



# Outline

- 1 Motivation: The Short Run and the Long Run
  - What is High Frequency Data?
  - Connection between Short and Long Run
- 2 Statistical Background
  - Confidence Intervals and Testing
  - Asymptotics
- 3 High Frequency Data
  - Setting for inference
  - Parametric inference
  - Nonparametric inference
  - Themes and challenges, Web references
  - Limit Theory vs Data; The Hidden Semimartingale Model

# Welcome to Fin Math 33170/Stat 33910

## **Financial Statistics: Time Series, Forecasting, Mean Reversion, and High Frequency Data FINM 33170 and STAT 33910**

- Instructor: Per Mykland
- Teaching assistants:
  - Ahmed Bou-Rabee
  - Yi Wang
- Further and updated information:
- <http://www.stat.uchicago.edu/~mykland/33170S21/index.html>

◀ ◻ ▶ ◀ ◻ ◻ ▶ ◀ ≡ ≡ ▶ ◀ ≡ ≡ ▶ ≡ ≡ ≡ ↺ 🔍 ↻

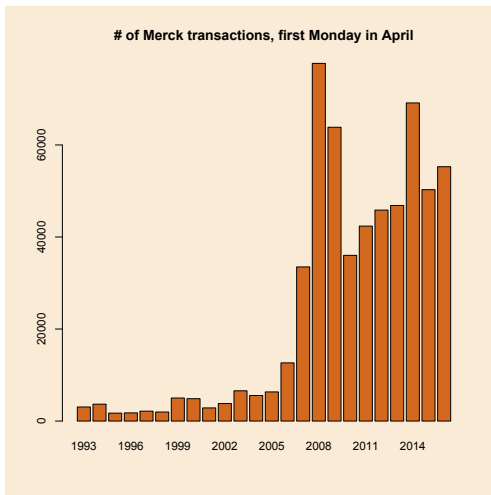
# The Short Run and the Long Run

- **Short run: high frequency data (time measured in seconds or less)**
  - Volatilities (short run variances), covariances, regression, ANOVA, factor analysis, leverage effect
  - market microstructure
  - "Low latency" trading
  - The shortest of frequencies: the price jump
- **Long run: days, months, year, centuries**
  - Time series, forecasting
  - Mean reversion and momentum, cointegration
- No grand unified theory (yet), but...
  - ... maybe the day will come
  - Meanwhile: Useful both separately and in combination
- For whom is this course?
  - **Private sector:** The Trader, the Risk manager, the Investor
  - **Public sector:** The Regulator, the Central Bank
  - **The Observers:** The Academic, the Journalist

# High Frequency Data

- financial prices, volumes, number of trades, order time
- Intra-day:
  - transactions tick-by-tick, from TAQ, Reuters, CME
  - quotes - bid, ask - same sources
  - limit order books, harder to get but more information
  - stocks, bonds, futures, currencies, ...
  - low latency data

# Evolution of Data Size per Day



Note: Merck represents a medium-density data. A liquid stock has more than 200,000 trades per day.

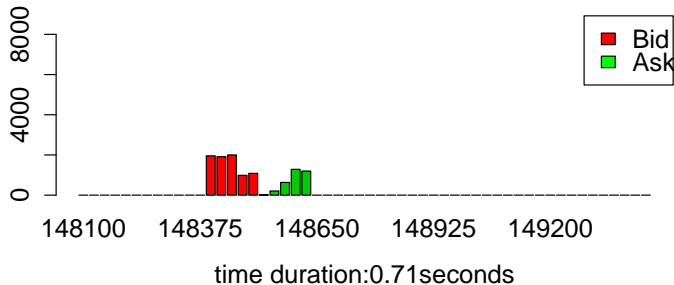
# Intraday Trading: almost time continuous

	Time	Size	Price
Apple April 2, 2012  Number of Trades 102,986	9:00:05.897	100	601.740
	9:00:11.257	100	601.700
	9:00:11.340	100	601.730
	9:00:12.190	100	601.700
	9:00:12.393	500	601.700
	9:00:12.807	200	601.700
	9:00:13.060	100	601.700
	9:00:13.460	100	601.650
	9:00:14.240	100	601.700
	9:00:14.913	100	601.700
	9:00:14.913	200	601.700
	9:00:15.310	100	601.700
	9:00:18.380	100	601.530
	⋮	⋮	⋮

Observation times are : (1) up to milli-seconds per trade,  
 (2) non-equidistant, (3) could be **endogenous**.



# Snapshot of Limit Order Book for E-mini S&P 500



Snapshot from May 1, 2007: horizontal line shows the five best bid prices (red) and five best ask prices (green), while vertical line shows the volume of each quote.

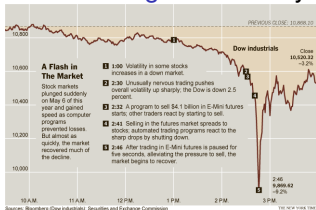
# High Dimension

- Equity Cross Section: over 4000 stocks are traded at NYSE. Each day, NYSE has about 1 billion shares being traded.
- Options: contracts with varying exercise prices, contracts with varying maturity times
- Order Book: varying depth

# Price movement almost path-continuous, but ...

- Flash Crash on May 6 2010: All major US stock indices plunged and rebounded within about 30 minutes. Dow Jones Industrial Average plunged 998.5 points (about 9%), most within minutes.
- Twitter Flash Crash on Tuesday April 23, 2013: Dow quickly plunged 140 points (about 1%) after a false tweet. The S&P 500 lost \$121 billion of its value within minutes.
- 2017 Gold Flash Crash: On Monday June 26, 2017, around 2 billion dollars worth (1.85 millions oz) of Gold futures were sold in the early morning, which triggered the price suddenly plunge by \$18 an ounce (1.6%) before bouncing back \$10 an ounce a minute later.
- Swiss Franc: On 15 January 2015, the SNB scrapped the peg to the Euro

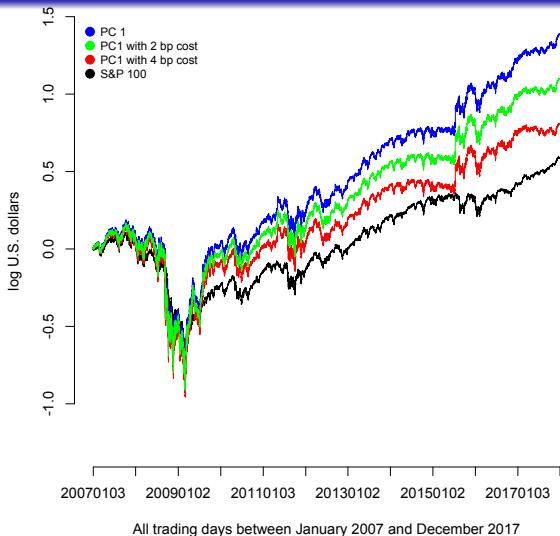
Figure: Intraday Sudden Price Movement



# Example of Short Run Meets Long Run

- Principal Component Analysis (PCA): Find investment weights based on 2500 seconds of data
- S&P 100 index: Value Weighted. Based on long run considerations

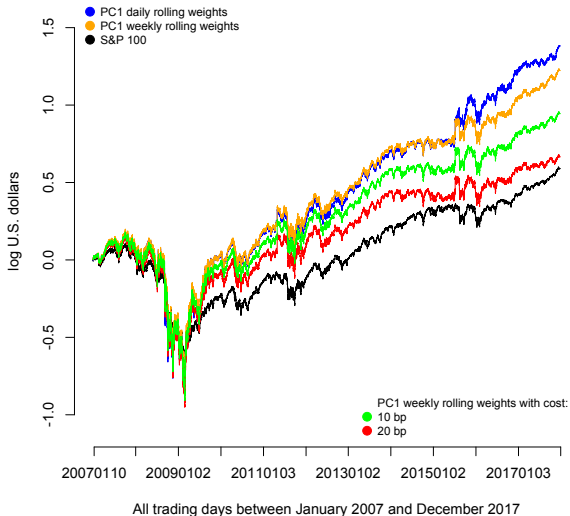
# Unsupervised Learning in Intraday Data: PC1 Portfolio vs. S&P 100: 1 Day Rolling Mean Eigenvector



# Why rolling Mean?

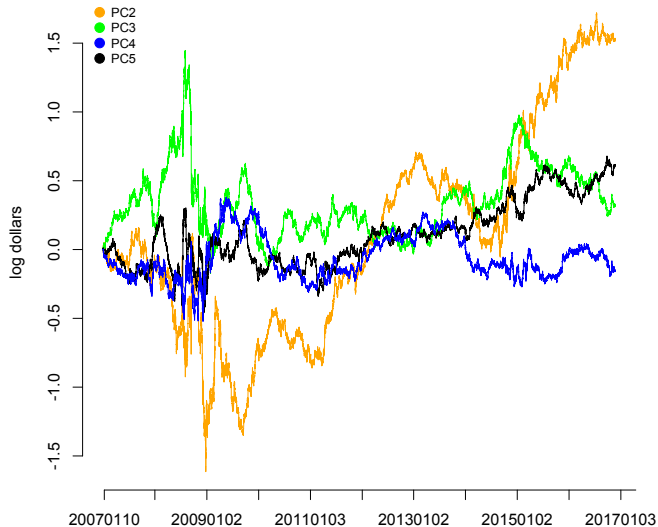
- 2500 seconds based estimators too variable relative to bias
- When trading: rolling mean  $\implies$  less trading cost:
  - 9 period (one day) rolling mean means that only about  $(1/9)^{\text{th}}$  of portfolio is updated every 2500 seconds
  - 45 period (one week) rolling mean means that only about  $(1/45)^{\text{th}}$  of portfolio is updated every 2500 seconds
- Both phenomena documented by plots (a few slides later)
- Overnight position uses same weights as period 1 next day (based on data from trading periods ending at 4 pm on the preceeding day)

# Allowing for higher Trading Cost in PC1: 5 Days Rolling Mean Eigenvector





# Higer Order PC Portfolios



# Plan for course

- Weave a little between short and long run
- Start with some background in statistics, and high frequency data
- Time series
- High frequency data

## Statistical Background

# Approaches to Data Analysis and Statistics

- Formal analysis
  - Testing: seeing whether a structure is present
  - Confidence intervals: setting error limits on estimates
  - Prediction intervals: setting error limits on future outcomes (risk management)
  - Bayesian methods
- Exploratory analysis
  - Finding good graphical representations, or descriptive statistics
  - “Inspirational” text: Edward R. Tufte: *The Visual Display of Quantitative Information*

# Hypothesis Testing and Confidence Intervals: Review

Typical setting (discussion in scalar case):

- $\beta$  is unknown parameter;  $\hat{\beta}_n$  is estimator based on  $n$  observations
- For example,  $(x_i, Y_i)$ ,  $i = 1, \dots, n$ , are observed, generated by

$$Y_i = x_i\beta + \epsilon_i, E(\epsilon) = 0 \quad (1)$$

- $\hat{\sigma}_n^2$  is estimator of  $\sigma_n^2 = \text{Var}(\sqrt{n}(\beta_n - \beta))$
- In setting (1), if the  $\epsilon_i$ 's are iid normal,

$$\frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\hat{\sigma}_n} \text{ has distribution } t_{n-1}$$

- $1 - \alpha$  confidence interval:

- Let  $t_{n-1,\alpha}$  be such that  $P(|T_{n-1}| > t_{n-1,\alpha}) = 1 - \alpha$

- Interval:  $\left| \frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\hat{\sigma}_n} \right| \leq t_{n-1,\alpha}$ , or

$$\beta \in [\hat{\beta}_n - t_{n-1,\alpha}\hat{\sigma}_n/\sqrt{n}, \hat{\beta}_n + t_{n-1,\alpha}\hat{\sigma}_n/\sqrt{n}]$$

# Confidence Intervals

- Property of confidence interval:

$$P(\beta \in \text{CI}) = 1 - \alpha \text{ independently of } \beta, \text{Var}(\epsilon)$$

- $1 - \alpha$  level **test** of  $\beta = \beta_0$  (say,  $\beta_0 = 0$ ):

accept  $H_0 : (\beta = \beta_0)$  if  $\beta \in \text{CI}$ , reject otherwise

- p-value: the  $\alpha$  which is such that the test is right on the border between acceptance or rejection
- Typical  $\alpha$ 's/p-values: 5% or smaller
- Purpose of test: determining presence or absence of certain structures. In the case of simple regression: whether  $Y$  depends on  $x$

# Theory faces reality

The above machinery depends on the following assumption:

*The law of  $\frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\hat{\sigma}_n}$  is independent of the unknown parameters*

This is true only in rare circumstances.

For example, it is true in regression when the errors  $\epsilon_j$  are i.i.d. normally distributed, but it is not true for general distributions of  $\epsilon_j$

What to do?

# Asymptotics

The theory is saved by the following “meta-theorem”:

**Central Limit “Theorem” (CLT).** *Under a variety of regularly conditions, the law of  $\frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\hat{\sigma}_n}$  approaches  $N(0,1)$  as  $n \rightarrow \infty$ . To be precise:*

$$P\left(\frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\hat{\sigma}_n} \leq z\right) \rightarrow P(N(0,1) \leq z) \text{ as } n \rightarrow \infty$$

## Convergence in law

We say that random variable  $Z_n$  *converges in law* to random variable  $Z$ , or

$$Z_n \xrightarrow{\mathcal{L}} Z$$

if  $P(Z_n \leq z) \rightarrow P(Z \leq z)$  at all continuity points of the function  $P(Z \leq z)$ . For  $N(0,1)$ , all points are continuity points



# Approximate (“asymptotically accurate”) CIs

An asymptotically accurate confidence interval is thus obtained by

- Let  $z_\alpha$  be such that  $P(|Z| > z_\alpha) = 1 - \alpha$  ( $z_\alpha$  is the “ $\alpha/2$  upper quantile of  $N(0,1)$ ).

- Asymptotics based Confidence Interval:

$$\left| \frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\hat{\sigma}_n} \right| \leq z_\alpha, \text{ or } \beta \in \text{CI}_n = [\hat{\beta}_n - z_\alpha \hat{\sigma}_n / \sqrt{n}, \hat{\beta}_n + z_\alpha \hat{\sigma}_n / \sqrt{n}]$$

- Property of approximate CI:

Under the conditions of the CLT,

$$P(\beta \in \text{CI}_n) \rightarrow 1 - \alpha \text{ as } n \rightarrow \infty$$

- terminology:

- $1 - \alpha$ : *nominal level, nominal coverage probability*
- $P(\beta \in \text{CI}_n)$ : *actual coverage probability*

- Test for  $\beta = \beta_0$ : check if  $\beta_0$  is in CI

# Relationship to exact small sample normal theory



$$T_n \xrightarrow{\mathcal{L}} N(0, 1)$$

- Therefore:  $t_{n-1,\alpha} \rightarrow z_\alpha$  as  $n \rightarrow \infty$ , and so
- Actual coverage probability is asymptotically the same for  $t$ - and  $z$ - confidence intervals
- In practice, most people use  $t$ -intervals also for general (non-normal) regression data, since slightly more conservative ( $t_{n-1,\alpha} > z_\alpha$ )
- For non-regression data, people usually use normal intervals

# Some background on the CLT: The simple case of a sum

Let  $X_1, \dots, X_n$  be iid random variables. Set

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Then the following main results are true:

- Law of large numbers: if  $E|X| < \infty$ :  $\bar{X}_n \rightarrow E(X)$  (“almost surely”, “in probability”):  $\bar{X}_n$  is *consistent* for  $E(X)$
- Central limit theorem: if  $E(X^2) < \infty$ : set  $\sigma^2 = \text{Var}(X) = E(X - E(X))^2 = E(X^2) - (EX)^2$ ; we have

$$\sqrt{n} (\bar{X}_n - E(X)) \xrightarrow{\mathcal{L}} N(0, \sigma^2) = \sigma \times N(0, 1)$$

# Application to estimation of the mean

Let  $X_1, \dots, X_n$  be iid random variables for which  $E(X^2) < \infty$ . In particular,  $\beta = E(X)$  and  $\sigma^2 = \text{Var}(X)$  exist.

- Estimate  $\beta$  by  $\hat{\beta}_n = \bar{X}_n$
- Estimate  $\sigma^2$  by  

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1} (\sum_{i=1}^n X_i^2 - n(\bar{X}_n)^2)$$
- Note that by LLN:

$$\begin{aligned} \hat{\sigma}_n^2 &= \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X}_n)^2 \right) \\ &\rightarrow E(X^2) - (EX)^2 = \sigma^2 \text{ as } n \rightarrow \infty \end{aligned}$$

In other words,  $\hat{\sigma}_n^2$  is consistent for  $\sigma^2$

# Slutsky's Theorem

**Theorem** Let  $Z_n$ ,  $U_n$  and  $V_n$  be sequences of random variables, so that  $Z_n \xrightarrow{\mathcal{L}} Z$ ,  $U_n \rightarrow u$ ,  $V_n \rightarrow v$  as  $n \rightarrow \infty$ , where  $u$  and  $v$  are nonrandom. Then  $U_n Z_n + V_n \xrightarrow{\mathcal{L}} uZ + v$ .

Consequence for estimation of the mean:

- $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \sigma \times N(0, 1)$
- $\hat{\sigma}_n^2 \rightarrow \sigma^2$

Therefore:

$$\begin{aligned} \frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\hat{\sigma}_n} &= \frac{1}{\hat{\sigma}_n} \times \sqrt{n}(\hat{\beta}_n - \beta) \\ &\xrightarrow{\mathcal{L}} \frac{1}{\sigma} \times \sigma \times N(0, 1) \\ &= N(0, 1) \end{aligned}$$

# Computational Illustration of the CLT

This will use our computer package:

- Splus (proprietary software)
- R (freeware)
- generically: R

R will also be used in other courses

Help with R:

- Get a book about it (such as Krause and Olson, or Venables and Ripley)
- the help command in R

We now open R, and get...

# Open R

R version 3.3.2 (2016-10-31) - "Sincere Pumpkin Patch" Copyright (C) 2016 The R Foundation for Statistical Computing

R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type `'license()'` or `'licence()'` for distribution details. Natural language support but running in an English locale

R is a collaborative project with many contributors Type `'contributors()'` for more information and `'citation()'` on how to cite R or R packages in publications.

Type `'demo()'` for some demos, `'help()'` for on-line help, or `'help.start()'` for an HTML browser interface to help. Type `'q()'` to quit R.

>

# Example: Binomial Distribution

- $X_i$  is 0 or 1
- $P(X_i = 1) = p$ ,  $P(X_i = 0) = q = 1 - p$
- The distribution of  $S_n = X_1 + \dots + X_n$  is called the *binomial distribution* with parameters  $(n, p)$ , or simply  $b(n, p)$  (see Chapter 2.1.2 in Rice)
- $\beta = E(X_i) = 1 \times P(X_i = 1) = p$
- $\sigma^2 = \text{Var}(X_i) = E(X_i^2) - (EX_i)^2 = E(X_i) - (EX_i)^2 = p - p^2 = pq$
- The LLN and CLT hold for  $\hat{\beta}_n = \bar{X}_n = S_n/n$

Let's see in R if this is true...



# To find out about the binomial distribution

```
> help(rbinom)
```

```
[...]
```

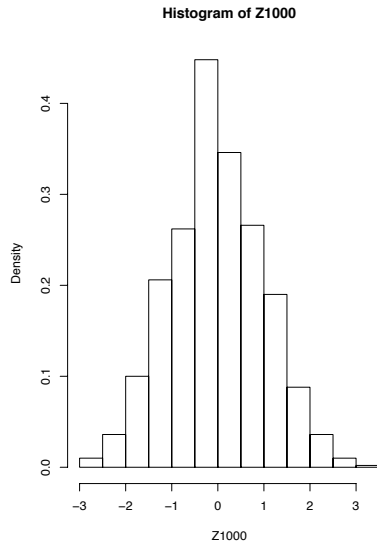
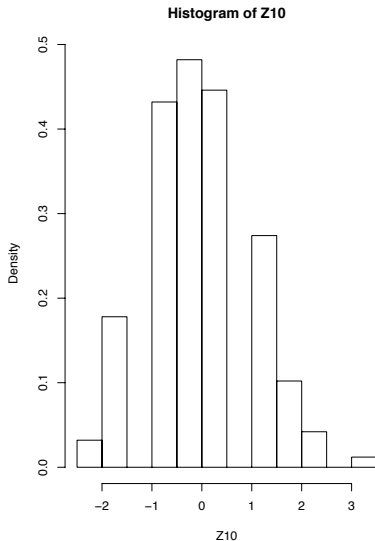
```
Usage:
```

- `dbinom(x, size, prob, log = FALSE)`
- `pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)`
- `qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)`
- `rbinom(n, size, prob)`

```
[...]
```

```
M<-1000 # number of simulations
n<-10
p<- .33
S<- rbinom(M,n,p) #"help" is wrong
Xbar<-S/n
sigma<- sqrt(p*(1-p))
Z10 <- sqrt(n)*(Xbar -p)/sigma
par(mfrow=c(1,2)) # check this command out!!!
hist(Z10,freq=F)
# try again with a larger n
n<-1000
S<- rbinom(M,n,p) #"help" is wrong
Xbar<-S/n
Z1000 <- sqrt(n)*(Xbar -p)/sigma
hist(Z1000,freq=F)
```

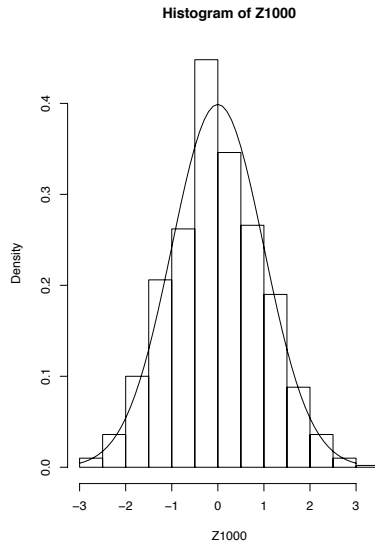
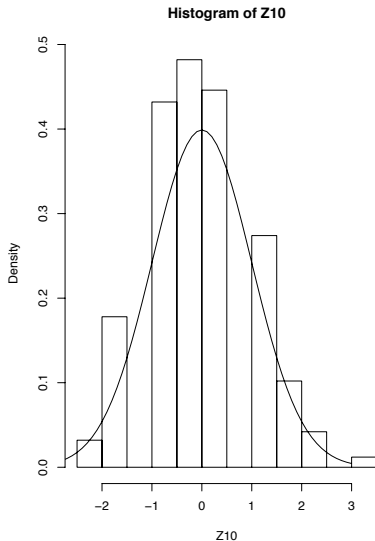
# The distribution of $Z_n$ stabilizes



# Superimposing the normal distribution on the histogram

```
par(mfrow=c(1,2))
hist(Z10,freq=F)
# compare to normal distribution
xpoints <- c(-30:30)/10
density <- dnorm(xpoints,mean=0,sd=1)
lines(xpoints,density)
# try again with larger n hist(Z1000,freq=F)
lines(xpoints,density)
```

# Normal curve superimposed on histograms



# More General Quantities than the Mean

- For many (but not all) well chosen estimators,  $\sqrt{n}(\hat{\beta}_n - \beta)$  has a CLT, i.e.,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} N(0, \sigma^2)$$

- (We shall discuss how to find estimators next time)
- If we have a consistent estimator  $\hat{\sigma}_n$  for  $\sigma$ , Slutsky's Theorem then yields that

$$\frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\hat{\sigma}_n} \xrightarrow{\mathcal{L}} N(0, 1)$$

- We can then use the normal distribution to set confidence intervals, tests

Similar results apply in the vector case

# How High Frequency Data differ from Low Frequency Data

## Volatility Estimation without Microstructure:

### Parametric and Non-parametric Approaches

# Models for Inference in High Frequency Data

- Natural to use **same model as in quantitative finance**: the Itô process:

$$\text{log securities price: } X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dB_s$$

$B_t$  is Brownian motion;  $\mu_t$  and  $\sigma_t$  can be random processes

- Model can also include **jumps**

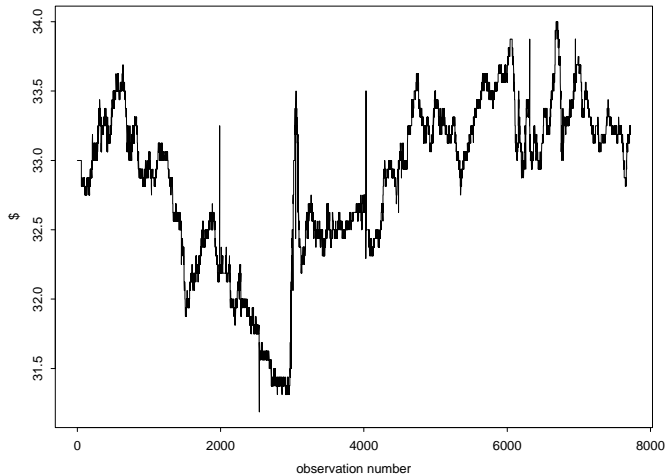
High frequency data formalism:

- Up to several transactions per second, sampling times  $0 = t_0 < t_1 < \dots < t_n = T$
- Typical time period of analysis  $[0, T]$ : one day (or 5 minutes)
- Can also combine results for several days
- Typical asymptotics as  $n \rightarrow \infty$ ,  $T$  fixed



# Model vs. Data: An Intra-day Time Series

Alcoa Aluminium (AA), first 4 days of 2001



# Brownian motion and Geometric Brownian motion

$X_t = \log S_t =$  the logarithm of the stock price  $S_t$  at time  $t$ .

The Geometric Brownian motion (GBM) model is now that

$$X_t = X_0 + \mu t + \sigma W_t,$$

where  $\mu$  and  $\sigma$  are constants,  $W_t$  is a *Brownian Motion (BM)*.

The process  $(W_t)_{0 \leq t \leq T}$  is a Brownian motion provided

- (1)  $W_0 = 0$  ("time zero" is an arbitrary reference time);
- (2)  $t \rightarrow W_t$  is a continuous function of  $t$ ;
- (3)  $W$  has independent increments: if  $t > s > u > v$ , then  $W_t - W_s$  is independent of  $W_u - W_v$ ;
- (4) for  $t > s$ ,  $W_t - W_s$  is normal with mean zero and variance  $t - s$  ( $N(0, t-s)$ ).

# Estimation in the GBM model: Parametric inference

- Equal spacing:  $t_{n,i} = i\Delta t_n = iT/n$
- Observations:  $X_{t_{n,i}}$ , or  
 $\Delta X_{t_{n,i+1}} = X_{t_{n,i+1}} - X_{t_{n,i}}, i = 0, \dots, n-1$
- The  $\Delta X_{t_{n,i+1}}$  are iid with law  $N(\mu\Delta t_n, \sigma^2\Delta t_n)$ .
- Natural estimators are:

$$\hat{\mu}_n = \frac{1}{n\Delta t_n} \sum_{i=0}^{n-1} \Delta X_{t_{n,i+1}} = (X_T - X_0)/T \text{ MLE and UMVU}$$

$$\hat{\sigma}_{n,MLE}^2 = \frac{1}{n\Delta t_n} \sum_{i=0}^{n-1} (\Delta X_{t_{n,i+1}} - \overline{\Delta X}_{t_n})^2 \text{ MLE; or}$$

$$\hat{\sigma}_{n,UMVU}^2 = \frac{1}{(n-1)\Delta t_n} \sum_{i=0}^{n-1} (\Delta X_{t_{n,i+1}} - \overline{\Delta X}_{t_n})^2 \text{ UMVU.}$$

# Behavior of parametric estimators: $\hat{\mu}$

- $\mu$  *cannot be consistently estimated* for fixed  $T$
- $\hat{\mu}_n$  does not depend on  $n$ , but only on  $T, X_0, X_T$
- If  $T \rightarrow \infty$ , then  $\mu$  *can* be estimated consistently:  
 $(X_T - X_0)/T \xrightarrow{p} \mu$  as  $T \rightarrow \infty$ . This is because  
 $\text{Var}((X_T - X_0)/T) = \sigma^2/T \rightarrow 0$ .

# Behavior of parametric estimators: Consistency of $\hat{\sigma}$

- $\sigma^2$  can be estimated consistently for fixed  $T$ , as  $n \rightarrow \infty$ :  
 $\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2$  as  $n \rightarrow \infty$ .
- Set  $U_{n,i} = \Delta X_{t_n,i} / (\sigma \Delta t_n^{1/2})$ , then: Then the  $U_{n,i}$  are iid with distribution  $N((\mu/\sigma)\Delta t_n^{1/2}, 1)$ . Set  $\bar{U}_{n,\cdot} = n^{-1} \sum_{i=0}^{n-1} U_{n,i}$ .
- From considerations for normal random variables:

$$\sum_{i=0}^{n-1} (U_{n,i} - \bar{U}_{n,\cdot})^2$$

is  $\chi^2$  distributed with  $n - 1$  df (and independent of  $\bar{U}_{n,\cdot}$ )

- For the UMVU estimator,

$$\hat{\sigma}_n^2 = \sigma^2 \Delta t_n \frac{1}{(n-1)\Delta t_n} \sum_{i=0}^{n-1} (U_{n,i} - \bar{U}_{n,\cdot})^2 \stackrel{\mathcal{L}}{=} \sigma^2 \frac{\chi_{n-1}^2}{n-1}$$

# Behavior of parametric estimators: Consistency of $\hat{\sigma}$ (cont'd)

- $\hat{\sigma}_n^2 = \sigma^2 \frac{\chi_{n-1}^2}{n-1}$
- It follows that

$$E(\hat{\sigma}_n^2) = \sigma^2 \text{ and } \text{Var}(\hat{\sigma}_n^2) = \frac{2\sigma^4}{n-1},$$

since  $E\chi_m^2 = m$  and  $\text{Var}(\chi_m^2) = 2m$ .

- Hence  $\hat{\sigma}_n^2$  is consistent for  $\sigma^2$ :  $\hat{\sigma}_n^2 \rightarrow \sigma^2$  in probability as  $n \rightarrow \infty$ .

# Behavior of parametric estimators: Asymptotic normality of $\hat{\sigma}$

- $\hat{\sigma}_n^2 = \sigma^2 \frac{\chi_{n-1}^2}{n-1}$
- Since  $\chi_{n-1}^2$  is the sum of  $n-1$  iid  $\chi_1^2$  random variables, by the central limit theorem we have the following convergence in law:

$$\frac{\chi_{n-1}^2 - E\chi_{n-1}^2}{\sqrt{\text{Var}(\chi_{n-1}^2)}} = \frac{\chi_{n-1}^2 - (n-1)}{\sqrt{2(n-1)}} \xrightarrow{\mathcal{L}} N(0, 1),$$

- and so

$$n^{1/2}(\hat{\sigma}_n^2 - \sigma^2) \sim (n-1)^{1/2}(\hat{\sigma}_n^2 - \sigma^2)$$

$$\stackrel{\mathcal{L}}{=} \sqrt{2}\sigma^2 \frac{\chi_{n-1}^2 - (n-1)}{\sqrt{2(n-1)}} \xrightarrow{\mathcal{L}} \sigma^2 N(0, 2) = N(0, 2\sigma^4).$$

# Confidence intervals for $\sigma$

- $n^{1/2}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{\mathcal{L}} \sigma^2 N(0, 2) = N(0, 2\sigma^4)$
- Intervals:  $\sigma^2 = \hat{\sigma}_n^2 \pm 1.96 \times \frac{\sqrt{2}\hat{\sigma}_n^2}{\sqrt{n}}$  is asymptotic 95 % confidence interval for  $\sigma^2$ .
- Since  $\hat{\sigma}_{n,MLE}^2 = \frac{n-1}{n} \hat{\sigma}_{n,UMVU}^2$ , the same asymptotics applies to the MLE



# Non-Centered Estimators

$$\hat{\sigma}_{n,nocenter}^2 = \frac{1}{n\Delta t_n} \sum_{i=0}^{n-1} (\Delta X_{t_n,i+1})^2.$$

For MLE version of  $\hat{\sigma}_n$ :

$$\begin{aligned} \hat{\sigma}_{n,MLE}^2 &= \frac{1}{n\Delta t_n} \sum_{i=0}^{n-1} (\Delta X_{t_n,i+1} - \overline{\Delta X}_{t_n})^2 \\ &= \frac{1}{n\Delta t_n} \left( \sum_{i=0}^{n-1} (\Delta X_{t_n,i+1})^2 - n(\overline{\Delta X}_{t_n})^2 \right) \\ &= \hat{\sigma}_{n,nocenter}^2 - \Delta t_n \hat{\mu}_n^2 \\ &= \hat{\sigma}_{n,nocenter}^2 - \frac{T}{n} \hat{\mu}_n^2 \end{aligned}$$

$$\text{Hence: } n^{1/2} \left( \hat{\sigma}_{n,MLE}^2 - \hat{\sigma}_{n,nocenter}^2 \right) \xrightarrow{p} 0.$$

# The Classical Nonparametric Case: Integrated Volatility (IV)

- Classical target: Integrated volatility:

$$\langle X, X \rangle = \int_0^T \sigma_t^2 dt = \lim_{\Delta t \rightarrow \infty} \sum_{t_{i+1} \leq T} (X_{t_{i+1}} - X_{t_i})^2$$

- Purpose of Estimating IV

- Asset management, portfolio optimization
- Options hedging
- Risk management
- Model dynamics
- Prediction interval based hedging of options

# The Classical case: Realized Volatility as Measure of Integrated Volatility

High frequency data: up to several transactions per second

Chance to estimate  $\langle X, X \rangle_T$  very accurately

Usual estimator:  $[X, X]_T = \sum_{t_{i+1} \leq T} (X_{t_{i+1}} - X_{t_i})^2$  “realized volatility”

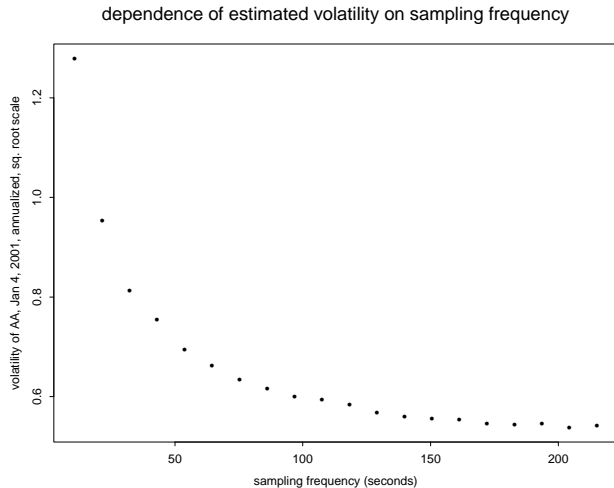
- a standard way to measure volatility (Andersen, Bollerslev, many others)
- consistent:  $[X, X]_T \xrightarrow{P} \langle X, X \rangle_T$  as  $\Delta t \rightarrow 0$  (stoch calc)
- asymptotically mixed normal, variance  $\frac{2T}{n} \int_0^T \sigma_t^4 dt$  (Barndorff-Nielsen & Shephard, Jacod & Protter, Mykland & Zhang)
- can estimate variance by  $\frac{2}{3} [X, X, X, X]_T$  (Barndorff-Nielsen & Shephard)

# Other Quantities that can be Estimated in Data from One Day

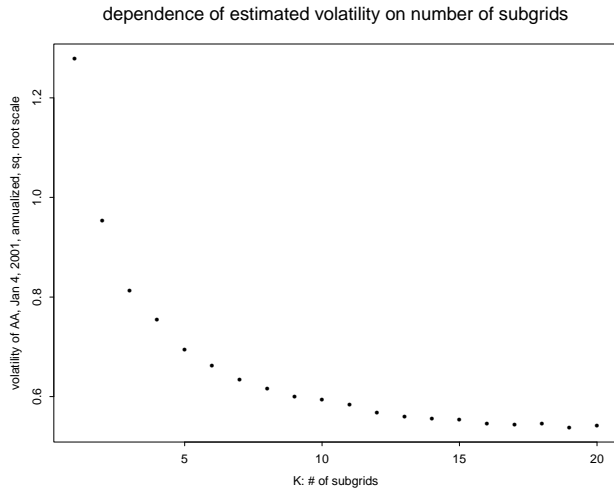
- Other powers of volatility:  $\int_0^T \sigma_t^p dt$
- Leverage effect:  $\langle \sigma^2, X \rangle_T$ , or corresponding correlation, relates to **skewness** and **volatility risk**
- Volatility of volatility  $\langle \sigma^2, \sigma^2 \rangle_T$ , related to kurtosis
- Regression of one process on another, integrated alphas and betas, ANOVA, related to **systematic risk**, **options hedging**, and **model testing**
- Same quantities, but instantaneously
- Nonparametric trading strategies
- Liquidity; time to execution; dark pools

## Inference in the presence of market microstructure

# RV as One Samples More Frequently



# RV vs Sampling Interval



# The Failure of Realized Volatility

**The realized volatility methodology does not work so well in ultra high frequency data**

- In real data, when  $\Delta t \rightarrow 0$ ,  $[X, X]_T$  does not converge
- Theory usually illustrated with  $\Delta t = 5\text{-}15$  minutes

Why? Our candidate explanation:

- existence of microstructure (e.g. bid-ask spread, strategic trading, limited market depth, price impact, discrete price changes, ...)
- transaction as measurement device

**The realized volatility methods form part of a more general theory which can also handle noise**



# Microstructure Noise & Hidden Semimartingale model

- observed log stock price:  $Y_{t_i} = X_{t_i} + \epsilon_i$
- $X_t$  is latent log price, semimartingale, say, Ito process

$$dX_t = \mu_t dt + \sigma_t dB_t$$

$B_t$  is Brownian motion;  $\mu_t$  and  $\sigma_t$  can be random processes

- Model in  $X$  can also include jumps
- Microstructure  $\epsilon_i$  is iid, or fast mixing

## Challenges

- Dependent noise  $\epsilon_i$  can have short run dependence
- Irregular spacings
- Endogenous times
- Epps effect
- Time scrambling
- Relationship to trading

