
# First look at some data

Yvan Lengwiler
Faculty of Business and Economics
University of Basel

yvan.lengwiler@unibas.ch

File: prep-01.tex

**Contents**

# 1 Computer infrastructure

## 1.1 *Coding*

We will work with data in this course. the aim is to help you become potentially productive employees in the finance industry. We will use the R programming language for all the coding we do. Of course, there are many different hammers that can drive in a nail, but our hammer will be R.

So you have to have R installed on your computer. The R languge is available from R Core Team (2022) at `https://www.r-project.org/`.

Personally, I also use `Rstudio`, which is a development environment specially for the R language. You do not necessarily need that, but many people do. `Rstudio` from `https://posit.co/download/rstudio-desktop/`. If you use Rstudio, make sure to install the free version to avoid paying money.

## 1.2 *A word about git*

The files you need for this lecture are available on ADAM. They are also available on Github. Collecting them from there is probably easier if you have git installed on your system. You can then simply say

```
git clone --depth 1 https://github.com/yvan-lengwiler/Finance-Uni.git
```

If you have never used git, this looks foreign to you. git is version control software, and if you have never used it, I would advise you to learn git and start to use it for all your projects. Moreover, you should also use Github or another online repo store.

Why should you do that? It is a fact that most finance jobs today cannot be neatly separated from computer science. Even if you will not be a code developer, people working professionally in finance today are expected to code at least to some extent, and that also means sharing code and contributing to the codebase of your employer. In most modern finance houses, this means R, Python, C/C++, and, yes, git as well. (In other companies, it just means Excel :-)

The separate handout called *git* should help you get up to speed.

# 2 Types of assets

Financial assets are the things that are traded in financial markets.

**1 Question.** What types of financial assets do you know? What are their defining properties?

Your answer here:

## 3 Downloading some data

Equities are shares of ownership of companies and represent productive capital. They are the main investment vehicle that drives any capitalist economy. Equities of many large companies are publictly traded on organized exchanges and we we can thus easily track their market prices. Let us download a time series of a significant company from a website that provides such data for free (`finance.yahoo.com`) using the `R` programming language. Please study the supplied program `'explore-an-equity-unfinished.r'` to see how I do this. There is also an R Markdown version `'.Rmd'` of this file if you prefer that format.[1] Here are the most important parts of the code:

```
# **** preparations ************************************************
# install and load some packages if they are missing
packages <- c('yfR','scales','here','curl')
missing_packages <- !(packages %in% rownames(installed.packages()))
if (any(missing_packages)) {install.packages(packages[missing_packages])}
invisible(lapply(packages, library, character.only = TRUE))
# select location of this file as working directory
setwd(here())
```

[1] Plain R scripts and R markdown notebooks are functionally similar. The notebook format has the advantage that you can mix and match 'blobs' of different versions and your plots are rendered together with your code. The disadvantage is the overhead you have for defining code and text sections. I will mostly use plain scripts, but you are free to work with the notebook.

```
# **** parameters ****************************************************
symbol    <- 'GM'               # General Motors
interval  <- 'monthly'          # daily, weekly, monthly
from_date <- '2010-12-01'
to_date   <- '2023-12-31'
# **** number of observations per year ********************************
# **** acquire data from finance.yahoo.com ****************************
data <- yf_get(tickers = symbol, do_cache = TRUE, be_quiet = TRUE,
    freq_data = interval, first_date = from_date, last_date = to_date)
price <- as.numeric(data$price_adjusted)
dates <- as.Date(data$ref_date)
# **** plot it ********************************************************
plot(dates, log(price), main = symbol, type='l')
# **** compute returns ************************************************
# annualized return rate from one observation to the next
factor <- switch(
    interval,
        'monthly' = 12,          # number of trading months
        'weekly'  = 365.25/7,    # number of trading weeks
        'daily'   = 365.25*5/7   # number of trading days
    )
yield <- diff(log(price)) * factor
# **** plot it ********************************************************
bullet_size <- sqrt(150/length(yield))
plot(dates[-1], yield, pch=20, cex=bullet_size,
     main = paste("annualized", interval_name, "return of", symbol))
```

**2 Task.** Study this code. Please go through the sections of this code and mark each line you do not understand. (The code in the "Preparations" section might look cryptic, but that is not the important part you need to understand.)

---

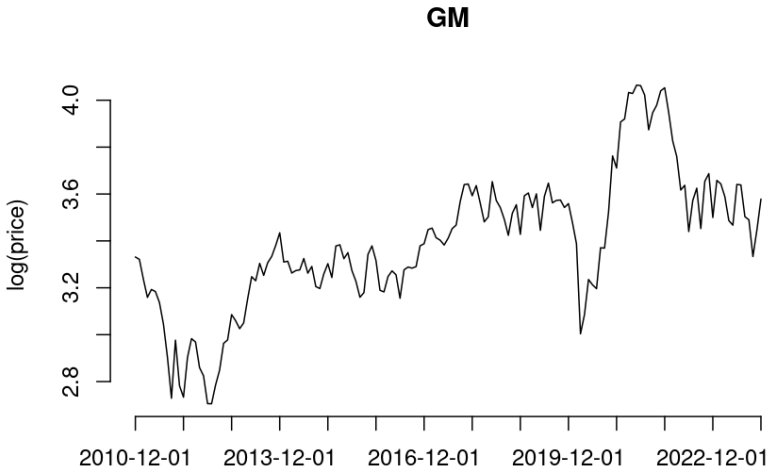
Your answer here:

---

## 4 Equities as stochastic processes

### 4.1 *Visual inspection*

The result of the skript above is two charts. the first chart is the one below.




**3 Question.** Try to describe the most relevant features of this time series. is it random? is there a trend? etc.

---

Your answer here:

---

**4 Question.** The chart is represented on a logarithmic scale. What are the two main advantages of the log scale here?

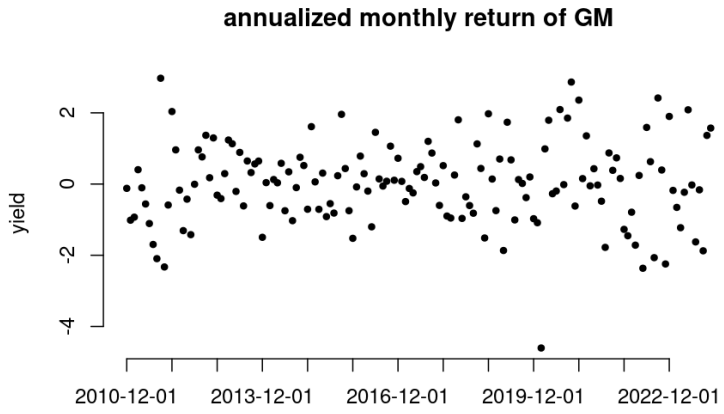
*note*: If you are interested in such things, check out `https://youtu.be/OjIwCOevUew?si=BtzwSN5DI6NdE6CN` for a discussion of the history of the logarithm and an intuitive explanation of its properties.

---

Your answer here:

---

Instead of studying the level of the share price, we can also study the *yield* or *return rate* that this path entails. doing this for our data produces




The stochastic properties of yield time series are central in finance, so it is worth to study this in some depth.

**5 Question.** What can you observe with the naked eye?

---

Your answer here:

---

## 5 Distribution of the return rates

Just as any financial asset, equity prices are to some extent stochastic. The interpretation what this means is a philosophical question. one interpretation could be that these prices are fundamentally driven by some randomness in nature — like quantum particles. But this is unlikely in my opinion. More likely, equities appear stochastic to us because first of all, we do not fully understand what drives these prices, and moreover, we do not have all the information of the elements that we know do drive these prices. So we lack information as well as a complete theory.

Given this state of affairs, we have to deal with equities as random objects. And we can investigate their stochastic properties.

### 5.1 *Kernel estimates and q-q plots*

A histogram is a simple method to estimate the distribution of a random variable. We partition the range of possible values into intervals of equal sizes and then count the number of observations in each interval. The relative counts then provide information about the relative frequency of different values of the random variable.

A kernel estimate (or density estimate) is the continuous analog of the histogram. Instead of having a fixed set of non-overlapping intervals, we have one fixed-size interval that we drag along the range of possible values. At each point, we again count the number of observations and record this count as the value of the density estimation at the center of the interval. This way, we produce a continuous analog of the histogram. Please read the article in wikipedia `https://en.wikipedia.org/wiki/kernel_density_estimation` for a much deeper explanation. In `r`, we can compute the kernel with the `density` command, see `explore-an-equity.r`

We often work with the assumption of normally distributed returns in finance, so it is worthwile to check if this assumption is valid empirically. We should explore if this

assumption is empirically justified. One way to do this is to generate a Q-Q plot. This is a scatter plot of all our observations. The vertical axis shows the data scaled according to their quantiles, the horizontal axis shows the location of the quantiles that would result if the distribution was Gaussian (normal). Deviations of the empirical distribution from the normal become apparent as a deviation of the scatter plot from a straight line. (Again, consult the following link if you do not understand what I try to explain here, `https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot`.)

**6 Task.** Expand our R skript so that it produces a kernel estimate and a Q-Q plot of the returns. The following R functions are useful for this: `density`, `qqnorm`, and `qqline`. You can find the manual of any R function with `?command`, so for instance `?density`. You are free to work in groups.
The aim is to produce kernel estimates and Q-Q plots for the mobthly returns of the GM stock.

---

Your answer here:

## 5.2 *Comparing data at different frequencies*

**7 Task.** Create kernel estimates and Q-Q plots for the S$P500 index (`symbol <- '^GSPC'`) for monthly and for daily data. Feel free to explore other equities as well. Describe your observations. Is there a qualitative difference between the distributions of the monthly and the daily returns?

---

Your answer here:

⋆ ⋆ ⋆

I expect to get to this point in the first lecture. For next week, go though the rest of this handout, as well as through next week's handout as preparation for the lecture. That lecture will be online via Zoom.

**Have this ready for the Zoom meeting.** Please prepare notes, charts, programs etc so that you can share the in Zoom. If you failed to complete a task, that is not a problem. It happens to all of us. But do prepare an explanation that makes clear where exactly you struggled.

⋆ ⋆ ⋆

## 6 Short-term vs longer-term return rates

Let $p(t)$ be the price of the equity at the end of day $t$. The yield from $t-1$ to $t$ is the logarithmic difference,[2]

$$r_d(t) = \ln p(t) - \ln p(t-1).$$

The annualized daily return ($r_D$) is the same except we compute the change that would result if each day of a trading year yielded the same result. This is simply

$$r_D(t) := r_d(t) \cdot \mathit{factor}, \qquad \text{with} \quad \mathit{factor} = 365.25 \cdot 5/7 \approx 261.$$

*factor* here is the approximate number of trading days in a year.

**8 Question.** Let $\mu_d$ and $\sigma_d^2$ denote the mean and the variance of the day-to-day yields. Compute the the mean, the variance, and the volatility (standard deviation) of the annualized daily return $r_D$?

---

Your answer here:

---

[2]The logarithmic difference produces the *continuous yields*, i.e. the return rate at any moment in time, assuming continuous compounding. In practice, we often see *discrete yields* $(p(t)/p(t-1))-1$. These are mathematically more cumbersome but often used in practical work. Annualizing a discrete yield involves taking geometric means and not the much simpler arithmetic means, for instance. Also, gains and losses are not symmetric: a discrete loss of 50% requires a subsequent gain of 100% to compensate. This asymmetry is not present in the continuous yields.

The monthly return is the logarithmic difference of the level from one month to the next. We measure only trading days. Each year has 12 months and there are 261 trading days, so each month has about 22 trading days on average (more precisely, 21.75). The yield from one month to the next is therefore approximately

$$r_m(t) = \ln p(t) - \ln p(t-22).$$

Likewise, the annualized monthly yield is

$$r_M(t) := r_m(t) \cdot \mathit{factor}, \qquad \text{with} \quad \mathit{factor} = 12.$$

**9 Question.** Assume that the day-to-day return is not serially correlated. What is the mean, variance, and volatility of the month-to-month yields, $r_m$, expressed as a function of $\mu_d$ and $\sigma_d^2$? Furthermore, what is the mean, variance, volatility of the annualized monthly returns, $r_M$?

---

Your answer here:

**10 Question.** What can you say about the form of the distribution of monthly return rates, compared to the form of the distribution of daily return rates (independently of their different volatilities)? What do you expect for the distribution of annual or multi-year return rates? How does this relate to your findings in Task 6?

---

Your answer here:

---

## 7 Preparation for next week

Please work through the part of this handout we have not covered in class. Also work through the handout for next week.

Have your notes, charts, programs etc. ready for the Zoom lecture, so that you can share them with us and we can discuss them. If you failed to complete a task, that is not a problem. It happens to all of us. But do prepare an explanation that makes clear where exactly you struggled.

### References

**R Core Team** (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.