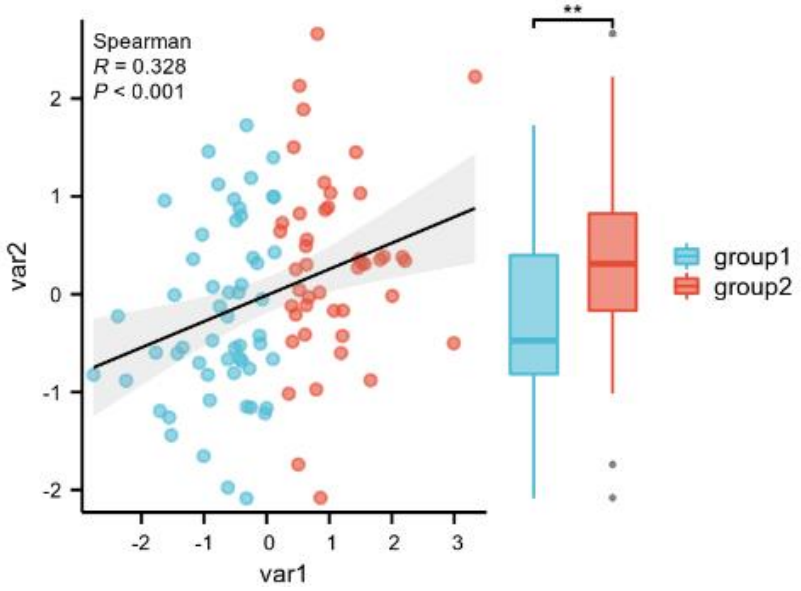
基础绘图 - [关系情况] - 相关性散点图-两组比较



网址：https://www.xiantao.love

更新时间：2023.10.07

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-两组比较：根据分组信息，将点分成两组单个维度的比较，用箱式图展示
- 两组比较统计方法：
  - T test，亦称 student t 检验（Student's t test），主要用于两组之间的比较， 两组需要满足正态性和方差齐性的要求。
  - Welch's test，又称不等方差检验，即当两组仅满足正态而不满足方差齐性的要求时，可以选择用该方法进行两组的比较。
  - Wilcoxon rank sum test，也叫 Mann-Whitney U test（曼-惠特尼 U 检验），或者 Wilcoxon-Mann-Whitney test。秩和检验是一个非参数的假设检验方法，一般用于两组不满足正态性的情况。

应用场景

- 相关性散点图常用来进行数据的对比
- 两组比较图能够比较两组数据之间的差异

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）
  - 至少提供 2 列数据；至少需要 6 行， 最多 5000 行；每组数据最少需要 3 行数据，第 1、2 列必须为数值类型，对应用于相关性分析的变量 1、变量 2，按照第 1 列的中位值大小分成两组进行比较
  - 如果需要自定义分组信息，则必须提供 3 列数据，数据第 1 列作为分组信息，必须为字符类型，只支持 2 分类（2 组），数据第 2、3 列必须为数值类型，对应用于相关性分析的变量 1、变量 2，通过分组信息的内容进行比较

	A	B	C
1	type	var1	var2
2	group1	0.1020305	-0.663584
3	group2	0.46421611	-0.20844003
4	group2	0.50614989	-1.74054472
5	group1	-1.52132525	-1.44109924
6	group1	-0.40090029	0.09681601
7	group1	-0.32205095	-2.08802513
8	group2	2.17045728	0.37926024
9	group2	0.52650206	2.13031809
10	group2	1.42700632	1.45112923
11	group1	-0.21581337	0.37145807
12	group1	-0.61734349	-1.97720369
13	group1	-0.00021546	-1.16543814
14	group2	3.33051527	2.22366236
15	group1	-1.76803376	-0.5976789
16	group1	-1.3416139	-0.54591428

- 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
- 分组只支持 2 分类

➢ 分析：

- 相关性分析 – 统计描述
  - ◆ 对变量进行常见统计描述指标统计分析

相关性分析 - 统计描述

各个组对应常见「统计描述指标」

组别	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)	标准误(SE)
var1	100	-2.764	3.3305	0.050908	1.4037	-0.60632	0.79734	0.08983	1.1229	0.11229
var2	100	-2.088	2.6627	-0.012691	1.3199	-0.65649	0.66339	0.012546	0.96118	0.096118

- 相关性分析 – 异常值分析
  - ◆ 检查数据中是否有离群值和异常值

相关性分析 - 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	离群值	异常值
var1	3.33051527361684,...	
var2	2.66270453551476	

各组离群值和异常值如上所示，如数据确认非人为记录错误，可不进行处理

- 相关性分析 – 正态性检验
  - ◆ 对变量进行正态性检验（Shapiro-Wilk normality test）

相关性分析 - 正态性检验(Shapiro-Wilk normality test)

组别	自由度(df)	统计量	p值
var1	99	0.99393	0.9371
var2	99	0.99182	0.8079

正态性检验结果显示，提供的变量均接近正态分布(P > 0.05)，建议选择用 参数检验方法(Pearson)

- 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

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	3.2515	0.31205	0.12318 - 0.4791	0.0016
Spearman	var1	var2	98	1.12e+05	0.32769		0.0009

相关系数为正，说明两个变量之间存在正相关关系; 相关系数为负，说明两个变量之间存在负相关关系;
相关系数绝对值代表相关程度，0-0.3代表弱或者不相关; 0.3-0.5代表弱相关; 0.5-0.8代表中等程度相关; 0.8-1代表强相关
相关是否有统计学意义还需要结合p值来查看

- 统计描述
  - 对分组结果进行常见统计描述指标统计分析

统计描述

各个组常见「统计描述指标」

组别	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)	标准误(SE
group1	55	-2.088	1.7273	-0.47407	1.2151	-0.81529	0.39981	-0.2216	0.90871	0.12253
group2	45	-2.0818	2.6627	0.31082	0.99344	-0.16898	0.82446	0.29873	0.9555	0.14244

- 统计描述
  - 检查分组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	离群值	异常值
group2	-1.74054471828801...	

各组离群值和异常值如上所示，如数据确认非人为记录错误，可不进行处理

- 正态性分析
  - 检查分组数据是否满足正态性检验

正态性检验

检验方法: Shapiro-Wilk normality test

组别	自由度(df)	统计量	p值
group1	54	0.97306	0.2516
group2	44	0.97735	0.5165

正态性检验结果显示，观测变量在各组内接近正态分布(P > 0.05)，建议选择用 参数检验的方法

- 方差齐性检验

- ◆ 检查被比较的两组数那

方差齐性检验

检验方法: Levene's test

· Base on Mean

自由度1(df1)	自由度2(df2)	统计量	p值
1	98	0.24315	0.6230

方差齐性检验显示，各组观测变量的方差相等(P > 0.05)

- ■ 独立样本 T 检验

- ◆ 两组数据比较的检验结果（不同的统计方法会有不一样的统计检验的表格。）

独立样本T检验

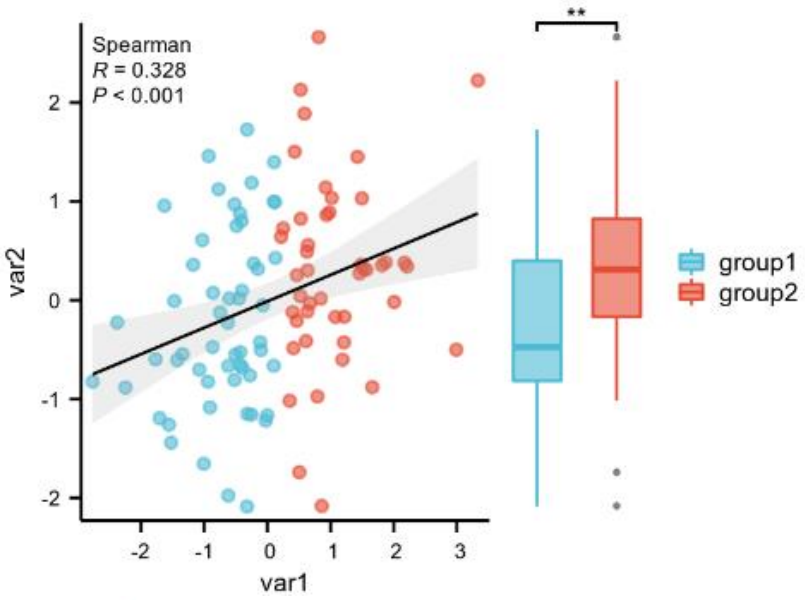
应用条件: 两组独立数据, 满足正态性检验和方差齐性检验

组别I	组别J	自由度(df)	统计量t	差值(J-I)	置信区间(95%CI)	p值
group1	group2	98	2.7834	0.52033	0.14935 - 0.8913	0.0065

p值满足<0.05时，可认为两组存在统计学上差异

- ➢ 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
- 散点图中的线为拟合线，拟合线周围的阴影部分为置信区间
- 箱式图是根据分组信息，绘制的两组比较结果，图中的灰色点为离群点
- 图中左上角为标注：
  - “Spearman”表示变量间进行相关性分析的方法
  - “R”表示变量间的相关性系数
  - “P”表示变量间的统计学 p 值
  - “**”表示两组数据之间的差异结果的显著性

数据格式

相关性散点图-两组比较

	A	B	C
1	type	var1	var2
2	group1	0.1020305	-0.663584
3	group2	0.46421611	-0.20844003
4	group2	0.50614989	-1.74054472
5	group1	-1.52132525	-1.44109924
6	group1	-0.40090029	0.09681601
7	group1	-0.32205095	-2.08802513
8	group2	2.17045728	0.37926024
9	group2	0.52650206	2.13031809
10	group2	1.42700632	1.45112923
11	group1	-0.21581337	0.37145807
12	group1	-0.61734349	-1.97720369
13	group1	-0.00021546	-1.16543814
14	group2	3.33051527	2.22366236
15	group1	-1.76803376	-0.5976789
16	group1	-1.3416139	-0.54591428

数据要求：

- 至少提供 2 列数据，至少 6 行，当提供 2 列时，第 1-2 列均需要是数值类型数据，分析第 1 列和第 2 列（变量 1 和变量 2）的相关性；分组差异分析，按第一列的中位值大小分组；提供 3 列时，第 1 列需要是字符型的分组信息，每个分组至少有 3 行数据（只支持 2 分类），第 2-3 列需要是数值类型数据，分析第 2-3 列（对应变量 1 和变量 2）的相关性和分组的差异。
- 上传数据至少需要 2 列，最少 6 行数据，最多支持 5000 行，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

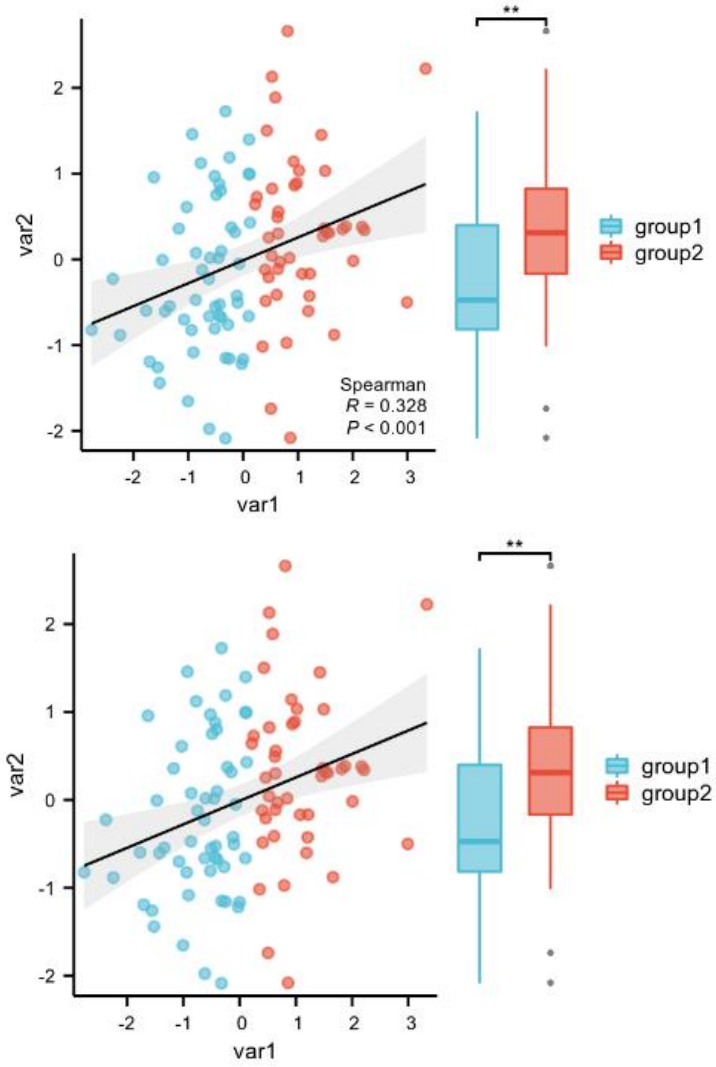
- 数据每一列列名不能重复，不能有空值，不能有不识别的字符
- 每组数据不能完全为一个值
- 第一列分类变量中的分组数量只支持 2 分类（即 2 个组）

参数说明

（说明：标注了颜色的为常用参数。）

相关性分析

- 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(P)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

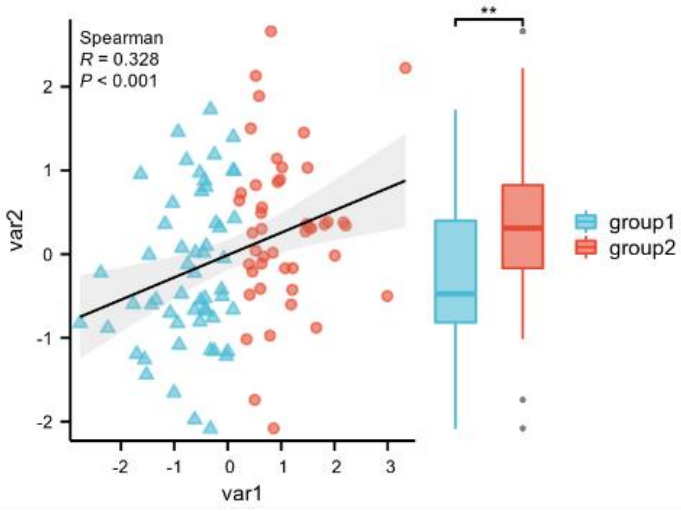
分组比较

- 统计方法：统计方法默认为 auto（自动选择），当上传数据验证成功并点击确认后，会自动替换成适合于上传数据的统计方法，之后可以自行选择和修改别的统计方法。统计方法的选择依据可以参考“基本概念”中统计方法的说明
- 分组比较：统计学差异标注的分组，默认为 group1 vs group2（标注分组）。当上传数据验证成功并点击确认后，会自动替换成对应上传数据的分组。之后可以自行选择想要保留和去掉的比较。（如果分组不满足>3 个观测以及标准差>0 的情况，则可能不会出现在此处。）允许都去掉，即不标注分组比较的内容
- 显著性显示的类型：可选择星号或者 p 值以及其他，影响分组比较中显著性标注，默认为星号。可以根据需要进行修改。

- 显著性大小：可以修改显著性的大小。

点

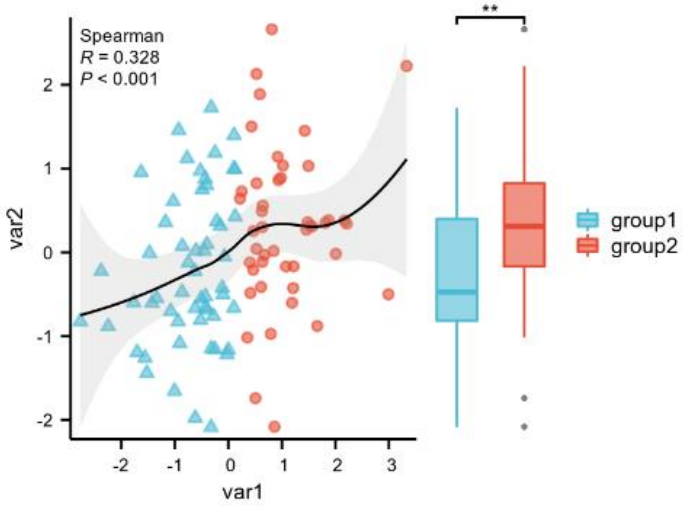
- 填充色：可以修改图中各点的填充颜色，受配色方案全局性修改。
- 描边色：可以修改图中各点的描边颜色，受配色方案全局性修改。
- 样式：可以修改图中各点的样式（形状），默认为圆形，可选择 圆形、正方形、菱形、三角形、倒三角，默认为圆形。多选，多选后不同的分组/分类中的点的类型也会有相应变化，循环取该参数值。如下：



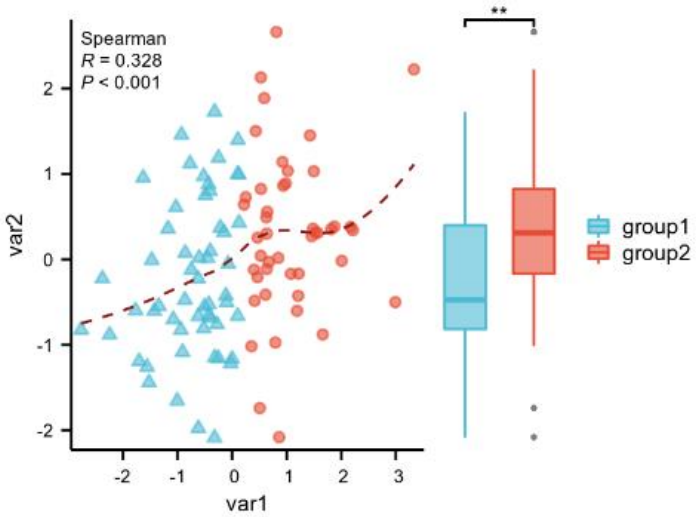
- 大小：可以修改图中各点的大小比例，默认为 1
- 不透明度：可以修改拟合线线条的不透明度，1 表示完全不透明

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。

箱

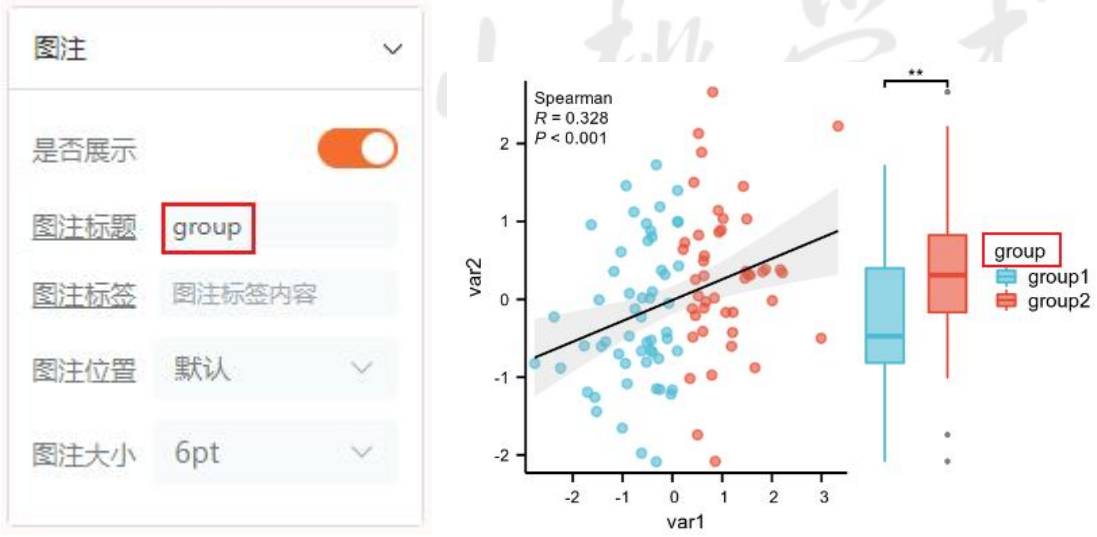
- 填充色：可以修改图中箱子的填充颜色，受配色方案全局性修改。
- 描边色：可以修改图中箱子的描边颜色，受配色方案全局性修改。
- 描边粗细：箱子描边的粗细，默认为 0.75p。
- 展示离群点：可选是否展示。
- 箱子宽度：箱子的宽度。
- 不透明度：箱子的透明度。0 为完全透明，1 为完全不透明。

标题文本

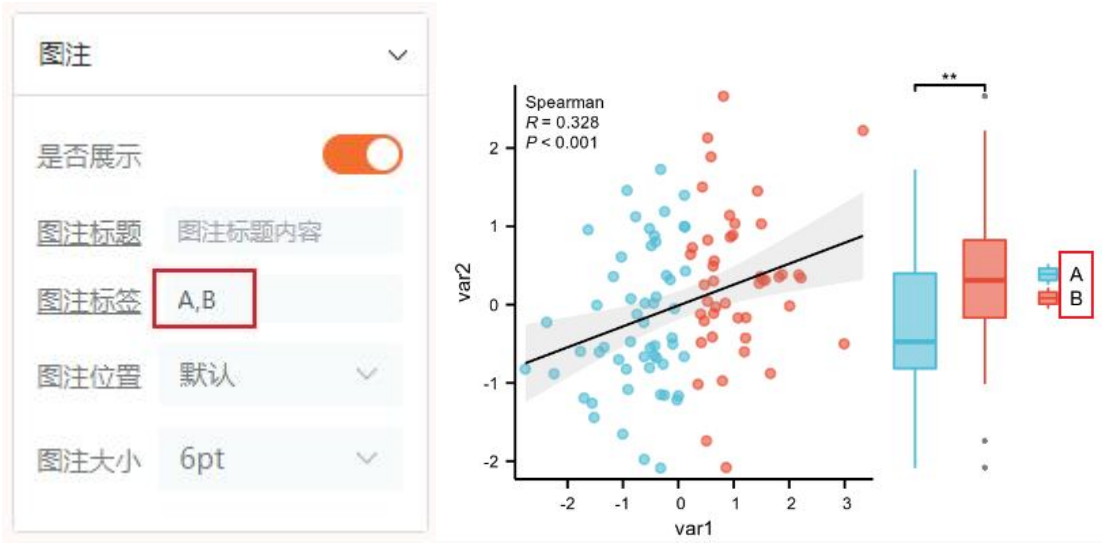
- 大标题：大标题文本。
- x 轴标题：x 轴标题文本。
- y 轴标题：y 轴标题文本。
- 补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如{{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如[[2]]

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：



- 图例标签：可以修改图注中分组标签的名字，如果有多个名字要修改，则需要把这些名字以逗号的形式合并成一个，例如：A，B



- 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

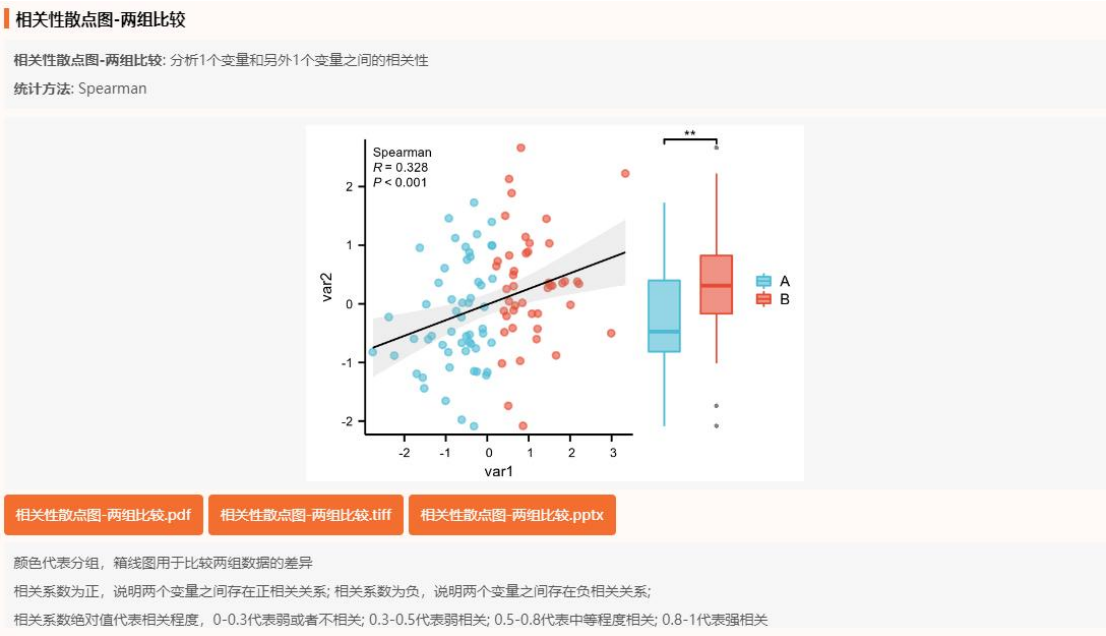
- 边框：可以选择是否进行添加图形边框的操作
- 网格：可以选择是否进行添加图形内网格的操作
- 文字大小：控制整体文字大小，默认为 6pt

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、stats 包和 car 包（用于统计计算）

处理过程：

(1) 根据数据格式特征情况选择合适的统计方法进行统计(stats 包以及 car 包)(如果不满足统计要求将不会进行统计分析)，进而分析数据变量之间相关性和两组之间的，用 gglot2 可视化结果

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. 能否超过 2 个组？

答：该模块就是两个组之间的比较，如果有多个组，建议使用不带分组比较的【相关性散点图-分组】模块

5. 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