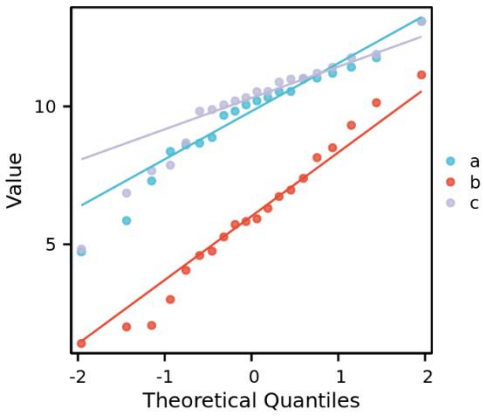
基础绘图 - 正态分析



网址：https://www.xiantao.love

更新时间：2023.09.24

目录

基本概念

- 正态分析：展示样本数据的分布情况，并验证数据是否服从正态分布

应用场景

正态分析常用来验证数据是否服从正态分布

分析过程

上传数据 → 数据处理(清洗) → 正态分析 → 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
  - 数据中不能含有非数值及其他非法字符
  - 数据中不能一列都为一种数值，即方差为 0 的列
  - ……

	A	B	C
1	a	b	c
2	4.72591	1.40529	4.82595
3	5.85786	2.0045	6.85752
4	7.29838	2.05869	9.884
5	8.36719	2.99921	10.884
6	8.60107	4.05812	11.884
7	8.66616	4.59349	7.66616
8	8.86973	4.75245	7.86973
9	9.67506	5.2682	8.67506
10	9.8286	5.71786	9.8286
11	10.0563	5.82617	10.0563
12	10.2001	5.92447	10.2001
13	10.3193	6.30299	10.3193
14	10.534	6.73115	10.534
15	10.5438	6.96759	10.5438
16	10.9896	7.39144	10.9896
17	11.0264	8.143	11.0264
18	11.2021	8.50166	11.2021
19	11.4215	9.31763	11.4215
20	11.7645	10.1337	11.7645

- 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	均值(Mean)	标准差(SD)	中位数(Median)	最小值	最大值	下四分位	上四分位	标准误(SE)
a	20	9.6517	2.0069	10.128	4.7259	13.086	8.6499	10.999	0.44875
b	20	5.962	2.6992	5.8753	1.4053	11.143	4.4596	7.5793	0.60357
c	20	9.9759	1.9414	10.427	4.8259	13.086	9.5402	11.07	0.43411

统计描述.xlsx

  - 异常值分析
    - 对变量进行异常值分析

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	数目	离群值	异常值
a	20	4.72590959254947	
b	20		
c	20	4.825947, 6.85751559	4.825947

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

  - 正态性检验(Shapiro-Wilk normality test)
    - 对变量进行正态性检验

正态性检验(Shapiro-Wilk normality test)

组别	自由度(df)	p值	统计量	偏度(Skewness)	超值峰度(Kurtosis)
a	19	0.2276	0.9388	-0.90961	0.98013
b	19	0.9280	0.97954	0.080125	-0.51219
c	19	0.0757	0.91393	-1.1173	1.4235

正态性检验.xlsx

若不呈现出显著性(P>0.05)，说明符合正态分布，反之说明不符合正态分布

偏度(Skewness)评估标准：

· 偏度 ≈ 0 表示其数据分布形态与正态分布的偏斜程度相同

· 偏度 > 0 表示其数据分布形态与正态分布相比为正偏, 呈现为右偏态

· 偏度 < 0 表示其数据分布形态与正态分布相比为负偏, 呈现为左偏态

· 偏度的绝对值数值越大表示其分布形态的偏斜程度越大

超值峰度(Kurtosis)评估标准：

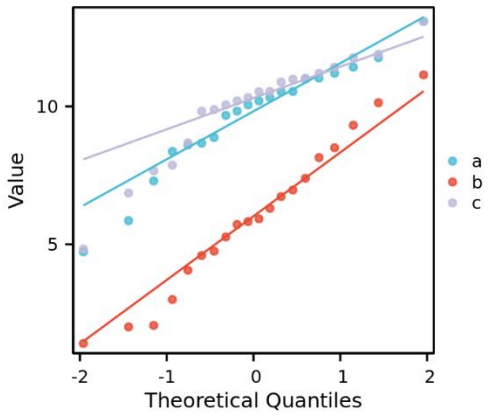
· 超值峰度 ≈ 0 表示该总体数据分布与正态分布的陡缓程度相同

· 超值峰度 > 0 表示该总体数据分布与正态分布相比较为陡峭，为尖顶峰

· 超值峰度 < 0 表示该总体数据分布与正态分布相比较为平坦，为平顶峰

· 超值峰度的绝对值数值越大表示其分布形态的陡缓程度与正态分布的差异程度越大

结果解读



- 图中点的位置由数据决定，X 轴表示假定正态分布时数据对应的分位数，Y 轴表示输入数据的实际数值
- 线条表示拟合线，散点与直线重合度越高表示数据越服从正态分布

可视化默认采用 R 包 ggplot2 中 geom_qq 内置计算方法，图片结果可以直接使用。

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

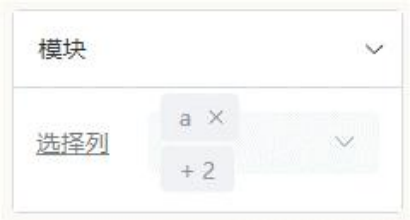
数据要求：

- 至少 1 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 10 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
  - 数据中不能含有非数值及其他非法字符
  - 数据中不能含有 Inf
  - 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。）

模块



- 选择列：默认是选择上传数据的所有列，可选择列进行分析

类型

➢ 可视化类型：可以修改可视化类型 QQ 图或 PP 图，默认是 QQ 图。

PP 图和 QQ 图都是用来观察变量是否服从正态分布, PP 图是用分布的累计比, QQ 图是用分布的分位数来做检验。

点

- 填充色：可以修改图中各变量点的填充颜色
- 描边色：可以修改图中各变量点的描边颜色
- 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
- 大小：可以修改图中各点的大小，默认为 0.8
- 不透明度：可以修改图中各点的不透明度，1 表示完全不透明，0 表示完全透明，默认为 0.8

线

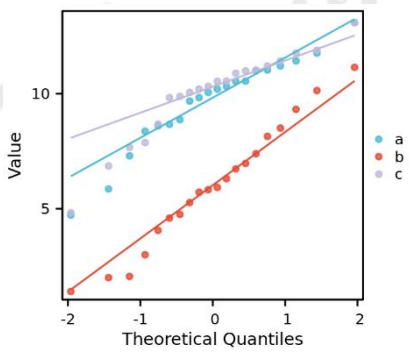
- 线条类型：可以选择图中线相关部分的线条类型，默认为实线，还可以选择虚线
- 颜色：可以修改图中线相关部分的线条颜色
- 线条粗细：可以选择修改图中线相关部分的线条粗细，默认为 0.75pt
- 不透明度：可以修改图中线相关部分的线条不透明度，默认为 1，表示完全不透明，0 表示完全透明

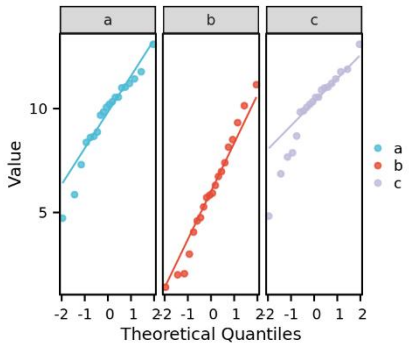
置信区间

- 是否展示：可以选择是否进行展示置信区间的操作，默认为不展示
- 填充颜色：可以修改图中置信区间的填充颜色，只有展示置信区间时才有作用
- 不透明度：可以修改图中置信区间的不透明度，1 表示完全不透明，0 表示完全透明，默认为 0.2，只有展示置信区间时才有作用

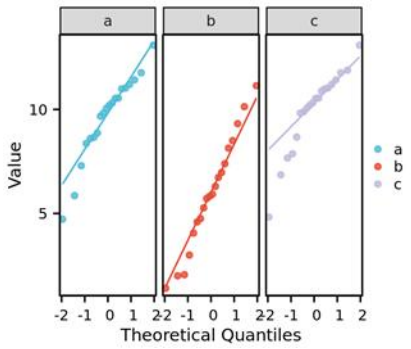
分面

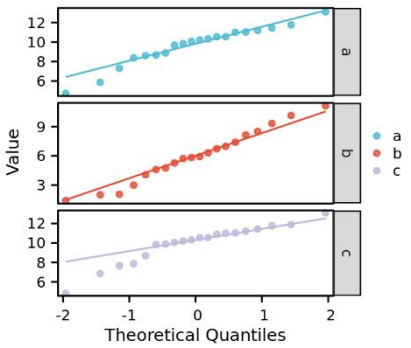
- 分面映射：可以选择是否对图形进行分面映射，默认为不映射，如下：





- 分面方向：可以修改分面的方向，默认为按列进行，还可以选择按行，如下：





- 分面颜色：可以修改分面图形的分面颜色
- 文字大小：可以选择并修改分面文字的大小，默认为 6pt
- 空间大小：可以选择分面的空间大小，默认为自适应（表示跟随图形变化），还可以选择固定（表示不随图形变化）

标题

- 大标题：大标题内容
- x 轴标题：x 轴标题内容
- y 轴标题：y 轴标题内容

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

- 是否展示：可以选择是否展示图注，默认展示
- 图注标题：首先选择展示图注，则可以修改需要上传的图注标题信息
- 图注位置：可以选择图注的位置，默认展示在右侧

坐标轴

- 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不能可视化数据范围 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不能可视化数据范围 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

图片

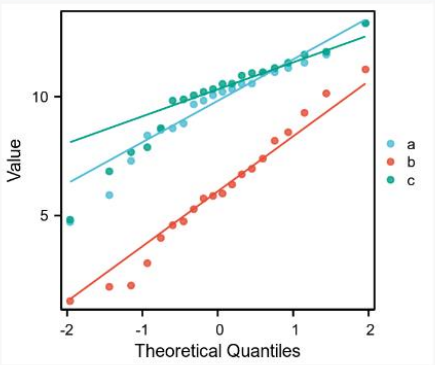
- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

结果说明

主要结果

正态分析

正态分析: 展示样本数据的分布情况, 并验证数据是否服从正态分布



正态分析.pdf 正态分析.tiff

· 图中点的位置由数据决定, X轴表示假定正态分布时数据对应的分位数, Y轴表示输入数据的实际数值

· 线条表示拟合线, 散点与直线重合度越高表示数据越服从正态分布

注意: 可视化默认采用R包ggplot2中geom_qq内置计算方法，图片结果可以直接使用

提供 PDF 和 TIFF 格式图片下载

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	均值(Mean)	标准差(SD)	中位数(Median)	最小值	最大值	下四分位	上四分位	标准误(SE)
a	20	9.6517	2.0069	10.128	4.7259	13.086	8.6499	10.999	0.44875
b	20	5.962	2.6992	5.8753	1.4053	11.143	4.4596	7.5793	0.60357
c	20	9.9759	1.9414	10.427	4.8259	13.086	9.5402	11.07	0.43411

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	数目	离群值	异常值
a	20	4.72590959254947	
b	20		
c	20	4.825947, 6.85751559	4.825947

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 - 正态性检验(Shapiro-Wilk normality test)

正态性检验(Shapiro-Wilk normality test)

组别	自由度(df)	p值	统计量	偏度(Skewness)	超值峰度(Kurtosis)
a	19	0.2276	0.9388	-0.90961	0.98013
b	19	0.9280	0.97954	0.080125	-0.51219
c	19	0.0757	0.91393	-1.1173	1.4235

正态性检验.xlsx

若不呈现出显著性(P>0.05)，说明符合正态分布，反之说明不符合正态分布

偏度(Skewness)评估标准：

· 偏度 ≈ 0 表示其数据分布形态与正态分布的偏斜程度相同

· 偏度 > 0 表示其数据分布形态与正态分布相比为正偏，呈现为右偏态

· 偏度 < 0 表示其数据分布形态与正态分布相比为负偏，呈现为左偏态

· 偏度的绝对值数值越大表示其分布形态的偏斜程度越大

超值峰度(Kurtosis)评估标准：

· 超值峰度 ≈ 0 表示该总体数据分布与正态分布的陡缓程度相同

· 超值峰度 > 0 表示该总体数据分布与正态分布相比较为陡峭，为尖顶峰

· 超值峰度 < 0 表示该总体数据分布与正态分布相比较为平坦，为平顶峰

· 超值峰度的绝对值数值越大表示其分布形态的陡缓程度与正态分布的差异程度越大

这里提供各变量的正态性检验

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：

对数据用ggplot2包绘制正态图，默认采用geom_qq内置计算方法

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．QQ 图和 PP 图有什么区别?

答:

PP 图和 QQ 图都是用来观察变量是否服从正态分布，PP 图是用分布的累计比，QQ 图是用分布的分位数来做检验。

2. 数据结果是否准确?

可视化默认采用 R 包 ggplot2 中 geom_qq 内置计算方法，图片结果可以直接使用。