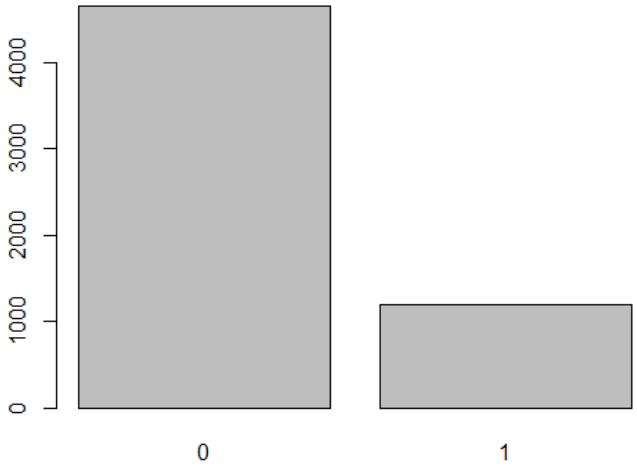
数据探索----描述性统计、BI

描述分类变量的分布

频数表

	频次	百分比
正常	4648	79.50
违约	1197	20.40

柱形图

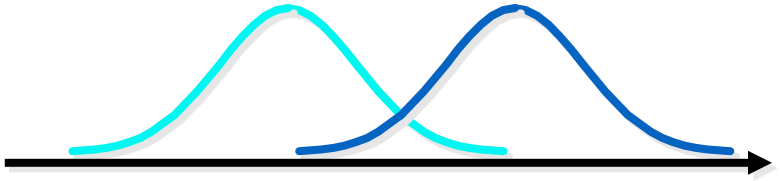


描述连续变量的分布

- 需要对**变量进行分布探索**,并了解以下情况:

集中趋势

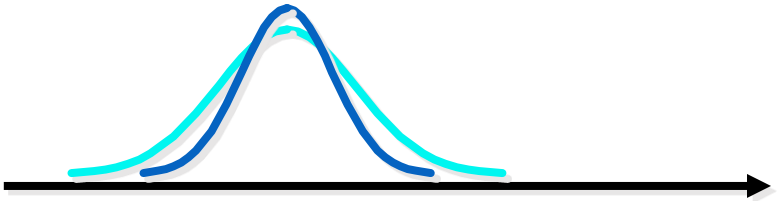
(位置)



什么统计量可以概括这个变量?

离中趋势

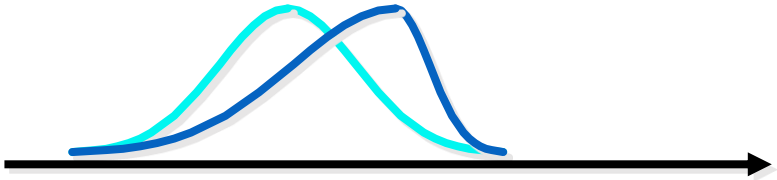
(分散程度)



这个统计量的概括能力有多强?

偏态和峰态

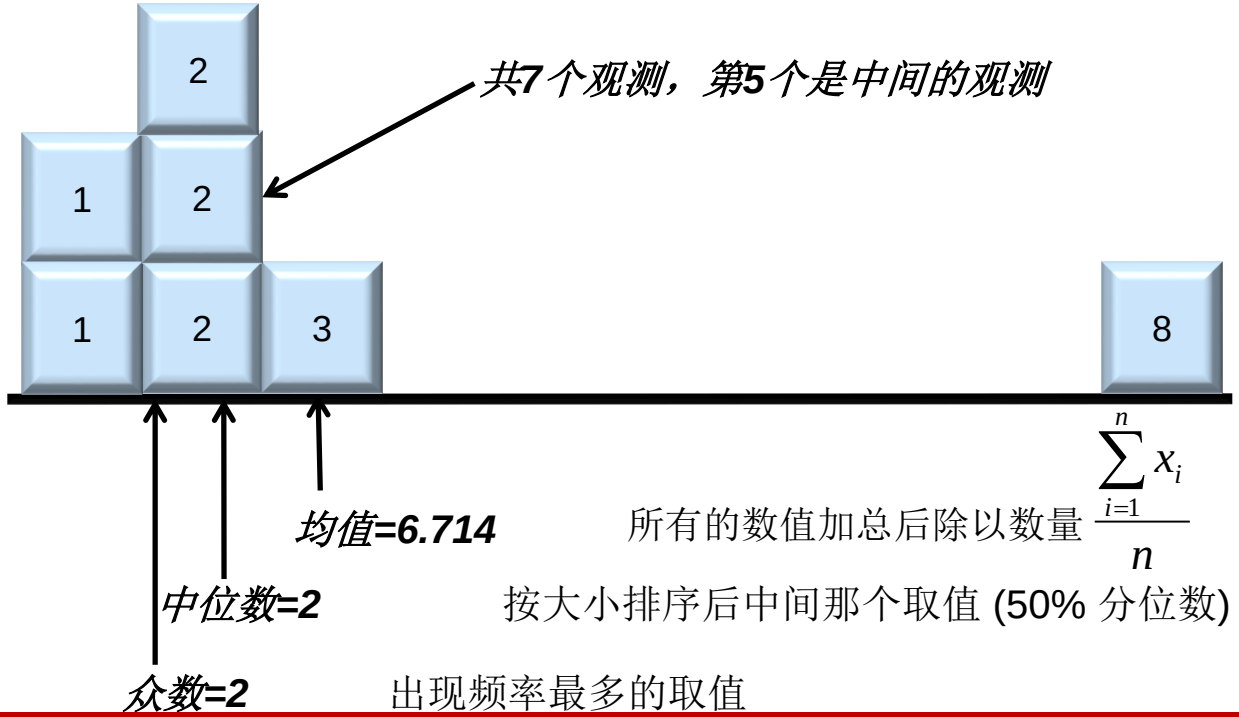
（形状）



选择哪个统计量更”科学”?

数据的位置

中心的度量– 均值、中位数、众数



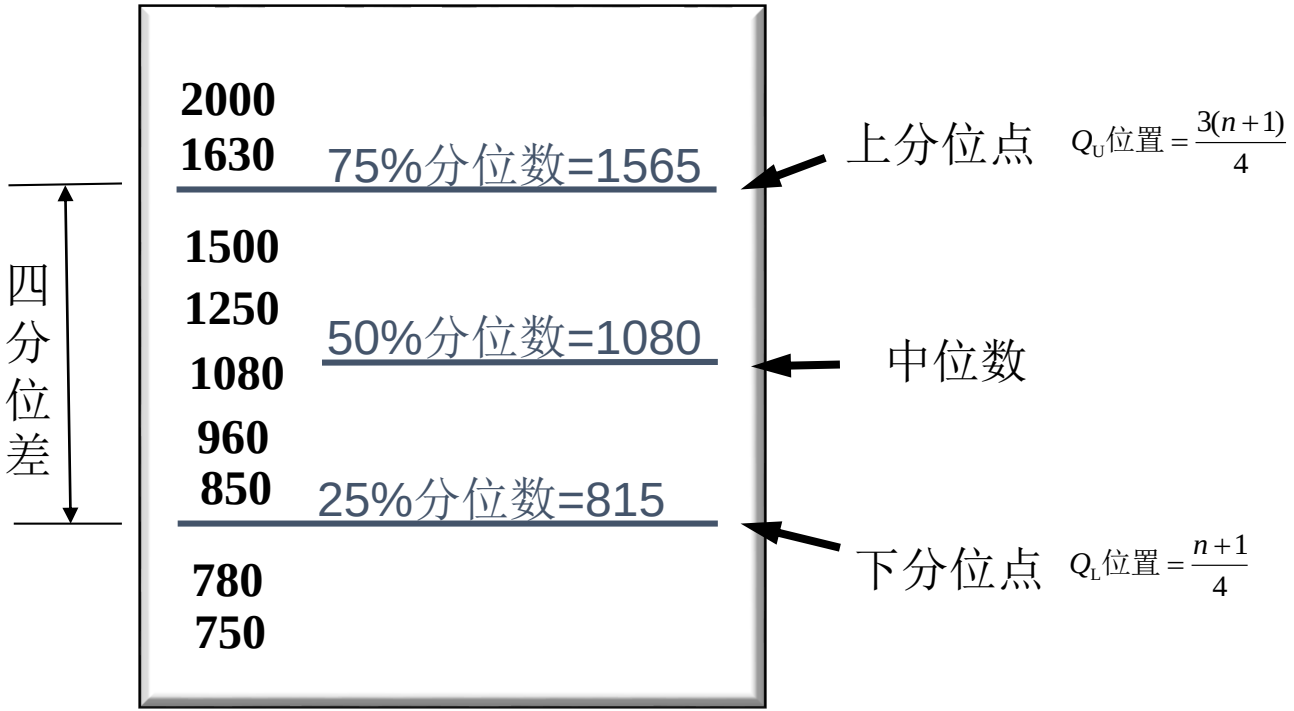
$\frac{\sum_{i=1}^{n} x_i}{n}$ 所有的数值加总后除以数量

按大小排序后中间那个取值 (50% 分位数)

出现频率最多的取值

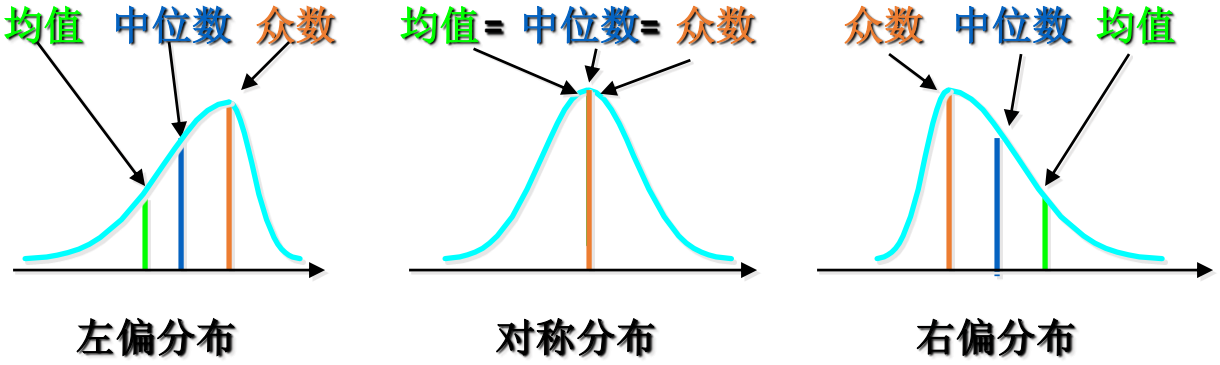
数据的位置

百分位数


2000
1630
75%分位数=1565
1500
1250
50%分位数=1080
1080
960
850
25%分位数=815
780
750
四分位差
上分位点
Q_U位置 $= \frac{3(n+1)}{4}$
中位数
下分位点
Q_L位置 $= \frac{n+1}{4}$

数据的位置

- 众数、中位数和平均数的关系



数据的离散程度

极差(Range)

极差=最大值-最小值

平均绝对偏差(Mean Absolute Deviation)

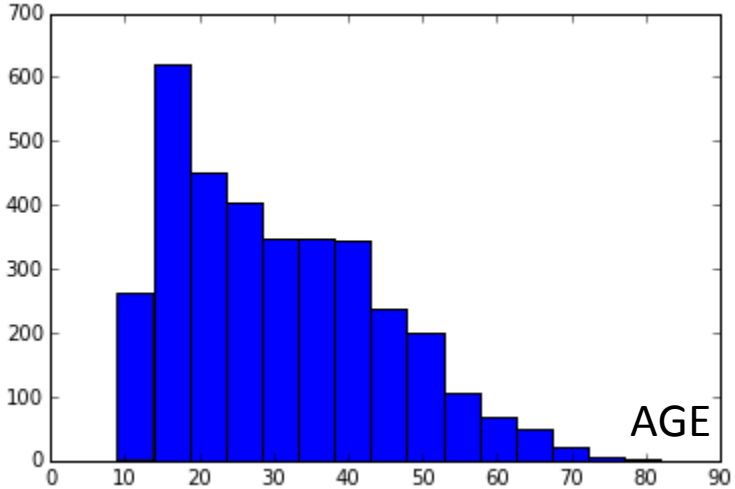
$$MAD = \frac{1}{n}\sum_{i-1}^{n}\left|x_i - \bar{x}\right|$$

方差(Variance)和标准差(Standard Deviation)

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

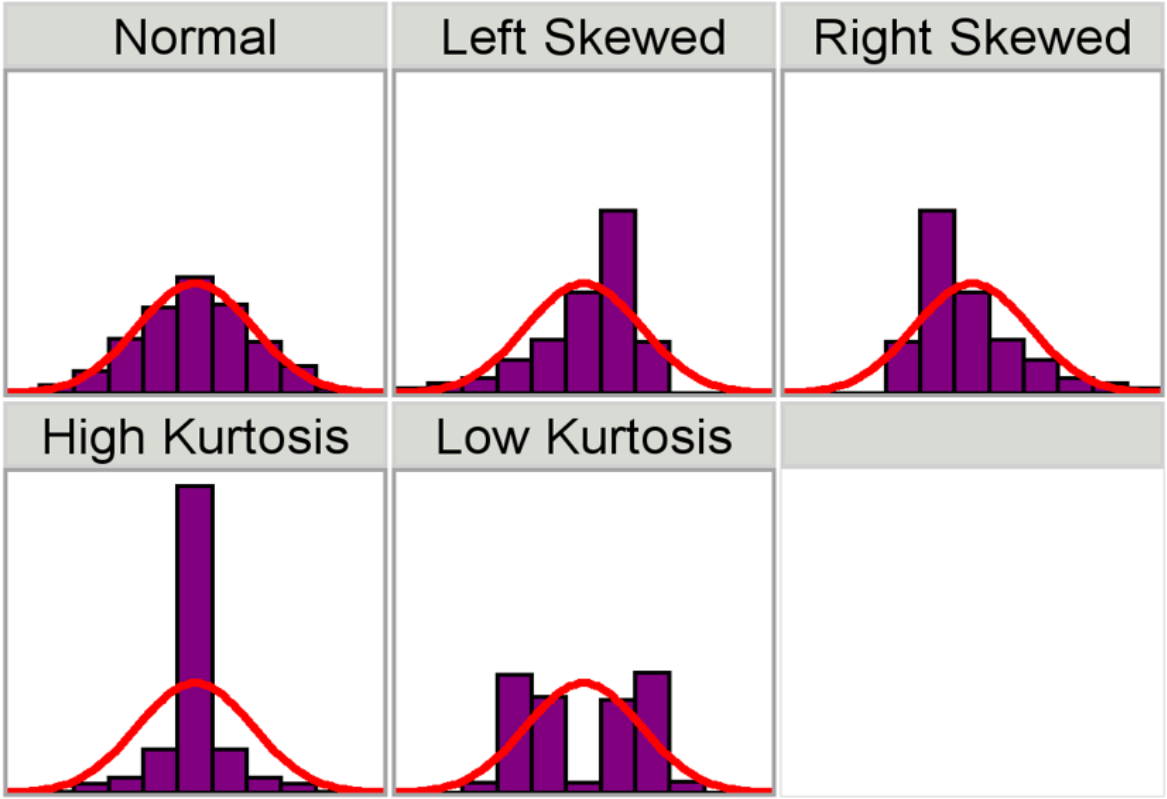
$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

直方图

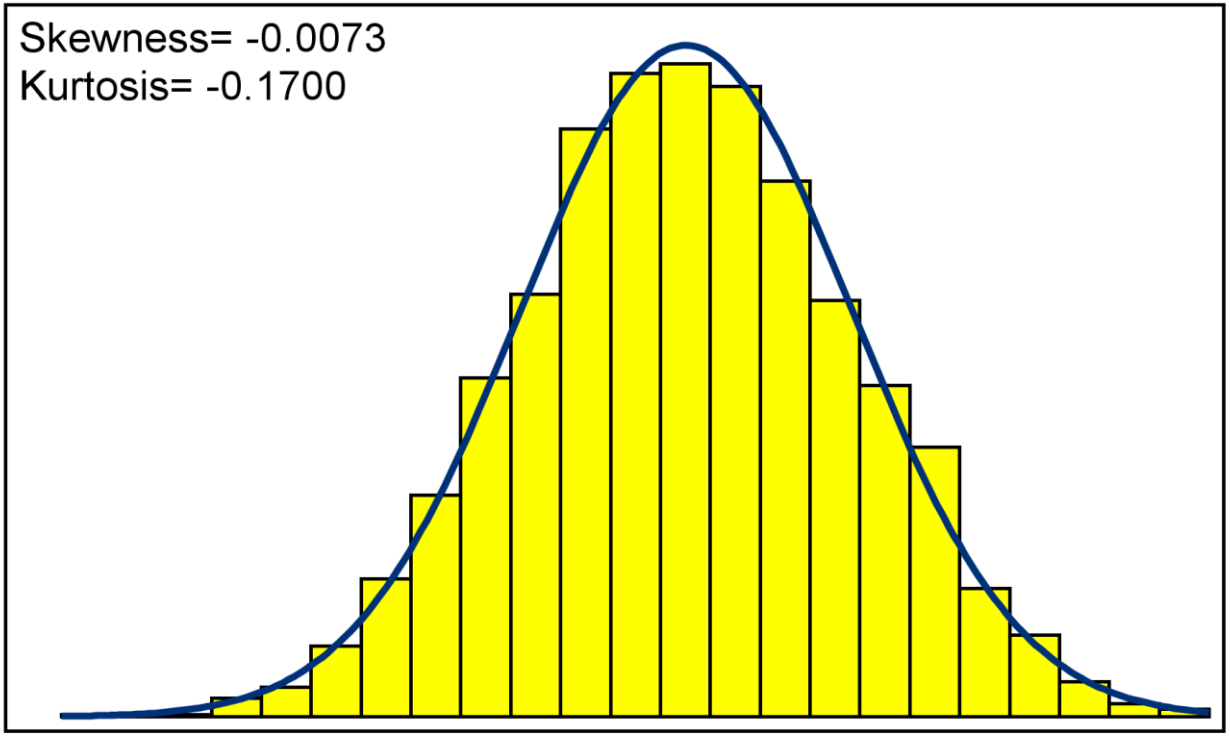


- 直方图常用于了解数据的分布形状
- **一般情况下**，横轴为连续变量的分段进行等宽离散后的值，纵轴为频次；
- 每个柱的宽度可不相同，纵轴也可以不是频次，通过bins和normed参数可进行相应设置

数据的偏态与峰度

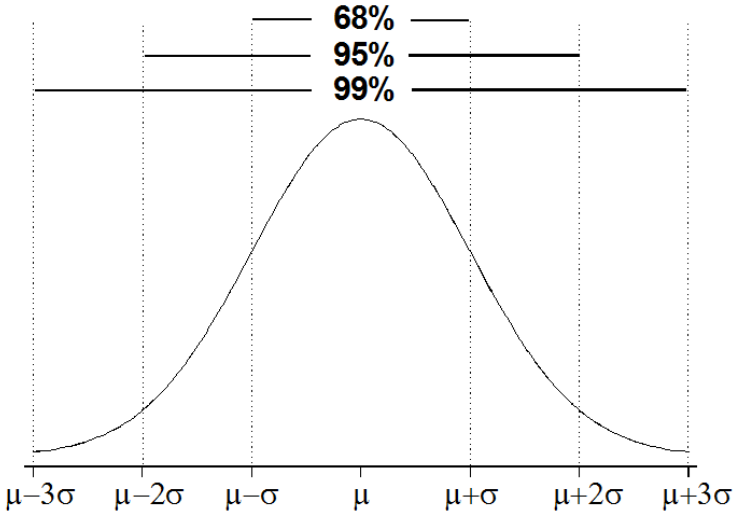

Normal
Left Skewed
Right Skewed
High Kurtosis
Low Kurtosis

正态分布的偏度与峰度



常见分布

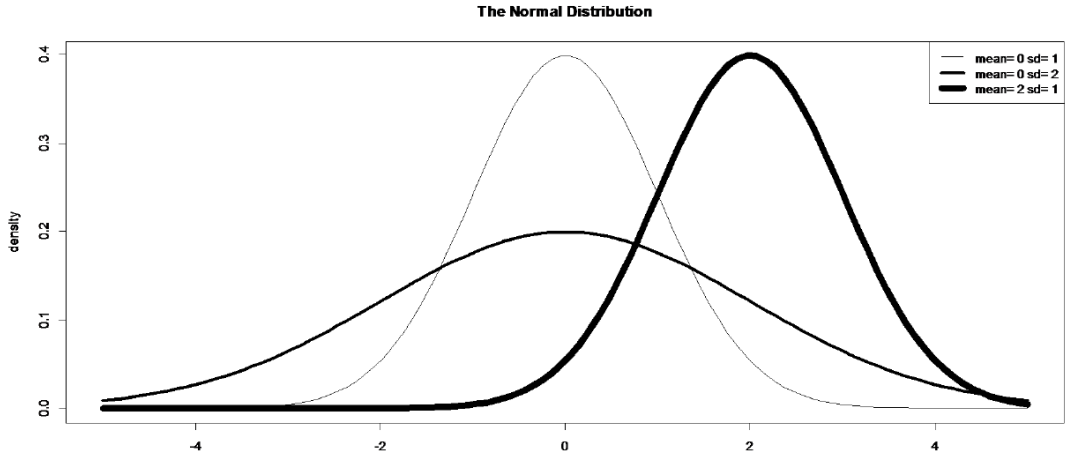
正态分布



对称（symmetric）. 关于均值左右对称分布.
均值和标准差的代表性（ ***fully characterized***） 只要知道其均值和标准差，这个变量的分布情况就完全知道了.
倒钟形.
均值= 中位数 =众数.

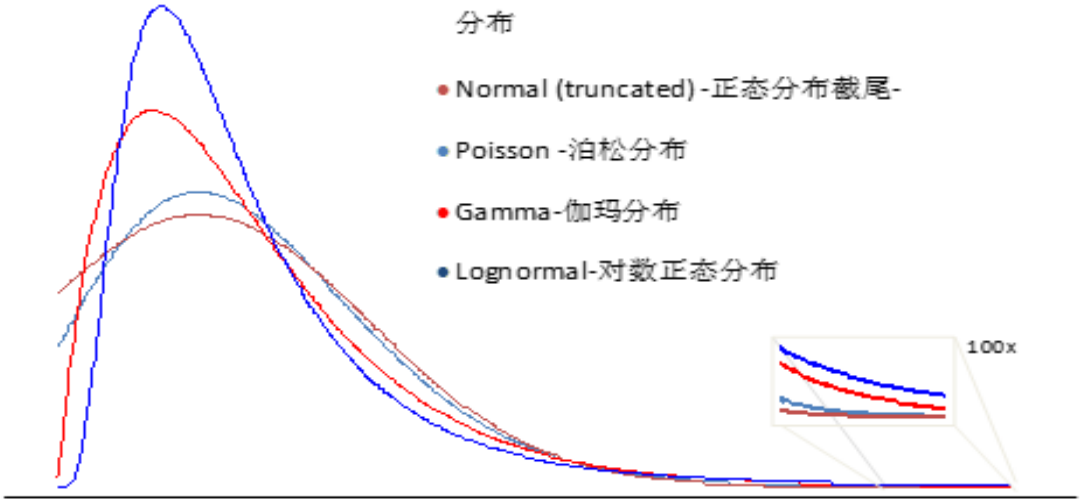
常见分布

- 正态分布



不同参数下的正态分布

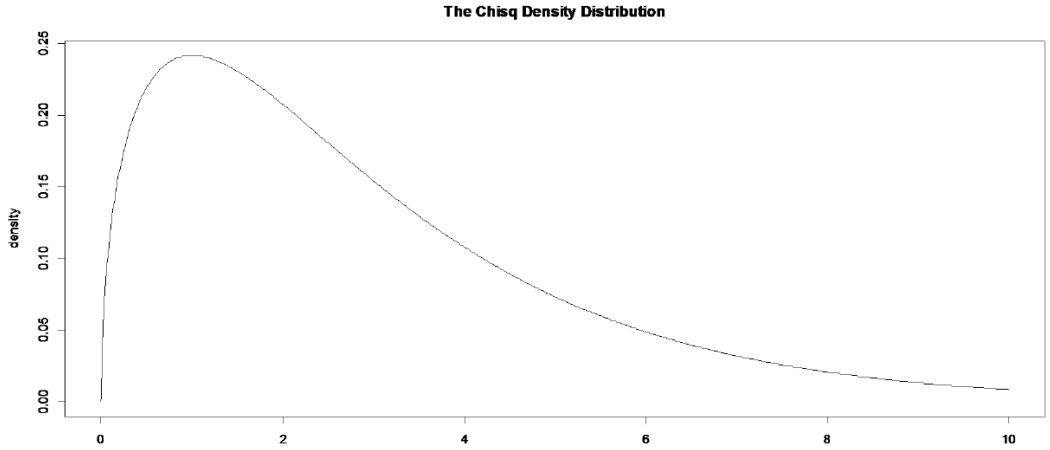
其它常见分布形式



其中对数正态分布在统计分析中运用最为广泛，顾名思义，这种类型的分布在取对数之后服从正态分布。因为其具有这样的良好属性，在精确度要求并不严格的统计分析中，经常对偏态分布首先进行对数转换。

常见统计量的分布

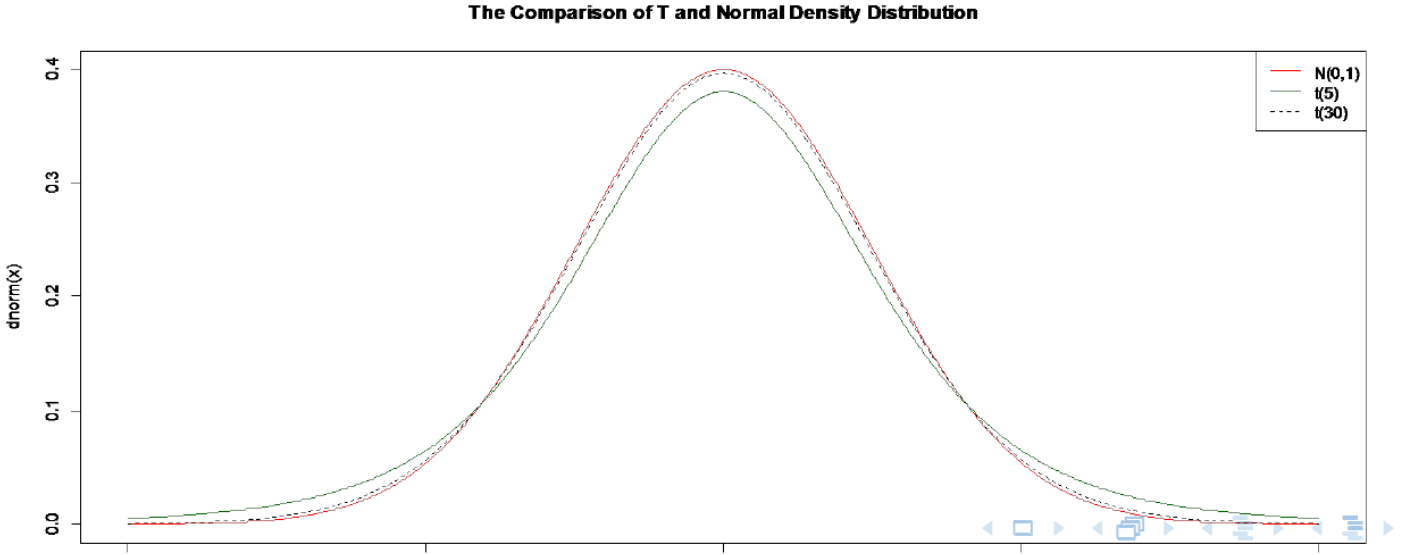
- 卡方分布



卡方分布(自由度为3)

常见统计量的分布

- T分布



T分布(实线自由度为5，虚线自由度为30，红线为标准正态分布)

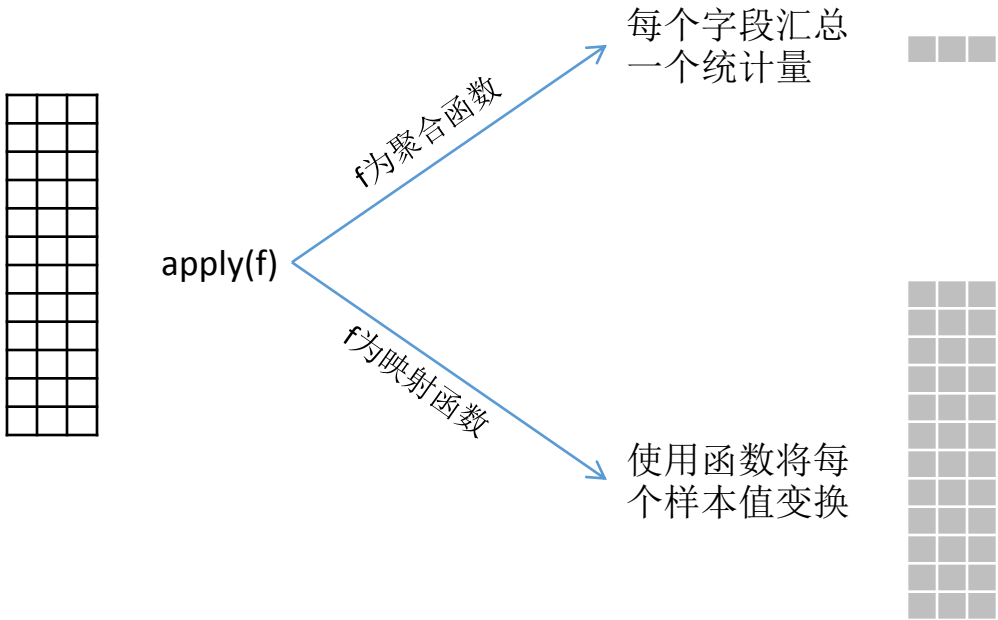
4.2 apply\map\groupby及其它相关功能

常用的汇总分析

- 日常的数据分析当中经常需要生成报表，其应用广泛，结果简单易懂：
  - 聚合(汇总)：pandas提供了比reduce更强大的汇总方法-apply
  - 映射：使用“广播”或使用与列表、数组相类似的方法-map
  - 分组汇总：使用groupby按字段分组，再使用aggregate进行汇总
  - 交叉表：多个字段交叉，汇总频次、均值等
  - 其它：transform、agg等等

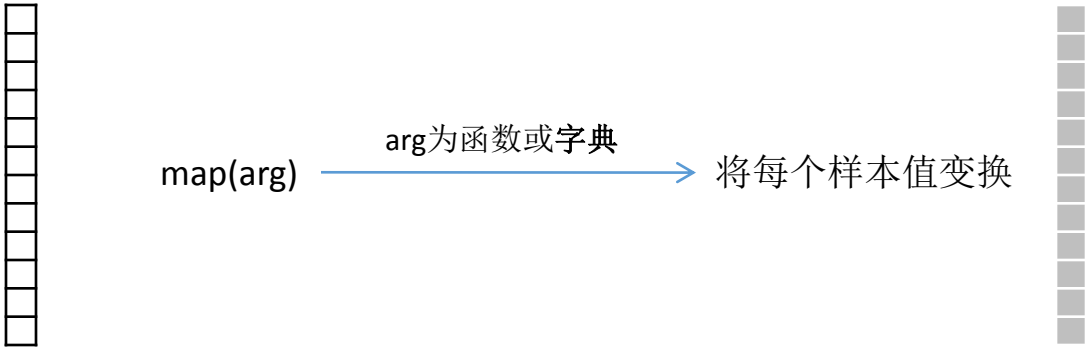
apply

- 可将函数应用到每个字段，根据函数的不同可以用于“聚合”数据或“映射”数据



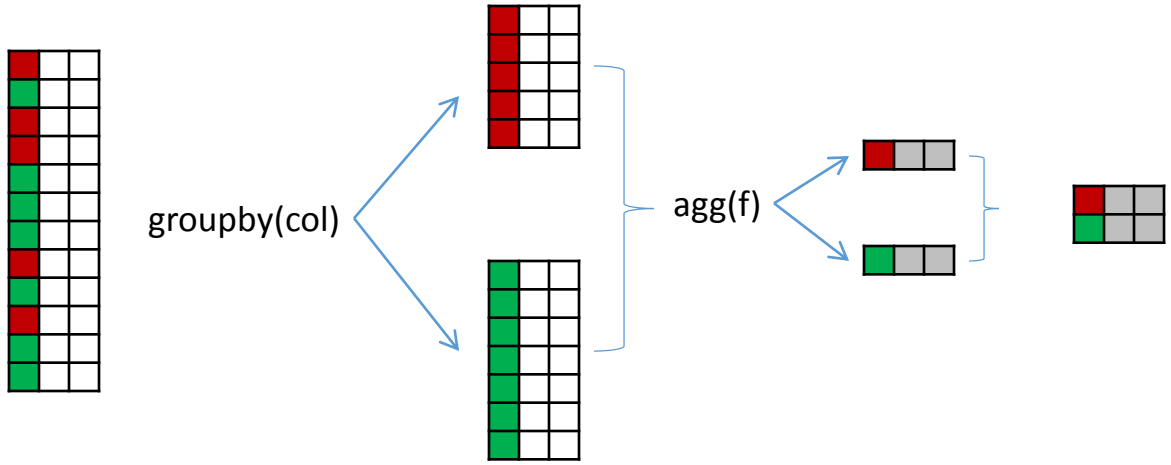
map

- map方法可以将某个字段的每个值使用函数进行变换，类似于内建的map方法，仅能对单独字段（series）使用
- 可以使用字典进行map，例如将“物品编号-物品名”字典传入，可将编号字段映射成名称字段



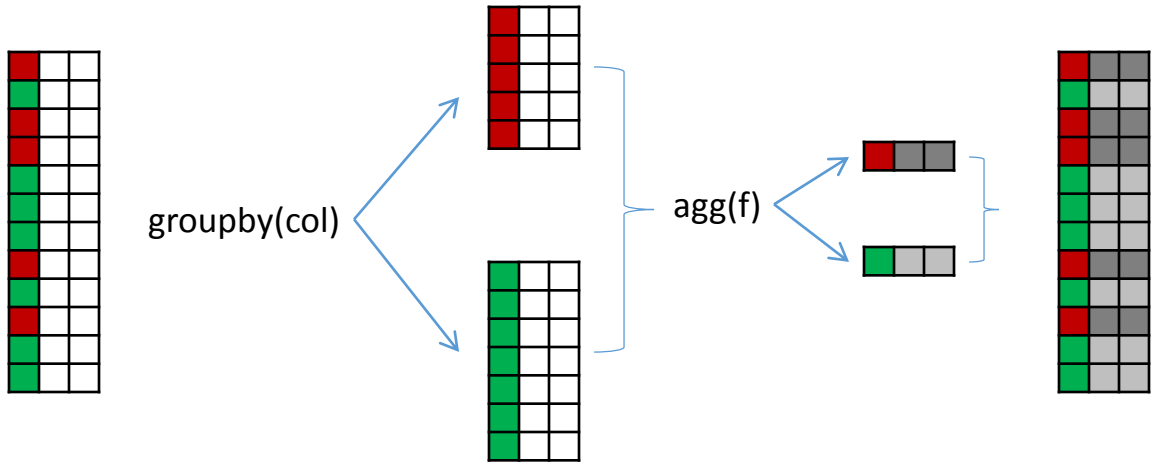
groupby和aggregate

- 使用groupby将数据分组，使用aggregate将每个分组进行聚合，类似还可以使用agg和apply
- 用于分组的字段需要与待聚合的字段有同样长度，可以**按多个字段进行分组**，也可以一次聚合多个汇总字段



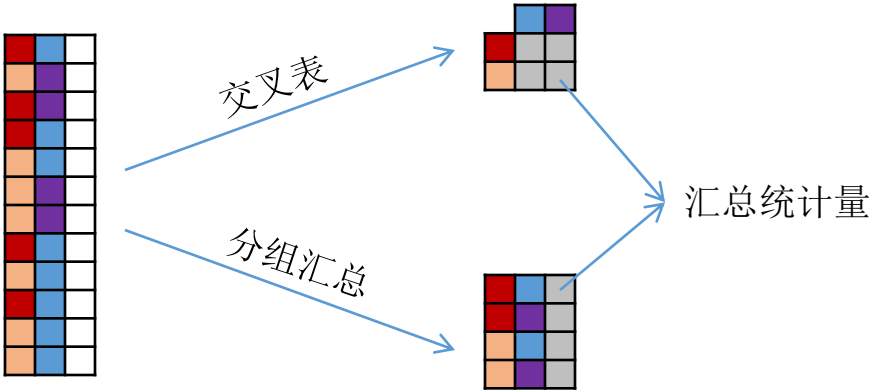
transform

- 使用groupby将数据分组，使用transform将每个分组进行聚合，聚合的结果返回到原数据集中



交叉表

- 按照多个字段汇总统计量可以使用交叉表，其结果的可读性高于分组汇总表
- 可以交叉两个或多个字段



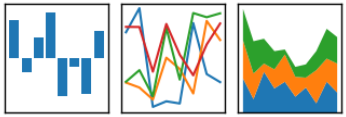
4.3 Python绘图

Python绘图功能

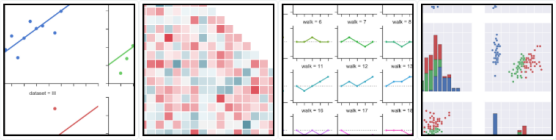
matplotlib

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Seaborn: statistical data visualization

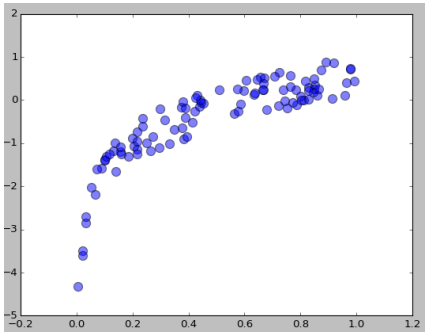


pandas.plot是matplotlib.pyplot.plot的简单包装

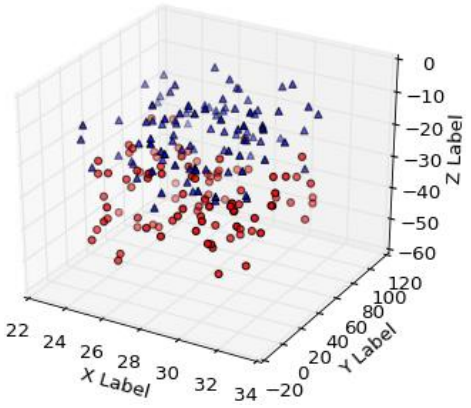
seaborn在matplotlib的基础上丰富了绘图样式

简单散点图

- 散点图（scatter diagram）用于观察变量之间关系，例如应变量随自变量变化的大致趋势



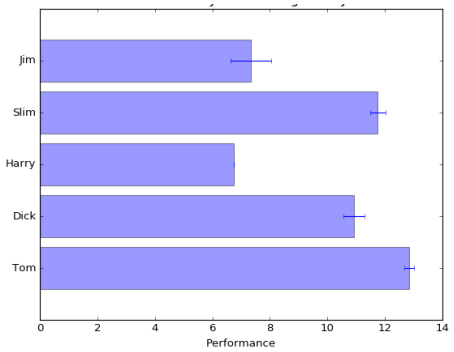
二维散点图



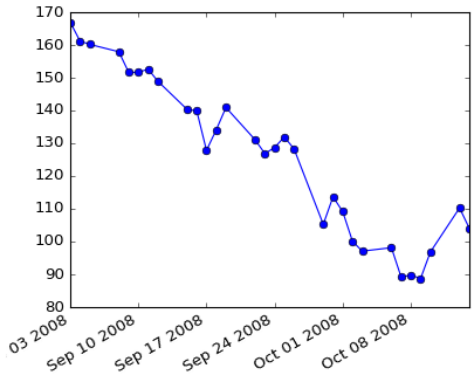
三维散点图

柱图、折线图

- 柱图（Bar）、折线图（Lines）显示随自变量变化而变化的连续数据，因此非常适用于显示在相等时间间隔下数据的趋势。
- 类别数据沿水平轴均匀分布，所有值数据沿垂直轴均匀分布。



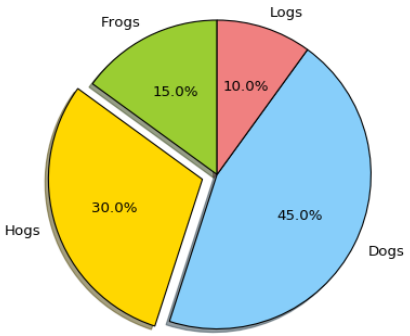
条形图（柱图）

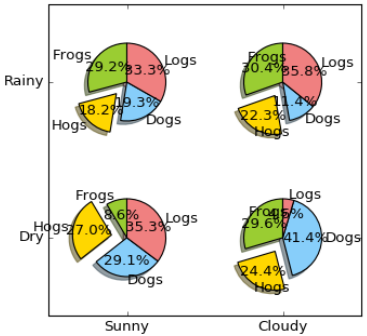


折线图

饼图

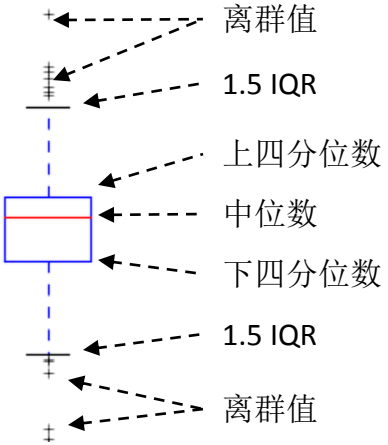
- 饼图（pie chart）用于显示一个变量各项的大小，在图中各数据点大小对应面积
- 常用于分组汇总后的数据在各组之间的比较，例如比例、份额等

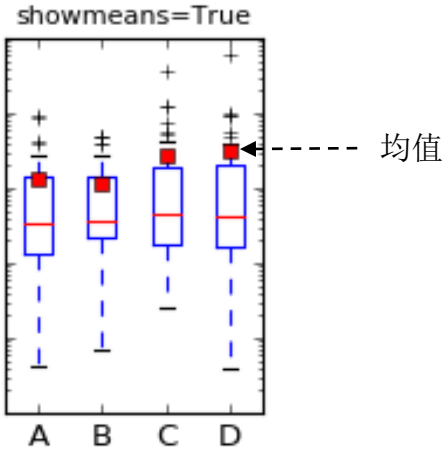




箱线图

- 也叫盒须图（Box-plot），用于显示一组数据分散情况的统计图，可以显示中位数、均值、四分位数、离群值等信息
- 常用于多组之间数据分布的比较





盒须图/箱线图

盒须图能够提供某变量分布以及异常值的信息，其通过分位数来概括某变量的分布信息从而比较不同变量的分布。盒须图的基本元素包括：

- IQR：变量上下四分位数之间的数据，这个范围代表了数据中间50%的数据。
- 中位数位置：中位数位置即代表变量中位数在总体分布中的位置。
- 1.5IQR：上下1.5IQR表示上下1.5倍IQR范围的数据，其能够提供中位数左右95%的置信区间的数据。可以直观的从盒须图中看出超出95%置信区间范围的数据，即异常值。
- 不同变量的盒须图比较时，可以通过中位数位置来比较两变量数据的中位数差异状况。