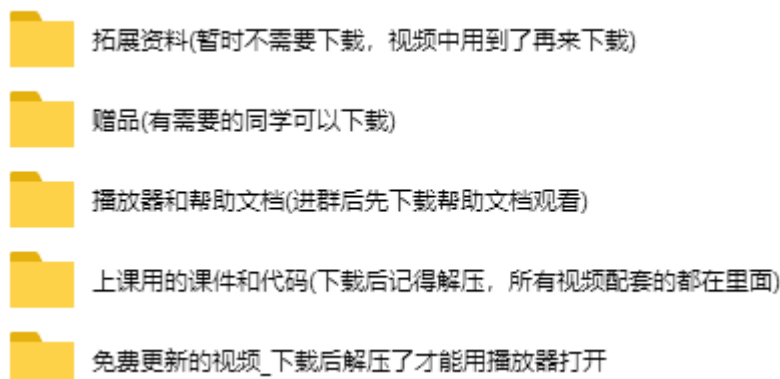


# 第五讲:相关系数

本讲我们将介绍两种最为常用的相关系数: 皮尔逊pearson相关系数和斯皮尔曼spearman等级相关系数。它们可用来衡量两个变量之间的相关性的<sup>大小</sup>, 根据数据满足的不同条件, 我们要选择不同的相关系数进行计算和分析 (建模论文中最容易用错的方法)。

## 温馨提示

- (1) 视频中提到的附件可在**售后群的群文件**中下载。  
包括讲义、代码、我视频中推荐的资料等。



(2) 关注我的**微信公众号《数学建模学习交流》**，后台发送“**软件**”两个字，可获得常见的建模软件下载方法；发送“**数据**”两个字，可获得建模数据的获取方法；发送“**画图**”两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

(3) **购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送“**买**”这个字即可进入店铺进行购买。

(4) 视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

## 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。

## 总体皮尔逊Pearson相关系数

回顾《概率论与数理统计》中的数理统计部分:

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

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

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

**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。

# 总体皮尔逊Pearson相关系数

回顾《概率论与数理统计》中的数理统计部分:

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

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

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

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

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

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

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。

注: 为什么绝对值小于1的证明见本节拓展资料: 文件A

# 样本皮尔逊Pearson相关系数

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

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

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

$$\text{样本Pearson相关系数: } r_{XY} = \frac{\text{Cov}(X, Y)}{S_X S_Y}$$

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

注: 为什么样本方差分母是n-1见本节拓展资料: 文件B

## 相关性可视化

通过绘制散点图可以很容易地判定两个数据对象 $x$ 和 $y$ 之间的相关性。

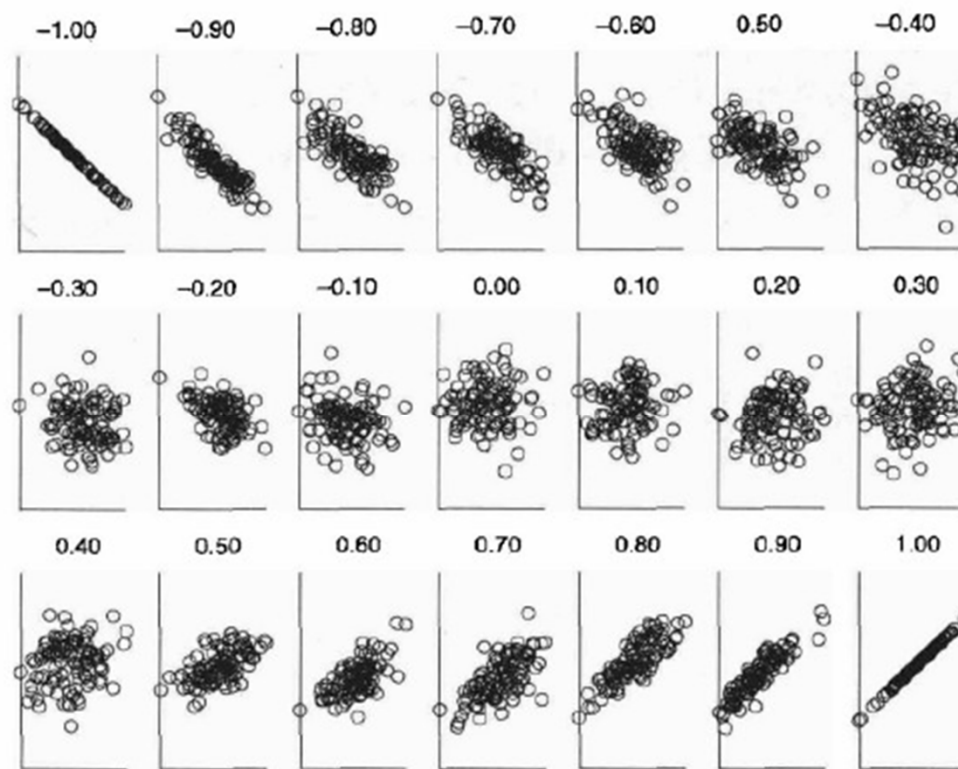
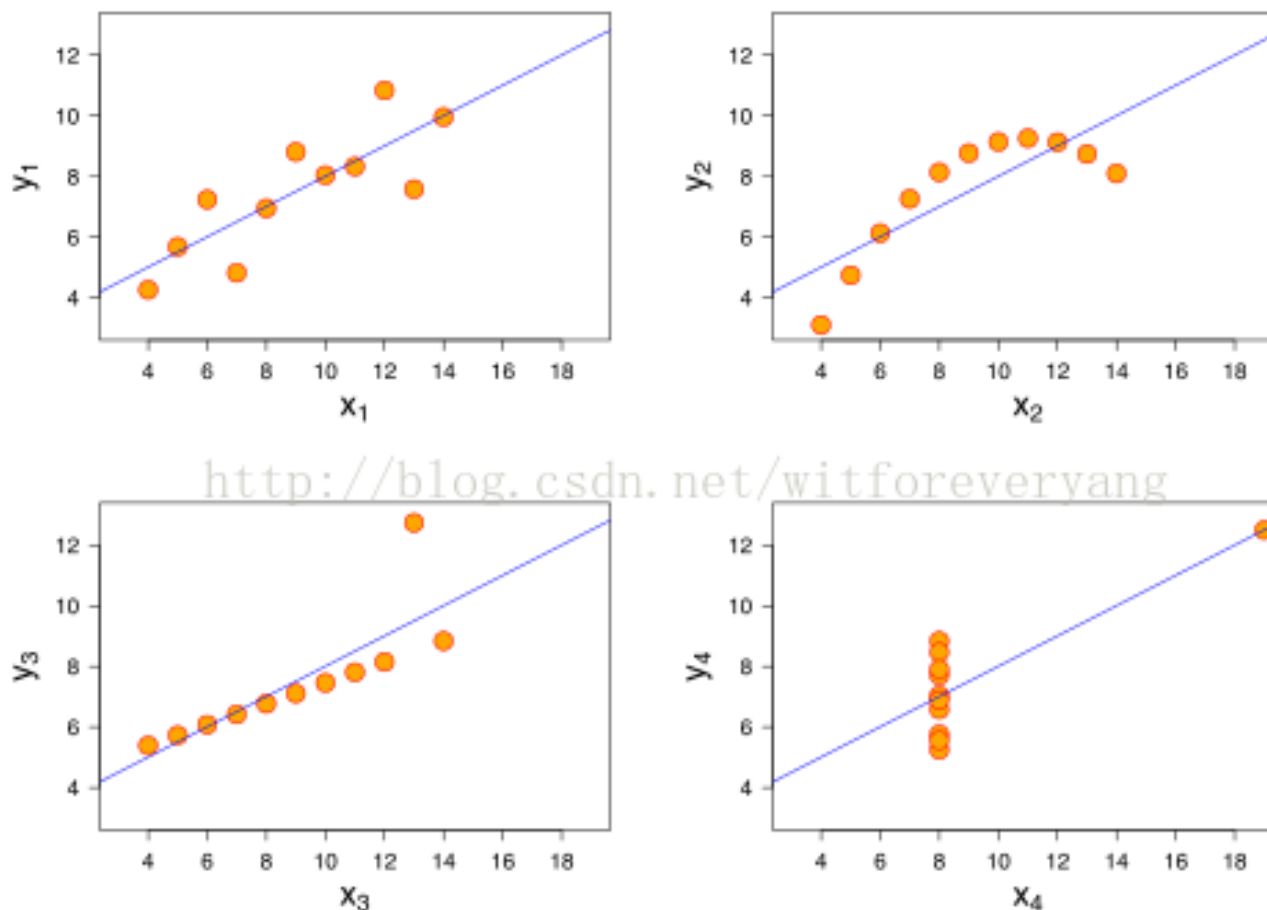


图 2-17 解释相关度从-1 到 1 的散布图

图片来源: [美]作者Pang-Ning Tan 《数据挖掘导论》

## 关于皮尔逊相关系数的一些理解误区

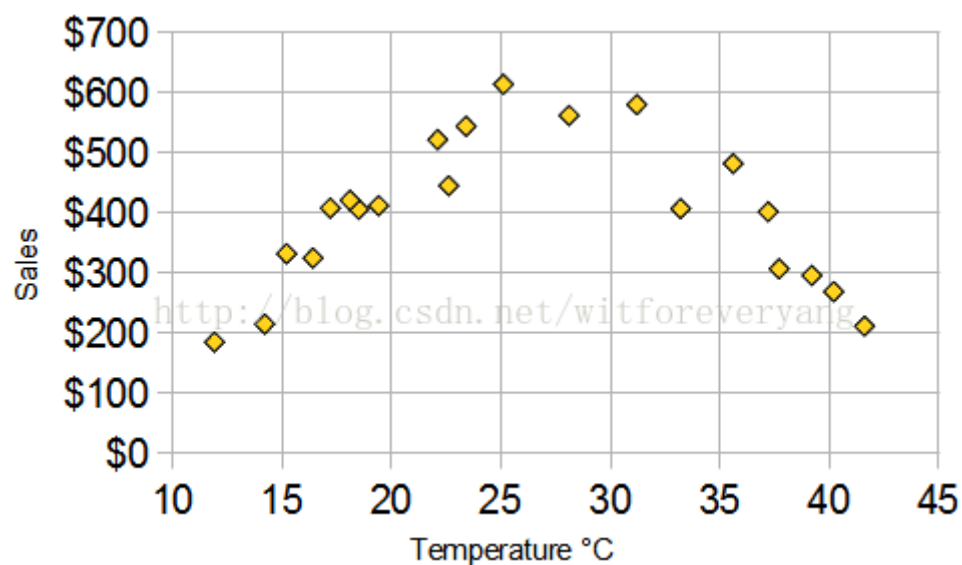


上面四个散点图对应的数据的皮尔逊相关系数均为0.816



## 关于皮尔逊相关系数的一些理解误区

冰激凌的销量和温度之间的关系:



相关系数计算结果为0

## 关于皮尔逊相关系数的一些理解误区

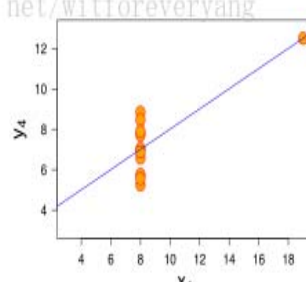
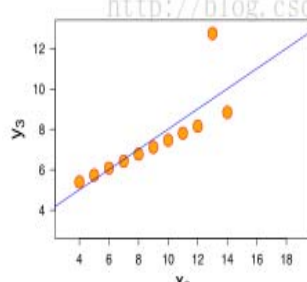
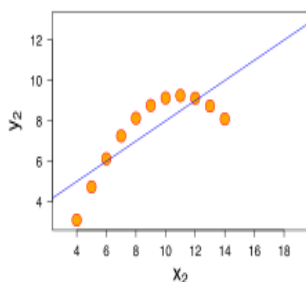
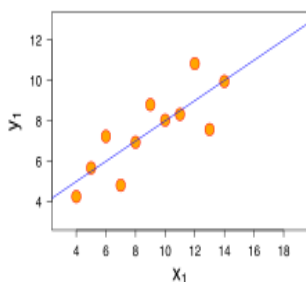
In statistics, the Pearson product-moment correlation coefficient is a measure of the **linear** correlation (dependence) between two variables  $X$  and  $Y$ , giving a value between  $+1$  and  $-1$  inclusive, where  $1$  is total positive correlation,  $0$  is no correlation, and  $-1$  is total negative correlation. It is widely used in the sciences as a measure of the degree of **linear** dependence between two variables.

--from wiki维基百科

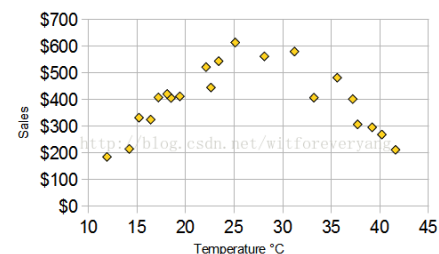
注意红色标注的"linear"：

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。

## 容易忽视和犯错的点



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。

## 两点总结

可别轻易下结论哦



(1) 如果两个变量本身就是线性的关系, 那么皮尔逊相关系数绝对值大的就是相关性 强, 小的就是相关性弱;

(2) 在不确定两个变量是什么关系的情况 下, 即使算出皮尔逊相关系数, 发现很大, 也不能说明那两个变量线性相关, 甚至不能 说他们相关, 我们一定要画出散点图来看才 行。

## 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
强相关性	-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)**

## 例题

现有某中学八年级所有女学生的体测样本数据, 请见下表, 试计算各变量之间的皮尔逊相关系数。

身高	体重	肺活量	50米跑	立定跳远	坐位体前屈
155	51	1687	9.7	158	9.3
158	52	1868	9.3	162	9.6
160	59	1958	9.9	178	9.5
163	59	1756	9.7	183	10.1
165	60	1575	9	156	10.4
151	47	1700	9.1	154	11.1
150	45	1690	9.7	164	12.5
147	43	1888	8.9	178	11.2
158	42	1949	12.1	168	10.6
161	51	1548	11.1	180	9.6
162	47	1624	10.1	191	9.8
165	47	1657	9.8	193	7.8
157	45	1574	9.6	190	8.7
154	41	1544	9.2	187	9.8
149	40	1687	9	167	9.7
...	...	...	...	...	...

## 描述性统计

Matlab中基本统计量的函数（一般用标粗的）：

函数名	功能
<b>min</b>	数组的最小元素
mink	计算数组的 k 个最小元素
<b>max</b>	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
<b>mean</b>	数组的均值
<b>median</b>	数组的中位数值
mode	数组的众数
<b>skewness</b>	数组的偏度
<b>kurtosis</b>	数组的峰度
<b>std</b>	标准差
var	方差

这些函数根据参数的不同有多种用法，我们这里用到的只是其最简单的功能，但这对建模已经足够了，需要用到高级功能的同学可以百度这些函数的用法。

## 结果演示

```

MIN = min(Test); % 每一列的最小值
MAX = max(Test); % 每一列的最大值
MEAN = mean(Test); % 每一列的均值
MEDIAN = median(Test); % 每一列的中位数
SKEWNESS = skewness(Test); % 每一列的偏度
KURTOSIS = kurtosis(Test); % 每一列的峰度
STD = std(Test); % 每一列的标准差
RESULT = [MIN;MAX;MEAN;MEDIAN;SKEWNESS;KURTOSIS;STD]
%将这些统计量放到一个矩阵中表示

```

将计算结果复制到EXCEL表格中

变量 - RESULT						
RESULT						
7x6 double						
	1	2	3	4	5	6
1	135	16	1450	7.8000	52	0.5000
2	171	65	3272	15	205	17.5000
3	156.0034	46.7834	2.3332e+...	10.7920	166.8257	9.4966
4	157	47	2391	10.7000	167	9.6000
5	-0.2954	-0.3607	-0.2852	0.7095	-0.8369	-0.2250
6	2.7427	9.4356	2.7520	3.2994	8.4313	2.7550
7	7.3894	5.0315	350.4362	1.3109	16.8136	2.9382

J	K	L	M	N	O	P
描述性统计						
	身高	体重	肺活量	50米跑	立定跳远	坐位体前屈
最小值	135	16	1450	7.8	52	0.5
最大值	171	65	3272	15	205	17.5
均值	156.0034	46.78342	2333.234	10.79201	166.8257	9.496616
中位数	157	47	2391	10.7	167	9.6
偏度	-0.29539	-0.36069	-0.28523	0.709546	-0.83687	-0.22497
峰度	2.742704	9.435585	2.751974	3.299414	8.431329	2.75503
标准差	7.38941	5.031473	350.4362	1.310873	16.81359	2.938186

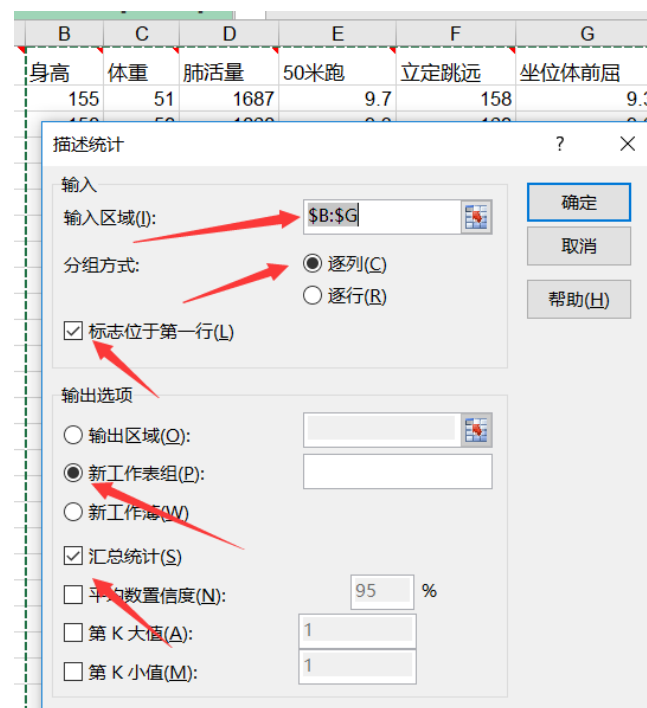
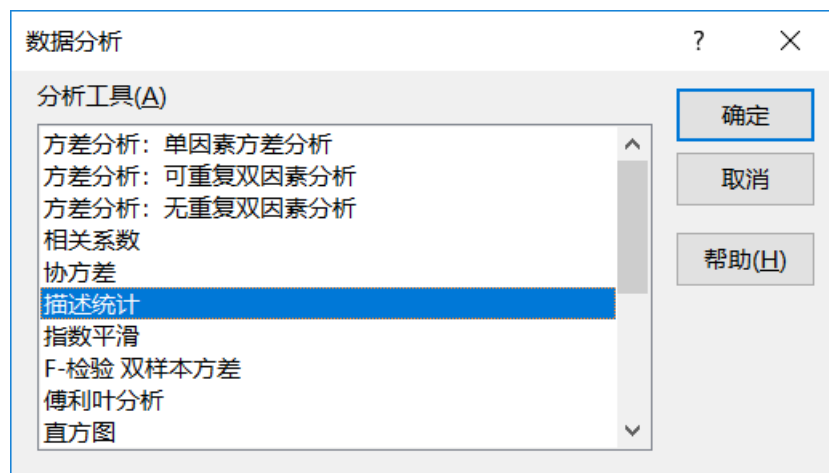


# EXCEL的数据分析工具



标题栏: 数据 – 数据分析

如果没有找到, 请百度: **Excel数据分析功能在哪里?**



## EXCEL描述性统计结果

	A	B	C	D	E	F	G	H	I	J	K	L
1	身高		体重		肺活量		50米跑		立定跳远		坐位体前屈	
2												
3	平均	156.0034	平均	46.78342	平均	2333.234	平均	10.79201	平均	166.8257	平均	9.496616
4	标准误差	0.30396	标准误差	0.206967	标准误差	14.41502	标准误差	0.053922	标准误差	0.691619	标准误差	0.120861
5	中位数	157	中位数	47	中位数	2391	中位数	10.7	中位数	167	中位数	9.6
6	众数	160	众数	50	众数	2400	众数	10.9	众数	160	众数	9.8
7	标准差	7.38941	标准差	5.031473	标准差	350.4362	标准差	1.310873	标准差	16.81359	标准差	2.938186
8	方差	54.60338	方差	25.31572	方差	122805.5	方差	1.718388	方差	282.6967	方差	8.632938
9	峰度	-0.24926	峰度	6.500587	峰度	-0.23992	峰度	0.312184	峰度	5.487782	峰度	-0.23683
10	偏度	-0.29615	偏度	-0.36161	偏度	-0.28595	偏度	0.711352	偏度	-0.839	偏度	-0.22554
11	区域	36	区域	49	区域	1822	区域	7.2	区域	153	区域	17
12	最小值	135	最小值	16	最小值	1450	最小值	7.8	最小值	52	最小值	0.5
13	最大值	171	最大值	65	最大值	3272	最大值	15	最大值	205	最大值	17.5
14	求和	92198	求和	27649	求和	1378941	求和	6378.08	求和	98594	求和	5612.5
15	观测数	591	观测数	591	观测数	591	观测数	591	观测数	591	观测数	591

不要直接粘贴这张表格到你的论文中，需要精简

## SPSS描述性统计结果

描述: 选项

☒ 平均值(M) ☐ 总和(S)

离散

☒ 标准差(I) ☒ 最小值(N)

☒ 方差(V) ☒ 最大值(X)

☐ 范围(R) ☐ 标准误差平均值(E)

分布

☒ 峰度(K) ☒ 偏度(W)

显示顺序

☒ 变量列表(B)

☐ 字母(A)

☐ 按平均值的升序排序(C)

☐ 按平均值的降序排序(D)

变量(V):

身高  
体重  
肺活量  
50米跑 [ @50米跑 ]  
立定跳远  
坐位体前屈

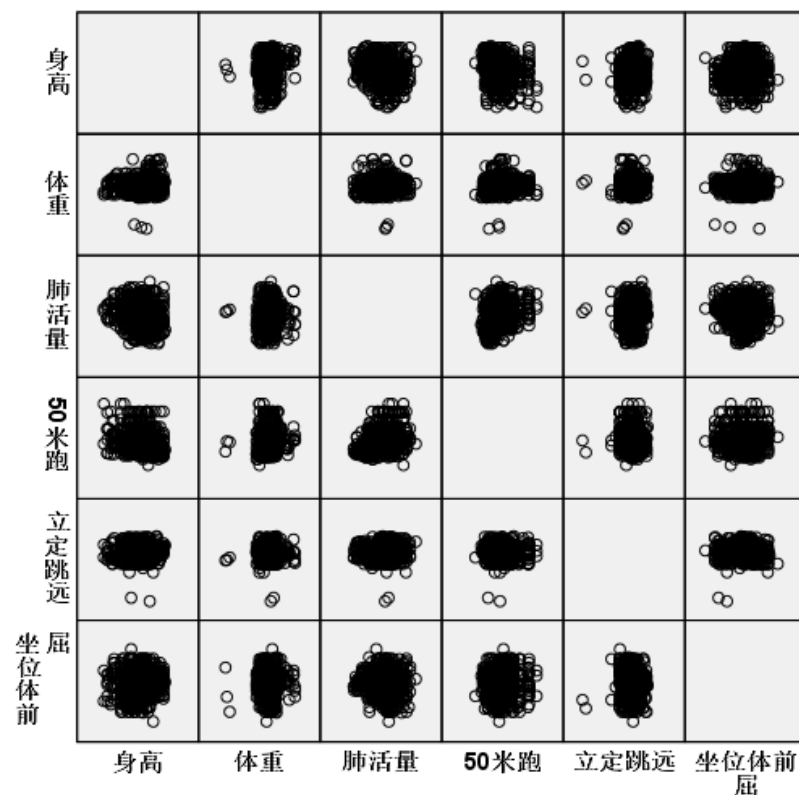
☒ 将标准化值另存为变量(Z)

确定 粘贴(P) 重置(R) 取消 帮助

描述统计										
	个案数 统计	最小值 统计	最大值 统计	平均值 统计	标准差 统计	方差 统计	偏度		峰度	
							统计	标准误差	统计	标准误差
身高	591	135	171	156.00	7.389	54.603	-.296	.101	-.249	.201
体重	591	16	65	46.78	5.031	25.316	-.362	.101	6.501	.201
肺活量	591	1450	3272	2333.23	350.436	122805.498	-.286	.101	-.240	.201
50米跑	591	7.800000000	15.00000000	10.79201354	1.310872852	1.718	.711	.101	.312	.201
立定跳远	591	52	205	166.83	16.814	282.697	-.839	.101	5.488	.201
坐位体前屈	591	.5000000000	17.50000000	9.496615905	2.938186121	8.633	-.226	.101	-.237	.201
有效个案数（成列）	591									

## 矩阵散点图

在计算皮尔逊相关系数之前,一定要做出散点图来看两组变量之间是否有线性关系  
这里使用Spss比较方便: 图形 - 旧对话框 - 散点图/点图 - 矩阵散点图



注意: 这个数据看起来特别奇怪, 是因为数据是我随机生成的。。。实际建模中遇到的数据应该不会这么奇怪~~~

## 皮尔逊相关系数的计算

corrcoef函数: [correlation coefficient](#)相关系数

$R = \text{corrcoef}(A)$

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)。

$R = \text{corrcoef}(A,B)$

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:

$R = \text{corrcoef}(\text{Test});$

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

## 如何美化相关系数表

苹果前置：

	A	B	C	D	E	F	G
1		身高	体重	肺活量	50米跑	立定跳远	坐位体前屈
2	身高	1	0.066531	-0.21766	-0.192	0.043973	0.0950686
3	体重	0.066531	1	0.095375	0.06854	0.027943	-0.0160892
4	肺活量	-0.21766	0.095375	1	0.289751	0.024827	-0.074931
5	50米跑	-0.192	0.06854	0.289751	1	-0.05868	-0.0018764
6	立定跳远	0.043973	0.027943	0.024827	-0.05868	1	-0.0174066
7	坐位体前屈	0.095069	-0.01609	-0.07493	-0.00188	-0.01741	1
8							

OPPO前置：

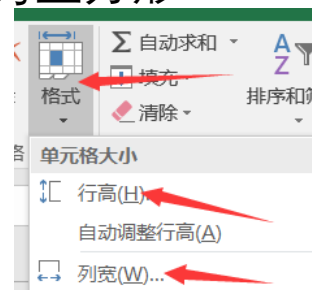
	身高	体重	肺活量	50米跑	立定跳远	坐位体前屈
身高	1	0.06653149	-0.2176628	-0.1920015	0.0439729	0.09506865
体重	0.06653149	1	0.09537485	0.06854	0.02794285	-0.0160892
肺活量	-0.2176628	0.09537485	1	0.28975123	0.02482733	-0.074931
50米跑	-0.1920015	0.06854	0.28975123	1	-0.0586811	-0.0018764
立定跳远	0.0439729	0.02794285	0.02482733	-0.0586811	1	-0.0174066
坐位体前屈	0.09506865	-0.0160892	-0.074931	-0.0018764	-0.0174066	1

注：其实R语言和Python中有很多类似的包，这里我们用Excel模仿

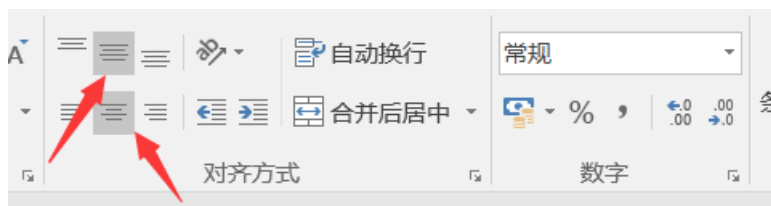


## 操作步骤总结

第一步: 在开始-格式中调整每个单元格的格式为正方形



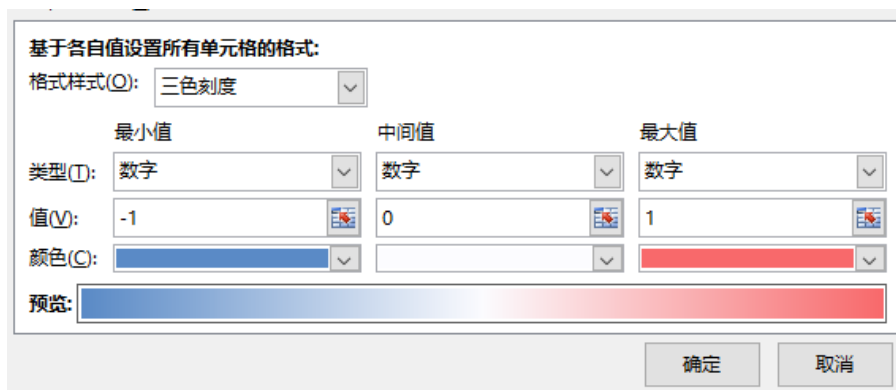
第二步: 在对齐方式中设置两个居中



第三步: 选中相关系数表, 开始-条件格式-色阶-选中红-白-蓝那个



第四步: 选中相关系数表, 选择条件格式-管理规则-编辑规则



## 对皮尔逊相关系数进行假设检验

第一步: 提出原假设 $H_0$ 和备择假设 $H_1$  (两个假设是截然相反的哦)

假设我们计算出了一个皮尔逊相关系数 $r$ , 我们想检验它是否显著的异于0.

那么我们可以这样设定原假设和备择假设:  $H_0: r = 0$  ,  $H_1: r \neq 0$

第二步: 在原假设成立的条件下, 利用我们要检验的量构造出一个符合某一分布的统计量

(注1: 统计量相当于我们要检验的量的一个函数, 里面不能有其他的随机变量)

(注2: 这里的分布一般有四种: 标准正态分布、 $t$ 分布、 $\chi^2$ 分布和 $F$ 分布)

对于皮尔逊相关系数 $r$ 而言, 在满足一定条件下, 我们可以构造统计量:

$$t = r\sqrt{\frac{n-2}{1-r^2}}, \text{ 可以证明 } t \text{ 是服从自由度为 } n-2 \text{ 的 } t \text{ 分布}$$

(注: 这里的条件在后面会提到, 证明很复杂, 可参考以下文献)

N.A Rahman, A Course in Theoretical Statistics; Charles Griffin and Company, 1968

注: 在数理统计中, 这里的原假设和备择假设中的 $r$ 应该改为 $\rho$ , 其中 $\rho$ 为未知的总体相关系数, 实际上我们关心的是总体的统计特征。但为了方便大家理解, 在这里我们做了简化, 非统计专业的同学理解到这个程度就足够了。



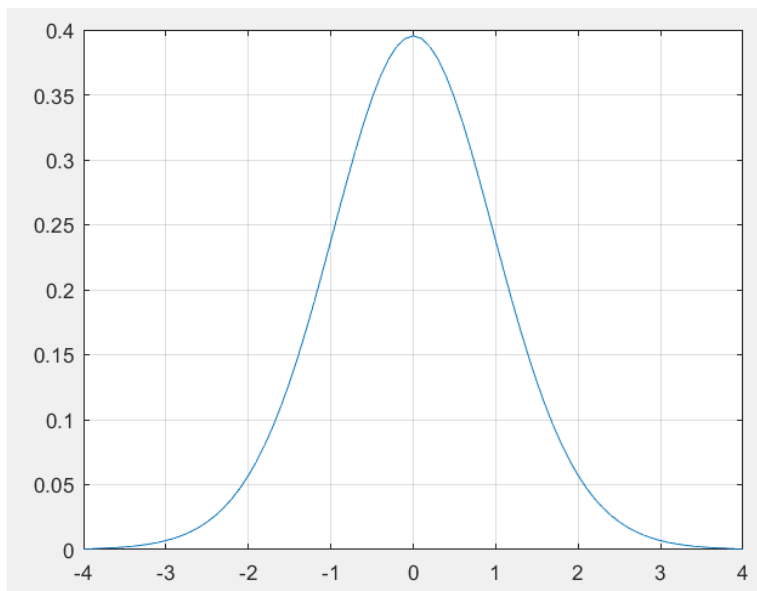
## 对皮尔逊相关系数进行假设检验

第三步: 将我们要检验的这个值带入这个统计量中, 可以得到一个特定的值 (检验值)。

假设我们现在计算出来的相关系数为0.5, 样本为30, 那么我们可以得到  $t^* = 0.5 \sqrt{\frac{30-2}{1-0.5^2}} = 3.05505$

第四步: 由于我们知道统计量的分布情况, 因此我们可以画出该分布的概率密度函数 $pdf$ , 并给定一个置信水平, 根据这个置信水平查表找到临界值, 并画出检验统计量的接受域和拒绝域。

例如, 我们知道上述统计量服从自由度为28的 $t$ 分布, 其概率密度函数图形如下:



```
x = -4:0.1:4;  
y = tpdf(x,28);  
plot(x,y,'-')  
grid on % 在画出的图上加上网格线
```

## 对皮尔逊相关系数进行假设检验

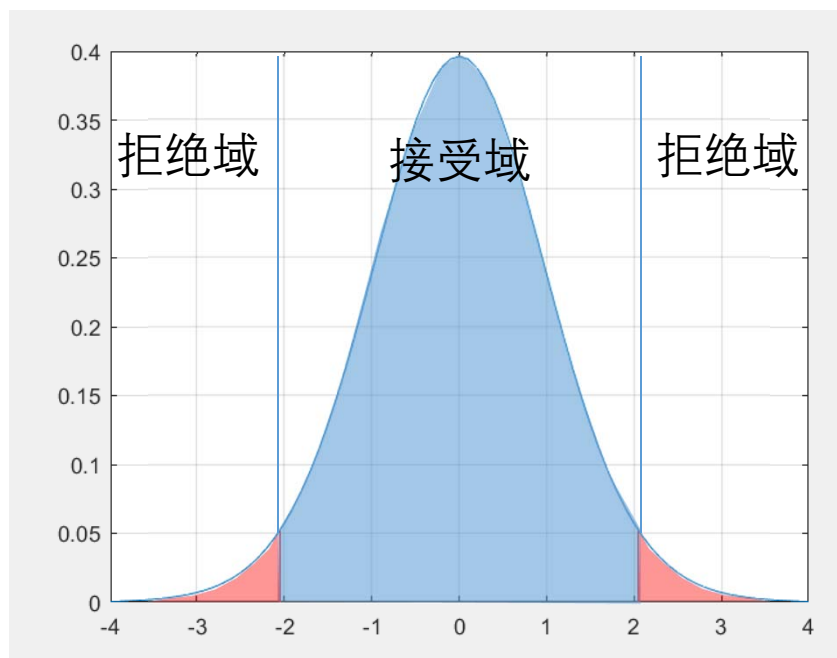
第四步: 由于我们知道统计量的分布情况, 因此我们可以画出该分布的概率密度函数 $pdf$ , 并给定一个置信水平, 根据这个置信水平查表找到临界值, 并画出检验统计量的接受域和拒绝域。

常见的置信水平有三个: 90%, 95%和99%, 其中95%是三者中最为常用的。

因为我们这里是双侧检验, 所以我们需要找出能覆盖0.95概率的部分

t分布表: <https://wenku.baidu.com/view/d94dbd116bd97f192279e94a.html>

查表可知, 对应的临界值为2.048, 因此我们可以做出接受域和拒绝域。

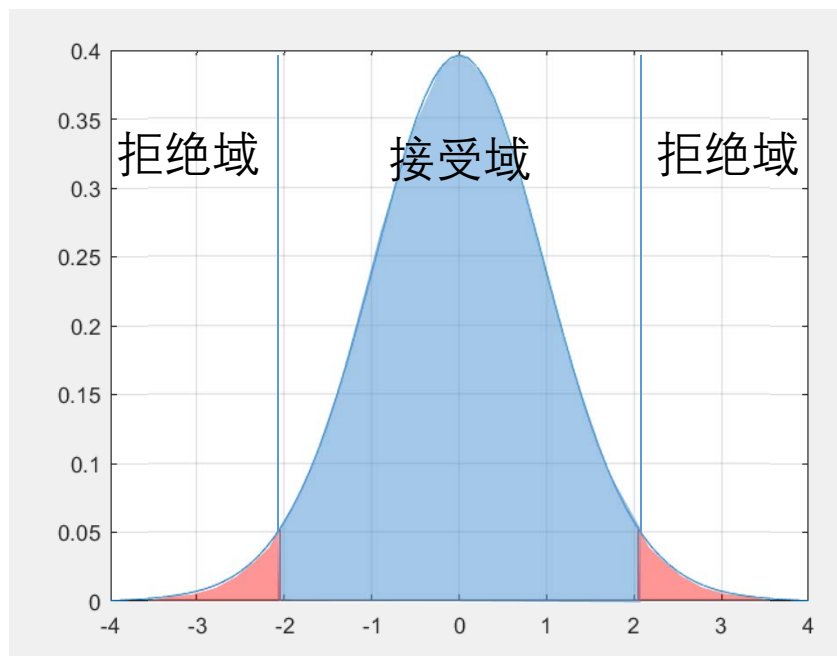


## 对皮尔逊相关系数进行假设检验

第五步: 看我们计算出来的检验值是落在了拒绝域还是接受域, 并下结论。

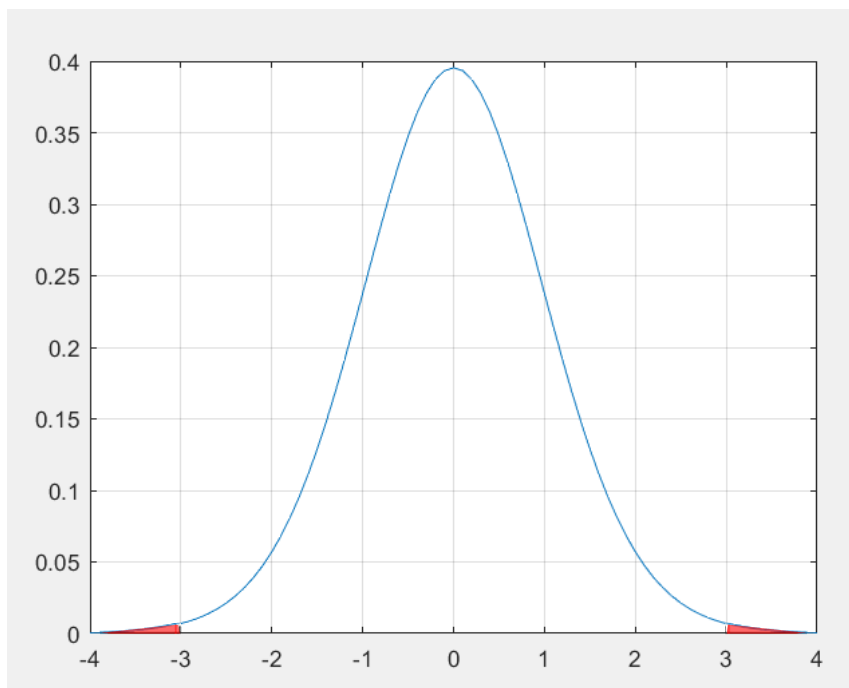
因为我们得到的 $t^* = 3.05505 > 2.048$ , 因此我们可以下结论:

在95%的置信水平上, 我们拒绝原假设 $H_0: r = 0$ , 因此 $r$ 是显著的不为0的。



## 更好用的方法: p值判断法

我们得到的检验值  $t^* = 3.05505$ , 根据这个值, 我们可以计算出其对应的那个概率



```
disp('该检验值对应的p值为: ')\ndisp((1-tcdf(3.055,28))*2)\n%双侧检验的p值要乘以2
```

注意这里的函数是tcdf: 累积分布函数

最后我们计算得到的p值为: 0.0049

$p < 0.01$ , 说明在99%的置信水平上拒绝原假设;

$p < 0.05$ , 说明在95%的置信水平上拒绝原假设;

$p < 0.10$ , 说明在90%的置信水平上拒绝原假设;

$p$  值  $> 0.01$ , 说明在99%的置信水平无法拒绝原假设;

$p > 0.05$ , 说明在95%的置信水平上无法拒绝原假设;

$p > 0.10$ , 说明在90%的置信水平上无法拒绝原假设;

在本例中, 拒绝原假设意味着皮尔逊相关系数显著的异于0。

小补充: 0.5、0.5\*、0.5\*\*、0.5\*\*\*的含义是什么? (显著性标记)

# 计算各列之间的相关系数以及p值

一行代码: `[R,P] = corrcoef(Test)`

R返回的是相关系数表, P返回的是对应于每个相关系数的p值

```
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

P =
    1.0000    0.1061    0.0000    0.0000    0.2859    0.0208
    0.1061    1.0000    0.0204    0.0960    0.4978    0.6963
    0.0000    0.0204    1.0000    0.0000    0.5469    0.0687
    0.0000    0.0960    0.0000    1.0000    0.1542    0.9637
    0.2859    0.4978    0.5469    0.1542    1.0000    0.6728
    0.0208    0.6963    0.0687    0.9637    0.6728    1.0000
```

%% 计算各列之间的相关系数以及p值

`[R,P] = corrcoef(Test)`

% 在EXCEL表格中给数据右上角标上显著性符号吧

`P < 0.01` % 标记3颗星的位置

`(P < 0.05) .* (P > 0.01)` % 标记2颗星的位置

`(P < 0.1) .* (P > 0.05)` % 标记1颗星的位置

	身高	体重	肺活量	50米跑	立定跳远	坐位体前屈
身高	1.0000	0.0665	-0.2177***	-0.192***	0.0440	0.0951**
体重	0.0665	1.0000	0.0954**	0.0685*	0.0279	-0.0161
肺活量	-0.2177***	0.0954**	1.0000	0.2898***	0.0248	-0.0749*
50米跑	-0.192***	0.0685*	0.2898***	1.0000	-0.0587	-0.0019
立定跳远	0.0440	0.0279	0.0248	-0.0587	1.0000	-0.0174
坐位体前屈	0.0951**	-0.0161	-0.0749*	-0.0019	-0.0174	1.0000

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

Matlab计算的是双侧检验的p值, 如果需要单侧的话只需要除以2即可。

## 皮尔逊相关系数假设检验的条件

**第一，实验数据通常假设是成对的来自于正态分布的总体。**因为我们在求皮尔逊相关性系数以后，通常还会用t检验之类的方法来进行皮尔逊相关性系数检验，而t检验是基于数据呈正态分布的假设的。

**第二，实验数据之间的差距不能太大。**皮尔逊相关性系数受异常值的影响比较大。

**第三：每组样本之间是独立抽样的。**构造t统计量时需要用到。

如何检验数据是否是正态分布？

## 正态分布JB检验 (大样本 $n > 30$ )

### 雅克-贝拉检验(Jarque-Bera test)

对于一个随机变量  $\{X_i\}$ , 假设其偏度为 $S$ , 峰度为 $K$ , 那么我们可以构造 $JB$ 统计量:

$$JB = \frac{n}{6} \left[ S^2 + \frac{(K-3)^2}{4} \right]$$

可以证明, 如果  $\{X_i\}$  是正态分布, 那么在大样本情况下  $JB \sim \chi^2(2)$  (自由度为2的卡方分布)

注: 正态分布的偏度为0, 峰度为3

那么进行假设检验的步骤如下:

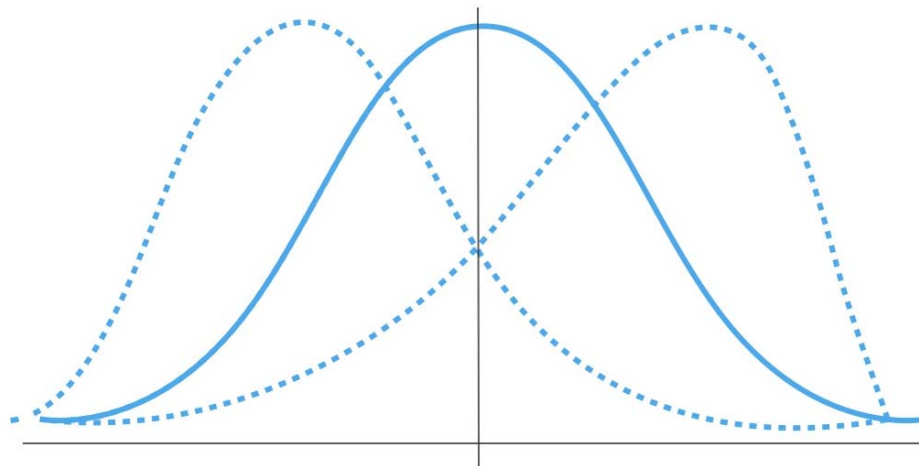
$H_0$ : 该随机变量服从正态分布  $H_1$ : 该随机变量不服从正态分布

然后计算该变量的偏度和峰度, 得到检验值 $JB^*$ , 并计算出其对应的 $p$ 值

将 $p$ 值与0.05比较, 如果小于0.05则可拒绝原假设, 否则我们不能拒绝原假设。

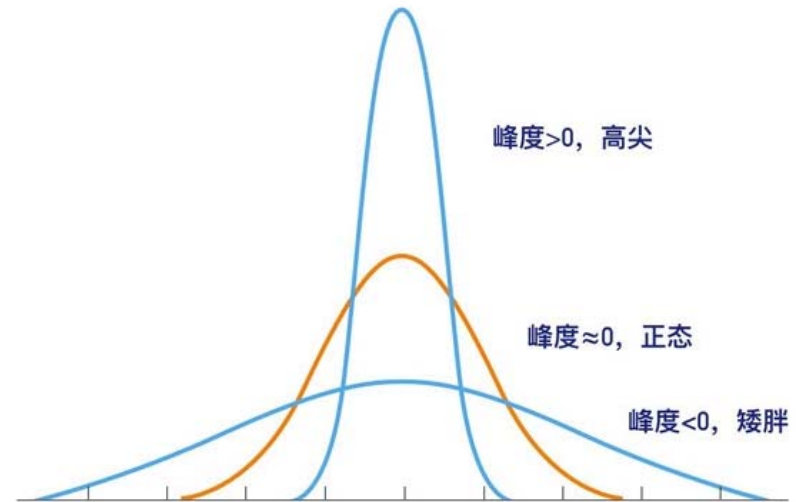
## 偏度和峰度

偏度>0, 正偏态      偏度≈0, 正态      偏度<0, 负偏态



$$E\left[\left(\frac{X-u}{\sigma}\right)^3\right]$$

正态分布的偏度为0



$$E\left[\left(\frac{X-u}{\sigma}\right)^4\right] \text{ 或 } E\left[\left(\frac{X-u}{\sigma}\right)^4\right] - 3$$

正态分布的峰度为3  
(有些地方定义的正态分布峰度为0)  
Matlab软件中使用的是第一种定义

```
x = normrnd(2,3,100,1);  
% 生成100*1的随机向量, 每个元素是均值为2, 标准差为3的正态分布  
skewness(x) %偏度  
kurtosis(x) %峰度
```



## MATLAB结果

MATLAB中进行JB检验的语法:  $[h,p] = jbtest(x,alpha)$

当输出 $h$ 等于1时, 表示拒绝原假设;  $h$ 等于0则代表不能拒绝原假设。

$alpha$ 就是显著性水平, 一般取0.05, 此时置信水平为 $1-0.05=0.95$

$x$ 就是我们要检验的随机变量, 注意这里的 $x$ 只能是向量。

```
%% 正态分布检验
% 检验第一列数据是否为正态分布
[h,p] = jbtest(Test(:,1),0.05)

% 用循环检验所有列的数据
n_c = size(Test,2); % number of column 数据的列数
H = zeros(1,6);
P = zeros(1,6);
for i = 1:n_c
    [h,p] = jbtest(Test(:,i),0.05);
    H(i)=h;
    P(i)=p;
end
disp(H)
disp(P)
```

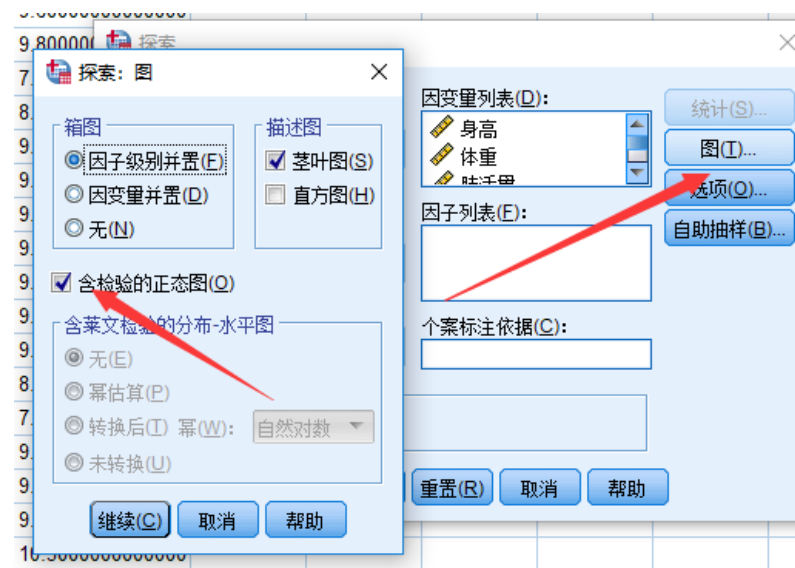
## 小样本 $3 \leq n \leq 50$ : Shapiro-wilk 检验

### Shapiro-wilk 夏皮洛-威尔克检验

$H_0$ : 该随机变量服从正态分布  $H_1$ : 该随机变量不服从正态分布

计算出威尔克统计量后, 得到相应的  $p$  值

将  $p$  值与 0.05 比较, 如果小于 0.05 则可拒绝原假设, 否则我们不能拒绝原假设。



## 另一种常见的方法: Q-Q图

在统计学中, Q-Q图 (Q代表分位数Quantile) 是一种通过比较两个概率分布的分位数对这两个概率分布进行比较的概率图方法。

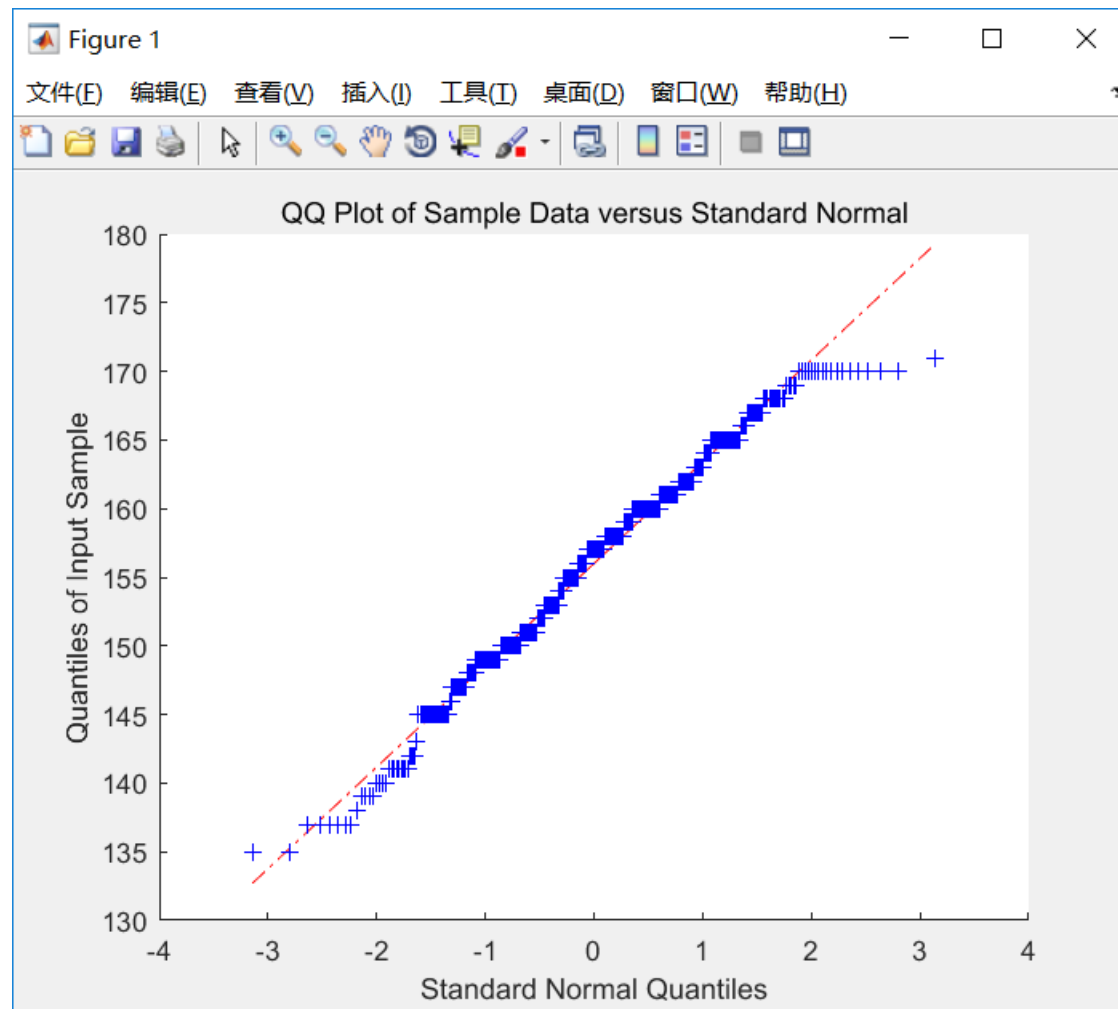
首先选定分位数的对应概率区间集合, 在此概率区间上, 点 $(x,y)$ 对应于第一个分布的一个分位数 $x$ 和第二个分布在和 $x$ 相同概率区间上相同的分位数。

这里, 我们选择正态分布和要检验的随机变量, 并对其做出QQ图, 可想而知, 如果要检验的随机变量是正态分布, 那么QQ图就是一条直线。

要利用Q-Q图鉴别样本数据是否近似于正态分布, 只需看Q-Q图上的点是否近似地在一条直线附近。(要求数据量非常大)

## 第一列数据和正态分布的Q-Q图

```
qqplot(Test(:,1))
```



## 斯皮尔曼spearman相关系数

定义:  $X$ 和 $Y$ 为两组数据, 其斯皮尔曼(等级)相关系数:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

其中,  $d_i$ 为 $X_i$ 和 $Y_i$ 之间的等级差。

(一个数的等级, 就是将它所在的一列数按照从小到大排序后, 这个数所在的位置)

可以证明:  $r_s$ 位于 $-1$ 和 $1$ 之间。

X	Y	X的等级	Y的等级	等级差	等级差的平方
3	5	2	1	1	1
8	10	5	4.5	0.5	0.25
4	8	3	3	0	0
7	10	4	4.5	-0.5	0.25
2	6	1	2	-1	1

注: 如果有的数值相同, 则将它们所在的位置取算术平均。

# 斯皮尔曼spearman相关系数

X	Y	X的等级	Y的等级	等级差	等级差的平方
3	5	2	1	1	1
8	10	5	4.5	0.5	0.25
4	9	3	3	0	0
7	10	4	4.5	-0.5	0.25
2	6	1	2	-1	1

根据公式:  $r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$  可得:

$$X \text{ 和 } Y \text{ 的斯皮尔曼相关系数 } r_s = 1 - \frac{6 \times (1 + 0.25 + 0.25 + 1)}{5 \times 24} = 0.875$$

## 另一种斯皮尔曼spearman相关系数的定义

斯皮尔曼相关系数被定义成等级之间的皮尔逊相关系数。

X	Y	X的等级	Y的等级	等级差	等级差的平方
3	5	2	1	1	1
8	10	5	4.5	0.5	0.25
4	9	3	3	0	0
7	10	4	4.5	-0.5	0.25
2	6	1	2	-1	1

```
%% MATLAB求解皮尔逊相关系数
RX = [2 5 3 4 1]
RY = [1 4.5 3 4.5 2]
R = corrcoef(RX,RY)
```

R =

```
1.0000    0.8721
0.8721    1.0000
```

和之前的结果有微小差别。

# MATLAB中计算斯皮尔曼相关系数

两种用法

(1) `corr(X, Y, 'type', 'Spearman')`

这里的X和Y必须是列向量

(2) `corr(X, 'type', 'Spearman')`

这时计算X矩阵各列之间的斯皮尔曼相关系数

X	Y
3	5
8	10
4	9
7	10
2	6

```
X = [3 8 4 7 2]'; % 一定要是列向量哦, 一撇'表示求转置
Y = [5 10 9 10 6]';
coeff = corr(X, Y, 'type', 'Spearman')
```

`coeff =`

0.8721

这说明Matlab使用的是第二种计算方法



## 两种相关系数计算结果的对比

```
>> 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

皮尔逊相关系数

```
>> R = corr(Test, 'type', 'Spearman')
```

R =

1.0000	0.0301	-0.2430	-0.1990	0.0624	0.1099
0.0301	1.0000	0.1305	0.0898	0.0216	-0.0488
-0.2430	0.1305	1.0000	0.2626	0.0219	-0.0801
-0.1990	0.0898	0.2626	1.0000	-0.0910	-0.0029
0.0624	0.0216	0.0219	-0.0910	1.0000	-0.0399
0.1099	-0.0488	-0.0801	-0.0029	-0.0399	1.0000

斯皮尔曼相关系数

# 斯皮尔曼相关系数的假设检验

分为小样本和大样本两种情况:

小样本情况, 即 $n \leq 30$ 时, 直接查临界值表即可。

表B.7 斯皮尔曼等级相关的临界值\*

\*样本相关系数 $r$ 必须大于等于表中的临界值, 才能得出显著的结论。

$n$	单尾检验的显著水平			
	.05	.025	.01	.005
	双尾检验的显著水平			
	.10	.05	.02	.01
4	1.000			
5	0.900	1.000	1.000	
6	0.829	0.886	0.943	1.000
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.600	0.700	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.536	0.618	0.709	0.755
12	0.503	0.587	0.671	0.727
13	0.484	0.560	0.648	0.703
14	0.464	0.538	0.622	0.675

$$H_0: r_s = 0$$

$$H_1: r_s \neq 0$$

## 斯皮尔曼相关系数的假设检验

大样本情况下, 统计量  $r_s \sqrt{n-1} \sim N(0, 1)$

$$H_0: r_s = 0, \quad H_1: r_s \neq 0$$

我们计算检验值  $r_s \sqrt{n-1}$ , 并求出对应的  $p$  值与 0.05 相比即可。

R =

1.0000	0.0301	-0.2430	-0.1990	0.0624	0.1099
0.0301	1.0000	0.1305	0.0898	0.0216	-0.0488
-0.2430	0.1305	1.0000	0.2626	0.0219	-0.0801
-0.1990	0.0898	0.2626	1.0000	-0.0910	-0.0029
0.0624	0.0216	0.0219	-0.0910	1.0000	-0.0399
0.1099	-0.0488	-0.0801	-0.0029	-0.0399	1.0000



[1.0000, 0.0301, ...  
591x6 double  
1x81 double

检验值  $z^* = 0.0301 \sqrt{591-1} = 0.731126$

```
>> disp((1-normcdf(0.7311))*2)
0.4647
```

$p$  值大于 0.05, 因此我们无法拒绝原假设。(和 0 没有显著的差异)

## 斯皮尔曼相关系数的假设检验

% 直接给出相关系数和p值

```
[R,P]=corr(Test, 'type' , 'Spearman' )
```

R =

1.0000	0.0301	-0.2430	-0.1990	0.0624	0.1099
0.0301	1.0000	0.1305	0.0898	0.0216	-0.0488
-0.2430	0.1305	1.0000	0.2626	0.0219	-0.0801
-0.1990	0.0898	0.2626	1.0000	-0.0910	-0.0029
0.0624	0.0216	0.0219	-0.0910	1.0000	-0.0399
0.1099	-0.0488	-0.0801	-0.0029	-0.0399	1.0000

P =

1.0000	0.4647	0.0000	0.0000	0.1295	0.0075
0.4647	1.0000	0.0015	0.0290	0.5996	0.2362
0.0000	0.0015	1.0000	0.0000	0.5944	0.0517
0.0000	0.0290	0.0000	1.0000	0.0270	0.9436
0.1295	0.5996	0.5944	0.0270	1.0000	0.3330
0.0075	0.2362	0.0517	0.9436	0.3330	1.0000

## 两个相关系数的比较

### 斯皮尔曼相关系数和皮尔逊相关系数选择:

- 1.连续数据, 正态分布, 线性关系, 用pearson相关系数是最恰当, 当然用spearman相关系数也可以, 就是效率没有pearson相关系数高。
- 2.上述任一条件不满足, 就用spearman相关系数, 不能用pearson相关系数。
- 3.两个定序数据之间也用spearman相关系数, 不能用pearson相关系数。

**定序数据**是指仅仅反映观测对象等级、顺序关系的数据, 是由定序尺度计量形成的, 表现为类别, 可以进行排序, 属于品质数据。

例如: 优、良、差;



我们可以用1表示差、2表示良、3表示优, 但请注意, 用2除以1得出的2并不代表任何含义。定序数据最重要的意义代表了一组数据中的某种逻辑顺序。

注: 斯皮尔曼相关系数的适用条件比皮尔逊相关系数要广, 只要数据满足单调关系(例如线性函数、指数函数、对数函数等)就能够使用。

## 课后作业

(1) 写一篇文章, 分析男生体测数据各指标之间的相关性, 并与女生的数据得到的结论进行对比。

数据见文件:

 八年级男生体测数据.xls	2019/07/08 19:04	Microsoft Excel ...	86 KB
 八年级女生体测数据.xls	2019/07/08 16:16	Microsoft Excel ...	88 KB

要求: 要说明选择哪一种相关系数的原因, 并要求做出散点图。

(可以自己动手试试相关系数矩阵可视化。因为数据是我随机生成的, 所以可能效果看起来很奇怪, 这里只供大家练手, 实际比赛中按照这个流程做就好了)

(2) 自己尝试编程实现斯皮尔曼 (只考虑样本数 $n>30$ ) 第一种方法的计算。

要求写一个函数:

```
function [ R , P ]= fun_spearman(X, kind)
```

其中,  $X$ 是数据矩阵,  $kind$ 用来区分是单侧还是双侧检验 ( $kind=1$ 表示单侧检验;  $kind=2$ 表示双侧检验), 如果用户没有输入 $kind$ 值, 则默认为双侧检验。

要求计算出 $X$ 各列之间的相关系数矩阵 $R$ , 并求出对应的 $P$ 值。

提示: 设置默认参数可百度Matlab中`nargin`的用法

关于函数的定义可参见第二节的内容, 大家可设置子函数简化步骤