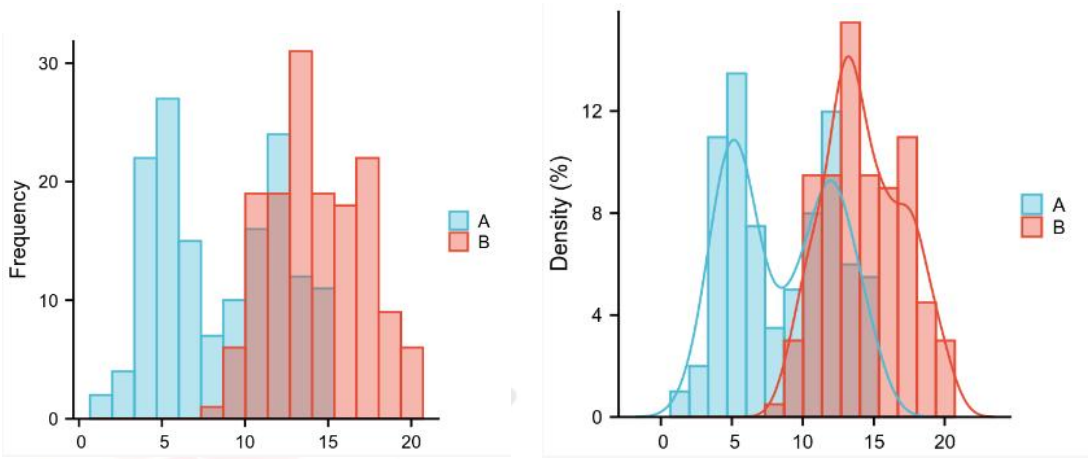
基础绘图 – [数据分布] – 频数直方图



网址：https://www.xiantao.love

更新时间：2023.06.13

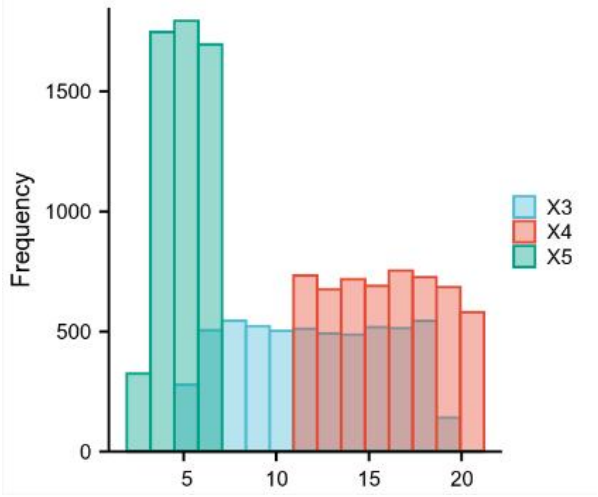
目录

基本概念

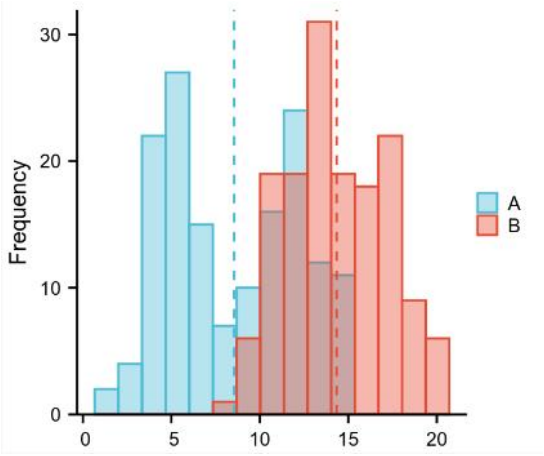
- 直方图：将数据分布情况用柱子的高低来表示。数据按照固定的区间间隔分割数据，并将结果以柱状图的形式展现来描述数据的分布情况，在一定区间内数据越多，柱子越高。
- 分布山峦：将数据分布情况用峰的高低来表示，分布越密集的区间，峰越高。
- 频数分布直方图：是统计了每个区间内的数据个数并使用直方图进行可视化。
- 频率分布直方图：是计算了每个区间内的数据频率，可以使用直方图（或可以添加山峦分布）进行可视化。

应用场景

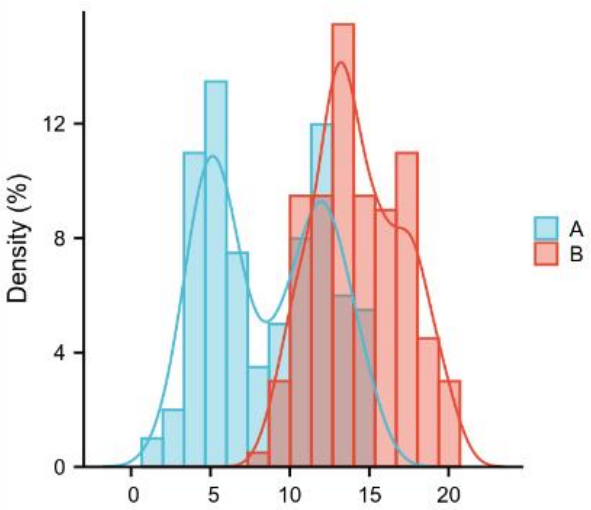
- 查看数据分布情况和特点
- 比较两组或者多组的数据分布情况，如下：



- 展示数据的频数分布情况，包括数据的均值或中值位置，如下：



- 展示数据的频率分布情况，如下：



- 其他…

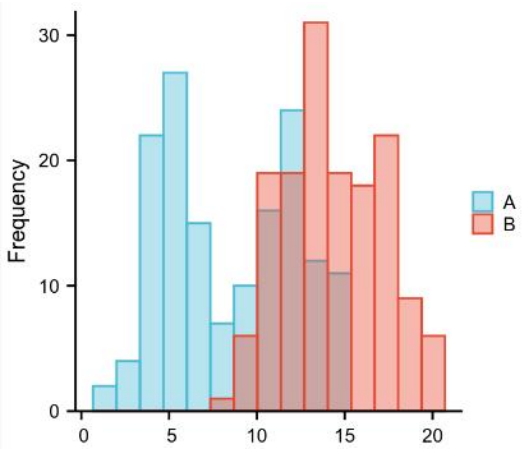
分析过程

上传数据 ⟹ 数据处理(清洗)⟹可视化

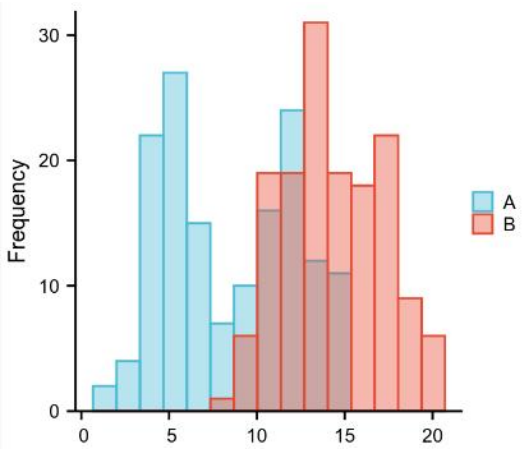
- 数据格式：（具体数据格式要求可以看后面过程的"数据格式"部分）
  - 数据第 1 列必须为数值类型，对应第一组直方图
  - 数据第 2 列必须为数值类型，对应第二组直方图
  - 数据第 3 列及以后必须均是数值类型

1	A	B
2	4.332394	13.15688
3	4.301052	10.50695
4	4.843305	10.03155
5	7.112503	13.13563
6	3.02587	12.47698
7	5.367813	8.970119
8	6.807115	11.28535
9	4.33895	14.35695
10	5.63424	11.18921
11	6.447295	11.11568
12	6.51166	13.64312
13	5.075572	10.99032
14	5.915049	12.16442
15	3.399208	13.59889
16	5.452444	12.24917

- 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
  - 数值类型数据只能是纯数值类型数据，不能包含非数值与不规则的值
  - ……
- 可视化：将清洗后的数据进行 ggplot2 包可视化（图形默认转置后的）



结果解读



- 频数直方图横向坐标表示样本数据的范围大小，对应数据第 1、2 列
- 频数直方图纵向坐标表示样本数据在各区间中的计数
- 频数直方图的柱子颜色表示不同样本，对应数据的第 1、2 列的列名
- 可以直观比较不同样本数据情况

数据格式

一维频数分布直方图

	A
1	A
2	4.332394
3	4.301052
4	4.843305
5	7.112503
6	3.02587
7	5.367813
8	6.807115
9	4.33895
10	5.63424
11	6.447295
12	6.51166
13	5.075572
14	5.915049
15	3.399208
16	5.452444

数据要求：

- 一列代表 1 个变量（一个组直方图的数据），数据至少需要 1 列，2 行，每一列均需要是数值类型。
- 纵坐标（y 轴）是根据选择可视化中的分布类型（频数分布/频率分布），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。
- 数据最多支持 10000 行，6 列，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。
  - 数值类型数据只能是纯数值类型数据，不能包含非数值与不规则的值
- 数据每一列列名不能重复

三维频数分布直方图

	A	B	C
1	X3	X4	X5
2	14.59161	11.46371	3.22493
3	17.19045	14.70113	5.4854
4	6.62201	18.15868	6.3791
5	13.87005	14.11478	5.62329
6	11.50642	11.26155	5.95598
7	5.177	12.70742	6.48615
8	14.34853	19.47203	6.34862
9	11.78924	13.35394	6.37359
10	11.09413	11.22307	6.27067
11	8.51632	14.50618	4.27223
12	9.1326	16.97389	4.09665
13	9.8947	16.14888	4.36482
14	7.8373	13.86424	3.14554
15	6.37554	12.34579	4.58128
16	17.30645	19.93294	6.83041

数据要求：

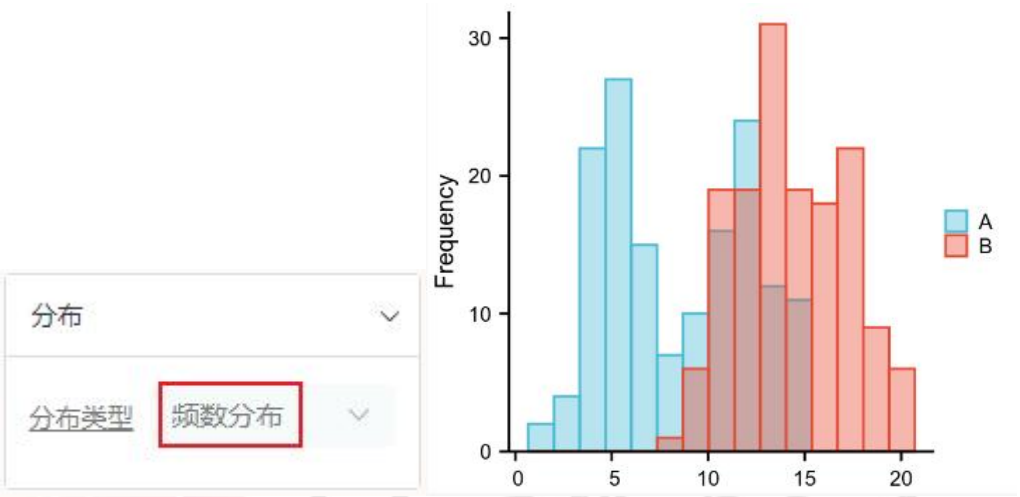
- 一列代表 1 个变量（一个组直方图的数据），数据至少需要 1 列，2 行，每一列均需要是数值类型。
- 纵坐标（y 轴）是根据选择可视化中的分布类型（频数分布/频率分布），若需要调整图中组的顺序，需要在上传数据内进行调整，然后再上传数据。
- 数据最多支持 10000 行，6 列，若验证数据时返回报错，需要在上传数据内进行相应的调整，然后再上传数据。
  - 数值类型数据只能是纯数值类型数据，不能包含非数值与不规则的值
- 数据每一列列名不能重复

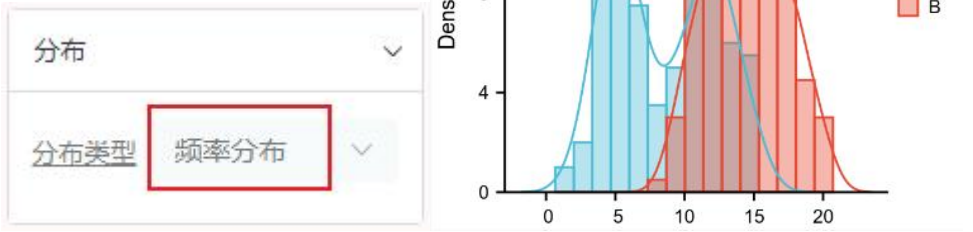
参数说明

（说明：标注了颜色的为常用参数。）

分布

- 样式：可选择频数分布或频率分布。各种样式见“主要结果”部分的展示。

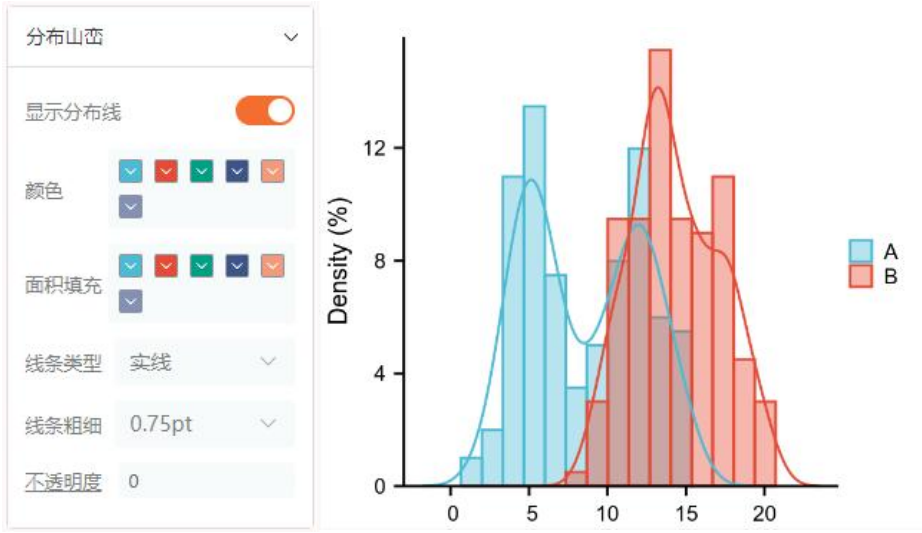




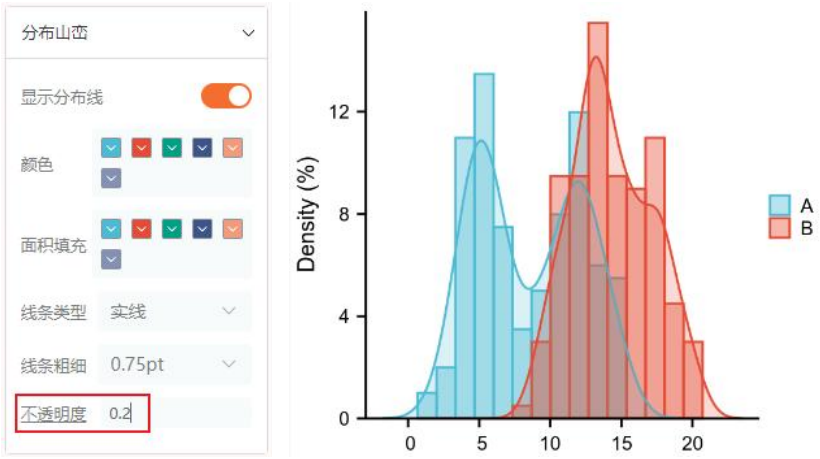
直方图柱子

- 指定柱子的数量：2-60 之间，数字代表指定直方图中所有的柱子数量；若不指定，默认绘制 15 个柱子，若指定的数字超出的范围时，默认绘制 30 个柱子。
- 填充色：柱子的填充色颜色选项，有多少个变量（数据列）会提取多少个颜色，最多支持修改 6 个颜色。受配色方案全局性修改。
- 描边色：柱子的描边色颜色选项，有多少个变量（数据列）会提取多少个颜色，最多支持修改 6 个颜色。受配色方案全局性修改。
- 描边粗细：柱子描边的粗细，默认为 0.75pt。
- 不透明度：柱子的透明度。0 为完全透明，1 为完全不透明。

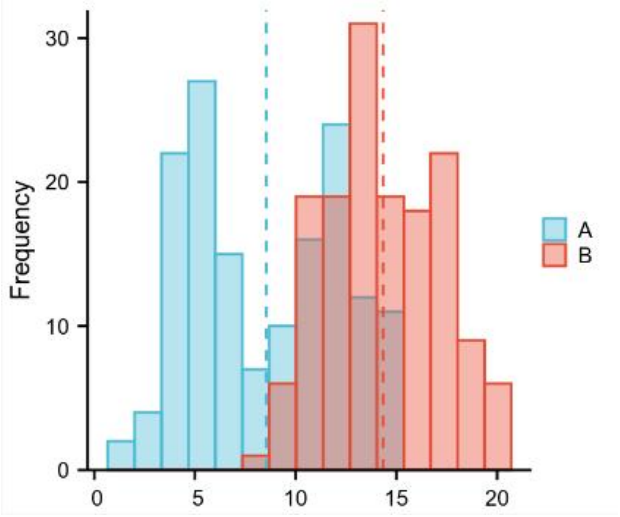
分布山峦



- 显示分布线：选择展示分布线
- 颜色：分布线的描边色颜色选项，有多少个变量（数据列）会提取多少个颜色，最多支持修改 6 个颜色。受配色方案全局性修改。
- 面积填充：分布山峦的填充色颜色选项，有多少个变量（数据列）会提取多少个颜色，最多支持修改 6 个颜色。受配色方案全局性修改。
- 线条类型：分布线的线条类型，默认为实线。
- 描边粗细：分布线描边的粗细，默认为 0.75pt。
- 不透明度：分布山峦的透明度。0 为完全透明，1 为完全不透明。设置大于 0 的值，可以显示分布山峦的填充。

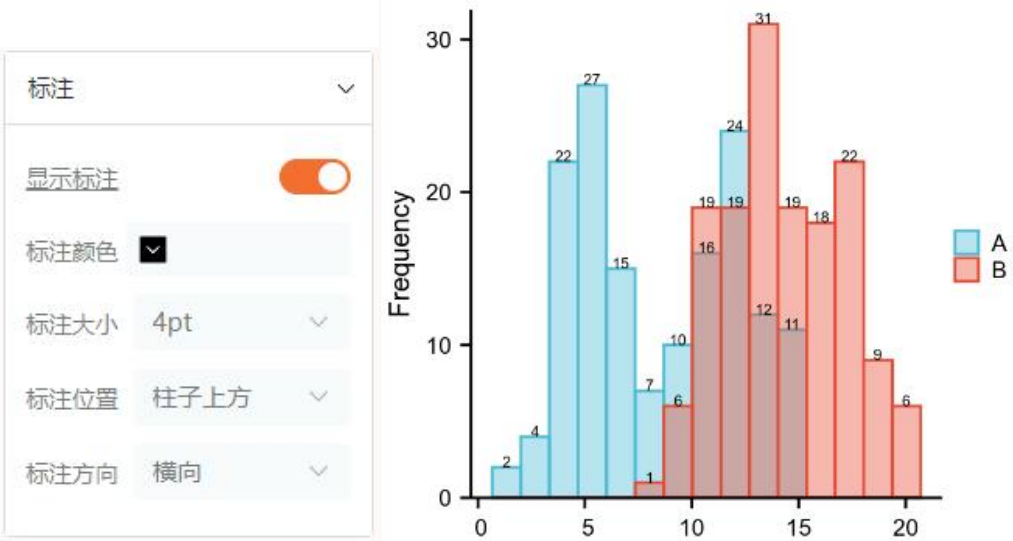


竖线



- 统计类型：可选择均值线或中值线，计算变量的均值或中值统计量。默认不展示。
- 颜色：竖线的颜色选项，有多少个变量（数据列）会提取多少个颜色，最多支持修改 6 个颜色。受配色方案全局性修改。
- 线条类型： 竖线的线条类型，默认为实线。
- 线条粗细：竖线的粗细，默认为 0.75pt。

标注



- 显示标注：选择即显示标注（若是分布类型是频数分布，则展示计数；若是频率分布，则展示频率百分比）
- 标注颜色：可以选择并修改标注的颜色
- 标注大小：可以选择并修改标注字体的大小
- 标注位置：可以选择并修改标注的位置，可选柱子上方、柱子中间和柱子底部
- 标注方向：可以选择并修改标注字体的方向，可选横向和纵向

坐标轴

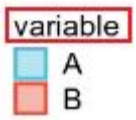
- 是否显示 x 轴：选择即展示 x 轴
- 是否显示 y 轴：选择即展示 y 轴
- x 轴标注旋转：可以选择设置 x 轴标注的倾斜角度
- x 轴范围+刻度：可以控制 x 轴范围和刻度，可只提供 2 个值来控制范围。形如 0.1，0.1，0.2，0.3（最小值和最大值不能超过可视化数据范围 20%，如果调整过大可能会无作用）
- y 轴范围+刻度：可以控制 y 轴范围和刻度，可只提供 2 个值来控制范围。形如 0.1，0.1，0.2，0.3（最小值和最大值不能超过可视化数据范围 20%，如果调整过大可能会无作用）。

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：



- 图注位置：可选择默认、右、上、右上、左上。

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网格线的操作
- xy 颠倒：可以选择是否进行 xy 轴颠倒的操作
- 文字大小：控制整体文字大小，默认为 7pt

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

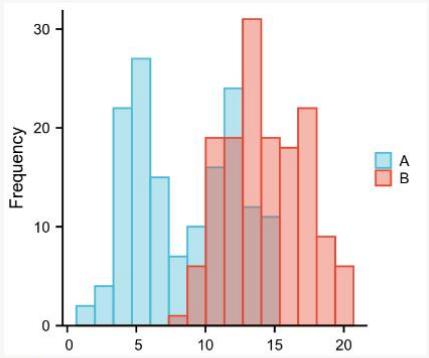
结果说明

主要结果

频数直方图

频数分布图: 用于直观展示1组(或2组)连续变量的分布情况

分布类型: 频数分布



频数分布图.pdf 频数分布图.tiff 频数分布图.pptx

方法学

软件：R (4.2.1)版本

R 包：ggplot2 包（用于可视化）

处理过程：

(1) 对数据进行统计描述，并用 ggplot2 包绘制频数（或频率）分布图

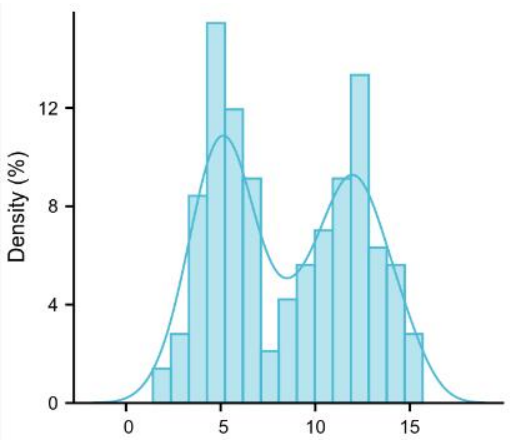
如何引用

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）。

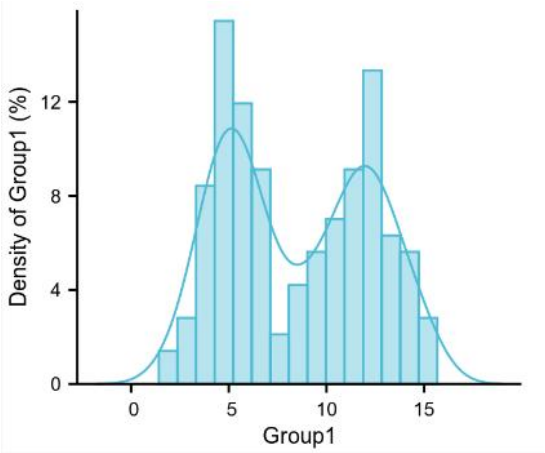
方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

常见问题

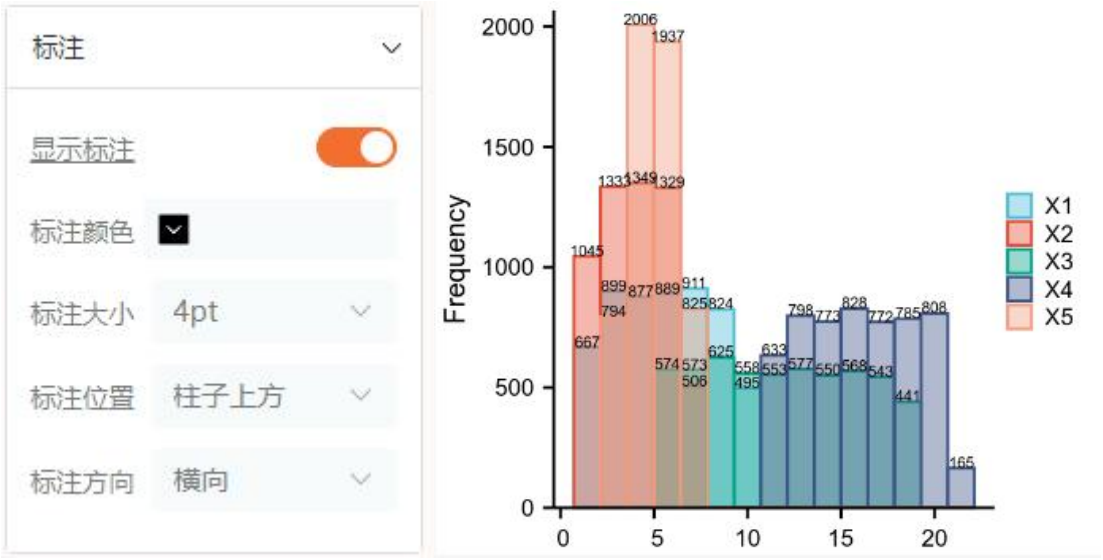
1.为什么绘制一组变量的直方图没有显示图注



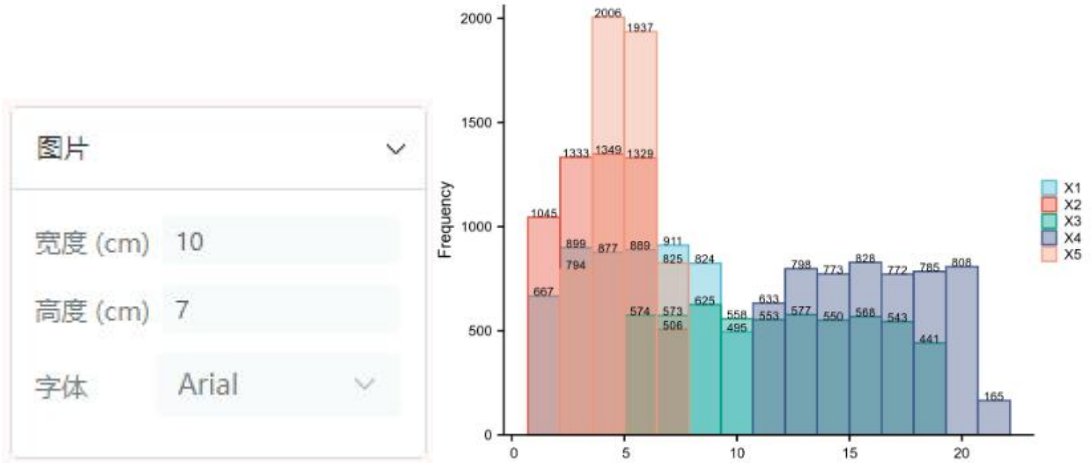
答：当进行一组直方图数据绘制的时候，默认是不显示图注（legend）的，如果需要在图中展示图注，可以用如下方式代替：添加 x 轴的标题，如需修改 y 轴的标题，对应修改即可，再次点击确认



2.展示标注时，由于柱子数过多导致标注有重叠，应该怎么调整



答：当需要调整标注重叠的问题，可以从以下两个方面进行调整：一是调整标注的字体大小（调小），二是调整图片的宽度（调宽），若出现上下标注重叠，则调整图片高度



3.如果需要展示频率分布的分布山峦的面积填充颜色，应该怎么调整

答：当进行频率分布的分布山峦绘制的时候，面积填充颜色默认是完全透明的，如果需要在图中展示，可以设置透明度大于零，例如 0.2，再次点击确认即可