2015 南开大学小学期数学建模竞赛

承　诺　书

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛章程和参赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛章程和参赛规则，以保证竞赛的公正、公平性。如有违反竞赛章程和参赛规则的行为，我们将受到严肃处理。

我们参赛选择的题号是（从 A/B/C 中选择一项填写）：B

所属学院（请填写完整的全名）：数学学院 商学院 经济学院

参赛队员及学号 (打印并签名) ：1. 姚智元　1310100

2. 于　淼　1313219

3. 王　睿　1313274

指导教师或指导教师组负责人　(打印并签名)：

（论文纸质版与电子版中的以上信息必须一致，只是电子版中无需签名。以上内容请仔细核对，提交后将不再允许做任何修改。如填写错误，论文可能被取消评奖资格。）

日期：2015　年 7 月 19　日

评阅编号（由组委会评阅前进行编号）：

股票首破时分布问题及其实践意义研究

摘要：本文选取不同年份牛市、熊市下的指数、蓝筹股和中小盘股构造标的数据集，计算出各支股票在各交易日一定时段的均值、方差及其衍生数据，从中探究出一定的变化规律。利用 mathematica 软件，求出各支股票超出均值上下 3 倍标准差时完成的波动周期数——首破时，并由此推导出其相关统计特征及分布情况。结果发现，股价与均值的交点多发生于突破前，一旦突破 3 倍标准差的界限，便很难再回到均值水平。另外，突破界限前的周期数一般很少。基于以上发现，本文提出了若干建议以供操盘手在实际操作中应用。

关键词：股票；首破时；分布；标准差

一、问题重述

从数学的角度看，如果从 10：00 之后股价在$m\pm 3\sigma_0$ 水平带中完成一个周期的概率记为p，如果突破之前共计完成了ξ个周期。那么就称ξ为首破时（这与概率论中首次打开锁的模型中直到第 k+1 次才打开的概率$b_k=(1-p)^k p$中的k类似，但也有不同，开锁模型最终必然打开，但股价在水平盘整阶段可能整天都不突破）。显然，我们能够通过大样本，对这个随机变量ξ进行建模分析。请回答以下问题：

a) 通过选取尽可能多的年代中熊市、牛市情况下的指数、大蓝筹股、中小盘股的分时数据，构造一个标的（benchmark）数据集（越多越好）。并用统计语言说明选取的样本数据的无偏性。

b) 基于 benchmark 集，通过计算前半小时的σ_0和全天的σ，求出比值$\left\lceil \frac{\sigma}{\sigma_0} \right\rceil$（取整）之后的变化范围和对应各值的概率。

c) 求首破时ξ（就是$b_k=(1-p)^k p$ 中的 k）的取值空间、概率列(分布函数)、均值、方差。

d) 针对上述三小题的结果，给出一份可供操盘手操作的报告。

二、问题分析

（一）问题(a)分析

1、标的（benchmark）数据集的构造

由题目要求出发，我们从同花顺网站收集了相应指数、股票的分时数据，它们分别为上证指数，长安汽车、招商银行、中国卫星三家蓝筹股以及汉麻产业、鸿达兴业、中航机电三家中小盘股。由于数据可得性的限制，上证指数样本开始时间为 2001 年 12 月，其他个股则由其上市时间为样本选取的起始时间。

由于所采用数据集过于庞大，此处不将其详细列出，仅描述数据的选取时间，具体的标的数据集见附件“标的数据集”。样本的构成情况如表（1）所示：

表（1）样本构成情况

样本名称	时间范围及市场形势
上证指数	2001-2003（熊） 2004-2005（熊） 2006-2007（牛） 2014-2015（牛）
蓝筹股-长安汽车	2002-2003（熊） 2004-2005（熊） 2006-2007（牛） 2014-2015（牛）
蓝筹股-招商银行	2002-2003（熊） 2004-2005（熊） 2006-2007（牛） 2014-2015（牛）
蓝筹股-中国卫星	2002-2003（熊） 2004-2005（熊） 2006-2007（牛） 2014-2015（牛）
中小盘股-汉麻产业	2004-2005（熊） 2006-2007（牛） 2008-2009（熊） 2014-2015（牛）
中小盘股-鸿达兴业	2004-2005（熊） 2006-2007（牛） 2008-2009（熊） 2014-2015（牛）
中小盘股-中航机电	2004-2005（熊） 2006-2007（牛） 2008-2009（熊） 2014-2015（牛）

注：表中给出的是时间段，具体样本仅包含对应时间段内的若干交易日，未能涵盖每个交易日的情况，详见附件。数据来源：同花顺网

2、样本数据无偏性的证明

无偏性，指的是样本估计量的统计特征（平均值、变异数、方差等）与其母体真值一致。

样本均值$\bar{Y}$是总体均值 μ 的一个无偏统计量，这是因为不管其背后的总体分布是怎样的，利用期望值的相关性质，总有：

$$E(\bar{Y})=E[(1/n)\sum_{i=1}^{n} Y_i]=(1/n)E(\sum_{i=1}^{n} Y_i)=(1/n)[\sum_{i=1}^{n} E(Y_i)]$$

$$=(1/n)(\sum_{i=1}^{n} \mu)=(1/n)(n\ \mu)=\mu$$

故所选取的标的数据集具有无偏性。

（二）问题(b)分析

这一问只需要计算数值即可，利用 excel 的函数功能，很容易求出样本中各交易日前半小时（即 9:30-10:00 这一时段）股价的标准差σ_0和全天的标准差σ，

从而计算比值$\left\lceil \frac{\sigma}{\sigma_0} \right\rceil$（向上取整）的值，进而得到这一整数值的取值范围及其对应概率。部分计算结果如表（3）所示，完整结果见附件中的 excel 文件“第二题结果”。

表（3）

类别	个股名称	时间	σ_0	σ	比值取整	频率（保留两位）
上证指数	—	20011214	2.878097	5.147087	2	0.39
上证指数	—	20020611	1.155895313	4.62392716	5	0.04
上证指数	—	20030324	0.580634717	1.752096274	4	0.08
上证指数	—	20041210	0.655068905	6.417151588	10	0.02
上证指数	—	20050413	4.011609704	3.318292141	1	0.16
上证指数	—	20070302	7.415223829	15.71032891	3	0.28
上证指数	—	20141202	5.54724	31.4607	6	0.02
上证指数	—	20150603	10.5476	29.8028	3	0.28
蓝筹股	长安汽车	20030630	0.126002	0.118505	1	0.16
蓝筹股	长安汽车	20040722	0.0205991	0.172934	9	0.01
蓝筹股	招商银行	20141121	0.0085293	0.0829886	10	0.02
蓝筹股	招商银行	20150324	0.0434124	0.0418088	1	0.16
蓝筹股	中国卫星	20020913	0.026423745	0.021821197	1	0.16
蓝筹股	中国卫星	20021211	0.004082483	0.039470175	10	0.02
中小盘股	汉麻产业	20050120	0.016089007	0.088047107	6	0.02
中小盘股	汉麻产业	20081030	0.096834613	0.080347955	1	0.16
中小盘股	鸿达兴业	20040809	0.043085232	0.292684334	7	0.01
中小盘股	鸿达兴业	20080909	0.066151342	0.039738946	1	0.16
中小盘股	中航机电	20060918	0.018393558	0.213246125	12	0.01
中小盘股	中航机电	20150424	0.818512526	0.304458815	1	0.16

数据来源：同花顺网

在选取要列出的计算结果时，我们特意包含了各个指数及个指计算结果的最大值和最小值，以便直观显示取整后的变化范围。在我们选取的样本中，各个主体所求的变化范围如表（4）所示：

表（4）

类别	个股名称	$\left\lceil \frac{\sigma}{\sigma_0} \right\rceil$取值范围
上证指数	—	[1,10]
蓝筹股	长安汽车	[1,9]

蓝筹股	招商银行	[1,10]
蓝筹股	中国卫星	[1,10]
中小盘股	汉麻产业	[1,6]
中小盘股	鸿达兴业	[1,7]
中小盘股	中航机电	[1,12]

数据来源：同花顺网

（三）问题(c)分析

按照题述要求，当股价超过 $m\pm3\sigma_0$ 水平带时，已完成的周期数 ξ 为首破时，而其完成一个周期的概率为 p。因为股价的波动往往极其复杂，难以通过图像直观算出首破时，故需借助特定的数学软件来实现首破时的求解。Mathematica 是一款功能强大、应用广泛的科学计算软件，可帮助我们处理数据集数据并得出结果。采用 Mathematica 9.0 版，我们将构造的数据集导入软件，编程实现对于首破时的求解，具体程序代码见附件，计算原理如下所示：

若$\forall$i， m-3σ_0<Y_i<m+3σ_0，则ξ=0；

否则，当 m-3σ_0<Y_x<m+3σ_0 且 $Y_{x+1}\leq m-3\sigma_0$ 或$Y_{x+1}\geq m+3\sigma_0$ 时，

求当 i≤x 时，满足$(Y_i-m)(Y_{i+1}-m)<0$ 的 i 的个数 j，

最终$\xi=\lfloor j\rfloor$。

其中，m 表示前半小时股价的均值，σ_0 为前半小时股价的标准差。

上述公式通过求股价超出水平带前与平均值 m 的交点个数来得出周期个数，在处理时还要将交点数除以 2 并向下求整，从而排除掉不完整的周期，得出题目中定义的首破时 ξ。

用 excel 很容易获得各类首破时的数值在样本总体中出现的概率，以概率为因变量，首破时为自变量，利用 matlab 进行函数拟合，可以得到首破时的概率函数，以此为模型估计总体的统计特征。

三、模型假设

根据题设条件及模型简化的需要，我们做出以下假设：

1、假设首破时的分布服从几何分布

2、由于取样时间间隔最短为 5 秒，故假设一个周期最短时长为 10 秒，相应

的首破时也有最大值限制，这一数值为 1440。

3、假设取样数据均匀，可以前 1/8 时间段表示前半小时。

四、符号说明

符号	意义
m	前半小时的股价均值
σ_0	前半小时股价的标准差
k/ξ	首破时
b_k	首破时为 k 的概率
p	股价在题设水平带内完成一个周期的概率

五、模型的构建与求解

如题所述，首破时的分布服从几何分布，故我们的模型设定为：

$$b_k = (1-p)^k p$$

其中各符号的表示意义已于前文介绍，k 和 b_k 的获得方法也已在问题分析中论述，此处不再赘述。以此模型为基础，利用 matlab 软件对整理过的数据进行拟合，得到指数、蓝筹股、中小盘股票各自的概率密度函数。 均值、方差及取值区间则可通过简单的计算得出。计算及拟合结果如下表所示：

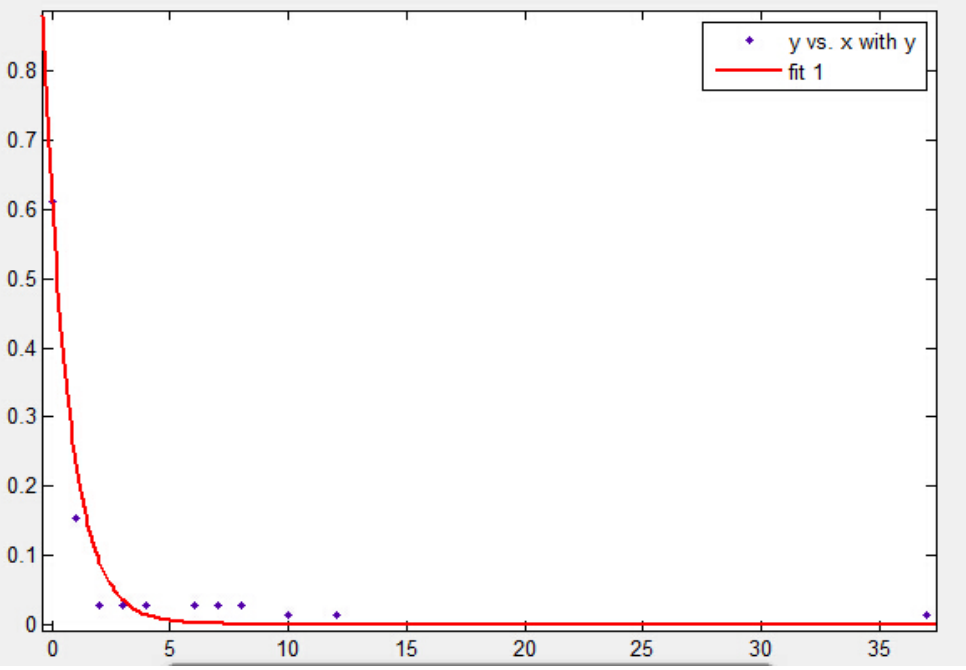
表（5）

	p	EX	DX	取值空间
上证	0.3831	2.610285	4.203301	[0,1,2,3,4,6,7,8,10,12,37]
中小盘	0.4939	2.024701	2.074714	[0,1,2,3,4,6,7,8,10,12,17]
蓝筹股	0.4552	2.196837	2.629254	[0,1,2,4,5,6,7,8,11,12,14,15,18,20,21,22]

数据来源：同花顺网

图（1）—图（3）给出了拟合的图像及分布函数：

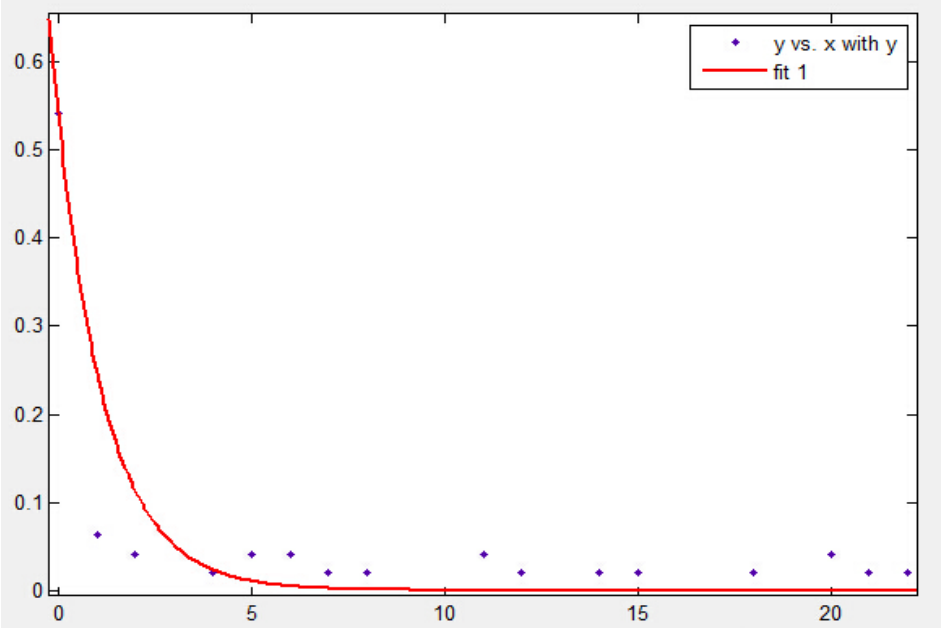
图（1）上证指数首破时分布情况图



概率密度：b_k=0.62^{k}0.38

分布函数：$F(x)=1-0.62^x$

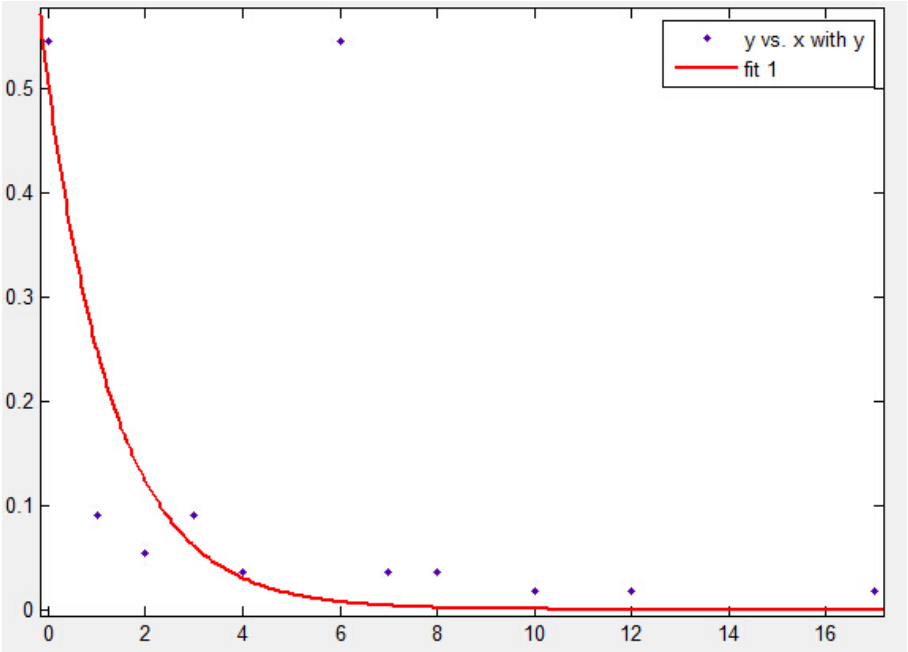
图（2）蓝筹股首破时分布情况图



概率密度：b_k=0.55^{k}0.45

分布函数：$F(x)=1-0.55^x$

图（3）中小盘首破时分布情况图



概率密度：$b_k=0.51^k0.49$

分布函数：$F(x)=1-0.51^x$

注：以上分布函数均采取复旦大学出版教材中的定义，令 $F(x)=P\{k<x\}$。

依据分布函数算出的具体分布列如表（6）所示：

表（6）

上证指数		中小盘		蓝筹股	
x	概率	x	概率	x	概率
0	0	0	0	0	0.000000
1	0.6169	1	0.5061	1	0.544800
2	0.853234	2	0.756063	2	0.792793
3	0.943774	3	0.879519	4	0.957065
4	0.97846	4	0.940495	5	0.980456
6	0.996839	6	0.985484	6	0.991104
7	0.998789	7	0.992831	7	0.995950
8	0.999536	8	0.996459	8	0.998157

10	0.999932	10	0.999136	11	0.999826
12	0.99999	12	0.999789	12	0.999921
37	1	17	0.999994	14	0.999984
				15	0.999993
				18	0.999999
				20	1.000000
				21	1.000000
				22	1.000000

六、结果分析

在问题(b)中，前半小时股价的标准差与全天股价标准差的比值向上取整后的值（即$\left\lceil \frac{\sigma}{\sigma_0} \right\rceil$）等于1的概率约为0.16，这意味着原比值大于1的概率约为0.84，也就是说，全天的股价波动在大多数情况下是会大于前半小时的波动的，因而对于操盘者来说，上个交易日的股价不及交易日当天前半小时股价稳定，宜采用前半小时的均值作为后续分析判断的基准而较少采用历史交易日的股价均值。

在问题(c)中，由于我们通过交点个数来计算首破时，一个新的发现产生于此：在股价突破$m\pm 3\sigma_0$水平带后，便很少再有机会回落或上涨到与均值 m 相交。这意味着，一旦股价突破$m\pm 3\sigma_0$这个界限，就有很强势的继续上涨或下跌的趋势，操盘手应抓住这一时机进行对应的买入或卖空的操作。

另外，我们还发现，在各类股票中，首破时等于 0 的概率都大于 0.5，而其在 3 以内的概率均达到了 0.9 以上。相比于股价复杂的波动，严格符合题设要求的周期极其微小，这意味着股价不大可能会想“白噪音”那样无规变动，而大多是有一个较长时间的上涨或下跌的过程，仅在特定时期出现由涨转跌或由跌转涨的拐点。对于操盘者来说，关键在于拐点的把握，在股价刚开始上升时进行买入操作，在股价刚开始下降时进行卖空操作是有一定的事实依据为支撑的。

七、附录

（一）使用软件

Excel 2013、Mathematica 9.0、Matlab 7.11.0

（二）Mathematica 代码及输出数据说明

```
originaldata =
m = Mean[Table[
    originaldata[[i, 2]], {i, 1, IntegerPart[Length[originaldata]/8]}]]
sigma = StandardDeviation[
  Table[originaldata[[i, 2]], {i, 1,
    IntegerPart[Length[originaldata]/8]}]]
If[(Max[Table[
      originaldata[[i, 2]], {i, 1, Length[originaldata]}]] > (m +
      3*sigma)) || (Min[
     Table[originaldata[[i, 2]], {i, 1,
       Length[originaldata]}]] < (m - 3*sigma)), true, false]
For[q = IntegerPart[Length[originaldata]/8],
 q < Length[originaldata], q++,
 If[(originaldata[[q, 2]] > (m + 3*sigma)) || (originaldata[[q,
       2]] < (m - 3*sigma)), Break[], test]]
Print[q]
Length[originaldata]
n = 0
For[j = IntegerPart[Length[originaldata]/8], j < q, j++,
 If[(originaldata[[j, 2]] - m)*(originaldata[[j + 1, 2]] - m) < 0, n++,
   n = n]]
Print[n]
If[(Max[Table[
      originaldata[[i, 2]], {i, 1, Length[originaldata]}]] > (m +
      3*sigma)) || (Min[
     Table[originaldata[[i, 2]], {i, 1,
```

```
        Length[originaldata]}]] < (m - 3*sigma)),
  result=Insert[result, n, Length[result] + 1],
  result = Insert[result, false, Length[result] + 1]]
n = 0
For[j = IntegerPart[Length[originaldata]/8],
  j < Length[originaldata], j++,
  If[(originaldata[[j, 2]] - m)*(originaldata[[j + 1, 2]] - m) < 0, n++,
     n = n]]
Print[n]
```

返回：

均值 m

标准差 sigma

是否突破

突破条数

总条数

n=0

10 点以后到首破交点个数

数组 result 储存交点个数

n=0

10 点至 3 点的交点个数