相关分析

授课人：徐杨

相关分析

- 两个变量间线性关系的方向和强度
- 协方差
- 相关系数

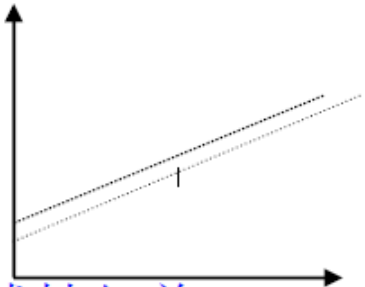
相关分析

相关分析是研究现象之间是否存在某种依存关系，并对具体有依存关系的现象探讨其相关方向以及相关程度，是研究随机变量之间的相关关系的一种统计方法。

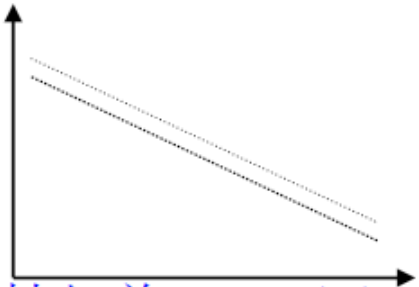
（1）按变量的多少划分：①单相关 ②复相关
（2）按表现形态划分：①直线相关 ②曲线相关
（3）从变动的方向划分：①正相关 ②负相关
（4）按相关的程度不同分：①完全相关 ②统计相关 ③完全无关

通过散点图来描述相关

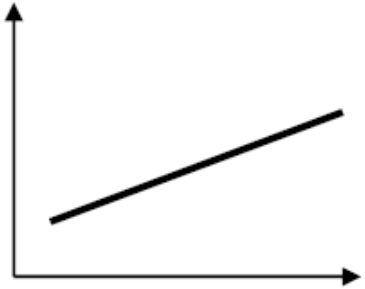
散点图是描述变量之间相关关系的一种直观方法。我们用横坐标中代表自变量x,纵坐标代表因变量y，每组数据（x,y）在坐标系中用一个点表示，n组数据在坐标系形成的点称为散点，这样的图称为散点图。散点图描述了两个变量之间的大致关系，从中可以直观地看出变量间的关系形态及关系强度。图7.1就是不同形态的散点图。



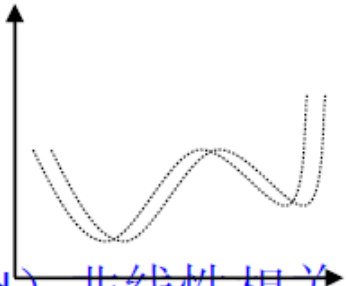
(a)正线性相关



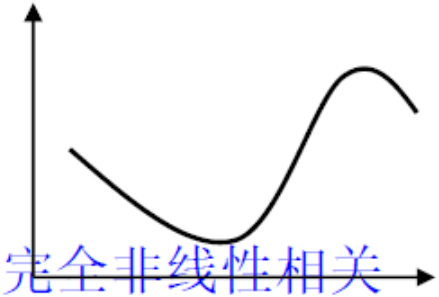
（b）负线性相关



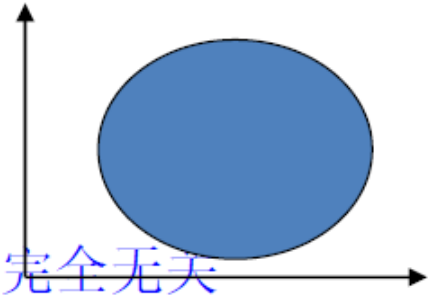
（c）完全线性相关



（d）非线性相关



（e）完全非线性相关



（f）完全无关

协方差

期望值分别为$E[X]$与$E[Y]$的两个实随机变量X与Y之间的**协方差**$Cov(X,Y)$定义为：

$$\begin{aligned} Cov(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY]-2E[Y]E[X]+E[X]E[Y] \\ &= E[XY]-E[X]E[Y] \end{aligned}$$

如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值时另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值；如果两个变量的变化趋势相反，即其中一个变量大于自身的期望值时另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。

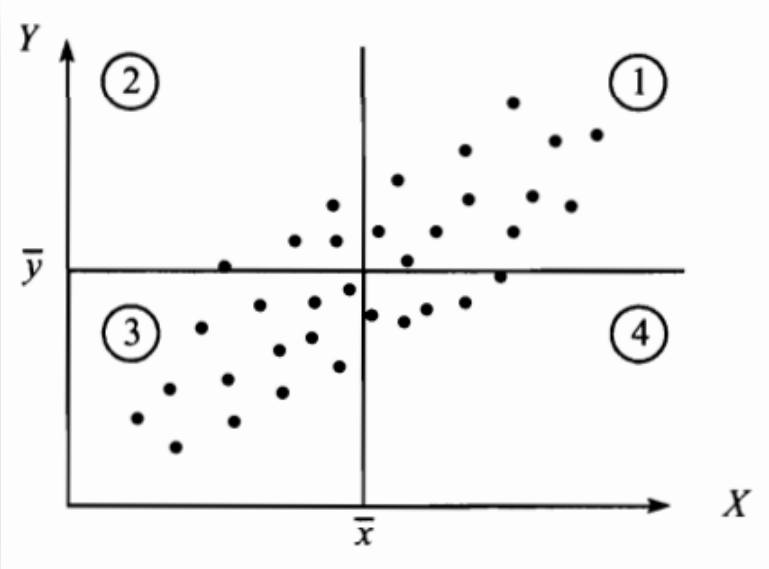
如果X与Y是统计独立的，那么二者之间的协方差就是0，因为两个独立的随机变量满足$E[XY]=E[X]E[Y]$。

协方差一般只能描述变化趋势，无法直观描述变化程度。

协方差

- 已知 n 组观测（X,Y）

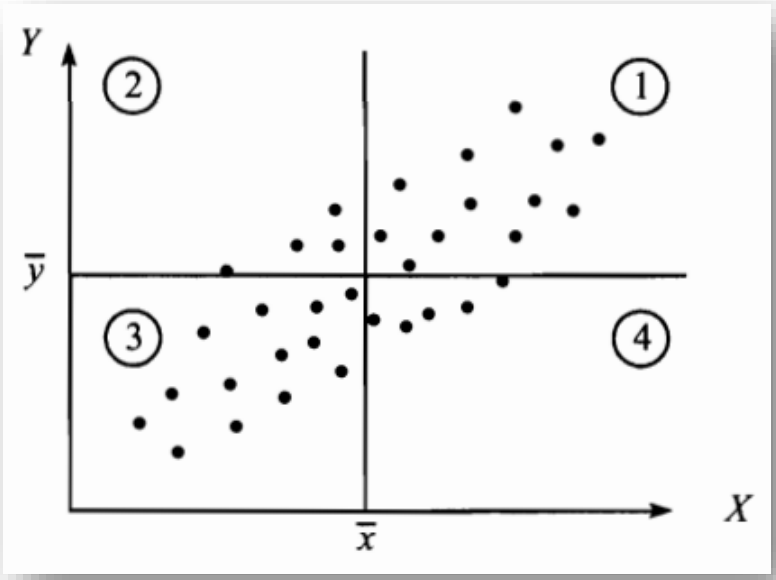
观察序号	自变量 X	因变量 Y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
n	x_n	y_n



- $\bar{x}$ x 的均值
- $\bar{y}$ y 的均值

协方差

象限	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	+	+	+
2	+	-	-
3	-	-	+
4	-	+	-



协方差

- Y 和 X 的协方差

$$\mathrm{Cov}(Y,X) = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{n-1}$$

- Cov(Y,X) > 0　　Y 和 X 正相关
- Cov(Y,X) < 0　　Y 和 X 负相关

- 受度量单位的影响（不能反映变量间线性关系的强弱）

相关系数

- Y 和 X 的相关系数
  - 经过标准化后的 Y 和 X 的协方差

$$\mathrm{Cor}(Y,X) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{s_y}\right)\left(\frac{x_i-\bar{x}}{s_x}\right)$$

相关系数

$$\mathrm{Cor}(Y,X)=\frac{\mathrm{Cov}(Y,X)}{s_y s_x}$$

$$=\frac{\sum(y_i-\overline{y})(x_i-\overline{x})}{\sqrt{\sum(y_i-\overline{y})^2\sum(x_i-\overline{x})^2}}$$

- Cov 协方差
- Cor 相关系数

相关系数

- 性质
  - $\mathrm{Cor}(Y,X) = \mathrm{Cor}(X,Y)$
  - $-1 \leq Cor(Y,X) \leq 1$
    - 不受单位影响
    - 值　　Y 和 X 之间线性关系的强度
    - 符号　　Y 和 X 之间线性关系的方向

相关系数的解读

- ① r 的取值范围为-1≤r≤+1
- ② |r|越接近1,表明相关关系越密切；越接近于0,相关关系就越不密切.
- ③ r=+1或r =-1,表明两现象完全相关
- ④ r=0,两变量无直线关系.
- ⑤ r>0现象呈正直线关系;r<0,现象呈负相关.
- ⑥在说明两个变量之间线性关系的密切程度时,根据经验可将相关程度分为以下几种情况：当|r|≥0.8时，视为高度相关：0.5≤|r|<0.8时，视为中度相关；0.3≤|r|<5时，视为低度相关；|r|<0.3时，说明两个变量之间的相关程度极弱，可视为不相关。但这种说明必须建立在相关系数通过显著性检验的基础之上。

各相关系数的区别

皮尔逊相关系数：一般用来计算两个连续型变量的相关系数。

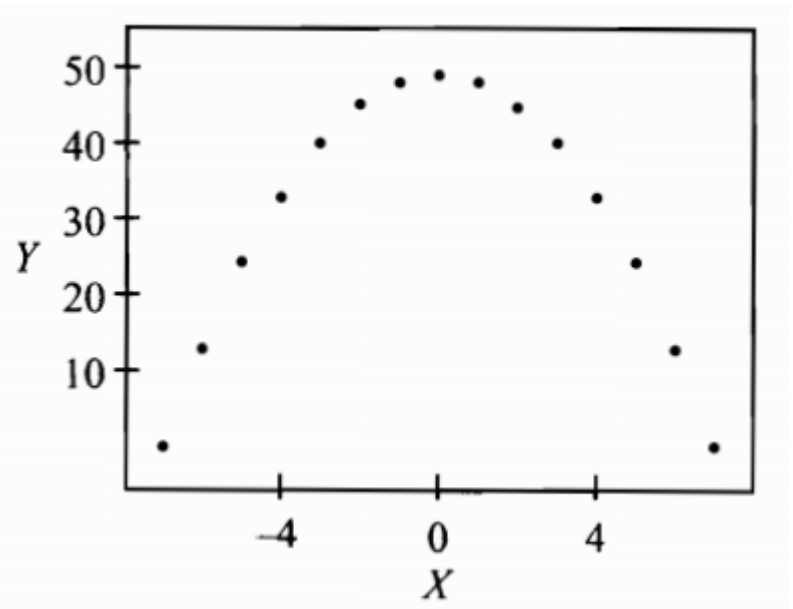
肯德尔相关系数：一个连续一个分类（最好是定序变量）

斯皮尔曼相关系数：2个变量无论连续还是分类都可以，但斯皮尔曼是非参数的，会损失信息，尽量不用

相关系数

- $\mathrm{Cor}(Y, X) = 0$
  - Y 和 X 之间没有线性相关性

Y	X	Y	X
1	-7	49	1
14	-6	46	2
25	-5	41	3
34	-4	34	4
41	-3	25	5
46	-2	14	6
49	-1	1	7
50	0		



- $\mathrm{Y} = 50 - x^2 \quad and \quad Cor(\mathrm{Y}\ ,\ \mathrm{X}) = 0$

相关系数

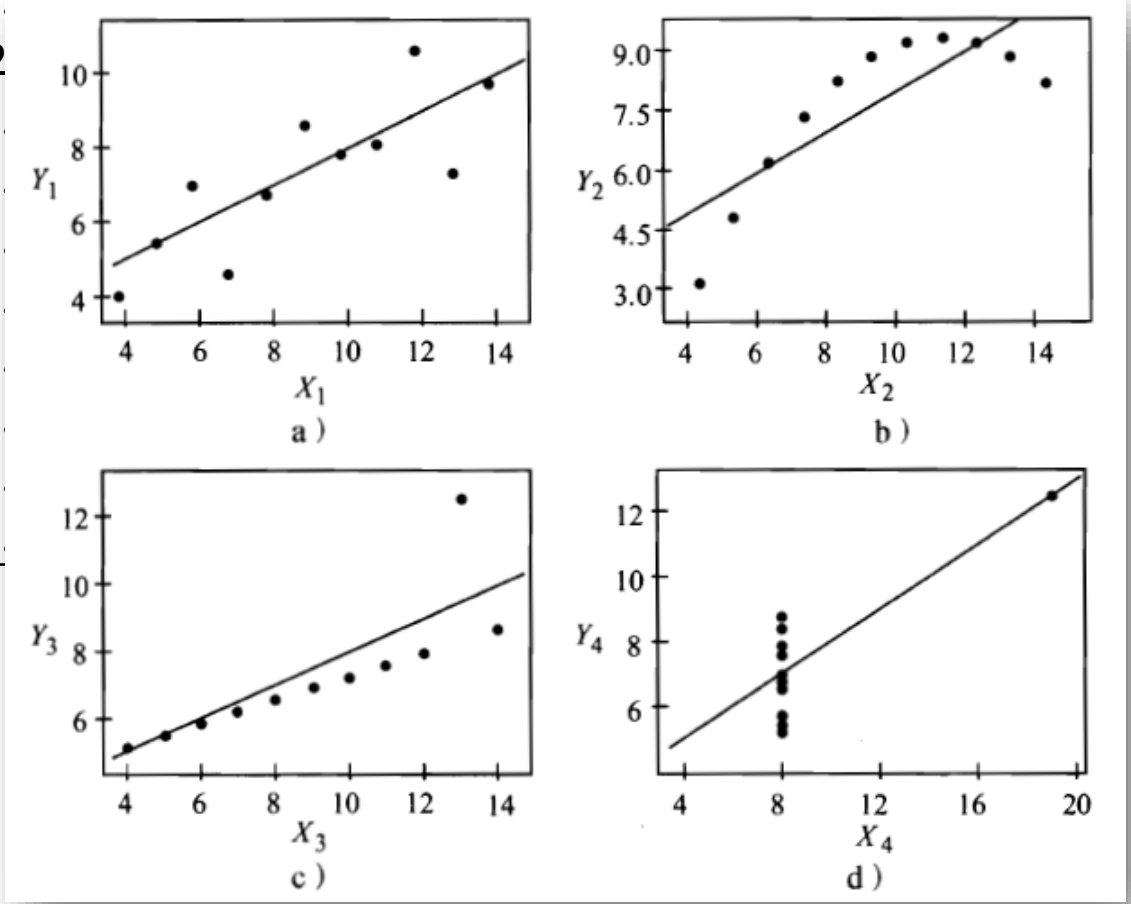
- 相关系数易受到离群值的影响

Y1	X1	Y2	X2	Y3	X3	Y4	X4
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.1	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.1	4	5.39	4	12.5	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

- $Cor(Y_1, X_1) = Cor(Y_2, X_2) = Cor(Y_3, X_3) = Cor(Y_4, X_4) \approx 0.8$

相关系数

Y1	X1	Y2	X2	Y3	X3	Y4	X4
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12			
8.81	9	8.77	9	7			
8.33	11	9.26	11	7			
9.96	14	8.1	14	8			
7.24	6	6.13	6	6			
4.26	4	3.1	4	5			
10.84	12	9.13	12	8			
4.82	7	7.26	7	6			
5.68	5	4.74	5	5			



协方差和相关系数

- 数据集：计算机维修时间
  - 维修时间（分钟）
  - 维修元件个数
- 维修时间与元件个数之间的关系
  - 计算协方差
  - 计算相关系数
  - 绘制散点图

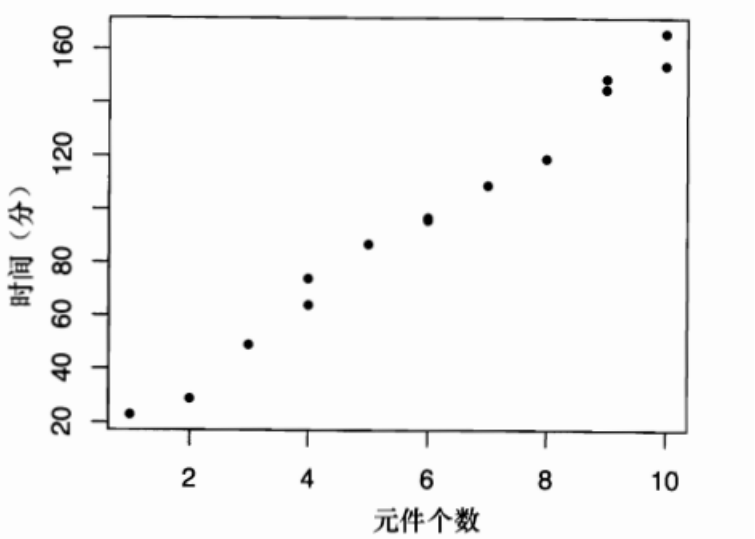
Minutes	Units
23	1
29	2
49	3
64	4
74	4
87	5
96	6
97	6
109	7
119	8
149	9
145	9
154	10
166	10

协方差和相关系数

i	y_i	x_i	$y_i-\overline{y}$	$x_i-\overline{x}$	$(y_i-\overline{y})^2$	$(x_i-\overline{x})^2$	$(y_i-\overline{y})(x_i-\overline{x})$
1	23	1	−74.21	−5	5 507.76	25	371.07
2	29	2	−68.21	−4	4 653.19	16	272.86
3	49	3	−48.21	−3	2 324.62	9	144.64
4	64	4	−33.21	−2	1 103.19	4	66.43
5	74	4	−23.21	−2	538.90	4	46.43
6	87	5	−10.21	−1	104.33	1	10.21
7	96	6	−1.21	0	1.47	0	0.00
8	97	6	−0.21	0	0.05	0	0.00
9	109	7	11.79	1	138.90	1	11.79
10	119	8	21.79	2	474.62	4	43.57
11	149	9	51.79	3	2 681.76	9	155.36
12	145	9	47.79	3	2 283.47	9	143.36
13	154	10	56.79	4	3 224.62	16	227.14
14	166	10	68.79	4	4 731.47	16	275.14
合计	1 361	84	0.00	0	27 768.36	114	1 768.00

协方差和相关系数

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{84}{14} = 6$$



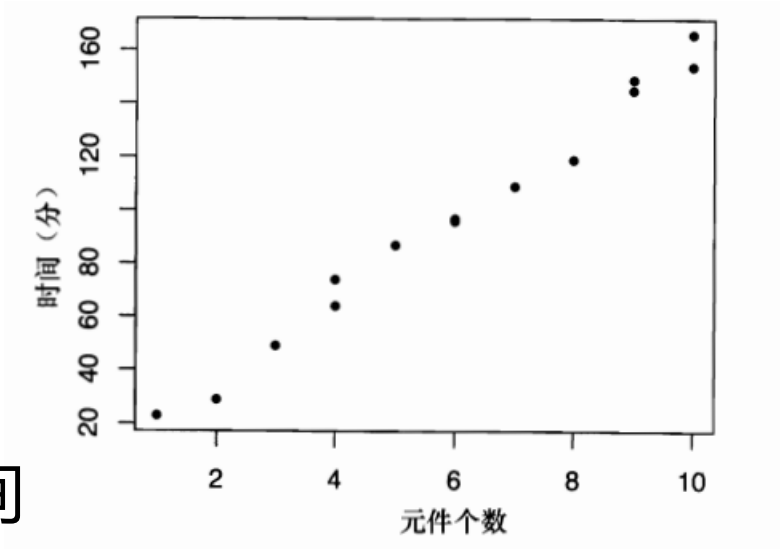
$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1\,361}{14} = 97.21$$

$$\mathrm{Cov}(Y,X) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n-1} = \frac{1\,768}{13} = 136$$

$$\mathrm{Cor}(Y,X) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}} = \frac{1\,768}{\sqrt{27\,768.36 \times 114}} = 0.996$$

协方差和相关系数

- 数据集：计算机维修时间



- $\mathrm{Cor}(Y, X) = 0.996$
  - 强正线性相关性
  - 无法根据元件数预测维修时间
  - 解决方法
    - 回归模型