统计方法分析　作业 1

这次作业使用 matlab 2016 版完成。

问题 1：

问题重述：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：通过 matlab 的函数可得到求得上诉数据。

Average: 14.270977

Median: 12.825000

0.25 Quantile: 9.985000

0.75 Quantile: 16.785000

Variance: 43.483923

Standard deviation: 6.594234

Coefficient of variation: 46.207306

Range: 42.850000

Four fraction range: 6.800000

Skewness: 1.800776

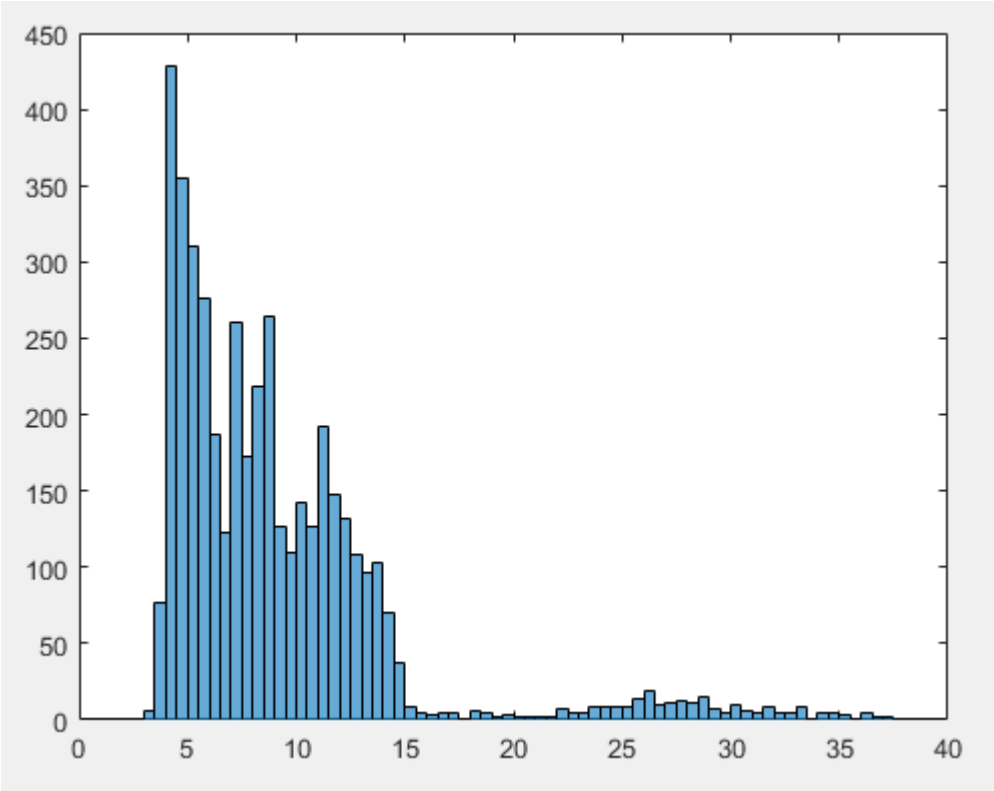
Kurtosis: 4.419173

问题 2：

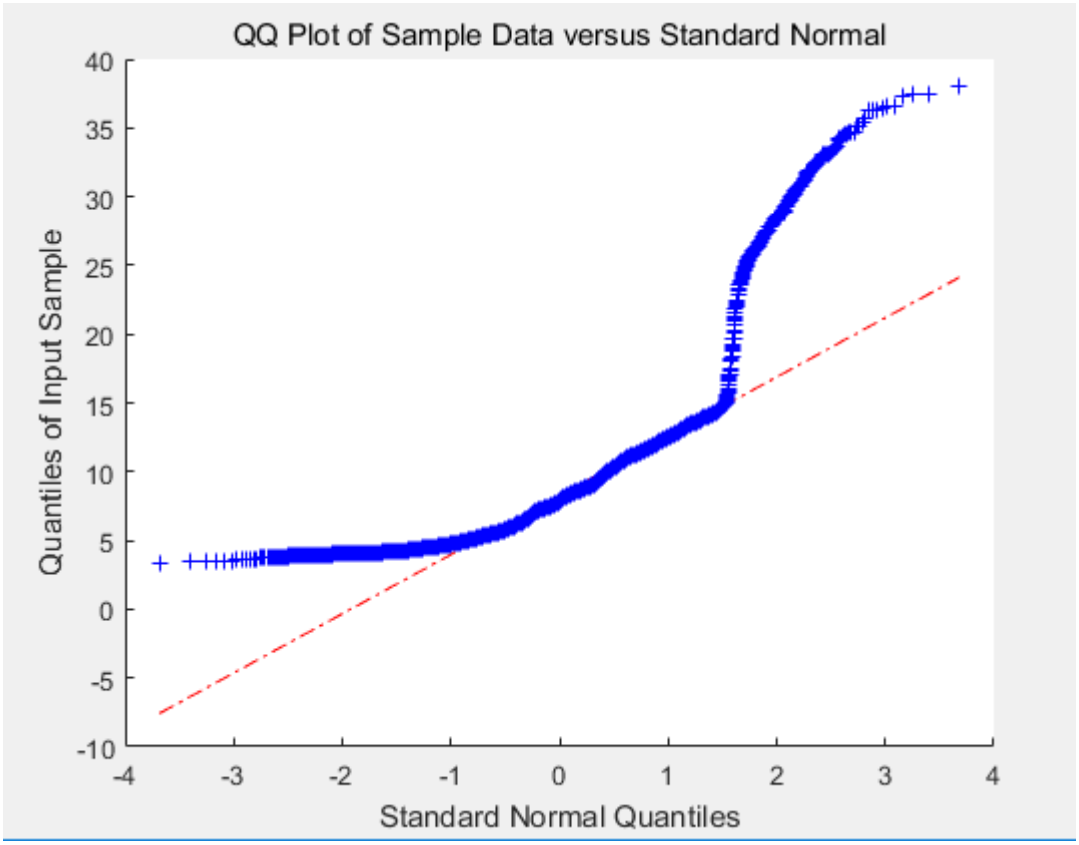
问题重述：对股票 000006 股价进行分析，选取合适组距，进行统计，画出的直方图（价格-频率）和正态 QQ 图，直观判断数据是否来自正态分布总体，给出简要的判断依据。如果对 000006 股价的差值（相邻两个日期的股价差值，忽略缺失日期，例如有 t_1, t_3, t_4,则差值为: t_3- t1, t_4 - t_3）,同理计算差值的直方图和正态 QQ 图,判断差值是否服从正态分布,给出简要的判断依据。

解：根据要求，对 000006 股价进行分析，有 4341 个数据，数据极差 34.760，

一般直方图的分组个数是$\sqrt{4341} \approx 66$，因此组距取$\frac{34.76}{66} \approx 0.5$。从而得出直方图：

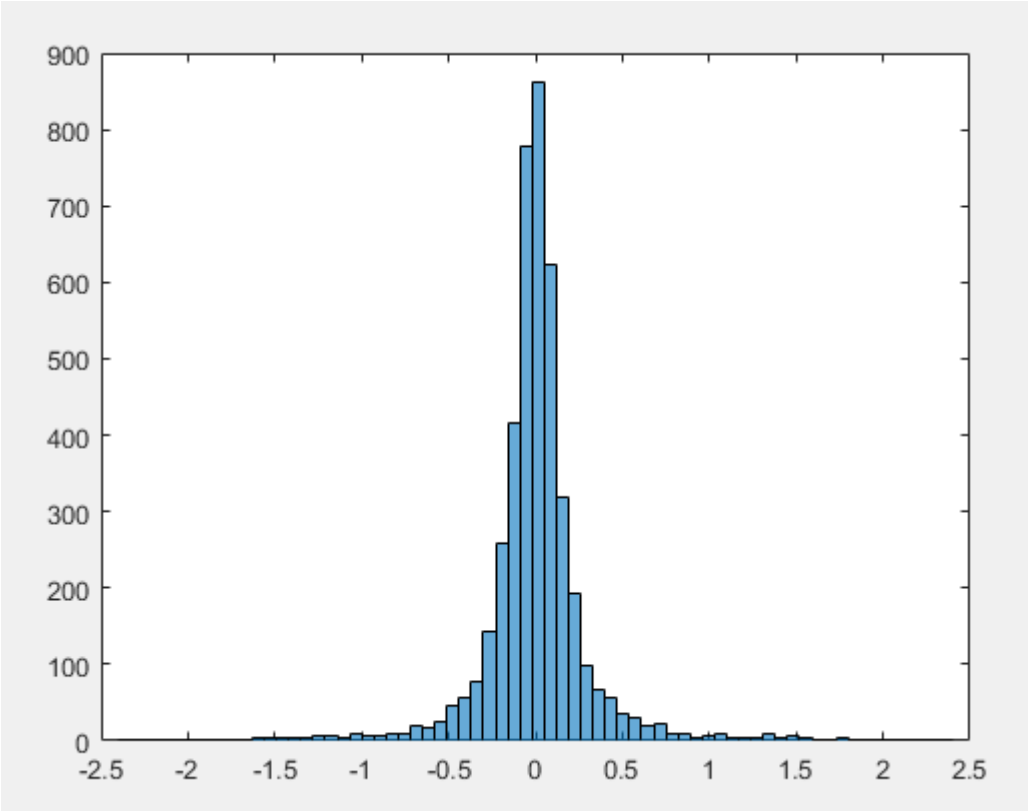


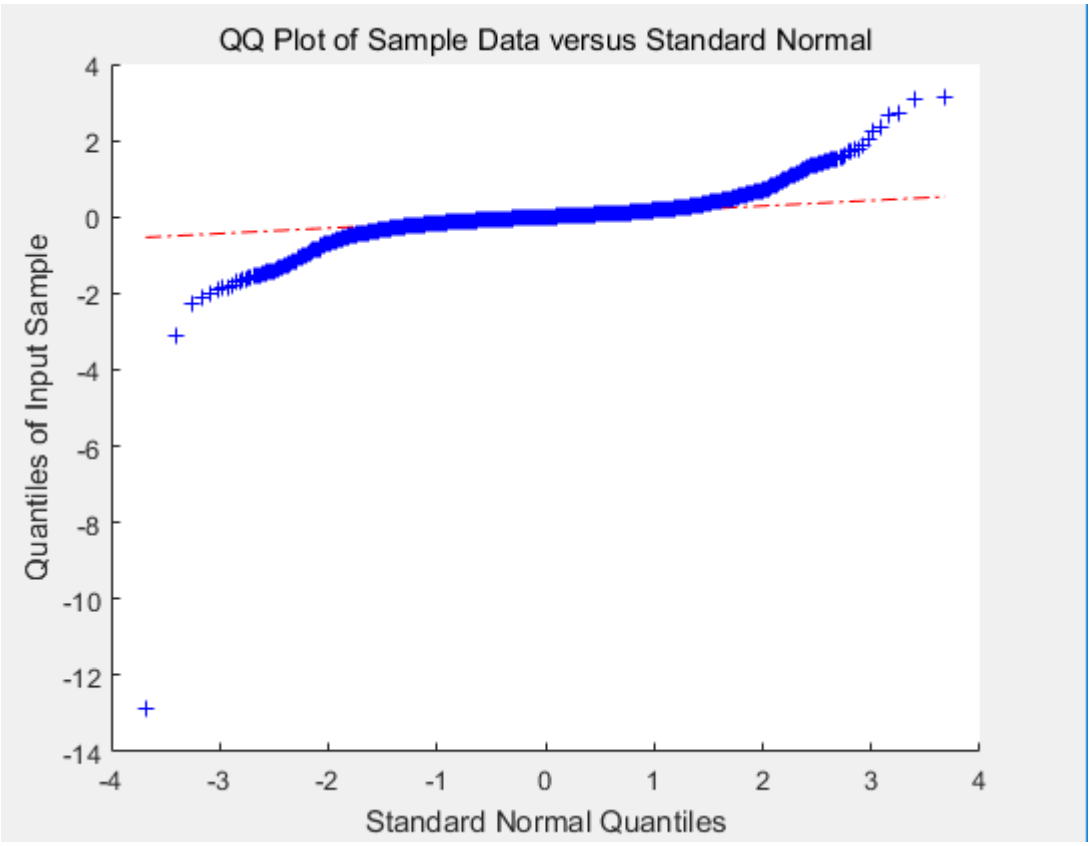
然后作出正态的 QQ 图：



分析：从直方图和 QQ 图中可以观察到，直方图的形状与正太分布非常的不相似，从 QQ 图中也能看出有很多数据偏离了$y = \sigma x + \mu$直线，因此数据不是来自正太分布总体。

接下来我们对 000006 估计就啊的差值进行分析，在这些数据的 4340 个值中，有-3.1350， 3.0900， 2.7200， 2.6700，-12.8850，-3.1300 不在区间[-2.4，2.4]中.其它大多数值都在该区间中，因此这 6 个值就先忽略，因此极差约等于 4.8，分组仍然接近$\sqrt{4340} \approx 66$，因此组距为$\frac{4.8}{66} \approx 0.07$。同样画出直方图和 QQ 图：





分析：直方图类似于正态分布，从 QQ 图可以看出，虽然有不少的点偏离了 $y = \sigma x + \mu$直线，但是它们都是在 0.25 分位数和 0.75 分位数之外的点，而且总体偏离值也不算大，去除少数极端值后，可以推断出 000006 股价的差值服从正态分布。

问题 3：

问题重述：对股票 000012 进行分析，求股价和成交量的 Pearson，Spearman 相关系数。

解：依题意，首先根据数据求出股价和对应的成交量，然后用 matlab 函数求出两者之间的 Pearson，Spearman 相关系数。

Pearson correlation coefficient: 0.029816

Spearman correlation coefficient: -0.018696

问题 4：

问题重述：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 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)。

解：先对股票 000001 和股票 000006 的股价进行相关分析，得出它们的 Pearson 相关系数和 Spearman 相关系数，如下：

Pearson correlation coefficient: 0.777194

Spearman correlation coefficient: 0.509898

依题意，计算 100 支不同股票的 Pearson 和 Spearman 相关矩阵（数量太大放在附件中）。根据 Pearson 相关矩阵，得出相关性最强的 5 对股票如下：

000708 000059 Pearson correlation coefficient: 0.906339 p value:0.000000

000069 000006 Pearson correlation coefficient: 0.915083 p value:0.000000

000567 000025 Pearson correlation coefficient: 0.927097 p value:0.000000

000046 000006 Pearson correlation coefficient: 0.946432 p value:0.000000

000069 000046 Pearson correlation coefficient: 0.949496 p value:0.000000

Spearman 矩阵中的相关矩阵得出的相关性最强的 5 支股票如下：

Spearman Max

000702 000421 Spearman correlation coefficient: 0.930884 p value:0.000000

000661 000423 Spearman correlation coefficient: 0.931183 p value:0.000000

000418 000025 Spearman correlation coefficient: 0.936061 p value:0.000000

000567 000025 Spearman correlation coefficient: 0.943159 p value:0.000000

000661 000028 Spearman correlation coefficient: 0.961674 p value:0.000000

Pearson 相关矩阵，相关性最弱的 5 对股票如下：

000661 000521 Pearson correlation coefficient: 0.000554 p value:0.956478

000632 000525 Pearson correlation coefficient: 0.000841 p value:0.984932

000425 000036 Pearson correlation coefficient: 0.001079 p value:0.912675

000090 000049 Pearson correlation coefficient: 0.001347 p value:0.966216

000598 000062 Pearson correlation coefficient: 0.001428 p value:0.883596

Spearman 矩阵中的相关矩阵得出的相关性最弱的 5 支股票如下：

000088 000001 Spearman correlation coefficient: 0.000092 p value:0.964823

000661 000012 Spearman correlation coefficient: 0.000264 p value:0.917798

000667 000567 Spearman correlation coefficient: 0.000544 p value:0.980600

000538 000078 Spearman correlation coefficient: 0.000630 p value:0.979225

000717 000525 Spearman correlation coefficient: 0.000823 p value:0.920794

相关性强的两个股票之间的 p 值都是 0, 表明拒绝两个列之间不存在相关性的假设。相关性不强的 p 值都是非 0,这里的相关系数的值都是取了绝对值的。

程序说明：

question1.m, question2.m, question3.m, question4.m,分别表示问题 1 到问题 4 的解题代码。question1.txt, question3.txt, question4.txt,分别是问题 1，3，4 的结果。而且问题 4 的输出结果在 question4.txt, Pearson.xlsx,Spearman.xlsx 三个文件中，而 question4.txt 放的是问题 4 中的股票 000001 和股票 000006 的股价的相关分析。Pearson.xlsx 和 Spearman.xlsx 分别放了 100 个股票的 pearson 相关矩阵和 spearman 相关矩阵，然后 10 对股票的 p 值在此实验报告中，运行 question4.m 文件中的相关代码能把 p 值生成在变量 PMaxh 和 PMin 中。

运行数据为老师提供的数据，放在同一个目录里一个名为 data_selected 的

文件夹中。

可能是因为 matlab 的问题，生成 pearson 和 spearman 矩阵要花 30 分钟左右，因此这里附上 matlab 生成的两个矩阵的变量 R_Pearson_Table 和 R_Spearman_Table。