Lecture 13:
Financial Time Series Data

Steven Skiena

Department of Computer Science
State University of New York
Stony Brook, NY 11794–4400

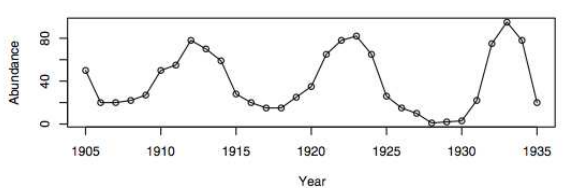
http://www.cs.sunysb.edu/∼skiena

Time Series Analysis

A *time series* consists of the values of a function sampled at different points in time.

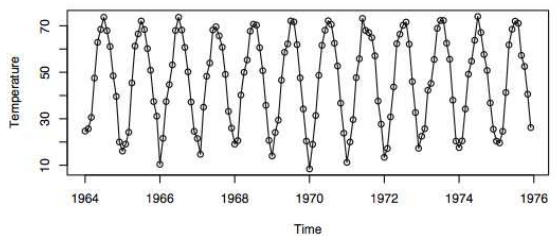
Time series data arises throughout the natural and physical sciences, as growth curves, statistical measurements of activity, . . .

Exhibit 1.5 Abundance of Canadian Hare




```
> win.graph(width=4.875, height=2.5,pointsize=8)
> data(hare); plot(hare,ylab='Abundance',xlab='Year',type='o')
```

Exhibit 1.7 Average Monthly Temperatures, Dubuque, Iowa



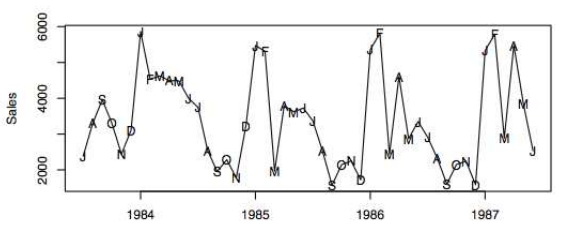

```
> win.graph(width=4.875, height=2.5,pointsize=8)
> data(tempdub); plot(tempdub,ylab='Temperature',type='o')
```

Financial Time Series Data

With respect to financial data, the price of any asset as a function of time naturally gives a time series.

Many relevant statistics (such as the unemployment rate or index of leading economic indicators) can also be thought of as time series data.

Exhibit 1.9 Monthly Oil Filter Sales with Special Plotting Symbols



J=January (and June and July),
F=February, M=March (and May), and so forth

```
> plot(oilfilters,type='l',ylab='Sales')
> points(y=oilfilters,x=time(oilfilters),
    pch=as.vector(season(oilfilters)))
```

Approaches to Time Series Analysis

A wide variety of mathematical and statistical tools have been developed for working with time series data.

Adherents to *technical analysis* argue that insight into future price movements follow from the analysis of a given asset's price time series.

Regardless, the analysis of financial time series is important in developing/evaluating any investment strategy, risk modeling, and arbitrage.

Issues of Time Series Data

- Is it sampled at equally-spaced intervals?
- Is there noise or error in the data?
- How long/rapidly growing is the available series?
- Are there any missing values?

Different answers will result when working with stock prices, tick data, sales data, polling data, and government statistics like the unemployment rate.

Has anyone here ever worked with time series data in some context?

The Discrete Nature of Stock Prices

When the price of a share of stock gets so expensive it is unwieldy, each share *splits* into equal-sized pieces which sum to the the original value.

Reverse splits combine several shares into a single more expensive share.

Such games are played for psychological reasons, but also to set an meaningful lower bound on the minimum amount prices can change. Since *decimalization*, the minimum change is typically \$0.01, but used to be \$0.125 (one eighth). Reducing this minimum change in principle enables buyers and sellers to get fairer prices.

Adjusted Price Time Series

Computing an *adjusted price* time series corrects for splits and dividends, but requires recomputing all past history on each new event.

PRICES

Date	Open	High	Low	Close	Avg Vol	Adj Close*
Oct-08	9.63	9.90	4.00	6.43	35,175,500	6.43
Sep-08	10.47	14.31	8.51	9.45	32,449,000	9.45
Aug-08	10.42	11.88	9.68	10.00	24,325,700	10.00
Jul-08	11.02	16.35	8.81	11.07	31,934,000	11.07
Jun-08	17.73	18.18	10.57	11.50	31,822,600	11.50
14-May-08	$ 0.25 Dividend					
May-08	23.10	24.04	16.87	17.10	20,311,500	17.10
Apr-08	19.53	24.24	18.72	23.20	23,239,800	22.91
Mar-08	23.67	23.67	17.47	19.05	22,666,900	18.81
13-Feb-08	$ 0.25 Dividend					
Feb-08	28.20	29.28	22.96	23.28	17,741,700	22.99

Dividends are best adjusted for by adding them back to the stock price, as if immediately used to purchase more stock.

Asset Returns

The *price* of an asset as a function of time is perhaps the most natural financial time series, but it is not the best way to manipulate the data mathematically.

The price of any reasonable asset will increase *exponentially* with time, but mathematical tools (e.g. correlation, regression) work most naturally with linear functions.

The *mean* value of an exponentially-increasing time series has no obvious meaning.

The *derivative* of an exponential function is exponential, so day-to-day changes in price have the same unfortunate properties.

Simple Net Returns

Much better is to represent the data as a *simple net return*:

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

Negative returns means the asset declined in value, positive returns means it increased, zero returns means it is unchanged.

The return is a complete and scale-free summary of investment performance.

Multiperiod Returns

A nice property of returns is that multiplying them gives the return over a longer period:

$$1 + R_t[k] = \frac{P_t}{P_{t-k}} = \prod_{j=0}^{k-1} \frac{P_{t-j}}{P_{t-j-1}} = \prod_{j=0}^{k-1} (1 + R_{t-j})$$

Normally, returns are discussed in annualized terms, so over k years the *annualized return* is computed by its *geometric mean*:

$$\{R_t[k]\} = \left(\prod_{j=0}^{k-1} (1 + R_{t-j}) \right)^{1/k} - 1$$

Geometric vs. Arithmetic Mean

Why use the geometric mean ($(\Pi_n x_i)^{1/n}$) instead of the arithmetic mean ($(\Sigma_n x_i)/n$)?

Because k years at the annualized rate of return gives exactly the same payoff as the given return time series.

Thus the geometric mean of a funds return over n years is more meaningful than the arithmetic mean.

If the price ever goes to zero, you never recover!

The arithmetic mean never smaller than the geometric mean, so which appears in mutual fund ads?

Consider $1, 2, \ldots, 100$. The geometric mean is 37.99 vs. 50.5.

Logarithmic and Other Returns

The mathematical complexities of multiplying returns can be eliminated by dealing with *continuously compounded returns* or *log returns*:

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

The multiperiod log return is simply the sum of the log returns.

Returns of assets paying dividends must include the value of the dividend payments at the time they are issued. Ignoring dividend payments with respect to returns only messes up data points on dividend days, instead of invalidating the entire time series.

Excess Return

The *excess return* of an asset at time t is the difference between its return and that of a reference asset, typically the risk-free rate.

The excess return is the payoff of a portfolio going long in the asset and short on the reference.

Excess return is often measured in *basis points*, or $(1/100)$th of a percent.

A strategy of returning 5.5% with a risk-free rate of 5.0% earns an excess return of 50 basis points.

Moments of Distributions

The *lth moment* of a continuous random variable X is defined

$$m_l' = E(X^l) = \int_{-\infty}^{\infty} x^l f(x) dx$$

where $E(X)$ is the expected value and $f(x)$ is the probability density function of X.

The first moment is the *mean* or *expectation* of X, μ_x.

The *lth central moment* of a continuous random variable X is defined

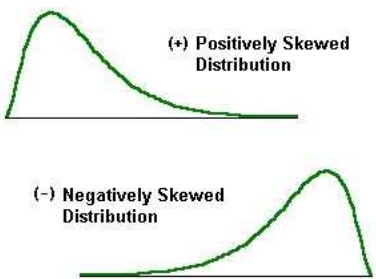
$$m_l' = E(X^l) = \int_{-\infty}^{\infty} (x - \mu_x)^l f(x) dx$$

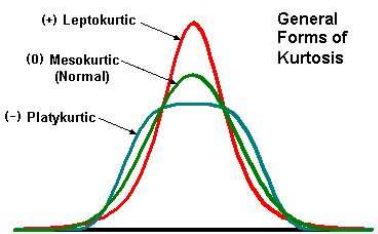
The second central moment is the *variance* $(\sigma_x)^2$ where σ_x is the *standard deviation*. The *variance* measures how much the random variable jumps around from the mean.

Higher Moments

The third central moment is the skewness of the random variable, a measure of the extent of symmetry.

The fourth central moment is the kurtosis, a measure of how much mass in the *tails* of the distribution.





Properties of Asset Returns

If we consider the returns of a volatile asset, such the daily return on a stock, we would expect:

- The expected value will be small, probably near zero. Think about 10% return divided by 365 days.
- There should be a gross symmetry between negative returns and positive returns.
- Smaller absolute returns will occur more frequently than larger absolute returns.

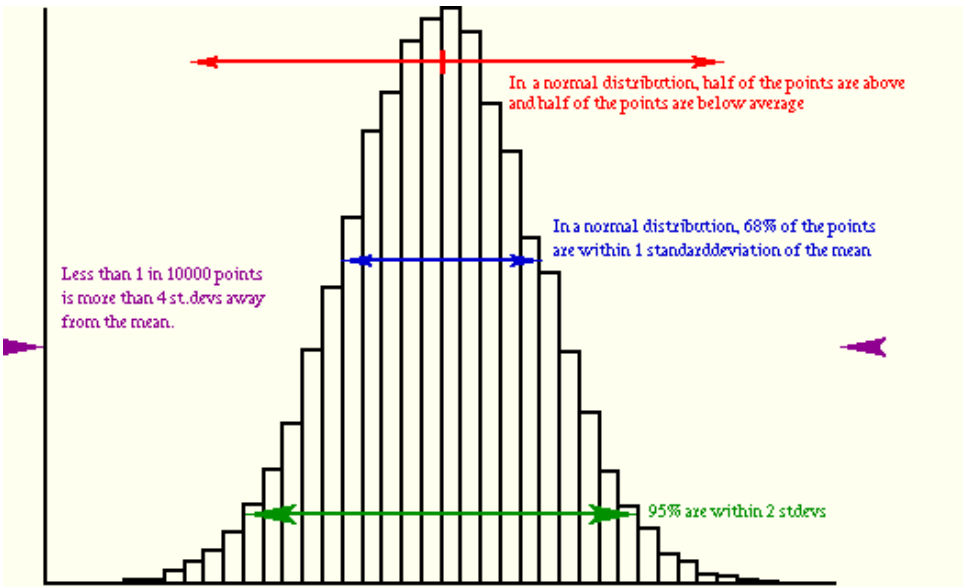
All of these suggest some time of *bell-shaped curve*, but which one. . .

The Normal Distribution

The classic bell-shaped curve is that of the *normal distribution*, whose probability density function is:

$$f(x) = \frac{e^{((-(x-\mu)^2)/(2\sigma^2))}}{(\sigma\sqrt{2\pi})}$$

where $\sigma > 0$, $-\infty < \mu < \infty$, and $-\infty < x < \infty$.



This is centered around the mean, symmetrical, and has tails which go out to infinity in each direction.

The normal distribution is completely parameterized by the mean and standard deviation.

Thinking Normally

Approximately 2/3 of the probability mass of a normal distribution lies within one standard deviation from the mean.
Approximately 95% of the probability mass of a normal distribution lies within two standard deviations from the mean.
Thus the probability of being far from the mean decreases rapidly – less than one in 10,000 points is more than four standard deviations from the mean.
Six Sigma denotes a process with 99.9997% reliabilty.

What's Normal?

Human heights and weights seem to be fit reasonably well by normal distributions, although the observed distributions do not have tails which go to infinity.

Consider compare this to the distribution of incomes. It is much rarer to find someone twice as tall as the mean than twice as rich as the mean.

The tails of the income distribution go out much further than is supported by a normal distribution.

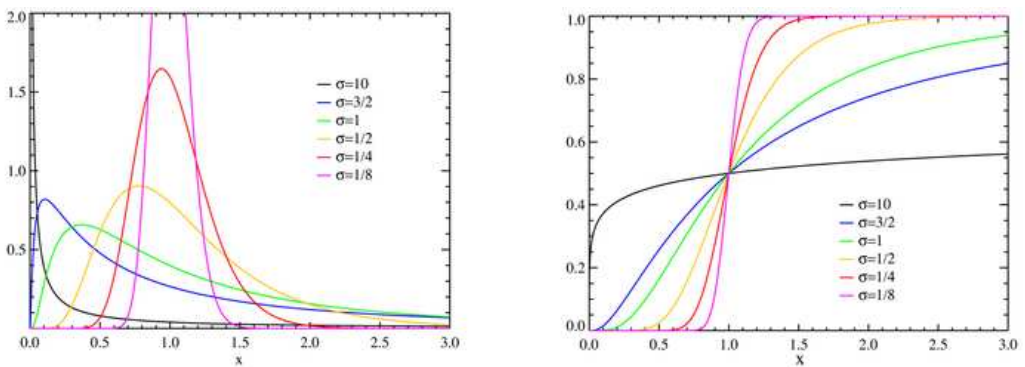
Stock Returns are not Normal

Stock returns are not completely modeled by normal distributions because:

- a lower bound on return is -1, with no upper bound,
- if daily returns were normally distributed, then multi-period returns are not normal (they would be the product of normals),
- Empirical data suggests that returns show leptokurtosis, fatter tails than expected with a normal distribution.

The Lognormal Distribution

A better assumption is that the *log returns* r_t are normally distributed with mean μ and variance σ^2.



Since the sum of a finite number of independent normal random variables is normal, the conceptual problem with multiperiod returns is eliminated

Mixture Models

Creating a mixture (not sum) of two normal distributions with identical mean but different variance can produce fatter tails:

$$r_t \approx (1-\alpha)N(\mu, {\sigma_1}^2) + \alpha N(\mu, {\sigma_2}^2)$$

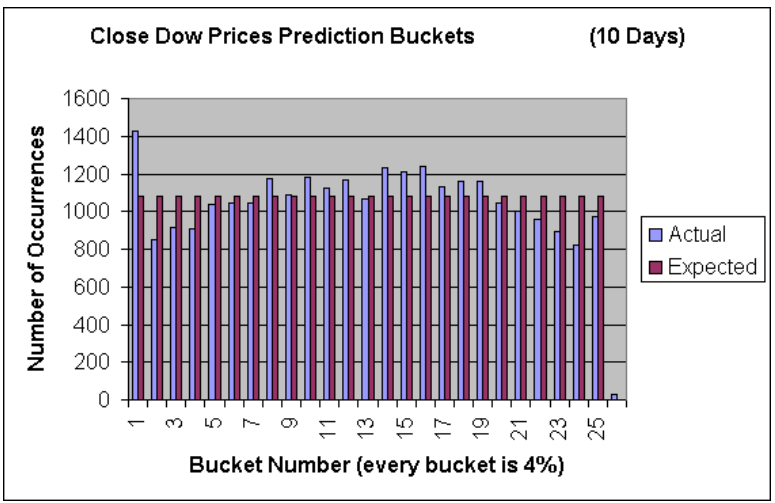
However, adding parameters requires more data to fit accurately, and are less satisfying theoretically unless you can explain the need for two distributions.

Other distributions, including *stable distributions* and the *Cauchy distribution* have been proposed to model returns.

Certain simulations draw return values sampled from historical data instead of distributions.

Empirical Properties of Returns

Extreme events are more likely to be crashes than explosions.



Empirical density functions are taller, skinnier, and wider than normal or lognormal (leptokuretic).

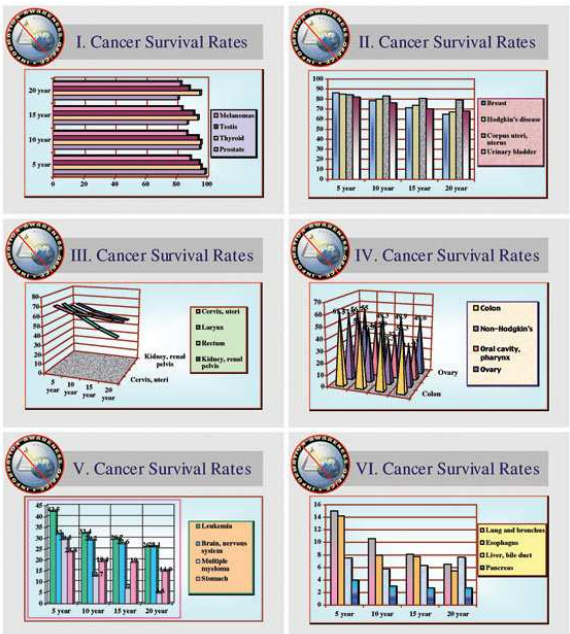
Visualizing Statistical Data

The first step to understanding your data is to visualize it, usually using software.
Be aware of good design principles for mathematical graphics:

- Minimize ratio of non-data ink.
- Meaningfully label axes / curves and display scales.
- Avoid visual clutter.

Read Tufte's books on graphic design to develop an appropriate sense of style.

Dying from Cancer / Graphic Design



Recovering from Bad Design