南京大学
NANJING UNIVERSITY

第16章 相关分析

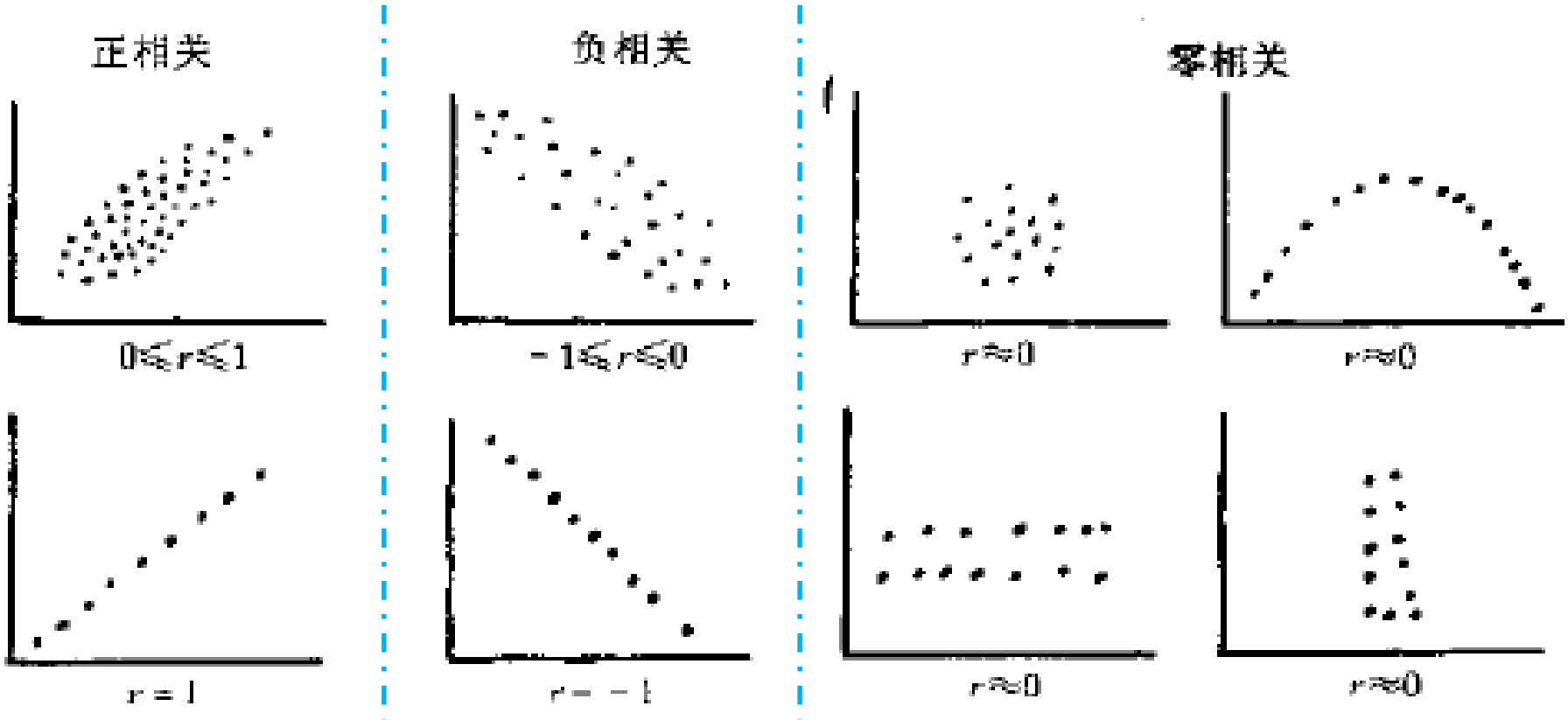
学习目标

- 了解相关分析的一些指标
- 能用SPSS做相关分析

主要内容

- 16.1 相关分析简介
- 16.2 简单相关分析
- 16.3 偏相关分析
- 16.4 距离计算

16.1 相关分析简介


正相关
负相关
零相关
0≤r≤1
-1≤r≤0
r≈0
r≈0
r=1
r=-1
r≈0
r≈0

相关分析的指标体系

- 连续变量的相关指标
  - Pearson相关系数（积差相关系数）：只适用于两变量呈线性相关时，表示相关性的大小，其数值介于-1～1之间。
  - Spearman等级相关系数：两变量非线性相关时使用，表示相关性的大小，其数值介于-1～1之间。

相关分析的指标体系

- 有序变量的相关指标
  - Gamma统计量：描述有序分类数据的联系强度，取值介于-1～1之间。
  - Kendall’s Tau-b系数：描述有序分类数据的联系强度，取值介于-1～1之间。
  - Kendall’s Tau-c系数：描述有序分类数据的联系强度，取值介于-1～1之间。
  - Somers’ D(C|R)系数：描述有序分类数据的联系强度，取值介于-1～1之间。

相关分析的指标体系

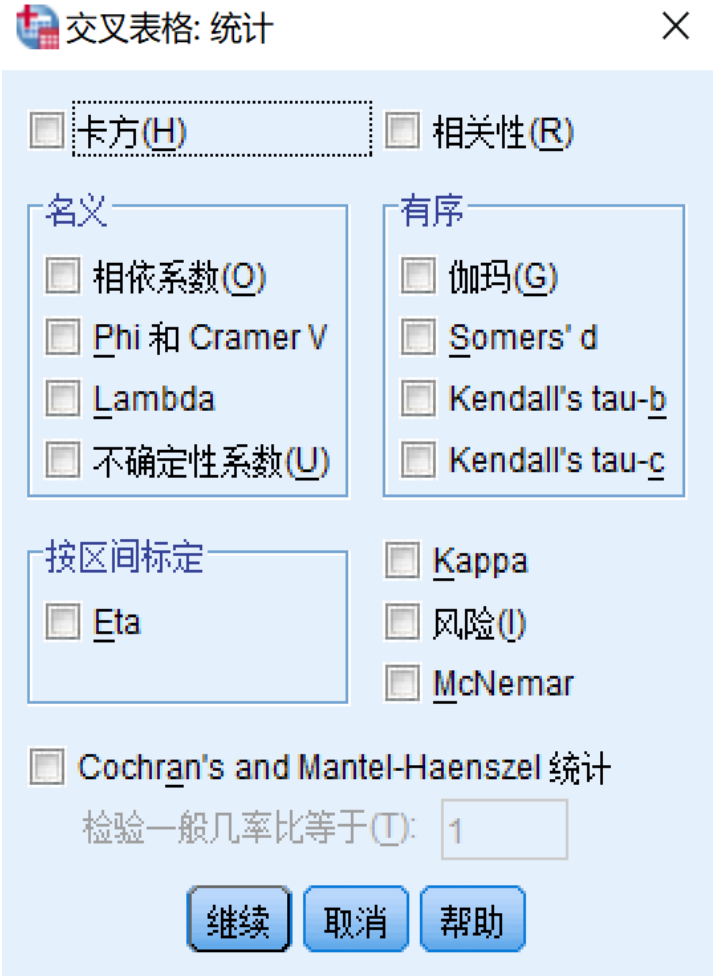
- 名义变量的相关指标
  - 列联系数：取值介于0～1之间，取值越大表明两变量间相关性越强。
  - Cramer’s V系数：取值介于0～1之间，取值越大表明两变量间相关性越强。
  - λ系数：用于反映自变量对因变量的预测效果，取值介于0～1之间。值为1时表明知道了自变量就可以完全确定因变量取值，值为0时表明自变量对因变量完全无预测作用。
  - 不确定系数：取值介于0～1之间，与λ系数类似。

相关分析的指标体系

- 其他特殊指标
  - Eta：反映一个名义变量和一个连续变量间的相关程度，取值介于0～1之间，取值越大表明相关性越强。
  - Kappa、OR、RR：15章中介绍的反映分类变量相关程度的指标

SPSS中的相应功能

- “分析”→“描述统计”→“交叉表格”→“统计”



SPSS中的相应功能

- “相关性”复选框：适用于两个连续变量的分析，计算行、列变量的Pearson相关系数和Spearman等级相关系数
- “按区间标定”复选框：适用于对一个名义变量和一个连续变量的分析，计算Eta值
- “有序”复选框组：适用于对两个有序分类变量的分析，计算Gamma统计量等
- “名义”复选框组：适用于对两个无序分类变量的分析，计算列联系数等
- “Kappa”复选框，适用于两个分类变量的一致性分析，计算Kappa值
- “风险”复选框：计算OR值和RR值

SPSS中的相应功能

- “分析”→“相关”
  - “双变量”过程：对两个变量的相关性进行分析
  - “偏相关”过程：如果需要分析的两个变量其取值均受到其他变量的影响，就可以利用偏相关分析对其他变量进行控制，输出其他变量控制后的相关系数
  - “距离”过程：计算个案间的距离或者变量间的距离

16.2 简单相关分析

方法原理

- 一些基本概念
  - 直线相关：两个变量呈线性共同增大或呈线性一增一减
  - 曲线相关：两变量存在相关趋势，但非线性，例如：指数、对数、幂
  - 正相关、负相关
  - 完全相关：一个变量的取值能被另一个变量的取值准确推算

方法原理

积差相关系数（Pearson相关系数）的计算

$$l_{xx}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$l_{yy}=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$l_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

积差相关系数 $r=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$

因为 l_{xx}、l_{yy}和l_{xy}都是离均差积和

所以 r 又称为积差相关系数

注：积差相关系数严格上仅适用于两变量呈线性相关时。

方法原理

- 积差相关系数（person相关系数）的特点
  - 是一个无单位的量值，且-1 < r < 1
  - r > 0为正相关，r < 0为负相关
  - |r|越接近于1，说明相关性越强，|r|越接近于0，说明相关性越弱

方法原理

- 积差相关系数的检验方法
  - 样本相关系数 r 是总体相关系数 ρ 的估计值，需对 ρ 进行假设检验。

    H_0：ρ＝0，两变量间无直线相关关系

    H_1：ρ≠0，两变量间有直线相关关系

  - 在SPSS中，直接给出最终的 P 值。

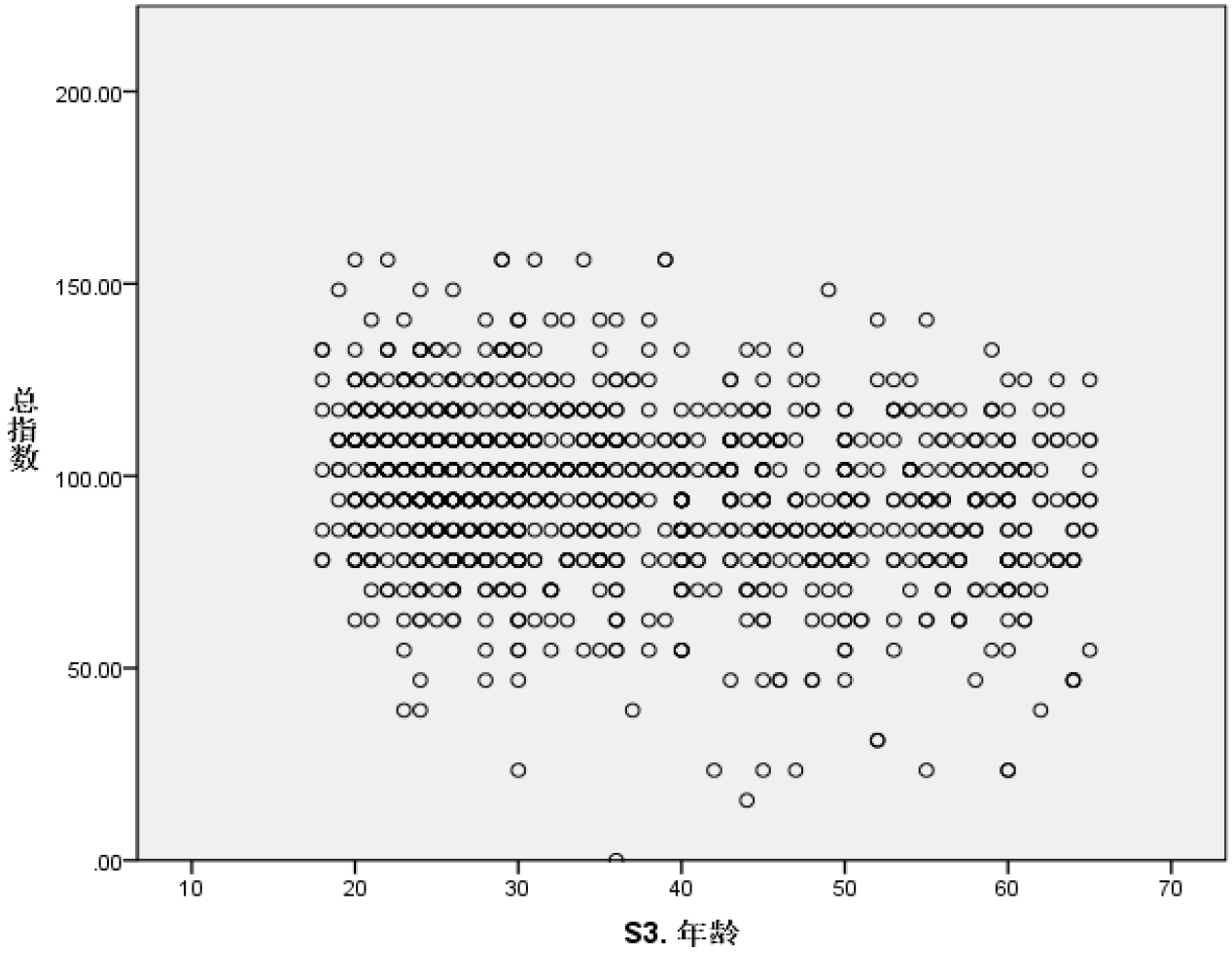
方法原理

- 积差相关系数的适用条件
  - 积差相关系数适用于两变量线性相关的情形，且服从双变量正态分布（可放宽条件）。
  - 样本中存在的极端值对积差相关系数的计算影响极大，要慎重考虑和处理，必要时可以对其进行剔除，或者加以变量变换。
  - 工具：散点图

案例

- 数据文件: **CCSS_Sample.sav**
- 要求：利用相关分析考察总信心指数值和年龄的相关性
- 具体过程
  - 绘制散点图，观察两连续变量是否存在线性相关
  - “分析”→“相关”→“双变量”
  - 将“**index1**”和“年龄”选入变量框
  - 相关系数默认用“**pearson**”相关系数

案例



案例


双变量相关性
Qa3
Qa4
Qa8
Qa9
Qa10
Qa16
现状指数 [index1a]
预期指数 [index1b]
time = 200704 (Fl...
变量(V):
S3. 年龄 [s3]
总指数 [index1]
选项(O)...
样式(L)...
Bootstrap...
相关系数
Pearson
Kendall's tau-b
Spearman
显著性检验
双尾检验(T)
单尾检验(L)
标记显著性相关(F)
确定
粘贴(P)
重置(R)
取消
帮助
双变量相关性: 选项
Statistics
平均值和标准差(M)
叉积偏差和协方差(C)
缺失值
按对排除个案(P)
按列表排除个案(L)
继续
取消
帮助

案例

Descriptive Statistics

	Mean	Std. Deviation	N
S3. 年龄	36.36	12.861	1147
总指数	95.8935	20.99710	1147

Correlations

		S3. 年龄	总指数
S3. 年龄	Pearson Correlation	1	-.219**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	189541.728	-67796.202
	Covariance	165.394	-59.159
	N	1147	1147
总指数	Pearson Correlation	-.219**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	-67796.202	505246.298
	Covariance	-59.159	440.878
	N	1147	1147

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

Pearson相关系数

P值

I_{xx}

I_{xy}

I_{yy}

秩相关系数（Spearman等级相关系数）

- 适用情况：不服从双变量正态分布、分布未知、等级资料。
- Spearman等级相关用 r_s 表示两变量相关关系的密切程度及相关方向。
- 基本思想：将两变量分别从小到大编秩，对秩次进行Pearson相关分析。

秩相关系数（Spearman等级相关系数）

- 数据文件: **CCSS_Sample.sav**
- 要求：利用Spearman相关分析考察总信心指数值和年龄的相关性
- 具体过程
  - “分析”→“相关”→“双变量”
  - 将“index1”和“年龄”选入变量框
  - 相关系数用“Spearman”相关系数

秩相关系数（Spearman等级相关系数）

秩相关系数（Spearman等级相关系数）

Correlations

			S3. 年龄	总指数
Spearman's rho	S3. 年龄	Correlation Coefficient	1.000	-.213**
		Sig. (2-tailed)	.	.000
		N	1147	1147
	总指数	Correlation Coefficient	-.213**	1.000
		Sig. (2-tailed)	.000	.
		N	1147	1147

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

Kendall等级相关系数

- 适用情况：两个变量均为有序分类变量。
- 数据文件：CCSS_Sample.sav
- 要求：利用Kendall等级相关分析考察总信心指数值和年龄的相关性
- 具体过程
  - “分析”→“相关”→“双变量”
  - 将“index1”和“年龄”选入变量框
  - 相关系数用“Kendall’s tau-b”相关系数

Kendall等级相关系数

案例

Correlations

			S3. 年龄	总指数
Kendall's tau_b	S3. 年龄	Correlation Coefficient	1.000	-.152**
		Sig. (2-tailed)	.	.000
		N	1147	1147
	总指数	Correlation Coefficient	-.152**	1.000
		Sig. (2-tailed)	.000	.
		N	1147	1147

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

16.3 偏相关分析

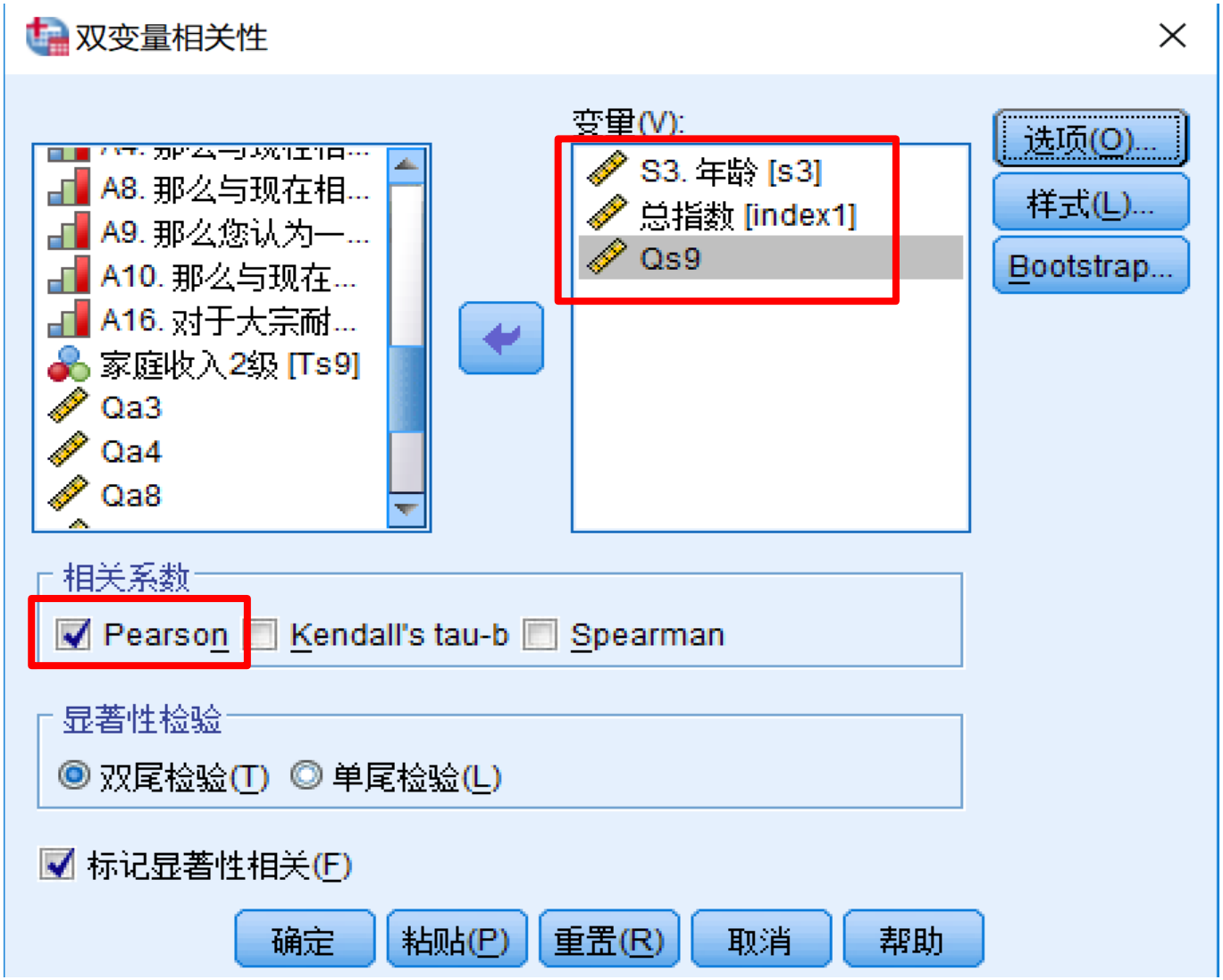
方法原理

- 控制其它变量影响的情况下，分析两个变量之间的相关关系。
- 偏相关系数：揭示两变量之间的真实联系。

案例

- 数据文件：CCSS_Sample.sav
- 要求：控制家庭收入的影响之后考察年龄和总信心指数的相关关系
- 具体过程
  - 首先考察“总信心指数”、“年龄”、“家庭收入”两两之间的相关性
  - “分析”→“相关”→“偏相关”
  - 将“index1”和“年龄”选入变量框
  - 将“Qs9”选入控制框

案例


双变量相关性
变量(V):
S3. 年龄 [s3]
总指数 [index1]
Qs9
A8. 那么与现在相...
A9. 那么您认为一...
A10. 那么与现在...
A16. 对于大宗耐...
家庭收入2级 [Ts9]
Qa3
Qa4
Qa8
选项(O)...
样式(L)...
Bootstrap...
相关系数
Pearson
Kendall's tau-b
Spearman
显著性检验
双尾检验(T)
单尾检验(L)
标记显著性相关(F)
确定
粘贴(P)
重置(R)
取消
帮助

案例

Correlations

		S3. 年龄	总指数	Qs9
S3. 年龄	Pearson Correlation	1	-.219**	-.138**
	Sig. (2-tailed)		.000	.000
	N	1147	1147	992
总指数	Pearson Correlation	-.219**	1	.084**
	Sig. (2-tailed)	.000		.008
	N	1147	1147	992
Qs9	Pearson Correlation	-.138**	.084**	1
	Sig. (2-tailed)	.000	.008	
	N	992	992	992

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

结论：“总信心指数”和“年龄”均与“家庭收入”存在统计学上的相关关系

案例


偏相关
A16. 对于大宗耐...
家庭收入2级 [Ts9]
Qa3
Qa4
Qa8
Qa9
Qa10
Qa16
现状指数 [index1a]
变量(V):
S3. 年龄 [s3]
总指数 [index1]
控制(C):
Qs9
选项(O)...
Bootstrap...
显著性检验
双尾检验(T)
单尾检验(N)
显示实际显著性水平(D)
确定
粘贴(P)
重置(R)
取消
帮助

案例

Correlations

Control Variables			S3. 年龄	总指数
Qs9	S3. 年龄	Correlation	1.000	-.216
		Significance (2-tailed)	.	.000
		df	0	989
	总指数	Correlation	-.216	1.000
		Significance (2-tailed)	.000	.
		df	989	0

结论：控制了“家庭收入”后，“总信心指数”和“年龄”之间的偏相关系数为 -0.216，具有统计学意义，认为两者之间存在负相关关系。

16.4 距离计算

- “距离计算”可对个案间的距离或变量间的距离进行计算，是因子分析、聚类分析、多维尺度分析的预分析，可以帮助用户了解复杂数据集的内在结构，为进一步分析做准备。

指标体系

- 距离测量（非相似性测量）
  - 欧几里得距离：以两变量差值平方和的平方根为距离
  - 欧式平方距离：以两变量差值平方和为距离
  - 切比雪夫距离：以两变量绝对差值的最大值为距离
  - **Block距离：**以两变量绝对差值之和为距离
  - 明可夫斯基距离：以两变量绝对差值 **p** 次幂之和的 **p** 次根为距离
  - 自定义距离公式：以两变量绝对差值 **p** 次幂之和的 **q** 次根为距离

指标体系

- 相似性测量
  - 积差相关系数
  - 余弦值

案例

- 数据文件：distance.sav
- 要求：计算7个基因间的距离
- 具体过程
  - “分析”→“相关”→“距离”
  - 将7个基因选入变量框
  - 选择“计算距离”中的“变量间”
  - 其余默认

案例


距离
id
变量(V):
FPGS
ELF3
CDK2AP1
GFRA2
TCEB1
标注个案(L):
计算距离
个案间(C)
变量间(B)
测量
非相似性(D)
相似性(S)
测量(M)...
Euclidean 距离
确定
粘贴(P)
重置(R)
取消
帮助

案例

Proximity Matrix

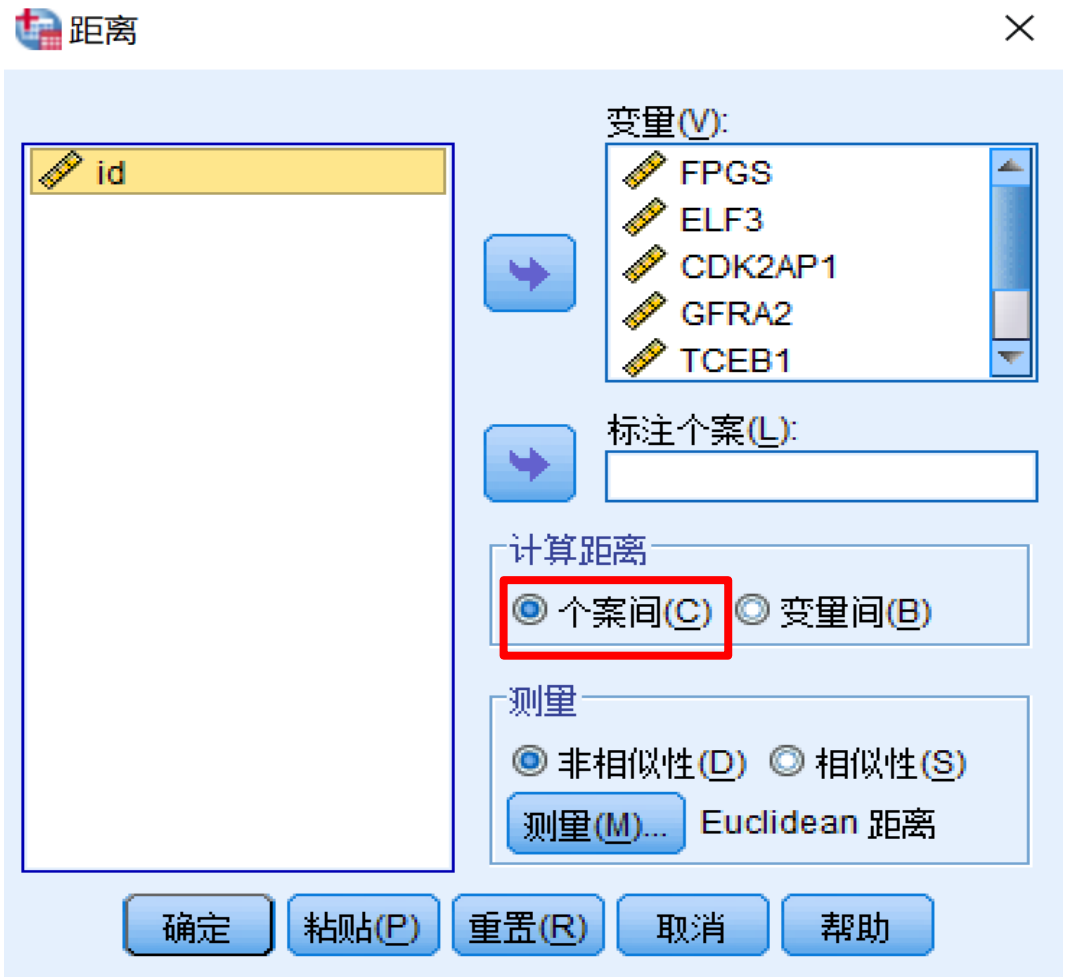
	Euclidean Distance						
	FPGS	ELF3	CDK2AP1	GFRA2	TCEB1	NFE2	IRF2
FPGS	.000	.779	2.416	.749	1.006	.781	1.424
ELF3	.779	.000	1.749	.804	1.106	.933	1.578
CDK2AP1	2.416	1.749	.000	2.106	2.480	2.349	2.784
GFRA2	.749	.804	2.106	.000	1.312	.521	1.085
TCEB1	1.006	1.106	2.480	1.312	.000	1.400	1.864
NFE2	.781	.933	2.349	.521	1.400	.000	.962
IRF2	1.424	1.578	2.784	1.085	1.864	.962	.000

This is a dissimilarity matrix

案例

- 数据文件：distance.sav
- 要求：计算13个个案间的距离
- 具体过程
  - “分析”→“相关”→“距离”
  - 将7个基因选入变量框
  - 选择“计算距离”中的“个案间”
  - 其余默认

案例


距离
id
变量(V):
FPGS
ELF3
CDK2AP1
GFRA2
TCEB1
标注个案(L):
计算距离
个案间(C)
变量间(B)
测量
非相似性(D)
相似性(S)
测量(M)...
Euclidean 距离
确定
粘贴(P)
重置(R)
取消
帮助

案例

Proximity Matrix

	Euclidean Distance												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	.000	.899	.933	1.019	1.45	.741	.35	.640	.439	.402	.615	.840	.972
2	.899	.000	.698	1.268	2.03	.688	1.0	.745	1.15	.744	1.26	1.26	.951
3	.933	.698	.000	1.066	1.87	.646	.99	.806	1.15	.896	1.31	1.14	1.06
4	1.02	1.268	1.066	.000	1.04	.849	.79	1.36	.967	1.00	.875	.490	1.24
5	1.45	2.025	1.866	1.041	.000	1.718	1.2	2.01	1.33	1.49	1.05	.791	1.64
6	.741	.688	.646	.849	1.72	.000	.73	.767	.859	.636	.979	.975	.993
7	.346	1.028	.991	.794	1.23	.726	.00	.826	.307	.483	.371	.609	1.06
8	.640	.745	.806	1.358	2.01	.767	.83	.000	.836	.795	1.14	1.33	1.31
9	.439	1.151	1.155	.967	1.33	.859	.31	.836	.000	.630	.401	.833	1.26
10	.402	.744	.896	1.004	1.49	.636	.48	.795	.630	.000	.641	.804	.678
11	.615	1.264	1.308	.875	1.05	.979	.37	1.14	.401	.641	.000	.607	1.16
12	.840	1.262	1.138	.490	.791	.975	.61	1.33	.833	.804	.607	.000	1.02
13	.972	.951	1.065	1.243	1.64	.993	1.1	1.31	1.26	.678	1.16	1.02	.000

This is a dissimilarity matrix

THE END