第三章
金融数据及其特征

上节内容回顾

- 如何安装R软件和Rstudio软件
- R与Rstudio界面介绍
- R的基本操作
- 如何安装和使用R工具包(package)

■ 本章内容

1、资产收益率

2、金融数据的例子

3、收益率的分布特征

4、金融数据的特征

资产收益率的定义

- 假设P_t是t时刻的资产价格，下面给出一些资产收益率的定义，我们暂时假设资产不支付红利。
- **单期简单收益率(One-Period Simple Return)**
  - 单期毛收益率：$$1+R_t=\frac{P_t}{P_{t-1}} \quad \text{or} \quad P_t=P_{t-1}(1+R_t).$$
  - 单期净收益率：$$R_t=\frac{P_t}{P_{t-1}}-1=\frac{P_t-P_{t-1}}{P_{t-1}}$$
- **多期简单收益率(Multi-Period Simple Return)**
  - 多期毛收益率：$$\begin{aligned}1+R_t[k]&=\frac{P_t}{P_{t-k}}=\frac{P_t}{P_{t-1}}\times\frac{P_{t-1}}{P_{t-2}}\times\cdots\times\frac{P_{t-k+1}}{P_{t-k}}\\&=(1+R_t)(1+R_{t-1})\cdots(1+R_{t-k+1})\\&=\prod_{j=0}^{k-1}(1+R_{t-j}).\end{aligned}$$
  - 多期净收益率：$R_t[k]=(P_t-P_{t-k})/P_{t-k}.$
- 一般而言，所提到的收益率均指净收益率

例子

- 下表给出了苹果公司股票的每日收盘价。

TABLE 1.1. Daily Closing Prices of Apple Stock from December 2 to 9, 2011

Date	12/02	12/05	12/06	12/07	12/08	12/09
Price($)	389.70	393.01	390.95	389.09	390.66	393.62

- 从12月8日到12月9日的1天毛收益率为$1 + R_t = 393.62/390.66 = 1.0076$。这一天的净收益率为$393.62/390.66-1 = 0.76\%$。

- 从12月2日到12月9日刚好一周，这一周的毛收益率为$1 + R_t [5]= 393.62/389.70 = 1.0101$，这一周的净收益率为$393.62/389.70-1 = 1.01\%$

年化收益率

- **资产收益率**和时间区间有关，例如1天收益率，1周收益率，1个月收益率等等。那么不同收益率如何做比较呢？
- 若时间区间没有给出，则惯例为时间区间为一年。如果持有资产的期限为K年，则**年化收益率**的定义为：

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

- 实际上，这是由K个单期毛收益率几何平均得到的，也可以表示成：

$$\text{Annualized}\{R_t[k]\} = \exp\left[\frac{1}{k}\sum_{j=0}^{k-1}\ln(1+R_{t-j})\right] - 1,$$

- 考虑到单期收益率一般很小，所以我们可以用一阶泰勒展开来近似表示**年化收益率**，即

$$\text{Annualized}\{R_t[k]\} \approx \frac{1}{k}\sum_{j=0}^{k-1}R_{t-j}.$$

连续复合收益率

- **连续复合收益率**又称为对数收益率(log-return)，是资产的简单毛收益率的自然对数。即：

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

其中$p_t = \log(P_t)$。

- 对于多期连续复合收益率，我们有

$$\begin{aligned} r_t[k] &= \ln(1+R_t[k]) = \ln[(1+R_t)(1+R_{t-1})\cdots(1+R_{t-k+1})] \\ &= \ln(1+R_t) + \ln(1+R_{t-1}) + \cdots + \ln(1+R_{t-k+1}) \\ &= r_t + r_{t-1} + \cdots + r_{t-k+1}. \end{aligned}$$

- 相比于简单收益率R_t，连续复合收益率的**优势**：
  - 多期连续复合收益率是单期连续复合收益率之和；
  - 连续复合收益率具有更容易处理的统计特性。

例子

- 延续之前苹果公司股票的每日收盘价的例子

TABLE 1.1. Daily Closing Prices of Apple Stock from December 2 to 9, 2011

Date	12/02	12/05	12/06	12/07	12/08	12/09
Price($)	389.70	393.01	390.95	389.09	390.66	393.62

- 12月8日-12月9日的日对数收益率为$r_t = \ln(393.62) - \ln(390.66) = 7.5\%$.
- 12月2日-12月9日的周对数收益率为: $r_t(5) = \ln(393.62) - \ln(389.7) = 1\%$.
- 12月5日、12月6日、12月8日、12月8日、12月9日的日对数收益率为

12月5日	12月6日	12月7日	12月8日	12月9日	求和
0.85%	-0.53%	-0.48%	0.40%	0.75%	1.00%

- 显然，周对数收益率是该周5个日对数收益率之和。

关系小结

- 简单收益率R_t与连续复合收益率r_t的关系为

$$r_t = \ln(1+R_t), \quad R_t = e^{r_t} - 1.$$

- 如果收益率R_t与r_t均用百分比表示，则

$$r_t = 100\ln\left(1+\frac{R_t}{100}\right), \quad R_t = 100\left(e^{r_t/100} - 1\right).$$

- 如果把单期收益率进行时间累加，则有

$$1 + R_t[k] = (1+R_t)(1+R_{t-1})\cdots(1+R_{t-k+1}),$$
$$r_t[k] = r_t + r_{t-1} + \cdots + r_{t-k+1}.$$

- 如果连续复合年利率为r，则资产现值与资产的未来价值之间的关系为：

$$A = C\exp(r\times n), \quad C = A\exp(-r\times n).$$

■ 本章内容

1、资产收益率

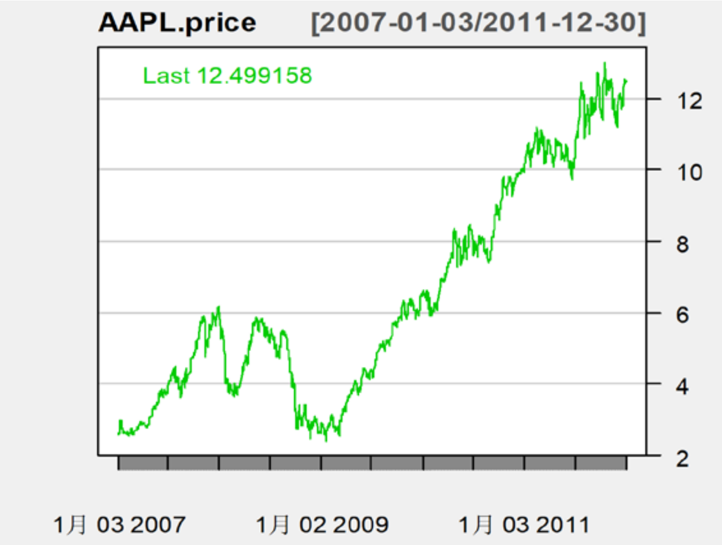
2、金融数据的例子

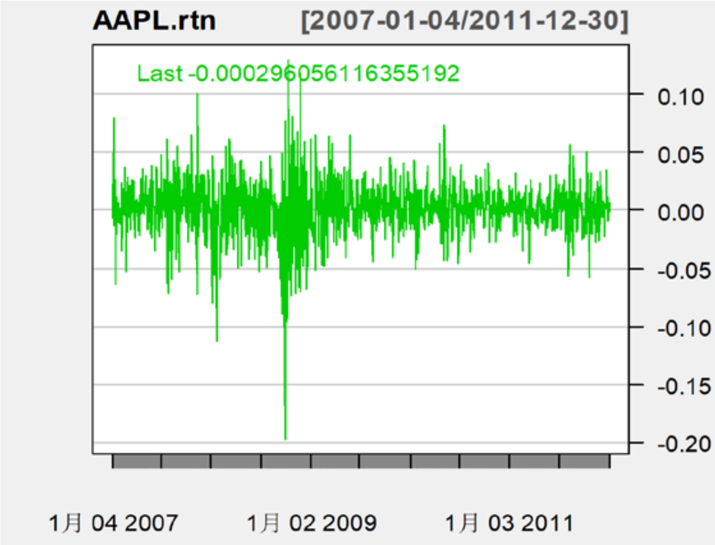
3、收益率的分布特征

4、金融数据的特征

金融数据的例子

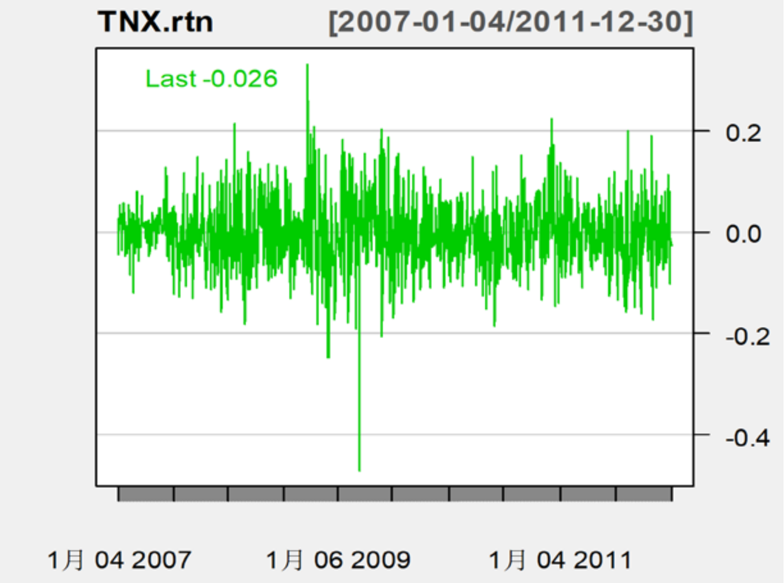
- **例1：**下图给出了2007年1月4日至2011年12月31日苹果公司（AAPL）股票的价格走势图和日对数收益率的时序图。从下图可以看到（1）收益率存在**厚尾现象**，例如在2008年金融危机期间；（2）收益率有**波动集聚**（volatility clustering）,即收益率在某些时期波动率很大，而在其他时候相对稳定。





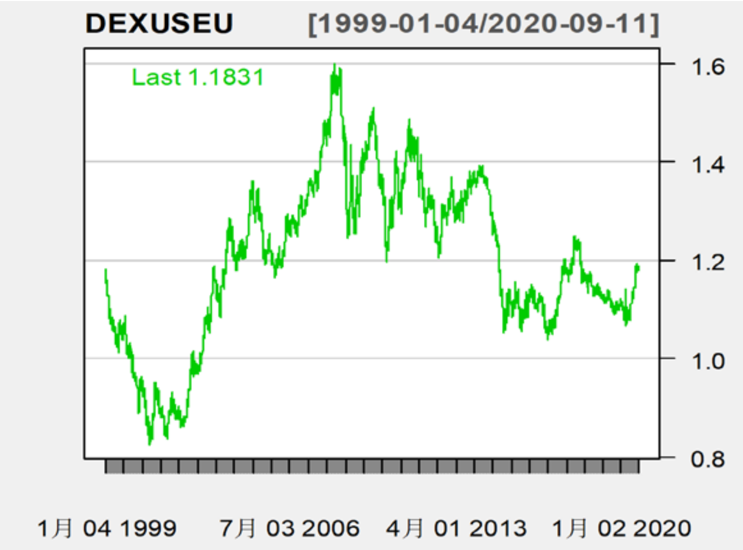
金融数据的例子

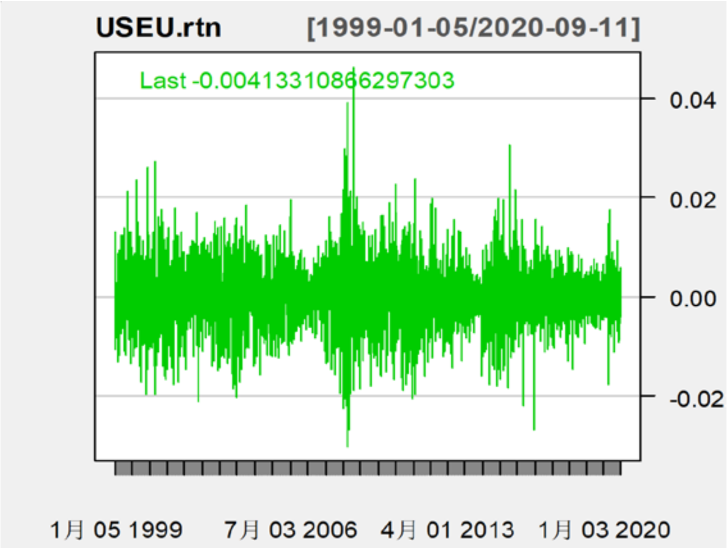
- **例2：** 下图给出了2007年1月4日至2011年12月31日10年期国债到期收益率和日变化的。时序图。从下图可以看到10年期国债到期收益率显示出和苹果公司股票日收益率类似的特征。



金融数据的例子

- **例3：**下图给出的是欧元/美元的汇率走势图和汇率的日对数收益率的走势图。从中也可以看到（1）**厚尾现象**；（2）**波动集聚**（volatility clustering）。
- **厚尾现象和波动集聚在金融数据中具有普遍性。**





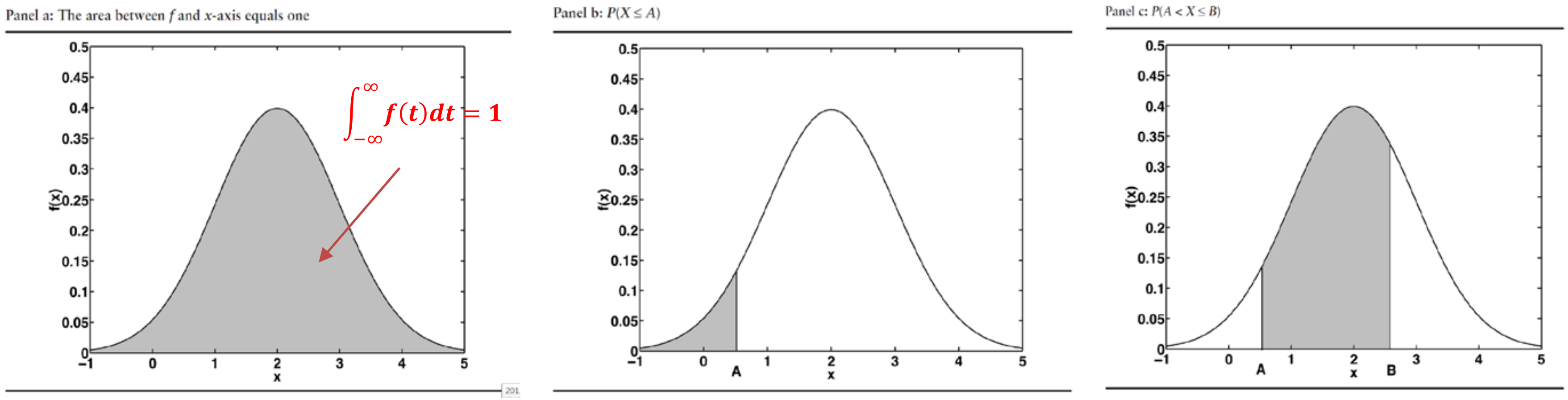
■ 本章内容

分布函数回顾

- **分布函数的定义：**对于随机变量X，其分布函数可以定义为

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

其中$f(x)$为密度函数，$F(x)$为分布函数。



分布函数的矩

- **矩的定义：**一个连续型随机变量X的l阶矩定义为

$$m'_\ell = E(X^\ell) = \int_{-\infty}^{\infty} x^\ell f(x)\mathrm{d}x,$$

其中E表示期望（expectation）。一阶矩称为X的均值（mean）或者**期望（expectation）**。均值被用来度量分布的中心位置，通常记为μ_x。

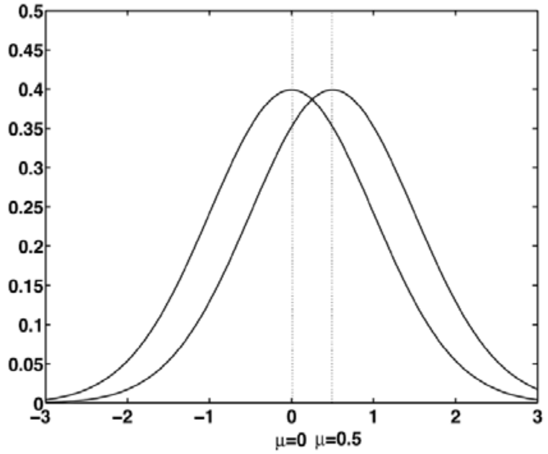
- 第l阶中心矩的定义为

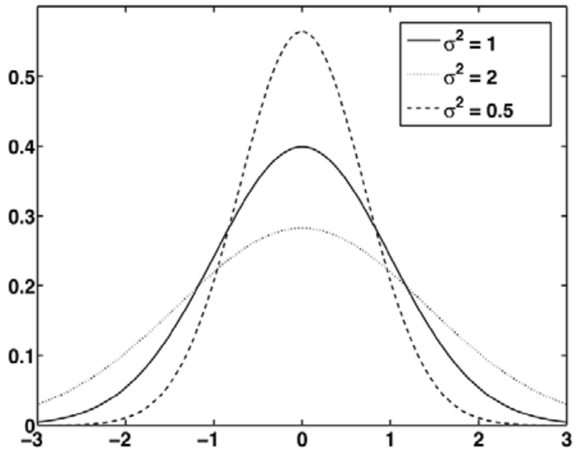
$$m_\ell = E[(X-\mu_x)^\ell] = \int_{-\infty}^{\infty} (x-\mu_x)^\ell f(x)\mathrm{d}x$$

- 二阶中心矩可以度量X偏离其均值的程度，称为X的**方差（variance）**，通常记为σ_x^2，方差的平方根σ_x，被称为X的标准差（standard deviation）.
- 对于资产收益率，方差（标准差）是衡量不确定性的指标，因此常被用于资产的风险管理和度量中。
- 正态分布可以由它的前两阶矩确定，而对于其他分布，可能需要更高阶的矩。

正态分布

- **下图左**：方差为1，均值分别为0和0.5的正态分布密度函数比较
- **下图右**：均值为0，方差分别为1,2,0.5的正态分布密度函数比较





峰度和偏度

- 标准化的三阶中心矩称为**偏度**（skewness），标准化的四阶中心矩称为**峰度**（kurtosis），**偏度和峰度**分别用来描述随机变量的非对称程度和尾部厚度。
- 随机变量X的**偏度和峰度**分别定义为：

$$S(x) = E\left[\frac{(X-\mu_x)^3}{\sigma_x^3}\right], \quad K(x) = E\left[\frac{(X-\mu_x)^4}{\sigma_x^4}\right].$$

- **超额峰度**：$K(x) - 3$，正态分布的峰度为3，所以正态分布的超额峰度为0。若一个分布具有正的超额峰度，则称为具有厚尾性，这样的分布被称为**尖峰的**（leptokurtic）；反之，若一个分布具有负的超额峰度，则称为薄尾分布。
- 金融资产收益率序列是往往具有尖峰厚尾特性的，详见后面的例子。

样本矩

- 在实际应用中，通常能获得金融资产收益率的一系列观察值，我们面临的实际问题是：
  - 金融资产收益率的分布是什么？
  - 金融资产收益率的分布特征？例如，均值、标准差（波动率）、偏度、峰度、**尾指数**等
- 庆幸的是，得益于统计理论的发展，我们可以通过这些观察值去**估计**均值、方差、偏度和峰度。
- 样本均值和样本方差可以表示成：

$$\hat{\mu}_x = \frac{1}{T}\sum_{t=1}^{T} x_t, \qquad \hat{\sigma}_x^2 = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^2,$$

- 样本偏度和样本峰度可以表示成：

$$\hat{S}(x) = \frac{1}{(T-1)\hat{\sigma}_x^3}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^3, \qquad \hat{K}(x) = \frac{1}{(T-1)\hat{\sigma}_x^4}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^4.$$

正态性检验

- 在金融数据分析中，我们通常需要检验数据的正态性，下面是几个常用的正态性检验。
- **偏度检验**：正态分布的偏度$S(x)=0$，偏度检验的备择假设为$S(x)\neq 0$。在正态假设的情况下，我们有

$$S^*=\frac{\hat{S}(x)}{\sqrt{6/T}}\sim N(0,1)$$

  判别准则：若$|S^*|>Z_{1-\alpha/2}$或者p-值小于α则视为拒绝原假设。
- **峰度检验**：正态分布的峰度$K(x)=3$，偏度检验的备择假设为$K(x)-3\neq 0$。在正态假设的情况下，我们有

$$K^*=\frac{\hat{K}(x)-3}{\sqrt{24/T}}\sim N(0,1)$$

  判别准则：若$|K^*|>Z_{1-\alpha/2}$或者p-值小于α则视为拒绝原假设。

正态性检验（续）

- 偏度与峰度的联合检验：**JB检验**。Jarque and Bera (1987)提出了如下统计量

$$\mathrm{JB} = \frac{\hat{S}^2(r)}{6/T} + \frac{(\hat{K}(r) - 3)^2}{24/T},$$

- JB统计量渐近服从自由度为2的卡方分布$\chi^2(2)$。判别准则：若$|JB| > \chi^2_\alpha(2)$或者p-值小于α则视为拒绝原假设。
- 正态性检验案例（R语言实现）

```
> library(fBasics) # Load package
> da=read.table("d-mmm-0111.txt",header=T) # Load data
> # header=T means the first row of data file contains names.
> # deafult is no names.
> head(da) #Show the first 6 rows of data
> mmm=da[,2]  # Obtain 3m simple returns
> basicStats(mmm) #Compute summary statistics
> mean(mmm)  # mean return
> var(mmm)   # variance
> stdev(mmm) # standard deviation
> t.test(mmm)  # Testing mean return = 0
> s3=skewness(mmm)
> T=length(mmm) # Sample size
> t3=s3/sqrt(6/T) # Skewness test
> pp=2*(1-pnorm(t3)) # Compute p-value
> pp
> s4=kurtosis(mmm)
> t4=s4/sqrt(24/T) # Kurtosis test
> normalTest(mmm,method='jb') # JB-test
```

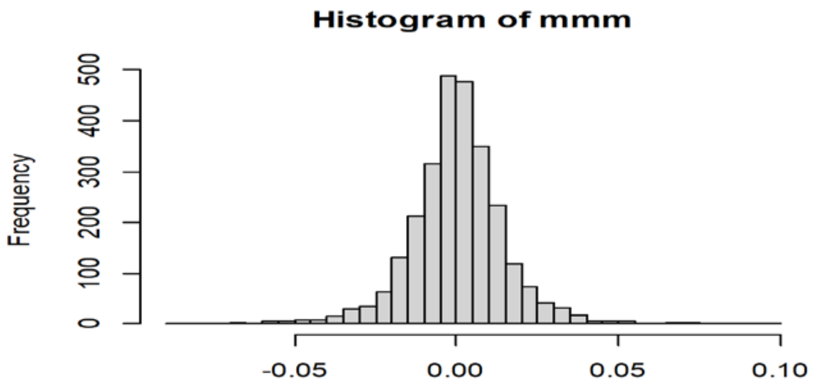

head(da) #Show the first 6

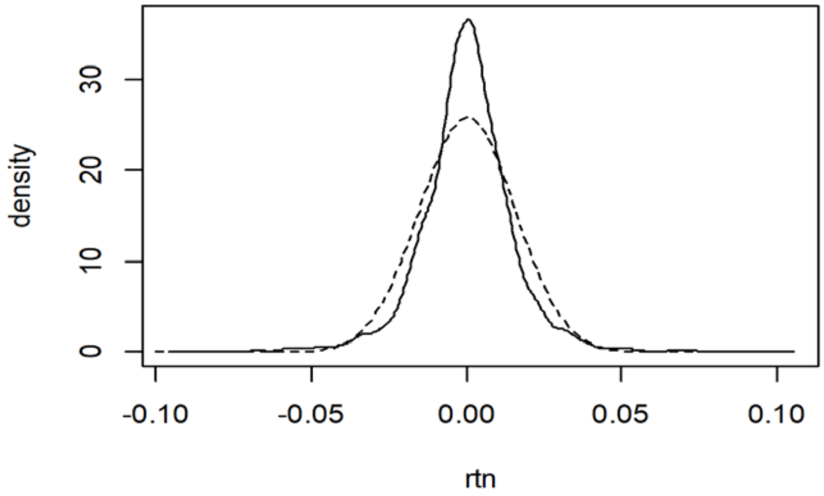
■ 本章内容

金融数据初探

- **例1：** mmm公司股票价格收益率

```
> library(fBasics)
> da=read.table("d-mmm-0111.txt",header=T)    # Load data
> mmm=da[,2] # Locate 3M simple returns
> hist(mmm,nclass=30) # Histogram
> d1=density(mmm)  # Obtain density estimate
> range(mmm)  # Range of 3M returns
> x=seq(-.1,.1,.001) # Create a sequence of x with increment 0.001.
> y1=dnorm(x,mean(mmm),stdev(mmm))
> plot(d1$x,d1$y,xlab='rtn',ylab='density',type='l')
> lines(x,y1,lty=2)
```

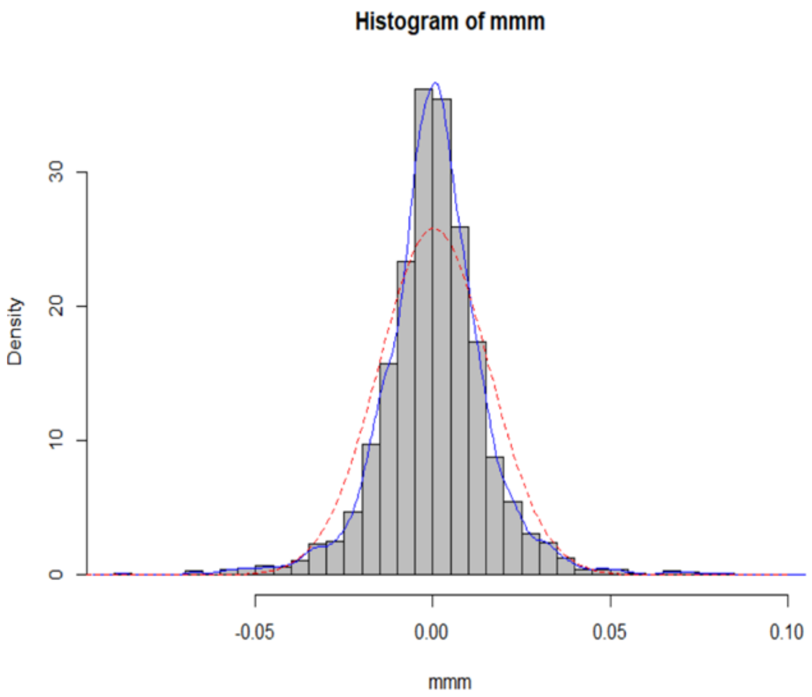




金融数据初探

- **例1：** mmm公司股票价格收益率
- **为了便于比较，将两个图并成一个图**

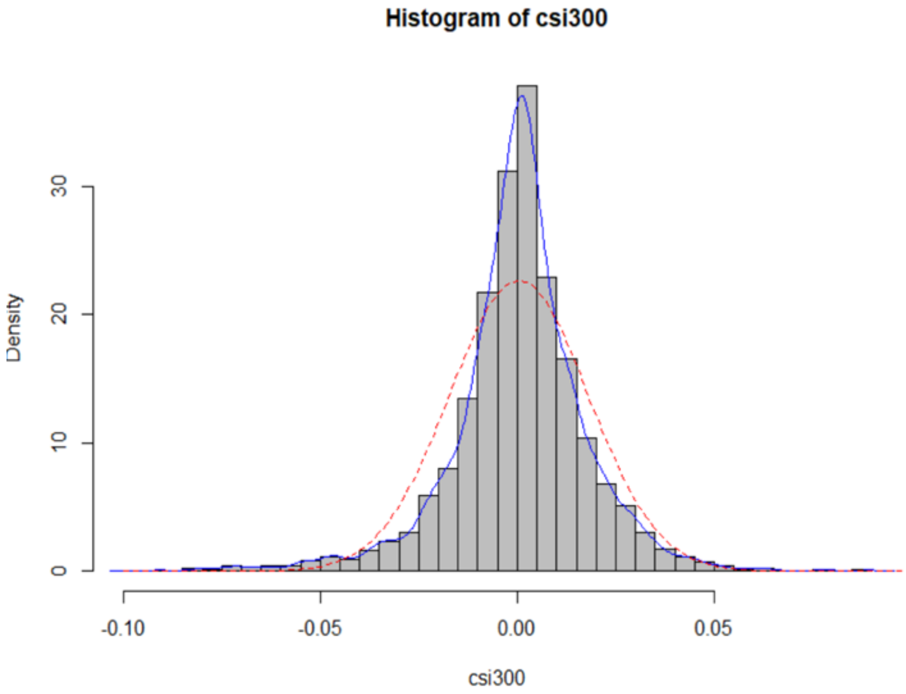
```
> library(fBasics)
> da=read.table("d-mmm-0111.txt",header=T)  # Load data
> mmm=da[,2] # Locate 3M simple returns
> hist(mmm,nclass=30,freq=F,col="grey")  # Histogram
> d1=density(mmm)  # Obtain density estimate
> range(mmm)  # Range of 3M returns
> x=seq(-.1,.1,.001)  # Create a sequence of x with increment 0.001.
> y1=dnorm(x,mean(mmm),stdev(mmm))
> lines(d1$x,d1$y,lty=1,col="blue")
> lines(x,y1,lty=2,col="red")
```



金融数据初探

- **例2：** 沪深300指数日收益率

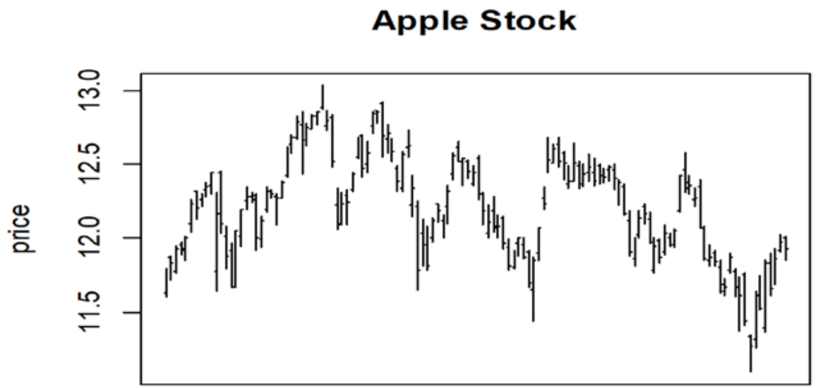
```
>   ## example, csi 300 index return
>   code = "000300.SH"
>   da = read.csv(paste(code,"_ret.csv",sep=""),
    sep=',',header=TRUE)
>   head(da)  #See the first 6 rows of the data
>   csi300=da[,3]  # log return series
>   d2=density(csi300)  # Obtain density estimate
>   range(csi300)  # Range of csi 300 returns
>   x=seq(-.1,.1,.001) # Create a sequence of x with increment
    0.001.
>   y2=dnorm(x,mean(csi300),stdev(csi300))
>   # get the plot
>   hist(csi300,nclass=30,freq=F,col="grey") # Histogram
>   lines(d2$x,d2$y,lty=1,col="blue")
>   lines(x,y2,lty=2,col="red")
```

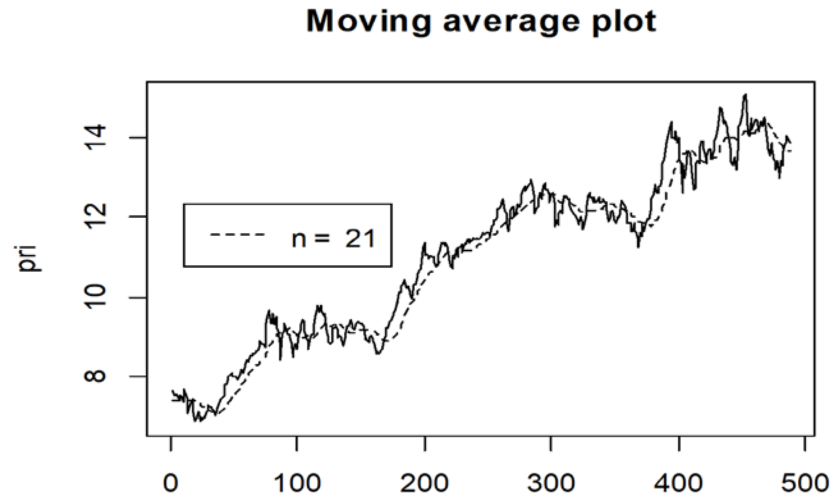


金融数据初探

- **例3：**股票价格条形图和移动平均线图

```
> library(quantmod)
> getSymbols("AAPL",from="2011-01-03",to="2011-06-30")
> X=AAPL[,1:4] # Locate open, high, low, and close prices
> xx=cbind(as.numeric(X[,1]),as.numeric(X[,2]),as.numeric(X[,3]),as.numeric(X[,4]))
> source("ohlc.R") # Compile the R script
> ohlc_plot(xx,xl="days",yl="price",title="Apple Stock")
> source("ma.R")  # Compile R script
> getSymbols("AAPL",from="2010-01-02",to="2011-12-08")
> x1=as.numeric(AAPL$AAPL.Close) # Locate close price
> ma(x1,21)
```

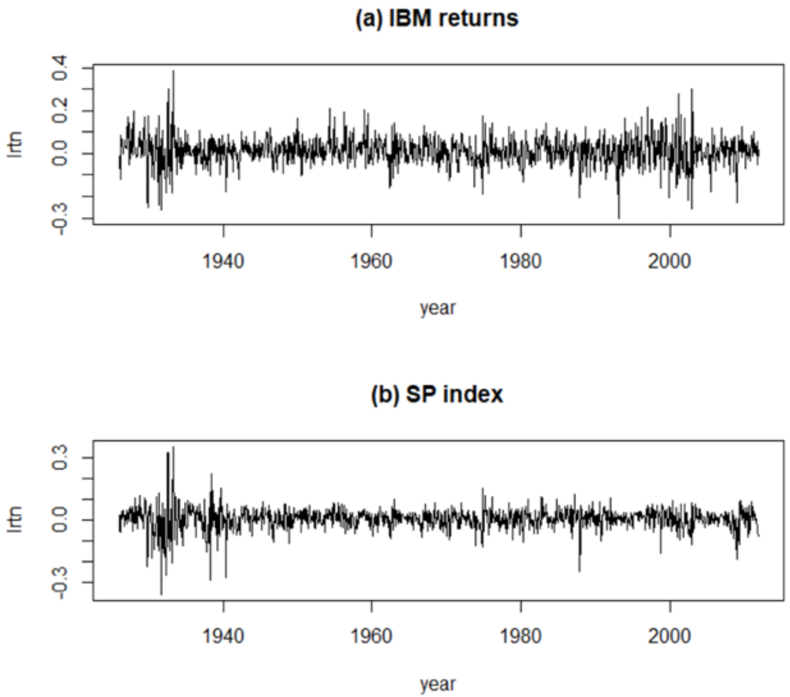




金融数据初探

- **例4：** 两个资产收益率的比较

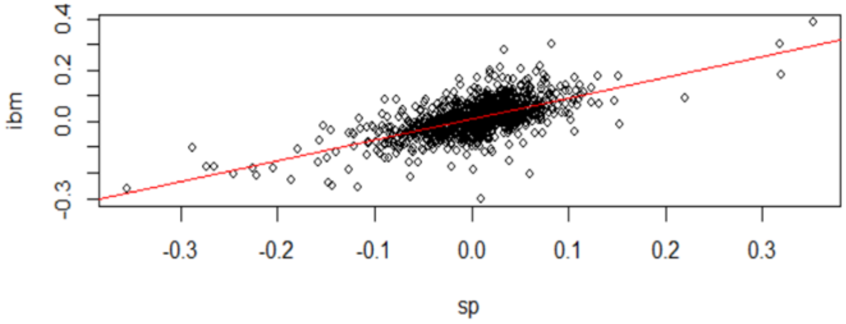
```
> da=read.table("m-ibmsp-2611.txt",header=T)
> head(da)
> ibm=log(da$ibm+1) # Transform to log returns
> sp=log(da$sp+1)
> tdx=c(1:nrow(da))/12+1926 # Create time index
> par(mfcol=c(2,1))
> plot(tdx,ibm,xlab='year',ylab='lrtn',type='l')
> title(main='(a) IBM returns')
> plot(tdx,sp,xlab='year',ylab='lrtn',type='l') # X-axis first.
> title(main='(b) SP index’)
> cor(ibm,sp)  # Obtain sample correlation
> m1=lm(ibm~sp)  # Fit the Market Model (linear model)
> summary(m1)
> plot(sp,ibm,cex=0.8)  # Obtain scatter plot
> abline(0.008,.807,col="red")  # Add the linear regression line
```

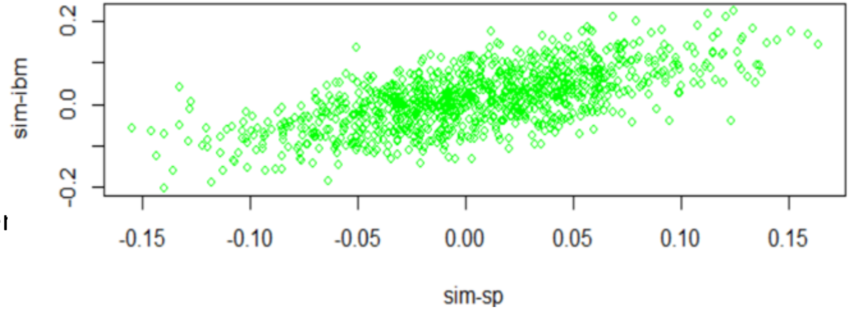


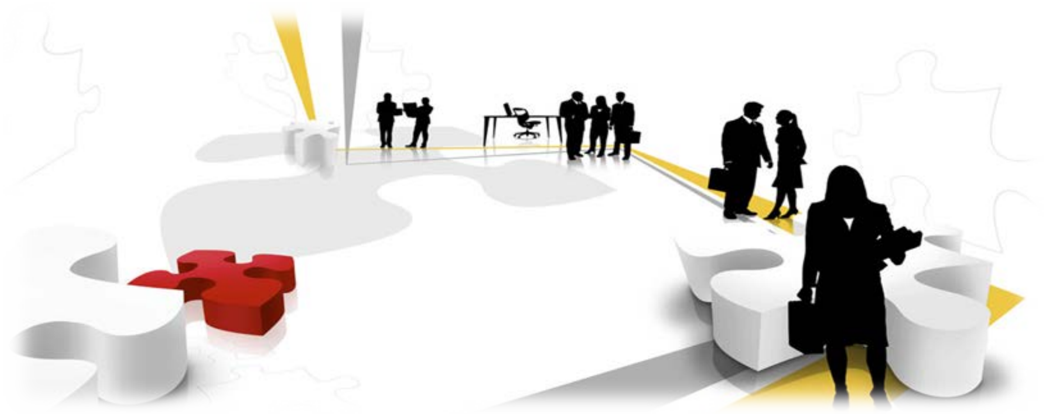
金融数据初探

- **例4：**两个资产收益率的比较（续）

```
> da=read.table("m-ibmsp-2611.txt",header=T) #Load  data
> dim(da)
> ibm=log(da$ibm+1) # Compute log returns
> sp=log(da$sp+1)
> rt=cbind(ibm,sp) # Obtain bivariate returns
> m1=apply(rt,2,mean) # Obtain sample means
> v1=cov(rt) # Obtain sample covariance matrix
> corr = cor(rt) # obtain sample correlation matrix
> m1
> v1
> corr
> library(mnormt) # Load package
> x=rmnorm(1029,mean=m1,varcov=v1) # Simulation
> dim(x)
> plot(x[,2],x[,1],xlab='sim-sp',ylab='sim-ibm',cex=0.8,col="greer
```







感谢大家的聆听！