MIT MANAGEMENT SLOAN SCHOOL

15.455x Mathematical Methods of Quantitative Finance

Week 2: Introduction to Discrete-Time Stochastic Processes

Paul F. Mende
MIT Sloan School of Management

Finance at MIT
Where ingenuity drives results

The Random Walk

Finance at MIT
Where ingenuity drives results

Time series models

- A **stochastic process** is a time-dependent random variable
  - Continuous time, *S(t)*
  - Discrete time, $S_1, S_2, S_3, \ldots S_t$
  - Time series models sample the variable at uniform intervals
    - Integer indices
    - Equal spacing (weekdays?)
    - Time zero origin

- Discrete time processes can be constructed by adding increments
  - Example: $S_t = S_{t-1} + x_t$

$$= x_0 + x_1 + x_2 + \cdots + x_t$$

  - Increments can be recovered by differencing

$$x_t = S_t - S_{t-1}$$

Time series models

Time series are used to model processes that evolve or are observed discretely in time.

- Ordered steps are identified with endpoints of time intervals.
- Often defined, recursively, in terms of prior variables plus an "innovation."
- The model is "solved" when we can describe or forecast the distribution of values.

Examples:

- Cumulative income $I_1 + I_2 + I_3 + \cdots$
- Daily stock price values $P_t = P_0 e^{r_1 + r_2 + \cdots + r_t}$

logarithmic return of a stock price

- Random walk

Random walk model

The random walk is the simplest of all time-series models

$$S_T = z_1 + z_2 + \cdots + z_T$$

- Each increment is an random IID variable
  - Independent and identically distributed
  - No dependence on past history independent
  - Uniform time evolution identically: evolution in time is uniform (stationarity: if we look at the values between any two points in time, the properties depend only on the differences in the point in time, not on some absolute point in time)
- Universality
  - The essential features are independent of many step-level details
  - Easy to generalize for financial applications
  - Appears frequently in other contexts
  - Building block for more complex models

Random walk model

We can often express results in terms of "standard" random variables:

- Mean = 0
- Variance = 1
- Correlation = 0 for different times

$$\mathrm{E}[z_t] = 0$$
$$\mathrm{E}[z_t^2] = 1$$
$$\mathrm{E}[z_t z_{t'}] = 0 \text{ if } t \neq t'$$

Examples:

- Coin toss (+1 or –1 with equal probability) $z_t = \pm 1, \qquad p = 1/2$

- Gaussian random variable $z_t \sim \mathcal{N}(0,1), \quad p(z_t) = \frac{1}{\sqrt{2\pi}} e^{-z_t^2/2}$

Random walk model

Now consider a **sum** of these identical random variables.

- What is the probability distribution of the sum?
  - In general, it can be a complicated (given by a convolution of the individual distributions).
  - However, there is an especially simple result for the **mean** and the **variance** of the sum.

$$\mathrm{E}[S_T] = \mathrm{E}[z_1] + \mathrm{E}[z_2] + \ldots + \mathrm{E}[z_T] = 0$$

$$\begin{aligned}\mathrm{Var}(S_T) &= \mathrm{E}[S_T^2] = \mathrm{E}\left[(z_1 + z_2 + \ldots + z_T)^2\right] \\ &= \sum_{t=1}^{T} \mathrm{E}[z_t^2] + 2\sum_{t<t'} \mathrm{E}[z_t z_{t'}] = T\end{aligned}$$

- The *T*-step random walk has a variance that **grows linearly with *T***.
- The **standard deviation** (its square root) grows as square root of *T.*

Random walk model

- **Summary**
  - The elementary **random walk model** is constructed as a sum of IID random variables

$$S_T = z_1 + z_2 + \cdots + z_T$$

  - The random walk is a simple example of a discrete-time stochastic process
  - Its moments can be computed by using the linearity of the expectation operator
  - Its mean is zero and its variance is proportional to the number of steps
    - These results hold for any "standard" increments, whether continuous or discrete

$$\begin{aligned} \mathrm{E}[z_t] &= 0 \\ \mathrm{E}[z_t^2] &= 1 \\ \mathrm{E}[z_t z_{t'}] &= 0 \text{ if } t \neq t' \end{aligned}$$

Generalized random walk model

Let's generalize this model by introducing two constant parameters,

$$r_t = \sigma z_t + \mu$$

- Random variable scaled by σ
- Constant additive piece μ
- Independent steps in each time period
- These two parameters will be measures of **risk** and **return**.

Application: asset price returns

- Consider a series of regularly observed stock prices, whose logarithmic returns are defined by

$$r_t \equiv \log\left(P_t / P_{t-1}\right)$$

- Then

$$P_T = P_0 e^{r_1 + r_2 + \cdots + r_T}$$

Generalized random walk model

Since the expectation operator is **linear**, we can compute the expected return, variance, and covariance immediately from our previous results.

$$r_t \equiv \sigma z_t + \mu \implies \begin{aligned} &\mathrm{E}[r_t] = \mu \\ &\mathrm{E}[(r_t - \mu)^2] = \sigma^2 \\ &\mathrm{E}[(r_t - \mu)(r_{t'} - \mu)] = 0 \text{ if } t \neq t' \end{aligned}$$

Generalized random walk model

Now consider the stochastic process generated by the sum

$$X_T \equiv r_1 + r_2 + \ldots + r_T = \sum_{t=1}^{T} r_t$$

and again use the linearity of the expectation operator to find the mean and variance of this sum:

- Mean: $\mathrm{E}[X_T] = \mathrm{E}[r_1] + \mathrm{E}[r_2] + \ldots + \mathrm{E}[r_T] = T\mu$

- Variance:
$$\begin{aligned}\mathrm{Var}(X_T) &= \mathrm{E}\left[(X_T - T\mu)^2\right] \\ &= \mathrm{E}\left[\Big((r_1 - \mu) + (r_2 - \mu) + \cdots + (r_T - \mu)\Big)^2\right] \\ &= T\sigma^2\end{aligned}$$

Generalized random walk model

- **Summary**
  - The **generalized random walk model** is constructed as a sum of IID random variables

$$X_T \equiv r_1 + r_2 + \ldots + r_T$$

  - Each term has a scale and an offset, which can be used to model the volatility and mean return of asset price dynamics

$$r_t = \sigma z_t + \mu$$

  - The moments can be computed by using the linearity of the expectation operator
  - The mean and variance of the sum are **linear** in time. They are *T* times the parameter values of each individual step.

Linear Time Series Models

Finance at MIT
Where ingenuity drives results

Time series models

To capture causality and **correlation across time**, more complex models can be built by combining elements of the random walk.

- Example: MA(1) moving average model is not IID

moving average

$$r_t \equiv \mu + \sigma z_t + \phi z_{t-1}$$

already been observed (constant): different time, different distribution

Notice that the last term refers to an earlier-time random variable whose value will be known by time *t*.

- In other models, such as GARCH, the distribution may itself be time-varying

$$\begin{aligned} r_t &\equiv \mu + \sigma_t z_t \\ &= \mu + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma_t^2) \end{aligned}$$

Time series models

Another way to model causal influences is to have past values of the random variable itself on the right hand side of a recursive definition.

autoregressive model

- Example: the AR(*p*) autoregressive model of order *p:*

$$R_t = c_0 + c_1 R_{t-1} + \cdots + c_p R_{t-p} + \sigma z_t, \quad z_t \sim \text{IID}(0,1)$$

- Example: the ARMA(*p,q*) model combines both of the above:

$$R_t = c_0 + c_1 R_{t-1} + \cdots + c_p R_{t-p} + \sigma z_t + \phi_1 z_{t-1} + \cdots + \phi_q z_{t-q}$$

- Linear structure
- Past observations determine likelihood of future outcomes
- Coefficients to be determined

Time series models

The distribution of future outcomes can be determined using linearity, recursion, and stationarity.

- A time series process is **stationary** if the joint distribution of all of its values is invariant under time translation $t \to t + s$ for any positive and negative s
- It is **weakly stationary** if the first and second moments (i.e., means and covariances) are invariant.

To see how this works, let's solve for a special case in detail: AR(1)

$$R_t = c_0 + c_1 R_{t-1} + \sigma z_t$$

Time series models

Solving AR(1)

- Suppose that both *t* and *t-1* lie in the future. Apply the (unconditional) expectation operator to both sides of the equation

$$\begin{aligned} R_t &= c_0 + c_1 R_{t-1} + \sigma z_t \\ \implies \mathrm{E}\left[R_t\right] &= c_0 \mathrm{E}\left[1\right] + c_1 \mathrm{E}\left[R_{t-1}\right] + \sigma \mathrm{E}\left[z_t\right] \\ &= c_0 + c_1 \mathrm{E}\left[R_{t-1}\right] \end{aligned}$$

- Stationarity means that $\mathrm{E}\left[R_t\right] = \mathrm{E}\left[R_{t-1}\right]$, and substituting above gives the mean value in terms of the model parameters:

$$\mathrm{E}\left[R_t\right] = \frac{c_0}{1 - c_1}.$$

- For convenience and clarity, let's swap out these parameters for two new ones,

$$\mu = \frac{c_0}{1 - c_1}, \qquad \lambda = -c_1$$

Time series models

In terms of these parameters, AR(1) can be written in this suggestive form:

$$R_t - \mu = -\lambda\left(R_{t-1} - \mu\right) + \sigma z_t, \qquad |\lambda| < 1$$

This model is often used to model **mean reversion**.

- The change in the random variable relative to its mean μ is related to the **previous** period's deviation from the mean.
- For positive values of the coefficient λ, an excess value in one period leads to a change in the opposite direction, toward the mean, in the following period.
- The variance can also be computed from stationarity, as can the lagged **autocovariances**.

Time series models

To solve for the variance, substitute in the "equation of motion" and expand:

$$\begin{aligned}\gamma_0 = \mathrm{Var}[R_t] &= \mathrm{E}\left[(R_t - \mu)^2\right] \\ &= (-\lambda)^2 \mathrm{E}\left[(R_{t-1} - \mu)^2\right] + \sigma^2 \mathrm{E}\left[z_t^2\right] \\ &= \lambda^2 \gamma_0 + \sigma^2\end{aligned}$$

since the cross term drops out due to independence of the z_t from all earlier-time values, $\mathrm{E}\left[z_t(R_{t-k} - \mu)\right] = 0$

Therefore

$$\boxed{\gamma_0 = \frac{\sigma^2}{1 - \lambda^2}}$$

lambda is telling us how much of the previous periods result is going to contribute to the next periods result. And when that's less than 1, it means that affects that shocks tend to die off over time in the absence of the new shocks that are arriving via the new zt. However, a case for lambda greater than 1 into the shocks get amplified, they get bigger and bigger. So that once a shock enters the system, it runs away and it takes over the system.

Time series models

Now consider the covariance of observations taken at two different times. From stationarity, it depends on *k* **only**, not *t*.

$$\begin{aligned}\gamma_k &= \mathrm{E}\left[(R_t - \mu)(R_{t-k} - \mu)\right] \\ &= -\lambda \mathrm{E}\left[(R_{t-1} - \mu)(R_{t-k} - \mu)\right] \\ &= -\lambda \gamma_{k-1}\end{aligned}$$

So for any *k > 0*

$$\gamma_k = (-\lambda)^k \gamma_0 = \frac{(-\lambda)^k}{1-\lambda^2}\sigma^2$$

influence is dying off as K gets larger and larger

This is known as the lag-*k* autocovariance coefficient. It relates the influence of an excess return value at one point in time with values *k* periods in the past.

Time series models

- **Summary**
  - Linear time series models can be constructed out of standard random variables, lagged observations, and constant coefficients.
  - They can exhibit temporal correlations useful in modeling the propagation of information and influence over time.
  - An example is the AR(1) model, used for **mean reversion**:

$$R_t - \mu = -\lambda (R_{t-1} - \mu) + \sigma z_t$$

  - By applying **weak stationarity**, which is the assumption that first and second moments of distributions are invariant in time, models can be solved in terms of their basic parameters.
    - Unconditional expectations are taken with respect to future values
    - Conditional expectations, taken at a fixed point in time, treat past observations as known values rather than as random variables.

unconditional: different periods are all unobserved

Monte Carlo simulation

Finance at MIT
Where ingenuity drives results

Monte Carlo simulation

Use random number generators to **simulate** stochastic process

- This provides a test lab, with **"best case"** data since it comes from a known distribution In the real world we don't know the true data generating process assume a particular model
- Typically generate an ensemble with a **large number** of hypothetical realizations of a financial process or market In the real world we only get to see history once
- Approximate exact results by computing statistics over the ensemble.
  - Results become **more precise** as simulation size increases
  - Numerical results exist even when close-form results do not
  - Testbed for code and algorithms designed for real-world data

do statistical calculations against the empirical data set that's been generated

- Applications:
  - Asset price dynamics
  - Option pricing
  - Portfolio and risk management

Monte Carlo simulation

Most computer languages include functions to generate individual random numbers from various distributions. The R language includes these:

- **sample** – returns discrete values with specified probabilities
- **runif** – returns real number uniformly distributed on [0,1]
- **rnorm** – returns real number with normal distribution of mean zero, variance one

- Note that they are **not truly random**, since the machine is deterministic, and also are just **approximations** to the distribution. For instance, range of **rnorm** (approximation to Gaussian) is bounded, and Prob($X = 1/2$) for **runif** (approximation to uniform) is non-zero.

  we can't do real numbers on a computer. We can only do finite precision.

R Monte Carlo

Rescale these functions to get useful ranges, e.g.,

for efficiency, run many simulations in parallel.

Np as a number of different possible realizations of price paths

`z <- matrix(runif(Nt*Np),nrow=Nt)`	Generates *Nt* x *Np* independent pseudo-random draws
`x <- sign(p-z)`	Generates +1 with probability *p* and -1 with probability *1-p*
`u <- (x+1)/2`	Generates +1 with probability *p* and 0 with probability *1-p*
return: normal `r <- matrix(rnorm(Nt*Np,mean=mu,sd=sigma),nrow=Nt)`	Generates *Nt* x *Np* independent pseudo-random draws with mean *mu* and standard deviation *sigma*
price: lognormal `S <- exp(apply(r,2,cumsum))`	Generate sequences *S* corresponding to lognormal distribution

cumulative summation function to aggregate returns and exponentiate the cumulative sums and returns to get the actual asset price paths.

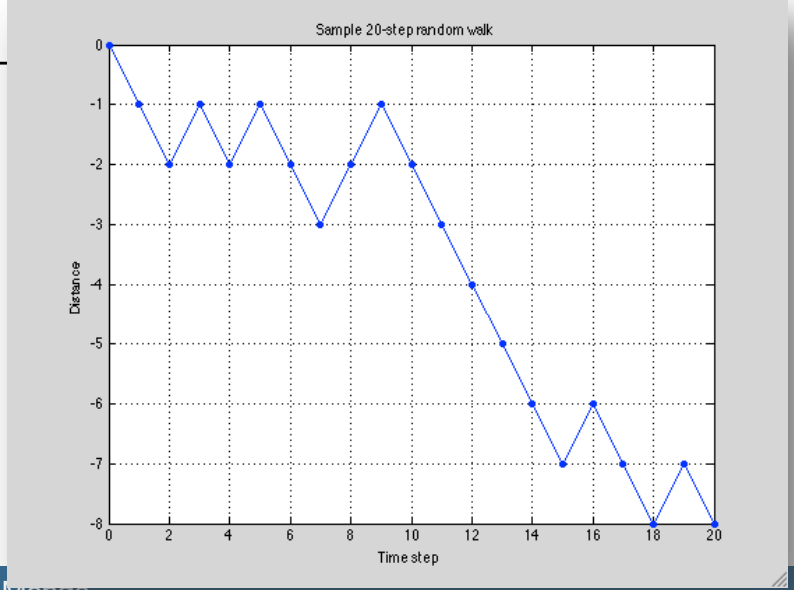
The potential returns of a stock can be graphed in a normal distribution. The prices of the stock, however, can be graphed in a lognormal distribution

lognormal: a random variable whose logarithm is normally distributed

Simple 20-step random walk

```
p  <- 0.5;q <- 1-p;       # Set probability of "success" and "failure"
Nt <- 20;                 # Number of time steps
Np <- 1;                  # Number of sample paths
z  <- matrix(runif(Nt*Np),nrow=Nt)   # Generate a set of uniform random draws
x  <- sign(p-z);          # Transform to binomial random variable +/- 1
s  <- matrix(0,Nt+1,Np)   # Initial value for random walk.
                          # Note that R indices start with 1,not zero.
for (t in 1:Nt) {
 s[t+1,] <- s[t,]+x[t,]   # New location equals previous plus a new random step
}
plot(s,type="b")          # Plot resulting path
```

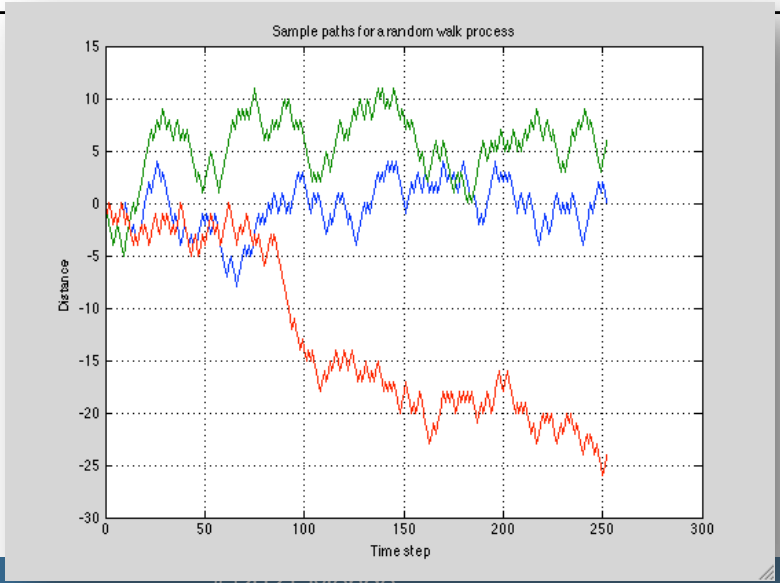
start at time 0 with a fixed initial value, need one more placeholders for state variable values.



Many simulations of a one-year daily walk

```
Nt <- 252;                      # Number of trading days in a year
Np <- 1e4;   accuracy: 1 %      # Reasonably large number of simulations
z  <- matrix(runif(Nt*Np),nrow=Nt)   # Generate a set of uniform random draws
x  <- sign(p-z);                # Transform to binomial random variable +/- 1
s  <- matrix(0,Nt+1,Np)         # Initialize and reserve space for random paths.
for (t in 1:Nt) {
 s[t+1,] <- s[t,]+x[t,]         # New location equals previous plus a new random step
}
matplot(s[,1:3],type="b")               # Plot a few of the resulting paths
```

the accuracy of our results typically is going to scale with 1 over the square root of the number of simulations that we do



Application: asset price dynamics

Asset prices are often modeled as **lognormal** variables, i.e., where their continuously compounded (log) returns are drawn from a normal distribution whether or not this is a good model is an empirical question. We need to look at actual data.

$$r_t = \log(P_t/P_{t-1}) \sim \mathcal{N}(\mu, \sigma^2), \qquad P_t = P_{t-1}e^{r_t},$$
$$= P_{t-2}e^{r_t + r_{t-1}}, \ldots$$

So over *T* periods, the sum of log returns is Gaussian, and the price is a function of that sum:

$$P_T = P_0 e^{r(T)} = P_0 \exp(r_1 + r_2 + \cdots + r_T)$$
$$r(T) = r_1 + r_2 + \cdots + r_T \sim \mathcal{N}(T\mu, T\sigma^2)$$

Simulating a lognormal price process

1. Determine parameters for underlying distribution
2. Scale parameters appropriately for sampling interval
3. Draw random numbers from standard normal distribution
4. Generate simulated returns for each time period
5. Construct ensemble of price paths
6. Compute and plot appropriate analytics and curves

a stock might have a 1 0 % annual return, with the volatility of 3 0 % . But if our individual time steps are on a one day level, then we need to change the parameters in order that they make sense for the one day period

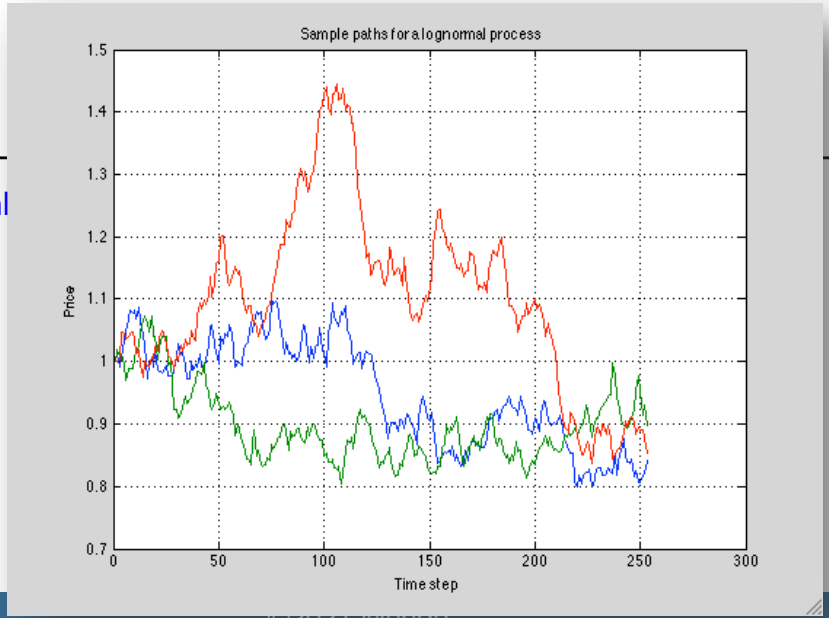
e.g. variance

Simulating a lognormal price process

```
sigma <- 0.3       # Set annualized volatility to 30%
mu    <- 0.1       # Set annualized drift/return to 10%
dt    <- 1/252     # Set time step scale factor to one day, (252 trading days per year)
z  <- matrix(rnorm(Nt*Np),nrow=Nt)  # Generate a set of standard normal random draws
r  <- mu*dt + z*sigma*sqrt(dt)      # Draw daily return from scaled N(mu,sigma^2)
r  <- matrix(rnorm(Nt*Np, mean=mu*dt, sd=sigma*sqrt(dt)),nrow=Nt) # 1-step generation
s  <- matrix(0,Nt+1,Np)
for (t in 1:Nt) {
  s[t+1,] <- s[t,] + r[t,]
}
P <- exp(s)
matplot(P[,1:3],type="l")
```

scale parameters

initial value of price is 1 because initial return is 0 and price is exponential of return



Simulating a lognormal price process

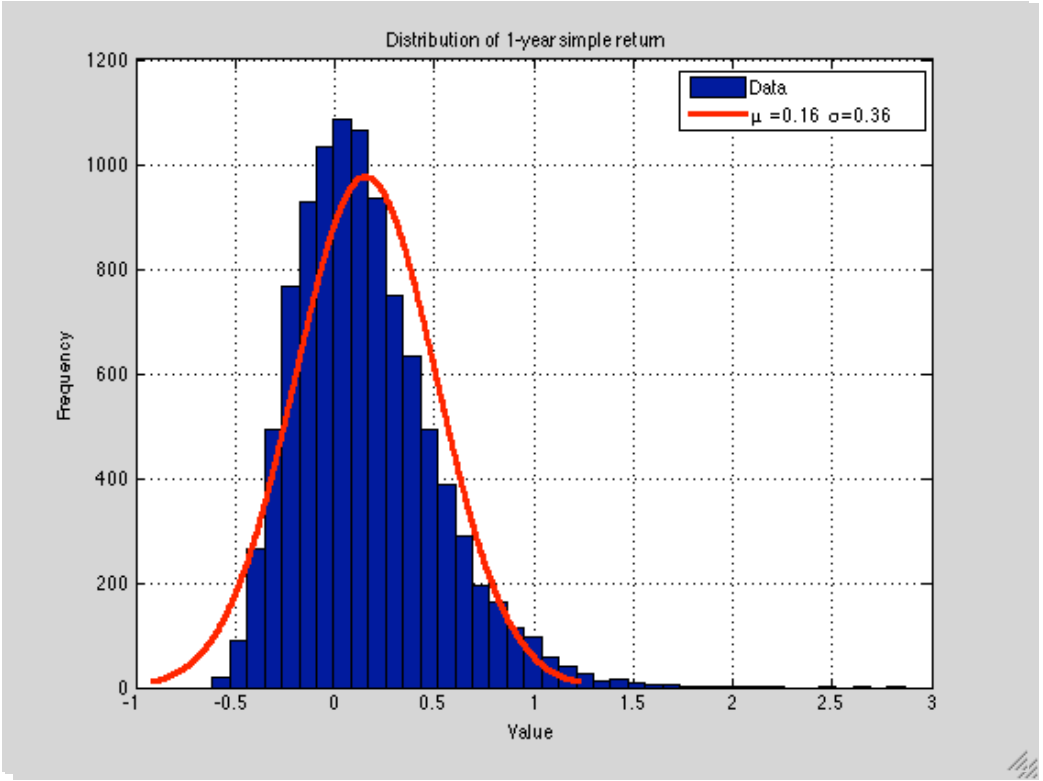
simple return is not normal

```
> R <- P[Nt+1,] - 1

> mean(R)
[1] 0.1547869
> exp(mu+sigma^2/2)-1
[1] 0.1560396

> sd(R)
[1] 0.3563872
> sqrt(exp(2*mu+sigma^2)*(exp(sigma^2)-1))
[1] 0.3547634

hist(R,breaks=50)
hist(log(1+R))
plot.ecdf(1+R)
barplot(sort(1+R))
qqnorm(1+R)
```



$$R = P_T/P_0 - 1$$

$$\mathrm{E}\left[R\right] = \mathrm{E}\left[e^{r(T)} - 1\right] = e^{\mu+\sigma^2/2} - 1$$

$$\mathrm{Var}(R) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

Simulating a lognormal price process

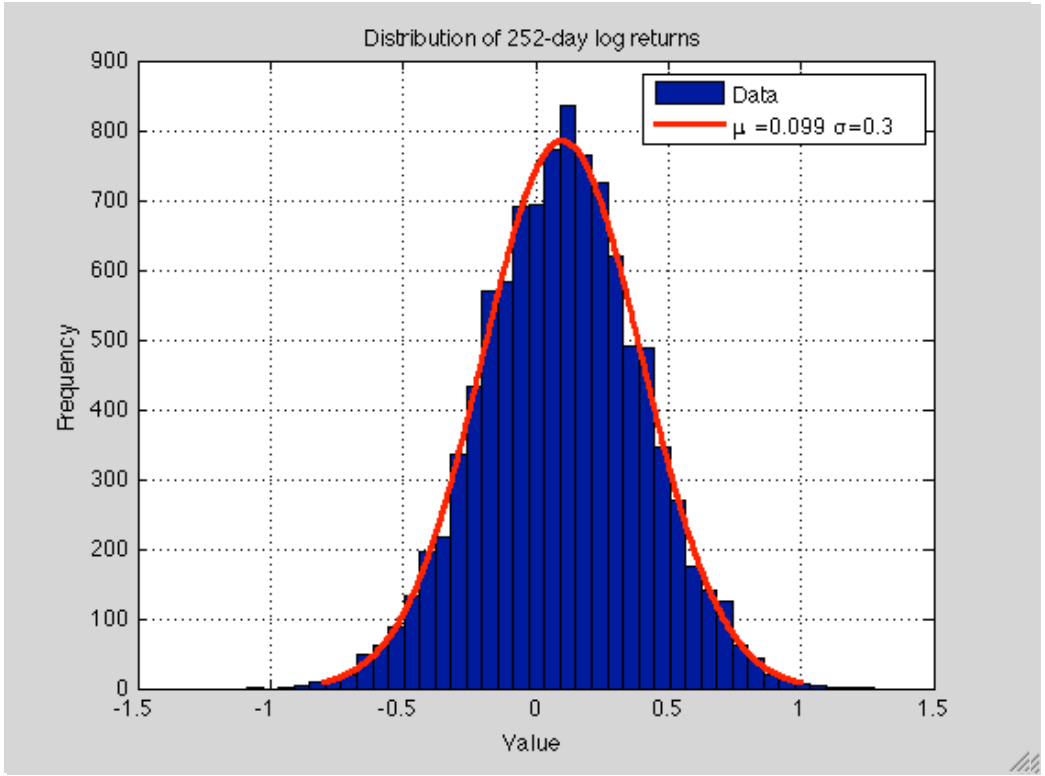
simple return is lognormal

```
> R <- P[Nt+1,] - 1

> mean(R)
[1] 0.1547869
> exp(mu+sigma^2/2)-1
[1] 0.1560396

> sd(R)
[1] 0.3563872
> sqrt(exp(2*mu+sigma^2)*(exp(sigma^2)-1))
[1] 0.3547634

hist(R,breaks=50)
hist(log(1+R))
plot.ecdf(1+R)
barplot(sort(1+R))
qqnorm(1+R)
```



$$R = P_T/P_0 - 1$$

$$\mathrm{E}\left[R\right] = \mathrm{E}\left[e^{r(T)} - 1\right] = e^{\mu+\sigma^2/2} - 1$$

$$\mathrm{Var}(R) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

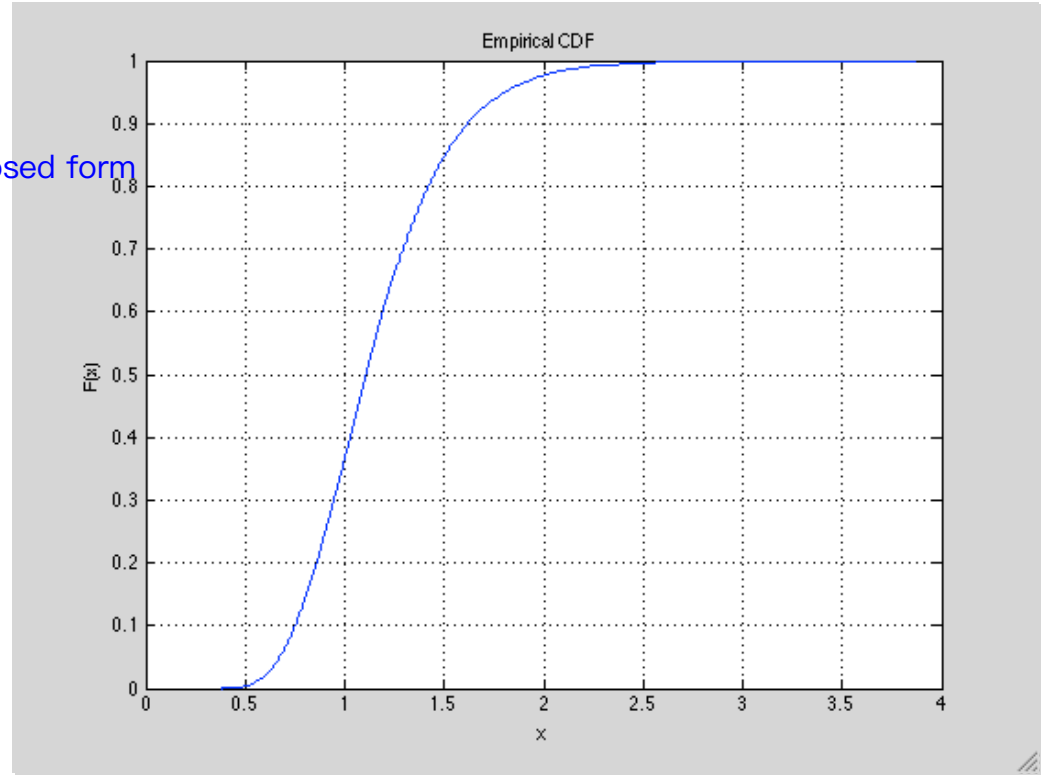
Simulating a lognormal price process

```
> R <- P[Nt+1,] - 1

> mean(R) empirical
[1] 0.1547869
> exp(mu+sigma^2/2)-1 theoretical: sometimes no closed form
[1] 0.1560396

> sd(R)
[1] 0.3563872
> sqrt(exp(2*mu+sigma^2)*(exp(sigma^2)-1))
[1] 0.3547634

hist(R,breaks=50)
hist(log(1+R))
plot.ecdf(1+R)
barplot(sort(1+R))
qqnorm(1+R)
```



$$R = P_T/P_0 - 1$$

$$\mathrm{E}[R] = \mathrm{E}\left[e^{r(T)} - 1\right] = e^{\mu+\sigma^2/2} - 1$$

$$\mathrm{Var}(R) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

theoretical moments are computed by Gaussian intervals

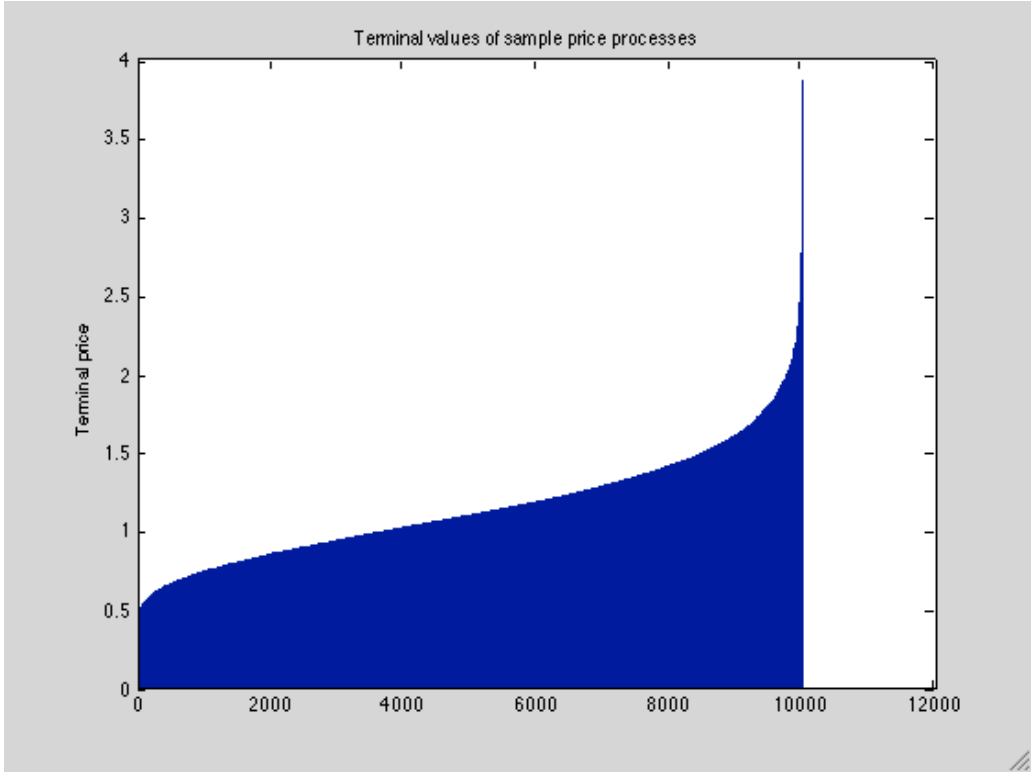
Simulating a lognormal price process

```
> R <- P[Nt+1,] - 1

> mean(R)
[1] 0.1547869
> exp(mu+sigma^2/2)-1
[1] 0.1560396

> sd(R)
[1] 0.3563872
> sqrt(exp(2*mu+sigma^2)*(exp(sigma^2)-1))
[1] 0.3547634

hist(R,breaks=50)
hist(log(1+R))
plot.ecdf(1+R)
barplot(sort(1+R))
qqnorm(1+R)
```



$$R = P_T/P_0 - 1$$

$$\mathrm{E}\left[R\right] = \mathrm{E}\left[e^{r(T)} - 1\right] = e^{\mu+\sigma^2/2} - 1$$

$$\mathrm{Var}(R) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

Q–Q plot typically lets us visualize a comparison of two probability distributions

Simulating a lognormal price process

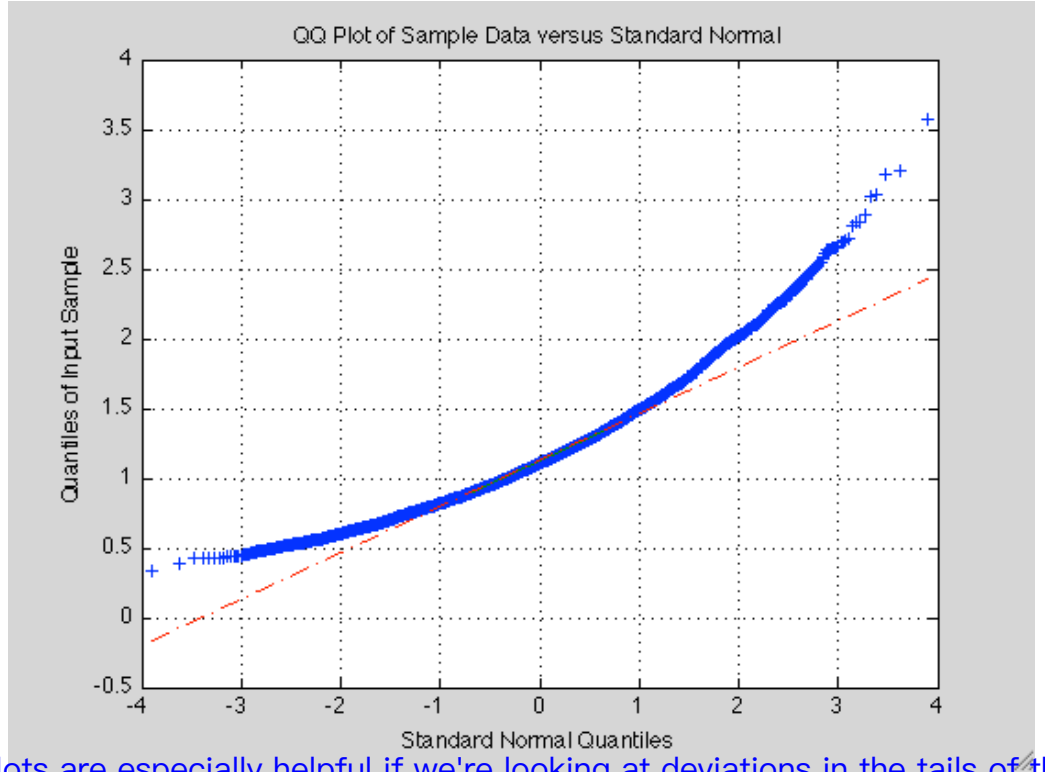
```
> R <- P[Nt+1,] - 1

> mean(R)
[1] 0.1547869
> exp(mu+sigma^2/2)-1
[1] 0.1560396

> sd(R)
[1] 0.3563872
> sqrt(exp(2*mu+sigma^2)*(exp(sigma^2)-1))
[1] 0.3547634

hist(R,breaks=50)
hist(log(1+R))
plot.ecdf(1+R)
barplot(sort(1+R))
qqnorm(1+R)
```

A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y–coordinate) plotted against the same quantile of the first distribution (x–coordinate). This defines a parametric curve where the parameter is the index of the quantile interval



$$R = P_T/P_0 - 1$$

$$\mathrm{E}[R] = \mathrm{E}\left[e^{r(T)} - 1\right] = e^{\mu+\sigma^2/2} - 1$$

$$\mathrm{Var}(R) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

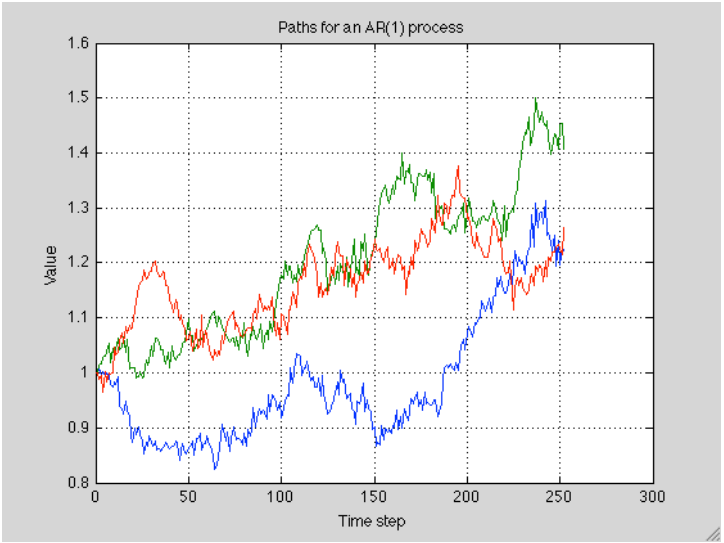
Q–Q plots are especially helpful if we're looking at deviations in the tails of the distribution. A histogram is not so great for looking at outliers in the tails, because the tails are small. Q–Q plot emphasizes the tails and it keeps the boring stuff in the middle (look like a straight line)

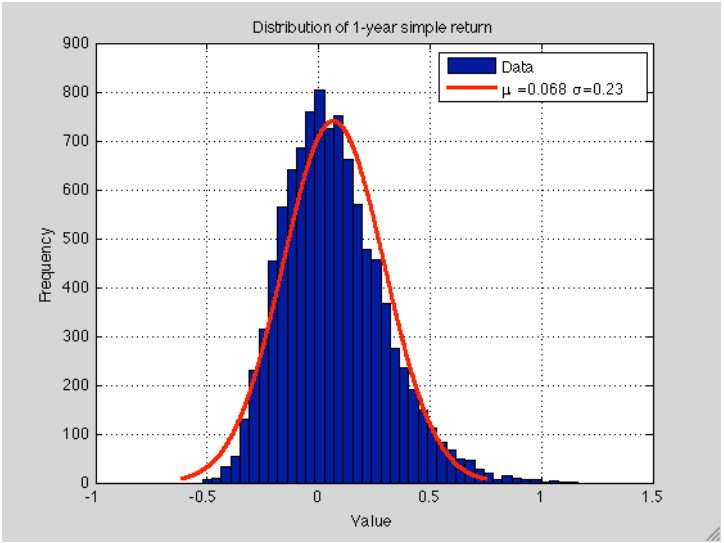
Simulating an AR(1) process

defined recursively (depend on prior history), not once

```
lambda   <- 0.4; mu <- 0.1     # Set mean-reversion strength and drift
R        <- matrix(0,Nt,Np)    # Reserve space for sample processes
epsilon  <- matrix(rnorm(Nt*Np,sd=sigma*sqrt(dt)),nrow=Nt) # Simulate noise
for (t in 2:Nt)
{                          # Run simulation process forward in time
  R[t,] <- (1+lambda)*(mu*dt) - lambda*R[t-1,] + epsilon[t,]
}
r <- log(1+R)              # Interpret R as simple return, r as continuous return
acf(R[,1])                 # Show autocorrelation coefficients as function of lag
```

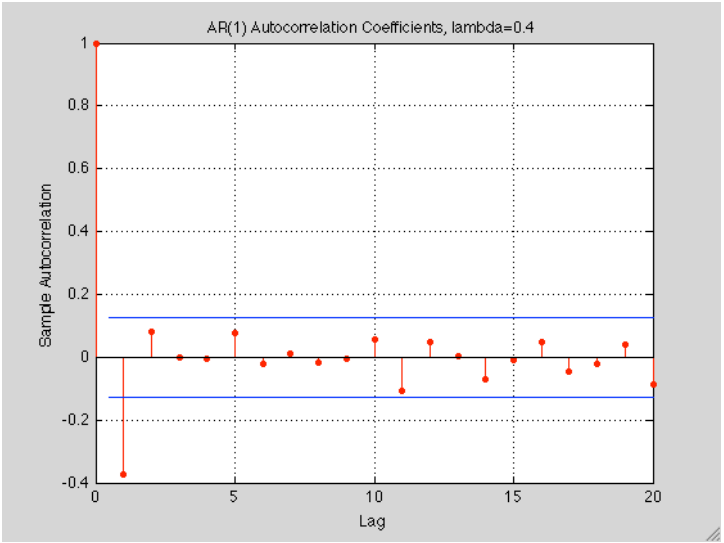
returns of random walk are all independent, I didn't need to do the for loop. I could have used some matrix algebra to do everything once

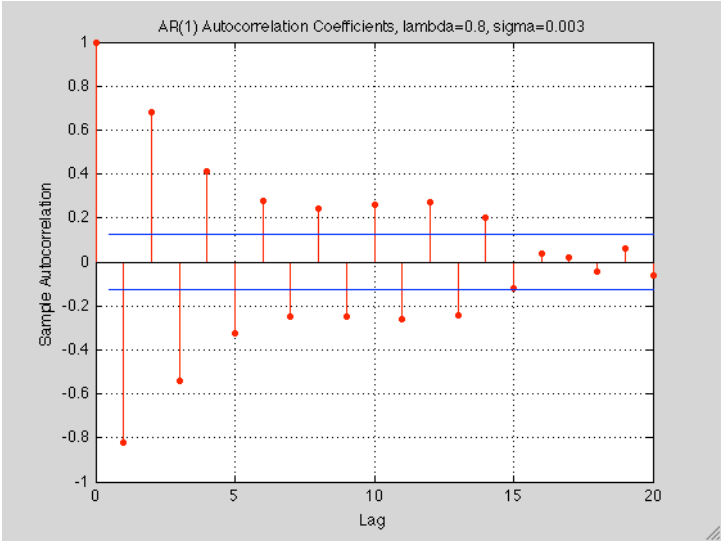




Simulating an AR(1) process

```
lambda   <- 0.4; mu <- 0.1     # Set mean-reversion strength and drift
R        <- matrix(0,Nt,Np)    # Reserve space for sample processes
epsilon  <- matrix(rnorm(Nt*Np,sd=sigma*sqrt(dt)),nrow=Nt) # Simulate noise
for (t in 2:Nt)
{                          # Run simulation process forward in time
  R[t,] <- (1+lambda)*(mu*dt) - lambda*R[t-1,] + epsilon[t,]
}
r <- log(1+R)              # Interpret R as simple return, r as continuous return
acf(R[,1])                 # Show autocorrelation coefficients as function of lag
```





Monte Carlo Methods

MIT MANAGEMENT SLOAN SCHOOL

Summary

- Monte Carlo sampling uses computer random number generators to simulate data that follows a given model, process, or distribution

- Simulations provide an idealized testing environment for theory, statistics, and financial analytics. They provide a "best case" setting since
  - The "true" data-generating process is rarely known, if it even exists
  - Real financial history happens only once and can never be repeated

- Asset price paths are simulated by drawing successive period returns from chosen distributions.
  - Closed-form analytics are replaced by statistical approximations.
  - Results are subject to sampling error, machine limitations.
  - Advanced techniques can greatly improve efficiency and accuracy.

Testing the Random Walk

Finance at MIT
Where ingenuity drives results

A random walk down Wall Street

Do stock prices follow random walks?

- Model for returns drawn from lognormal distribution

$$X_t = X_{t-1} + r_t, \qquad X_t \equiv \log P_t/P_0, \quad r_t = X_t - X_{t-1} = \log P_t/P_{t-1}$$
$$r_t = \mu + \epsilon_t, \qquad \epsilon_t = \sigma z_t \sim \mathcal{N}(0, \sigma^2)$$

- More generally, random shocks could be IID, drawn from other distributions

Random walks, efficient markets, and the real world

How **do** stock prices behave?
How **should** stock prices behave?

- Deterministic?
- Noisy?
- Random?

- Are markets **efficient**?
- Are asset prices **predictable**?

These are empirical questions.
Let's look at the data.

Random Walks in Stock Market Prices

by Eugene F. Fama

FOR MANY YEARS economists, statisticians, and teachers of finance have been interested in developing and testing models of stock price behavior. One important model that has evolved from this research is the theory of random walks. This theory casts serious doubt on many other methods for describing and predicting stock price behavior — methods that have considerable popularity outside the academic world. For example, we shall see later that if the random walk theory is an accurate description of reality, then the various "technical" or "chartist" procedures for predicting stock prices are completely without value.

In general the theory of random walks raises challenging questions for anyone who has more than a passing interest in understanding the behavior of stock prices. Unfortunately, however, most discussions of the theory have appeared in technical academic journals and in a form which the non-mathematician would usually find incomprehensible. This article describes, briefly and simply, the theory of random walks and some of the important issues it raises concerning the work of market analysts. To preserve brevity some aspects of the theory and its implications are omitted. More complete (and also more technical) discussions of the theory of random walks are available elsewhere; hopefully the introduction provided here will encourage the reader to examine one of the more rigorous and lengthy works listed at the end of this article.

Common Techniques for Predicting Stock Market Prices

In order to put the theory of random walks into perspective we first discuss, in brief and general terms, the two approaches to predicting stock prices that are commonly espoused by market professionals. These are (1) "chartist" or "technical" theories and (2) the theory of fundamental or intrinsic value analysis.

The basic assumption of all the chartist or technical theories is that history tends to repeat itself, i.e., past patterns of price behavior in individual securities will tend to recur in the future. Thus the way to predict stock prices (and, of course, increase one's potential

Eugene F. Fama is Assistant Professor of Finance, Graduate School of Business, The University of Chicago. The author is indebted to his colleagues William Alberts, David Green, Merton Miller, and Harry Roberts for their helpful comments and criticisms.

This article has been reprinted as paper No. 16 in the current series of Selected Papers of the Graduate School of Business, The University of Chicago.

SEPTEMBER-OCTOBER 1965

gains) is to develop price behavior in or recurrence.

Essentially, then, knowledge of the p predict the probabl statistician would c suming that succes securities are depen theories assume that to any given day is change for that day.

The techniques of rounded by a certain as a result most ma suspect. Thus it is chartist is relatively Rather the typical a as fundamental ana The assumption of is that at any point an intrinsic value (c equilibrium price) w tial of the security. depends in turn on of management, out omy, etc.

Through a carefu the analyst should, whether the actual p its intrinsic value. I intrinsic values, the trinsic value of a se diction of its future predictive procedur

The Theory of Ran

Chartist theories analysis are really sional and to a larg torically, however, academic people, pr who adhere to a rad analysis—the theory prices. The remaind a discussion of this t

*Probably the best proach to predicting

Proof That Properly Anticipated Prices Fluctuate Randomly

Paul A. Samuelson
Massachusetts Institute of Technology

The Enigma Posed

"In competitive markets there is a buyer for every seller. If one could be sure that a price will rise, it would have already risen." Arguments like this are used to deduce that competitive prices must display price changes over time, $X_{t+1} - X_t$, that perform a random walk with no predictable bias.

Is this a correct fact about well-organized wheat or other commodity markets? About stock exchange prices for equity shares? About futures markets for wheat or other commodities, as contrasted to the movement of actual "spot prices" for the concrete commodity.

Or is it merely an interesting (refutable) hypothesis about actual markets that can somehow be put to empirical testing?

Or is it a valid deduction (like the Pythagorean Theorem applicable to Euclidean triangles) whose truth is as immutable as $2 + 2 = 4$? Does its truth follow from the very definition of "free, competitive markets?" (If so, can there fail to exist in New York and London actual stock and commodity markets with those properties; and must any failures of the "truism" that turn up be attributable to "manipulation," "thinness of markets," or other market imperfections?)

The more one thinks about the problem, the more one wonders what it is that could be established by such abstract argumentation. Is the fact that American stocks have shown an average annual rise of more than 5 per cent over many decades compatible with the alleged "fair game" (or martingale property) of an unbiased random walk? Is it an exception that spot wheat prices generally rise (presumably because of storage costs) from the July harvest time to the following spring and drop during June? Is the fact that the price of next July's future shows much less strong seasonal patterns a confirmation of the alleged truism? If so, what about the alleged Keynes-Hicks-Houthakker-Cootner pattern of "normal backwardation," in which next July's wheat future could be expected to rise in price a little from July harvest to, say, the following March (as a result of need of holders of the crop to coax out, at a cost, risk-disliking speculators with whom to make short-hedging transactions); and what about the Cootner pattern in which, once wheat stocks become low in March, processors wishing to be sure of having a minimum of wheat to process, seek short-selling speculators with whom to make long-hedging transactions, even at the cost of having the July quotation dropping a little in price in months like April and May?

Consideration of such prosaic and mundane facts raises doubt that there is anything much in celestial *a priori* reasoning from the axiom that what can be perceived about the future must already be "discounted" in current price quotations. Indeed, suppose that all the participants in actual markets are necessarily boobs when it comes to foreseeing the unforeseeable future. Why should "after-the-fact" price changes show *any* systematic pattern, such as non-bias? Are the very mathematical notions of probability of any relevance to actual market quotations? If so, how could we decide that this is indeed so?

Model predictions

Two-parameter model, as written, holds for all time periods with same parameter values.

$$r_t = \mu + \sigma z_t$$

too long period will lead to different parameter values in different parts

increase number of samples not by longer period, but by increasing observation frequency

So re-estimation in different periods or with different data should yield same parameter values. How significant are variations?

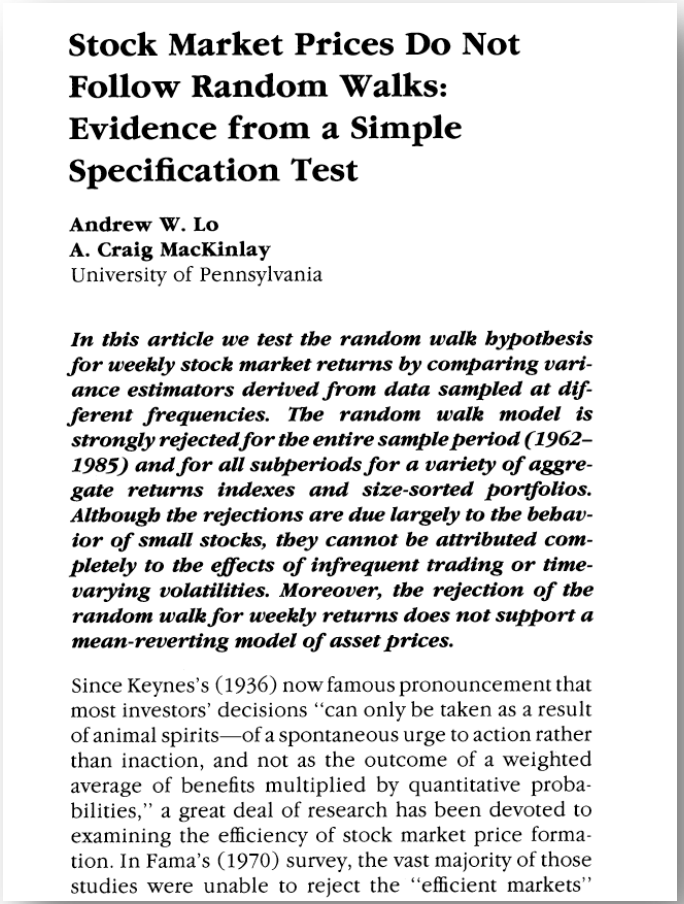
Model predicts aggregation properties over time.

- **Scaling relationships**
- Case study: Tootsie Roll (TR)
  - Founded: 1896
  - Found by reddit: 2020
  - "How many licks?"

Testing the random walk

- Scaling behavior
- Variance ratio test
- Frequency-dependence
- Long-range behavior
- Alternative models
- Serial correlation
- Implications for market efficiency

And the idea was that not only might random walks provide a good model for the way stock prices seem to behave in the market. It would explain why there's so much randomness and noise and why not everybody can get rich trading stocks.



Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test

Andrew W. Lo
A. Craig MacKinlay
University of Pennsylvania

In this article we test the random walk hypothesis for weekly stock market returns by comparing variance estimators derived from data sampled at different frequencies. The random walk model is strongly rejected for the entire sample period (1962–1985) and for all subperiods for a variety of aggregate returns indexes and size-sorted portfolios. Although the rejections are due largely to the behavior of small stocks, they cannot be attributed completely to the effects of infrequent trading or time-varying volatilities. Moreover, the rejection of the random walk for weekly returns does not support a mean-reverting model of asset prices.

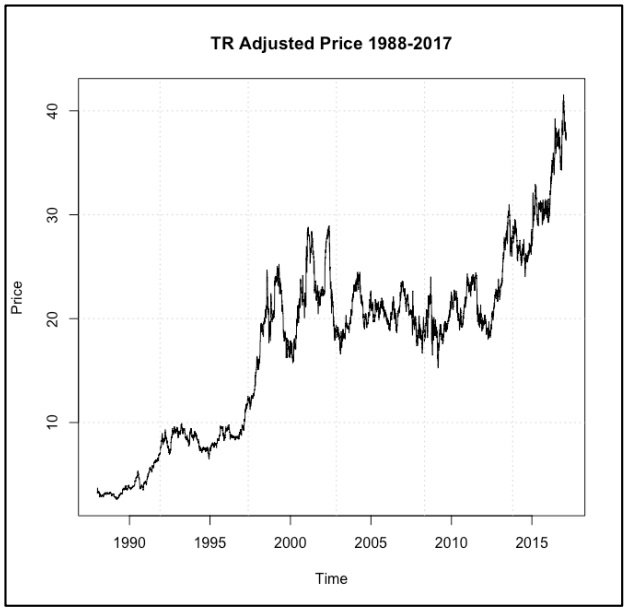
Since Keynes's (1936) now famous pronouncement that most investors' decisions "can only be taken as a result of animal spirits—of a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of benefits multiplied by quantitative probabilities," a great deal of research has been devoted to examining the efficiency of stock market price formation. In Fama's (1970) survey, the vast majority of those studies were unable to reject the "efficient markets"

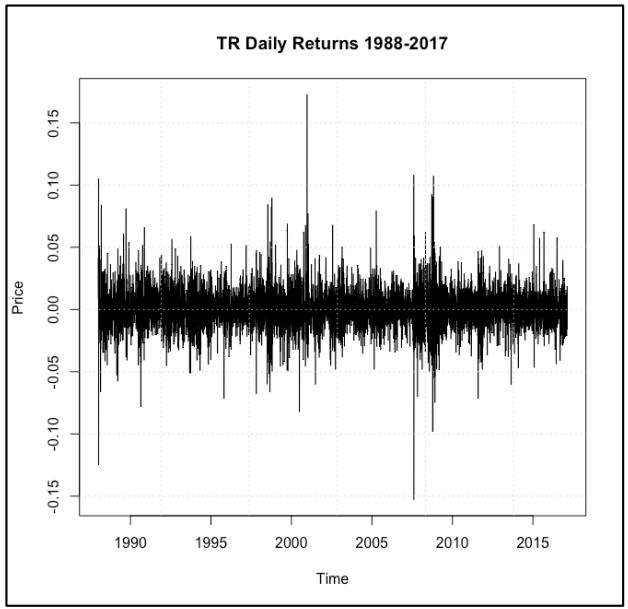
we have this notion that if markets were efficient in the sense that investors are taking into account the available information, then the only reason prices should change is the arrival of new information that hadn't been anticipated. That's sort of the definition of news. And in such a world, prices would behave randomly. So somewhat paradoxically, randomness of prices would be a sign of markets operating efficiently.

Exploratory data analysis

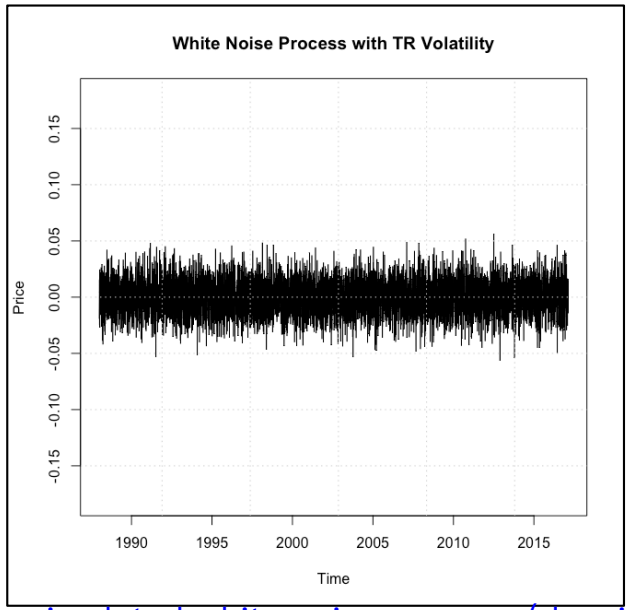
data from Yahoo Finance
in R using the tidyquant package and the tq_get() function

In the case of Tootsie Roll, a quick plot of the return series appears to show non-constant variance.





log returns from the successive price differences on a daily level



simulated white noise process (drawing from a random, normal distribution with constant volatility)

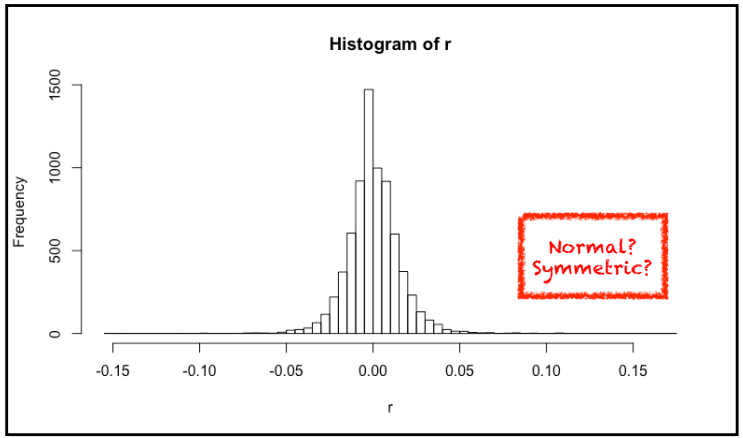
Exploratory data analysis

Summarize vital statistics

- Length of series
- Max, min, sum, mean, std. deviation, etc.
- Example: Tootsie Roll (TR): mean return and volatility: 8.5% and 24.4%

Histogram price levels are not important to investors

- Useful for *returns*, not prices (Why?)




```
> summary(tr) # Summary of imported Yahoo data for TR
      Date                 Open             High             Low
Min.   :1987-12-30   Min.   :19.55   Min.   :19.66   Min.   :19.00
1st Qu.:1995-04-03   1st Qu.:28.95   1st Qu.:29.17   1st Qu.:28.64
Median :2002-07-18   Median :33.00   Median :33.26   Median :32.61
Mean   :2002-07-18   Mean   :38.72   Mean   :39.07   Mean   :38.37
3rd Qu.:2009-10-29   3rd Qu.:42.00   3rd Qu.:42.50   3rd Qu.:41.75
Max.   :2017-02-15   Max.   :83.50   Max.   :84.50   Max.   :82.50
      Close            Volume          Adj Close
Min.   :19.46   Min.   :      0   Min.   : 2.587
1st Qu.:28.94   1st Qu.:  34800   1st Qu.: 8.797
Median :32.94   Median :  68300   Median :19.457
Mean   :38.74   Mean   :  90813   Mean   :17.417
3rd Qu.:42.12   3rd Qu.: 114750   3rd Qu.:22.523
Max.   :83.75   Max.   :5819300   Max.   :41.550

> P <- tr$`Adj Close` # Define price using split&div adjusted values
> N <- length(P); N
[1] 7342
> r <- diff(log(P)) #Define log returns from successive daily prices
> summary(r)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.1527000 -0.0078470  0.0000000  0.0003375  0.0082100  0.1726000
> mean(r)*252    # Mean return for TR (annualize by 252 days/year)
[1] 0.08506148
> sd(r)*sqrt(252) # Volatility of TR (annualize with square root!)
[1] 0.2442396
> hist(r, breaks=50)
```

Flavors of Random Walk

used

RW1: IID – Independent and identically distributed increments (or "innovations")

X: logarithm of the price

$$X_t = X_{t-1} + \mu + \epsilon_t, \quad \epsilon_t \sim \text{IID}(0, \sigma^2), \quad \mathbb{E}[\epsilon_t \epsilon_{t'}] = \sigma^2 \delta_{tt'}$$
$$\mathbb{E}[X_t | X_0] = X_0 + \mu t,$$
$$\text{Var}(X_t | X_0) = \sigma^2 t$$

RW2: INID – Independent but non-identically distributed increments

- For example, each day's return is independently lognormal, but volatility parameter varies each day.

RW3: Dependent, non-identical innovations, but increments are uncorrelated

- For example, volatility clustering modeled with correlated **squares** of increments

$$\mathbb{E}[\epsilon_t \epsilon_{t'}] = 0, \quad \mathbb{E}[\epsilon_t^2 \epsilon_{t'}^2] \neq 0, \quad t \neq t'$$

dependent at a higher order

uncorrelated without them being independent

Variance ratios

- Start with returns at some base frequency (weekly, daily, hourly, etc.)
- Aggregate to form time series of 2-period, 3-period, q-period returns with common terminal observation

$$r_t^{(2)} = r_t + r_{t-1} = \log(P_t/P_{t-2}), \quad r_t^{(q)} = \sum_{i=1}^{q} r_{t-i+1} = \log(P_t/P_{t-q})$$

- If returns are **uncorrelated**, then variance computed from each series is proportional to its length

$$\mathrm{Var}(r_t^{(q)}) = q\mathrm{Var}(r_t)$$

- Therefore test whether or not the **variance ratio**

$$\frac{\mathrm{Var}(r_t^{(q)})}{q\mathrm{Var}(r_t)} = 1$$

Variance and ratios

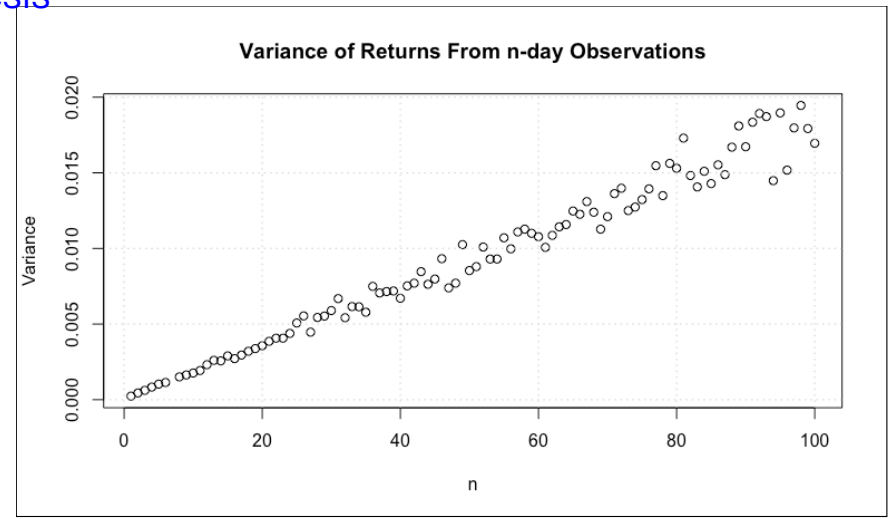
Random walk model predicts that **variance** has simple **linear scaling** with sampling interval

- Example: Tootsie Roll data set 1988-2017 from Yahoo Finance
  - Close to linear? yes! fit random walk hypothesis
  - Close enough?

```
Variance <- var(diff(log(P)))

for (n in 2:100) {
  Variance[n] <- var(diff(log(P[seq(from=n, to=N, by=n)])))
}

plot(Variance,xlab="n",main="Variance of Returns From n-day
Observations");grid()
```



Variance ratio estimators and data

- Need to match length of time series
- Longer base periods have fewer contributions
- Asymptotic distribution to determine expected size of deviations from unity
- Bias corrections
- Overlapping returns

$$\widehat{\mu} = \frac{1}{T}\sum r_t,$$

$$\widehat{\sigma}_a^2 = \frac{1}{T}\sum (r_t - \widehat{\mu})^2$$

$$\widehat{\sigma}_b^2(q) = \frac{1}{T}\sum (r_t - q\widehat{\mu})^2$$

$$\sqrt{T}\left(\widehat{\sigma}_b^2(q) - \sigma^2\right) \sim \mathcal{N}(0, q\sigma^4)$$

$$\widehat{\sigma}_c^2(q) = \frac{1}{qT}\sum_{k=q}^{nq} (\log(S_{t_k}/S_{t_{k-q}}) - q\widehat{\mu})^2,$$

$$T = nq$$

Variance ratio test

Test performed to see if ratio equals one

$$\widehat{VR}(q) = \frac{\hat{\sigma}_b^2(q)}{q\hat{\sigma}_a^2}, \quad \sqrt{T}\left(\widehat{VR}(q) - 1\right) \sim \mathcal{N}(0, 2(q-1))$$

Define test statistic to have normalized distribution of deviations under null

$$z(q) = \sqrt{\frac{T}{2(q-1)}}\left(\widehat{VR}(q) - 1\right) \sim \mathcal{N}(0,1)$$

if the random walk holds, we should expect z to lie between plus or minus 2 , 9 5 % of the time. if we get values that are much larger than that, then that would be evidence for rejecting a random walk

Variance ratio tests: as easy as *a, b, c…*

Definitions refined to improve power, remove bias, and enable comparison of multiple window sizes.

- Initially consider $T = nq + 1$ sequential observations and group them q at a time for q-period windows.
- Final form uses overlapping windows, which improves power.

we don't need to compute the intermediate returns, because we can just look at the endpoints

$$\hat{\mu} \equiv \frac{1}{T}(X_T - X_0)$$

$$\hat{\sigma}_a^2 \equiv \frac{1}{T}\sum_{t=1}^{T}(X_t - X_{t-1} - \hat{\mu})^2$$ base frequency

$$\hat{\sigma}_b^2(q) \equiv \frac{1}{nq}\sum_{k=1}^{nq}(X_{qk} - X_{qk-q} - q\hat{\mu})^2$$ q period observations

$$\hat{\sigma}_c^2(q) \equiv \frac{1}{nq^2}\sum_{t=q}^{nq}(X_t - X_{t-q} - q\hat{\mu})^2$$ overlapping periods

Variance and ratios

- Lo and MacKinlay add further refinements and consider whether variance ratios over different *q*-day window sizes are statistically identical
- The idea is to construct a **sampling statistic** *z(q)* that follows a standard Normal distribution **if** the random walk hypothesis holds.
- Large values of *z(q)* are unlikely. The *p*-value gives the probability of observing a value at least as large.

```
variance.c <- function(X, q) {
# Compute variance statistic from overlapping q-period windows
# See Lo & MacKinlay (1988), p. 47, Eq. 12

  T  <- length(X) - 1
  mu <- (X[T+1] - X[1])/T
  m  <- (T-q)*(T-q+1)*q/T
  sumsq <- 0
  for (t in q:T) {
  sumsq <- sumsq + (X[t+1] - X[t-q+1] - q*mu)^2
  }
  return(sumsq/m)
}

z <- function(X, q) {
# Compute sampling statistic for variance ratio
# See Lo & MacKinlay (1988), p. 47, last line (after Eqs. 12-14)
  T <- length(X) - 1
  c <- sqrt(T*(3*q)/(2*(2*q-1)*(q-1)))
  M <- variance.c(X,q)/variance.c(X,1) - 1
  z <- c*M
  return(z)
}

Vc      <- 0; for (q in 1:100) {Vc[q] <- variance.c(log(P),q)}
zstats  <- 0; for (q in 2:100) {zstats[q] <- z(log(P),q) }
pValues <- 2*pnorm(-abs(zstats)) probabilities of extreme results
barplot(zstats)
```

Variance ratio test

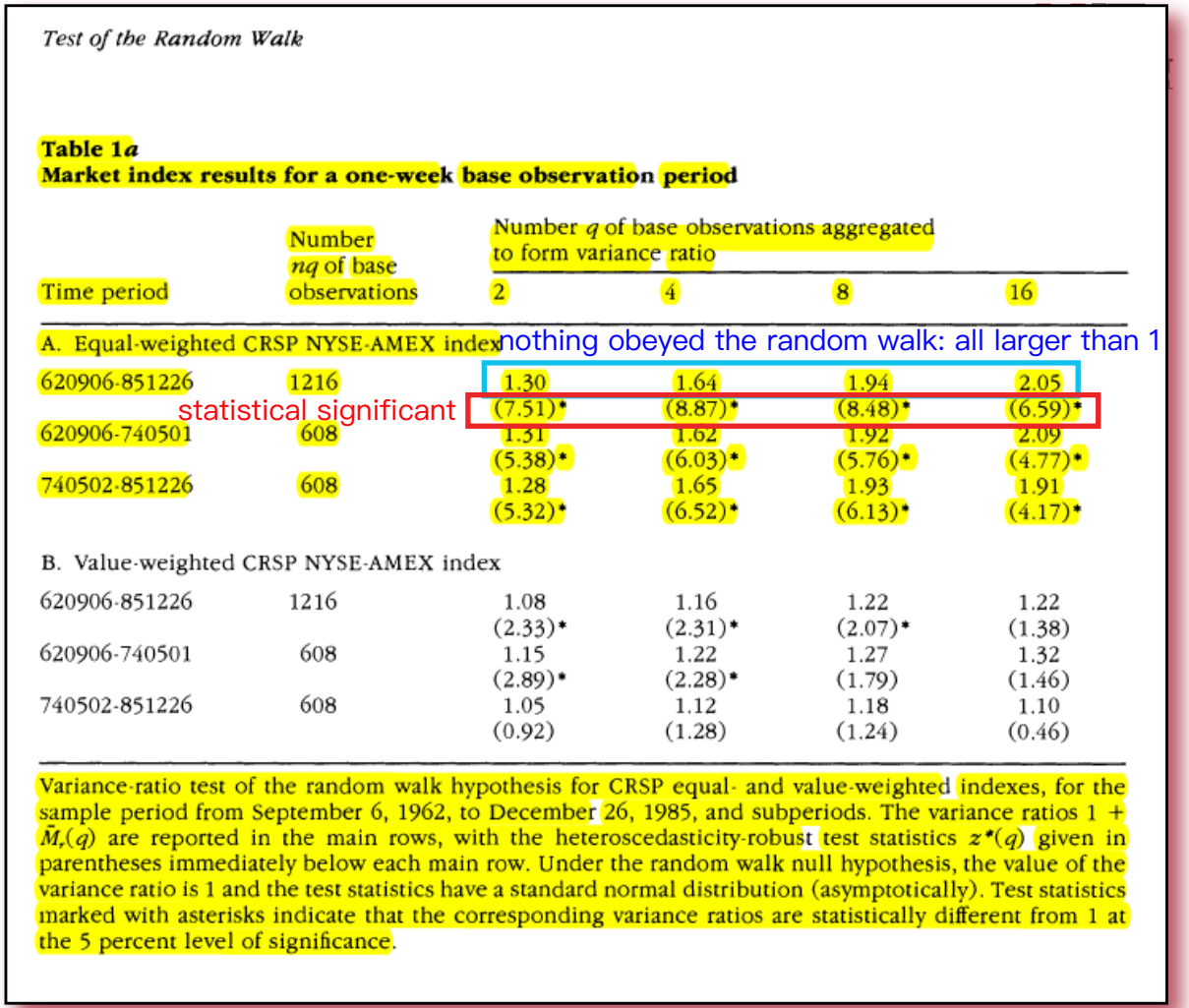
Bias correction and improved statistics from overlapping terms

$$\bar{\sigma}_c^2 = \frac{\hat{\sigma}_c^2}{(1 - 1/n + 1/nq)(1 - 1/n)}, \quad \sqrt{nq}\left(\overline{VR}(q) - 1\right) \sim \mathcal{N}(0, 2(2q-1)(q-1)/3q)$$

$$z(q) = \sqrt{nq}\left(\overline{VR}(q) - 1\right)\sqrt{\frac{3q}{2(2q-1)(q-1)}}$$

Variance ratio test

- Evaluation for stocks, indices
- Period and frequencies
- Baseline reference period
- Data presentation



Test of the Random Walk

Table 1*a*
Market index results for a one-week base observation period

Time period	Number nq of base observations	Number q of base observations aggregated to form variance ratio: 2	4	8	16
A. Equal-weighted CRSP NYSE-AMEX index					
620906-851226	1216	1.30	1.64	1.94	2.05
		(7.51)*	(8.87)*	(8.48)*	(6.59)*
620906-740501	608	1.31	1.62	1.92	2.09
		(5.38)*	(6.03)*	(5.76)*	(4.77)*
740502-851226	608	1.28	1.65	1.93	1.91
		(5.32)*	(6.52)*	(6.13)*	(4.17)*
B. Value-weighted CRSP NYSE-AMEX index					
620906-851226	1216	1.08	1.16	1.22	1.22
		(2.33)*	(2.31)*	(2.07)*	(1.38)
620906-740501	608	1.15	1.22	1.27	1.32
		(2.89)*	(2.28)*	(1.79)	(1.46)
740502-851226	608	1.05	1.12	1.18	1.10
		(0.92)	(1.28)	(1.24)	(0.46)

Variance-ratio test of the random walk hypothesis for CRSP equal- and value-weighted indexes, for the sample period from September 6, 1962, to December 26, 1985, and subperiods. The variance ratios $1 + \hat{M}_r(q)$ are reported in the main rows, with the heteroscedasticity-robust test statistics $z^*(q)$ given in parentheses immediately below each main row. Under the random walk null hypothesis, the value of the variance ratio is 1 and the test statistics have a standard normal distribution (asymptotically). Test statistics marked with asterisks indicate that the corresponding variance ratios are statistically different from 1 at the 5 percent level of significance.

nothing obeyed the random walk: all larger than 1

statistical significant

Variance and ratios

Random walk model predicts that **volatility** has **square-root scaling** with sampling interval

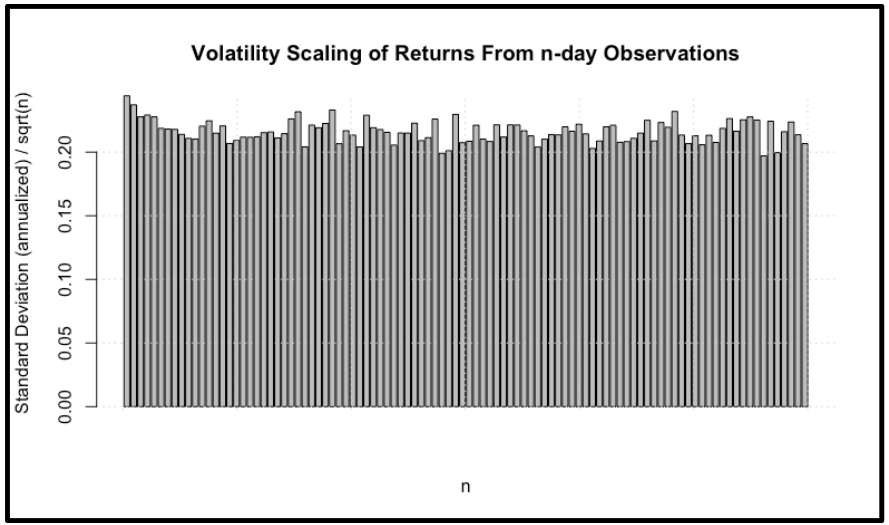
Example: Tootsie Roll 1988-2017

Let's present data so that it appears **constant** if the model is correct:

- Compute annualized volatility from every day, every other day, every third day, etc.
- Divide each volatility by sqrt(*n*)
- Does the data support the model? How significant are the variations?

```
> sigma <- sqrt(252)*sd(diff(log(P)))

> for (n in 2:100) {
  sigma[n] <- sqrt(252/n)*sd(diff(log(P[seq(from=n, to=N, by=n)])))
  }> barplot(sigma,xlab="n",ylab="Standard Deviation (annualized) /
sqrt(n)",main="Volatility Scaling of Returns From n-day
Observations");grid()
```



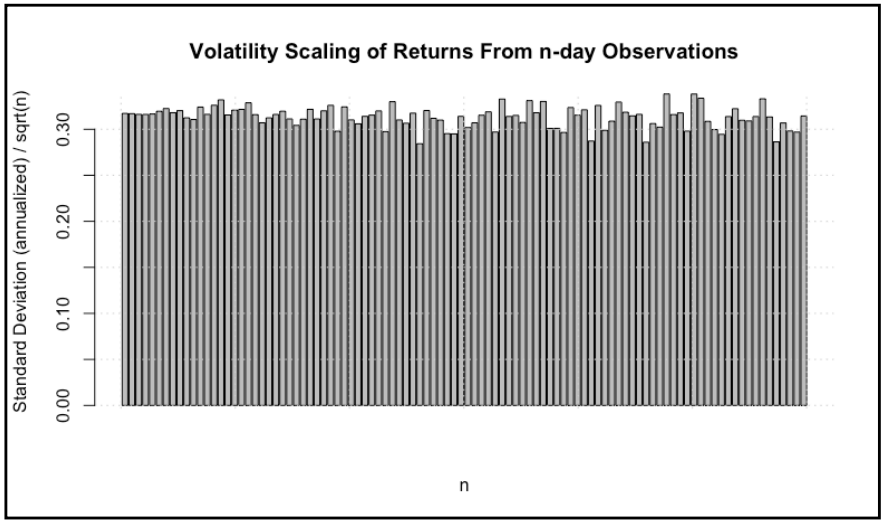
Variance and ratios

How **significant** are the variations?

- Quick & dirty visualization: compare with **simulated** price path from a pure random-walk process that is subject to statistical error
- How can we minimize statistical error? How do we identify departures that violate the model's expected behavior?

```
> P.MC <- exp(cumsum(rnorm(N)*0.02)) # Monte Carlo returns 32% vol
> sigma.MC <- sqrt(252)*sd(diff(log(P.MC)))

> for (n in 2:100) {
  sigma.MC[n] <- sqrt(252/n)*sd(diff(log(P.MC[seq(from=n, to=N,
by=n)])))
  }
> barplot(sigma.MC,xlab="n",ylab="Standard Deviation (annualized) /
sqrt(n)",main="Volatility Scaling of Returns From n-day
Observations");grid()
```



we see the same general qualitative behavior on this plot as we saw in the previous one.So it's possible that what we're looking at could just be statistical error, or it's possible that there could be systematic departures from the random walk when we do the detailed analysis

Testing the Random Walk

Summary

- The random walk model predicts that variances scale linearly with time.
- The variance ratio test of Lo & MacKinlay analyzes how variances scale as the observation frequency changes. It can be applied to individual stocks, portfolios of stocks, or dynamic trading strategies – anything with an empirical time series of returns.
- The test is not dependent on the distribution of returns, only their independence. It has been refined to handle other complexities.
- Rejecting the random walk opens new questions
  - Is there a better model? Or no model at all?
  - In what sense, if any, are markets efficient?
  - Are asset prices predictable?

References

- Books
  - Campbell, Lo, and MacKinlay (1977) – "*Econometrics of Financial Markets,*" Princeton
  - Feller, William (1968) – "*Introduction to Probability and Its Applications (3rd ed.),*" Wiley
  - Glasserman, Paul (2004) – "*Monte Carlo Methods in Financial Engineering,*" Springer
  - Tsay, Ruey S. (2010) – "*Analysis of Financial Time Series (3rd ed.),*" Wiley

- Articles
  - Fama, Eugene (1965) – "*Random Walks in Stock Market Prices,*" Financial Analysts Journal, Vol. 21, No. 5, pp. 55-59
  - Samuelson, Paul. (1965) – "*Proof that Properly Anticipated Prices Fluctuate Randomly,*" Industrial Management Review, 6:2, 41-49
  - Lo, Andrew W. & A. Craig MacKinlay (1988) – "*Stock Market Prices do not Follow Random Walks: Evidence from a Simple Specification Test,*" Review of Financial Studies, Vol. 1, No. 1 (Spring, 1988), pp. 41-66