

散点图

目录

关于 Bins.....	1
步骤 1.....	2
步骤 2.....	2
步骤 3.....	2
步骤 4.....	2
步骤 5.....	2
步骤 6.....	2
关于线性趋势.....	2

散点图 (Scatter plot) 用于展示两个定量变量之间的关系，即 X 与 Y 之间的关系可视化。在 SPSSAU 中支持：

- ✓ 最多一次性放入 20 个 X，仅支持 1 个 Y，即最多 20 个 X 与 1 个 Y 的关系；
- ✓ 支持颜色区分 (可选项)，即 X 与 Y 的关系时，放入第 3 个分类项 (比如性别)，区分不同分类项时 X 与 Y 的关系情况；



- ✓ SPSSAU 支持‘矩阵排列图’，当 X 个数过多时，可通过‘矩阵排列图’直观展示各个 X 与 Y 之间的有关系情况；
- ✓ 支持输出结果中加‘线性趋势’；
- ✓ 散点图直观展示各个点之间的关系情况，如果点非常多，那么很难观察到 X 与 Y 的关系情况，因而 SPSSAU 提供 bins 功能；

关于 Bins

在散点图中，“bins”指将数据点分入不同的区间 (bins)，并且记录下不同 bins 的集中趋势 (SPSSAU 取 bins 的平均值) 来表示该 bins，比如某个 bins 包括 100 个数据，取该 100 个数据的平均值来表示该 bins。SPSSAU 中进行 bins 处理的算法如下：

设有一组散点数据 $\{(x_i, y_i)\}_{i=1}^n$, 其中 n 为数据点的总数。

步骤 1

设定 bins 的数量为 b , 当数据点个数大于 1000 时, SPSSAU 自动设置 bins 为 100, 当然研究者自行设置 bins 个数 (可选为 25、50、100 和 1000 个), 或者要求不进行 bins 处理。

步骤 2

计算每个 bins 的宽度: $binWidth = \frac{n}{b}$

步骤 3

对 x 进行升序排序, 得到有序序列 $\{x_{(i)}\}_{i=1}^n$ 。bins 的取值边界 $edges$ 为:

$$edges = \{x_{(1)}, x_{(1)} + binWidth, x_{(1)} + 2 \times binWidth, \dots, x_{(1)} + n \times binWidth\}$$

步骤 4

对 $edges$ 进行四舍五入, 得到整数索引, 四舍五入后的边界取闭区间。设边界为 $\{e_j\}_{j=1}^{b+1}$, 其中 $e_1 = \lfloor x_{(1)} \rfloor$, $e_{b+1} = \lceil x_{(n)} \rceil$ 。

步骤 5

对于每个 bins: $j = 1, 2, \dots, b$, 计算:

该 bins 内 x 坐标的点集合:

$$S_j = \{(x_i, y_i) | e_j \leq x_i < e_{j+1}\}$$

该 bins 内 x 坐标的平均值:

$$\bar{x}_j = \frac{1}{|S_j|} \sum_{(x_i, y_i) \in S_j} x_i$$

该 bins 内 y 坐标的平均值:

$$\bar{y}_j = \frac{1}{|S_j|} \sum_{(x_i, y_i) \in S_j} y_i$$

步骤 6

输出 $\{\bar{x}_j, \bar{y}_j\}_{j=1}^b$, 并且绘制散点图。

关于线性趋势

SPSSAU 输出散点图带‘线性趋势’功能, 其原理是利用 X 与 Y 之间的线性回归拟合计算得到, 具体算法说明, 可查阅‘线性回归’文档。