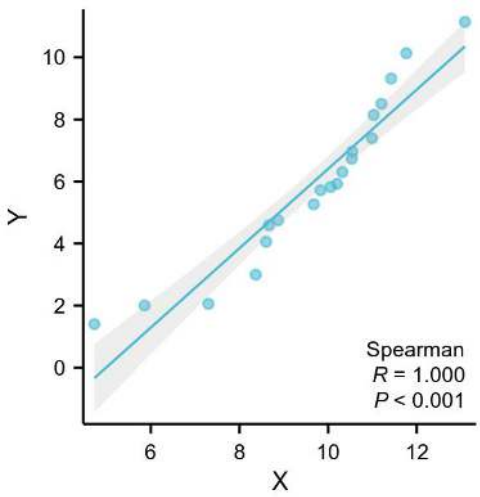
基础绘图 - 相关性散点图



网址：https://www.xiantao.love

更新时间：2023.02.23

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 个或者 2 个变量之间的相关性

应用场景

相关性散点图常用来进行数据的对比

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）
  - 数据每一列都代表一个变量/样本，都需要是数值类型的数据
    - 数据第 1、2 列都是数值类型数据，这时候第 1 列对应到相关性散点图的 x 轴，第 2 列对应到散点图的 y 轴
    - 如果数据上传有 3 列（上传数据最多支持 3 列数据），这时候第 1 列对应到相关性散点图的 x 轴，第 2、3 列都对应到散点图的 y 轴
  - 数据中不能含有非数值及其他非法字符
  - ……

	A	B
1	X	Y
2	4.72591	1.40529
3	5.857858	2.004501
4	7.298382	2.05869
5	8.367185	2.999211
6	8.601069	4.058117
7	8.666156	4.593494
8	8.869726	4.752448
9	9.675064	5.268196
10	9.828603	5.717856
11	10.05634	5.826166
12	10.20007	5.924474
13	10.31928	6.302995
14	10.53395	6.731148
15	10.54377	6.967591
16	10.98956	7.391441
17	11.02636	8.143003
18	11.20207	8.501662

- 数据处理
  - 对数据中每一列非数值类型的数据进行处理
    - 所有变量/列都需要纯数值类型的数据
    - 不能有非数值，特殊值(特殊符号等)
    - 每一个变量不能都是一个值
- 分析：
  - 统计描述
    - 对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	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
X	20	4.7259	13.086	10.128	2.3489	8.6499	10.999	9.6517	2.0069	0.44875
Y	20	1.4053	11.143	5.8753	3.1197	4.4596	7.5793	5.962	2.6992	0.60357

统计描述.xlsx

  - 正态性检验
    - 对第 1、2 列或者第 1、2、3 列数据进行正态性检验
  - 异常值分析
    - 对变量进行异常值分析

- 相关性分析
  - 将处理(清洗)后的数据进行相关性分析
    - 主要变量（数据第 1 列）与其它变量（数据第 2 列开始的列）之间
      - 如果有数据上传有 3 列，则将第 1 列分别与第 2、3 列数据进行相关性分析，如果只有 2 列数据，则将会对第 1、2 列进行分析
    - 相关性分析表
      - 包含不同方法（Pearson、Spearman）计算的相关性系数值与统计学 p 值等

相关性分析

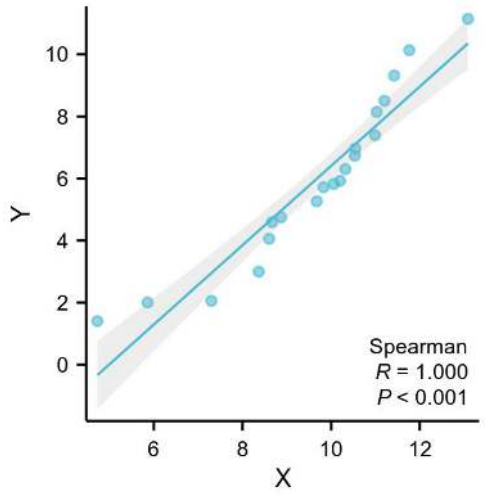
同时提供Pearson和Spearman统计方法，可以根据需要选择标注在图中的方法

方法	组别I	组别J	自由度(df)	统计量	相关系数	置信区间(95%CI)	p值
Pearson	X	Y	20	12.95	0.9503	0.87628 - 0.9805	1.47e-10
Spearman	X	Y	20	0	1		5.98e-06

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

- 将分析后得到的相关性系数与 p 值进行后续的相关性热图可视化

结果解读



- 横坐标表示第 1 列变量
- 纵坐标表示第 2 列变量（如果数据上传 3 列数据，则表示第 2、3 列的值）
- 图中的线为拟合线，拟合线周围的阴影部分为置信区间
- 图中右下角
  - “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

数据格式

	A	B
1	X	Y
2	4.72591	1.40529
3	5.857858	2.004501
4	7.298382	2.05869
5	8.367185	2.999211
6	8.601069	4.058117
7	8.666156	4.593494
8	8.869726	4.752448
9	9.675064	5.268196
10	9.828603	5.717856
11	10.05634	5.826166
12	10.20007	5.924474
13	10.31928	6.302995
14	10.53395	6.731148
15	10.54377	6.967591
16	10.98956	7.391441
17	11.02636	8.143003
18	11.20207	8.501662

数据要求：

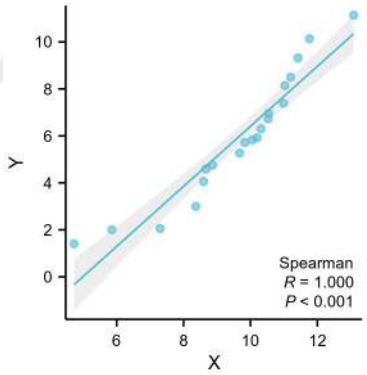
- 至少 2 列数据，每列至少 3 个观测（即至少 3 行数据），数值类型，最多支持 3 列和 5000 行数据
  - 数据每一列都代表一个变量/样本，都需要是数值类型的数据
    - 数据第 1、2 列都是数值类型数据，这时候第 1 列对应到相关性散点图的 x 轴，第 2 列对应到散点图的 y 轴
    - 如果数据上传有 3 列，这时候第 1 列对应到相关性散点图的 x 轴，第 2、3 列都对应到散点图的 y 轴
  - 数据中不能含有非数值及其他非法字符
  - 每一个变量不能都是一个值
- 变量名（列名）不能重复

参数说明

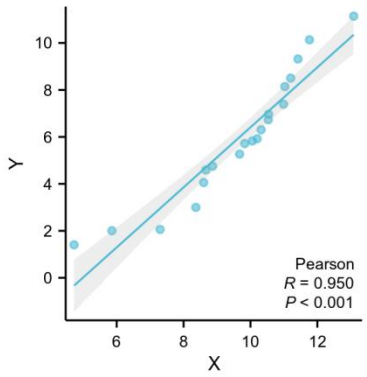
（说明：标注了颜色的为常用参数。）

数据处理

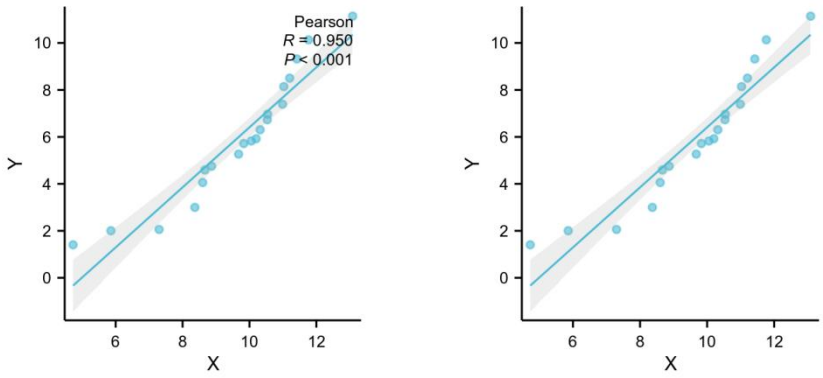
- 统计：可以选择主要变量与其他变量间进行相关性分析的方法
  - spearman: Spearman(默认)为非参数检验方法，数据可以不需要满足正态性



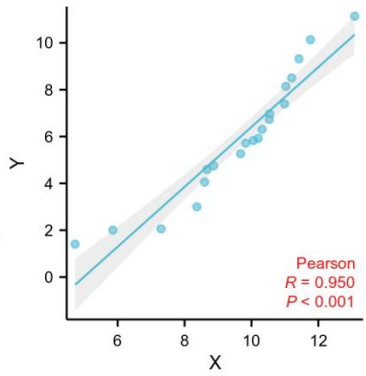
  - pearson: Pearson 为参数检验方法，数据需要满足双正态



- 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，默认在图形的右下，还可以选择左下、左上、右上、无(不进行标注)，如下：左侧为右上，右侧为无

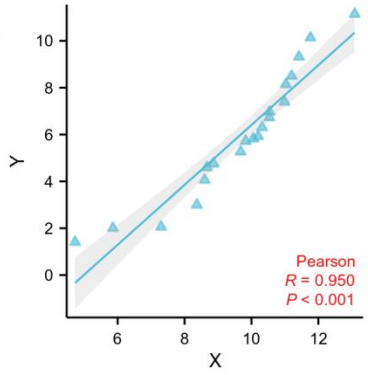


- 标注颜色：当图形中进行标注的时候，可以修改标注的颜色，如下：



点

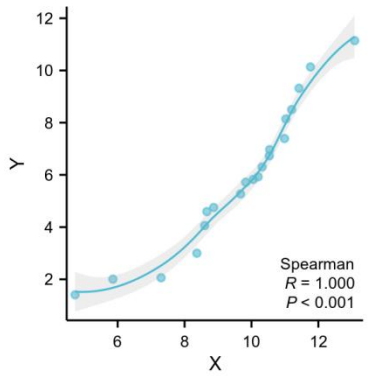
- 填充色：可以修改图中各点的填充颜色
- 描边色：可以修改图中各点的描边颜色
- 样式：可以修改图中各点的样式（形状），默认为圆形，还可以选择正方形、菱形、三角形、倒三角形，如下：



- 大小：可以修改图中个点的大小比例，默认为 1
- 不透明度：可以修改图中各点不的透明度，1 表示完全不透明,0 表示完全透明

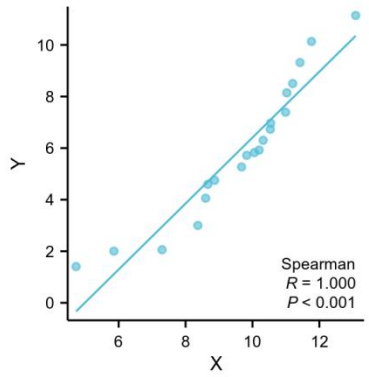
拟合线

- 展示：可以选择是否进行展示拟合线的操作，默认为展示
- 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，还可以选择虚线
- 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：可以修改拟合线线条的不透明度，1 表示完全不透明，0 表示完全透明

仙桃学术

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

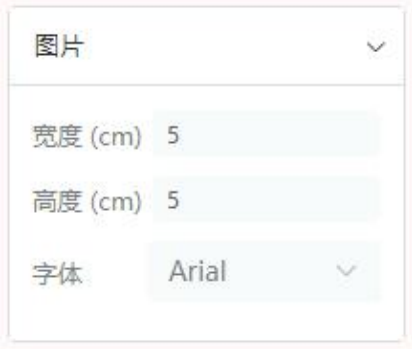
补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如｛{2}｝；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]

风格

- 边框：可以选择是否展示图片边框，默认展示
- 网格：可以选择是否展示网格，默认不展示
- 文字大小：控制整体文字大小，默认为 7pt

仙桃学术

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

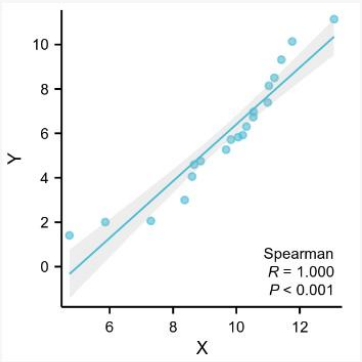
结果说明

主要结果

相关性散点图

相关性散点图: 分析1个变量和另外1个或者2个变量之间的相关性

统计方法: Spearman



相关性散点图.pdf　相关性散点图.tiff　相关性散点图.pptx

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

补充结果 - 统计描述

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
X	20	4.7259	13.086	10.128	2.3489	8.6499	10.999	9.6517	2.0069	0.44875
Y	20	1.4053	11.143	5.8753	3.1197	4.4596	7.5793	5.962	2.6992	0.60357

统计描述.xlsx

这里提供各个变量对应常见「统计描述指标」：最小值、最大值、中位数、标准差等

补充结果 - 异常值分析

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
X	4.72590959254947	

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

这里统计各变量的离群值、异常值情况

- 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)

补充结果 - 统计描述

正态性检验(Shapiro-Wilk normality test)

组别	自由度(df)	统计量	p值
X	20	0.9388	0.2276
Y	20	0.97954	0.9280

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)

这里提供各变量的正态性检验

- 变量接近正态分布(P ＞ 0.05)，建议选择用参数检验方法(Pearson)

补充结果 - 统计描述

相关性分析

提供Pearson和Spearman统计方法的结果

主变量	次变量	自由度(df)	统计量-Pearson	相关系数-Pearson	p值-Pearson	统计量-Spearman	相关系数-Spearman	p值-
Gene1	Gene2	48	3.99705	0.499724	0.0002	1.039e+04	0.500984	
Gene1	Gene3	48	4.98716	0.584217	8.42e-06	7818	0.624586	2
Gene1	Gene4	48	3.68768	0.469858	0.0006	1.131e+04	0.456711	
Gene1	Gene5	48	3.71781	0.472841	0.0005	1.076e+04	0.483409	
Gene1	Gene6	48	-2.65895	-0.358304	0.0106	2.778e+04	-0.333974	
Gene1	Gene7	48	-4.23575	-0.521615	0.0001	3.202e+04	-0.537479	7
Gene1	Gene8	48	-4.39615	-0.535773	6.08e-05	3.186e+04	-0.529988	9
Gene1	Gene9	48	-5.74211	-0.638124	6.2e-07	3.373e+04	-0.619592	7
Gene1	Gene10	48	-4.34108	-0.530962	7.28e-05	3.196e+04	-0.534886	8

相关性.xlsx

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;

这里提供相关性分析表：可以查看第 1 列（变量）与第 2 列(第 2、3 列)之间的相关系数与其对应的统计学 p 值

- 相关系数为正数，说明两个变量（主要变量与其他变量）之间可能存在正相关关系；相关系数为负数，说明两个变量可能存在负相关关系

- 相关系数绝对值在 0.8-1.0 之间，说明两个变量之间强相关
- 相关系数绝对值在 0.5-0.8 之间，说明两个变量之间中等程度相关
- 相关系数绝对值在 0.3-0.5 之间，说明两个变量之间相关程度一般
- 相关系数绝对值在 0.0-0.3 之间，说明两个变量之间弱相关或者不相关

➢ 相关是否有统计学意义还需要结合 p 值来查看

仙桃学术

方法学

统计分析和可视化均在R 4.2.1版本中进行

涉及的R包：ggplot2包（用于可视化）

处理过程：

(1) 对数据中主变量和次要变量之间进行相关性分析

(2) 分析结果用ggplot2包进行棒棒糖图可视化

仙桃学术

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。

仙桃学术

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 会要求数据是满足正态性，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