数学实践—

一、相关分析

主要内容

- 相关分析
- 偏相关分析

相关分析概述

(一)相关关系

(1)函数关系:事物间的一种一一对应的确定性关系.即:当一个变量x取一定值时,另一变量y可以依确定的关系取一个确定的值

- **如:销售额与销售量;圆面积和圆半径**

(2)统计关系:事物间的关系不是确定性的.即:当一个变量x取一定值时,另一变量y的取值可能有几个.一个变量的值不能由另一个变量唯一确定

- **如:收入和消费;身高的遗传.**

相关分析概述

- **统计关系的常见类型:**
  - **线性相关:正线性相关、负线性相关**
  - **非线性相关**
  - **统计关系不象函数关系那样直接,但却普遍存在,且有强有弱.如何测度?**
- **相关分析的研究对象:统计关系**
- **相关分析旨在测度变量间线性关系的强弱程度**

相关分析

(一)目的

通过样本数据,研究两变量间线性相关程度的强弱.

(二)基本方法

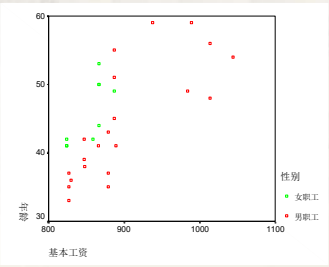
绘制散点图、计算相关系数

绘制散点图

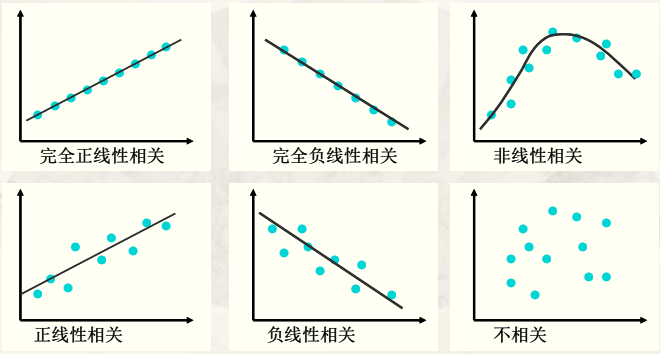
(一)散点图

- **将数据以点的形式绘制在直角平面上.比较直观,可以用来发现变量间的关系和可能的趋势.**

正相关趋势



绘制散点图



散点图的案例举例

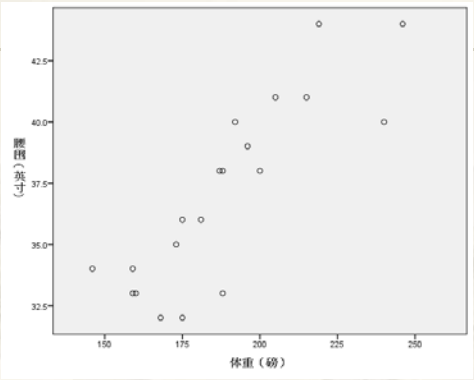
案例1：为研究腰围、体重和脂肪比重之间的关系，随机调查了20个人。具体数据在可供下载的压缩包中，文件名为“腰围和体重. sav"。

SPSS绘制散点图的基本操作

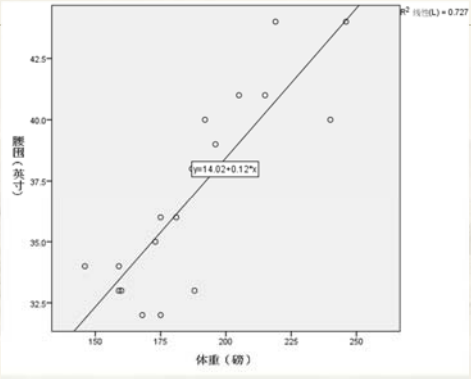
选择菜单：

【图形 (G) 】-【旧对话框(L)】-【散点 / 点状 CS) 】

体重与腰围的简单散点图



由图粗略可知， 体重与腰围之间存在较强的正相关关系。



计算相关系数

(一)相关系数

(1)作用:

- **以精确的相关系数(r)体现两个变量间的线性关系程度.**
- **r:[-1,+1]; r=1:完全正相关; r=-1:完全负相关; r=0:无线性相关; |r|>0.8:强相关; |r|<0.3:弱相关**

计算相关系数

(2)说明:

- 相关系数只是较好地度量了两变量间的线性相关程度,不能描述非线性关系.
- 如:x和y的取值为:(-1,-1) (-1,1) (1,-1) (1,1), r=0 但 $x^2+y^2=2$
- 数据中存在极端值时不好
- 如:(1,1)(2,2)(3,3),(4,4),(5,5),(6,1), r=0.33,但总体上表现出x=y,应结合散点图分析

计算相关系数

(3)种类:

- 简单线性相关系数(Pearson):针对定距数据.

  (如:身高和体重)

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

计算相关系数

- Spearman相关系数:用来度量定序变量间的线性相关关系(如:不同年龄段与不同收入段,职称和受教育年份)
  - 利用秩(数据的排序次序).认为:如果x与y相关,则相应的秩U_i、V_i也具有同步性.
  - 首先得到两变量中各数据的秩(U_i、V_i),并计算D_i^2统计量.
    - 若两变量存在强正相关性,则Di2应较小,秩序相关系数较大.若两变量存在强负相关性,则Di2应较大,秩序相关系数为负,绝对值较大
  - 计算Spearman相关系数,与简单相关系数形式完全相同.

$$\sum_{i=1}^{n}D_i^2=\sum_{i=1}^{n}(U_i-V_i)^2 \qquad R=1-\frac{6\sum D_i^2}{n(n^2-1)}$$

计算相关系数

- Kendall τ 相关系数:度量定序变量间的线性相关关系
  - 首先计算一致对数目(U)和非一致对数目(V)

  如: 对x和y求秩后为: x: 2 4 3 5 1

  y: 3 4 1 5 2

  x的秩按自然顺序排序后: x: 1 2 3 4 5

  y: 2 3 1 4 5

  一致对U:(2,3) (2,4)(2,5)(3,4)(3,5)(1,4)(1,5)(4,5); 非一致对V:(2,1)(3,1)
  - 然后计算Kendall相关系数
  - 若两变量存在强正相关性,则V较小,秩相关系数较大;若两变量存在强负关性,则V较大,秩相关系数为负,绝对值较大

$$\tau=(U-V)\frac{2}{n(n-1)}$$

计算相关系数

(二)相关系数检验

- 应对两变量来自的总体是否相关进行统计推断.
  - 原因:抽样的随机性、样本容量小等

(1)H_0:两总体零相关

(2)构造统计量

简单相关系数

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

t 统计量服从 $n-2$ 个自由度的 t 分布。

Spearman系数,大样本下,近似标准正态分布

$$Z=R\sqrt{n-1}$$

kendall系数,大样本下,近似标准正态分布

$$Z=\tau\sqrt{\frac{9n(n-1)}{2(2n+5)}}$$

计算相关系数

(二)相关系数检验

(3)计算统计量的值,并得到对应的相伴概率p

(4)结论:

- 如果p<=a,则拒绝H_0,两总体存在线性相关;
- 如果p>a,不能拒绝H_0.

计算相关系数

(三)基本操作步骤

(1)菜单选项:分析->相关->双变量

(2)选择计算相关系数的变量到变量框.

(3)选择相关系数.

(4)显著性检验

- **输出双尾检验概率P**
- **输出单尾检验概率P**

计算相关系数的应用举例

- 对于案例1，通过绘制散点图得知体重与腰围之间存在较强的正相关关系，为更准确地反映两者之间线性关系的强弱，采用计算相关系数的方法。由于这两个变量均为数值型变量，因此采用简单相关系数。

表　　案例数据的相关系数计算结果

相关性

		腰围（英寸）	体重（磅）
腰围（英寸）	Pearson 相关性	1	0.853**
	显著性（双侧）		0.000
	N	20	20
体重（磅）	Pearson 相关性	0.853**	1
	显著性（双侧）	0.000	
	N	20	20

**. 在 0.01 水平（双侧）上显著相关。

另外，表　　中相关系数上角的两个星号（**）表示显著性水平 α 为 0.01 时拒绝原假设。一个星号（*）表示显著性水平 α 为 0.05 时拒绝原假设。因此，两个星号比一个星号拒绝原假设犯错误的可能性更小。

偏相关分析

(一)偏相关系数

(1)含义:

在控制了其他变量的影响下计算两变量的相关系数

- **虚假相关.**
- **研究商品的需求量和价格、消费者收入之间的关系.因为:需求量和价格之间的相关关系包含了消费者收入对商品需求量的影响；收入对价格也产生影响，并通过价格变动传递到对商品需求量的影响中。**

- 偏相关分析也称净相关分析，它在控制其他变量的线性影响的条件下分析两变量间的线性关系，所采用的工具是偏相关系数。
- 控制变量个数为1时，偏相关系数称一阶偏相关；当控制两个变量时，偏相关系数称为二阶偏相关；当控制变量的个数为0时，偏相关系数称为零阶偏相关，也就是简单相关系数。

利用偏相关系数进行分析的步骤

- **第一，计算样本的偏相关系数**

假设有三个变量y、x1和x2，在分析x1和y之间的净相关时，当控制了x2的线性作用后，x1和y之间的一阶偏相关定义为：

$$r_{y1,2}=\frac{r_{y1}-r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}}$$

其中，r_{y1}、r_{y2}、r_{12}分别表示y和x_1的相关系数、y和x_2的相关系数、x_1和x_2的相关系数。

偏相关系数的取值范围及大小含义与相关系数相同。

- 第二，对样本来自的两总体是否存在显著的净相关进行推断
- 检验统计量为：

$$t = r\sqrt{\frac{n-q-2}{1-r^2}}$$

其中，r为偏相关系数，n为样本数，q为阶数。T统计量服从n-q-2个自由度的t分布。

偏相关分析

(二)基本操作步骤

(1)菜单选项:分析->相关->偏相关

(2)选择将参加计算的变量到变量框.

(3)选择控制变量到控制框。

(4)选项:

- **零阶相关系数:输出简单相关系数矩阵**

偏相关分析的应用举例

- 对于案例1, 已经分析了体重与腰围之间的相关性、直觉感觉这种相关性会受到体 内脂肪比重的影响。为此，可将脂肪比重作为控制变量，对体重和腰围作偏相关分析。

相关性

控制变量			腰围（英寸）	体重（磅）	%脂肪比重
-无-[a]	腰围（英寸）	相关性	1.000	.853	.887
		显著性（双尾）	.	.000	.000
		自由度	0	18	18
	体重（磅）	相关性	.853	1.000	.697
		显著性（双尾）	.000	.	.001
		自由度	18	0	18
	%脂肪比重	相关性	.887	.697	1.000
		显著性（双尾）	.000	.001	.
		自由度	18	18	0
%脂肪比重	腰围（英寸）	相关性	1.000	.709	
		显著性（双尾）	.	.001	
		自由度	0	17	
	体重（磅）	相关性	.709	1.000	
		显著性（双尾）	.001	.	
		自由度	17	0	

a. 单元格包含零阶（皮尔逊）相关性。

在脂肪比重作为控制变量的条件下， 体重和腰围的偏相关系数为 0.709，仍呈一定的正相关， 低于简单相关系数。