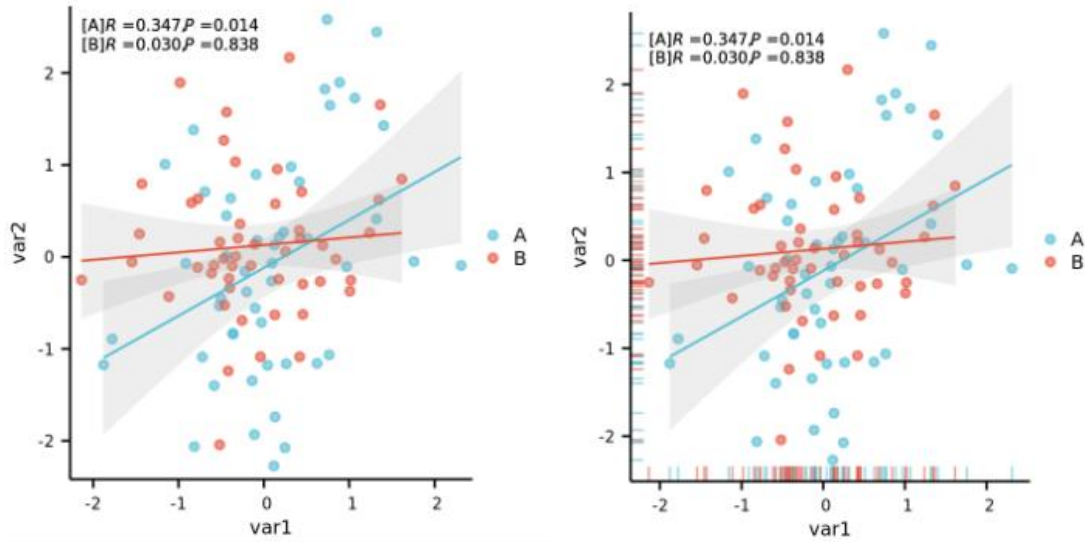
基础绘图 - [关系情况] - 相关性散点图-分组



网址：https://www.xiantao.love

更新时间：2023.09.21

目录

基本概念

- 散点图：通过点的形式来展示数据的分布情况
- 相关性散点图：分析 1 个变量和另外 1 个变量之间的相关性
- 相关性散点图-分组：根据分组信息，将点分成不同组进行展示

应用场景

- 相关性散点图常用来进行数据的对比
- 直观地观察到两个变量之间的关系

分析过程

上传数据 → 数据处理(清洗) → 相关性分析 → 可视化

- 数据格式：（具体数据格式要求可以看后面过程的"数据格式"部分）
  - 数据第 1 列必须为字符类型，作为分组信息，最多支持 10 组
  - 数据第 2 列必须为数值类型，对应用于相关性分析的变量 1
  - 数据第 3 列必须为数值类型，对应用于相关性分析的变量 2
  - 必须提供三列数据；至少需要 6 行，最多 50000 行；每组数据最少需要 3 行数据，最多支持 5000 行

1	group1	var1	var2
44	A	-1.77678	-0.89296
45	A	0.622867	-1.15757
46	A	-0.52228	-0.5303
47	A	1.322231	2.445683
48	A	-0.36344	-0.8325
49	A	1.319066	0.41352
50	A	0.043779	-1.17868
51	A	-1.87866	-1.17403
52	B	-0.33292	1.034686
53	B	1.363114	1.653503
54	B	-0.46915	-0.01795
55	B	0.842876	-0.0242
56	B	-1.45799	0.250247
57	B	-0.40031	-0.33712
58	B	-0.77642	-0.11335
59	B	-0.3693	-0.09888
60	B	1.240101	0.264087
61	B	-0.10743	0.138984
62	B	0.172594	-0.24227

- 数据处理：对每一列数值类型的数据及其他列数据进行相应处理
  - 分类类型数据只能是纯字符类型的数据，不能包含数值，缺失值与无法识别的值
  - 数值类型数据只能是纯数值类型数据，不能包含 0，负数、非数值与不规则的值
  - 分组中的每一个变量不能都是一个值
- 分析：
  - 统计描述
    - 对（每个分组的）变量进行常见统计描述指标统计分析

统计描述

各个组对应常见「统计描述指标」

组别1	组别2	数目	最小值	最大值	中位数(Median)	四分位距(IQR)	下四分位	上四分位	均值(Mean)	标准差(SD)
A	var1	50	-1.8787	2.3103	0.066833	1.0206	-0.42578	0.59486	0.081509	0.81901
A	var2	50	-2.2719	2.582	-0.081203	1.7132	-1.0215	0.69168	-0.075684	1.1923
B	var1	50	-2.1365	1.6152	-0.26768	0.92723	-0.50448	0.42275	-0.099787	0.79354
B	var2	50	-2.0418	2.1686	-0.0056045	0.84924	-0.2626	0.58664	0.12207	0.79117

  - 异常值分析

- ◆ 检查数据中是否有离群值和异常值

异常值分析

离群值 = Q1(下四分位) - 1.5*IQR(四分位间距) 或者 Q3(上四分位) + 1.5*IQR(四分位间距)
异常值 = Q1(下四分位) - 3.0*IQR(四分位间距) 或者 Q3(上四分位) + 3.0*IQR(四分位间距)

组别1	组别2	离群值	异常值
A	var1	2.31029682277906	
B	var1	-2.13649385561006	
B	var2	-2.04184985362182...	

- ■ 正态性检验
  - ◆ 对（每个分组的）变量进行正态性检验（Shapiro–Wilk normality test）

正态性检验(Shapiro-Wilk normality test)

组别1	组别2	自由度(df)	统计量	p值
A	var1	49	0.9838	0.7192
A	var2	49	0.98066	0.5799
B	var1	49	0.98263	0.6669
B	var2	49	0.96708	0.1755

- ■ 相关性分析
  - ◆ 包含不同方法（Pearson、Spearman）计算的分组变量相关性系数值与统计学 p 值等，补充了变量相关性表格

相关性分析

同时提供Pearson和Spearman统计方法，可以根据需要选择标注在图中的方法

表1：分组变量相关性

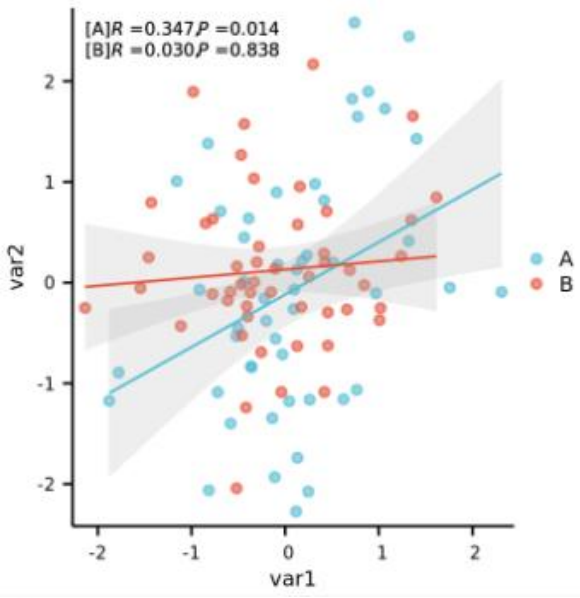
组别	方法	组别1	组别2	自由度(df)	统计量	相关系数	置信区间(95%CI)	p值
A	Pearson	var1	var2	48	2.6569	0.35806773	0.088545 - 0.57873	0.0107
A	Spearman	var1	var2		1.36e+04	0.34713085		0.0139
B	Pearson	var1	var2	48	0.56361	0.08108229	-0.20182 - 0.3515	0.5756
B	Spearman	var1	var2		2.021e+04	0.02962785		0.8378

表2：变量相关性

方法	组别1	组别2	自由度(df)	统计量	相关系数	置信区间(95%CI)	p值
Pearson	var1	var2	98	2.3572	0.2316401	0.036901 - 0.40943	0.0204
Spearman	var1	var2		1.328e+05	0.2033003		0.0426

- ➢ 可视化：数据清洗后，进行相关性分析，再用 ggplot2 包进行可视化

结果解读



- 横坐标表示第 1 列变量
- 纵坐标表示第 2 列变量
- 图中的线为拟合线，拟合线周围的阴影部分为置信区间
- 图中左上角为标注：
  - “[A]”中括号里是分组名
  - “R”表示变量间的相关性系数
  - “P”表示变量间的统计学 p 值

数据格式

相关性散点图-分组

1	group1	var1	var2
44	A	-1.77678	-0.89296
45	A	0.622867	-1.15757
46	A	-0.52228	-0.5303
47	A	1.322231	2.445683
48	A	-0.36344	-0.8325
49	A	1.319066	0.41352
50	A	0.043779	-1.17868
51	A	-1.87866	-1.17403
52	B	-0.33292	1.034686
53	B	1.363114	1.653503
54	B	-0.46915	-0.01795
55	B	0.842876	-0.0242
56	B	-1.45799	0.250247
57	B	-0.40031	-0.33712
58	B	-0.77642	-0.11335
59	B	-0.3693	-0.09888
60	B	1.240101	0.264087
61	B	-0.10743	0.138984
62	B	0.172594	-0.24227

数据要求：

- 第一列是用于分组的信息，第二列是第 1 个变量的观测值，第三列是第 2 个变量的观测值，数据只需要 3 列，每个分组至少 3 行数据，例如：两个分组至少 6 行数据，第一列均需要是字符类型，第二、三列需要是数值类型。
- 上传数据只需要 3 列，最多支持 50000 行，每个分组最少，若验证数据时返回报错，需要在上传数据内进行相应的调整，然后再上传数据。
  - 分类类型数据只能是纯字符类型的数据，不能包含数值，缺失值与无法

识别的值

- 数值类型数据只能是纯数值类型数据，不能包含 0、负数、非数值与不规则的值

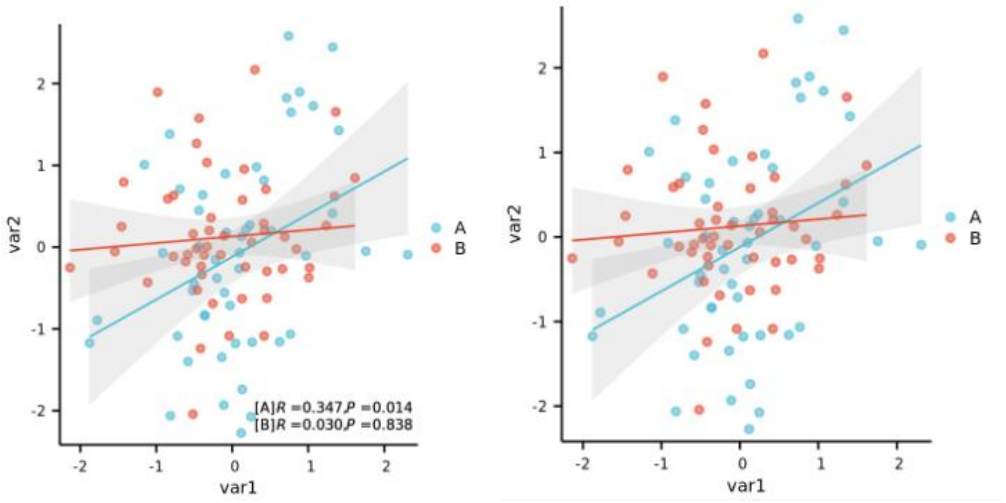
- 数据每一列列名不能重复，不能有空值，不能有不识别的字符
- 第一列分类变量中的分组数量最多支持 10 组

参数说明

（说明：标注了颜色的为常用参数。）

统计

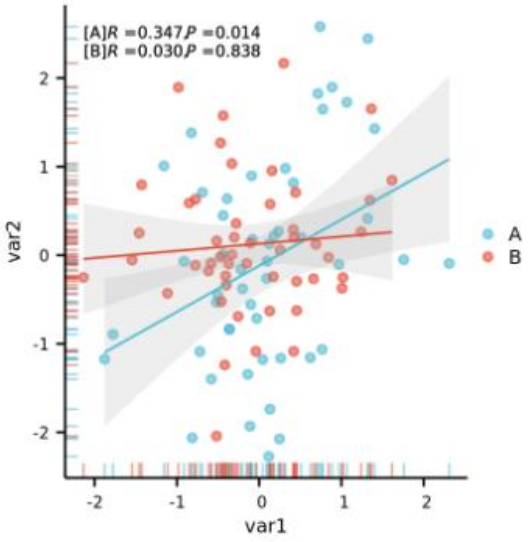
- 统计方法：可以选择变量 1 与变量 2 间进行相关性分析的方法
  - Spearman：非参数检验方法，默认使用该方法，数据可以不需要满足正态性
  - Pearson：参数检验方法，数据需要满足双正态
- 标注位置：可以修改图中相关性分析方法(Spearman)、相关性系数(R)，统计 学 p 值的位置，默认在图形的左上角，还可以选择左下、右上、右下、无(不进行标注)，如下：左侧为右下，右侧为无



- 标注颜色：当图形中有标注的时候，可以修改标注的颜色

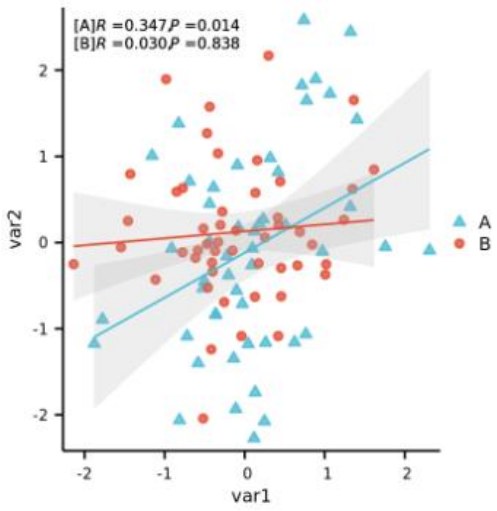
样式

- 样式：可以选择相关性散点图-分组图形展示的整体样式（结果），默认为经典，可选加分布竖线。例如：



点

- 填充色：可以修改图中各点的填充颜色，最多支持修改 10 个颜色，超出会使用随机颜色。受配色方案全局性修改。
- 描边色：可以修改图中各点的描边颜色，最多支持修改 10 个颜色，超出会使用随机颜色。受配色方案全局性修改。
- 样式：可以修改图中各点的样式（形状），默认为圆形，可选择 圆形、正方形、菱形、三角形、倒三角，默认为圆形。多选，多选后不同的分组/分类中的点的类型也会有相应变化，循环取该参数值。如下：

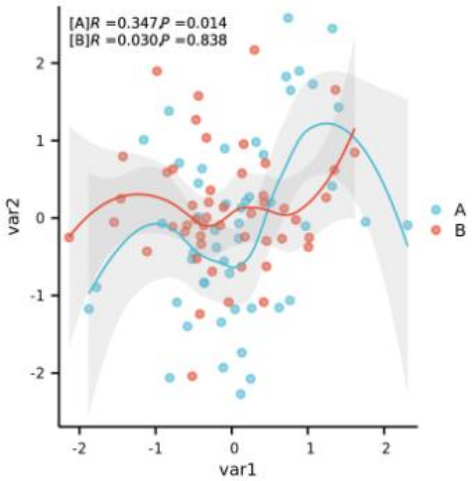


- 大小：可以修改图中各点的大小比例，默认为 1

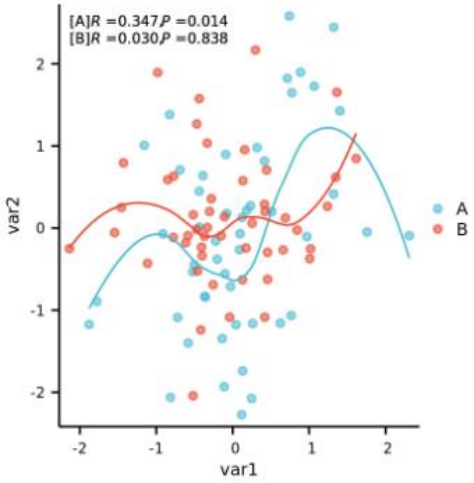
- 不透明度：可以修改拟合线线条的不透明度，1 表示完全不透明

拟合线

- 展示：可以选择是否进行展示拟合线的操作，默认展示。
- 拟合方法：可以修改图中拟合部分的拟合方法(类型)，默认为直线，还可以 选择曲线的形式，如下：



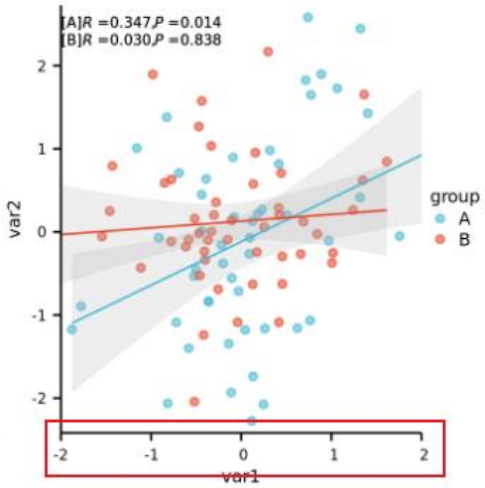
- 拟合线颜色：可以修改图中拟合线的颜色。
- 拟合线样式：可以修改图中拟合线的样式，默认为实线，可选择实线或虚线。
- 线条粗细：可以选择修改图中拟合线的线条粗细。
- 置信区间展示：可以选择是否展示拟合线的置信区间（阴影部分），默认为展示，还可以选择不展示，如下：



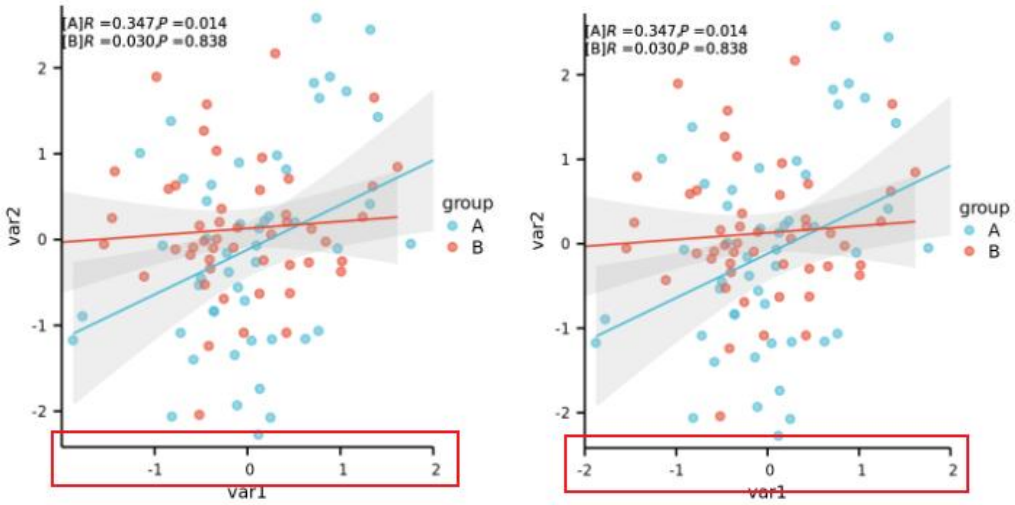
- 不透明度：波形的透明度。0 为完全透明，1 为完全不透明。

坐标轴

- 是否显示 x 轴：选择即展示 x 轴。
- 是否显示 y 轴：选择即展示 y 轴。
- x 轴标注旋转：可以选择设置 x 轴标注的倾斜角度。
- x 轴范围+刻度：可以控制 x 轴的范围和刻度（不能调整超过数据范围的 20%，如果调整 过大可能会无作用），可只提供 2 个值来控制范围。例如：-2,2;



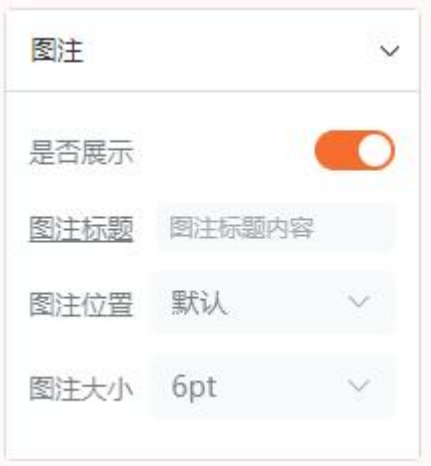
可以提供范围值和刻度值来控制范围可刻度。例如：-2,-1,0,1,2,2。注意，此时最小和最大值会被当做范围，不会作为刻度，如果需要作为刻度，重复写一次即可。


[A]R =0.347,P =0.014
[B]R =0.030,P =0.838
group
A
B
var1
var2

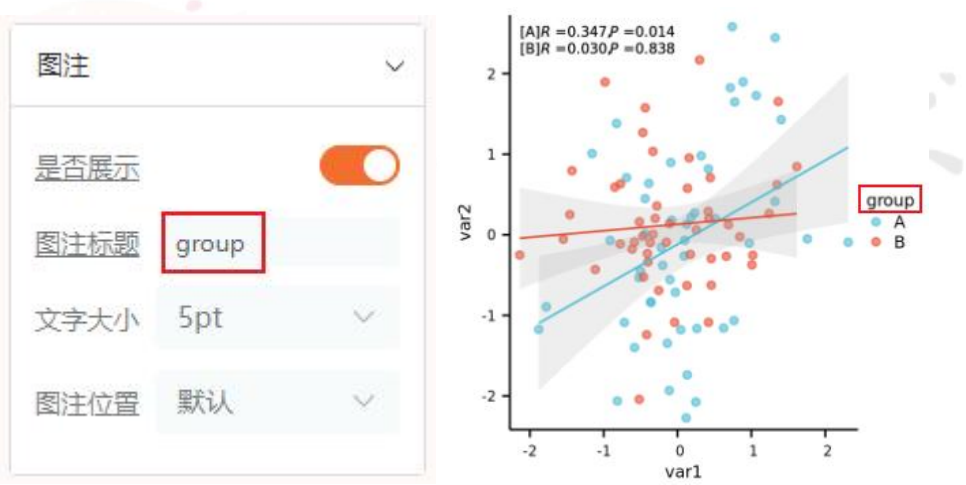
标题文本

- 大标题：大标题文本。
- x 轴标题：x 轴标题文本。
- y 轴标题：y 轴标题文本。
- 补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如{{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如[[2]]

图注（Legend）



- 是否展示：是否展示图注
- 图注标题：可以添加图注标题，如：

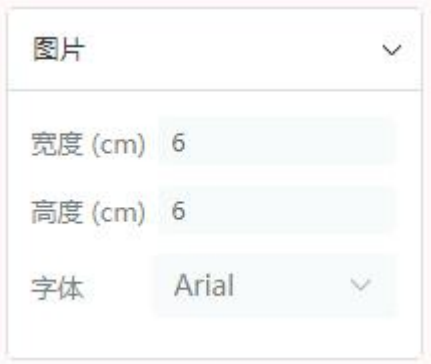


- 文字大小：图注标题文字的大小，默认为 6pt。
- 图注位置：可选择默认、右、上、下。

风格

- 边框：可以选择是否进行添加图形边框的操作
- 网格：可以选择是否进行添加图形内网格的操作
- 文字大小：控制整体文字大小，默认为 6pt

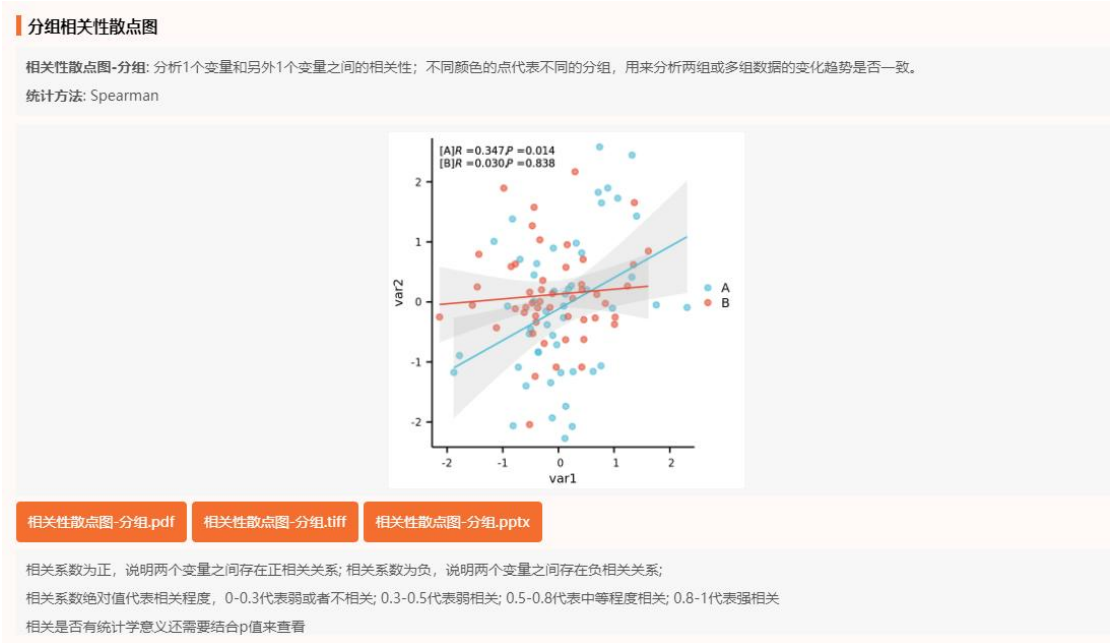
图片



- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

结果说明

主要结果



主要结果格式为图片格式，提供 PDF、TIFF 、PPTX 格式下载

方法学

软件：R（4.2.1)版本

R 包：ggplot2 包（用于可视化）、ggtext 包

处理过程：

(1）分析多组数据的两个变量之间相关性后，用 gglot2 可视化结果，进而展示各组数据中该变量的变化趋势

如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分（Acknowledge）致谢仙桃学术（www.xiantao.love）。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

常见问题

1．方法里面的 Spearman 和 Pearson 方法，应该选择哪一个?

答： 两种方法均可以选择。Pearson 要求数据满足正态性，Spearman 因为是非参数的方法，可以不需要满足。可以先选择非参数的 Spearman 相关进行尝试。

2．相关系数多少为好?

答：这个没有很统一的标准，可以参考以下：

- 相关系数强弱：
  - 绝对值在 0.8 以上：强相关
  - 绝对值在 0.5-0.8：中等程度相关
  - 绝对值在 0.3-0.5：相关程度一般
  - 绝对值在 0.3 以下：弱或者不相关
  - 正数表示正相关，负数表示负相关

3．每组的数据不一样多可以分析吗?

答：只要数据满足最低要求，就可以上传数据进行分析

4．数据中存在离群值和异常值的情况，怎么处理?

答：若【补充结果-异常值分析】表格中给出有离群值或异常值的情况，可以根据自己的研究情况进行取舍，如果是由一些试验误差等其他因素导致的，可以及时删除以保证数据的准确性