统计计算

方差分析

组员：闫超、苏浩然、黎思言

1 研究思路

在中国的 2000 多只 A 股中，每天都有一些股票上涨另一些股票下跌。那么，对于这些上涨和下跌的股票，它们的成交量是否有显著性差别呢？这是我们研究的问题。在本文中，我们用方差分析的思路来解决这个问题。

2 研究数据

我们的数据来自于另一个课程的课程作业。该数据为 2003 年 3 月到 2015 年 4 月，中国 A 股的所有股票每天的的开盘价、收盘价、最高价、最低价、成交量、涨跌幅、涨跌量等指标的面板数据，原始数据有 500 多万条。如下是该数据的数据结构：

表 1: 数据结构 1

字段	数据类型	单位	解释
datetime	character	日	日期
trade_code	character	无	股票代号
open	numeric	元	开盘价
high	numeric	元	最高价
low	numeric	元	最低价
close	numeric	元	收盘价
volume	numeric	元	交易规模
chg	numeric	元	涨跌数值
pct_chg	numeric	无	涨跌百分比
pct_chg2	factor	无	涨跌情况

我们的研究思路是，将上述数据切分为截面数据，分析截面数据中上涨的股票和下跌的股票在成交量上是否有显著性差别。我们选取的截面有九天，这九天分别为”2013-3-4”, ”2013-6-3”, ”2013-9-2”, ”2013-12-2”, ”2014-3-3”, ”2014-6-3”, ”2014-9-1”, ”2014-12-1”, ”2015-3-2”。我们选取截面的原则是，从 2013 年开始，每个季度最后一个月的第一个周一，如果第一个周一没有数据，就选取当天之后有记录的一天的数据。

每一个横截面的数据量如下所示：

表 2: 数据结构

时间	跌	涨
2013-3-4	2134	257
2013-6-3	1466	906
2013-9-2	714	1640
2013-12-2	2171	166
2014-3-3	272	2104
2014-6-3	1229	1077
2014-9-1	144	2160
2014-12-1	1559	732
2015-3-2	288	2087

如上表所示，每一个截面不同类别中的数据量都超过了 36，满足研究的需要。

3 方差分析

我们将要采取方差分析的方法研究上涨和下跌的股票成交量是否存在显著性差别。方差分析首先要满足如下四个基本假设。

（1）各处理条件下的样本是随机的；

（2）各处理条件下的样本是相互独立的；

（3）各处理条件下的样本分别来自正态分布总体；

（4）各处理条件下的样本方差相同。

对于双因素方差分析而言，各处理条件下的样本是相互独立的，满足了假设二，我们假设所有上涨的 A 股和所有下跌的 A 股的成交量是随机的，满足了假设一。我们需要验证假设三和假设四。

3.1 正态性假设检验

我们使用直方图和计算偏度系数的方法验证数据的正态性。如下是各水平下的 A 股成交量的直方图、偏度系数与峰度系数。

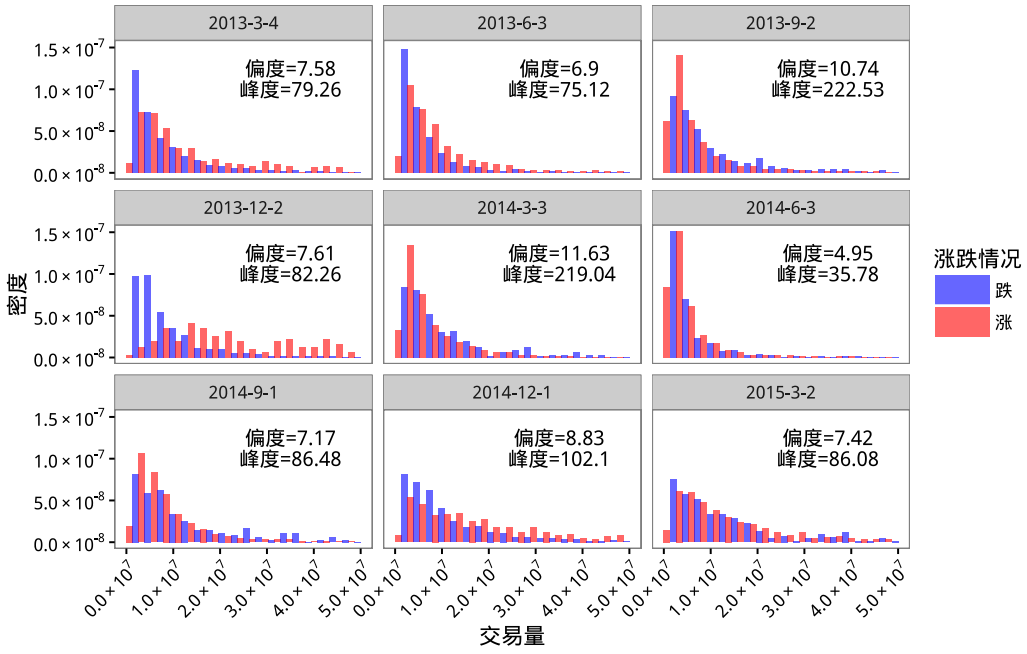

图 1: 成交量直方图

如上图所示，在九个截面数据中，各水平的成交量偏度系数都远大于 0，呈现出右偏分布的特征。说明我们的原始数据不满足正态性假设条件。

对于不满足正态性假设的情况，一般有两种解决措施，第一种方法是采取非参数检验的方式替代方差分析。第二种方法是采取适当的方法对原始数据做变换，使得数据满足正态性假设。

$Box-Cox$ 变换是统计建模中常用的一种数据变换，用于连续的响应变量不满足正态分布的情况。变换之后，可以一定程度修正数据的正态性水平。$Box-Cox$ 变换的公式如下所示：

$$f(y)=\begin{cases}\frac{y^{\lambda}-1}{\lambda} & \text{if}\lambda\neq 0,\\ log(y) & \text{otherwise.}\end{cases}$$

在本文中，我们采取的办法是使用 $Box-Cox$ 变换修正原始数据的正态性水平。

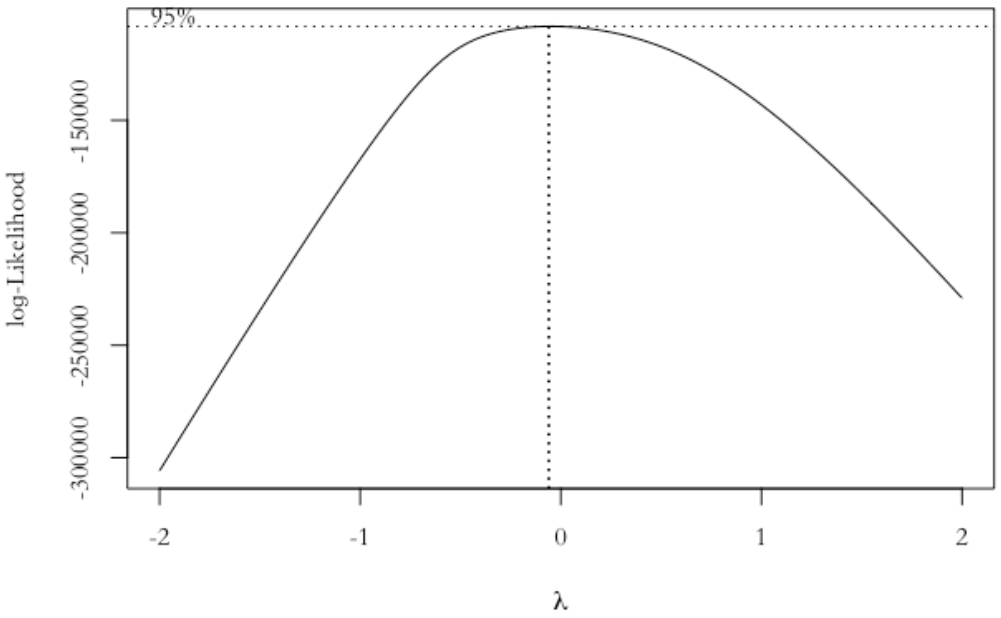

图 2: Box-Cox 变换

如上所示，我们选定 λ 为-0.1.

经过变换之后的成交量直方图以及偏度和峰度系数如下所示：

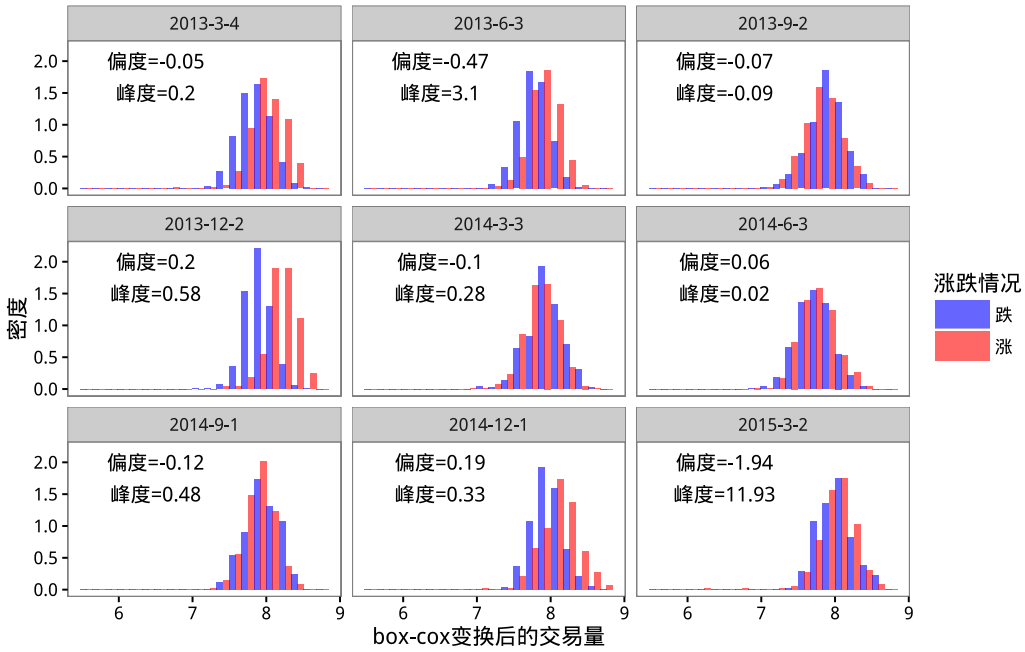

图 3: 成交量直方图 2

可知数据已经基本满足正态性假设条件。

3.2 方差齐性检验

我们计算了九天每个水平下的方差，如下表所示：

表 3: 方差齐性检验

时间	状态	标准差	状态	标准差
2013-3-4	涨	0.24	跌	0.23
2013-6-3	涨	0.21	跌	0.21
2013-9-2	涨	0.25	跌	0.24
2013-12-2	涨	0.2	跌	0.18
2014-3-3	涨	0.22	跌	0.24
2014-6-3	涨	0.24	跌	0.24
2014-9-1	涨	0.2	跌	0.22
2014-12-1	涨	0.25	跌	0.21
2015-3-2	涨	0.28	跌	0.24

如上表所示，每一天各水平下的方差相差不大，我们认为数据满足了方差齐性检验。可以使用方差分析。

3.3 方差分析

H_0 : 一天之内上涨的股票和下跌的股票成交量相等；

H_1 : 一天之内上涨的股票和下跌的股票成交量不相等。

表 4: 方差分析

时间		自由度	平方和	均方	F 统计量	p 值
2013-3-4	组内	1	4.03	4.03	76.42	10^{-16}
	组件	2389	125.84	0.053		
2013-6-3	组内	1	5.86	5.86	132.6	10^{-16}
	组间	2370	104.7	0.044		
2013-9-2	组内	1	7.0	7.0	115.6	10^{-16}
	组间	2352	77.75	0.033		
2013-12-2	组内	1	12.88	12.88	386.7	10^{-16}
	组间	2335	77.75	0.033		
2014-3-3	组内	1	1.93	1.93	38.65	5×10^{-10}
	组间	2374	118.33	0.0498		
2014-6-3	组内	1	0	0	0.072	0.789
	组间	2304	130.5	0.0566		
2014-9-1	组内	1	0.82	0.82	19.81	9×10^{-6}
	组间	2302	95.01	0.0413		
2014-12-1	组内	1	7.54	7.54	149.7	10^{-16}
	组间	2289	115.31	0.05		
2015-3-2	组内	1	0.11	0.11	1.466	0.226
	组间	2373	176.35	0.074		

如上表所示，在大部分的分组中，方差分析检验的结果都拒绝原假设，认为上涨和下跌的股票成交量有显著性差别，但是在 2014 年 6 月 3 日和 2015 年 3 月 16 日，方差分析不能拒绝原假设，在这两天，没有足够的理由认为上涨股票和下跌股票的成交量存在显著性差别。

4 代码

```
setwd("D:/Documents/作业/数据可视化作业/第二次作业")

library(ggplot2)
library(showtext)
library("latex2exp")
library(MASS)
library(fBasics)
showtext.auto(enable=T)
```

```
par(family="STSong")

load("dat2.rda")
dat2=dat2[dat2$volume>0,]
dat2$pct_chg[dat2$pct_chg>10]=10
dat2$pct_chg[dat2$pct_chg< -10]=-10
sam.time=unique(dat2$datetime)
dat2$pct_chg2=ifelse(dat2$pct_chg>=0,"涨","跌")

#正态性检验
tex1=data.frame(x=3.5e7,y=rep(1.25e-7,9),sk=sapply(sam.time,function(x)
    skewness(dat2$volume[dat2$datetime==x])),datetime=sam.time)
tex2=data.frame(x=3.5e7,y=rep(1e-7,9),ku=sapply(sam.time,function(x)
    kurtosis(dat2$volume[dat2$datetime==x])),datetime=sam.time)
ph1=ggplot(data=dat2,aes(x=volume))+
  geom_histogram(aes(y=..density..,fill=pct_chg2),
                 alpha=0.6,bins=20,position="dodge")+
  scale_fill_manual(values=c("blue","red"))+
  scale_x_continuous(limits=c(0,5e7),breaks=c(0,1e7,2e7,3e7,4e7,5e7),
                     labels=c(TeX("$\\0.0\\times10^{7}$"),
                              TeX("$\\1.0\\times10^{7}$"),
                              TeX("$\\2.0\\times10^{7}$"),
                              TeX("$\\3.0\\times10^{7}$"),
                              TeX("$\\4.0\\times10^{7}$"),
                              TeX("$\\5.0\\times10^{7}$")))+
  scale_y_continuous(breaks=c(0,5e-8,1e-7,1.5e-7,2e-7),
                     labels=c(TeX("$\\0.0\\times10^{-8}$"),
                              TeX("$\\5.0\\times10^{-8}$"),
                              TeX("$\\1.0\\times10^{-7}$"),
                              TeX("$\\1.5\\times10^{-7}$"),
                              TeX("$\\2.0\\times10^{-7}$")))+
  geom_text(data=tex1,aes(x=x,y=y,label=paste("偏度=",round(sk,2),sep="
    ")))+
  geom_text(data=tex2,aes(x=x,y=y,label=paste("峰度=",round(ku,2),sep="
    ")))+
  labs(x="交易量",y="密度",fill="涨跌情况")+
  theme_bw()+
  theme(panel.grid=element_blank(),
        axis.text.x=element_text(angle=45,hjust=1),
        legend.position="right",
        legend.key.width=unit(1.1,"cm"),
        legend.key=element_rect(colour='white',
```

```
                                        fill='white',
                                        size=1))+
  facet_wrap(~datetime,ncol=3)
print(ph1)
ggsave("成交量直方图.pdf",width=8,height=5)

#box-cox变换
mod1=aov(dat2$volume~dat2$pct_chg2)
summary(mod1)
bc=boxcox(mod1,lambda=seq(-2,2,0.1))
dat2$volume2=(dat2$volume^(-0.1)-1)/(-0.1)
tex1=data.frame(x=6.5,y=rep(2,9),sk=sapply(sam.time,function(x)skewness
    (dat2$volume2[dat2$datetime==x])),datetime=sam.time)
tex2=data.frame(x=6.5,y=rep(1.5,9),ku=sapply(sam.time,function(x)
    kurtosis(dat2$volume2[dat2$datetime==x])),datetime=sam.time)
ph2=ggplot(data=dat2,aes(x=volume2))+
  geom_histogram(aes(y=..density..,fill=pct_chg2),
                 alpha=0.6,position="dodge",bins=20)+
  geom_text(data=tex1,aes(x=x,y=y,label=paste("偏度=",round(sk,2),sep="
    ")))+
  geom_text(data=tex2,aes(x=x,y=y,label=paste("峰度=",round(ku,2),sep="
    ")))+
  scale_fill_manual(values=c("blue","red"))+
  labs(x="box-cox变换后的交易量",y="密度",fill="涨跌情况")+
  theme_bw()+
  theme(panel.grid=element_blank(),
        legend.position="right",
        legend.key.width=unit(1.1,"cm"),
        legend.key=element_rect(colour='white',
                                fill='white',
                                size=1))+
  facet_wrap(~datetime,ncol=3)
print(ph2)
ggsave("成交量直方图2.pdf",width=8,height=5)

#等方差检验
fun1=function(x){
  return(c(time=x[1],state=x[2],sigma=round(sd(dat2$volume2[dat2$
    datetime==x[1] & dat2$pct_chg2==x[2]]),2)))
}
tmp=data.frame(t(apply(cbind(rep(as.character(sam.time),each=2),c("涨",
    "跌")),1,fun1)))
```

```
83 write.table(tmp,file="方差检验.txt",sep="&",row.names=F)
84
85 for(i in sam.time){
86   d=dat2[dat2$datetime==i,]
87   print(summary(aov(volume2~pct_chg2,data=d)))
88 }
```