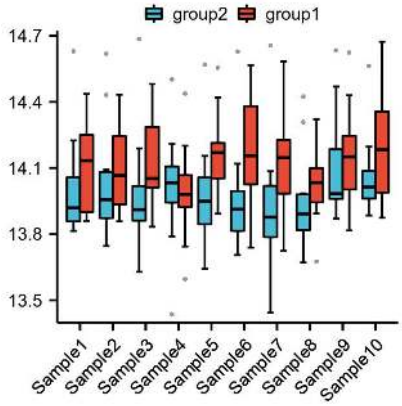
基础绘图 - 分组箱式图



网址：https://www.xiantao.love

更新时间：2023.02.06

目录

基本概念

- 箱式图：使用上边缘、上四分位数、中位数、下四分位数、下边缘五个数值描述数据的分布情况

	A	B	C	D	E	F	G	H	I	J	K
1	group	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
2	group1	14.43636188	14.43300025	14.4432824	14.43701085	14.41977269	14.46709796	14.48099994	14.32142543	14.43066523	14.45823265
3	group1	14.29299069	14.25202468	14.48143951	13.59690255	14.55496229	14.56642111	14.58414193	14.30456888	14.62427126	14.67283322
4	group1	14.27772185	14.22394436	14.23564697	14.20098905	14.21490688	14.22600957	14.24041456	14.10793312	14.27519818	14.25601645
5	group1	14.14295454	14.30086639	14.30292385	14.03514905	14.09055983	14.19273826	14.17351779	14.06926232	14.15765829	14.17658219
6	group1	13.92671864	13.93051056	14.02266363	13.99138622	14.04059609	14.07959465	14.02263273	13.99727614	14.08236834	14.05747679
7	group1	14.16758185	14.04544815	14.06230702	13.91239527	13.97869107	14.4317457	14.18716115	14.0728922	14.14509432	14.3896066
8	group1	14.12292216	14.08694506	14.04240961	13.74380647	14.17052301	14.11817669	13.97116976	13.89361581	14.15728607	14.18928802
9	group1	13.85938921	13.885389	13.83377858	14.07866266	14.16917806	13.91629465	13.72440026	13.95444866	13.8165355	13.96556813
10	group1	13.86449095	13.9474918	14.00836761	13.95490661	13.89266544	14.00858609	14.11948671	13.94115623	13.97614362	13.87459152
11	group1	13.8912195	13.85800071	13.85452968	13.96972217	14.20919283	13.7391234	13.84136325	13.67509655	13.82524619	13.93732931

应用场景

分组箱式图主要用来：展示各组数据的分布情况，比较各组之间的数据分布的差别，比如：

- 芯片数据中每个样本的探针数据分布情况，从而可以判断样本之间数据的校正情况
- 不同组之间数据直接的比较
- ...

分析流程

上传数据 ⟹ 数据清洗 ⟹ 数据处理 ⟹ 将各分组中的样本数据进行分析得到的箱式图结果 ⟹ 进行分组箱式图可视化

- 数据格式：xlsx / csv / txt 文件格式：
  - 第 1 列为分组信息（分类类型）

- 需要提供分组信息这 1 列，并且是分类类型的，不允许出现空的内容，除此之外第 1 列提供的分类变量的数量不能超过 8 个（不能出现 8 个分组及以上的情况，如下：为 2 个分组：group1、group2）

	A	B	C	D	E	F	G	H	I	J	K
1	group	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
2	group2	14.629414	14.618664	14.685406	14.502608	14.568888	14.628779	14.656772	14.424247	14.634583	14.562467
3	group1	14.436362	14.433	14.443282	14.437011	14.419773	14.467098	14.481	14.321425	14.430665	14.458233
4	group1	14.292991	14.252025	14.48144	13.596903	14.554962	14.566421	14.584142	14.304569	14.624271	14.672833
5	group1	14.277722	14.223944	14.235647	14.200989	14.214907	14.22601	14.240415	14.107933	14.275198	14.256016
6	group1	14.142955	14.300866	14.302924	14.035149	14.09056	14.192738	14.173518	14.069262	14.157658	14.176582
7	group2	14.225287	14.051152	14.189386	14.20989	14.155067	14.119611	13.90419	13.847288	14.220998	14.197158
8	group1	13.926719	13.930511	14.022664	13.991386	14.040596	14.079595	14.022633	13.997276	14.082368	14.057477
9	group1	14.167582	14.045448	14.062307	13.912395	13.978691	14.431746	14.187161	14.072892	14.145094	14.389607
10	group1	14.122922	14.086945	14.04241	13.743806	14.170523	14.118177	13.97117	13.893616	14.157286	14.189288
11	group2	14.085357	14.091892	13.95764	14.094817	14.061584	13.941858	14.041562	13.936374	14.077447	14.018555
12	group2	13.865153	13.959009	13.724416	13.789767	14.042269	13.821207	14.085423	13.841002	13.86998	14.102354
13	group1	13.859389	13.885389	13.833779	14.078663	14.169178	13.916295	13.7244	13.954449	13.816536	13.965568

- 第 2 列及以后为数值类型数据，表示每个变量/样本

- 数据清洗：对除了第 1 列外非数值的数据进行清洗
- 数据处理：
  - 根据上传数据第 1 列的分组信息，对所有的样本分成多个组（分组的意思就相当于：有 10 行样本数据，分成 2 组，每 1 个组中每 1 个样本有 5 行样本数据(随机/根据分组信息来分组)，如下所示：）

	A	B	C	D	E	F	G	H	I	J	K
1	group	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
2	group1	14.43636188	14.43300025	14.4432824	14.43701085	14.41977269	14.46709796	14.48099994	14.32142543	14.43066523	14.45823265
3	group1	14.29299069	14.25202468	14.48143951	13.59690255	14.55496229	14.56642111	14.58414193	14.30456888	14.62427126	14.67283322
4	group1	14.27772185	14.22394436	14.23564697	14.20098905	14.21490688	14.22600957	14.24041456	14.10793312	14.27519818	14.25601645
5	group1	14.14295454	14.30086639	14.30292385	14.03514905	14.09055983	14.19273826	14.17351779	14.06926232	14.15765829	14.17658219
6	group1	13.92671864	13.93051056	14.02266363	13.99138622	14.04059609	14.07959465	14.02263273	13.99727614	14.08236834	14.05747679
7	group1	14.16758185	14.04544815	14.06230702	13.91239527	13.97869107	14.4317457	14.18716115	14.0728922	14.14509432	14.3896066
8	group1	14.12292216	14.08694506	14.04240961	13.74380647	14.17052301	14.11817669	13.97116976	13.89361581	14.15728607	14.18928802
9	group1	13.85938921	13.885389	13.83377858	14.07866266	14.16917806	13.91629465	13.72440026	13.95444866	13.8165355	13.96556813
10	group1	13.86449095	13.9474918	14.00836761	13.95490661	13.89266544	14.00858609	14.11948671	13.94115623	13.97614362	13.87459152
11	group1	13.8912195	13.85800071	13.85452968	13.96972217	14.20919283	13.7391234	13.84136325	13.67509655	13.82524619	13.93732931

	A	B	C	D	E	F	G	H	I	J	K
1	group	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
2	group2	14.62941401	14.61866363	14.68540562	14.50260811	14.56888825	14.62877922	14.65677243	14.42424704	14.63458301	14.56246733
3	group2	14.2252866	14.05115246	14.18938608	14.2098896	14.15506733	14.119611	13.90419021	13.84728823	14.22099796	14.19715786
4	group2	14.0853574	14.09189224	13.95763962	14.09481741	14.06158385	13.94185844	14.04156182	13.93637372	14.07744685	14.01855481
5	group2	13.86515332	13.95900892	13.7244157	13.78976657	14.04226875	13.82120729	14.08542286	13.84100191	13.8699799	14.10235421
6	group2	13.85548481	14.43136771	13.62908878	13.43608582	13.64247968	13.70542239	13.44359655	14.30785916	14.46987378	13.91173129
7	group2	13.81541243	13.79427952	13.86380324	13.97760803	13.78246509	13.89230567	13.84161691	13.98431519	13.9621756	14.00978709
8	group2	13.90392537	13.87676998	13.98277894	14.08766028	13.92439892	14.01307924	13.95155947	13.81133974	13.97453862	14.04039036
9	group2	13.81437749	13.87131862	13.86183814	13.97340327	13.8199633	13.81265291	13.85050445	13.96873541	13.95930432	13.88495522
10	group2	13.97468602	13.74638191	14.0297873	14.11074404	13.97551197	13.93466404	13.76820782	13.67162205	13.99249957	13.97855697
11	group2	13.9351656	13.95471433	13.86147588	13.9338107	13.92237408	13.80226021	13.71717123	13.76362164	13.89824692	13.95631059

- 分组箱式图分析：
  - 将经过数据处理的各分组数据进行分析，得到各分组箱式图对应的各部分数据内容（最小值、下四分位数、中位数、上四分位数、最大值），如下：（结果数据可在模块的补充结果--统计描述.xlsx 进行下载）

分组:group1

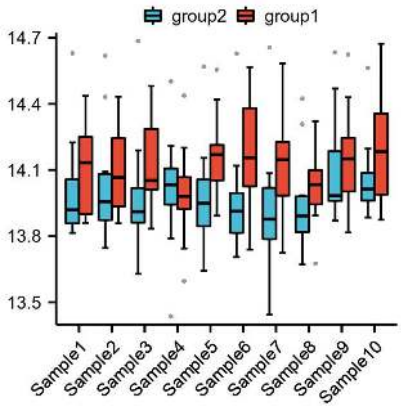
样本	最小值	下四分位	中位数	上四分位	最大值
Sample1	13.859	13.9	14.133	14.25	14.436
Sample2	13.858	13.935	14.066	14.245	14.433
Sample3	13.834	14.012	14.052	14.286	14.481
Sample4	13.597	13.923	13.981	14.068	14.437
Sample5	13.893	14.053	14.17	14.213	14.555
Sample6	13.739	14.026	14.155	14.38	14.566
Sample7	13.724	13.984	14.147	14.227	14.584
Sample8	13.675	13.944	14.033	14.099	14.321
Sample9	13.817	14.003	14.151	14.246	14.624
Sample10	13.875	13.989	14.183	14.356	14.673

分组:group2

样本	最小值	下四分位	中位数	上四分位	最大值
Sample1	13.814	13.858	13.92	14.058	14.629
Sample2	13.746	13.873	13.957	14.082	14.619
Sample3	13.629	13.862	13.911	14.018	14.685
Sample4	13.436	13.944	14.033	14.107	14.503
Sample5	13.642	13.846	13.95	14.057	14.569
Sample6	13.705	13.815	13.913	13.995	14.629
Sample7	13.444	13.787	13.877	14.019	14.657
Sample8	13.672	13.819	13.892	13.98	14.424
Sample9	13.87	13.96	13.984	14.185	14.635
Sample10	13.885	13.962	14.014	14.087	14.562

- 将得到的各分组结果数据进行分组箱式图可视化

结果解读



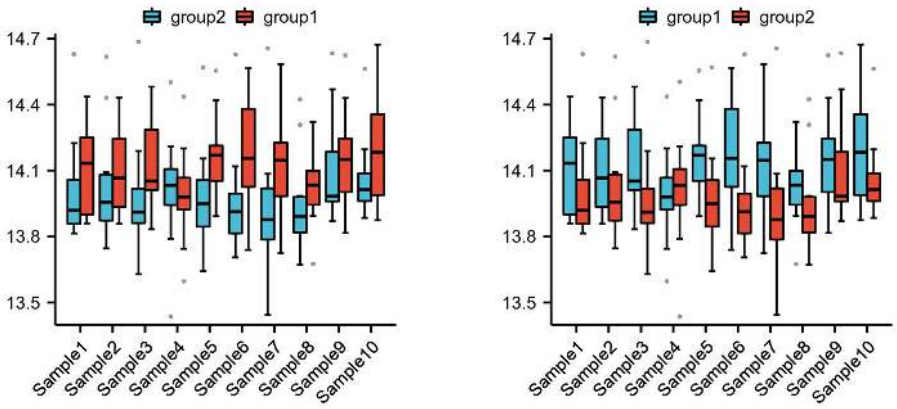
- 分组箱式图：
  - 横坐标表示样本，对应上传数据除第 1 列外的样本/变量/列
  - 纵坐标表示具体的数值（这里的数值不是上传数据中的值，而是经过处分析得到的分组箱式图各部分的值）
- 图中箱式图上边缘和下边缘的点代表离群点（异常值）
- 一种颜色表示一个分组。

数据格式

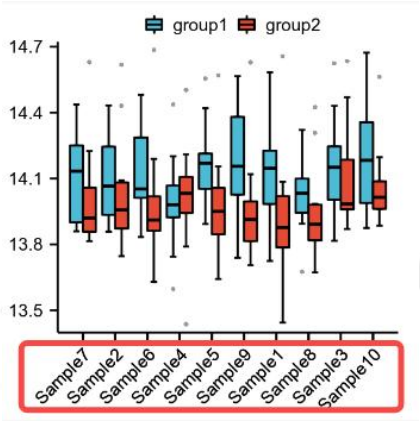
	A	B	C	D	E	F	G	H	I
1	group	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
2	group2	14.629414	14.618664	14.685406	14.502608	14.568888	14.628779	14.656772	14.424247
3	group1	14.436362	14.433	14.443282	14.437011	14.419773	14.467098	14.481	14.321425
4	group1	14.292991	14.252025	14.48144	13.596903	14.554962	14.566421	14.584142	14.304569
5	group1	14.277722	14.223944	14.235647	14.200989	14.214907	14.22601	14.240415	14.107933
6	group1	14.142955	14.300866	14.302924	14.035149	14.09056	14.192738	14.173518	14.069262
7	group2	14.225287	14.051152	14.189386	14.20989	14.155067	14.119611	13.90419	13.847288
8	group1	13.926719	13.930511	14.022664	13.991386	14.040596	14.079595	14.022633	13.997276
9	group1	14.167582	14.045448	14.062307	13.912395	13.978691	14.431746	14.187161	14.072892
10	group1	14.122922	14.086945	14.04241	13.743806	14.170523	14.118177	13.97117	13.893616
11	group2	14.085357	14.091892	13.95764	14.094817	14.061584	13.941858	14.041562	13.936374
12	group2	13.865153	13.959009	13.724416	13.789767	14.042269	13.821207	14.085423	13.841002
13	group1	13.859389	13.885389	13.833779	14.078663	14.169178	13.916295	13.7244	13.954449
14	group2	13.855485	14.431368	13.629089	13.436086	13.64248	13.705422	13.443597	14.307859
15	group1	13.864491	13.947492	14.008368	13.954907	13.892665	14.008586	14.119487	13.941156
16	group2	13.815412	13.79428	13.863803	13.977608	13.782465	13.892306	13.841617	13.984315

数据要求：（可 xlsx、csv、txt 格式）：

- 数据至少 2 列以上，每列至少 4 个观测。
  - 第 1 列为分类类型，表示分组，且提供的分组信息不能超过 8 个（最多 8 个不同分类）
  - 其他列为数值类型
- 最多支持 200 列和 70000 行数据
- 除了第 1 列之外，每 1 行表示一个观测；第 1 列作为分组名，其排列的顺序与上传数据中的顺序一致，如果需要调整，需要手动调整好再重新上传（只要不改动各分组对应的样本的值，那分组顺序没什么影响，只是箱式图各分组对应的颜色不一样）（也就是说，只要不改变第 1 行分组 group2 对应的各样本的值，将第 1 行换到其他行，结果都是没有变化的，唯一变化就是可能箱式图对应的颜色不同了），如下：右侧为改变部分分组所在行的顺序



- 第 2 列以及之后的列，为具体每个分组在两种因素下对应的数值情况。每个在每个因素下至少要有 3 个重复。每一列的列名会作为图中 x 轴标注的名字，如果需要修改 x 轴的名字，需要在上传文件进行修改



- 如果不同的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

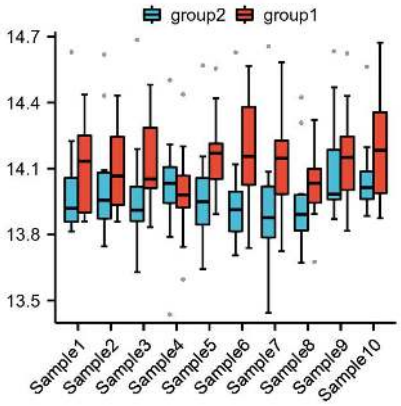
参数说明

（说明：标注了颜色的为常用参数。）

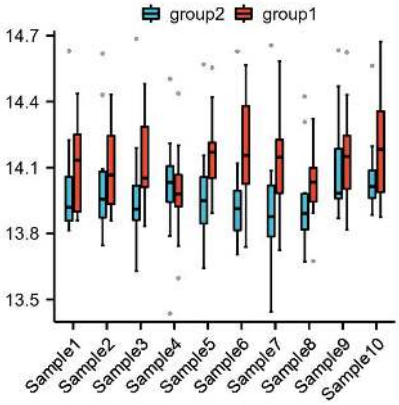
箱

- 填充色：箱子的填充色颜色选项，有多少个分组会提取多少个颜色。受配色方案全局性修改。如果上传数据的分类过多，则此处的配色会失效而自动采取系统默认配色（无法更改）。
- 描边色：箱子的描边色颜色选项，默认为纯黑，只提取第一个颜色。当第一个颜色和填充色的第一个颜色相同时，则不会绘制描边。
- 描边粗细：箱子描边的粗细，默认为 0.75pt，如果设置为 0，则不会有描边。
- 不透明度：箱子的透明度。0 为完全透明，1 为完全不透明。
- 宽度：箱子的宽度
- 组间宽度：同一个变量对应不同分组间的距离（间隔），默认为 0.8，组间宽度设置不能小于箱子宽度，如下：（第一组为默认 0.8，第二组为 0.5）


描边粗细
0.75pt
不透明度
1
宽度
0.8
组间宽度
0.8

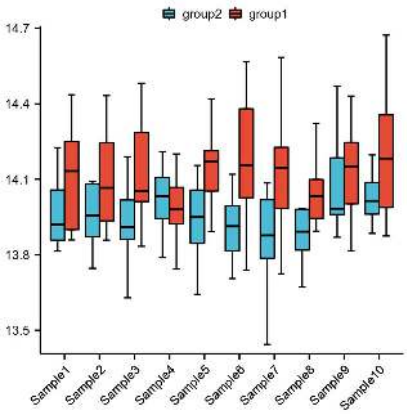

group2
group1
14.7
14.4
14.1
13.8
13.5
Sample1
Sample2
Sample3
Sample4
Sample5
Sample6
Sample7
Sample8
Sample9
Sample10


描边粗细
0.75pt
不透明度
1
宽度
0.4
组间宽度
0.5


group2
group1
14.7
14.4
14.1
13.8
13.5
Sample1
Sample2
Sample3
Sample4
Sample5
Sample6
Sample7
Sample8
Sample9
Sample10

离群点

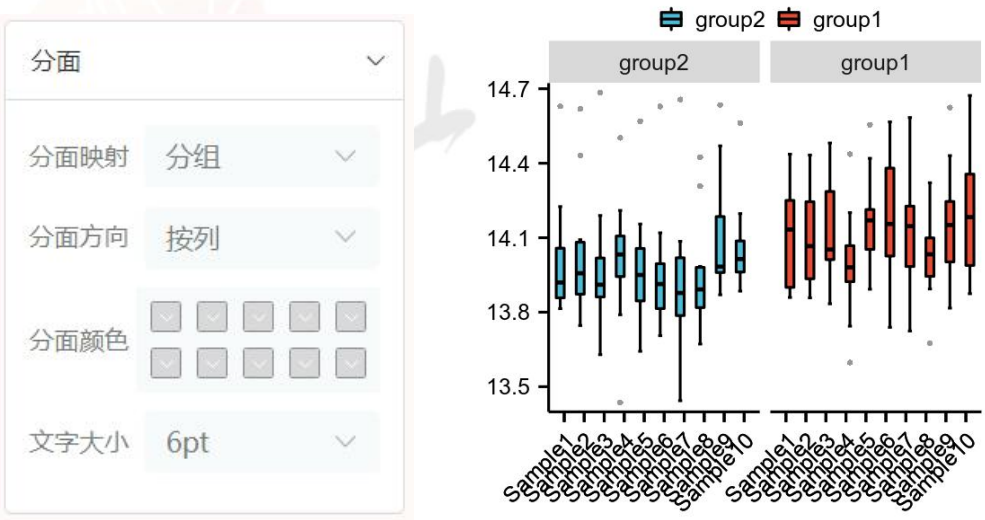
- 展示：可以选择是否展示离群点，默认展示，还可以选择不展示，如下：



- 描边色：在选择展示离群点时，可以修改离群点的颜色
- 大小：在选择展示离群点时，可以修改离群点的大小
- 不透明度：在选择展示离群点时，可以修改离群点的不透明度，默认为 0.4，1 表示完全不透明，0 表示完全透明

分面

- 分面映射：可以选择是否对箱式图进行分面映射操作，默认不映射，还可以选择对分组进行映射，如下：



- 分面方向：可以选择分面映射时，分面的方向，默认为按列进行分面，还可以选择按行进行分面
- 分面颜色：可以选择并修改进行分面映射时各分面的颜色
- 文字大小：可以选择并修改进行分面映射时各分面对应文本字体的大小

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）

- 展示：可以选择是否展示图注操作
  - 选择展示：将会展示图注
- 图注标题：首先选择展示，则可以修改需要上传的图注标题信息
- 图注位置：首先选择展示，则可以选择展示图注的位置

坐标轴

- x 轴标注旋转：可以选择 x 轴标注旋转的角度
- 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 0.1，0.2，0.3（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）

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
- 文字大小：控制整体文字大小，默认为 7pt

图片

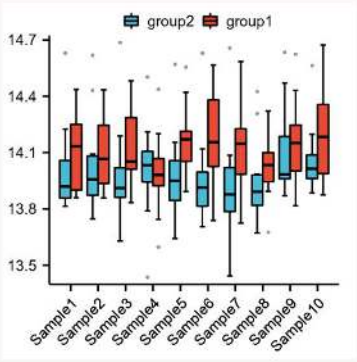
- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

结果说明

主要结果

分组箱式图

分组箱式图: 用箱子绘制每个样本对应的数据情况



分组箱式图.pdf　分组箱式图.tiff

箱子中间的横线代表中位数，箱子的上边代表上四分位，箱子的下边代表下四分位
如果箱子上下存在有黑点，代表此样本存在有离群值

补充结果

统计描述

各分组中各个样本对应下最小值、四分位、中位数、上四分位、最大值

分组:group2

样本	最小值	下四分位	中位数	上四分位	最大值
Sample1	13.814	13.858	13.92	14.058	14.629
Sample2	13.746	13.873	13.957	14.082	14.619
Sample3	13.629	13.862	13.911	14.018	14.685
Sample4	13.436	13.944	14.033	14.107	14.503
Sample5	13.642	13.846	13.95	14.057	14.569
Sample6	13.705	13.815	13.913	13.995	14.629
Sample7	13.444	13.787	13.877	14.019	14.657
Sample8	13.672	13.819	13.892	13.98	14.424
Sample9	13.87	13.96	13.984	14.185	14.635
Sample10	13.885	13.962	14.014	14.087	14.562

分组:group1

样本	最小值	下四分位	中位数	上四分位	最大值
Sample1	13.859	13.9	14.133	14.25	14.436
Sample2	13.858	13.935	14.066	14.245	14.433
Sample3	13.834	14.012	14.052	14.286	14.481
Sample4	13.597	13.923	13.981	14.068	14.437
Sample5	13.893	14.053	14.17	14.213	14.555
Sample6	13.739	14.026	14.155	14.38	14.566
Sample7	13.724	13.984	14.147	14.227	14.584
Sample8	13.675	13.944	14.033	14.099	14.321
Sample9	13.817	14.003	14.151	14.246	14.624
Sample10	13.875	13.989	14.183	14.356	14.673

统计描述.xlsx

这里提供不同分组对应各样本/变量的统计描述表

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) 将清洗后的分组。

(2) 将分组后的数据用ggplot2包进行分组箱式图绘制。

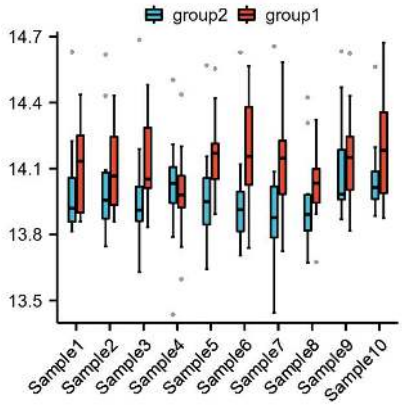
如何引用

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, 如果有很多的分组而图片宽度不够, 就有可能会导致坐标轴文字重叠。解决方案可以是:

① 增加图片宽度;

② 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