相关系数

介绍两种常用的相关系数，皮尔逊相关系数和斯皮尔曼等级相关系数。可以用来衡量两个变量之间的相关性的大小，根据数据满足的不同条件，选择不同的相关系数进行分析和计算。

基本概念

1.总体和样本

总体——所要考察对象的全部个体叫做总体.
我们总是希望得到总体数据的一些特征（例如均值方差等）

样本——从总体中所抽取的一部分个体叫做总体的一个样本.

计算这些抽取的样本的统计量来估计总体的统计量：

例如使用**样本均值、样本标准差**来估计**总体的均值（平均水平）和总体的标准差（偏离程度）**。

例子:

我国10年进行一次的人口普查得到的数据就是总体数据。
大家自己在QQ群发问卷叫同学帮忙填写得到的数据就是样本数据。

2.**总体皮尔逊相关系数**

如果两组数据$X:\{X_1,X_2,\cdots,X_n\}$和$Y:\{Y_1,Y_2,\cdots,Y_n\}$是总体数据（例如普查结果），

那么总体均值: $E(X)=\frac{\sum_{i=1}^{n}X_i}{n}$ ， $E(Y)=\frac{\sum_{i=1}^{n}Y_i}{n}$

总体协方差： $\mathrm{Cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-E(X))(Y_i-E(Y))}{n}$

直观理解协方差：如果X、Y变化方向相同，即当X大于（小于）其均值时，Y也大于（小于）其均值，在这两种情况下，乘积为正。如果X、Y的变化方向一直保持相同，则协方差为正；同理，如果X、Y变化方向一直相反，则协方差为负；如果X、Y变化方向之间相互无规律，即分子中有的项为正，有的项为负，那么累加后正负抵消。

注意：协方差的大小和两个变量的量纲有关，因此不适合做比较。

如果两组数据$X:\{X_1,X_2,\cdots,X_n\}$和$Y:\{Y_1,Y_2,\cdots,Y_n\}$是总体数据（例如普查结果），

那么总体均值: $E(X)=\dfrac{\sum_{i=1}^{n}X_i}{n}$ ， $E(Y)=\dfrac{\sum_{i=1}^{n}Y_i}{n}$

总体协方差：$\mathrm{Cov}(X,Y)=\dfrac{\sum_{i=1}^{n}(X_i-E(X))(Y_i-E(Y))}{n}$

总体*Pearson*相关系数：$\rho_{XY}=\dfrac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}=\dfrac{\sum_{i=1}^{n}\frac{(X_i-E(X))}{\sigma_X}\frac{(Y_i-E(Y))}{\sigma_Y}}{n}$

$\sigma_X(sigma\ X)$是X的标准差，$\sigma_X=\sqrt{\dfrac{\sum_{i=1}^{n}(X_i-E(X))^2}{n}}$，$\sigma_Y=\sqrt{\dfrac{\sum_{i=1}^{n}(Y_i-E(Y))^2}{n}}$

可以证明，$|\rho_{XY}|\le 1$，且当$Y=aX+b$时，$\rho_{XY}=\begin{cases}1, & a>0\\ -1, & a<0\end{cases}$

皮尔逊相关系数也可以看成是剔除了两个变量量纲影响，即将X和Y标准化后的协方差。

3.样本皮尔逊相关系数

假设有两组数据$X:\{X_1,X_2,\cdots,X_n\}$和$Y:\{Y_1,Y_2,\cdots,Y_n\}$（一般调查得到的数据均为样本数据）

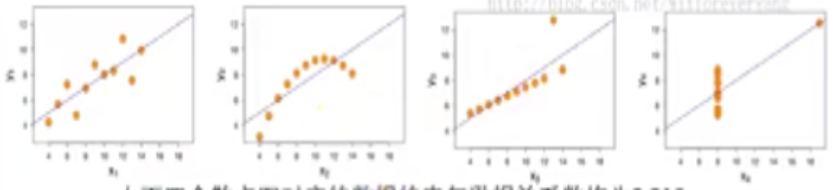
样本均值: $\bar{X}=\dfrac{\sum_{i=1}^{n}X_i}{n}$ ， $\bar{Y}=\dfrac{\sum_{i=1}^{n}Y_i}{n}$

样本协方差：$\mathrm{Cov}(X,Y)=\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$

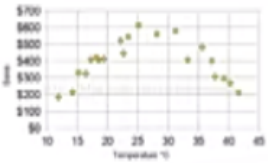
样本*Pearson*相关系数：$r_{XY}=\dfrac{\mathrm{Cov}(X,Y)}{S_XS_Y}$

其中：$S_X(sigma\ X)$是X的样本标准差，$S_X=\sqrt{\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}$，同理$S_Y=\sqrt{\dfrac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}$

这里的相关系数只是用来衡量两个变量线性相关程度的指标；
也就是说，你必须先确认这两个变量是线性相关的，然后这个相关系数才能告诉你他俩相关程度如何。



上面四个散点图对应的数据的皮尔逊相关系数均为0.816



相关系数计算结果为0

（1）非线性相关也会导致线性相关系数很大，例如图2。

（2）离群点对相关系数的影响很大，例如图3，去掉离群点后，相关系数为0.98。

（3）如果两个变量的相关系数很大也不能说明两者相关，例如图4，可能是受到了异常值的影响。

（4）相关系数计算结果为0，只能说不是线性相关，但说不定会有更复杂的相关关系（非线性相关），例如图5。

对相关系数大小的解释

相关性	负	正
无相关性	-0.09 to 0.0	0.0 to 0.09
弱相关性	-0.3 to -0.1	0.1 to 0.3
中相关性	-0.5 to -0.3	0.3 to 0.5
前相关性	-1.0 to 0.5	0.5 to 1.0

上表所定的标准从某种意义上说是武断的和不严格的。
对相关系数的解释是依赖于具体的应用背景和目的的。

事实上，比起相关系数的大小，我们往往更关注的是显著性。
（假设检验）

Matlab中基本统计量的函数（一般用标粗的）：

函数名	功能
min	**数组的最小元素**
mink	计算数组的 k 个最小元素
max	**数组的最大元素**
maxk	计算数组的 k 个最大元素
bounds	最小元素和最大元素
topkrows	按排序顺序的前若干行
mean	**数组的均值**
median	**数组的中位数值**
mode	数组的众数
skewness	**数组的偏度**
kurtosis	**数组的峰度**
std	**标准差**
var	方差

这些函数默认都是按列计算，如果令第二个参数为1，则变为按行计算

4.计算皮尔逊相关系数之前，首先要确定两个变量之间是线性关系。如果不是线性关系，计算出的皮尔逊相关系数无法说明变量之间的相关性。所以要先画出散点图，观察变量的线性关系。

5.皮尔逊相关系数的计算

Test矩阵即为题目中给出的数据。

corrcoef函数： correlation coefficient相关系数
R = corrcoef(A)
返回 A 的相关系数的矩阵，其中 A 的列表示随机变量（指标），行表示观测值（样本）。
R = corrcoef(A,B)
返回两个随机变量 A 和 B （两个向量）之间的系数。

我们要计算体测的六个指标之间的相关系数，只需要使用下面这个语句：
R = corrcoef(Test);

```
R =

    1.0000    0.0665   -0.2177   -0.1920    0.0440    0.0951
    0.0665    1.0000    0.0954    0.0685    0.0279   -0.0161
   -0.2177    0.0954    1.0000    0.2898    0.0248   -0.0749
   -0.1920    0.0685    0.2898    1.0000   -0.0587   -0.0019
    0.0440    0.0279    0.0248   -0.0587    1.0000   -0.0174
    0.0951   -0.0161   -0.0749   -0.0019   -0.0174    1.0000
```

6.首先在matlab中的变量区新建变量，再将excel中的数据复制到变量中，保存到相应的文件夹中。然后在matlab中load此mat文件即可。

7.对相关系数表进行美化

将结果R复制到excel中，选中数据，点击条件格式，选择色阶。