第三章　相关分析

主要内容

1. 各种测度水平变量之间的相关
   - 列联分析——定类与定类
   - 皮尔逊相关——数值型变量之间
   - Spearman 和kendall相关——定序与定序

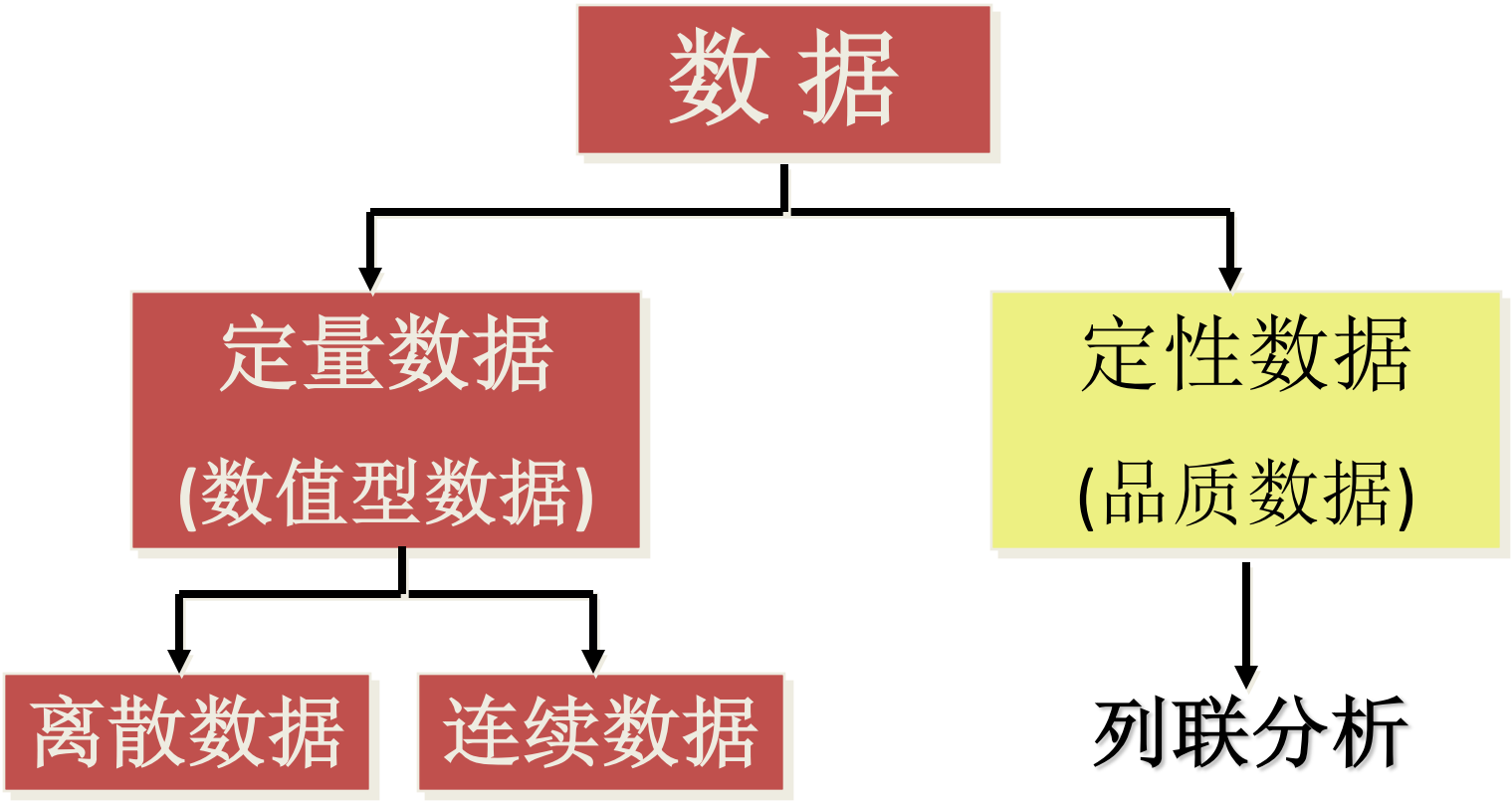
2. 偏相关分析

3. 典型相关分析（变量组对变量组，了解）

列联分析

1. 列联表的构造
2. 进行 χ^2 检验
   - 一致性检验
   - 独立性检验
3. 测度列联表中的相关性

数据的类型与列联分析



品质数据

1. 品质随机变量的结果表现为类别
   - 例如：性别 (男, 女)
2. 各类别用符号或数字代码来测度
3. 使用定类或定序尺度
   - 你吸烟吗?
     1.是；2.否
   - 你赞成还是反对这一改革方案?
     1.赞成；2.反对
4. 对品质数据的描述和分析通常使用列联表
   两个分类变量之间是否有关系。
5. 可使用 χ^2检验

列联表

1. 由两个以上的变量进行交叉分类的频数分布表
2. 行变量的类别用 r 表示， r_i 表示第 i 个类别
3. 列变量的类别用 c 表示， c_j 表示第 j 个类别
4. 每种组合的观察频数用 f_{ij} 表示
5. 表中列出了行变量和列变量的所有可能的组合，所以称为列联表
6. 一个 r 行 c 列的列联表称为 $r \times c$ 列联表

列联表的结构

一个2 × 2 列联表

列(c_j) / 行 (r_i)	列(c_j)		合计
	j =1	j =1	
i =1	f_{11}	f_{12}	$f_{11}+f_{12}$
i =2	f_{21}	f_{22}	$f_{21}+f_{22}$
合计	$f_{11}+f_{21}$	$f_{12}+f_{22}$	n

（$r \times c$ 列联表的一般表示）

r 行 c 列的列联表

列(c_j) / 行(r_i)	列(c_j)			合计
	$j=1$	$j=2$	…	
$i=1$	f_{11}	f_{12}	…	r_1
$i=2$	f_{21}	f_{22}	…	r_2
:	:	:	:	:
合计	c_1	c_2	…	n

f_{ij} 表示第 i 行第 j 列的观察频数

列联表实例

	老年	中年	青年
戏曲	*20*	*10*	2
歌舞	*5*	*20*	35
球赛	2	10	20

列联表实例2

【例】一个集团公司在四个不同的地区设有分公司，现该集团公司欲进行一项改革，此项改革可能涉及到各分公司的利益，故采用抽样调查方式，从四个分公司共抽取420个样本单位(人)，了解职工对此项改革的看法，调查结果如下表

	一分公司	二分公司	三分公司	四分公司	合计
赞成该方案	68	75	57	79	279
反对该方案	32	75	33	31	141
合计	100	120	90	110	420

观察值的分布（频数）

联合分布

1. 边缘分布
   - 行边缘分布

     行观察值的合计数的分布

     例如，赞成改革方案的共有279人，反对改革方案的141人
   - 列边缘分布

     列观察值的合计数的分布

     例如，四个分公司接受调查的人数分别为100人，120人，90人，110人
2. 条件频数
   - 变量 X 条件下变量 Y 的分布，或在变量 Y 条件下变量 X 的分布
   - 每个具体的观察值称为条件频数

观察值的分布

条件频数

行边缘分布

	一分公司	二分公司	三分公司	四分公司	合计
赞成该方案	68	75	57	79	279
反对该方案	32	75	33	31	141
合计	100	120	90	110	420

列边缘分布

百分比分布（概率）

1. 条件频数反映了数据的分布，但不适合进行对比
2. 为在相同的基数上进行比较，可以计算相应的百分比，称为百分比分布（概率分布）
   - 行百分比：行的每一个观察频数除以相应的行合计数（f_{ij}/r_i）
   - 列百分比：列的每一个观察频数除以相应的列合计数（f_{ij}/c_j）
   - 总百分比：每一个观察值除以观察值的总个数（f_{ij}/n）

百分比分布

行百分比　　列百分比　　总百分比

	一分公司	二分公司	三分公司	四分公司	合计
赞成该方案	24.4%	26.9%	20.4%	28.3%	66.4%
	68.0%	62.5%	63.35	71.8%	—
	16.2%	17.8%	13.6%	18.8%	—
反对该方案	22.7%	31.9%	23.4%	22.0%	33.6%
	32.0%	37.5%	36.7%	28.2%	—
	7.6%	10.7%	7.9%	7.4%	—
合计	23.8%	28.6%	21.4%	26.2%	100%

期望频数的分布

1. 假定行变量和列变量是独立的
2. 一个实际频数 f_{ij} 的期望频数 e_{ij}，是总频数的个数 n 乘以该实际频数 f_{ij} 落入第 i 行和第j列的概率，即

$$e_{ij} = n \cdot \left(\frac{r_i}{n} \right) \cdot \left(\frac{c_i}{n} \right) = \frac{r_i c_j}{n}$$

期望频数的分布

- 例如，第1行和第1列的实际频数为 f_{11} ,它落在第1行的概率估计值为该行的频数之和r_1除以总频数的个数 n ，即：r_1/n；它落在第1列的概率的估计值为该列的频数之和c_1除以总频数的个数 n ，即：c_1/n 。根据概率的乘法公式，该频数落在第1行和第1列的概率应为

$$\left(\frac{r_1}{n}\right)\cdot\left(\frac{c_1}{n}\right)$$

由于观察频数的总数为n ，所以f_{11} 的期望频数 e_{11} 应为

$$e_{11}=n\cdot\left(\frac{r_1}{n}\right)\cdot\left(\frac{c_1}{n}\right)=\frac{r_1c_1}{n}=\frac{279\times100}{420}=66.43\approx66$$

期望频数的分布

- 根据上述公式计算的前例的期望频数

		一分公司	二分公司	三分公司	四分公司
赞成该方案	实际频数	68	75	57	79
	期望频数	66	80	60	73
反对该方案	实际频数	32	75	33	31
	期望频数	34	40	30	37

χ^2 统计量

1. 用于检验列联表中变量之间是否存在显著性差异，或者用于检验变量之间是否独立
2. 计算公式为

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$$

其自由度为(r-1)(c-1)

式中：f_{ij}—列联表中第i行第j列类别的实际频数

e_{ij}—列联表中第i行第j列类别的期望频数

χ^2 统计量

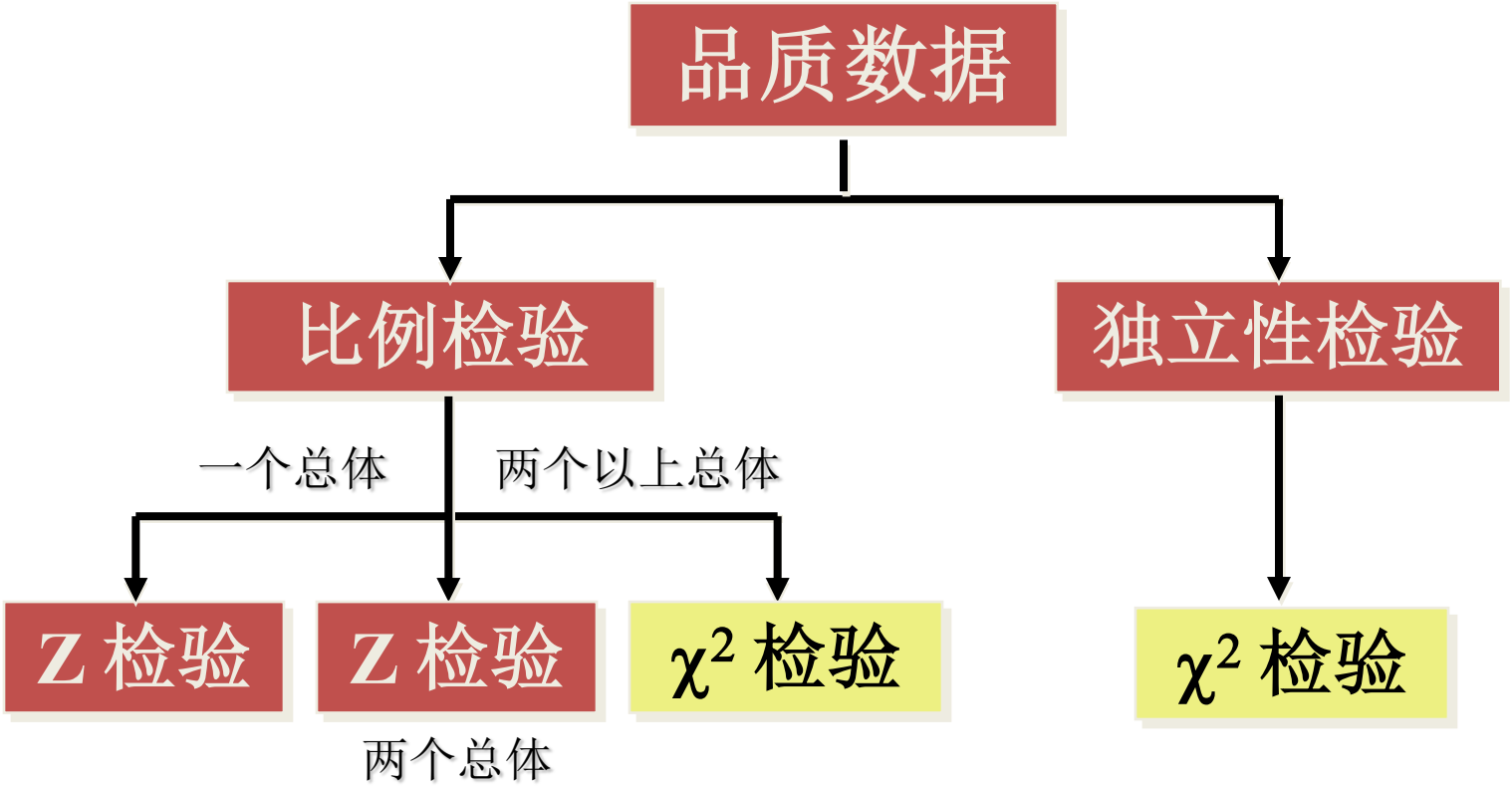
实际频数 (f_{ij})	期望频数 (e_{ij})	$f_{ij} - e_{ij}$	$(f_{ij} - e_{ij})^2$	$\frac{(f_{ij} - e_{ij})^2}{f}$
68	66	2	4	0.0606
75	80	-5	25	0.3125
57	60	-3	9	0.1500
79	73	6	36	0.4932
32	34	-2	4	0.1176
45	40	5	25	0.6250
33	30	3	9	0.3000
31	37	-6	36	0.9730

合计：**3.0319**

$$\chi^2 = \sum \frac{(f-e)^2}{e}$$

$$= 3.0319$$

品质数据的假设检验



一致性检验

1. 检验列联表中目标变量之间是否存在显著性差异
2. 检验的步骤为
   - 提出假设
     - H_0：$P_1 = P_2 = \ldots = P_j$ (目标变量的各个比例一致)
     - H_1：P_1，P_2，…，P_j 不全相等 (各个比例不一致)
   - 计算检验的统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

   进行决策
   - 根据显著性水平α和自由度$(r-1)(c-1)$查出临界值χ_α^2
   - 若$\chi^2 \geq \chi_\alpha^2$，拒绝$H_0$；若$\chi^2 < \chi_\alpha^2$，接受$H_0$

一致性检验

【例】续前例，检验职工的态度是否与所在单位有关？(α=0.1)

1. 提出假设
   - H_0：$P_1 = P_2 = P_2 = P_4$ (赞成比例一致)
   - H_1：P_1，P_2，P_3，P_4不全相等 (赞成比例不一致)
2. 计算检验的统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij} - e_{ij})^2}{e_{ij}} = 3.0319$$

根据显著性水平α＝0.1和自由度(2-1)(4-1)=3查出相应的临界值χ_α^2=6.251。由于χ^2=3.0319<χ_α^2=6.251，接受H_0

独立性检验

1. 检验列联表中的行变量与列变量之间是否独立
2. 检验的步骤为
   - 提出假设
     - H_0：行变量与列变量独立
     - H_1：行变量与列变量不独立
   - 计算检验的统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$$

   - 进行决策
     - 根据显著性水平α和自由度$(r-1)(c-1)$查出临界值χ_α^2
     - 若$\chi^2 \geq \chi_\alpha^2$，拒绝$H_0$；若$\chi^2 < \chi_\alpha^2$，接受$H_0$

独立性检验

【例】一种原料来自三个不同的地区，原料质量被分成三个不同等级。从这批原料中随机抽取500件进行检验，结果如下表。检验各地区与原料之间是否存在依赖关系（$\alpha=0.05$)

地区	一级	二级	三级	合计
甲地区	52	64	24	140
乙地区	60	59	52	171
丙地区	50	65	74	189
合计	162	188	150	500

独立性检验

1. 提出假设
   - H_0：地区与原料等级之间独立
   - H_1：地区与原料等级之间不独立
2. 计算检验的统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = 19.82$$

3. 根据显著性水平$\alpha=0.05$和自由度(3-1)(3-1)=4查出相应的临界值$\chi_\alpha^2=9.488$。由于$\chi^2=19.82>\chi_\alpha^2=9.448$，拒绝$H_0$

- 列联表通过频次（不是相对频次）来进行比较的。

- 性别（血型）与新冠肺炎
- 父辈职业是否与子辈职业有关？
- 饮食习惯与地区
- 年龄与对电影的评价
- 年级与求职意愿
- 性别与对音乐的偏好
- ……

列联表中的相关测量

1. 品质相关
   - 对品质数据(定类和定序数据)之间相关程度的测度
2. 列联表变量的相关属于品质相关
3. 列联表相关测量的指标主要有
   - φ 相关系数
   - 列联相关系数
   - V 相关系数

φ 相关系数

1. 测度 2×2列联表中数据相关程度的一个量
2. 对于2×2 列联表，φ系数的值在0～1之间
3. φ相关系数计算公式为

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

式中：$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$

n为实际频数的总个数，即样本容量

φ 相关系数
（原理分析）

一个简化的 2×2 列联表

因素 Y	因素 X		合计
	x_1	x_2	
y_1	a	b	$a+b$
y_2	c	d	$c+d$
合计	$a+c$	$b+d$	n

φ 相关系数
（原理分析）

1. 列联表中每个单元格的期望频数分别为

$$e_{11}=\frac{(a+b)(a+c)}{n} \qquad e_{21}=\frac{(a+c)(c+d)}{n}$$

$$e_{12}=\frac{(a+b)(b+d)}{n} \qquad e_{22}=\frac{(b+d)(c+d)}{n}$$

2. 将各期望频数代入 χ^2 的计算公式得

$$\chi^2=\frac{(a-e_{11})^2}{e_{11}}+\frac{(b-e_{12})^2}{e_{12}}+\frac{(c-e_{21})^2}{e_{21}}+\frac{(d-e_{22})^2}{e_{22}}$$

$$=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

φ 相关系数
（原理分析）

3. 将χ^2入φ相关系数的计算公式得

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

- ad 等于 bc， $\varphi = 0$，表明变量X 与 Y 之间独立
- 若 b=0 ，c=0，或a=0 ，d=0，意味着各观察频数全部落在对角线上，此时$|\varphi|$ =1, 表明变量X 与 Y 之间完全相关

4. 列联表中变量的位置可以互换，φ的符号没有实际意义，故取绝对值即可

列联相关系数

1. 用于测度大于2×2列联表中数据的相关程度
2. 计算公式为

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}$$

- C 的取值范围是 $0\le C<1$
- $C=0$表明列联表中的两个变量独立
- C 的数值大小取决于列联表的行数和列数，并随行数和列数的增大而增大
- 根据不同行和列的列联表计算的列联系数不便于比较

V 相关系数
（要点）

1. 计算公式为

$$V=\sqrt{\frac{\chi^2}{n\min\left[(r-1),(c-1)\right]}}$$

式中 $\min\left[(r-1),(c-1)\right]$ 表示取 $(r-1),(c-1)$ 中较小的一个

2. V 的取值范围是 $0\le V\le 1$
3. $V=0$ 表明列联表中的两个变量独立
4. $V=1$ 表明列联表中的两个变量完全相关
5. 不同行和列的列联表计算的列联系数不便于比较
6. 当列联表中有一维为2，$\min[(r\text{-}1),(c\text{-}1)]=1$，此时 $V=\varphi$

φ、C、V的比较

1. 同一个列联表，φ、C、V的结果会不同
2. 不同的列联表，φ、C、V的结果也不同
3. 在对不同列联表变量之间的相关程度进行比较时，不同列联表中的行与行、列与列的个数要相同，并且采用同一种系数

列联表中的相关测量

【例】一种原料来自三个不同地区，原料质量被分成三个不同等级。从这批原料中随机抽取500件进行检验，结果如下表。分别计算φ系数、C系数和V系数，并分析相关程度

地区	一级	二级	三级	合计
甲地区	52	64	24	140
乙地区	60	59	52	171
丙地区	50	65	74	189
合计	162	188	150	500

列联表中的相关测量

解： 已知n=500，根据前面的计算χ^2＝19.82，列联表为3×3

$$\varphi=\sqrt{\frac{\chi^2}{n}}=\sqrt{\frac{19.82}{500}}=0.199$$

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}=\sqrt{\frac{19.82}{19.82+500}}=0.195$$

$$V=\sqrt{\frac{\chi^2}{n\min\left[(r-1),(c-1)\right]}}=\sqrt{\frac{19.82}{500(2)}}=0.141$$

结论： 三个系数均不高，表明产地和原料等级之间的相关程度不高

练习

- 改革方案的调查结果

	一分公司	二分公司	三分公司	四分公司	合计
赞成	68	75	57	79	279
反对	32	45	33	31	141
合计	100	120	90	110	420

- 吸烟与患慢性气管炎的关系

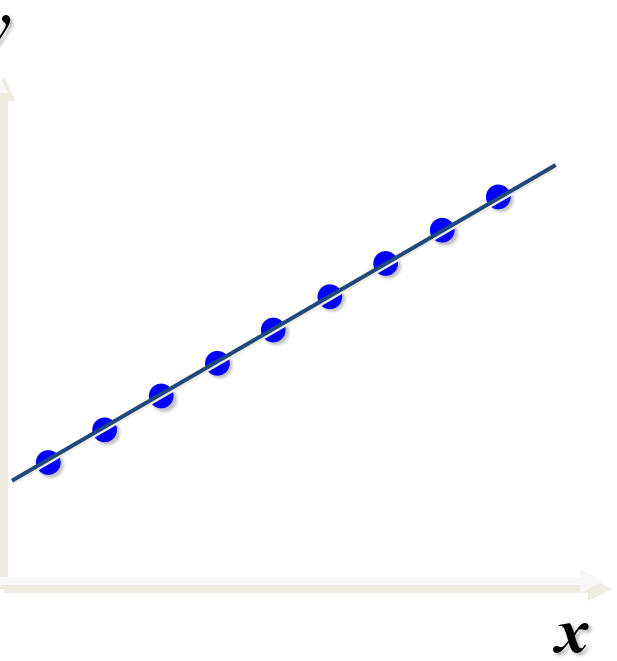
	患病	未患病
吸烟	43	162
不吸烟	13	121

皮尔逊相关系数

变量相关的概念
相关系数及其计算

变量间的关系
（函数关系）

1. 是一一对应的确定关系
2. 设有两个变量 x 和 y，变量 y 随变量 x 一起变化，并完全依赖于 x，当变量 x 取某个数值时，y 依确定的关系取相应的值，则称 y 是 x 的函数，记为 $\boldsymbol{y = f(x)}$，其中 x 称为自变量，y 称为因变量
3. 各观测点落在一条线上

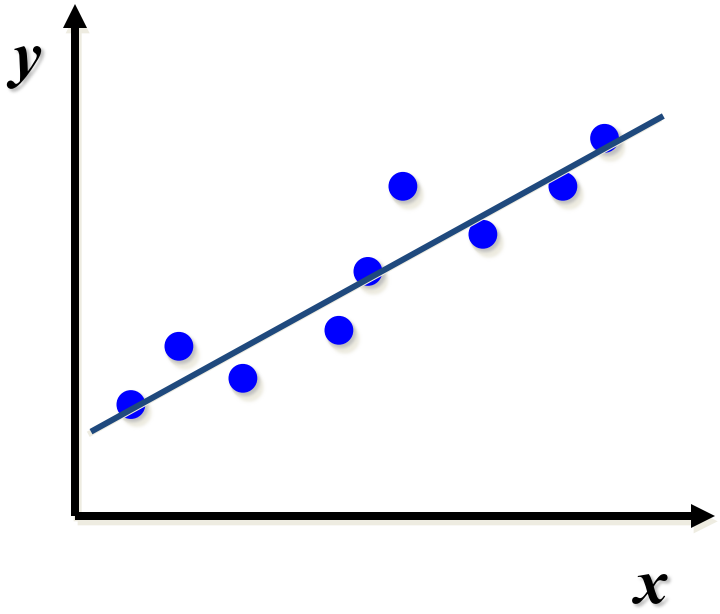


变量间的关系
（函数关系）

- 函数关系的例子
  - 某种商品的销售额(y)与销售量(x)之间的关系可表示为 $y = p\,x$ (p 为单价)
  - 圆的面积(S)与半径之间的关系可表示为 $S = \pi R^2$
  - 企业的原材料消耗额(y)与产量(x_1) 、单位产量消耗(x_2) 、原材料价格(x_3)之间的关系可表示为 $y = x_1 x_2 x_3$

变量间的关系
（相关关系）

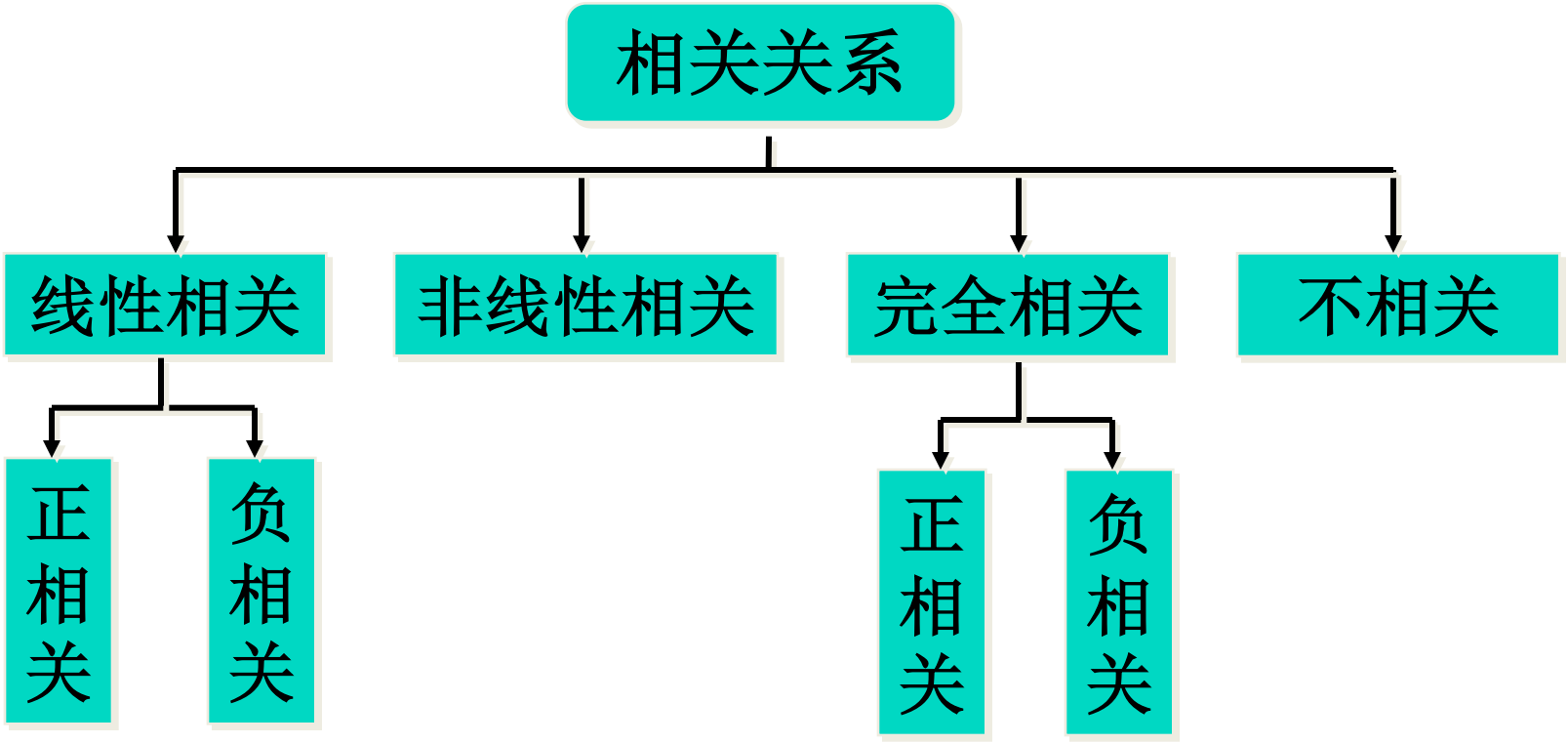
1. 变量间关系不能用函数关系精确表达
2. 一个变量的取值不能由另一个变量唯一确定
3. 当变量 x 取某个值时，变量 y 的取值可能有几个
4. 各观测点分布在直线周围



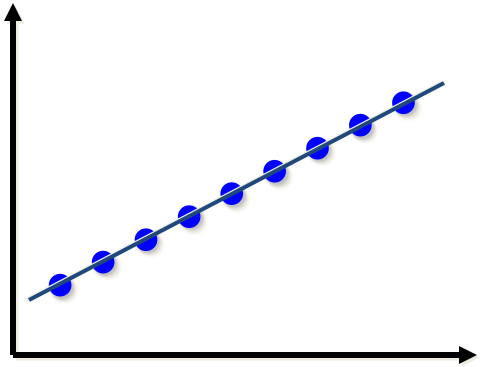
变量间的关系
（相关关系）

- 相关关系的例子
  - 商品的消费量(y)与居民收入(x)之间的关系
  - 商品销售额(y)与广告费支出(x)之间的关系
  - 粮食亩产量(y)与施肥量(x_1) 、降雨量(x_2) 、温度(x_3)之间的关系
  - 收入水平(y)与受教育程度(x)之间的关系
  - 父亲身高(y)与子女身高(x)之间的关系

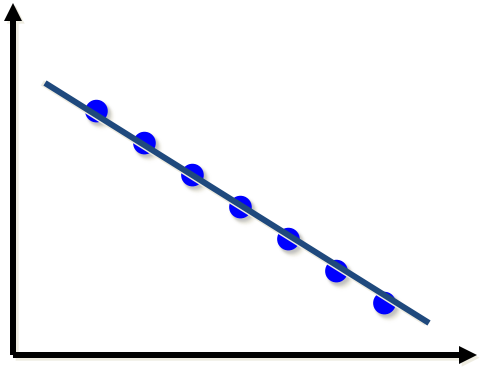
相关关系的类型



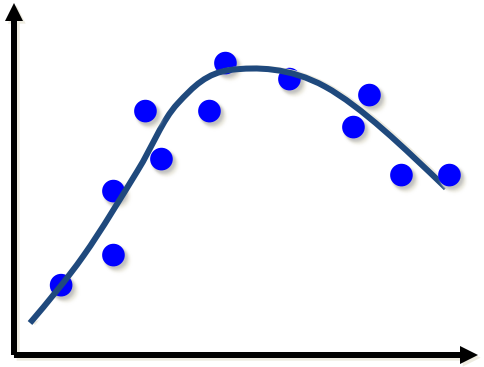
相关关系的图示



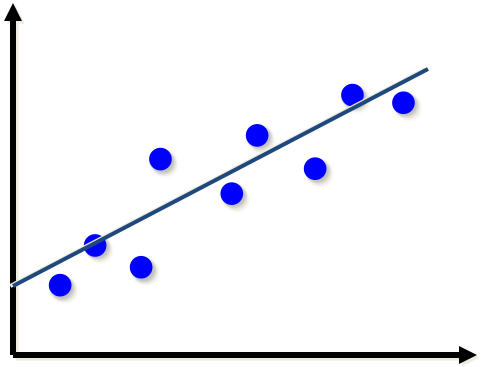
完全正线性相关



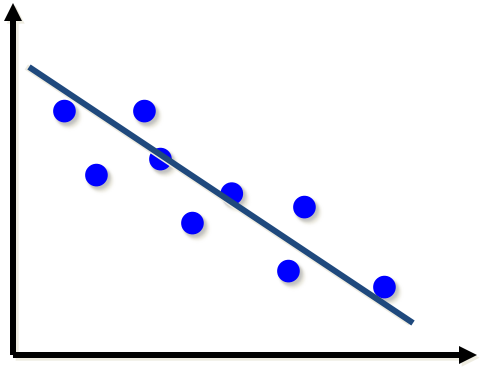
完全负线性相关



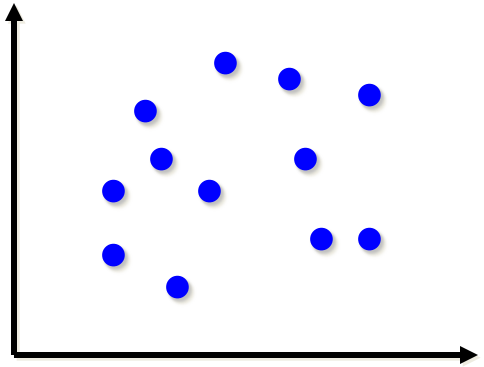
非线性相关



正线性相关



负线性相关



不相关

相关系数及其计算

相关关系的测度
（相关系数）

1. 对变量之间关系密切程度的度量
2. 对两个变量之间线性相关程度的度量称为简单相关系数
3. 若相关系数是根据总体全部数据计算的，称为总体相关系数，记为ρ
4. 若是根据样本数据计算的，则称为样本相关系数，记为 r

相关关系的测度

（相关系数：皮尔逊相关系数）

- 样本相关系数的计算公式

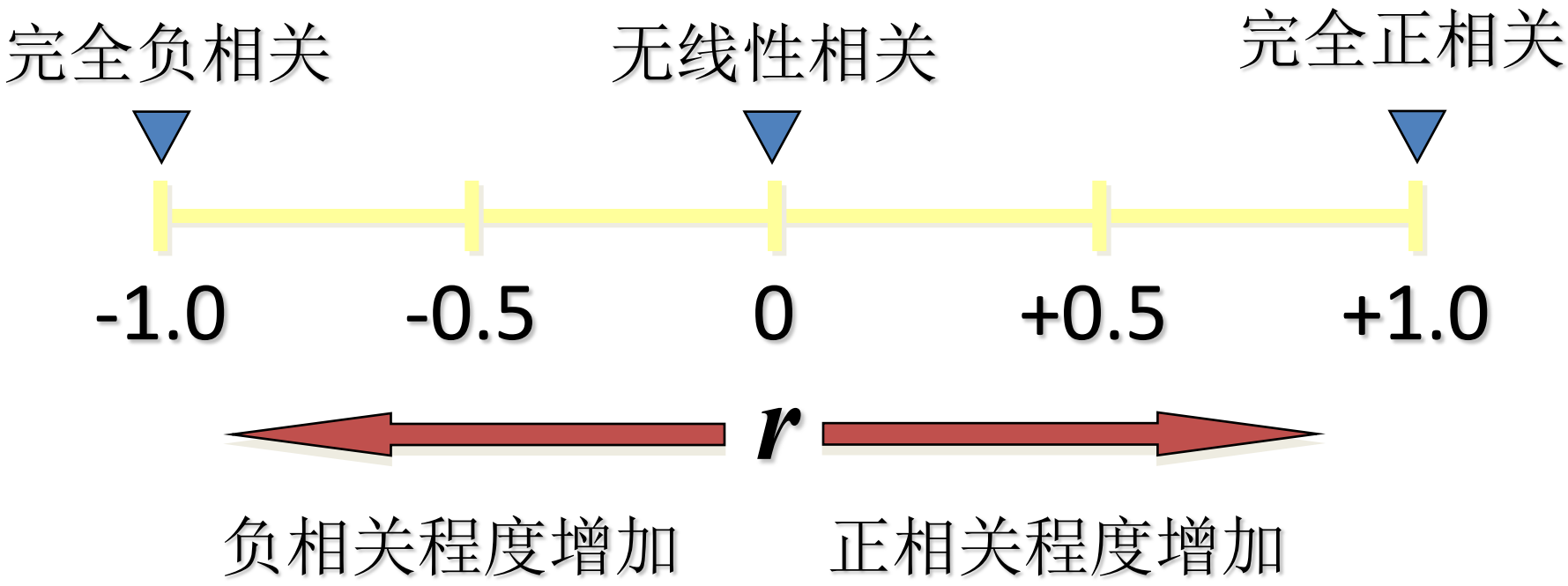
$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\cdot\sum(y-\bar{y})^2}}$$

或化简为 $$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\cdot\sqrt{n\sum y^2-\left(\sum y\right)^2}}$$

相关关系的测度

1. r 的取值范围是 [-1,1]
2. $|r|=1$，为完全相关
   - $r=1$，为完全正相关
   - $r=-1$，为完全负正相关
3. $r=0$，不存在线性相关关系
4. $-1\le r<0$，为负相关
5. $0<r\le 1$，为正相关
6. $|r|$越趋于1表示关系越密切；$|r|$越趋于0表示关系越不密切

相关关系的测度



相关关系的测度

【例1】在研究我国人均消费水平的问题中，把全国人均消费额记为y，把人均国民收入记为x。我们收集到1981～1993年的样本数据(x_i，y_i)，i=1,2,…，13，数据见表1，计算相关系数。

表1 我国人均国民收入与人均消费金额数据 单位:元

年份	人均国民收入	人均消费金额	年份	人均国民收入	人均消费金额
1981	393.8	249	1988	1068.8	643
1982	419.14	267	1989	1169.2	690
1983	460.86	289	1990	1250.7	713
1984	544.11	329	1991	1429.5	803
1985	668.29	406	1992	1725.9	947
1986	737.73	451	1993	2099.5	1148
1987	859.97	513			

（计算结果）

解：根据样本相关系数的计算公式有

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\cdot\sqrt{n\sum y^2-\left(\sum y\right)^2}}$$

$$=\frac{13\times 9156173.99-12827.5\times 7457}{\sqrt{13\times 16073323.77-\left(12827.5\right)^2}\cdot\sqrt{13\times 5226399-\left(7457\right)^2}}$$

$$=0.9987$$

人均国民收入与人均消费金额之间的相关系数为 **0.9987**

相关系数的显著性检验

1. 检验两个变量之间是否存在线性相关关系
2. 等价于对回归系数 β_1的检验
3. 采用 t 检验
4. 检验的步骤为
   - 提出假设：H_0：$\rho=0$；H_1：$\rho\neq 0$
   - 计算检验的统计量：$t=\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}\sim t(n-2)$
   - 确定显著性水平α，并作出决策
     - 若$|t|>t_{\alpha/2}$，拒绝H_0
     - 若$|t|<t_{\alpha/2}$，接受H_0

（实例）

- 对前例计算的相关系数进行显著性检(α=0.05)

1. 提出假设：H_0：$\rho = 0$；H_1：$\rho \neq 0$
2. 计算检验的统计量

$$t = \frac{0.9987\sqrt{13-2}}{\sqrt{1-0.9987^2}} = 64.9809$$

3. 根据显著性水平α=0.05，查t分布表得$t_{\alpha/2}(n-2)$=2.201
   - 由于$|t|$=64.9809>$t_{\alpha/2}$(13-2)=2.201，拒绝H_0，人均消费金额与人均国民收入之间的相关关系显著

（相关系数检验表的使用）

1. 若|*r*|大于表上的α=5%相应的值，小于表上α＝1%相应的值，称变量*x*与*y*之间有**显著**的线性关系
2. 若|*r*|大于表上α=1%相应的值，称变量*x*与*y*之间有**十分显著**的线性关系
3. 若|*r*|小于表上α=5%相应的值，称变量*x*与*y*之间没有**明显**的线性关系
4. 根据前例的$r=0.9987>\alpha=5\%(n-2)=0.553$，表明人均消费金额与人均国民收入之间有十分显著的线性相关关系

练习

- 皮尔逊相关系数的计算

等级相关（定序-定序）
(非参数检验的内容)

- 斯皮尔曼等级相关
- Kendall秩相关

一个例子

夫妻双方的家庭社会经济地位 是否 “门当户对”

例子：（妻子，丈夫）（1，2）、（2，3）、（3，4）、（4，5）、（5，1）

其他实例

- 评判员对参赛人名次的打分（是否相近）
- 学生活动能力与智商
- 婚姻美满与文化程度
- 主观指标与客观指标度量生活质量
- 交卷子的名次与考试成绩

交卷名次	1	2	3	4	5	6	7	8	9	10	11	12
考试成绩	90	74	74	60	68	86	92	60	78	74	78	64

- 生育率与受教育水平

…

非参数检验

秩相关及其检验

Spearman秩相关及其检验

Kendall秩相关及其检验

Spearman秩相关检验

- 对两个顺序变量之间相关程度的一种度量
- Spearman秩相关系数也称等级相关系数，记为r_s，计算公式为

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

  - r_s的取值范围为[-1,1]
  - r_s=1，两种排序之间完全相关；若-1<r_s<0，两种排序之间为负相关；若0<r_s<1，两种排序之间为正相关；若r_s=0，两种排序之间不相关
  - r_s越趋于1，相关程度越高；越趋于0，相关程度越低

Spearman秩相关检验

【例】在一项关于职业声望和可信赖程度的调查中，列举了12种职业，要求被调查者分别按声望高低和值得信赖程度进行排序，调查数据如表计算两种排序之间的Spearman秩相关系数，并进行检验。(α=0.01)

职业	声望排序	信赖程度排序
科学家	1	1
医生	2	2
工程师	6	4
政府官员	3	7
中小学教师	4	3
大学教师	5	5
新闻记者	7	8
律师	8	6
企业管理人员	9	12
银行管理人员	10	10
建筑设计人员	11	9
会计师	12	11

SPSS的输出结果

Correlations

			声望排序	信赖程度排序
Spearman's rho	声望排序	Correlation Coefficient	1.000	.860**
		Sig. (2-tailed)	.	.000
		N	12	12
	信赖程度排序	Correlation Coefficient	.860**	1.000
		Sig. (2-tailed)	.000	.
		N	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

☺ Spearman秩相关系数为0.860，两种排序之间有比较高的正相关，即职业声望越高，值得信赖的程度也就越高。双尾检验的P=0.000，拒绝原假设，表明声望排序与信赖程度排序之间存在显著的相关关系

Kendall秩相关检验

- 对两个序变量之间相关程度的一种度量
- Kendall秩相关系数记为τ，计算公式为

$$\tau = \frac{4U}{n(n-1)} - 1 \quad \text{或} \quad \tau = 1 - \frac{4V}{n(n-1)}$$

*U*表示 *y* 的一致对数目，*V*表示 *y* 的非一致对数目

同序对（变化方向相同）、异序对（变化方向相反）

- τ 的取值范围为[-1,1]
- 若τ ＝1，表明两组秩之间完全正相关
- 若τ ＝-1，表明两组秩之间完全正相关

SPSS的输出结果

Correlations

			声望排序	信赖程度排序
Kendall's tau_b	声望排序	Correlation Coefficient	1.000	.697**
		Sig. (2-tailed)	.	.002
		N	12	12
	信赖程度排序	Correlation Coefficient	.697**	1.000
		Sig. (2-tailed)	.002	.
		N	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

☺ 秩相关系数τ =0.679，两种排序之间有比较高的正相关。双尾检验的P=0.002，拒绝原假设，表明声望排序与信赖程度排序之间存在显著的相关关系

练习

秩相关系数的计算

偏相关分析

多个变量

- 偏相关分析：当控制了一个或几个另外的变量的影响条件下两个变量之间的相关性

例：控制年龄和工作经验两个变量的影响来估计工资收入与受教育程度之间的相关

计算偏相关系数的方法

1 从线性回归的角度计算变量间的偏相关系数，但是这样做很麻烦。

2 迭代法，可以认为简单相关系数为0阶偏相关系数，任何n阶偏相关都可以通过3个(n-1)阶偏相关系数计算出来。

3 相关矩阵求逆法，即首先计算出所有变量的相关性矩阵，然后求它的逆矩阵。这样可以求出任何两两变量之间的偏相关系数。

- 偏相关系数的检验（略）

练习

- 偏相关系数的计算

典型相关分析（了解）

典型相关分析

- 典型相关是一组x对一组y的相关。
- 模型：

$a_1y_1+a_2y_2+\ldots+a_ky_k=b_0+b_1x_1+b_2x_2+\ldots+b_px_p$

- 它们的关系不只一个，即：有多组系数ai,bj使方程成立。
- 组数=min{k,p}（精简信息）

典型相关变量

- 设：X=(x1,x2,…,xp)’

  Y=(y1,y2,…,yk)’
- 对X和Y可做因子分析：公因子是ui和vi。

  ui=ai1x1+ai2x2+…+aipxp=a’x

  vi=bi1y1+bi2y2+…+bikyk=b’y
- X与Y相关，可推出a’x与b’y相关。

求ai’和bi’

z(a1’x,b1’y)=max z(a’x,b’y)

var(a’x)=1

var(b’y)=1

这是一个LP model 可解。

a1’x和b1’y是x,y的第一对典型相关。

求出第一对后，还可求出第二、第三…，但各对之间不相关。

检验

- 每对典型变量的相关系数的绝对值是否显著地大于0？是，这对因子就有代表性。反之，就删除，以减少工作量。就可以通过少量典型变量的研究，代替原来两组变量间相关关系的研究。
- 数据经标准化，可用相关矩阵，简单。

结束