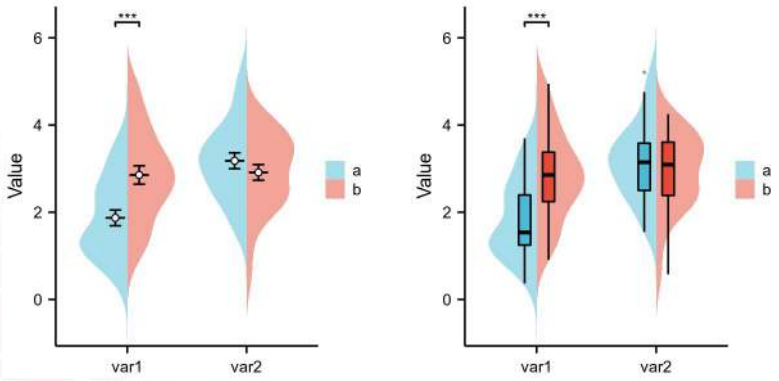
基础绘图 - 豆荚图



网址：https://www.xiantao.love

更新时间：2023.03.15

目录

基本概念

- 豆荚图：一种形似豆荚的图，分别用豆荚的一半代表一个分组下两种因素在数值上的差异。
- 统计方法：统计要求每组样本都要满足 3 个样本以上，并且每组样本的方差不能为 0，如果不满足条件，就不会进行统计分析。
  - 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。
  - Weltch`t 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

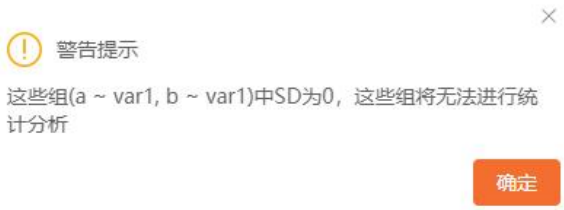
展示多个分组在两种因素下的差异情况：

比如：

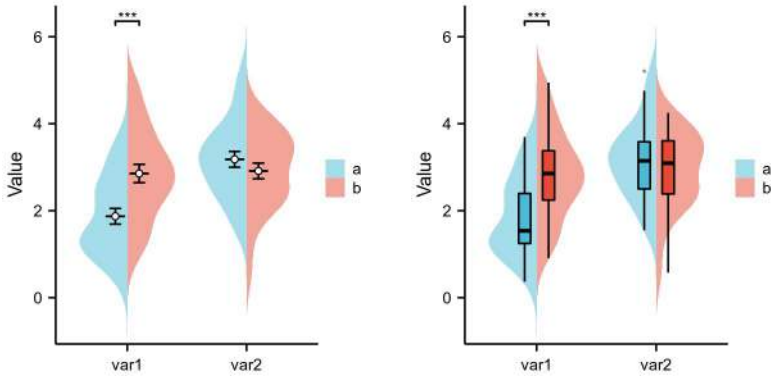
- 多个疾病（泛癌）的正常和疾病 2 组之间某个分子表达的差别
- 多个分子在单个疾病的正常和疾病组 2 组之间的表达差别

注意：如果不满足条件，就不会进行统计分析！如下：

	A	B	C
1	group	var1	var2
2	a	1	3.15096616
3	a	1	2.49989596
4	a	1	3.14322551
5	a	1	4.75935657
6	a	1	5.21034391
7	a	1	4.51257442
8	a	1	1.54841608
9	a	1	3.13828799



主要结果

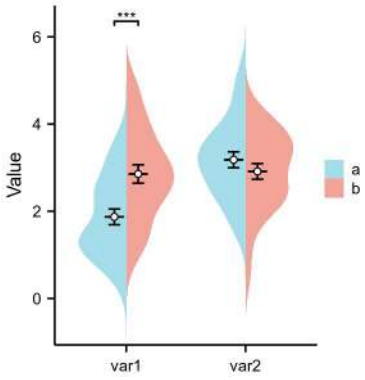


- 图中 x 轴代表不同的变量（第二列及以后）和分组（第一列），y 轴代表变量对应的数据。
- 图注（legend）代表分组，如上图 a 和 b 两种因素（情况）。
  - 研究目的：分别在 var1 和 var2 中 a 和 b 两种因素的数据差异情况。
  - 两个因素分别用一半的豆荚（一半的小提琴）区分。
- 组合可视化形式：需要选择对应参数中【展示】

  - 箱式图
    - 箱式图：常见分组比较图之一，箱子中间的横向代表中位数，箱子的上下边代表上四分位（75 百分位数）和下四分位（25 百分位数）。一般而言，箱子的上方和下方的线，如果分组内不存在离群值 (Q1-1.5*IQR or Q3+1.5*IQR，下四分位-1.5 倍四分位距)，那么线的最远位置就为最小值或者最大值。箱子的上方或者下方的点代表离群值的点。

数据格式

	A	B	C
1	group	var1	var2
2	a	3.2637216	3.15096616
3	a	3.03516659	2.49989596
4	a	0.76702972	3.14322551
5	a	1.30640926	4.75935657
6	a	2.3414782	5.21034391
7	b	3.73937622	2.10098765
8	b	4.68385315	2.38422873
9	b	2.12311822	3.33913133
10	b	1.45094024	3.63745763
11	b	2.45705866	3.09311945



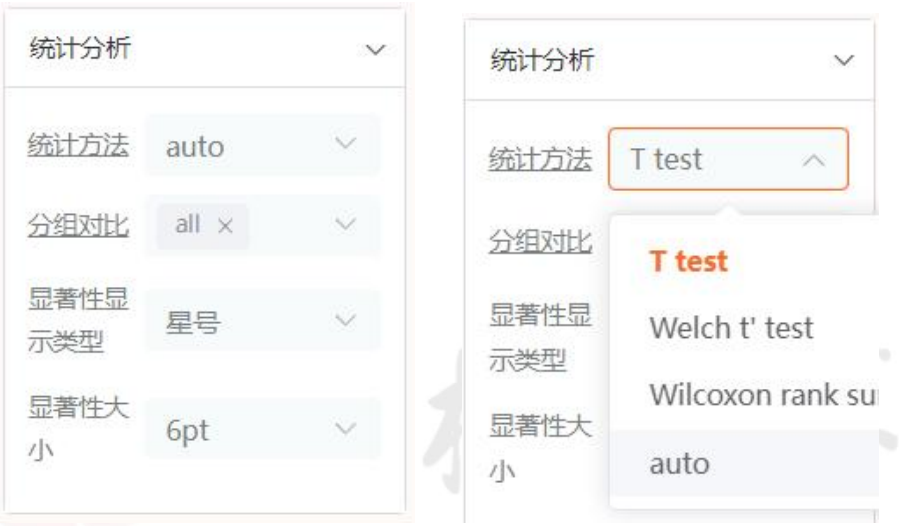
数据要求：

- 第 1 列数据为二分类变量，代表两种因素。两种因素至少需要有 3 行数据（3 个观测重复）。
- 第 2 列以及之后的列，为具体每个组（列）在两种因素下对应的数值情况，每一列均需要是数值类型。在每个因素下至少要有 3 个重复（如果每个因素不足 3 个，则对应的这个分组不会进行统计检验，不会有显著性 p 值的结果）。
- 图中的顺序（x 轴）与上传数据中的列名顺序保持一致，若需要调整图中组的顺序，需要在上传数据内进行调整，然后再上传数据。
- 如果不同的组（列）在两种因素下不是规整的，可以用空格代表每个分组在不同因素中的缺失，以满足数据整理的需要。
- 最多 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。

参数说明

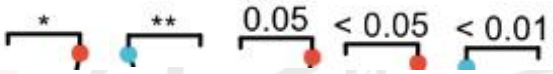
（说明：标注了颜色的为常用参数。）

统计分析



- 统计方法：统计方法默认为 auto（自动选择），当第一次点击确认分析后，会自动替换成适合于对应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。
- 分组对比：统计学差异标注的分组信息，默认为 all（全部都标注）。当第一次点击确认分析后，会自动替换成对应上传数据的分组（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！）。

- 显著性显示类型：影响分组比较中显著性标注，默认为星号。可选择星号或者 p 值以及其他形式，可以选 星号、p 值科学计数法、p 值数值(小于 0.05 自动<)、p 值数值(小于 0.001 自动<)、p = 科学计数、p = 数值(小于 0.05 自动<)、p = 数值(小于 0.001 自动<)、无。



- 显著性大小：可以修改显著性标注的大小。

间距设置

- 组间距离：两组之间的宽度，只有在二维数据(含 legend)的时候才会有效果。主要控制单个分子两组之间的距离。

豆荚

- 填充色：豆荚的填充色颜色选项，上传数据中第 1 列二分类的取色，最多支持修改 2 个颜色。受配色方案全局性修改。
- 描边色：豆荚的描边色颜色选项，上传数据中第 1 列二分类的取色，最多支持修改 2 个颜色。不受配色方案全局性修改。
- 描边粗细：豆荚描边的粗细，默认为 0.00pt。
- 不透明度：豆荚的透明度。0 为完全透明，1 为完全不透明。
- 宽度：豆荚的宽度控制，默认 0.8。

箱

- 展示：可选是否展示。
- 填充色：箱子的填充色颜色选项，上传数据中第 1 列二分类的取色，最多支持修改 2 个颜色。受配色方案全局性修改。
- 描边色：箱子的描边色颜色选项，上传数据中第 1 列二分类的取色，最多支持修改 2 个颜色，默认黑色。不受配色方案全局性影响。
- 描边粗细：箱子描边的粗细，默认为 0.75pt。
- 不透明度：箱子的透明度。0 为完全透明，1 为完全不透明
- 箱子宽度：箱子的宽度控制，默认 0.6。

误差线

误差线只有在没有箱式图时才会显示（箱式图本身自带类似误差线）。

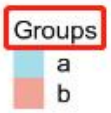
- 展示：可选是否展示。
- 类型：可选均值±标准差、均值±标准误、中位数~上下四分位，建议选择均值±标准差。
- 颜色：误差线颜色，默认为纯黑，不受配色方案全局性影响。
- 描边粗细：误差线粗细，默认为 0.75pt
- 宽度：误差线的宽度。

标题

- 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)

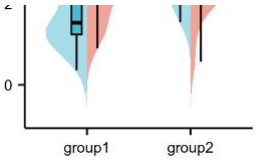
- 展示：是否展示图注
- 图注标题：可以添加图注标题，如下：



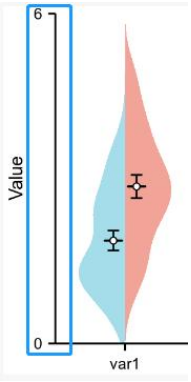
- 图注位置：可选右、上，默认为右。

坐标轴

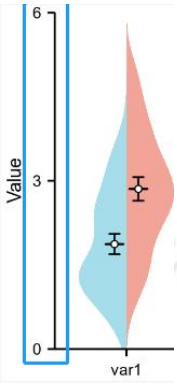
- X 轴分组名：支持直接修改 x 轴各个分组的名字，每个名字之间需要用英文输入法的逗号隔开，比如 group1, group2，即修改变量（列）名。这里支持换行，需要换行的位置可以插入\n



- X 轴标注旋转：支持对 x 轴文字进行旋转。适合于 x 轴文字过长的时候
- Y 轴范围+刻度：用于修改 y 轴范围以及刻度，如果需要分割，需要用小括号(英文输入法)隔开，数值间需要用逗号隔开，例如(1,1,2,5,5)。如果调整过大可能会无作用。
  - 如果只是想要修改范围，可以只输入两个范围值，比如 0,0,6,6：



- 如果同时想要修改范围+刻度，可以输入比如：0,0,3,6,6。注意，此时最大和最小值会被当做范围值，不会作为刻度，如果需要刻度，需要类似于 0 和 6 那样同时写两次：



风格

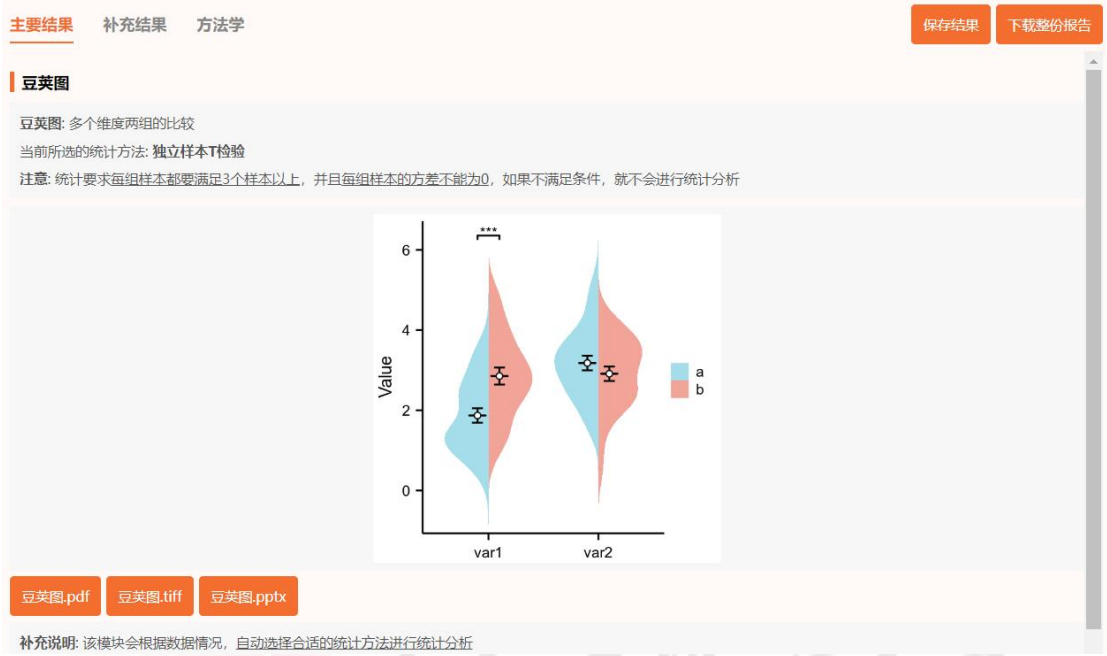
- 边框：是否添加外框
- 网格：是否添加网格
- xy 颠倒：可以颠倒 xy 轴
- 文字大小：针对图中所有文字整体的大小控制

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 格式下载，结果报告可以下载包括 pdf 以及说明文本的内容。

- 如果数据可以进行统计分析，将会进行统计分析。统计分析默认是根据数据情况选择合适的统计方法。统计要求每组样本都要满足 3 个样本以上，并且每组样本的方差不能为 0，如果不满足条件，就不会进行统计分析。

补充结果

统计描述

各个组常见「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)
var1	a	25	0.36674	3.6942	1.5358	1.1495	1.2469	2.3963	1.8713	0.90083
var1	b	25	0.90605	4.9463	2.8549	1.1323	2.2446	3.3769	2.855	1.0551
var2	a	25	1.5484	5.2103	3.1459	1.0837	2.4999	3.5836	3.1808	0.90899
var2	b	25	0.57402	4.2506	3.0931	1.222	2.3842	3.6062	2.9136	0.89828

统计描述.xlsx

此表格提供统计描述的结果，提供 EXCEL 格式下载。

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
var2	a	5.21034391087301	

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

此表格异常值情况表，可以判断数据是否存在异常值。

正态性检验

检验方法: Shapiro-Wilk normality test

组别1	组别2	自由度(df)	统计量	p值
var1	a	25	0.94865	0.2337
var1	b	25	0.98536	0.9670
var2	a	25	0.98182	0.9186
var2	b	25	0.94814	0.2276

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

此表格为正态性检验的结果。

方差齐性检验

检验方法: Levene's test

· Base on Mean

组别	自由度1(df1)	自由度2(df2)	统计量	p值
var1	1	48	0.10018	0.7530
var2	1	48	0.082579	0.7751

方差齐性检验显示，各组观测变量的方差相等(P > 0.05)

此表格为方差齐性的结果。

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	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
var1	a	b	48	3.5456	0.98377	0.42589 - 1.5417	0.0009
var2	a	b	48	-1.0457	-0.26727	-0.78117 - 0.24663	0.3009

p值满足<0.05时，可认为两组存在统计学上差异

此表格为 2 组比较统计检验的结果。

（注意：不同的统计方法会有不一样的统计检验的表格）

方法学

所有分析和可视化均在 R 4.2.1 中进行

涉及的 R 包：ggplot2 包（用于可视化）、stats、car（用于统计分析）

处理过程：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)，用 ggplot2 包对数据进行可视化。

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. 为什么图片中的内容被压缩了?

答:

由于文字不会被压缩，如果左侧的文字很多，就会压缩右侧图的内容而导致坐标轴文字重叠。解决方案可以是:

① 增加图片宽度;

② 颠倒 xy，同时增加图片高度。

2. 如何修改 x 轴文字内容或者顺序顺序?

答:

x 轴的文字的内容和顺序和上传数据每一列都是对应的，列名就是 x 轴的文字。所以，如果想要修改顺序，请在上传数据中进行修改。