数据挖掘导论作业一

学号：16340294

姓名：张星

**[Exercise 1]**

本题使用投点法，我使用matlab生成均匀随机分布的点，然后计算出落在1/4圆内的点数目与正方形内的数目之比，就可以得出pi的估计值。是否落在圆内使用距离判断。

代码：

length**=**1**;**

repeat **=** 100**;**

N**=[**20**,** 50**,** 100**,** 200**,** 300**,** 500**,** 1000**,** 5000**];**

mean **=** zeros**(**1**,**8**);**

variance **=** zeros**(**1**,**8**);**

**for** k **=** 1**:**8

**for** i **=** 1**:**repeat

circleTimes **=** 0**;**

mat**=**unifrnd**(**0**,**length**,**N**(**k**),**2**);**%n个长length的正方形中的随机点的坐标，均匀分布

temp **=** size**(**mat**,**1**);**

**for** j **=** 1**:**temp

**if** power**(**mat**(**j**,**1**),**2**)+**power**(**mat**(**j**,**2**),**2**)<=**1

circleTimes **=** circleTimes**+**1**;**

**end**

**end**

PI **=** **(**double**(**circleTimes**)/**temp**)\***4**;**

mean**(**k**)** **=** mean**(**k**)+**PI**;**

variance**(**k**)** **=** variance**(**k**)+**PI**\***PI**;**

**end**

mean**(**k**)** **=** mean**(**k**)/**100**;**

%variance = E(x^2) - E(x)^2

variance**(**k**)** **=** variance**(**k**)/**100 **-** mean**(**k**)\***mean**(**k**);**

%plot(mat(:,1),mat(:,2),'.');

**end**

x**=**0**:**714**:**5000**;**%x轴上的数据，第一个值代表数据开始，第二个值代表间隔，第三个值代表终止

plot**(**x**,**mean**,**'-\*b'**,**x**,**variance**,**'-or'**);** %线性，颜色，标记

axis**([**0**,**5000**,**0**,**4**])** %确定x轴与y轴框图大小

set**(**gca**,**'XTick'**,(**0**:**1000**:**5000**));**

set**(**gca**,**'YTick'**,(**0**:**1**:**4**));**

legend**(**'Mean'**,**'Variance'**);**%右上角标注

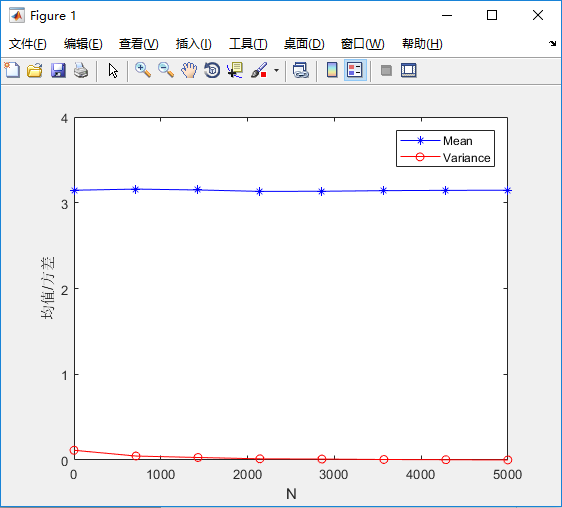
xlabel**(**'N'**)** %x轴坐标描述

ylabel**(**'均值/方差'**)** %y轴坐标描述

结果：每个N重复100次。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| N | 20 | 50 | 100 | 200 | 300 | 500 | 1000 | 5000 |
| 均值 | 3.146 | 3.159 | 3.149 | 3.133 | 3.134 | 3.141 | 3.144 | 3.146 |
| 方差 | 0.114 | 0.046 | 0.029 | 0.014 | 0.011 | 0.006 | 0.003 | 0.001 |

折线图：



**[Exercise 2]**

本题我使用了与均匀随机分布，效果很好，十分接近0.25。另外，使用随机分布与均匀随机分布的数据极其接近，所以我选择更合理的均匀随机分布。

代码：

length**=**1**;**

repeat **=** 100**;**

N**=[**0**,** 5**,** 10**,** 20**,** 30**,** 40**,** 50**,** 60**,** 70**,** 80**,** 100**];**

mean **=** zeros**(**1**,**11**);**

variance **=** zeros**(**1**,**11**);**

**for** k **=** 2**:**11

**for** i **=** 1**:**repeat

integralTimes **=** 0**;**

mat**=**unifrnd**(**0**,**length**,**N**(**k**),**2**);**%n个长length的正方形中的随机点的坐标，均匀分布

temp **=** size**(**mat**,**1**);**

**for** j **=** 1**:**temp

**if** mat**(**j**,**1**)** **<=** power**(**mat**(**j**,**2**),**3**)**

integralTimes **=** integralTimes**+**1**;**

**end**

**end**

result **=** integralTimes**/**temp**;**

mean**(**k**)** **=** mean**(**k**)+**result**;**

variance**(**k**)** **=** variance**(**k**)+**result**\***result**;**

**end**

mean**(**k**)** **=** mean**(**k**)/**repeat**;**

%variance = E(x^2) - E(x)^2

variance**(**k**)** **=** variance**(**k**)/**repeat **-** mean**(**k**)\***mean**(**k**);**

%plot(mat(:,1),mat(:,2),'.');

**end**

x**=**0**:**10**:**100**;**%x轴上的数据，第一个值代表数据开始，第二个值代表间隔，第三个值代表终止

plot**(**x**,**mean**,**'-\*b'**,**x**,**variance**,**'-or'**);** %线性，颜色，标记

axis**([**0**,**100**,**0**,**0.5**])** %确定x轴与y轴框图大小

set**(**gca**,**'XTick'**,(**0**:**10**:**100**));**

set**(**gca**,**'YTick'**,(**0**:**0.1**:**0.5**));**

legend**(**'Mean'**,**'Variance'**);**%右上角标注

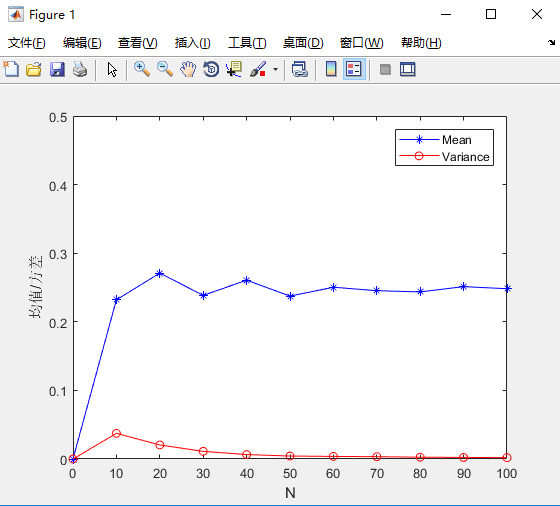
xlabel**(**'N'**)** %x轴坐标描述

ylabel**(**'均值/方差'**)** %y轴坐标描述

结果：每个N重复100次。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| N | 5 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 100 |
| 均值 | 0.232 | 0.271 | 0.239 | 0.261 | 0.238 | 0.250 | 0.246 | 0.244 | 0.251 | 0.249 |
| 方差 | 0.037 | 0.021 | 0.011 | 0.007 | 0.004 | 0.004 | 0.003 | 0.003 | 0.002 | 0.002 |

折线图：



**[Exercise 3]**

本题我依旧使用了均匀随机分布，方法上使用了平均值期望，结果数据很大，但是由于本题的函数我求不出原函数，导致无法验证。

代码：

length **=** 2**;**

repeat **=** 100**;**

N **=** **[**0**,** 10**,** 20**,** 30**,** 40**,** 50**,** 60**,** 70**,** 80**,** 100**,** 200**,** 500**];**

mean **=** zeros**(**1**,**12**);**

variance **=** zeros**(**1**,**12**);**

**for** p **=** 2**:**11

**for** i **=** 1**:**repeat

SumX **=** 0**;**

**for** j **=** 1**:**N**(**p**)**

mat**=**unifrnd**(**0**,**length**,**N**(**p**),**2**);**%n个长length的正方形中的随机点的坐标，均匀分布

mat**(:,**1**)** **=** mat**(:,**1**)-**1**;**%y坐标减1

mat**(:,**2**)** **=** mat**(:,**2**)+**2**;**%x坐标加2

SumY **=** 0**;**

**for** k **=** 1**:**N**(**p**)**

x **=** mat**(**j**,**2**);**

y **=** mat**(**k**,**1**);**

SumY **=** SumY **+** **(**power**(**y**,**2**)\***exp**(-**power**(**y**,**2**))+**power**(**x**,**4**)\***exp**(-**power**(**x**,**2**)))/(**x**\***exp**(-**power**(**x**,**2**)));**

**end**

SumY **=** SumY**/**N**(**p**);**

SumX **=** SumX**+(**4**-**2**)\***SumY**;**

**end**

result **=** SumX**/**N**(**p**);**

mean**(**p**)** **=** mean**(**p**)+**result**;**

variance**(**p**)** **=** variance**(**p**)+**result**\***result**;**

**end**

mean**(**p**)** **=** mean**(**p**)/**repeat**;**

%variance = E(x^2) - E(x)^2

variance**(**p**)** **=** variance**(**p**)/**repeat **-** mean**(**p**)\***mean**(**p**);**

%plot(mat(:,1),mat(:,2),'.');

**end**

x**=**0**:**42**:**500**;**%x轴上的数据，第一个值代表数据开始，第二个值代表间隔，第三个值代表终止

plot**(**x**,**mean**,**'-\*b'**);** %线性，颜色，标记

axis**([**0**,**500**,**0**,**100000**])** %确定x轴与y轴框图大小

set**(**gca**,**'XTick'**,(**0**:**42**:**500**));**

set**(**gca**,**'YTick'**,(**0**:**10000**:**100000**));**

legend**(**'Mean'**);**%右上角标注

xlabel**(**'N'**)** %x轴坐标描述

ylabel**(**'均值'**)** %y轴坐标描述

结果：每个N重复100次。

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| N | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 100 | 200 | 500 |
| 均值 | 5.784e+04 | 5.016e+04 | 5.257e+04 | 6.294e+04 | 5.552e+04 | 5.782e+04 | 5.621e+04 | 5.686e+04 | 5.795e+04 | 5.698e+04 | 5.799e+04 |
| 方差 | 2.032e+09 | 7.336e+08 | 5.228e+08 | 4.843e+08 | 4.120e+08 | 2.843e+08 | 2.532e+08 | 2.377e+08 | 1.652e+08 | 8.672e+07 | 4.070e+07 |

折线图：

