2022 年 7 月 A 股所有股票市值变化率统计分析

摘要

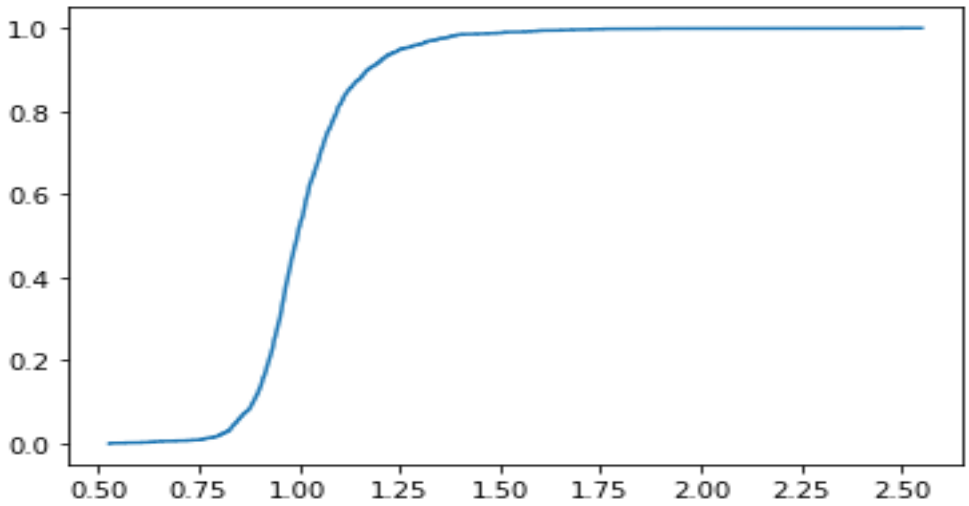
本文通过 python 程序借助开源网站 baostock.com 爬取了 2022 年 7 月份 A 股所有股票的市值，从而计算出在这个月内他们的变化率作为数据集。根据经济学基本知识，猜测应当服从对数正态分布，于是进行了一系列统计学分析，包括计算数字特征、估计参数与分布、检验、方差分析、回归分析等。同时由于具有良好的可扩展性，例如日期、时间跨度和股票过滤，故作为项目发布于 github.com

正文

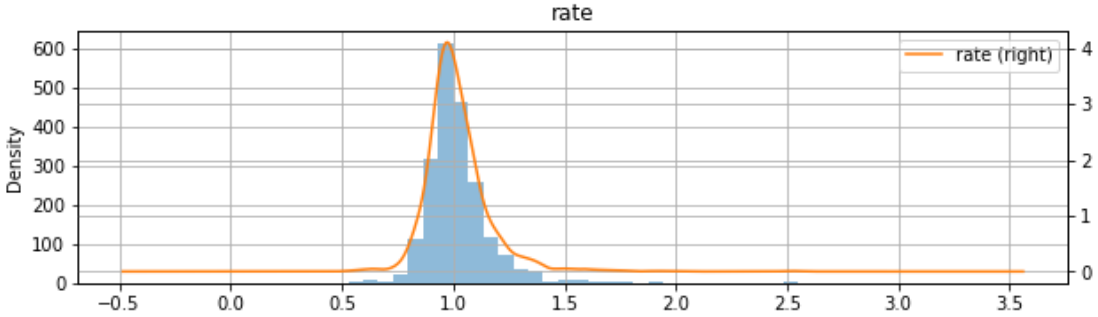
本文数据附于文章末尾。

变化率的均值为 1.0181518541654626，方差为 0.02050602513579442。

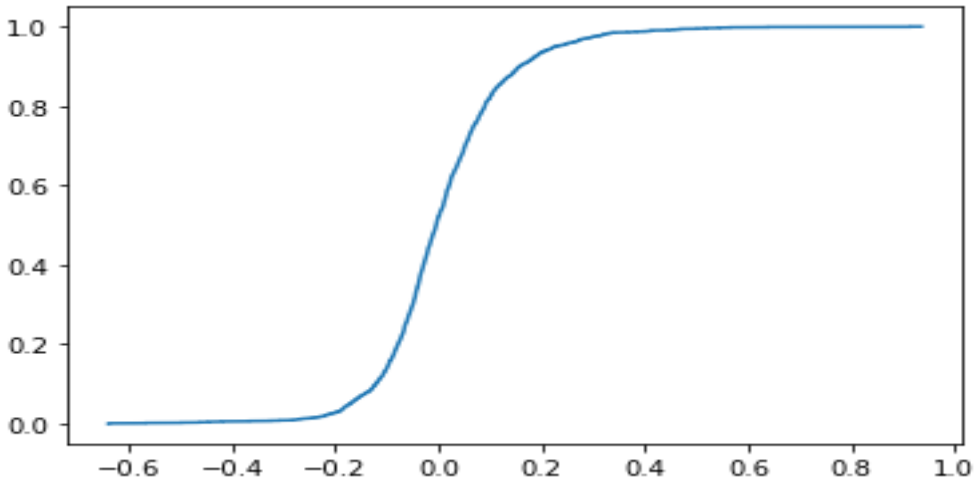
经验分布函数为：



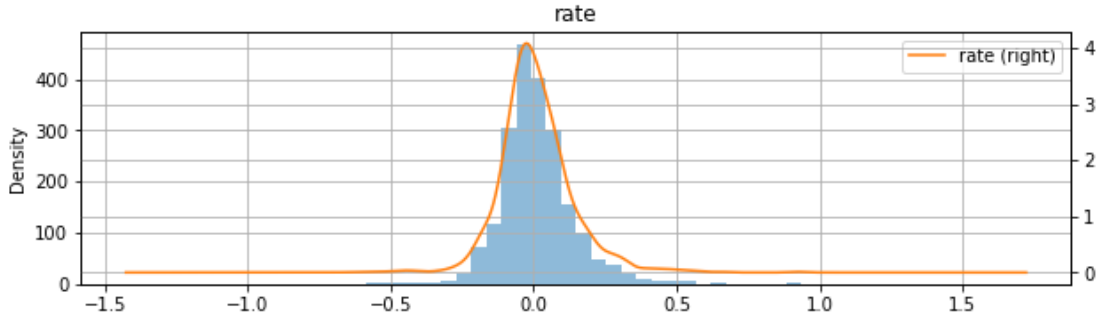
阶梯函数为（近似为经验密度函数）：



对数化之后经验分布函数为：



对数化之后阶梯函数为：

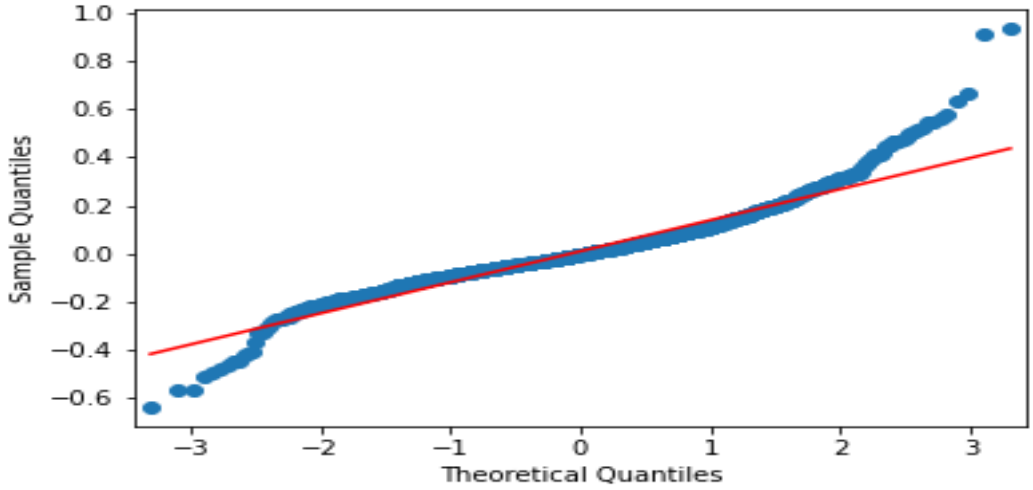


接下来进行分布估计，对对数化之后的变化率进行分析，根据对数化之后的经验函数可进行假设检验：

$$H_0: F(x) = F_0(x)\,,\ H_1: F(x) \neq F_0(x)$$

其中 F(x)是总体ξ的分布函数，$F_0(x)$是已知的分布函数，即正态分布$N(a,\sigma^2)$

绘制 Q-Q 图如下：



K-S 检验 p 值为 2.552216429613483e-294，明显小于 0.05，故拒绝原假设，认为变化率不服从对数正态分布。但实际上查阅相关文献可以指出，在面对样本量较大时，K-S 检验对于样本过于敏感，并且与数据可视化图表的观察结果不一致。

利用 W 检验得出的 p 值也为 1.5558938019348248e-29 < 0.05，故仍然拒绝原假设。但由于图像工具显示更具有说服力，故仍然相信本月变化率服从对数正态分布，并以对数化之后的变化率估计参数。

参数估计：

$$\hat{a} = \bar{\xi}, \quad \widehat{\sigma^2} = \tilde{S}^2$$

可得

$$\hat{\xi} = 0.009235628908394987$$

$$\widetilde{S^2} = 0.016787468092508993$$

区间估计，在置信度 0.95 的情况下，得出 a 的 0.95 置信区间为

(0.0036895621609148764, 0.014781695655875098)

σ^2的 0.95 置信区间为

(0.015816159567639192, 0.017851539757211463)

从而σ的 0.95 置信区间为

(0.1257623137813518, 0.13360965443115053)

做参数假设检验，构造枢轴量双侧检验均值与方差；构造统计量用于检验均值：

$$T = \frac{\bar{\xi} - a_0}{S/\sqrt{\{n-1\}}}$$

其否定域为：

$$\{|T| > t_{\{1-\frac{\alpha}{2}\}}(n-1)\}$$

构造K^2 统计量用于检验方差:

$$K^2 = (n-1)\widetilde{S^2}\,/\sigma_0^2$$

其否定域为:

$$\{K^2 < \chi^2_{\{\frac{\alpha}{2}\}}(n-1)\}$$

或

$$\{K^2 > \chi^2_{\{1-\frac{\alpha}{2}\}}(n-1)\}$$

计算时根据现实情况 n=2099,$\alpha = 0.95$,假设的参数使用参数估计时的统计量，但若直接使用毫无意义，应当分成多组，互相检验，计算得出分成两组时，均

落在否定域之外，即接受原假设，认为均值和方差为第一组（或者第二组）的统计量。

做回归分析，以总量为影响因素和变化量的绝对值为未知变量，得到模型观察截距项和系数，分别为：

$$\beta_0 = 1.021256527593252, \qquad \beta_1 = -7.43299e-13$$

预测值的影响因素就取为 2022 年 7 月所有 A 股股票的总量，其结果与真实数据非常接近，具体数据超过 2000 行，附于仓库内，文件位置为

data\pred_rate_2099.csv

读者可自行查看。

至于方差分析需要多得多的精力投入，寻找影响因子，分析对应数据，几乎已经是独立完成量化分析。如果能够给大作业更多的分数以及更长的时间或许会在兴趣的驱使下去完成这项有趣而富有挑战性的工作！然而迫于现实，数学系终究难像工科学生一样体会到写大作业和项目的快乐（不乏批判者，然而至少我觉得计网和 ics 的 pa 都很有意思），我得把精力放到备考上去了，数理统计大作业就此结束。

总结

本文数据来源真实可信，使用开源网站 api 自主编写代码爬取。可扩展性强，比如说方差分析部分可以作为现实世界的工具。同步发布于 github 项目，原创性强。但仍碍于时间不充足、作者水平局限具有相当瑕疵，望老师包涵。谢谢老师一个学期以来的教导，又学完了一门有趣而富有挑战性的课程。

参考文献

无