清华大学统计学辅修选修课程

金融统计

（课程号：40160743）

李　　东

清华大学统计学研究中心

第2讲 金融数据的统计分析

数据的清洗，问题的提出，模型的建立，统计的思想

1. 经验数据

1.1金融数据的搜集和分析中的结构变化

金融数据经验分析的变迁

- 1970年代以前，在基本运作中的数据，都是固定在大时间区间上的数据：年、季、月、周. 出现的概率统计模型：Random walk, AR, MA, ARMA等模型，都是线性的.

用于低频数据

- 1980年代，为进行日数据分析，出现了非线性模型，比如ARCH、GARCH模型等.
- 1990年代起，有可能进行一日内的数据分析，与此紧密相连的是电子计算机的进步以及信息技术的飞速发展，大大提高了统计信息的搜集、记录、存储和分析的效率，可以说，它们是以几乎连续的方式来进行的。

信息的发布

- 常规的途径：日版出版、电视屏幕、各种网站，……
- 信息代理机构：
  - 路透 (Reuters)
  - 德励 (Telerate)
  - 奈特瑞德 (Knight Ridder)
  - 彭博 (Bloomberg)
  - ……

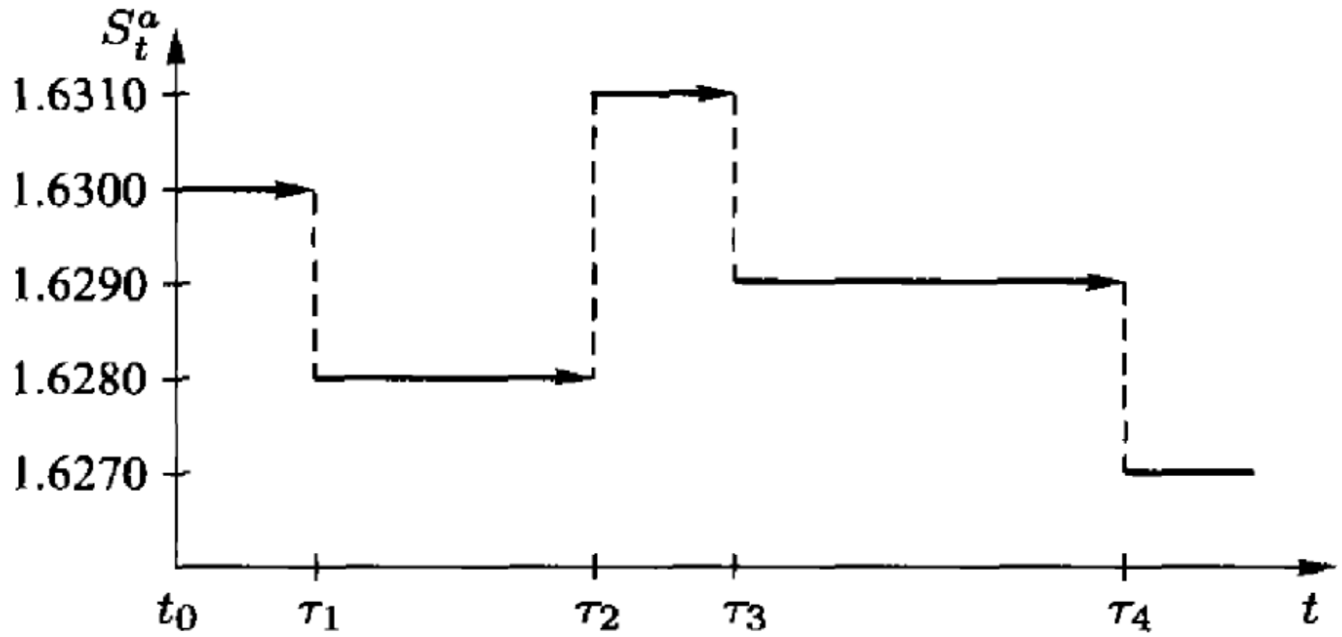
例子：路透社发布的**外汇**(与美元(USD)的比价)信息，从7:27以后，立即从显示屏上可以看到：

0727	DEM	RABO	RABOBANK	UTR	1.6290/00	DEM	1.6365	1.6270
0727	FRF	BUEX	UECIC	PAR	5.5620/30	FRF	5.5835	5.5588
0726	NLG	RABO	RABOBANK	UTR	1.8233/38	NLG	1.8309	1.8220

- ✓ 0726,0727是银行自己发布的格林威治时间(GMT)；
- ✓ DEM,FRF, NLG：外汇缩写 （德国马克、法国法郎、荷兰盾）；
- ✓ RABO,BUEX是分别在荷兰合作银行(RABOBANK,在乌得勒支，UTR)和UECIC银行(巴黎,PAR)的缩写；
- ✓ 1.6290是买入价(bid price),买入价后面的“00”意味着卖出价(ask price)是1.6300;
- ✓ 1.6365、1.6270是过去一天直到7:27的最高价和最低价；
- ✓ 第三行中的0726是指RABOBANK在7:26关于荷兰盾发布自己的报价为1.8233/38，至少在一分钟内没有一个银行(包括RABOBANK)发布新的报价.

如果取时刻$t_0 = 7{:}27$为时间起点，那么用**德国马克购买一美元**的价格（ask price）$S^a = (\mathrm{DEM/USD})^a$,

$$S_t^a = \left(\frac{\mathrm{DEM}}{\mathrm{USD}}\right)_t^a, \qquad t \geqslant t_0,$$



汇率 $S_t^a = \left(\frac{\mathrm{DEM}}{\mathrm{USD}}\right)_t^a$ $(t \geqslant t_0)$ 的性态

换句话说，(在区间$[t_0, \tau_1)$上的)某个时间，价格S_t^a“停留”在同一个水平上，即它不变；然后在时刻τ_1发生变换，或者说，产生**标记**(tick)，它是某个银行在时刻τ_1发布的新报价$s_{\tau_1}^a$，如此类推.

问题:

- (I) 标记之间的区间 $(\tau_{k+1} - \tau_k)$的长度的统计量是怎样的?
- (II)价格值变化(绝对量变化 $S_{\tau_{k+1}}^a - S_{\tau_k}^a$ 或相对量变化 $S_{\tau_{k+1}}^a / S_{\tau_k}^a$)的统计量是怎样的?

由所获得的数据来提取这些信息是汇率和其它金融指数演变**统计分析的首要问题**!

统计分析的目的

- 构造卖出价过程(S_t^a)，买入价过程(S_t^b)等的**概率统计模型**
- 理解金融指数的演变和价格形成的机制
- 建立对未来价格运动的预测

- 统计信息的搜集、记录、储存、处理，以至制作成便于高速提取的形式，是非常费事的，没有高技术是不可能达到的. 但是，同样明显的是，享有统计处理的结果并有可能运作运用它们，这对于在证券市场中的运营：构建各种项目、证券等的有效组合、合理投资，有着不可估量的好处.

- 当前的几乎不间断地搜集和获取统计信息的技术可能，揭示了金融指数性态的极为混乱不堪地随时间展开的**高频**特征。
- 这种高频特征在离散化时消失. 这样，在金融数学中出现“高频”争论，正是由这种几乎不间断获得统计信息的新可能所引起的，并且也正是这样搜集待处理的信息的现代技术与随时间变化的高频特征一起，允许揭示金融指数动态变化中的一系列专有的特点：

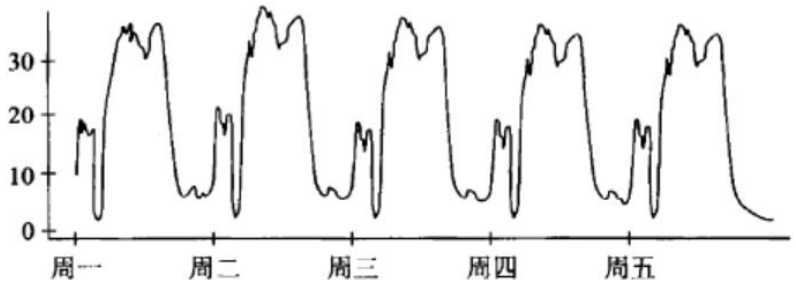
金融指数值形成的非线性特征以及表现为许多指数、价格等“记住”过去的后效.

- 高频性：出现**标记**次数的**强度**，同时也表达了统计数据的**海量**.
  - ✓ 1987.1.1-1993.12.31期间，(根据路透社的数据)DEM/USD汇率变化(**标记**，ticks)了 8 238 532 次. 其中 1 466 946次标记发生在1992.10.1-1993.9.30的一年左右. 同样在这一时期，对于汇率JPY/USD记录了570 814次标记.
  - ✓ 数据的高频特征：在典型的交易日时间中，汇率DEM/USD平均发生4.5千次标记，而JPY/USD的标记发生2千次.在1994年7月的某天，DEM/USD的标记量为9千次左右，每分钟发生15-20次标记.(平常一天的每分钟平均发生3-4次标记)
- 一般来说，在各种货币中，变化最激烈、频率最高的是汇率DEM/USD.各个代理机构所提出的报价(买卖价)不是具体交易时**真实的交易价**(transaction price).这种数据，以至交易量的数据，很难获得.
- 其它金融指数(包括股票、债券)或价格“标记”也具有类似的特征，其统计资料可以从很多机构获得，比如 NYSE等.

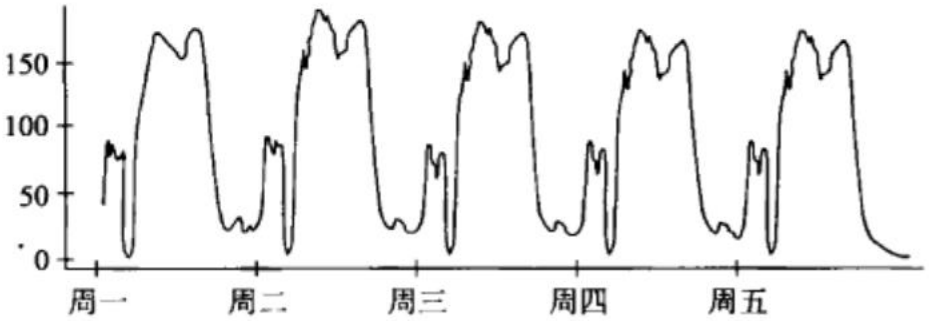
1.2汇率统计数据的“地理”特点

- 不同于(比如进行股票、债券、期货合约交易的，仅仅在“工作日”开盘的）交易所，**外汇市场**，或 **FX-市场** (Foreign Exchange, Forex)有许多专有的特点(“**周期性**”类型).
  - ✓ FX-市场按其实质而言是**国际市场**. 它不是地方化的，没有专门的地点，而是由广阔的遍布于全世界的银行、兑换点的网络分支所组成，它们都配备有良好的现代通信高技术手段.
  - ✓ FX-市场**在时间上是连续24小时工作的**. 在每星期的五天工作日中较为活跃，而其余时间以及节假日不太活跃.
  - ✓ 三个地理活跃带(按格林威治时间指示)：
    - ✓ 以东京为中心的东亚带， 21:00-7:00
    - ✓ 以伦敦为中心的欧洲带，7:00-13:00
    - ✓ 以纽约为中心的美洲带，13:00-20:00
- USD当作基本货币，最为活跃的重要货币：DEM, JPY, GBP, CHF(瑞士法郎).

- 汇率的活跃度：看标记的频率，或者汇率DEM/USD的图像来判断其**活跃性**.

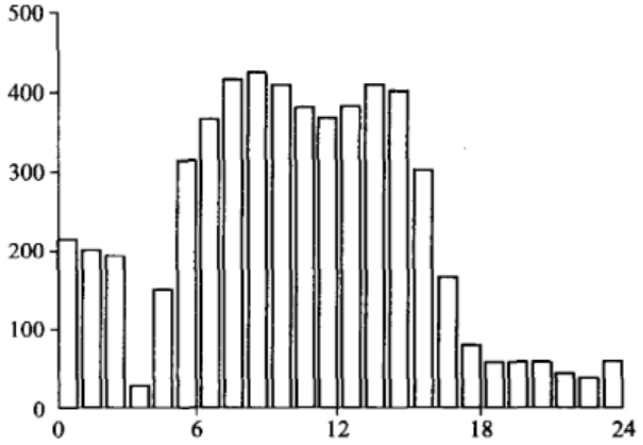


从周一到周五每隔 5 分钟记录的汇率 DEM/USD 的活跃程度特征的表示



每隔 20 分钟记录一次

- 纵轴记录了在这个时间区间中所发生的标记变化的**平均数**. 从图中清楚地看到昼夜上的**周期性**特点，它当然是取决于地球的自转和变化次数(标记)的**不均匀性**. 其中明显地可区分出三个**活跃峰**，他们与三个不同的地理带有关. 在欧洲和美洲的活跃峰大致是一样的. 在欧洲，活跃峰发生在下半日的前期，这时在美洲进入早晨，开始交易. 最不活跃的恰好发生在东京的早餐时间，这时，在欧美是深夜.



汇率 DEM/USD 在一个交易日 (24 小时) 内的平均 "标记" 数.

- 外汇市场是金融市场中最大的市场，根据国际结算银行的数据(1993年)，这个市场在1992年的日换手率到8320万亿美元！最著名的FX-市场的数据库之一属于“Olsen & Associates”根据这个数据库，(1987.1.1-1993.12.31期间)FX-市场的活跃度：

汇率	在数据库中的总标记数	一昼夜的平均标记数 (一年 52 周, 一周 5 个交易日)
DEM/USD	8 238 532	4500
JPY/USD	4 230 041	2300
GBP/USD	3 469 421	1900
CHF/USD	3 452 559	1900
FRF/USD	2 098 679	1150
JPY/DEM	190 967	630
FRF/DEM	132 089	440
ITL/DEM	118 114	390
GBP/DEM	96 537	320
NLG/DEM	20 355	70

1.3 金融指数演化的描述

- 在每一时刻t 都已知比如1美元(USD)值多少DEM的两个报价，卖出价(ask price)S_t^a和买入价(bid price) S_t^b.
- **价差**：$S_t^a - S_t^b$，它是市场状况的重要特征. 众所周知，价差和价格的波动率(这里就理解为通常的标准差)正相关. 从而波动率的增加，意味着与对价格运动的预测精度减小相联系的风险增加，它导致交易者增大价差，以作为补偿大风险的手段.

1.4 关于“标记”的统计

- **无条件**概率分布 Law$(\tau_1, \tau_2, \ldots)$什么是已知的问题. 对于 $k \geq 1$, 记

$$\Delta_k = \tau_k - \tau_{k-1},$$

并令 $\tau_0 = 0$.

- 有关{ Δ_k } 的模型
  - Random walk
  - Pareto 分布:

$$f_{\alpha b}(x) = \begin{cases} \dfrac{\alpha b^{\alpha}}{x^{\alpha+1}}, & x \geqslant b, \\ 0, & x < b. \end{cases}$$

  - ACD (Autoregressive Conditional Duration)模型

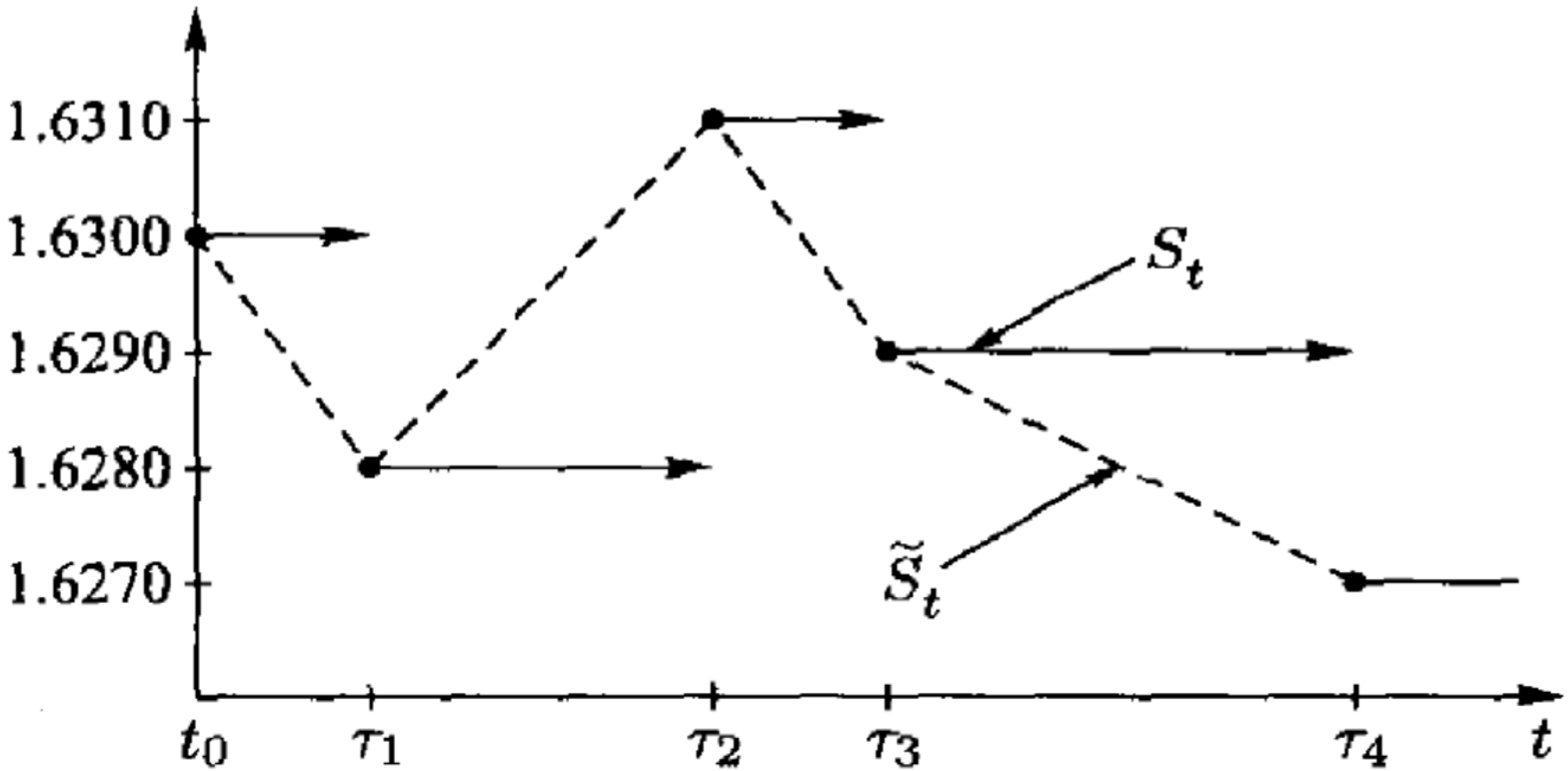
2. 一维分布的统计

2.1 统计数据的离散化

对于标记出现的强度和频率特征有了某种表示之后，同时也对间隔（$\tau_{k+1}-\tau_k$）的一维分布有了某种表示；现在转向价格变化性态的统计，即序列$(S_{\tau_k}-S_{\tau_{k-1}})$或者与此相联系的 $h_{\tau_k}=\ln(S_{\tau_k}/S_{\tau_{k-1}})$序列.

- “日数据”、“周数据”、“月数据”等，与“日内数据”之间的差别：通过等距有规则的时间间隔Δ来获得数据，如Δ可取为“一日”、“一周”等. 当转向一日之内的统计时，就会发生数据获得的**不规则性**，它们在偶然不均匀的时刻$\{\tau_k\}$, 以不同的间隔 $\{\Delta_k\}$来获得数据的，其中 $\Delta_k=\tau_k-\tau_{k-1}$.
- 这种不规则性对于应用已有的数据统计分析方法来说，带来一定的困难. 因此，统计数据先要经过某种预处理(离散化、剔除异常观测值、平滑化、分离出趋势项等).

- 离散化：
  - 固定某个“自然的”(实际物理)时间间隔Δ, 这个间隔不应该太“小”. 就汇率统计问题而言，必须使得这个间隔在下列意义下有**代表性**：其中必须落入显著的标记个数，或者， Δ显著大于两个标记之间的平均时间. 否则，所形成的离散化，统计“序列”将包含太多的“空白”数据. 在主要货币的汇率情形下，推荐$\Delta \geq 10$分钟.
  - 离散化的最简单的方法在于，选取Δ(比如10分钟、20分钟、24小时等)以后，取代有连续时间$t \geq 0$的按段常数过程$S = (S_t)_{t\geq 0}$, 而考察有离散时间$t_k = k\Delta$, $k = 0,1,\ldots$的新序列$S^{\Delta} = (S_{t_k})$.
  - 另外一种广泛流行的离散化方法：从**连续**修正常数过程.


1.6310
1.6300
1.6290
1.6280
1.6270
S_t
$\tilde{S}_t$
t_0
τ_1
τ_2
τ_3
τ_4
t

- 量子化：
  - 除了按时间离散化以为，统计数据也可以量子化，即可按相变量来**取整**. 通常如下进行：

    选取某个$\gamma > 0$, 并取代原来的过程$S = (S_t)_{t\geq 0}$，引入新过程$S(\gamma) = (S_t(\gamma))_{t\geq 0}$:

$$S_t(\gamma) = \gamma \left[\frac{S_t}{\gamma}\right].$$

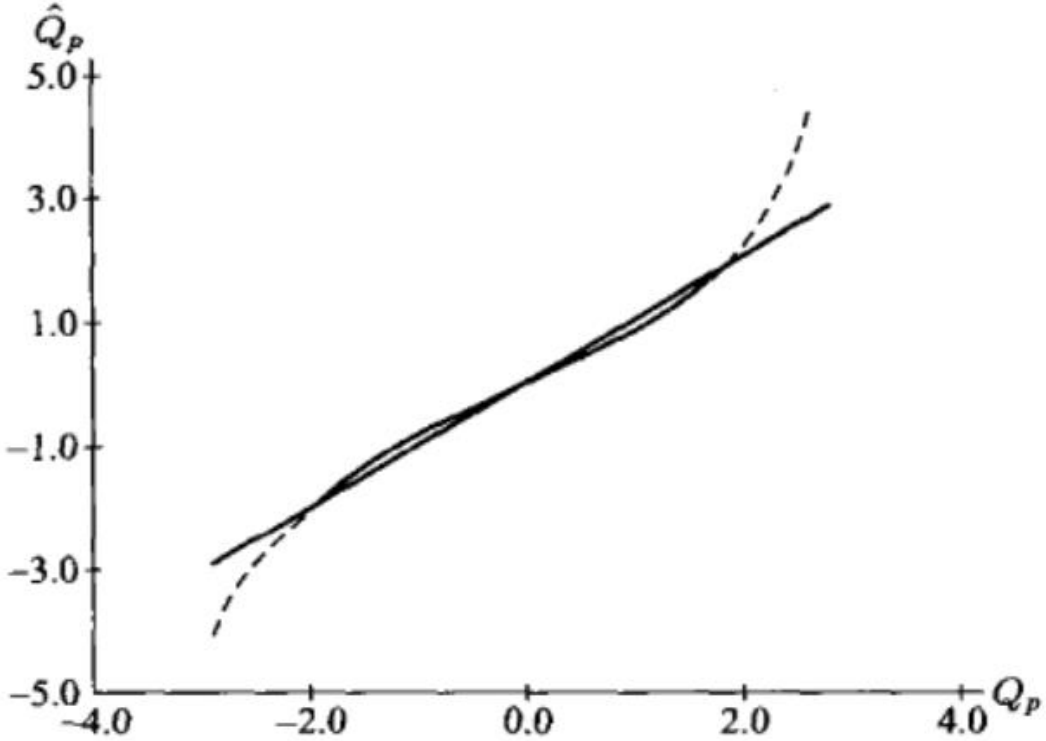
    $[\cdot]$为取整运算.

    如果先进行γ-量子化，然后再Δ-离散化，那么由(S_t)，我们可以得到 $S^{\Delta}(\gamma)$.

由于 $\gamma \to 0$ 时 $S_t(\gamma) \to S_t$, 这就发生一个重要问题: 应该用怎样的协调方式来选取 Δ 和 γ, 使得 $(S_{t_k}(\gamma))$ 在时刻 $t_k = k\Delta$ $(k = 0, 1, \cdots)$ 的值包含 "与 (S_t) 几乎同样的信息". 作为回答这一问题的第一种方法 (正如 G. Zhakod 所提出) 自然是要先阐明, 在怎样的 $\Delta \to 0$ 和 $\gamma \to 0$ 的收敛速度条件下, 使得对应的过程 $S^{\Delta}(\gamma)$, $\widetilde{S}^{\Delta}(\gamma)$ 的有限维分布收敛于极限过程 S.

2.2 相对价格变化的对数的 1-维分布

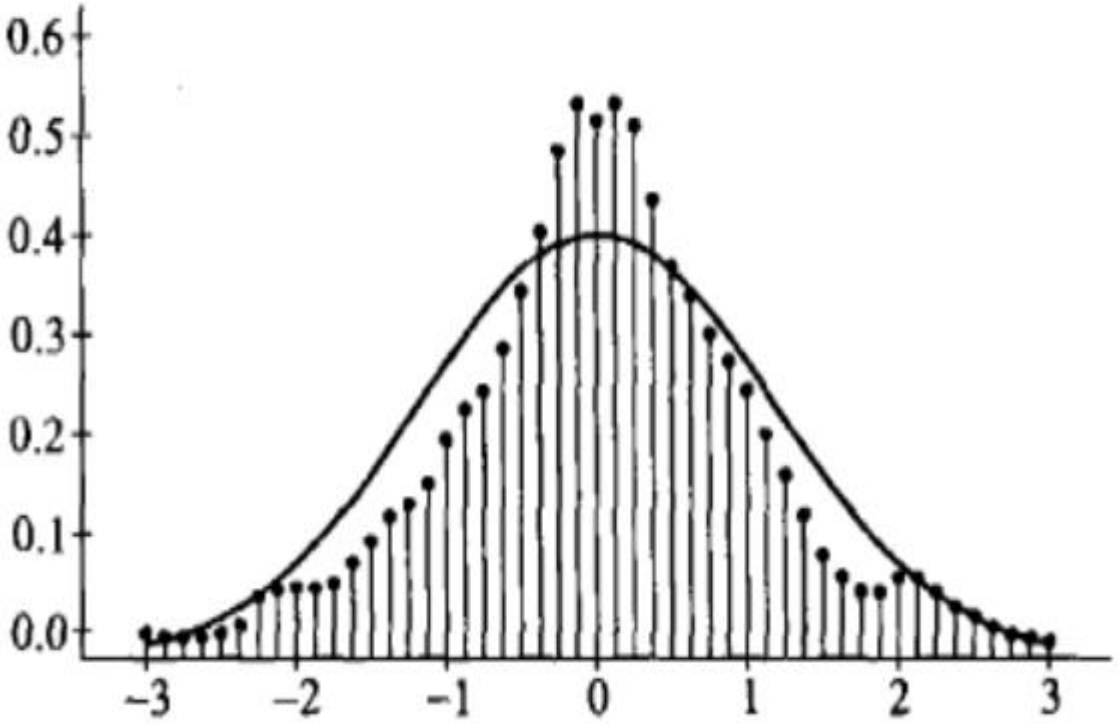
- 与高斯性质的偏差：重尾+尖峰



◆ 用Student t-distribution:

$$f_n(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

◆ Stable dist.



3. 价格中的波动率、相关依赖性和后效的统计

- 阅读《随机金融数学基础》(第一卷) (p.310-328)

4. 统计$\mathcal{R}/S$-分析

- **起源**：Hurst (1951)分析尼罗河逐年流量的统计数据存在**长记忆**和**自相似**现象. 促使Hurst创建了R/S-分析. 这种研究方法在统计实践中并不广为人知，这是因为相当稳健的Hurst方法可用来揭示统计数据中诸如序列值的聚集性、追随趋向方向的趋势性(倾向持续性, persistence)、强后效性、强记忆性、快速交替性(反持续性, anti-persistence)、分形性、具有周期和非周期的循环，噪声的“随机本性”和“混沌本性”的区分特征等等性质.

- 设 $S=(S_t)_{t\geq 0}$ 是某个金融指数，$h_n=\ln\left(\frac{S_n}{S_{n-1}}\right)$，$n\geq 1$. 令

$$H_n=h_1+\cdots+h_n,\quad n\geq 1.$$

$$\mathcal{R}_n=\max_{k\leqslant n}\left(H_k-\frac{k}{n}H_n\right)-\min_{k\leqslant n}\left(H_k-\frac{k}{n}H_n\right).$$

量 $\bar{h}_n\equiv\frac{H_n}{n}$ 是由样本 $(h_1,h_2,\cdots,h_n)$ 所构造的经验均值, 因而, $H_k-\frac{k}{n}H_n=\sum_{i=1}^{k}(h_i-\bar{h}_n)$ 是 H_k 与经验均值 $\frac{k}{n}H_n$ 的偏差量. 量 $\mathcal{R}_n$ 本身刻画的是这些偏差 $H_k-\frac{k}{n}H_n$ $(k\leqslant n)$ 的 "幅度".

$$\mathcal{S}_n^2=\frac{1}{n}\sum_{k=1}^{n}h_k^2-\left(\frac{1}{n}\sum_{k=1}^{n}h_k\right)^2=\frac{1}{n}\sum_{k=1}^{n}(h_k-\bar{h}_n)^2$$

$$Q_n \equiv \frac{\mathcal{R}_n}{S_n}$$

是积聚和 $H_k\ (k \le n)$的**规范幅度**，或者**调整幅度** (adjusted range)

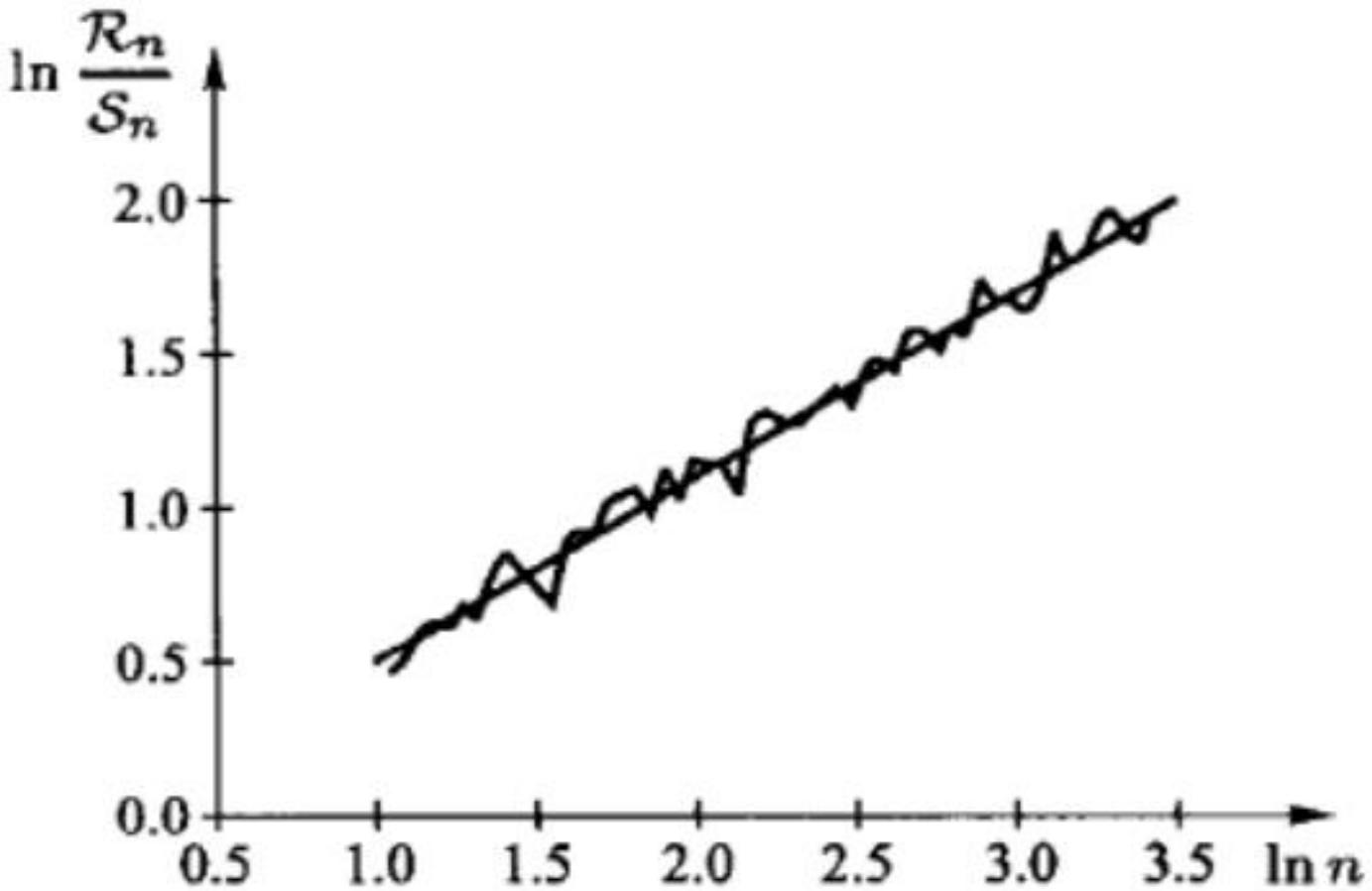
H_0：所考察的价格服从**random walk.**

如果H_0成立，那么对于充分大的 n,

$$\mathcal{R}_n/S_n \text{ 必定 “接近” 于 } \mathrm{E}_0\frac{\mathcal{R}_n}{S_n} \sim \sqrt{\frac{\pi}{2}n},$$

因此：

$$\ln\frac{\mathcal{R}_n}{S_n} \approx \ln\sqrt{\frac{\pi}{2}} + \frac{1}{2}\ln n.$$


$\ln \frac{R_n}{S_n}$
2.0
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
$\ln n$

上一图清楚地阐明了 R/S-分析方法：根据统计数据(按照对数尺度)列出 $(\ln n, \ln R_n/S_n)$, 再按最小二乘法估计直线 $\hat{a}_n + \hat{b}_n \ln n$. 如果发现$\hat{b}_n$以“显著方式”不同于1/2，那么应该拒绝$H_0$.

G. Hurst的研究的主要价值在于，他(用R/S-分析方法)发现，取代(对于尼罗河和其它河流)所期待的性质

$$\frac{\mathcal{R}_n}{\mathcal{S}_n} \sim cn^{1/2},$$

其实它是

$$\frac{\mathcal{R}_n}{\mathcal{S}_n} \sim cn^{\mathbb{H}},$$

其中Ⅱ显著大于1/2.

- 某些金融时间序列的$\mathcal{R}/S$-分析：
  - 道琼斯指数

Δ	观察值数 N	$\widehat{\mathbb{H}}_N$ 的估计值
1 天	12 500	0.59
5 天	2600	0.61
20 天	650	0.72

  - S&P500指数，取Δ =1分钟，5分钟和30分钟，

$$\widehat{\mathbb{H}} = 0.603,\ 0.590,\ 0.653.$$

清華大學 Tsinghua University · 统计学研究中心 CENTER FOR STATISTICAL SCIENCE

欢迎关注“水木数据派”