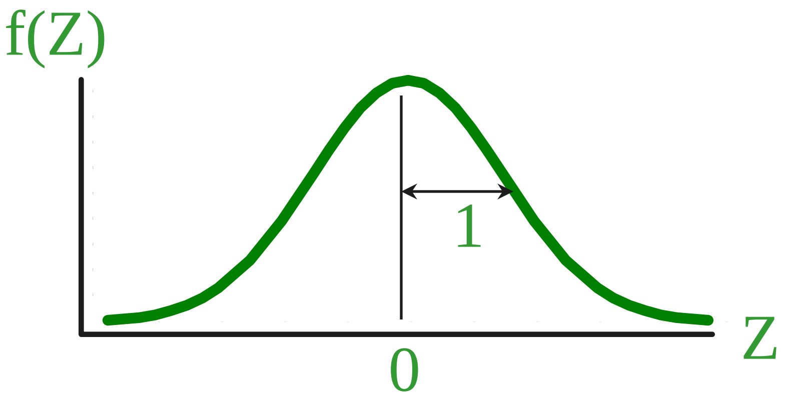
C-No, because the Jarque Bera test is not significant

D-No, because on the QQ plot, points are close to the bissector



The Standardized Normal Distribution

- Also known as the “Z” distribution
- Mean is 0
- Standard Deviation is 1



Values above the mean have positive Z-values

Values below the mean have negative Z-values

- To transform a normally distributed variable into a standard normal: subtract the mean and divide by the st. dev.: x~N(μ,σ) => (x-μ)/σ ~N(0,1)

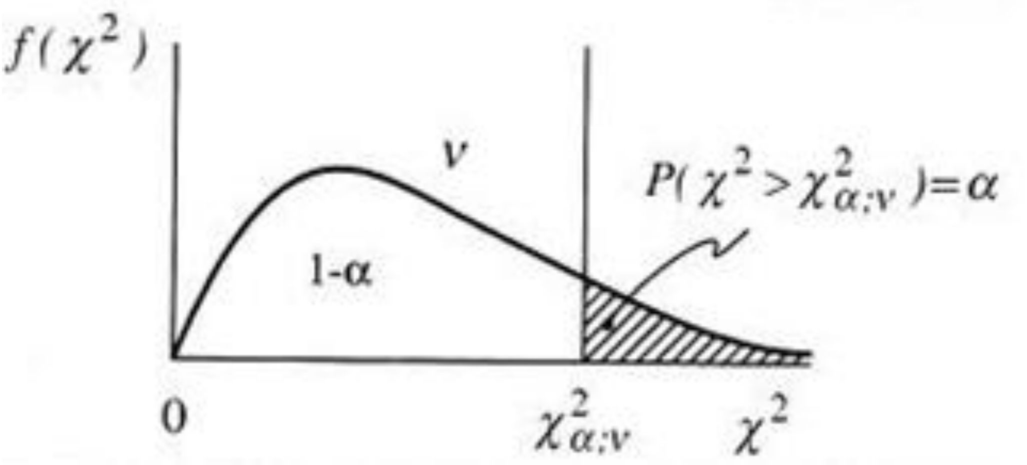
Chi square distribution

Definition:

Let (X_1,X_2 ,…,X_n) a sample with gaussian distribution N(0;1).

Then:

$$\sum_{i=1}^{n} X_i^{\ 2} \quad \text{is} \quad \chi^2(n)$$



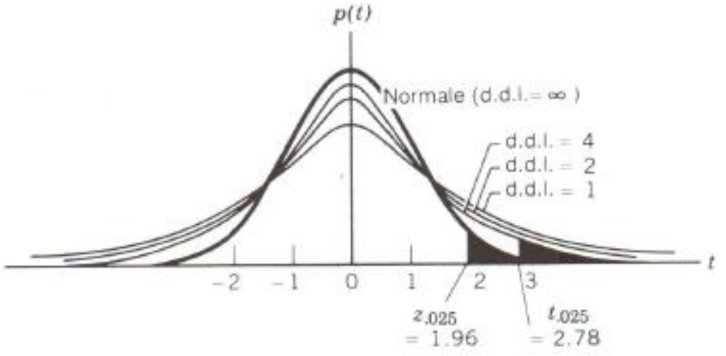
➔ Chi square distributed with n degrees of freedom (dof)

Student Distribution

Definition:

If X~N(0;1) and Y~χ^2(n) with X et Y independent variables, then:

$$t = \frac{X}{\sqrt{Y/n}} \quad \text{is} \quad t(n)$$



➔ Student distributed with n degrees of freedom (dof)

➔ When n>30 t(n)~N(0,1)

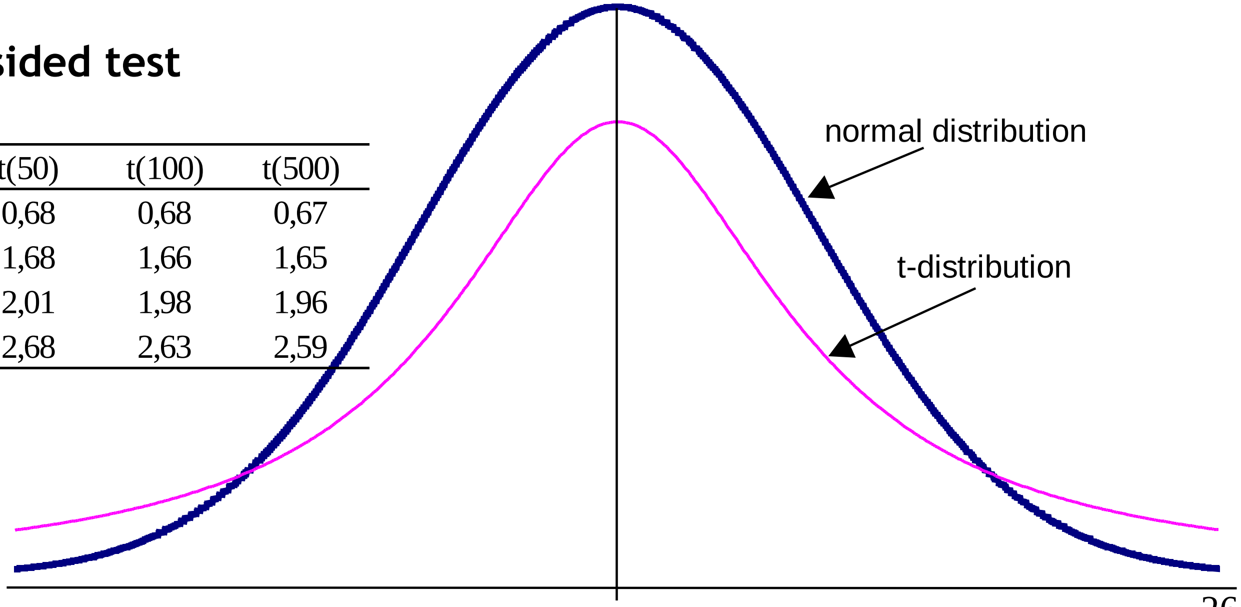
Normal and t-distribution

- *t*-distribution with an infinite number dof ≈ N(0;1)
- t- and the standard normal distribution : both are symmetrical and centred on zero. The t-distribution is characterized by another parameter: its degrees of freedom.

Value for $t_{\alpha/2}$: case of two-sided test

Example from statistical tables

Significance Level	N(0;1)	t(4)	t(50)	t(100)	t(500)
50%	0,67	0,74	0,68	0,68	0,67
10%	1,64	2,13	1,68	1,66	1,65
5%	1,96	2,78	2,01	1,98	1,96
1%	2,58	4,60	2,68	2,63	2,59

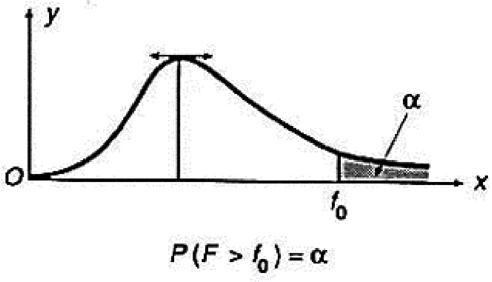


Fisher distribution

Definition:

Let X et Y two independent variables with X~χ^2(n) and Y~χ^2(p), then:

$$\frac{X/n}{Y/p}$$ is F(n;p)



➔ Fisher distributed with n and p degrees of freedom

TUTORIAL XLSTAT

1. Preliminary work on data
2. Descriptive Statistics

Tutorial

- Download data on Excel for the relevant variables:
  - ➔ Stock: Microsoft Corporation
  - ➔ Index: S&P 500
  - ➔ Risk free: Treasury Bill 3 Months

- Create times series of returns and excess returns
- Plot the series of prices and returns

(save as Excel Macro-Enabled Workbook)

Tutorial

- Using software XLSTAT (can be downloaded from Campus):

- **Compute the standard descriptive statistics** (mean, min, max, standard-deviation, skewness, excess kurtosis

- **Histogram**

- **Normality test (JB)**

=> Comments on the returns distributions? Are there outliers (extreme values)? Are the returns normally distributed?