Statistical Analysis of Financial Data
January 2017 – Session 01

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, January 16, 2017

Statistical Analysis of Financial Data
January 2017 – Session 01

Your Lecturer

Name: **Marcel Dettling**

Age: 42 Years

Civil Status: Married, 2 children

Education: Dr. Math. ETH

Position: **Lecturer** @ ETH Zürich and @ ZHAW
Researcher in Applied Statistics @ ZHAW

Connection: Research with industry: *hedge funds, insurance, …*
Academic research: *high-frequency financial data*

Statistical Analysis of Financial Data
January 2017 – Session 01

Topics of the Course

Session 01:

Financial data and their properties
Random Walk model with various distributions

Session 02:

The GARCH model for conditional heteroskedasticity
Risk measures for dealing with financial loss

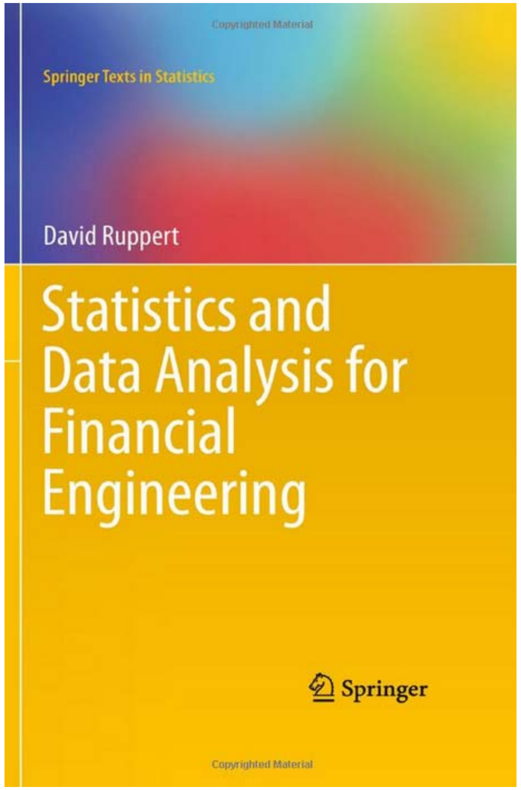
Sessions 03/04/05:

Multivariate Analysis: CAPM, Copulas, Factor Models

Statistical Analysis of Financial Data
January 2017 – Session 01

Resources

There is a book which is predominantly used as a guideline for the topics covered during the course:



Session 01: Chapters 2/4/5

Session 02: Chapters 18/19

Session 03: Chapters 11/7/8 ???

Session 04: Chapters 16/17 ???

Session 05: Chapters ???

→ *There is no urgent need to buy this book, notes/slides are sufficient.*

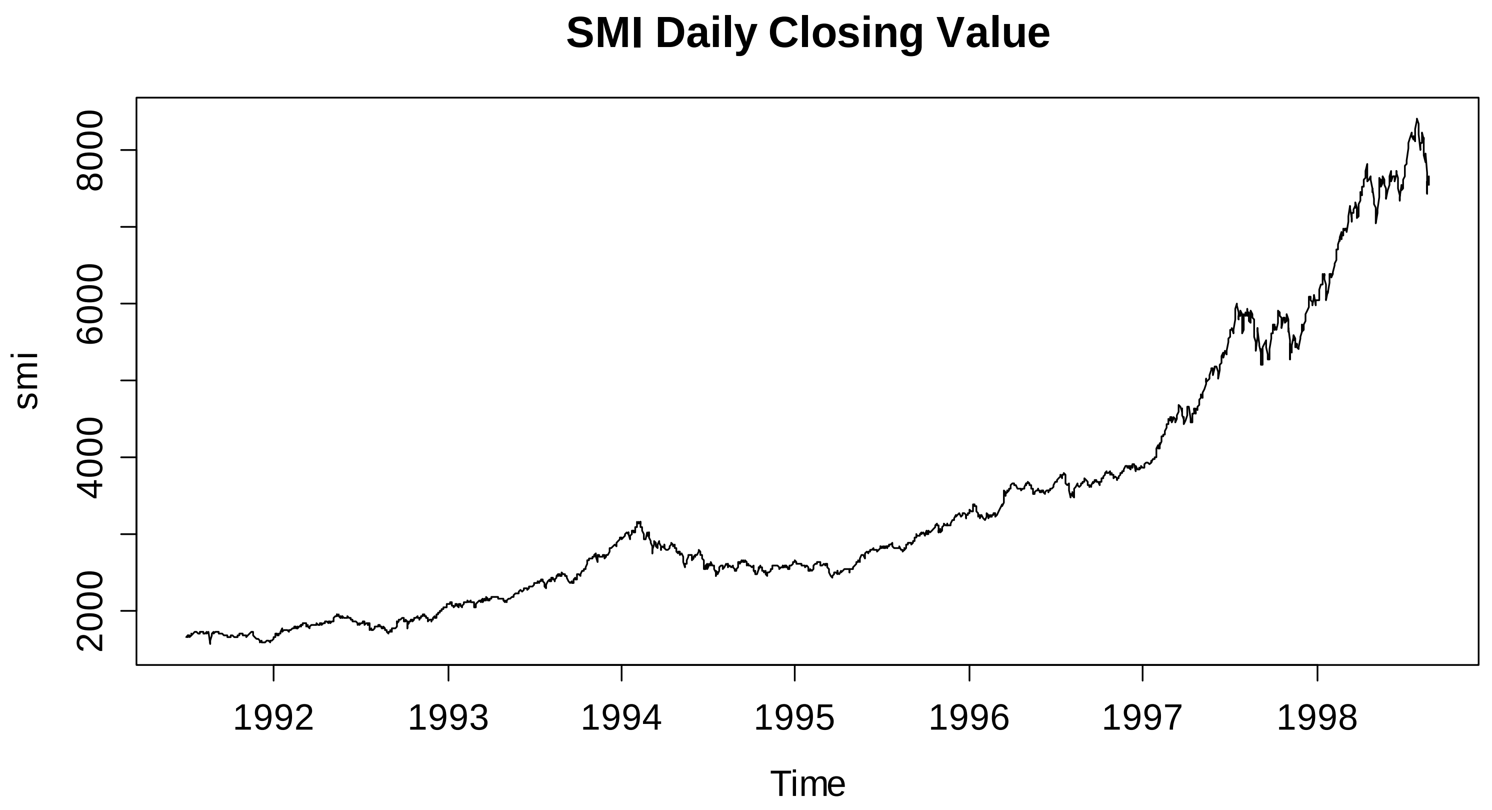
Example 1: Swiss Market Index

In R, we have daily values of the SMI over 8 years:

```
> data(EuStockMarkets)
> EuStockMarkets
Time Series:
Start = c(1991, 130)
End = c(1998, 169)
Frequency = 260
                 DAX    SMI    CAC   FTSE
1991.496 1628.75 1678.1 1772.8 2443.6
1991.500 1613.63 1688.5 1750.5 2460.2
1991.504 1606.51 1678.6 1718.0 2448.2
1991.508 1621.04 1684.1 1708.1 2470.4
1991.512 1618.16 1686.6 1723.1 2484.7
1991.515 1610.61 1671.6 1714.3 2466.8
```

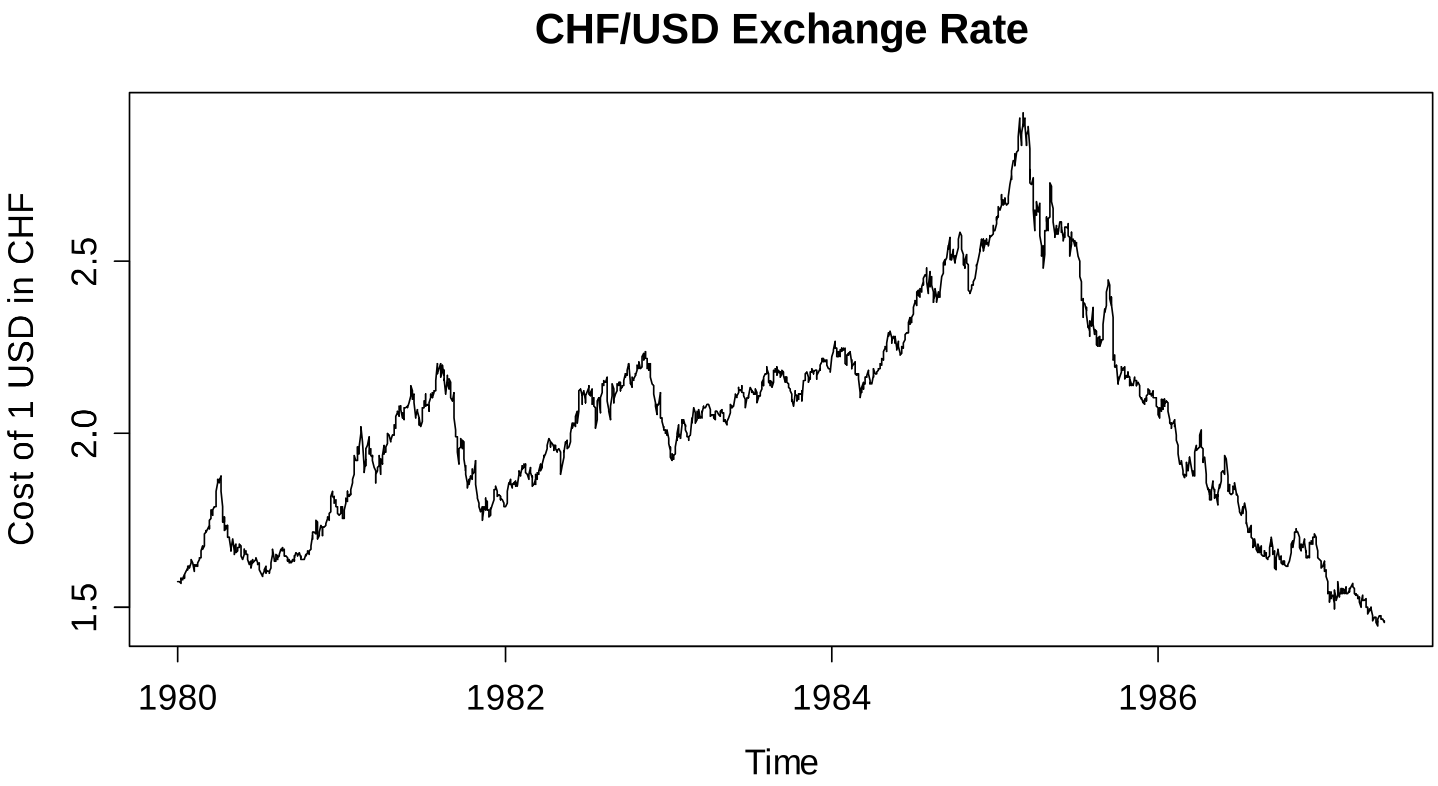
Example 1: Swiss Market Index

```
> smi <- ts(tmp, start=start(esm), freq=frequency(esm))
> plot(smi, main="SMI Daily Closing Value")
```



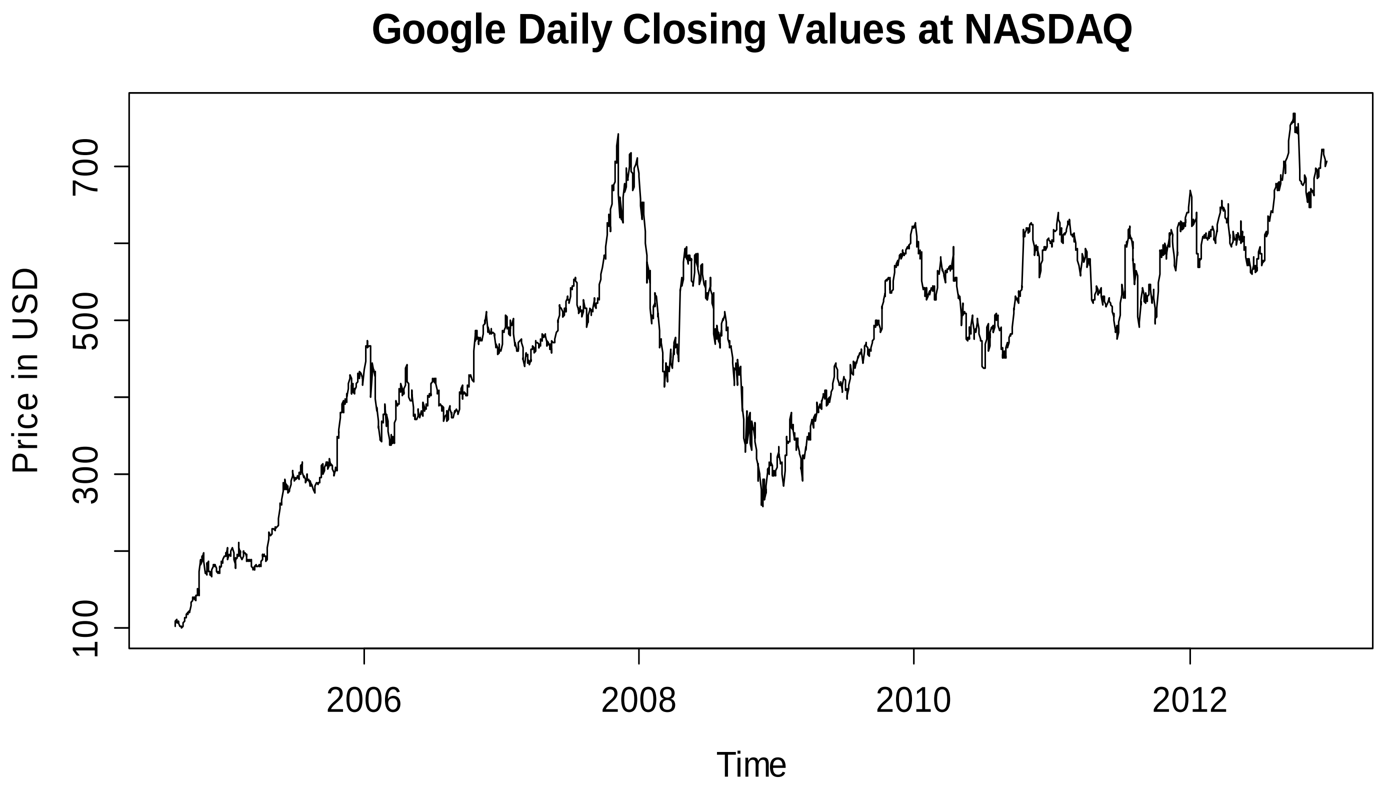
Example 2: CHF/USD Exchange Rate

```
> library(Ecdat)
> data(Garch); chf.usd <- ts(1/Garch$sf)
```



Example 3: Google Stock

Data taken from NASDAQ:
http://www.nasdaq.com/symbol/goog/historical



Statistical Analysis of Financial Data
January 2017 – Session 01

Introduction: What is a Time Series?

Time series process:

A set of random variables $\left\{X_t,\, t \in T\right\}$, where T is the set of time at which the process was (or can be) observed. We restrict ourselves cases where the set of times is discrete and finite. Also, the observations were made at fixed time intervals.

Observed time series:

An observed time series $\left\{x_t,\, t \in T\right\}$ is one single realization of the time series process. If we want to do statistics with it, there is no alternative than to assume additional structure.

Stationarity

For being able to do statistics with time series, we require that the series "doesn't change its probabilistic character" over time. This is mathematically formulated by **strict stationarity**.

Def: A time series $\{X_t, t \in T\}$ is strictly stationary, if the joint distribution of the random vector $(X_t, \ldots, X_{t+k})$ is equal to the one of $(X_s, \ldots, X_{s+k})$ for all combinations of t, s and k.

➔

$X_t \sim F$	all X_t are identically distributed
$E[X_t] = \mu$	all X_t have identical expected value
$Var(X_t) = \sigma^2$	all X_t have identical variance
$Cov(X_t, X_{t+h}) = \gamma_h$	the autocov depends only on the lag h

Simple Returns and Log Returns

→ Asset price time series are typically non-stationary!

If P_t is the price of an asset, we could consider **simple returns**:

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

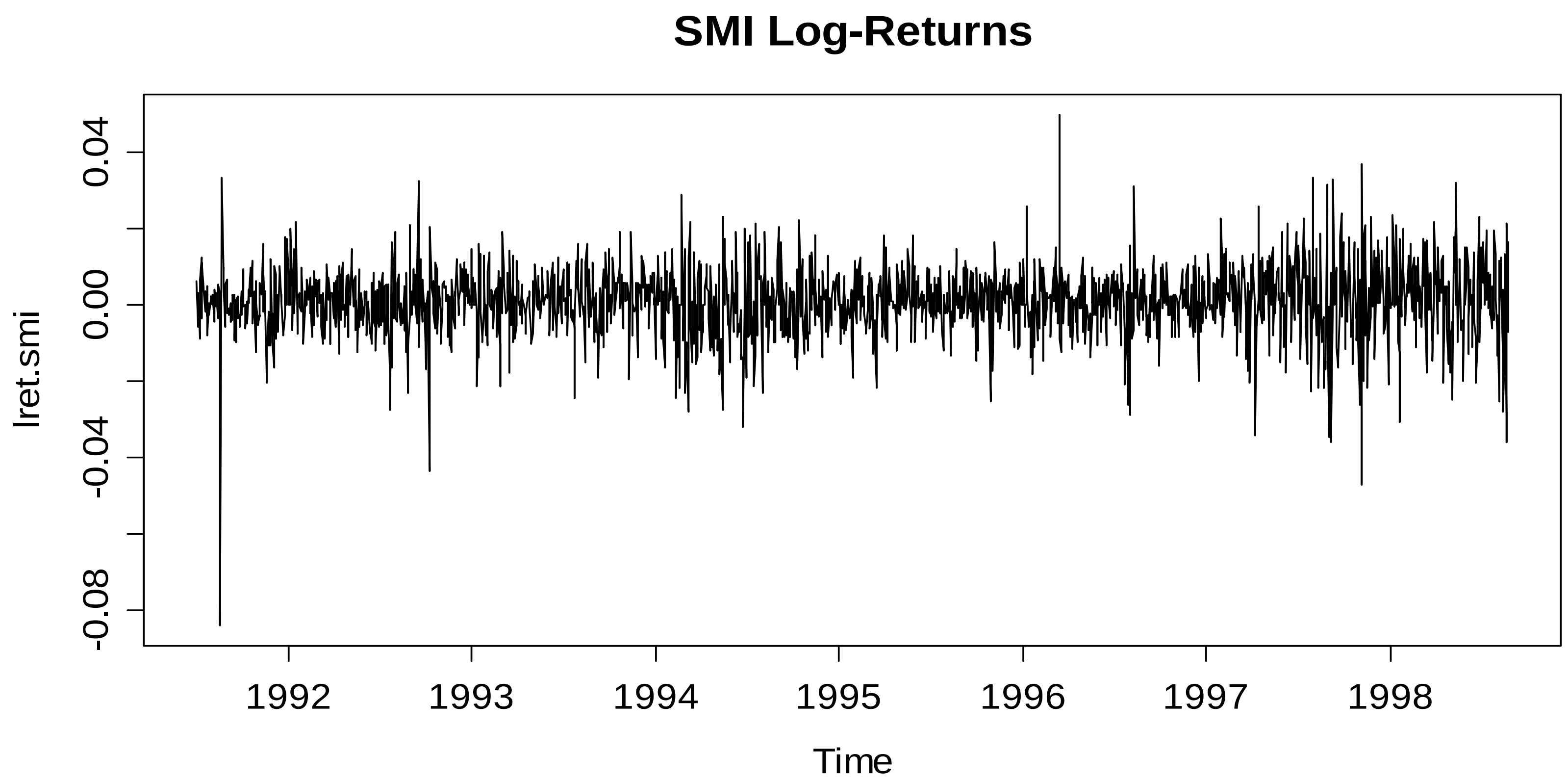
But instead, we prefer to work with **log returns**:

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

Example: **see next slide...**

Example: Log Returns of SMI

```
> lr.smi    <- diff(log(smi))
> plot(lr.smi, main="SMI Log-Returns")
```



Why Log Returns?

What is the rationale for working with log-returns?

- **SOP:** The prices are right-skewed and have a trend. Thus, we must log-transform and then take differences at lag 1.
- **Symmetry:** The minimum simple return is -100%, while the maximum increase is infinite. Log returns are symmetric.
- **Compounding:** The multi-period log returns are additive, i.e. are the sum over the single period log returns in that period.
- **Compatibility:** With the Random Walk model, see below.

Goals in Financial Data Analysis

What can we do and what can't we do?

- The prices P_t are non-stationary and thus usually not suitable for statistical analysis. That is why we lay focus on log returns.

- Empirical evidence and several economic theories suggest:

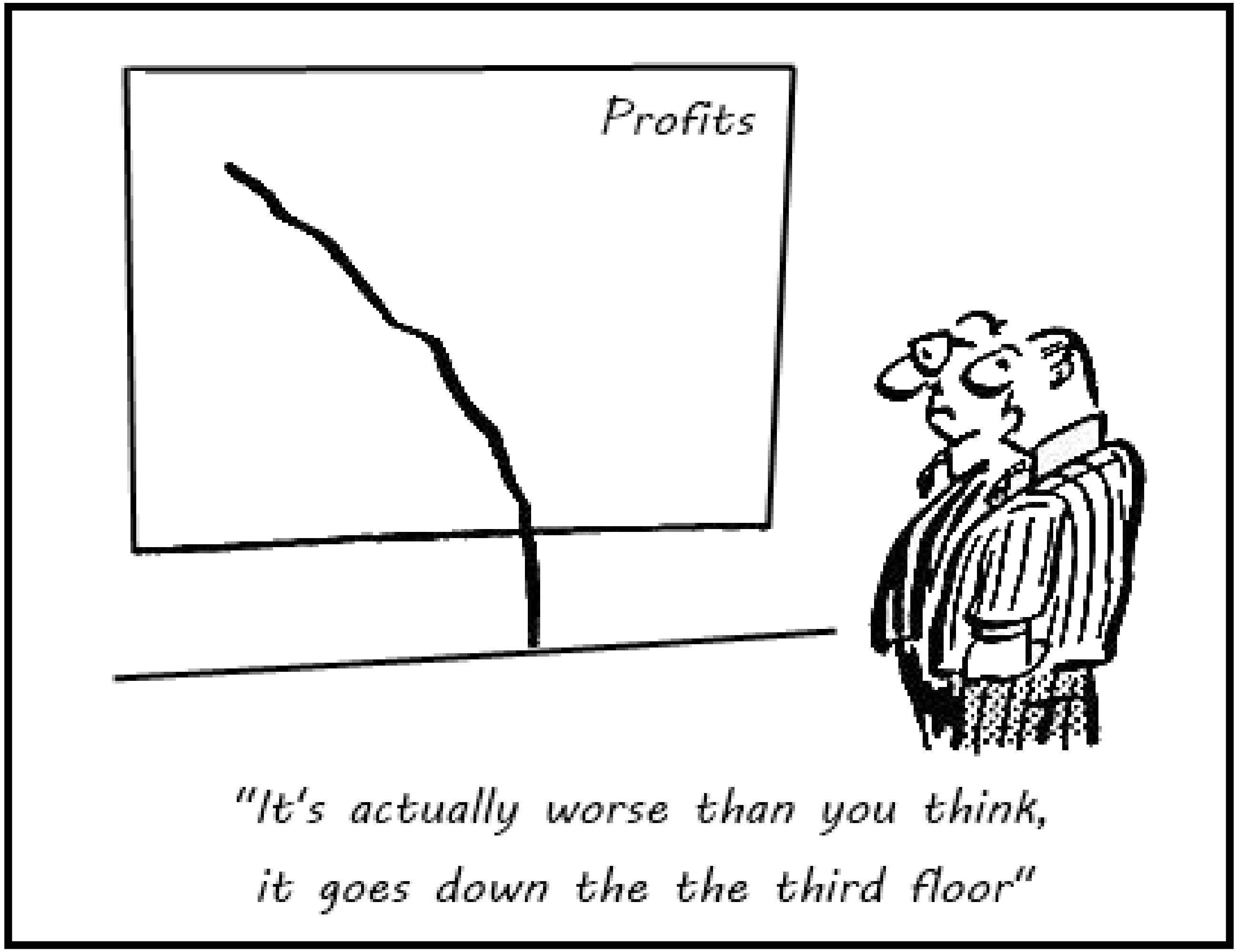
  $$E[r_t] \approx E[r_t \mid r_{t-1}, r_{t-2}, ...] = \mu \approx 0$$

  It is not realistic to make statements whether tomorrows return will be positive or negative, whether it is a good moment to invest in an asset, et cetera.

- There are aspects of the distributions of r_t and $r_t \mid r_{t-1}, r_{t-2}, ...$ which are interesting to study: *shape, scale, skewness, kurtosis, quantiles, tail distributions*, et cetera.

Statistical Analysis of Financial Data
January 2017 – Session 01

Goals in Financial Data Analysis



The Random Walk Model

From the compunding property of log returns, we derive:

$$r_{k,t} = r_{1,t} + ... + r_{1,t-k+1}, \text{ where } k = \text{horizon and } t = \text{time}$$

Assuming **normal returns** $r_{1,t} \sim N(\mu, \sigma^2)$ and **independence**:

$$r_{k,t} \sim N(k\mu, k\sigma^2)$$

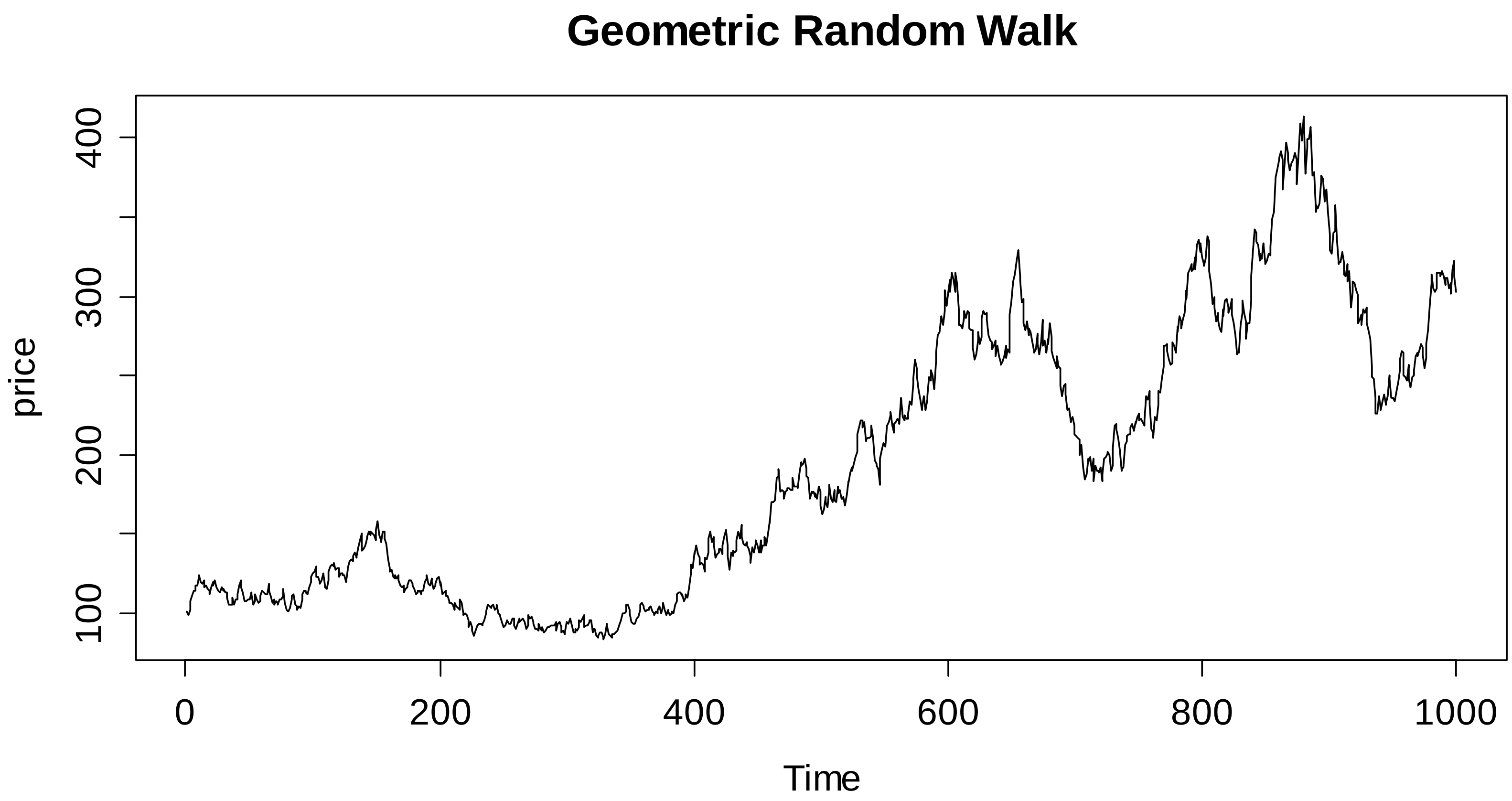
And the risk management problem would be solved. We can even derive the **price process** and its **distribution**:

$$P_t = P_0 \cdot \exp(r_{1,t} + ... + r_{1,1})$$

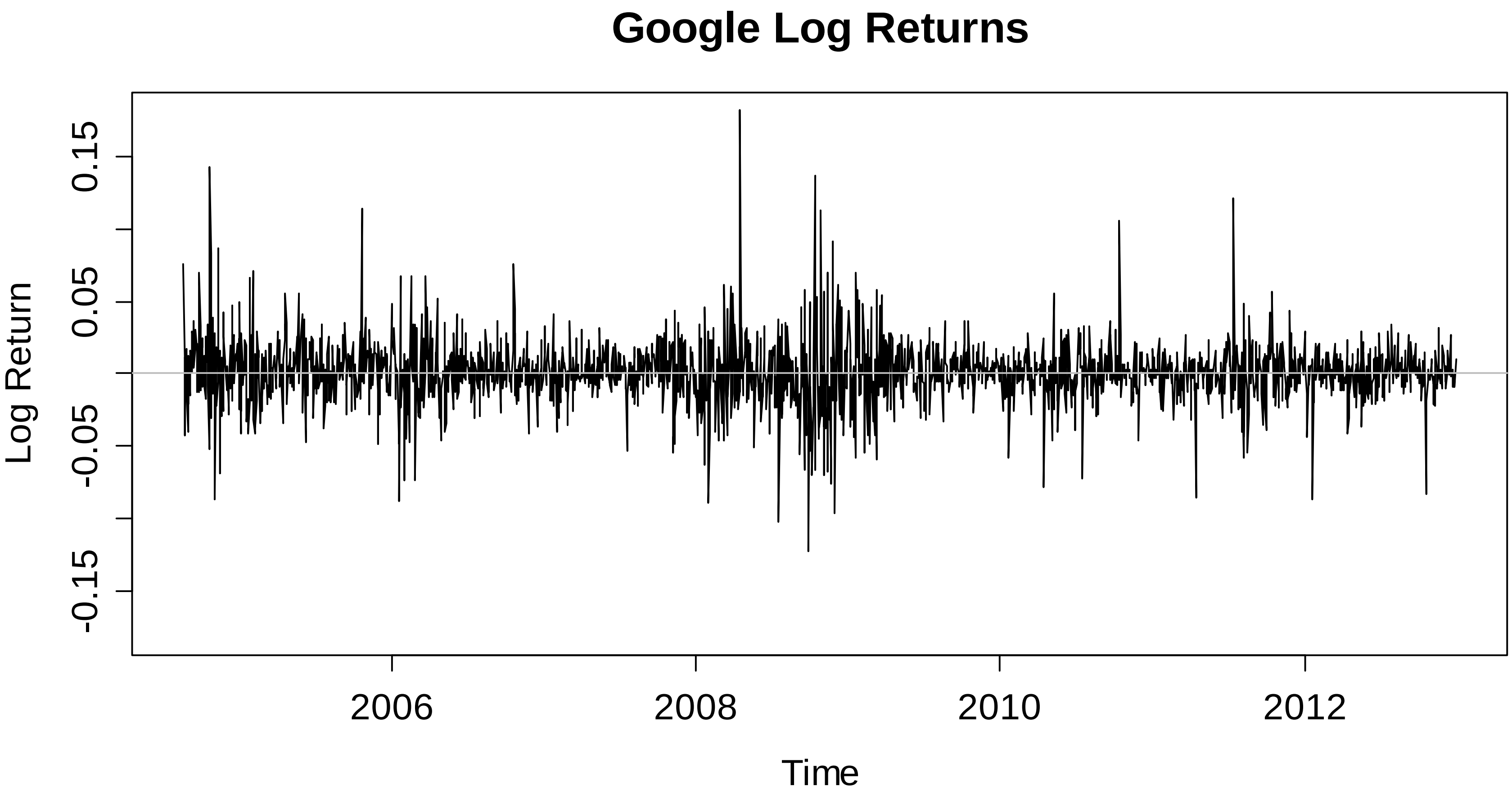
This is a *Geometric Random Walk*. The prices have a lognormal distribution at all times t.

Simulation Example

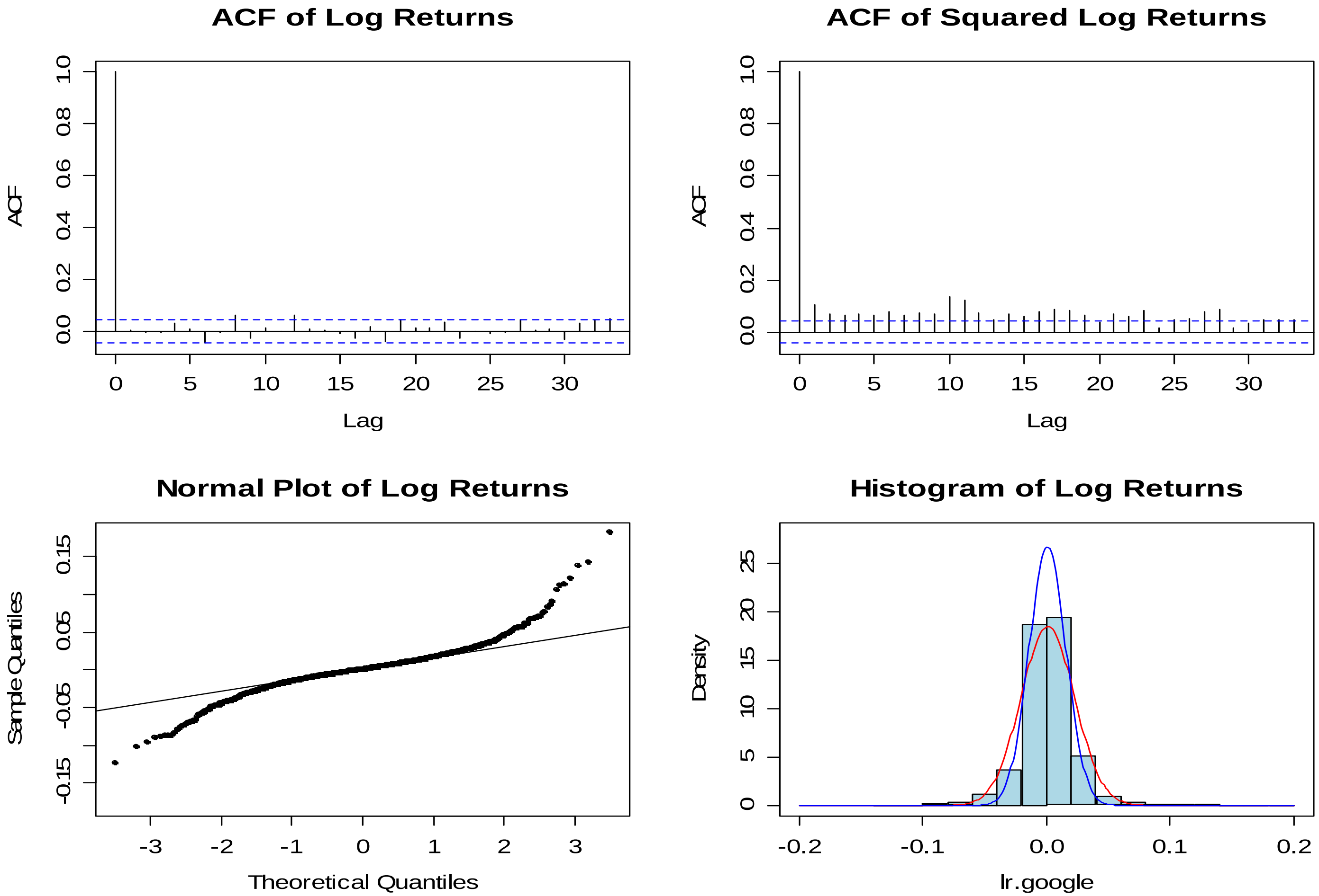
Let $P_0 = 100$ and $r_{1,t} \sim N(\mu = 0, \sigma^2 = 0.03^2)$: Realistic?!?



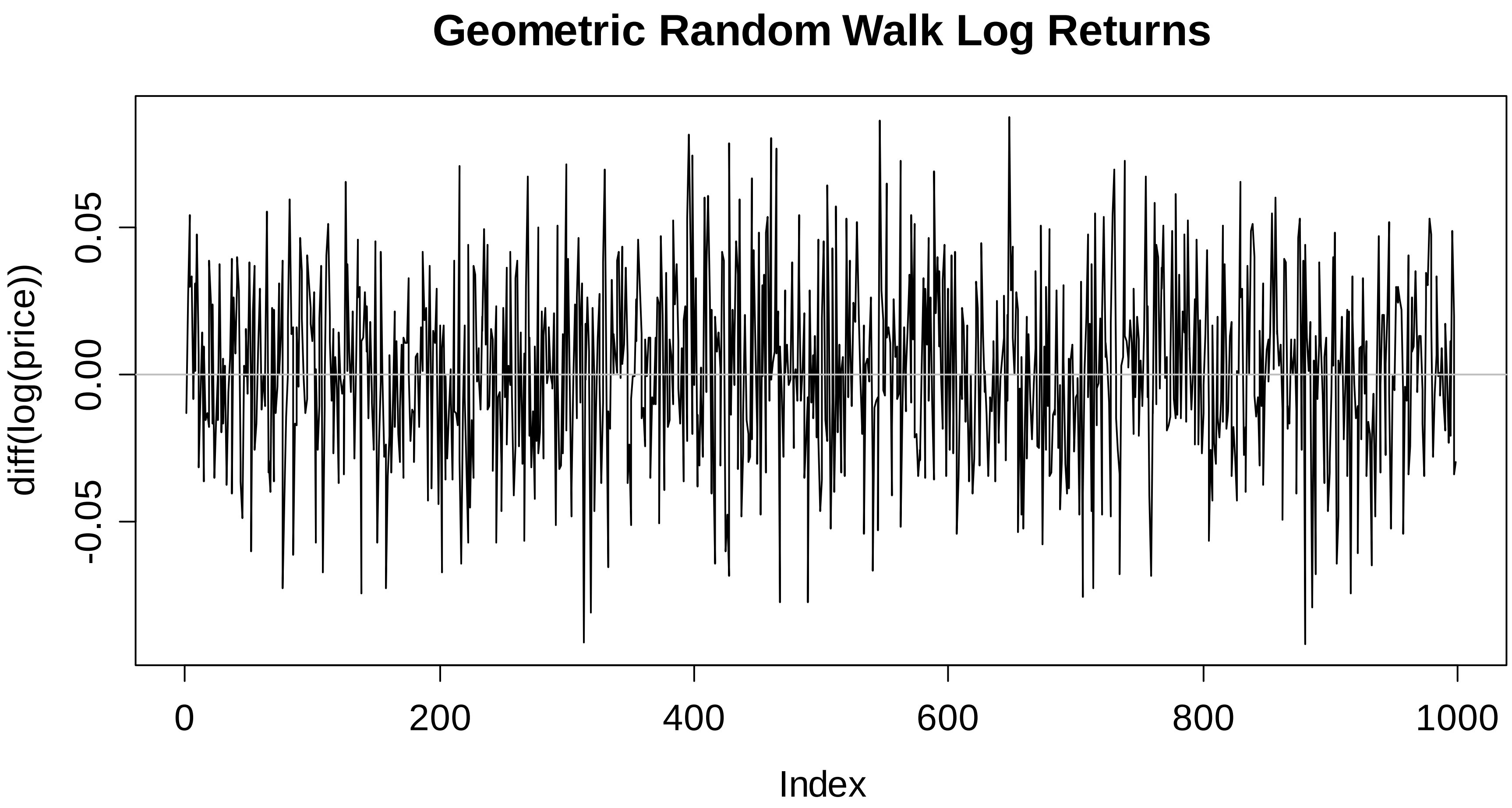
Analyzing Log Returns: Google



Analyzing Log Returns: Google



Analyzing Log Returns: Simulation



Properties of Log Returns

Practice shows that the following are always present:

- A mean close to zero
- Hardly any direct autocorrelation
- Clusters of volatility (high/low changes)
- Correlations among the squared returns
- Some extreme returns, longer tails than normal
- Stationarity! At least we will operate under this assertion

→ Log returns are not Gaussian. And while they are not directly correlated, they are still not independent / White Noise. Good models need to take that into account! The RanWalk does not!

Skewness and Kurtosis

In regular statistics, we seldom go beyond mean and variance. In financial statistics, there is interest in the third and fourth moment:

Skewness:

$$Skew = \frac{E[(X-\mu)^3]}{\sigma^3}, \text{ resp. } \hat{S}kew = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\hat{\sigma}}\right)^3$$

In R:

```
> library(timeDate)
> skewness(lr.google)
[1] 0.4340404
attr(,"method")
```

Skewness and Kurtosis

In regular statistics, we seldom go beyond mean and variance. In financial statistics, there is interest in the third and fourth moment:

Kurtosis:

$$Kurt = \frac{E[(X-\mu)^4]}{\sigma^4}, \text{ resp. } \hat{K}urt_{Ex} = -3 + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i - \overline{x})}{\sigma}\right)^4$$

In R:

```
> library(timeDate)
> kurtosis(lr.google)
[1] 7.518994
attr(,"method")
[1] "excess"
```

Testing Normality

It is usual to evaluate the log return distribution *de visu* using a Normal Plot. An experienced eye detects non-normality easily.

Jarque-Bera Test:

Tests the null hypothesis of a Gaussian distribution by comparing skewness and kurtosis to 0 and 3, respectively:

$$JB = \frac{n}{24}\left(4 \cdot \hat{S}kew^2 + \hat{K}urt_{Ex}^2\right) \sim \chi_2^2$$

In R:

```
> library(tseries)
> jarque.bera.test(lr.google)
> X-squared = 5040.39, p-value < 2.2e-16
```

Heavy-Tailed Distributions

Idea: *Use a heavy tailed distribution for the Random Walk model*

Most popular choice: t_ν - distribution

Take $Z \sim N(0,1)$ and $W \sim \chi^2_\nu$: $T = \sqrt{\nu} \cdot \frac{Z}{W} \sim t_\nu$

The parameter ν is called degrees of freedom and controls the shape of the distribution. It can take any positive real value. The smaller it is, the heavier the tails of the distribution are.

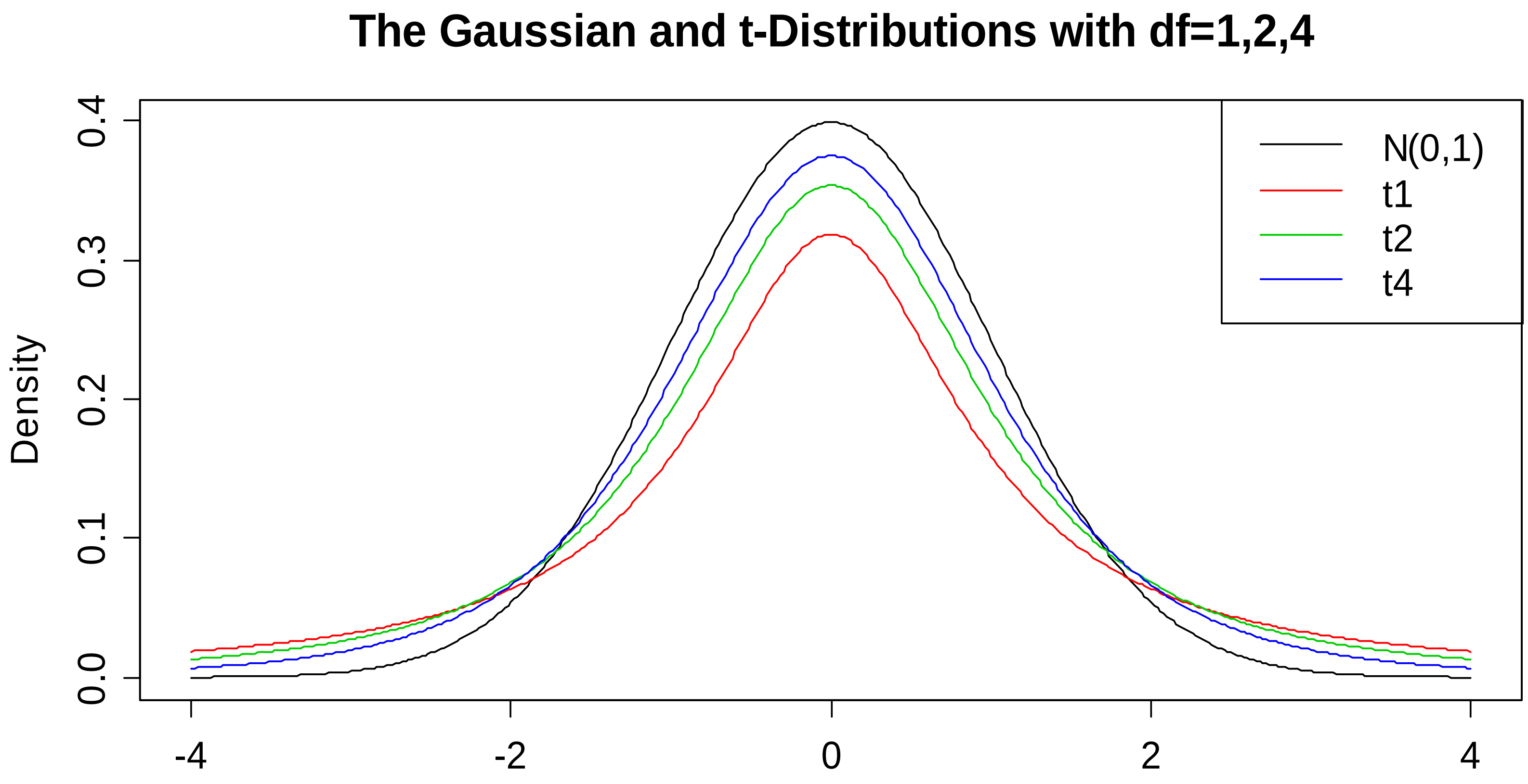
Also: $E[T] = 0$, exists if $\nu > 1$

$Var(T) = \nu / (\nu - 2)$, exists if $\nu > 2$

The third, fourth, fifth, … moment exist if $\nu > 3, 4, 5, \ldots$

Statistical Analysis of Financial Data
January 2017 – Session 01

The t-distribution



Enhancing with Location and Scale

While it seems that a t_ν-distribution can adapt well to financial log returns, that won't work well without location/scale parameters.

$S = \mu + \lambda T$ has a $t_\nu(\mu, \lambda^2)$-distribution with:
$E[S] = \mu$ and $Var(S) = \lambda^2 \cdot \nu / (\nu - 2)$

Important:

The tail behaviour remains the same, even if we add a location and a scale parameter. The decay is of polynomial order:

$f_{t_\nu(\mu,\lambda^2)}(x)$ goes to zero like $x^{-(\nu+1)}$ if $x \to \infty$

Note: The Gaussian tail decays exponentially with $\exp(-x^2)$.

Mixture Distributions

Goal: Mixture between 90% $N(0,1)$ and 10% $N(0,25)$

The probability density function can be written as:

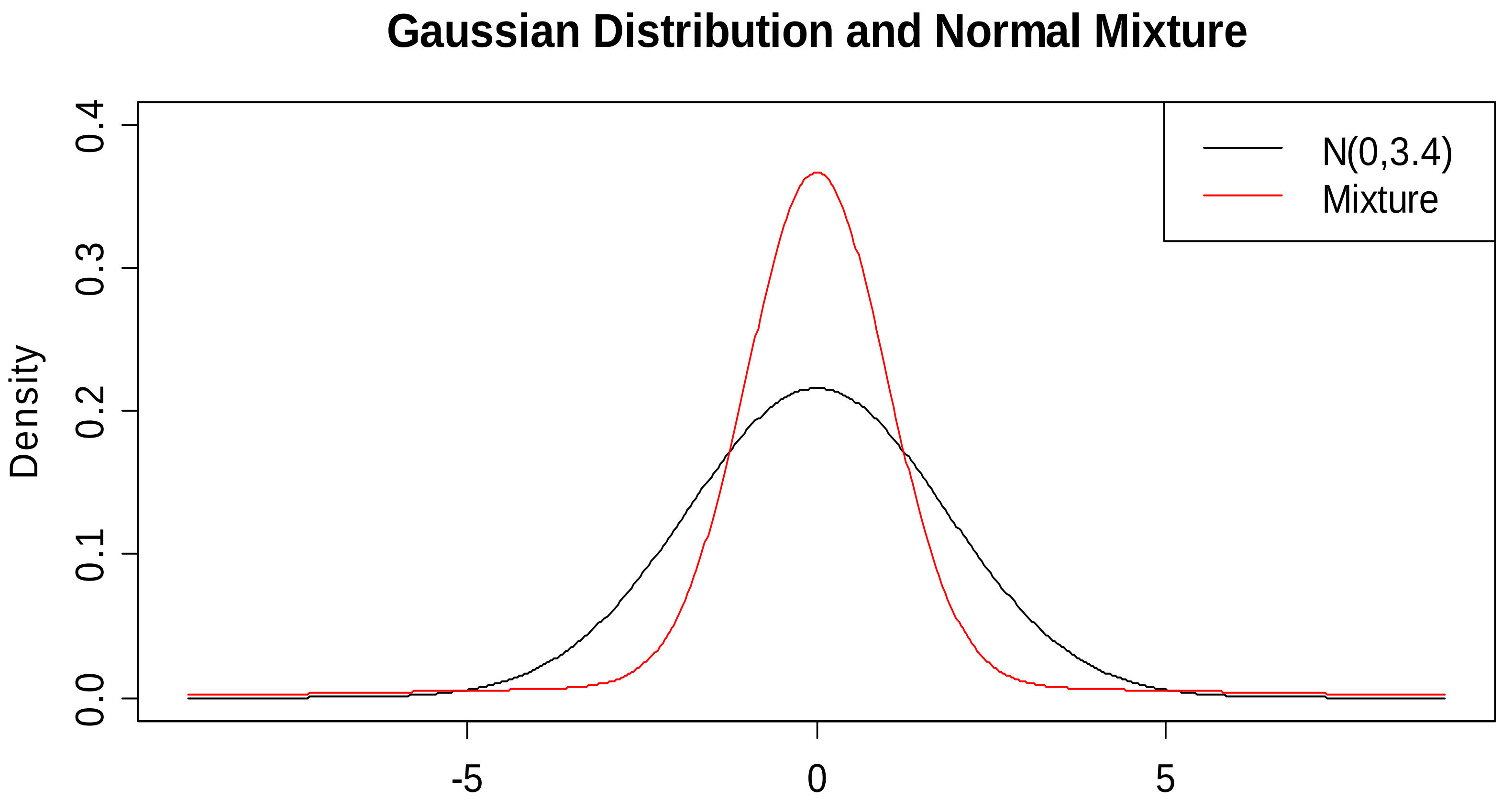
$$f_{mix}(x) = 0.9 \cdot f_{N(0,1)}(x) + 0.1 \cdot f_{N(0,25)}(x)$$

We can draw random variates of this distribution using a two-step approach, where we first determine from which Gaussian we have to simulate. While the mean of the mixture will remain at zero, the variance is:

$$Var(M) = 0.9 \cdot 1 + 0.1 \cdot 25 = 3.4$$

However, the mixture has more tail mass than a $N(0,3.4)$!!!

Mixture Distributions: Example



Mixture Distributions: Results

Comparing the ratio of extreme events:

```
> gauss <- 2*pnorm(-3*sd, 0, sd)
> mixt  <- 2*0.9*pnorm(-3*sd)+2*0.1*pnorm(-3*sd,0,5)
> mixt/gauss
[1] 9.948061
```

An extreme event is 10x more likely with the mixture distribution, and the kurtosis is 16.45. Does it mean that is a good approach?

→ *Not necessarily! Empirical evidence shows that the extreme events in real data come in clusters. Our mixture approach does not offer that. The GARCH model will…*

Random Walk with Heavy Tails

For obtaining a model that reflects the stylized facts of financial data more genuinely, we could use a *Random Walk with heavy-tailed increments*. The distributional choice is $t_v(\mu, \lambda^2)$.

- We require a routine for fitting the distribution to a set of observed one-day log returns.
- Multi-period risk management will no longer be as easy: the sum of independent heavy-tailed log-returns is no longer in the same distributional family and we urgently require a Monte Carlo simulation procedure.

→ **See next slides…**

Fitting t-Distributions to Data

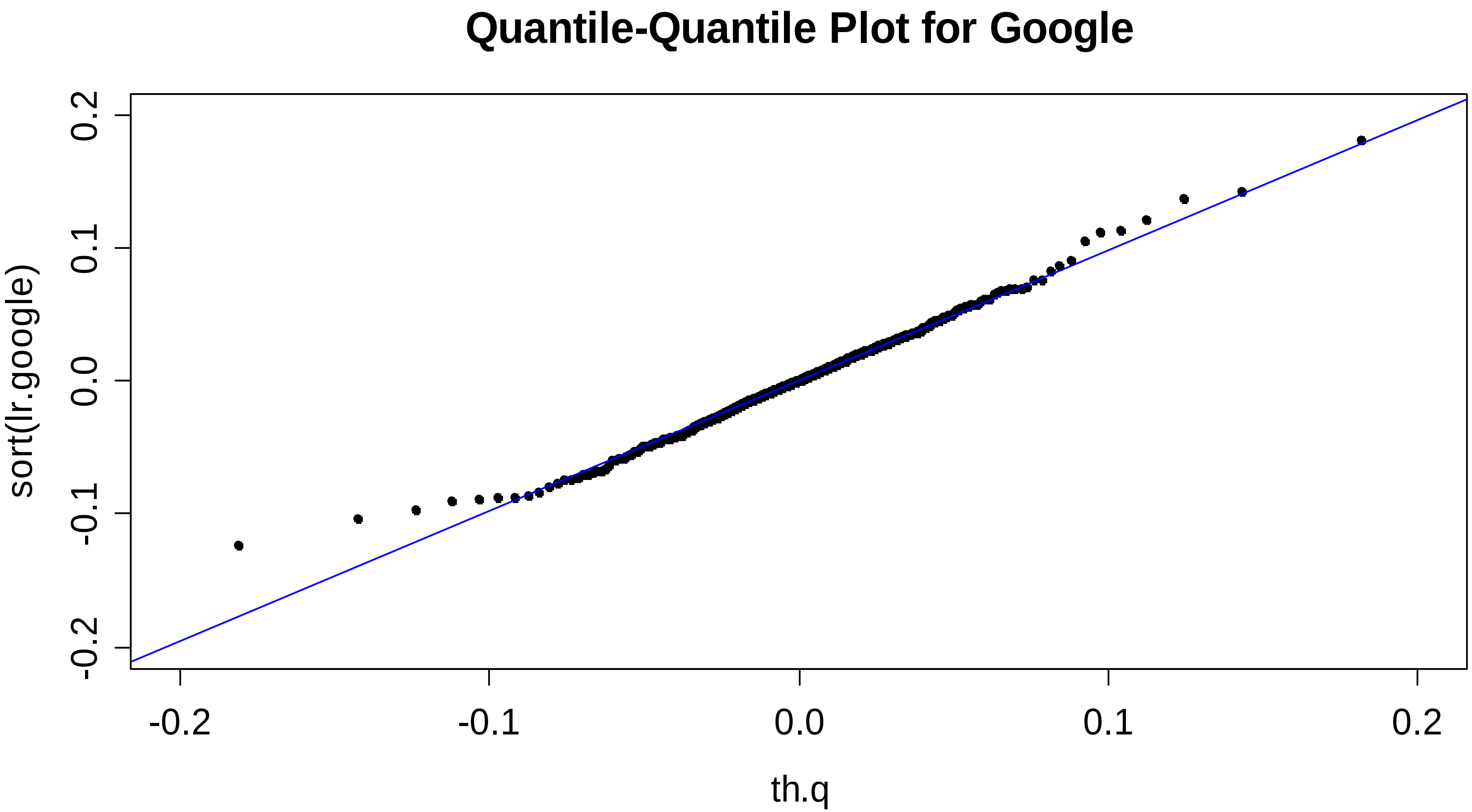
We can use a maximum likelihood approach to fit a $t_\nu(\mu, \lambda^2)$ to financial data. Numerical optimization is required.

In R:

```
> library(MASS)
> fitdistr(lr.google, "t")
        m               s               df
  0.0009455952   0.0133499234   2.9431358498
 (0.0003562337) (0.0003839158) (0.2159852731)
```

MLE theory says the estimates are asymptotically normal. Thus, we can construct approximate 95%-CI using the provided SEs.

Evaluating a t-Distribution



Risk Management with Heavy Tails

The 5%-quantile of the log return distribution turns out to be:.

```
> 0.000945595+0.0133499234*qt(0.05,2.9431358498)
[1] -0.03072064
```

If we aim for the 5%-quantile of the 20-day log return distribution, we have to resort to a simulation approach. It involves drawing many (i.e. 100’000x) sets of 20 single-day returns, before these are summed up and their empirical 5%-quantile is obtained:

```
> quantile(res, 0.05)
        5%
-0.1454524
```

Random Walk with Heavy Tails

Goal is computing the 1-day Value-at-Risk, i.e. the loss which is not exceeded with 95% probability:

Method	95%-VaR	99%-VaR
Gauss	-3.47%	-4.94%
T with 2.94 df	-3.07%	-6.06%
Empirical	-3.13%	-5.97%

If we are interested in a 20-day horizon, things are easy for the Gaussian distribution, but for the t-distribution…?

Method	95%-VaR	99%-VaR
Gauss	-14.07%	-20.67%
T with 2.94 df	-14.55%	-23.71%