**实验二 数据预处理常用方法与应用**

【实验目的】

1 掌握随机抽样的常用方法。

2 掌握变量规范化的常用方法。

3 掌握常见距离的定义及计算方法。

4 掌握二元变量相似度、连续变量余弦相似度计算方法。

5 掌握Mahalanobis距离的计算方法。

【实验类型】

设计型

【实验学时】

2学时

【实验环境】

Windows 7以上操作系统

Python3.0以上版本

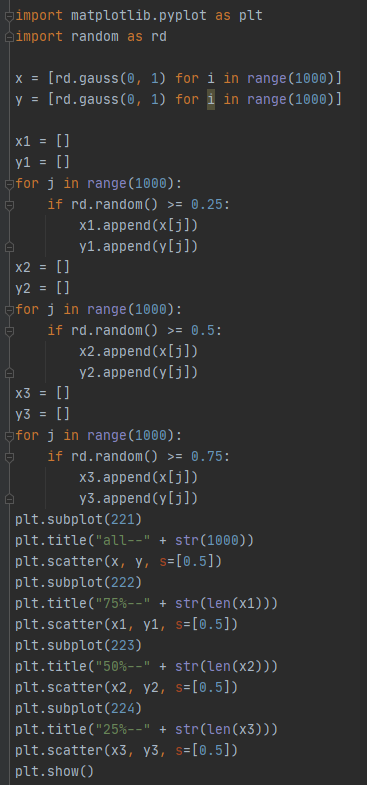
Pycharm开发环境

Spyder开发环境

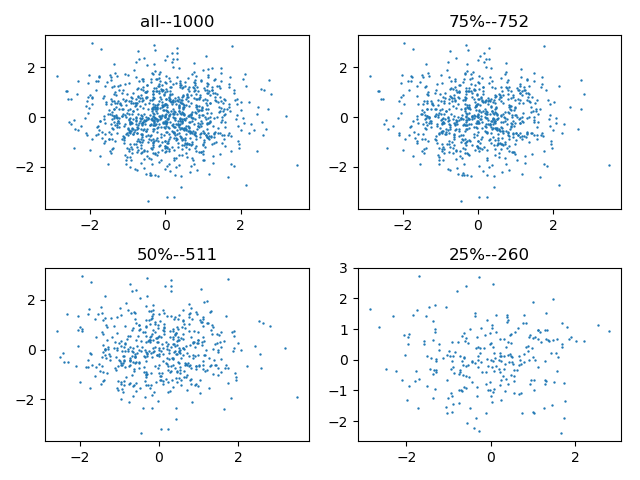
【实验要求】

一、请随机生成1000个二维数据点，从中分别随机抽取约75%，50%，25%的数据点，并绘图展示抽取数据点的情况。

**参考答案：**

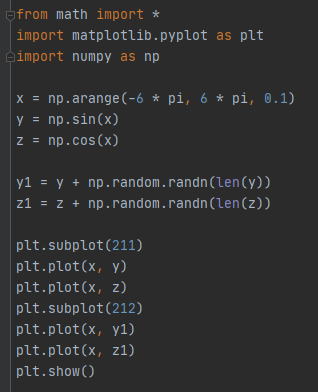


**参考输出结果:**

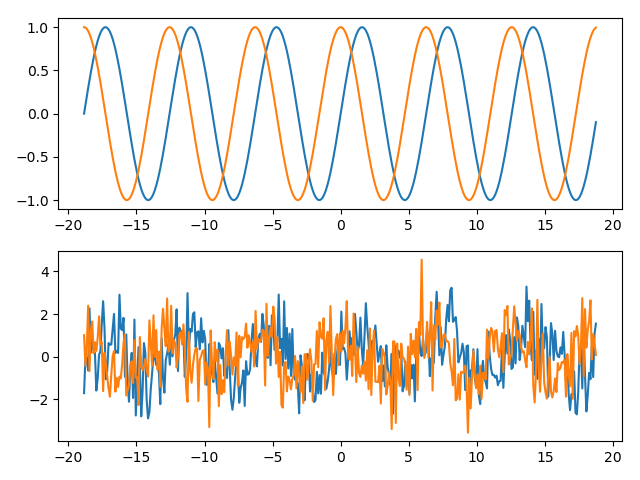


二、请计算生成正弦函数序列和余弦函数序列，并绘图展示。将正弦序列和余弦序列添加随机偏差（噪音）后再次绘图，观察噪音对数据的影响。

**参考答案：**

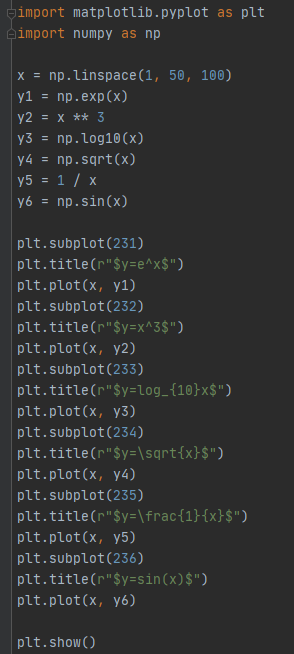


**参考输出结果:**

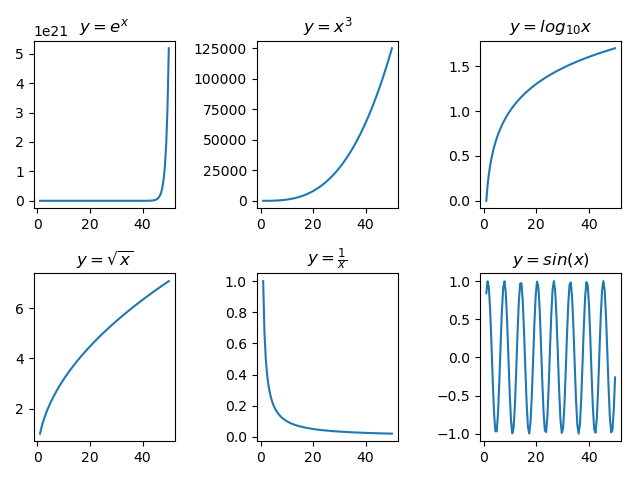


三、验证几类变量变换的效果，包括：指数变换、>1幂变换、对数变换、（0，1）幂变换、<0幂变换、正弦变换，绘图演示各类变换的效果。

**参考答案：**



**参考输出结果：**

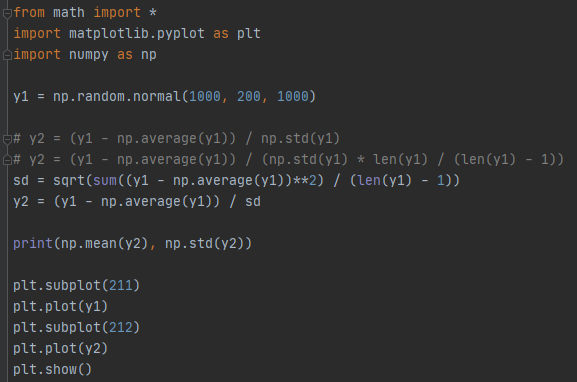


四、为了提出变量量纲对数据分析的影响，经常需要对变量进行一定的变换，其中常用的一种是标准化（normalization）变换（即将数据序列转换为期望值为0和方差为1的序列），转换公式为：

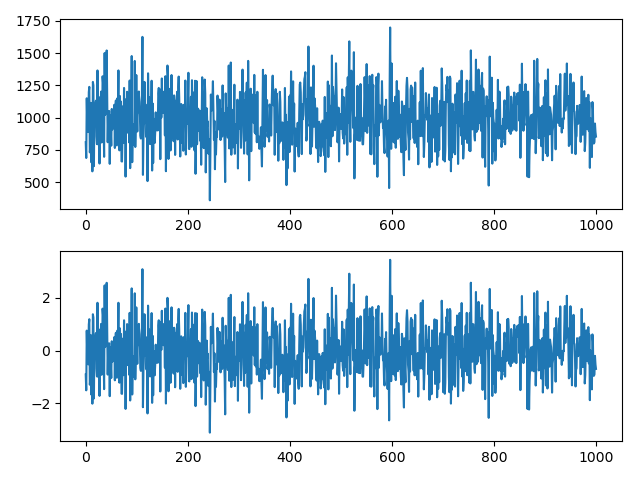
其中为序列的均值，为序列的样本标准差，即

请随机生成一个序列，并生成它的标准化序列，将两个序列通过绘图展示出来。

**参考答案：**

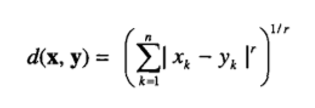


**参考输出结果（其中随机生成的序列为期望值1000，标准差200，长度1000的正态分布序列）：**



五、距离计算

向量间的闵可夫斯基距离的定义如下：



其中当r=1时成为街区距离，为不同分量差绝对值的和

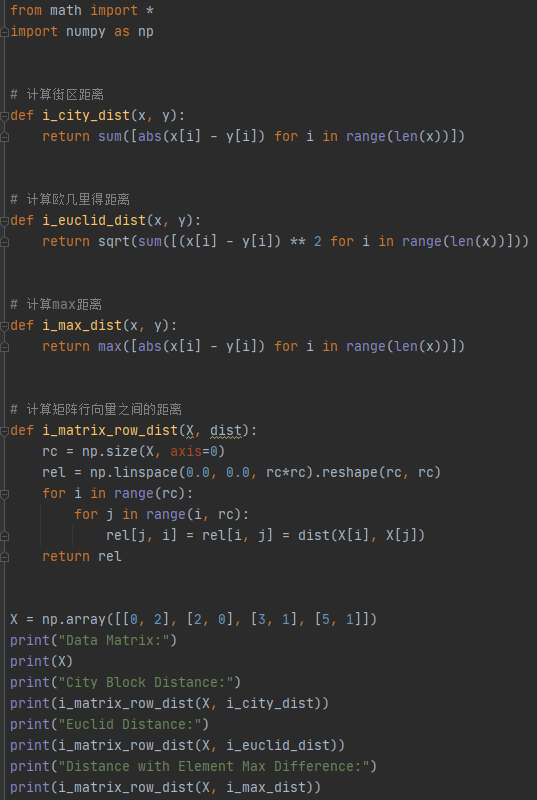
当r=2时即为欧几里得距离

当r=+infinity时，等于所有分量差绝对值中最大的一个

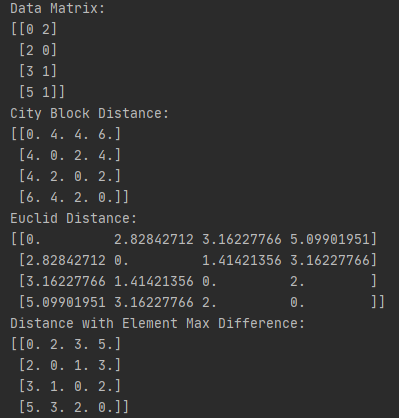
编写函数实现三种距离的计算，并对下面4个二维向量分别计算其三种距离定义的距离矩阵。



**参考答案：**

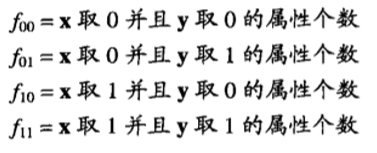


**参考输出结果如下：**

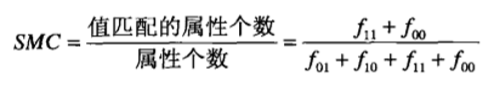


六、二元变量的相似度衡量可以使用简单匹配系数（SMC）或Jaccard系数

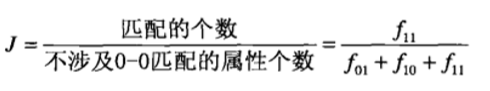
设x和y是两个对象，都由n个二元属性组成，这样两个对象可生成如下四个量（频率）：



简单匹配系数（Simple Matching Coefficient, SMC）定义为：

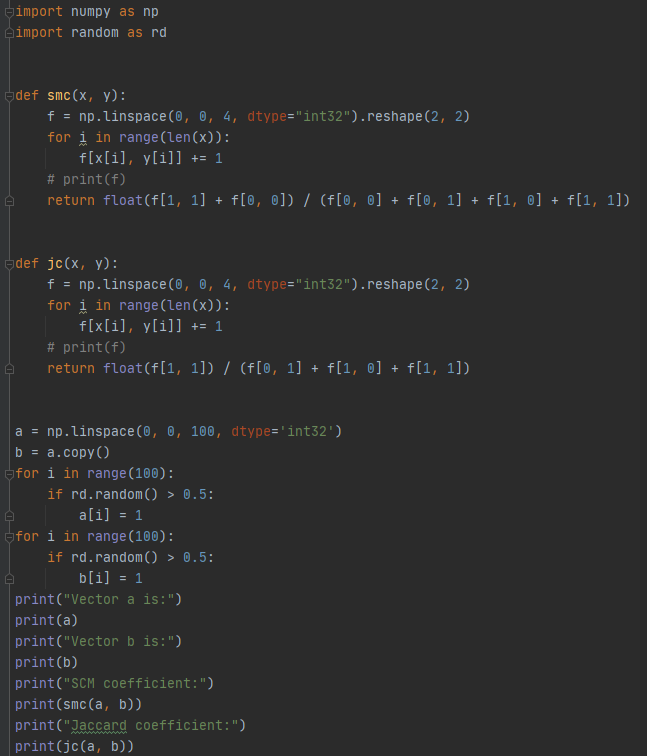


Jaccard系数（Jaccard Coefficient）的定义为：

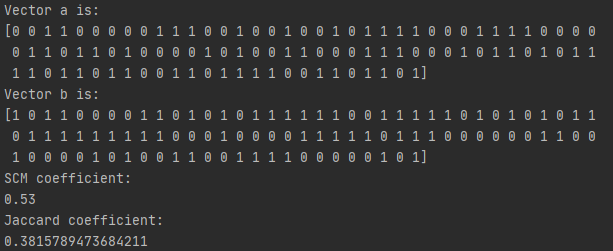


编写两个函数分别实现连个系数的计算，并生成两个随机序列调用函数计算。

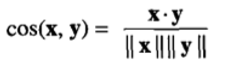
**参考答案：**



**参考输出结果：**



七、余弦相似度是衡量数值向量相似度的一种方法，设x和y是两个n元向量，则余弦相似度定义为





编写函数实现余弦相似度的计算，并调用函数计算下面两个向量的余弦相似度。



**参考答案：**

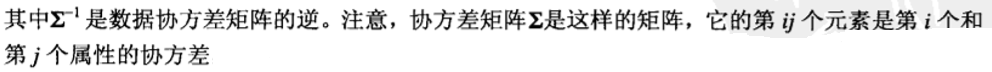


**参考输出结果：**

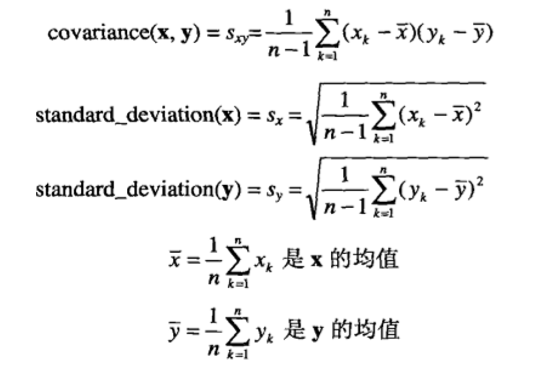


八、从统计学的观点看，计算多维向量的距离时，应该使用mahalanobis距离，其定义为：



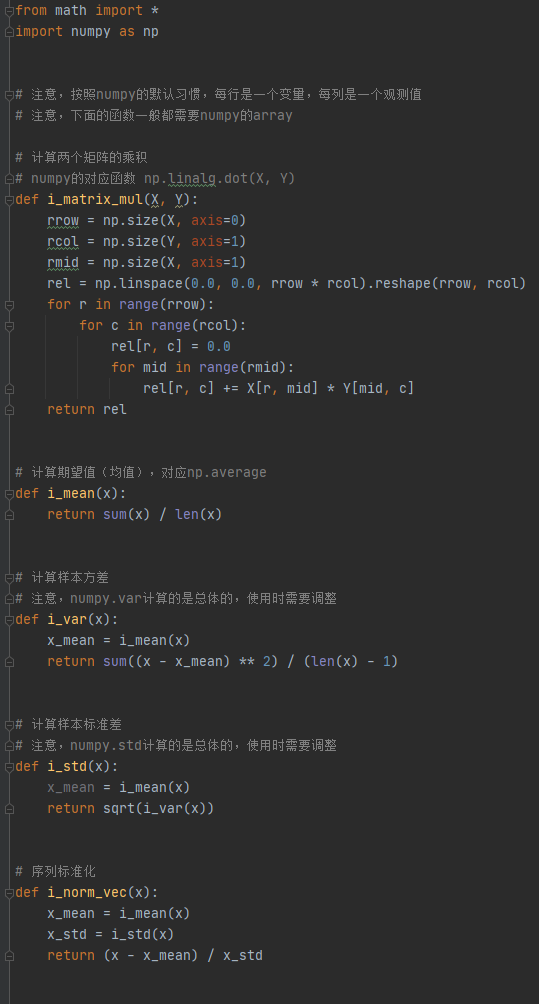


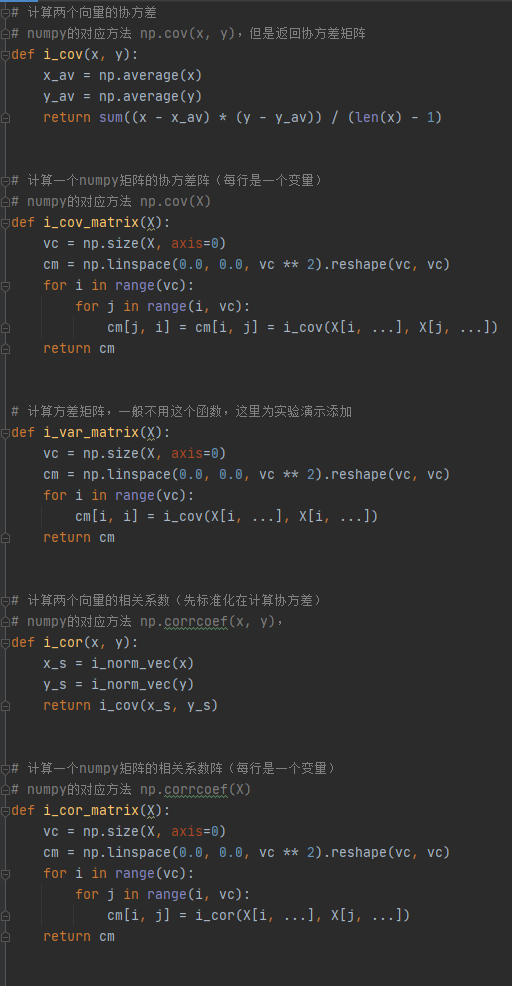
属性间协方差的定义如下所示：

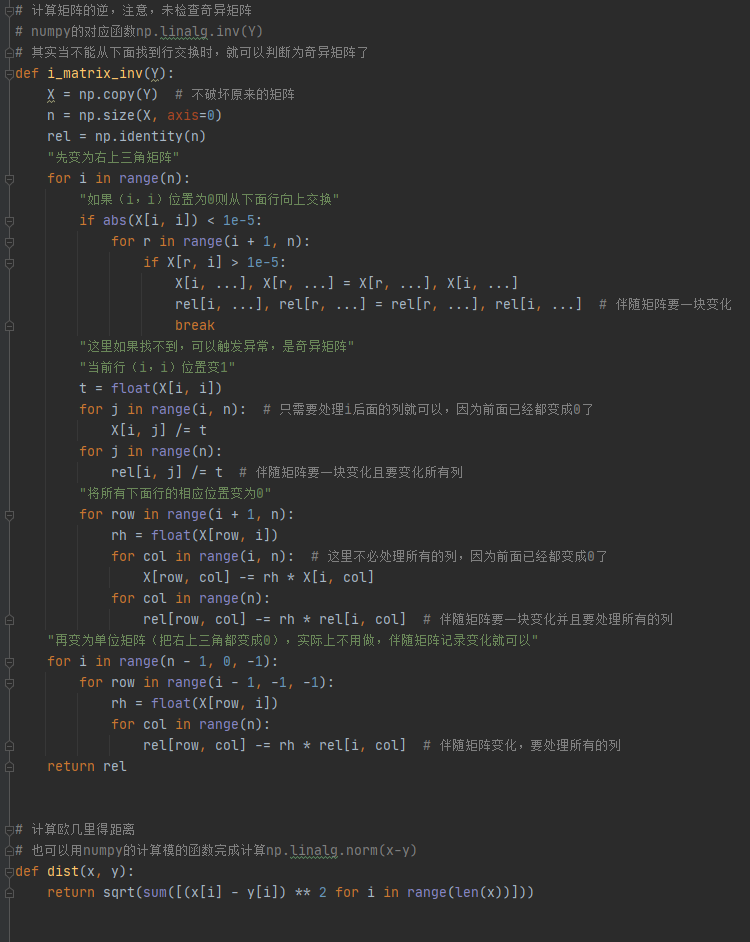


实现欧几里得距离、使用方差逆矩阵调整的距离、使用协方差逆矩阵调整的距离三类距离的计算函数，并构造两个随机向量验证三种距离的不同。

**参考答案：**









**参考输出结果：**

