R2 - 组织、可视化和描述数据

> 随着大数据和机器学习技术的兴起，投资从业者正在迎接一个海量、高速和种类繁多的数据时代——这使他们能够探索和利用这些丰富的信息来制定投资策略。

学习目标

1. 识别和比较数据类型
2. 描述如何组织数据以进行定量分析
3. 解释频率和相关分布
4. 解释列联表
5. 描述数据可视化的方式并评估特定可视化的使用
6. 描述如何在可视化类型中进行选择
7. 计算和解释集中趋势的度量

8. 评估均值的替代定义以解决投资问题
9. 计算分位数并解释相关的可视化
10. 计算和解释离散量度
11. 计算和解释目标下行偏差
12. 解释偏度
13. 解释峰态
14. 解释两个变量之间的相关性

数据类型

- 数据可以定义为数字面板数据、字符、单词和文本以及图像、音频和视频的集合，以原始或有组织的格式表示事实或信息。
- 将在三种不同的分类视角下讨论数据类型：
  - 数值数据与分类数据；
  - 横截面与时间序列与面板数据；
  - 结构化数据与非结构化数据。

数值数据与分类数据

从统计的角度来看，数据可以分为两大类：数值数据和分类数据

1. Numerical data数值数据是将测量或计数的数量表示为数字的值，也称为定量数据。
2. quantitative data数值（定量）数据可以分为两种类型：连续数据和离散数据。
3. Continuous data连续数据是可以测量的数据，可以取指定值范围内的任何数值。
4. 离散数据是计数过程产生的数值。

5. Categorical data分类数据（也称为quatitative data）是描述一组观察的质量或特征的值，因此可以用作标签，将数据集划分为组以进行总结和可视化。
6. Nominal data标称数据是不能按逻辑顺序组织的分类值。文本标签是表示标称数据的常用格式，但标称数据也可以使用数字标签进行编码。
7. Ordinal data序数数据是可以按逻辑排序或排序的分类值。

横截面数据与时间序列数据与面板数据

基于数据的收集方式，将数据分为三种类型：横截面、时间序列和面板。

1. Variable变量是可以测量、计数或分类的特征或数量，并且可能会发生变化。变量也可以称为字段、属性或特征。
2. Observation观察是在某个时间点或指定时间段内收集的特定变量的值。
3. Cross-sectional data横截面数据是在给定时间点从多个观察单位对特定变量的观察的列表。

4. Time-series data时间序列数据是针对特定变量的单个观测单位的一系列观测值，这些观测值随时间以离散且通常等间隔的时间间隔收集，
5. Panel data面板数据是财务分析和建模中经常使用的时间序列和横截面数据的混合体。

结构数据和非结构数据

1. Structured data结构化数据以预定义的方式高度组织，通常具有重复模式。
2. Unstructured data非结构化数据是不遵循任何常规组织形式的数据。

组织数据进行定量分析

将数据组织成一维数组或二维数组通常是数据分析和建模的第一步。

1. one-dimensional array一维数组是表示同一数据类型的数据集合的最简单格式，因此适合表示单个变量。
2. descriptive statistics描述性统计数据总结数据分布的集中趋势和分布变化的度量方法。
3. two-dimensional rectangular array (data table)

使用频率分布汇总数据

1. frequency distribution（one-way table）频率分布（也称为单向表）是数据的表格显示，通过按不同值或组计算变量的观察值或通过将数值变量的值汇总到一组数字排序的 bin（统计堆） 中来构建。
2. Absolute frequency绝对频率，是为变量的每个唯一值计算的实际观察次数。
3. Relative frequency相对频率是变量的每个唯一值的绝对频率除以观察总数。
4. Cumulative absolute frequency累积绝对频率会累积绝对频率。
5. Cumulative relative frequency累积相对频率是相对频率的部分和的序列。

使用列联表汇总数据

1. Contingency table列联表是一种表格格式，它同时显示两个或多个分类变量的频率分布，用于查找变量之间的模式。
2. Joint frequencies列联表单元格中的一项，表示将一行中的一个变量和列中的另一个变量连接起来，以统计观察值。
3. Marginal frequencies列联频率在列联表中跨行或跨列加联合频率所确定的和。
4. 评估分类模型的性能（在这种情况下，列联表称为混淆矩阵Confusion matrix）。
5. Chi-square test of independence独立性卡方检验研究两个分类变量之间的潜在关联。

数据可视化

1. Visualization可视化是以图形或图形格式呈现数据，目的是增加对数据的理解和深入了解数据。
2. Histogram直方图是一种图表，通过使用条形或柱形的高度来表示分布中每个 bin 或区间的绝对频率来呈现数值数据的分布。
3. frequency polygon频率多边形：用直线连接表示类频率的连续点而得到的频率分布图。
4. cumulative frequency distribution chart累积绝对频率或累积相对频率在y轴上以间隔的上限作图的图表，它使人们能够看到低于某一值的观测次数或百分比。
5. bar chart条形图绘制分类数据频率分布的图表，其中每个条形图代表一个不同的类别，每个条形图的高度与对应类别的频率成正比。
6. grouped bar chart显示两个分类变量的联合频率的柱状图(也称为clustered bar chart聚类柱状图)。

7. stacked bar chart表示两个分类变量的频率分布的另一种形式，其中代表子组的条被放置在彼此的顶部以形成单个条。
8. tree-map树形图由一组代表不同组的彩色矩形组成，每个矩形的面积与对应组的值成正比。
9. word cloud由从文本数据源中提取的单词组成。每个不同单词的大小与它在给定文本中出现的频率成正比(也称为tag cloud标签云)。
10. line chart折线图显示数据系列随时间的变化。bubble line chart气泡折线图用不同大小的气泡表示数据的第三维的折线图。
11. Scatter plot散点图是一种图形，用于可视化两个数值变量的联合变化。
12. scatter plot matrix散点图矩阵是一种有用的工具，可用于组织变量对之间的散点图，从而可以轻松地在一个组合视觉中检查所有成对关系。
13. heat map热图是一种以表格格式组织和汇总数据并使用色谱表示它们的图形。

集中趋势测量

1. measure of central tendency集中趋势的度量指定数据的中心位置。
2. 集中趋势的常用度量——算术平均值、中位数、众数、加权平均值、几何平均值和调和平均值。
3. measure of location位置测量不仅包括集中趋势测量，还包括说明数据位置或分布的其他测量。
4. 位置度量，包括quarties四分位数、quintiles五分位数、deciles十分位数和percentiles百分位数。
5. statistic统计量是一组观察值的汇总度量，描述性统计量总结了数据分布的集中趋势和分布变化。
6. 统计数据称为sample statistic样本统计数据。
7. population 总体 sample 样本

Arithmetic Mean算术平均值

- Sample Mean or average, $\bar{X}$
  $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

Outliers异常值

Median中位数

- 中位数是按升序或降序排序的一组项目的中间项目的值。

Mode众数

- 众数是分布中出现频率最高的值。
- 当分布具有最常出现的单个值时，该分布称为unimodal单峰分布。
- 如果一个分布有两个最常出现的值，则它有两种模式，称为bimodal双峰。
- 如果分布具有三个最常出现的值，则它是trimodal三峰的。
- modal interval/intervals区间

- Weighted Mean$\bar{X}_w = \sum_{i=1}^{n} w_i X_i, \sum_i w_i = 1.$
- Geometric Mean$\bar{X}_G = \left(X_1 X_2 \cdots X_{X_n}\right)^{1/n}, X_i \geq 0$
- Harmonic Mean$\bar{X}_H = \frac{n}{\sum_{i=1}^{n} (1/X_i)}, X_i \geq 0$
- 调和平均值可以被视为一种特殊类型的加权平均值，其中观察值的权重与其大小成反比。
- $Arithmetic\, mean \times Harmonic\, mean = Geometric\, mean^2$
- 调和均值小于几何均值，而几何均值又小于算术均值

分位数

- quantile (or fractile)表示数据的指定部分位于或低于该值。
- Quartiles四分位数将分布划分为四等分，Quintiles五分位数划分为五分之一，Deciles十分位数划分为十分之一，Percentiles百分位数划分为百分之一。
- interquartile range (IQR)四分位距(IQR) 是第三个四分位数和第一个四分位数之间的差。
- 一个百分数在有n个条目的数组中按升序排序的位置(或位置)公式为:
  - $L_y = (n+1)\frac{y}{100}$
  - y是我们除以分布的百分比，Ly是按升序排序的百分比(Py)在数组中的位置(L)。
- linear interpolation线性插值：在括号内的两个已知值的基础上，利用两个已知值之间的直线来估计一个未知值。
- box and whisker plot盒须图用于显示数据在四分位上的分散情况的图形。它由一个“盒子”和连接在盒子上的“胡须”组成。

离散度测量

- 离散度是围绕集中趋势的可变性。
- range, mean absolute deviation, variance, and standard deviation范围、平均绝对偏差、方差和标准偏差。这些都是absolute dispersion绝对分散的度量。
- Range $Range = Maximum\ value - Minimum\ value$
- mean absolute deviation $MAD = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$
- Variance 方差 sample variance $s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$
- Sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$
- 离散度及算术$\bar{X}$与几何$\bar{X}_G$均值的关系
  $\bar{X}_G \approx \bar{X} - \frac{s^2}{2}$

下行偏差和变异系数

- downside risk 下行风险：收益低于指定值的风险。
- target semideviation目标半偏差:以低于目标(也称为目标下行偏差)的观测的平方偏差的平均值的平方根计算。
  - $S_{Target} = \sqrt{\sum_{forall X_i \leq B}^{n} \frac{(X_i - B)^2}{n-1}}$
- Coefficient of Variation差异系数
- Relative dispersion 相对离差是相对于参考值或基准的离差量。
- Coefficient of Variation $CV = s / \bar{X}$

分布的形状

分布的形状：偏度

- skewness偏度：回报分布的对称程度
- 不对称的分布是skewed偏斜的。
- 具有正偏斜的回报分布经常会出现小额亏损和一些极端收益。
- 连续正偏态分布在其右侧有一条长尾。
- 具有负偏斜的收益分布经常会出现小幅收益和一些极端损失。
- 连续负偏态分布在其左侧有一条长尾。

- 连续的正偏态单峰分布，众数小于中位数，中位数小于均值。
- 连续的负偏态单峰分布，均值小于中位数，中位数小于众数。
- 投资者应该被正偏斜吸引，因为平均回报高于中位数
- Skewness偏斜度(有时称为相对偏斜度)的计算方法是对标准化的平均值的三次平均偏差除以三次标准偏差，以使测量不受尺度限制。
- sample skewness $skewness \approx (1/n) \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{s^3}$

分布的形状：峰度

- kurtosis峰度是分布尾部相对于分布其余部分的组合权重的量度，即总概率在平均值的 2.5 个标准差之外的比例。
- 尾部比正态分布更粗的分布称为leptokurtic或fat-tailed
- 尾部比正态分布更细的分布称为platykurtic或thin-tailed薄尾分布；
- 在尾部的相对权重方面类似于正态分布的分布称为mesokurtic中峰。
- 峰度的计算涉及找到平均偏差的平均值，然后将平均值除以四次方的标准偏差。

- Excess kurtosis 过度峰度是相对于正态分布的峰度。
- sample excess kurtosis样本过度峰度
  - $K_E = \left[(1/n)\frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4}\right] - 3$
- 正态分布的超峰态等于 0。
- 肥尾分布的超峰态大于 0，而细尾分布的超峰态小于 0。
- 超峰态为正的回报分布——肥尾回报分布——比正态分布更频繁地偏离均值的极大偏差。

两个变量之间的相关性

- Sample Covariance样本协方差
  - $s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$
- Sample Correlation Coefficient样本相关系数
  - $r_{XY} = \frac{s_{XY}}{s_X s_Y}$
- 当一个或两个变量中存在异常值时，相关性也可能是一种不可靠的度量。
- 相关性并不意味着因果关系。
- spurious correlation虚假相关性一词用于指代：
  i. 反映特定数据集中机会关系的两个变量之间的相关性；
  ii. 由将两个变量中的每一个与第三个变量混合的计算引起的相关性；
  iii. 两个变量之间的相关性不是源于它们之间的直接关系，而是源于它们与第三个变量的关系。