统计分析方法第一章作业

16337088　黄宏斌

一、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。

解：

假设天数为 n，每日最低价为 low_i，每日最高价为 $high_i$，(n=1,2,3,…,n)

日均值：$\frac{1}{n}\sum_{i=1}^{n}\frac{low_i+high_i}{2} = 14.270976945902762$

中位数：12.825

0.25 分位数：9.985

0.75 分位数：16.785

方差：$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 43.48392272115789$

标准差：$s = \sqrt{s^2} = 6.594234051135726$

变异系数：$CV = 100 \times \frac{s}{\bar{x}}(\%) = 46.20730645233751\%$

极差：42.849999999999994

四分位极差：6.800000000000001

偏度：$g_1 = \frac{n}{(n-1)(n-2)}\frac{1}{s^3}\sum_{i=1}^{n}(x_i-\bar{x})^3 = 1.8013931510673493$

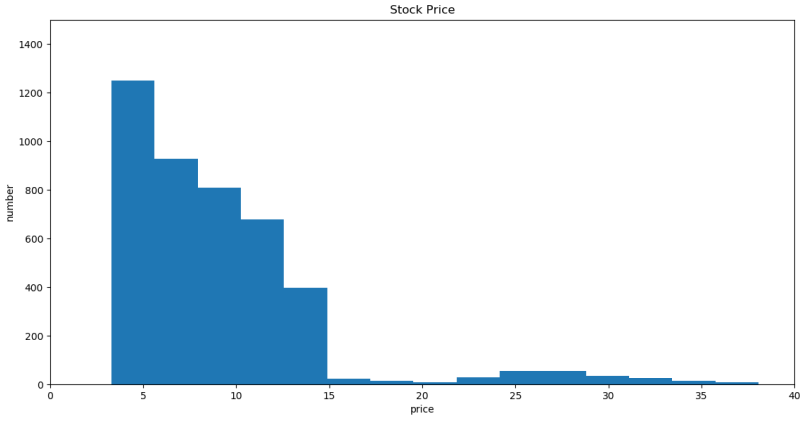
峰度：$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{1}{s^4}\sum_{i=1}^{n}(x_i-\bar{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)} = 4.4191734089528 79$

（代码见 problem1.py）

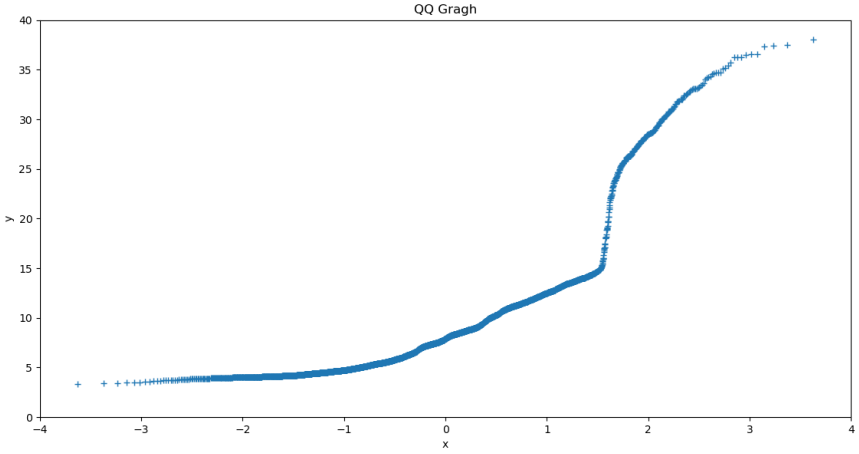
二、对股票 000006 股价进行分析，选取合适组距，进行统计，画出的直方图（价格-频率）和正态 QQ 图，直观判断数据是否来自正态分布总体，给出简要的判断依据。如果对 000006 股价的差值（相邻两个日期的股价差值，忽略缺失日期，例如有 t_1, t_3, t_4,则差值为: t_3- t1, t_4 - t_3）,同理计算差值的直方图和正态 QQ 图，判断差值是否服从正态分布，给出简要的判断依据。

解：

直方图（分为 15 组）：

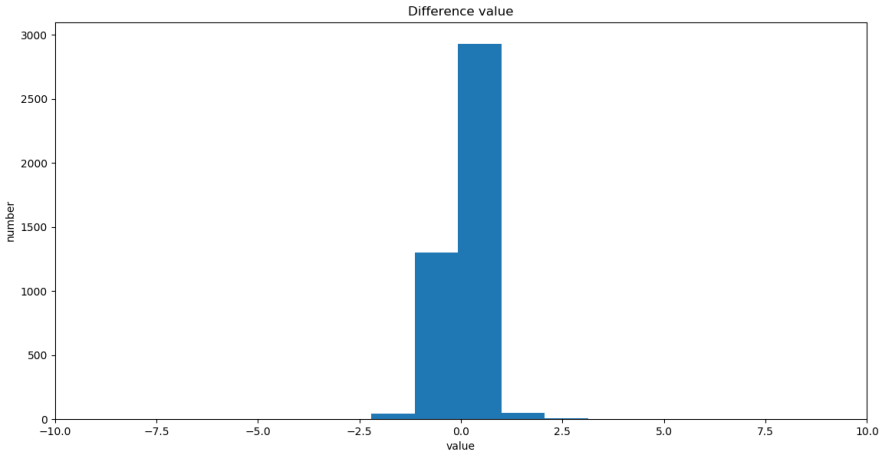


正态 QQ 图：

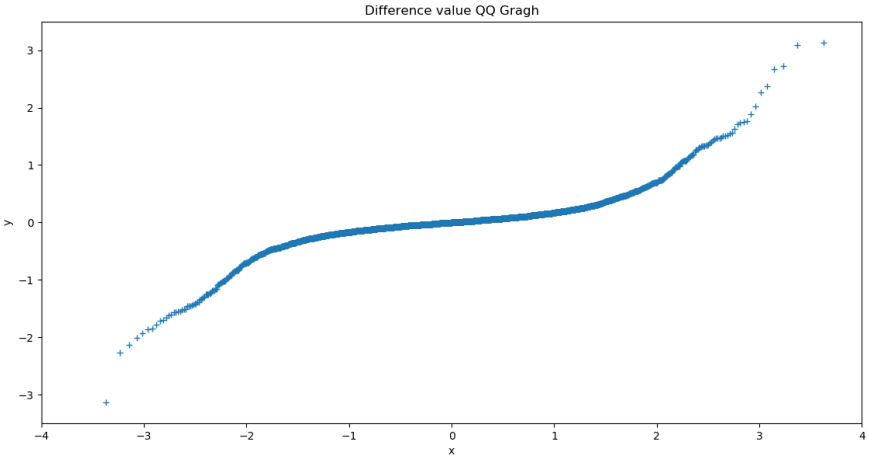


判断：不服从正态分布。
判断依据：直方图不近似钟形曲线的分布，且 QQ 图较弯曲，不近似在直线上。

差值直方图：



差值正态 QQ 图：



判断：不服从正态分布。

判断依据：直方图不近似钟形曲线的分布，且 QQ 图并不近似在直线上。（代码见 problem2.py）

三、对股票 000012 进行分析，求股价和成交量的 Pearson，Spearman 相关系数。

解：

Pearson 相关系数 r_{xy} 计算方法：

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}},$$

其中，

$$s_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$s_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Spearman 相关系数 q_{xy} 的计算方法：

$$q_{xy} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2$$

其中，

$d_i = R_i - S_i, i = 1,2,\dots,n$，Ri 和 Si 分别为 xi 和 yi 的秩统计量。

股票 000012 的股价和成交量的相关系数如下：

Pearson: 0.0298159151939864

Spearman: -0.018695776379890283

（代码见 problem3.py）

四、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, y2)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（100×100）。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 对股票，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)。

解：

P 值的计算：

① 提出假设：

H0：x_i 和 y_i 的相关系数 r=0

H1：总体 A 的相关系数 r≠0

② 计算检验的统计量：

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2), n = n_1 + n_2$$

其中，n_1 和 n_2 分别为 x 和 y 的个数。

③ 根据 t(n-2)计算 p 值。当 p<0.05 时，接受假设 H_0，否则拒绝 H_0。

对股票 000001 和股票 000006 的股价进行相关分析：
计算得到的相关系数如下：
Pearson: 0.777193654335734
Spearman: 0.5098981875846724

对 100 支股票两两进行分析：
计算得到的 Pearson 矩阵储存在文件"R_matrix.json"中；
计算得到的 Spearman 矩阵储存在文件"S_matrix.json"中。

其中，
按 Pearson 相关系数取前五：

股票 1	股票 2	Pearson 相关系数	p 值
000069	000046	0.9494846389143639	0.0
000046	000006	0.9464481049611131	0.0
000567	000025	0.9270517821307611	0.0
000069	000006	0.9150686814687823	0.0
000708	000059	0.9061597116938671	0.0

按 Pearson 相关系数取后五：

股票 1	股票 2	Pearson 相关系数	p 值
000632	000525	-0.00028548602317844473	0.5106557433813965
000090	000049	0.0006476817190809916	0.4761144782022949
000661	000521	-0.0008275803432526036	0.5307635433530755
000661	000601	0.0010438976200019272	0.4613204101930314
000425	000036	-0.001680314485977628	0.5616339824038029

按 Spearman 相关系数取前五：

股票 1	股票 2	Spearman 相关系数	p 值
000661	000028	0.9615927327084826	0.0
000567	000025	0.9431301195402602	0.0
000418	000025	0.9360251169020151	0.0
000661	000423	0.9311030852808425	0.0
000702	000421	0.9308496901011732	0.0

按 Spearman 相关系数取后五：

股票 1	股票 2	Spearman 相关系数	p 值
000667	000567	-0.0003671801722200385	0.5136945145192344

000538	000078	-0.00039850684350040133	0.5146794443646401
000088	000001	-0.000675695406699317	0.5248561509604942
000667	000418	0.0007424552672160578	0.472544761446966
000544	000425	-0.0013237493803239797	0.5484981027931313

（代码见 problem4.py）