数据探索

2021/9/28

目录

数据质量分析

- 数据质量分析是数据预处理的前提，是数据挖掘分析结论有效性和准确性的基础，其主要任务是检查原始数据中是否存在脏数据，脏数据一般是指不符合要求，以及不能直接进行相应分析的数据，在常见的数据挖掘工作中，脏数据包括：
  - 缺失值
  - 异常值
  - 不一致的值
  - 重复数据及含有特殊符号（如#、￥、*）的数据
- 本小节将主要对数据中的**缺失值**、**异常值**和**一致性**进行分析。

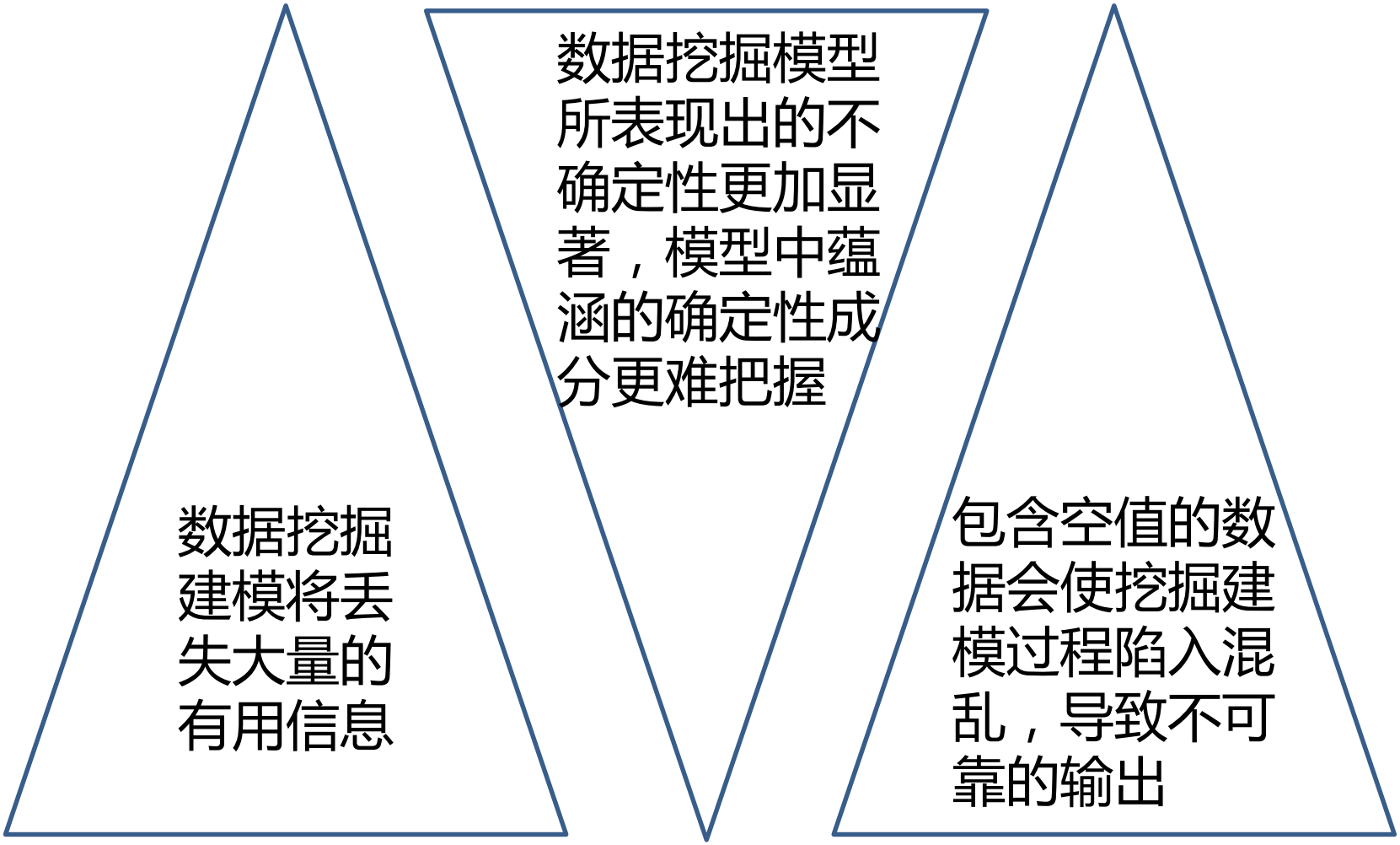
数据质量分析——缺失值产生的原因

- 有些信息暂时无法获取，或者获取信息的代价太大。

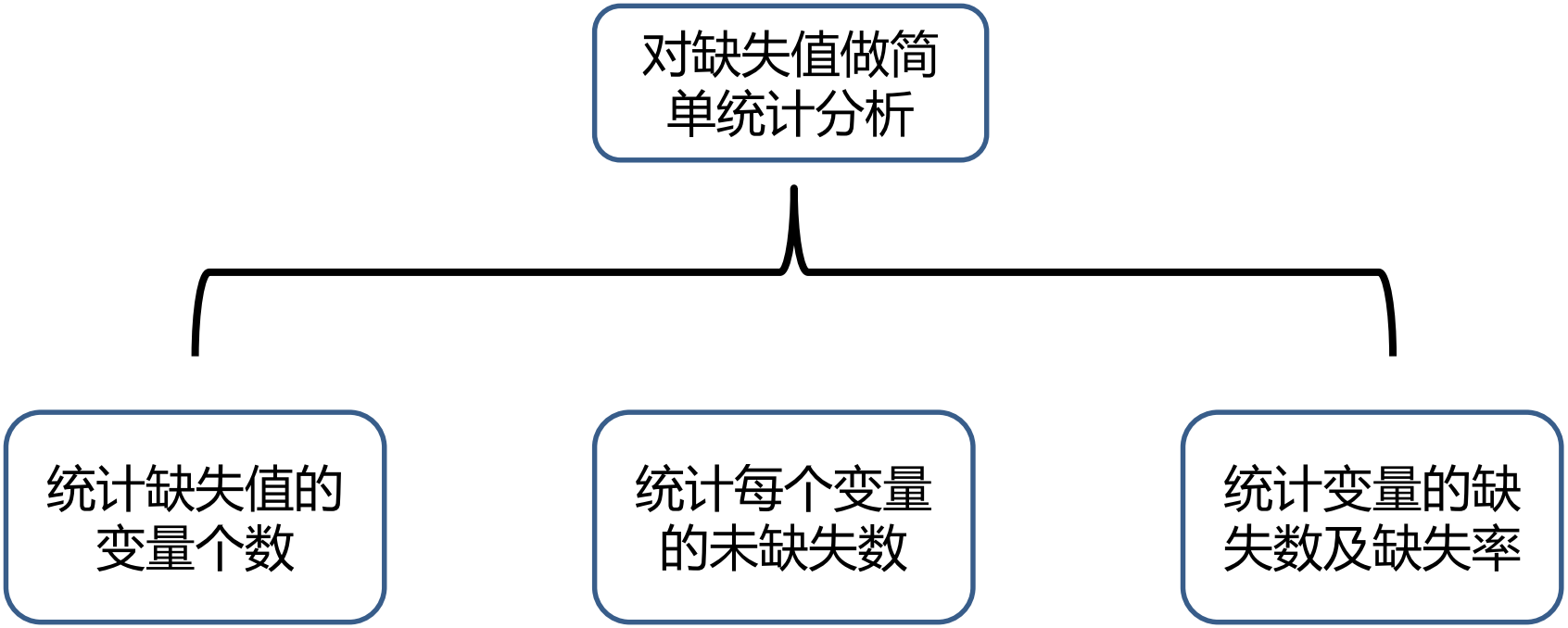
- 有些信息是被遗漏的。可能是因为输入时认为不重要、忘记填写或对数据理解错误等一些人为因素而遗漏，也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障等机械原因而丢失。

- 属性值不存在。在某些情况下，缺失值并不意味着数据有错误，对一些对象来说属性值是不存在的，如一个未婚者的配偶姓名、一个儿童的固定收入状况等。

数据质量分析——缺失值的影响



数据质量分析——缺失值分析



数据质量分析——异常值分析

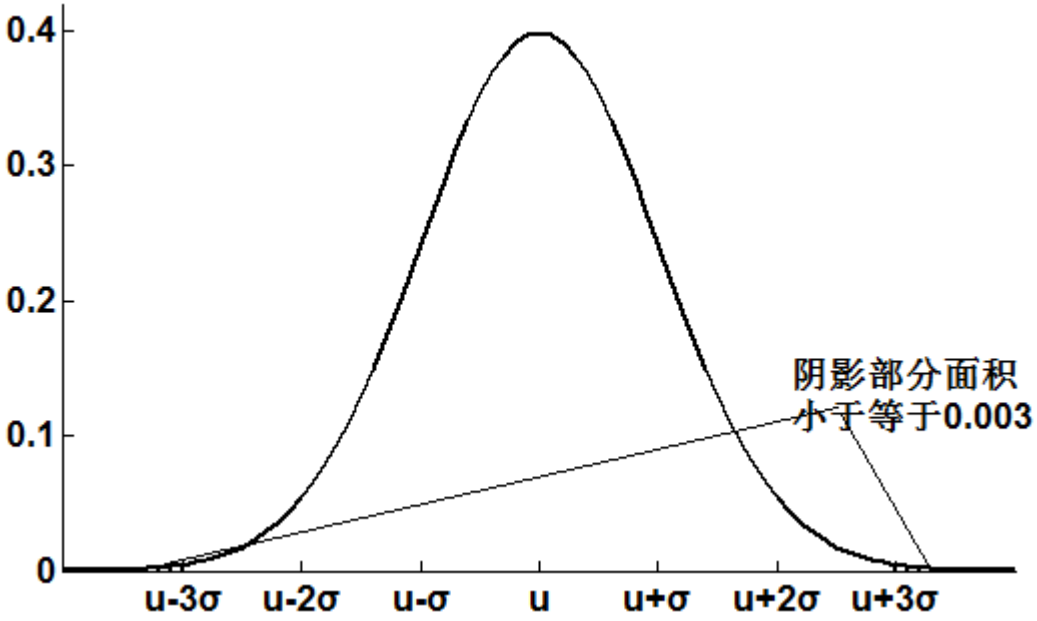
- 异常值分析是检验数据是否有录入错误以及含有不合常理的数据。忽视异常值的存在是十分危险的，不加剔除地把异常值包括进数据的计算分析过程中，对结果会带来不良影响；重视异常值的出现，分析其产生的原因，常常成为发现问题进而改进决策的契机。
- 异常值是指样本中的个别值，其数值明显偏离其余的观测值。异常值也称为离群点，异常值的分析也称为离群点的分析。
- 异常值分析方法主要有：简单统计量分析、3σ 原则、箱型图分析。

异常值分析——简单统计分析

- 可以先做一个描述性统计，进而查看哪些数据是不合理的。需要的统计量主要是最大值和最小值，判断这个变量中的数据是不是超出了合理的范围，如身高的最大值为5米，则该变量的数据存在异常。

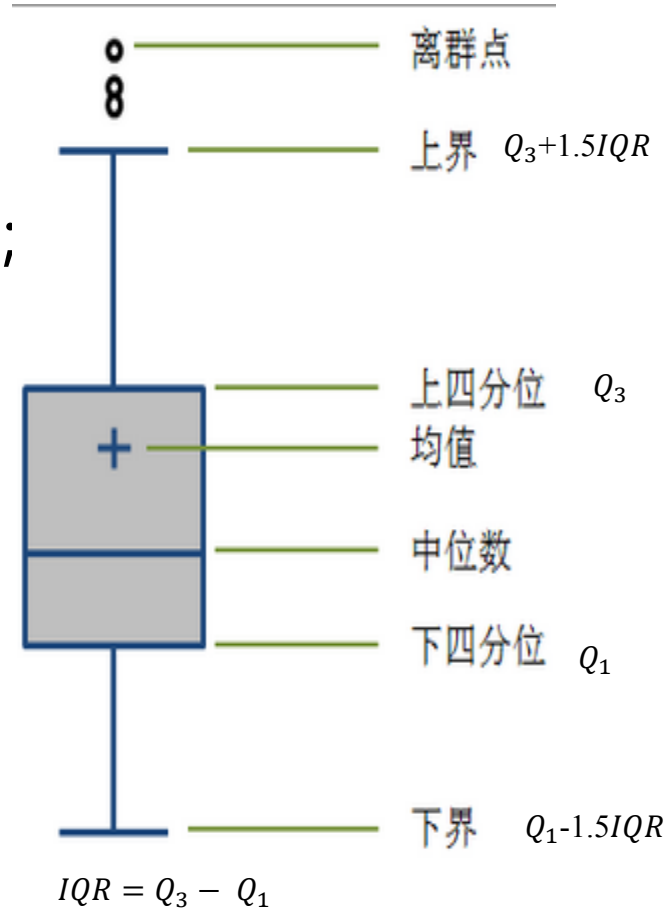
异常值分析——3σ 原则

- 如果数据服从正态分布，在3σ 原则下，异常值被定义为一组测定值中与平均值的偏差超过三倍标准差的值。在正态分布的假设下，距离平均值 3σ 之外的值出现的概率为 $P\left(\left|x-\mu\right|>3\sigma\right)\leq 0.003$，属于极个别的小概率事件。



异常值分析——箱型图分析

- 箱形图依据实际数据绘制，不需要事先假定数据服从特定的分布形式，没有对数据作任何限制性要求，它只是真实直观地表现数据分布的本来面貌；另一方面，箱形图判断异常值的标准以四分位数和四分位距为基础，四分位数具有一定的鲁棒性：多达25%的数据可以变得任意远而不会很大地扰动四分位数，所以异常值不能对这个标准施加影响，箱形图识别异常值的结果比较客观。由此可见，箱形图在识别异常值方面有一定的优越性。



一致性分析

- 数据不一致性是指数据的矛盾性、不相容性。直接对不一致的数据进行挖掘，可能会产生与实际相违背的挖掘结果。
- 在数据挖掘过程中，不一致数据的产生主要发生在数据集成的过程中，可能是由于被挖掘数据是来自于从不同的数据源、重复存放的数据未能进行一致性地更新造成的，比如两张表中都存储了用户的地址，在用户的地址发生改变时，如果只更新了一张表中的数据，那么这两张表中就有了不一致的数据。

目录

数据特征分析

- 对数据进行质量分析以后，接下来就是对数据做特征分析。一般可通过绘制图表、计算某些特征量等手段进行数据的特征分析。
- 这里主要介绍的特征方法有：
  - 分布分析
  - 对比分析
  - 统计量分析
  - 周期性分析
  - 贡献度分析
  - 相关性分析

数据特征分析——分布分析

- 分布分析能揭示数据的分布特征和分布类型，便于发现某些特大或特小的可疑值。对于定量数据，欲了解其分布形式，是对称的、还是非对称的，可做出频率分布表、绘制频率分布直方图、绘制茎叶图进行直观地分析；对于定性分类数据，可用饼图和条形图直观地显示分布情况。

定量数据的分布分析

- 对于定量变量而言，做频率分布分析时选择“组数”和“组宽”是主要的问题，一般按照以下步骤：
- 求极差
- 决定组距与组数
- 决定分点
- 列出频率分布表
- 绘制频率分布直方图

定量数据的分布分析

- 遵循的主要原则有：
- 各组之间必须是相互排斥的
- 各组必须将所有的数据包含在内
- 各组的组宽最好相等

定量数据分布分析——具体事例

- 下表是描述菜品捞起生鱼片在2014年第二个季度的销售数据，绘制销售量的频率分布表、频率分布图，对该定量数据做出相应的分析。

日期	销售额	日期	销售额	日期	销售额
2014/4/1	420	2014/5/1	1770	2014/6/1	3960
2014/4/2	900	2014/5/2	135	2014/6/2	1770
2014/4/3	1290	2014/5/3	177	2014/6/3	3570
2014/4/4	420	2014/5/4	45	2014/6/4	2220
2014/4/5	1710	2014/5/5	180	2014/6/5	2700
…	…	…	…	…	…
2014/4/30	450	2014/5/30	2220	2014/6/30	2700
		2014/5/31	1800		

定量数据分布分析——具体事例

第一步：求极差

极差 = 最大值 - 最小值 = 3960-45=3915

第二步：分组

这里根据业务数据的含义，可取组距为500。

组数 = 极差/组距 = 3915/500=7.83=8

第三步：决定分点，如下表：

[0，500）	[500，1000）	[1000，1500）	[1500，2000）
[2000，2500）	[2500，3000）	[3000，3500）	[3500，4000）

定量数据分布分析——具体事例

第四步：绘制频率分布直方图

根据分组区间得到如下表的频率分布表，见下表。其中，第1列把数据所在的范围分成的若干组段，第1个组段要包括最小值，最后一个组段要包括最大值，习惯上将各组段设为左闭右开的半开区间，如第一个分组为[0，500）。第2列组中值是各组段的代表值，由本组段的上、下限相加除以2得到。第3列和第4列分别为频数和频率。第5列是累计频率，是否需要该列，视情况而定。

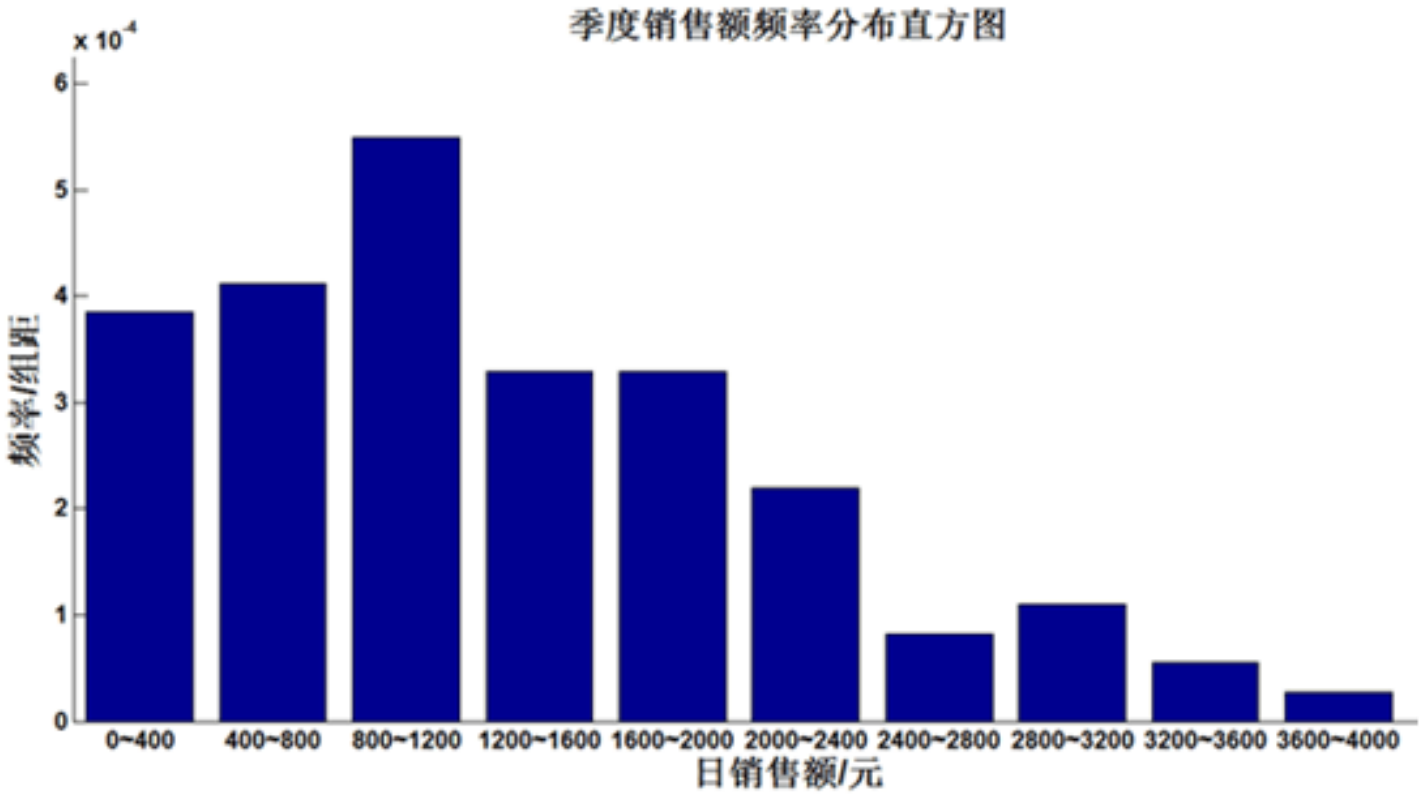
定量数据分布分析——具体事例

组段	组中值 x	频数	频率 f	累计频率
[0，500）	250	15	16.48%	16.48%
[500，1000）	750	24	26.37%	42.85%
[1000，1500）	1250	17	18.68%	61.54%
[1500，2000）	1750	15	16.48%	78.02%
[2000，2500）	2250	9	9.89%	87.91%
[2500，3000）	2750	3	3.30%	92.31%
[3000，3500）	3250	4	4.40%	95.60%
[3500，4000）	3750	3	3.30%	98.90%
[4000，4500）	4250	1	1.10%	100.00%

定量数据分布分析——具体事例

第五步：绘制频率分布直方图

若以2014年第二季度捞起生鱼片每天的销售额为横轴，以各组段的频率密度（频率与组距之比）为纵轴，表3-3的数据可绘制成频率分布直方图，见图：

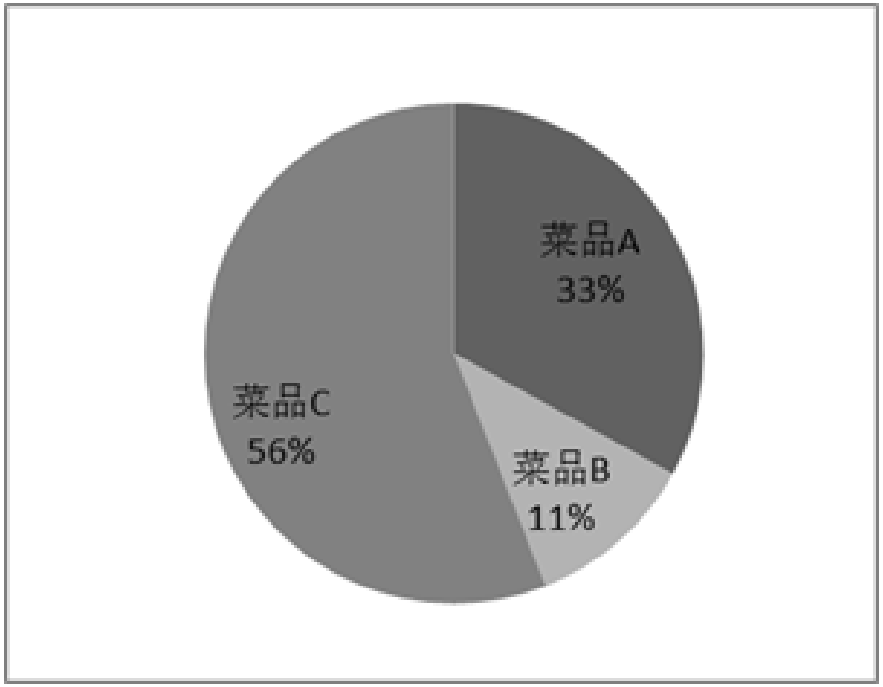


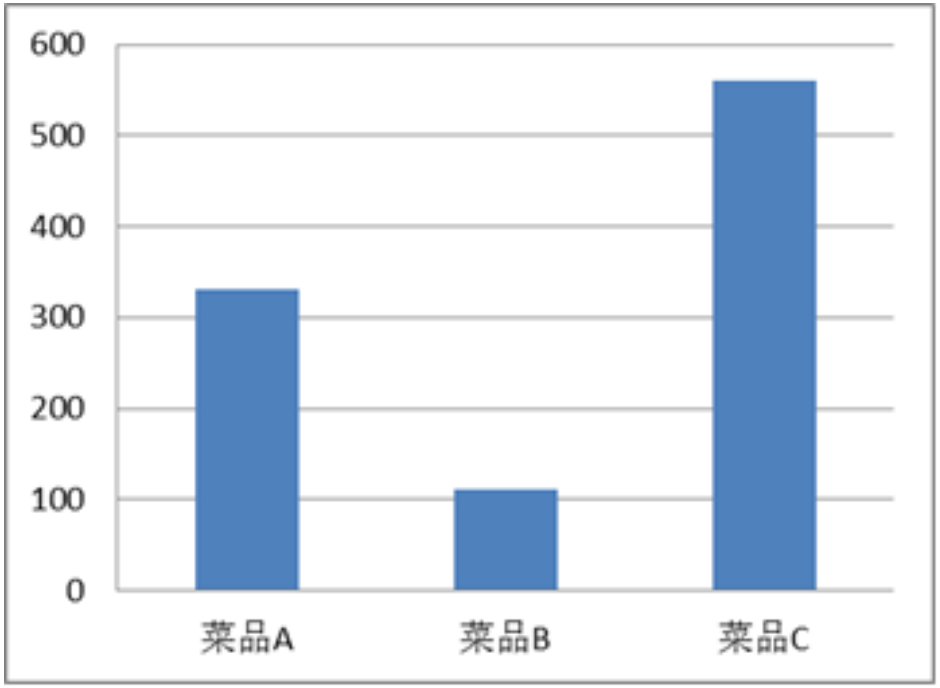
定性数据的分布分析

- 对于定性变量，常常根据变量的分类类型来分组，可以采用饼图和条形图来描述定性变量的分布。
- 饼图的每一个扇形部分代表每一类型的百分比或频数，根据定性变量的类型数目将饼图分成几个部分，每一部分的大小与每一类型的频数成正比；条形图的高度代表每一类型的百分比或频数，条形图的宽度没有意义。

定性数据的分布分析

下面左右两图分别是菜品A、B、C在某段时间的饼形和条形销售量分布图：





对比分析

- 对比分析是指把两个相互联系的指标数据进行比较，从数量上展示和说明研究对象规模的大小，水平的高低，速度的快慢，以及各种关系是否协调。特别适用于指标间的横纵向比较、时间序列的比较分析。在对比分析中，选择合适的对比标准是十分关键的步骤，选择得合适，才能做出客观的评价，选择不合适，评价可能得出错误的结论。

对比分析

- 对比分析主要有以下两种形式：
- 第一种:绝对数比较

  它是利用绝对数进行对比，从而寻找差异的一种方法。
- 第二种:相对数比较

  它是由两个有联系的指标对比计算的，用以反映客观现象之间数量联系程度的综合指标，其数值表现为相对数。由于研究目的和对比基础不同，相对数可以分为以下几种：

1)结构相对数
2)比例相对数
3)比较相对数
4)强度相对数
5)计划完成程度相对数
6)动态相对数

对比分析——相对数比较

1）结构相对数：

将同一总体内的部分数值与全部数值对比求得比重，用以说明事物的性质、结构或质量。如居民食品支出额占消费支出总额比重、产品合格率等。

2）比例相对数：

将同一总体内不同部分的数值对比，表明总体内各部分的比例关系，如人口性别比例、投资与消费比例等。

3）比较相对数：

将同一时期两个性质相同的指标数值对比，说明同类现象在不同空间条件下的数量对比关系。如不同地区商品价格对比，不同行业、不同企业间某项指标对比等。

对比分析——相对数比较

4）强度相对数：

将两个性质不同但有一定联系的总量指标对比，用以说明现象的强度、密度和普遍程度。如人均国内生产总值用“元/人”表示，人口密度用“人/平方公里”表示，也有用百分数或千分数表示的，如人口出生率用‰表示。

5）计划完成程度相对数：

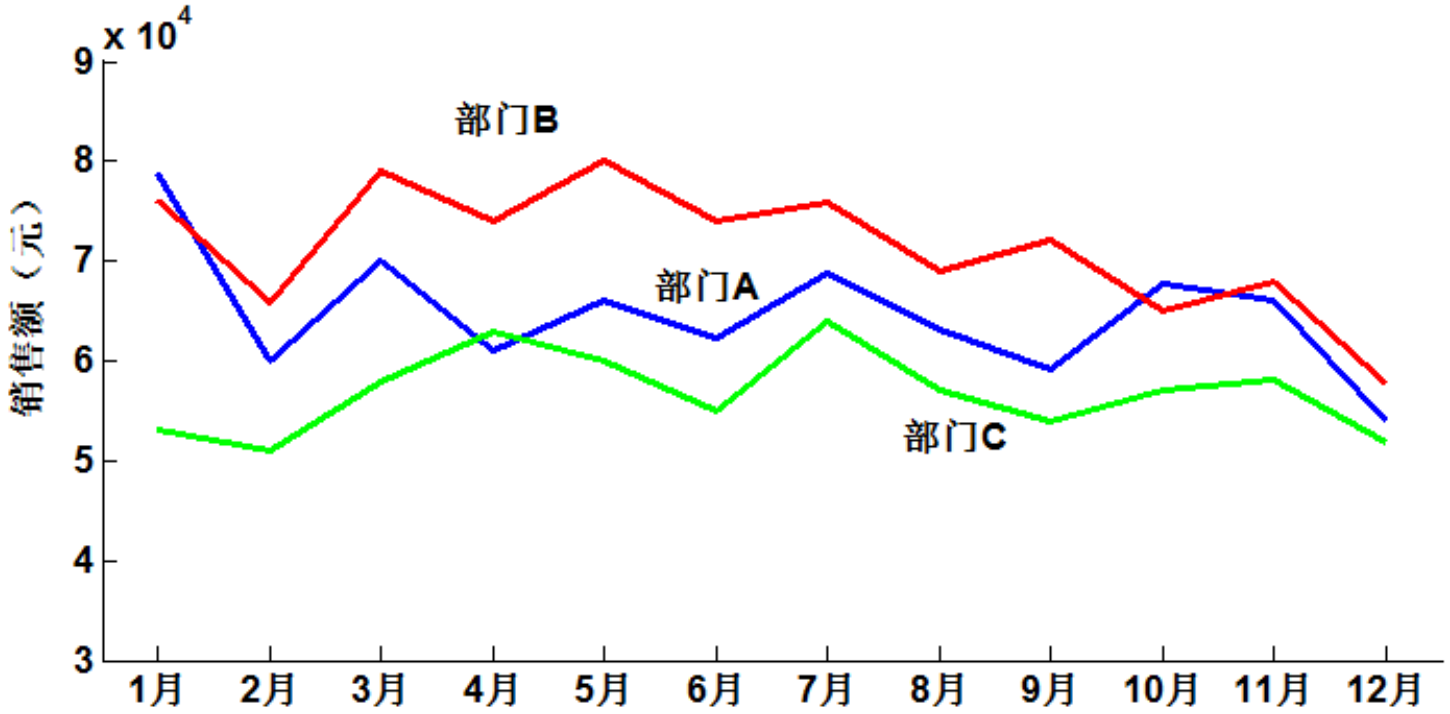
是某一时期实际完成数与计划数对比，用以说明计划完成程度。

6）动态相对数：

将同一现象在不同时期的指标数值对比，用以说明发展方向和变化的速度。如发展速度、增长速度等。

对比分析——具体事例

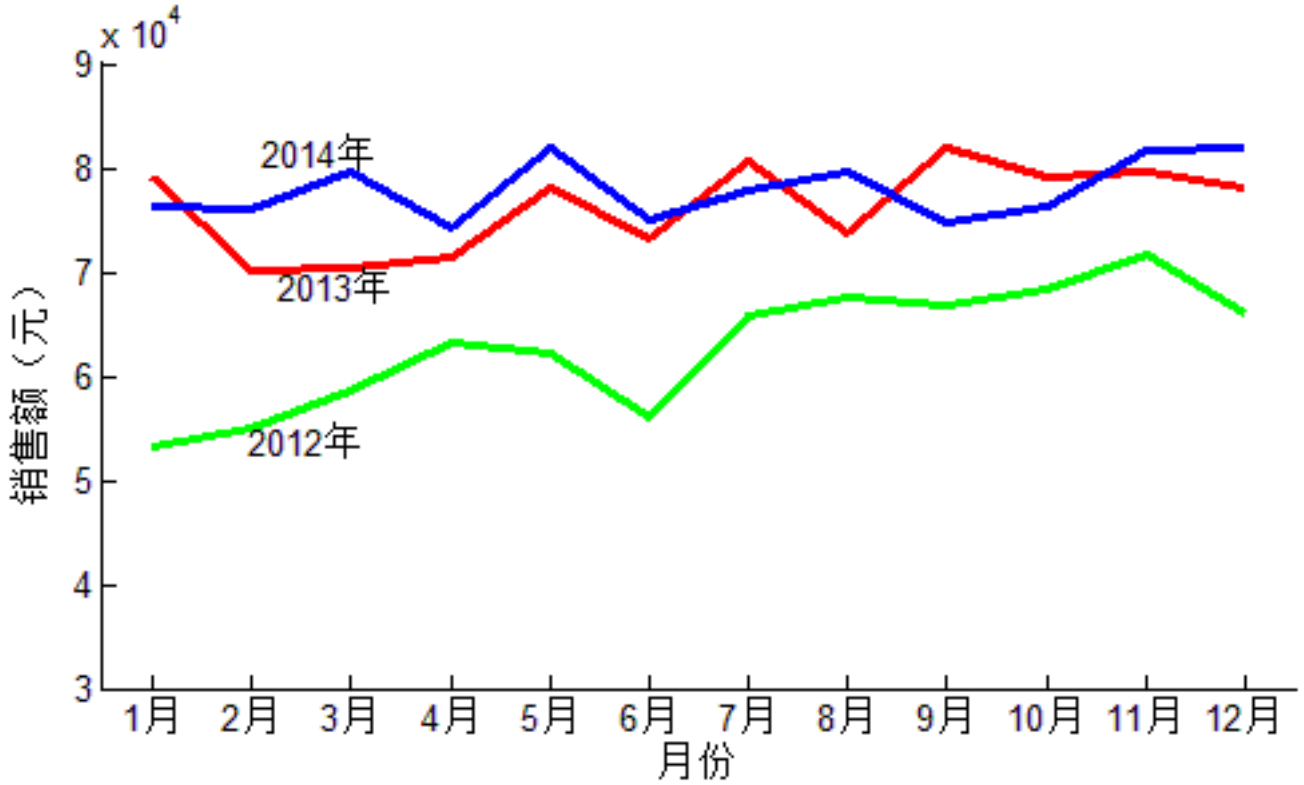
- 拿各菜品的销售数据来看，从时间的维度上分析，可以看到甜品部A、海鲜部B、素菜部C三个部门之间的销售金额随时间的变化趋势，了解在此期间哪个部门的销售金额较高，趋势比较平稳，如下图



- 从总体来看，三个部门的销售金额呈递减趋势；A部门和C部门的递减趋势比较平稳；B部门的销售金额在2月份骤降，可以进一步分析造成这种现象的业务原因，可能是原材料不足造成的。

定性数据的分布分析

- 也可以从单一部门（如海鲜部）做分析，了解各月份的销售对比情况，如下图：



- 从总体来看，2013年和2014年高于2012年，体现了相比2012年的增长趋势；相比2013年，2014年2-6月份有一定增长（特别是2、3月），但后几个月波动较大，相比2013年提升不大。

统计量分析

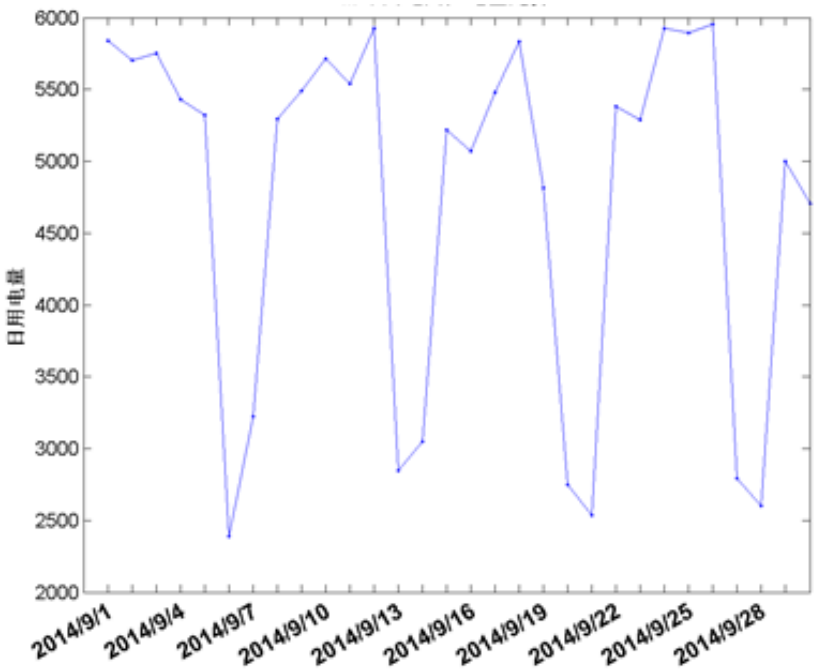
- 用统计指标对定量数据进行统计描述，常从**集中趋势**和**离中趋势**两个方面进行分析。
- 平均水平的指标是对个体集中趋势的度量，使用最广泛的是均值和中位数；反映变异程度的指标则是对个体离开平均水平的度量，使用较广泛的是标准差（方差）、四分位间距。
- 集中趋势度量主要有：均值、中位数、众数
- 离中趋势度量主要有：极差、标准差、变异系数

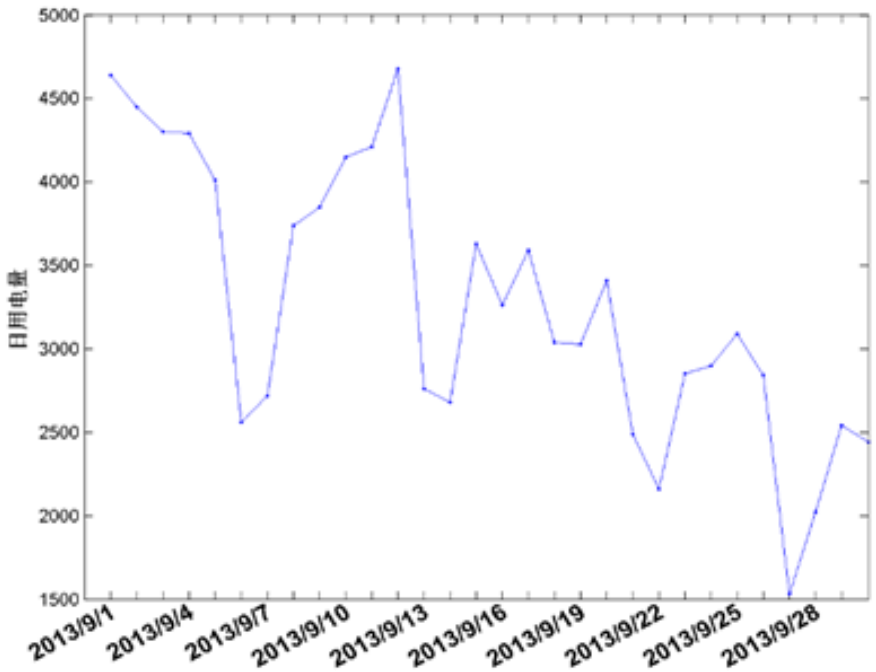
周期性分析

- 周期性分析是探索某个变量是否随着时间变化而呈现出某种周期变化趋势。周期性趋势相对较长的有年度周期性趋势、季节性周期趋势，相对较短的一般有月度周期性趋势、周度周期性趋势，甚至更短的天、小时周期性趋势。
- 如在做某用电单位用电量趋势预测过程中，可以先分析该用电单位日用电量的时序图，来直观地估计其用电量变化趋势。

周期性分析

- 下面两图分别是某用电单位A在2014年9月份和2013年9月份日用电量的时序图：





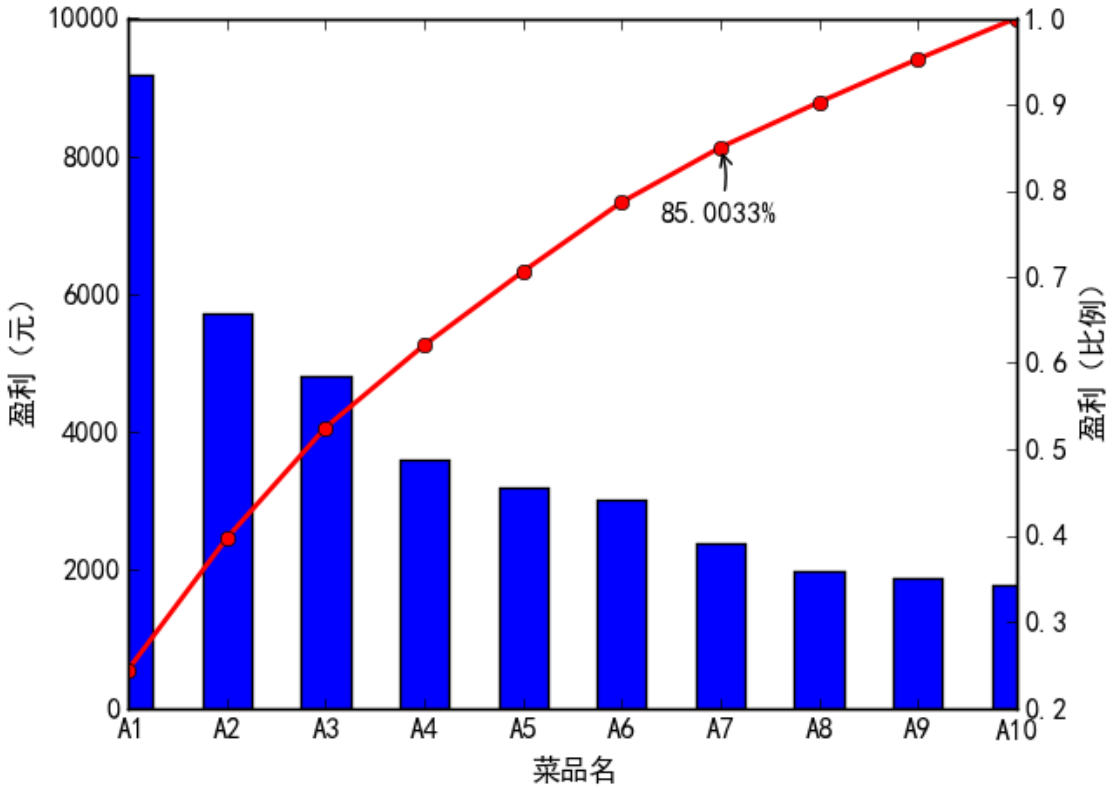
- 从总体来看用电单位A的2014年9月份日用电量呈现出周期性，以周为周期，因为周六周日不上班，所以周末用电量较低。工作日和非工作日的用电量比较平稳，没有太大的波动。
- 而2013年9月份日用电量总体呈现出递减的趋势，同样周末的用电量是最低的。

贡献度分析

- 贡献度分析又称帕累托分析，帕累托法则又称20/80定律。同样的投入放在不同的地方会产生不同的效益。比如对一个公司来讲，80%的利润常常来自于20%最畅销的产品；而其他80%的产品只产生了20%的利润。贡献度分析要求我们抓住问题的重点，找到那最有效的20%的热销产品、渠道或者销售人员，在最有效的20%上投入更多资源，尽量减少浪费在80%低效的地方。

贡献度分析

- 就餐饮企业来讲，可以重点改善盈利最高的80%的菜品，或者重点发展综合影响最高的80%的部门。这种结果可以通过帕累托分析直观的呈现出来，如下图：

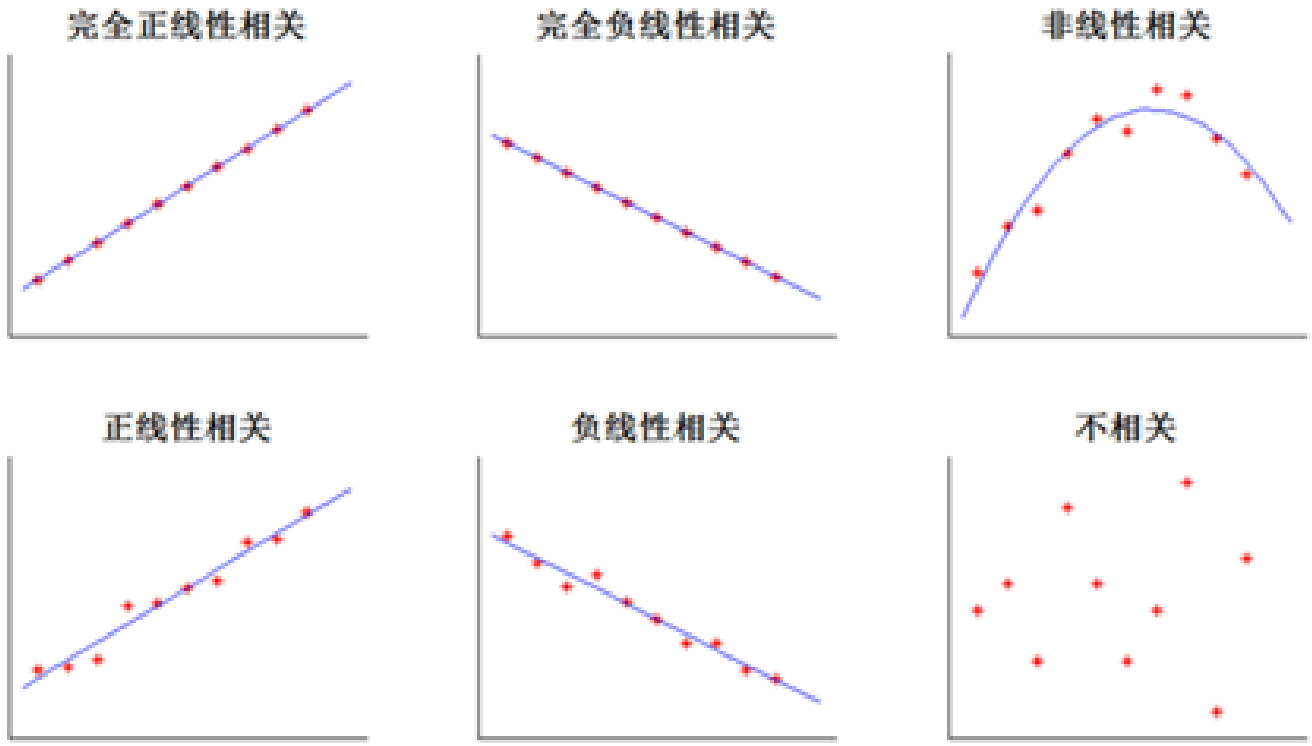


相关性分析

- 分析连续变量之间线性的相关程度的强弱，并用适当的统计指标表示出来的过程称为相关分析。
- 相关性分析方法主要有：
  - 直接绘制散点图
  - 绘制散点图矩阵
  - 计算相关系数

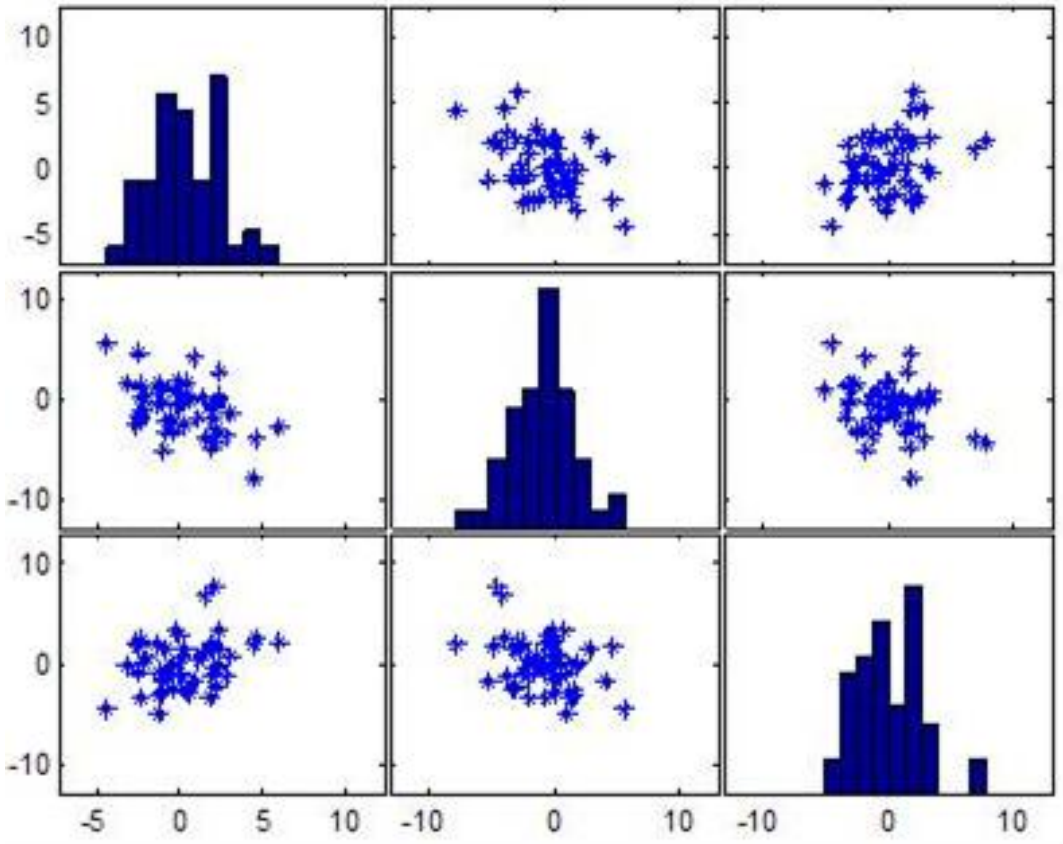
相关性分析——直接绘制散点图

- 判断两个变量是否具有线性相关关系的最直观的方法是直接绘制散点图，见下图：



相关性分析——绘制散点图矩阵

- 需要同时考察多个变量间的相关关系时，若一一绘制它们间的简单散点图，十分麻烦。此时可利用散点图矩阵来同时绘制各自变量间的散点图，这样可以快速发现多个变量间的主要相关性，这一点在进行多元线性回归时显得尤为重要。
- 散点图矩阵如下图所示：



相关性分析——计算相关系数

- 为了更加准确的描述变量之间的线性相关程度，可以通过计算相关系数来进行相关分析。在二元变量的相关分析过程中比较常用的如**Pearson相关系数**、**Spearman秩相关系数**和判定系数。

相关性分析——计算相关系数

- Pearson相关系数

一般用于对定距变量的数据进行计算，即分析两个连续性变量之间的关系，其计算公式如下：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

- Spearman秩相关系数

用于描述分类或等级变量之间、分类或等级变量与连续变量之间的关系。其计算公式如下：

$$r_s=1-\frac{6\sum_{i=1}^{n}(R_i-Q_i)^2}{n(n^2-1)}$$

其中R_i代表x_i中的秩，所谓秩指x_i在$(x_1,x_2,x_3...x_n)$中按照一定准则的排列顺序。Q_i代表y_i在$(y_1,y_2,y_3...y_n)$中的秩。

相关性分析——计算相关系数

特性

- 相关性系数取值 [-1,1]
- 接近0代表无相关性
- 接近1或-1代表强相关性
- 正数表示正相关
- 负数表示负相关

- 上述两种相关系数在实际应用计算中都要对其进行假设检验，使用t检验方法检验其显著性水平以确定其相关程度。研究表明，在正态分布假定下，Spearman秩相关系数与Pearson相关系数在效率上是等价的，而对于连续测量数据，更适合用Pearson相关系数来进行分析。

相关性分析——计算相关系数

- 餐饮销量数据和节假日、天气等因素都可能有关系，使用相关性分析可以得到餐饮销量数据和其他因素的相关性，其Python代码如下所示：

```
#-*- coding: utf-8 -*-
#餐饮销量数据相关性分析
from __future__ import print_function
import pandas as pd

catering_sale = '../data/catering_sale_all.xls'#餐饮数据，含有其他属性
data = pd.read_excel(catering_sale, index_col = u'日期')#读取数据，指定“日期”列为索引列

data.corr()#相关系数矩阵，即给出了任意两款菜式之间的相关系数
data.corr()[u'百合酱蒸凤爪']#只显示“百合酱蒸凤爪”与其他菜式的相关系数
data[u'百合酱蒸凤爪'].corr(data[u'翡翠蒸香茜饺']) #计算“百合酱蒸凤爪”与“翡翠蒸香茜饺”的相关系数
```

目录

统计特征函数

- 统计特征函数用于计算数据的均值、方差、标准差、分位数、相关系数、协方差等，这些统计特征能反映出数据的整体趋势。本小节所介绍的统计特征函数如下表所示。

方法名	函数功能	所属库
sum()	计算数据样本的总和（按列计算）	Pandas
mean()	计算数据样本的算术平均数	Pandas
var()	计算数据样本的方差	Pandas
std()	计算数据样本的标准差	Pandas
corr()	计算数据样本的 Spearman（Pearson）相关系数矩阵	Pandas
cov()	计算数据样本的协方差矩阵	Pandas
skew()	样本值的偏度（三阶矩）	Pandas
kurt()	样本值的峰度（四阶矩）	Pandas
describe()	给出样本的基本描述（基本统计量如均值、标准差等）	Pandas

统计特征函数

- sum

功能：计算数据样本的总和（按列计算）
使用格式：
D.sum()　按列计算样本D的总和，样本D可为DataFrame或者Series。

- mean

功能：计算数据样本的算术平均数
使用格式：
D.mean()　按列计算样本D的均值，样本D可为DataFrame或者Series。

- var

功能：计算数据样本的方差
使用格式：
D.var()　按列计算样本D的均值，样本D可为DataFrame或者Series。

- std

功能：计算数据样本的标准差
使用格式：
D.std()　按列计算样本D的均值，样本D可为DataFrame或者Series。

统计特征函数

- corr

功能：计算数据样本的Spearman（Pearson）相关系数矩阵
使用格式：
D.corr(method='pearson')　样本D可为DataFrame，返回相关系数矩阵，method参数为计算方法，支持pearson（皮尔森相关系数，默认选项）、kendall（肯德尔系数）、spearman（斯皮尔曼系数）；
S1.corr(S2, method='pearson')　S1、S2均为Series，这种格式指定计算两个Series之间的相关系数。

- cov

功能：计算数据样本的协方差矩阵
使用格式：
D.cov()　样本D可为DataFrame，返回协方差矩阵；
S1.cov(S2)　S1、S2均为Series，这种格式指定计算两个Series之间的协方差。

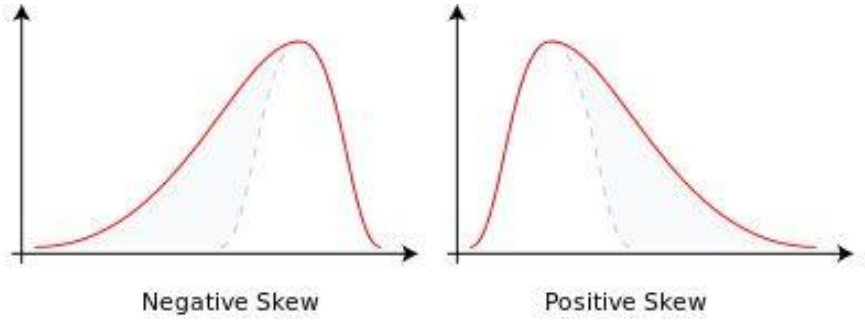
统计特征函数

- skew/kurt

功能：计算数据样本的偏度（三阶矩）/ 峰度（四阶矩）
使用格式：
D.skew() / D.kurt()　计算样本D的偏度（三阶矩）/ 峰度（四阶矩）。
样本D可为DataFrame或Series。

随机变量的偏态（衡量分布不对称性）定义为其三阶中心矩:

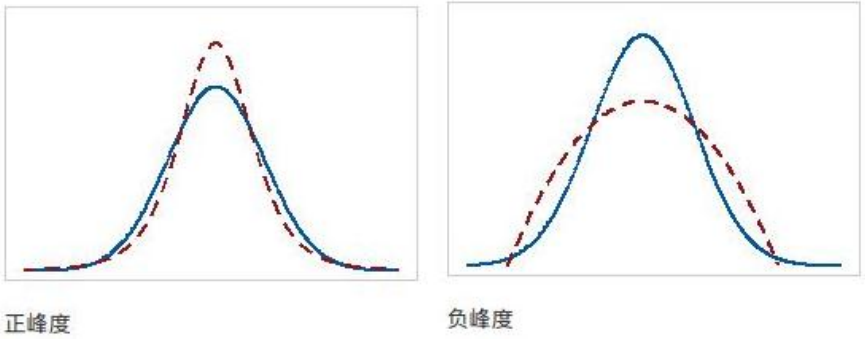
$$S(x) = \int_{-\infty}^{\infty} [x - E(x)]^3 \, f(x) \, dx$$



任何对称分布偏态为0，归一化三阶矩被成为偏斜度，向左偏斜（分布尾部在左侧较长）具有负偏度（失效率数据常向左偏斜，如极少量的灯泡会立即烧坏），向右偏斜分布（分布尾部在右侧较长）具有正偏度（工资数据往往以这种方式偏斜，大多数人所得工资较少）。

随机变量的峰度（衡量布的波峰和尾部与正态分布的区别）定义为其四阶中心矩-3:

$$K(x) = \frac{\int_{-\infty}^{\infty} [x - E(x)]^4 \, f(x) \, dx}{\sigma^2} - 3$$



完全符合正态分布的数据峰度值为0，且正态分布曲线被称为基线。如果样本峰度显著偏离0，就可判断此数据不是正态分布。

统计特征函数

- Describe

功能：直接给出样本数据的一些基本的统计量，包括均值、标准差、最大值、最小值、分位数等。

使用格式：

D.describe()　括号里可以带一些参数，比如percentiles = [0.2, 0.4, 0.6, 0.8]就是指定只计算0.2、0.4、0.6、0.8分位数，而不是默认的1/4、1/2、3/4分位数。

统计特征函数——实例

- 计算两个列向量的相关系数，采用Spearman方法

```
#生成样本 D，一行为 1~7，一行为 2~8
D = pd.DataFrame([range(1,8),range(2,9)])
#计算相关系数矩阵
D.corr(method='pearson')
#提取第一行
S1 = D.loc[0]
#提取第二行
S2 = D.loc[1]
#计算 S1、S2 的相关系数
S1.corr(S2, method='pearson')
```

统计特征函数——实例

- 计算6×5随机矩阵的协方差矩阵。

```
import numpy as np
D = pd.DataFrame(np.random.randn(6, 5)) #产生 6×5 随机矩阵
D.cov() #计算协方差矩阵
          0         1         2         3         4
0  1.745257 -0.299968  0.850216 -0.484931  1.068187
1 -1.453670  1.460928  0.347299  1.585089  0.595347
2 -0.751128  0.504498 -1.244944 -0.672183 -0.595296
3 -0.423802 -1.086470  0.637264  0.873043 -0.506736
4  0.969907  0.721997 -0.550993  1.033300 -0.903234
5 -0.705159  0.385077  0.120580  0.347470  2.036798
D[0].cov(D[1]) #计算第一列和第二列的协方差
0.5
```

统计特征函数

- 除了上述基本的统计特征外，Pandas还提供了另外一些非常方便实用的计算统计特征的函数，主要用累积计算（cum）和滚动计算（pd.rolling_）。

方法名	函数功能	所属库
cumsum()	依次给出前 1、2、…、n 个数的和	Pandas
cumprod()	依次给出前 1、2、…、n 个数的积	Pandas
cummax()	依次给出前 1、2、…、n 个数的最大值	Pandas
cummin()	依次给出前 1、2、…、n 个数的最小值	Pandas

- cum系列函数是作为DataFrame或Series对象的方法而出现的，命令格式为D.cumsum()。

```
D=pd.Series(range(0,20))
D.cumsum()
0   0
1   1
2   3
3   6
....
19  190
```

统计特征函数

- rolling_系列是pandas的函数，不是DataFrame或Series对象的方法，使用格式为pd.rolling_mean(D, k)，意思是每k列计算一次均值，滚动计算。

方法名	函数功能	所属库
rolling_sum()	计算数据样本的总和（按列计算）	Pandas
rolling_mean()	计算数据样本的算术平均数	Pandas
rolling_var()	计算数据样本的方差	Pandas
rolling_std()	计算数据样本的标准差	Pandas
rolling_corr()	计算数据样本的 Spearman（Pearson）相关系数矩阵	Pandas
rolling_cov()	计算数据样本的协方差矩阵	Pandas
rolling_skew()	样本值的偏度（三阶矩）	Pandas
rolling_kurt()	样本值的峰度（四阶矩）	Pandas

```
D=pd.Series(range(0,20))
pd.rolling_sum(D,2)
0   NaN
1   1.0
2   3.0
3   5.0
4   7.0
....
19  37.0
```

目录

统计作图函数

- 通过统计作图函数绘制的图表可以直观地反映出数据及统计量的性质及其内在规律，如盒图可以表示多个样本的均值，误差条形图能同时显示下限误差和上限误差，最小二乘拟合曲线图能分析两变量间的关系。如在做某用电单位用电量趋势预测过程中，可以先分析该用电单位日用电量的时序图，来直观地估计其用电量变化趋势。
- Python的主要作图库是Matplotlib，而Pandas基于Matplotlib并对某些命令作了简化，因此作图通常是Matplotlib和Pandas相互结合着使用。

统计作图函数

- Python中的常用作图函数。

作图函数名	作图函数功能	所属工具箱
plot()	绘制线性二维图，折线图	Matplotlib/Pandas
pie()	绘制饼型图	Matplotlib/Pandas
hist()	绘制二维条形直方图，可显示数据的分配情形	Matplotlib/Pandas
boxplot()	绘制样本数据的箱型图	Pandas
plot(logy = True)	绘制 y 轴的对数图形	Pandas
plot(yerr = error)	绘制误差条形图	Pandas

统计作图函数

- 在使用Python作图之前，我们通常要加载以下代码：

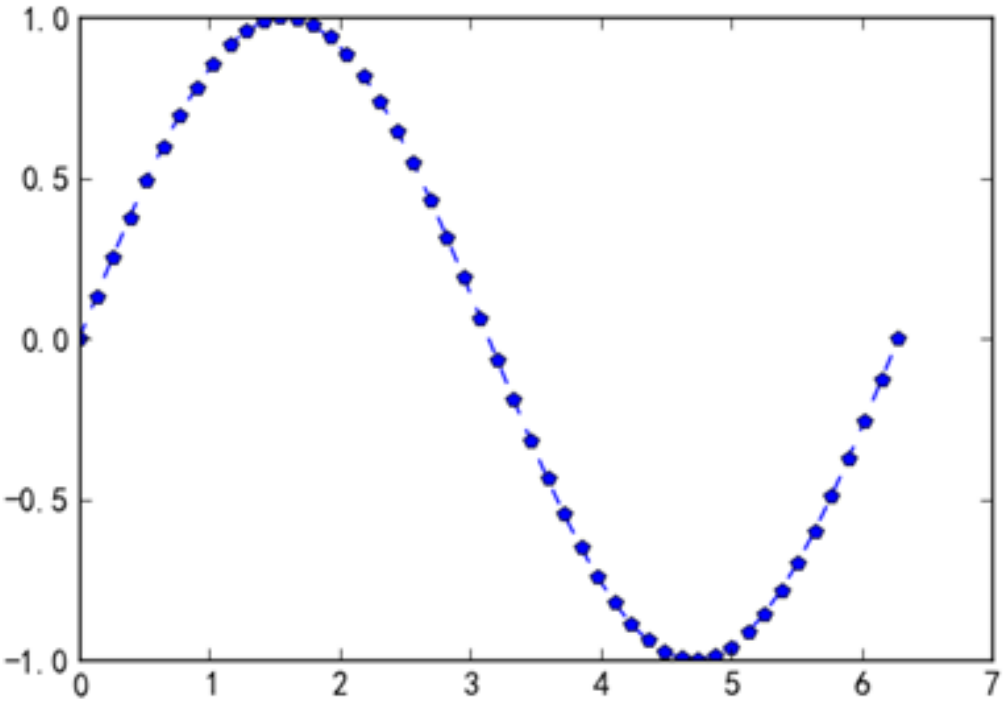
```
import matplotlib.pyplot as plt #导入作图库
plt.rcParams['font.sans-serif'] = ['SimHei'] #用来正常显示中文标签
plt.rcParams['axes.unicode_minus'] = False #用来正常显示负号
plt.figure(figsize = (7, 5)) #创建图像区域，指定比例
```

统计作图函数

- plot

功能：绘制线性一维图、折线图。

使用格式：



plt.plot(x, y x（即以x
为横轴的二 式和颜色
，常用的选 ’ o’ 为
圆圈、’ +’ 、y均为实
数同维向量
D.plot(kind 内置的方
法作图，默 通过kind
参数指定作 st（直方
图）、box（ 等，同时也
能够接受plt 为Pandas
中的对象，那么以这种方式作图是比较简洁的。

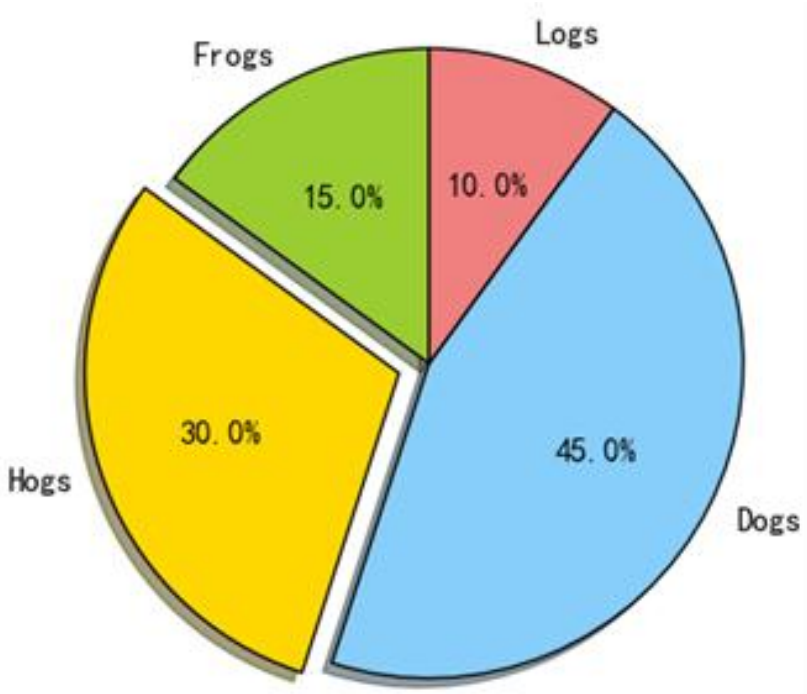
统计作图函数

- pie

功能：绘制饼型图。

使用格式：

plt.pie(size)　使用Matplotlib绘制饼图，其中size是一个列表，记录各个扇形的比例。



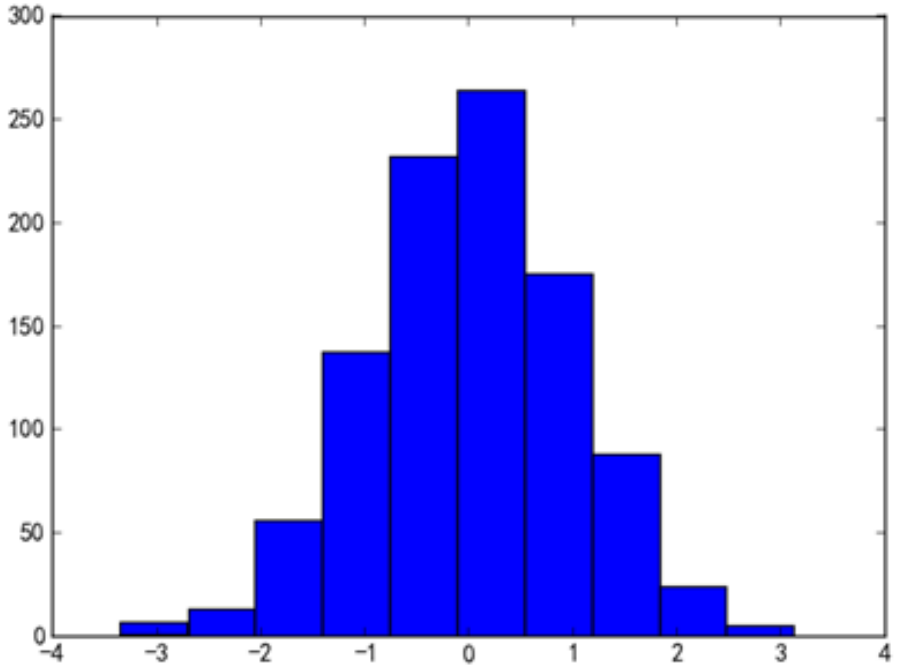
统计作图函数

- hist

功能：绘制二维条形直方图，可显示数据的分布情形。

使用格式：

Plt.hist(x, y)　其中x是待绘制直方图的一维数组，y可以是整数，表示均匀分为n组；也可以是列表，列表各个数字为分组的边界点（即手动指定分界点）。



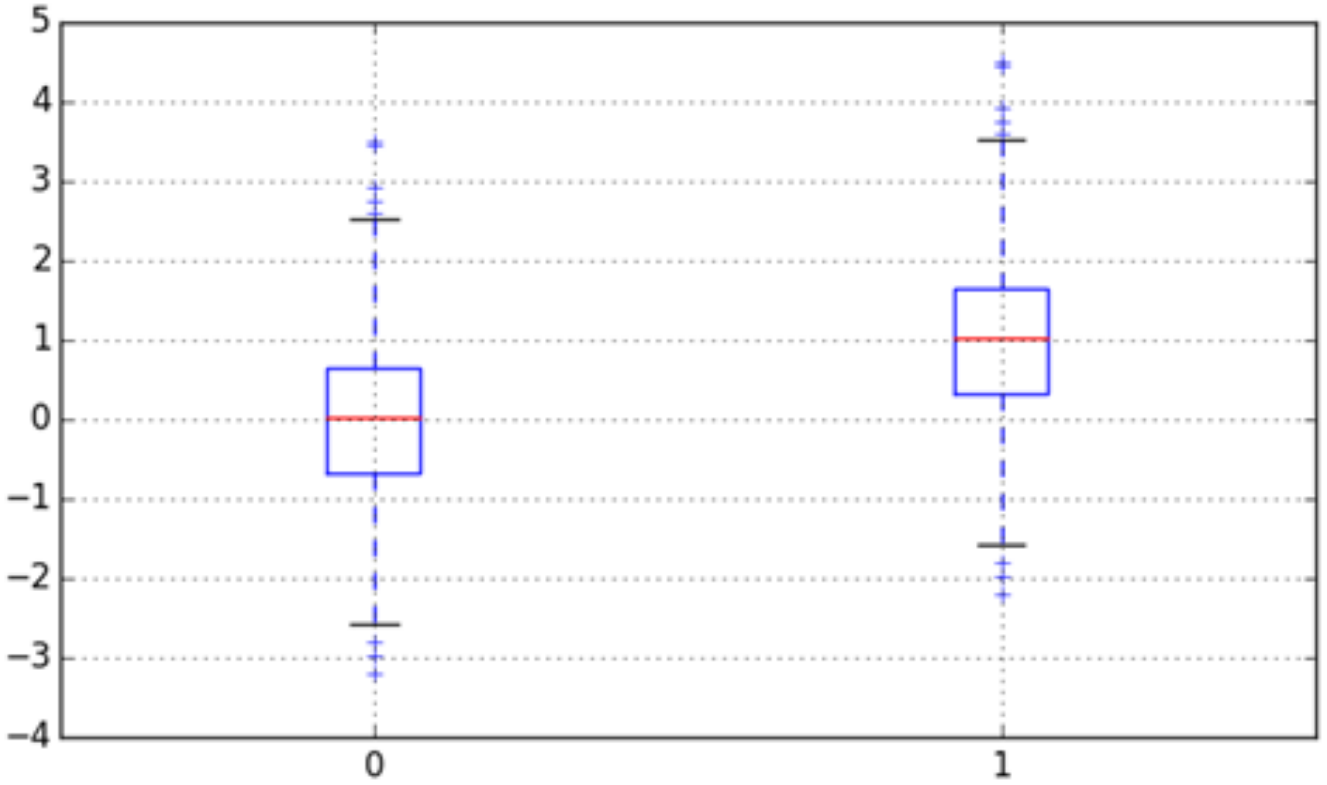
统计作图函数

- boxplot

功能：绘制样本数据的箱型图。

使用格式：

D.boxplot() / D.plot(kind = 'box')　有两种比较简单的方式绘制D的箱型图，其[illegible]种是调用Serie[illegible]（box）。其[illegible]延伸出去的直[illegible]在须的底部有-



统计作图函数

- plot(log

功能：绘制

使用格式：

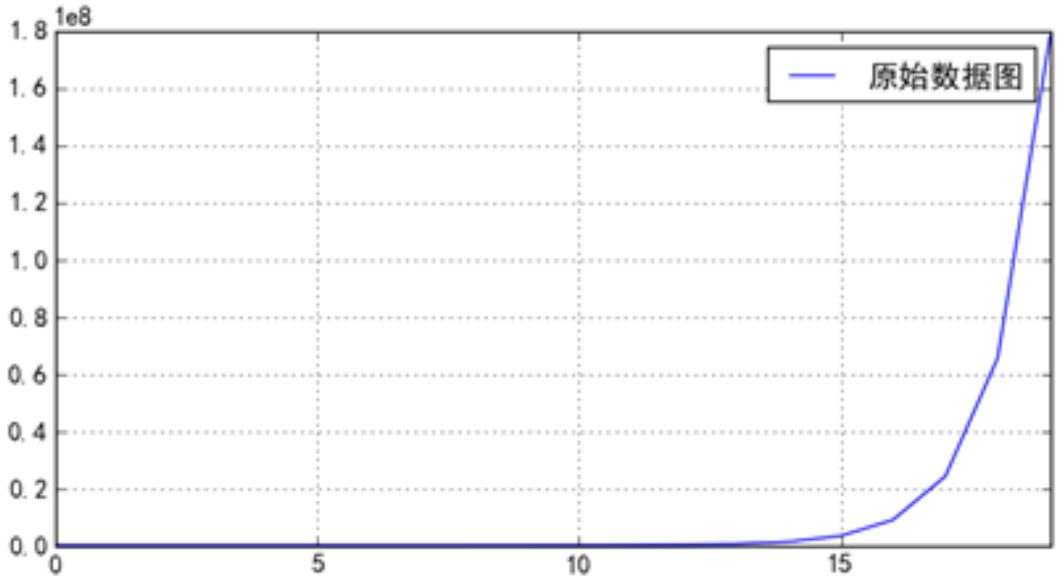
D.plot(log

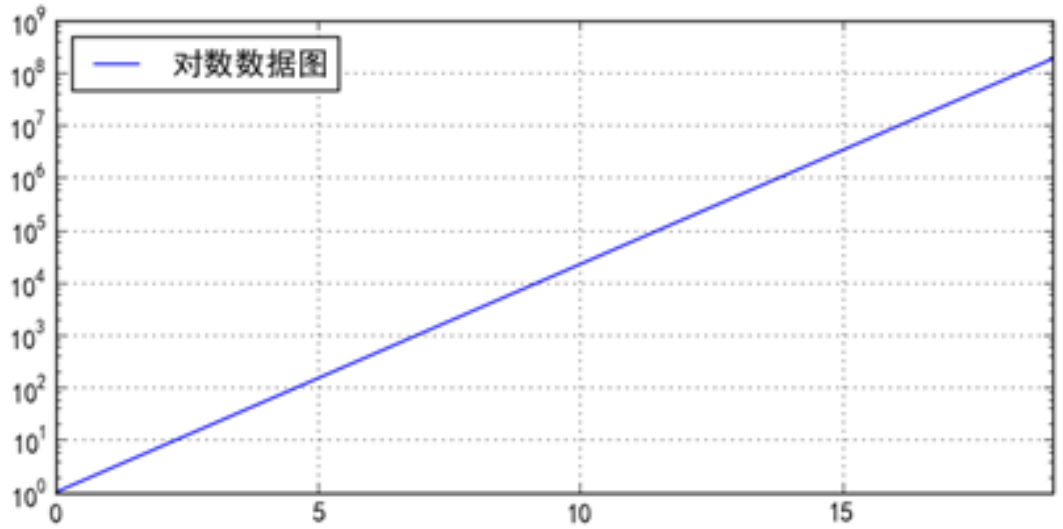
（以10为底

Pandas的D

数刻度

为





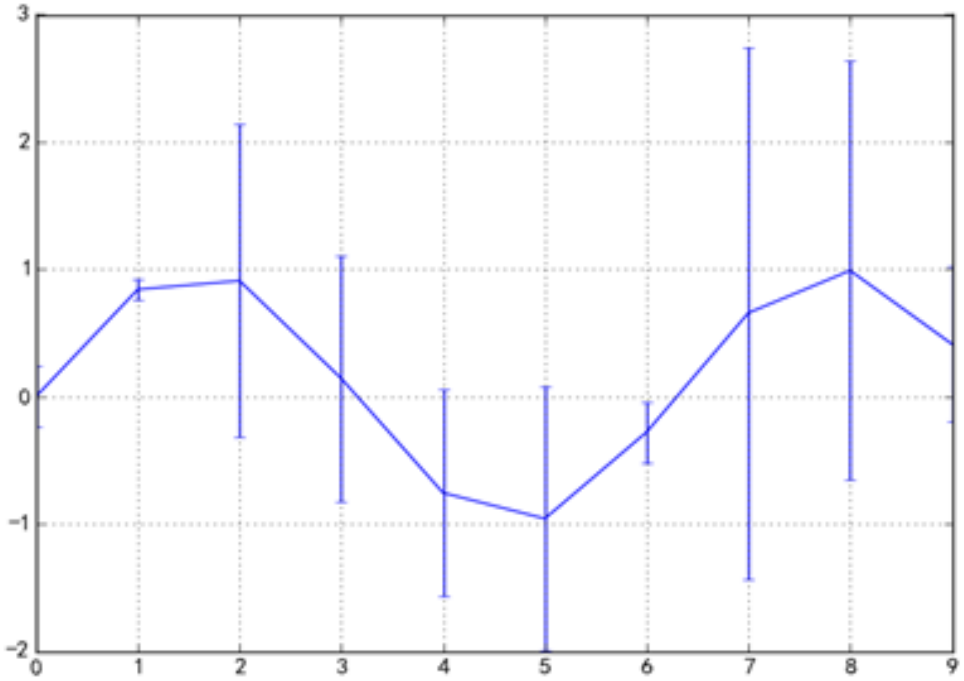
统计作图函数

- plot(yerr = error)

功能：绘制误差条形图。

使用格式：

D.plot(yerr = error)　绘制误差条形图。D为Pandas的DataFrame或Series，代表着均值数据列，而error则是误差列，此命令在y轴方向画出误差棒图；类似地，如果设置参数xerr = error，则在x轴方向画出误差棒图。



Thank You!