

# 第一章作业

---

## 背景

统计分析方法是广泛使用于金融大数据分析，其中股票分析是一个重要的部分。本次作业利用第一章学到的知识，对公开的股票行情数据（[Tushare](#)）进行分析。

## 题目

对沪深中100支股历史行情数据进行分析。历史行情数据的变量有：日期，开盘价，收盘价，最高价，最低价，成交量，股票代码。这里每日股价=（每日最低价+每日最高价）/2，数据见附件，请分析：

- 求股票000001（股票代码）的历史股价的日均值（所有天数的股价求平均）、中位数、0.25分位数、0.75分位数，方差，标准差，变异系数，极差，四分位极差，偏度，峰度。
- 对股票000006股价进行分析，选取合适组距，进行统计，画出的直方图（价格-频率）和正态QQ图，直观判断数据是否来自正态分布总体，给出简要的判断依据。如果对000006股价的差值（相邻两个日期的股价差值，忽略缺失日期，例如有 $t_1, t_3, t_4$ ，则差值为： $t_3 - t_1, t_4 - t_3$ ），同理计算差值的直方图和正态QQ图，判断差值是否服从正态分布，给出简要的判断依据。
- 对股票000012进行分析，求股价和成交量的Pearson，Spearman相关系数。
- 按照日期，对股票000001和股票000006的股价进行相关分析。例如股票000001在 $t_1, t_2, t_4, t_5$ 四个日期有记录 $x_1, x_2, x_4, x_5$ ；股票000006在 $t_2, t_3, t_4$ 三个日期有记录 $y_2, y_3, y_4$ ，那么我们选取有共同日期记录的值， $t_2, t_4$ 两个日期的记录，即 $(x_2, y_2)$ 和 $(x_4, y_4)$ 进行相关分析，而丢掉缺失数据（即 $t_1, t_3, t_5$ 日期的数据）。推广之，对100支股票两两进行分析，求100支不同股票股价的Pearson，Spearman相关矩阵（ $100 \times 100$ ）。根据相关矩阵，给出这100只股票中，相关性最强（绝对值接近1）的5对股票和相关性弱（绝对值最接近0）的5对股票，根据10支股票，求相关性假设的p值。（注意，Pearson，Spearman矩阵的元素排列依照股票代码，即，000001, 000006, 000012, ..., 000717）。

## 参考文献：

[Tushare官网](#)

[Tushare Github](#)

[Tushare文档](#)

