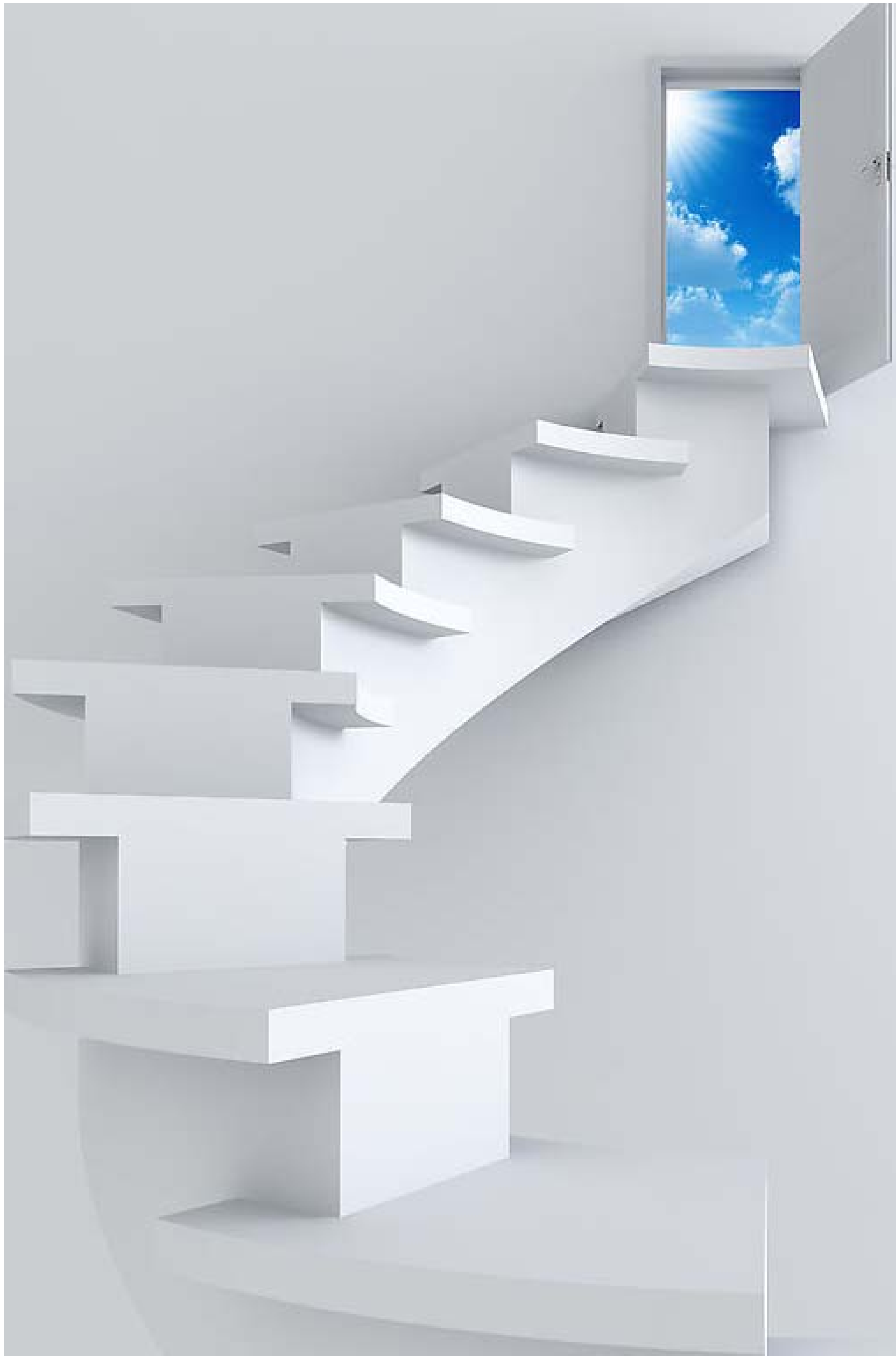


商务智能

Business Intelligence

认识数据(II)

朱卫平 博士

计算机学院

武汉大学

数据的基本统计描述

- 为了更好的做数据预处理，对数据有整体的了解很关键。基本的统计描述能鉴别数据，分辨出噪声和离群点。

中心性度量

- 问题
  - 假定有属性X的N个观察到的值， $x_1, x_2, \ldots . x_N$ 。
  - 如果我们画出它的分布图，绝大部分的值会落在哪里呢?这就是数据的中心性问题。
- 衡量中心性的测量有均值、中值、众数和中列数。

中心性度量：均值

- 最常用和最有效的测量是数据的（算术）均值 (mean)。计算公式是：

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

- 有时候，每一个x_i有一个关联的权重w_i, 权值表示相应值的重要性、显著性或者发生频率。这时候，平均值的计算公式为：

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_N x_N}{w_1 + w_2 + \ldots + w_N}$$

这称为加权算术平均值或者加权平均。

中心性度量：均值

- **例** 假设我们有salary的如下值（以千美元为单位），按递增次序显示：30,31,47,50,52,52,56,60,63,70,70,110，计算salary的平均值。

$$\bar{x} = \frac{30+31+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58$$

因此，salary的均值为58 000美元

中心性度量：均值

- 均值对极端值比较敏感。比如一个公司的员工平均薪水可能被少数高新的经理提高很多。
- 为了处理这种由少数极端值带来的效果，可以使用削减均值，即去掉极端大和极端小的值之后的平均值。比如，把薪水排序，然后去掉2%的最大值和最小值。
- 这样的均值称为截尾均值（trimmed mean）
- 应该避免削减太多（比如20%），这会导致数据信息的丢失。

中心性度量：中位数

- 对于偏斜（不对称）的数据，使用中位数（中值）(median)是更好的中心性测量。
- 中值是一系列排序好的数据的中点的值。该值把数据集分成2个部分，一半值大的，一半值小的。
- 中值一般用在数值型数据上。这里，中值可以扩展到次序属性上。
- 次序属性：将数据集的N个值按升序排列。如果N为奇数，中值即是排序集合的中点的值；如果N为偶数，中值可以是中点的2个值中的任意值。
- 数值型数据：传统上中值取两个中点数的均值。

中心性度量：中位数

例

假设我们有salary的如下值（以千美元为单位），按递增次序显示：30,31,47,50,52,52,56,60,63,70,70,110

找出其中的中位数。

该数据已经按递增序排序。有偶数个(12个)观测值，因此中位数可以是最中间两个值52和56（即列表中的第6和第7个值）中的任意值。

根据约定，我们指定这两个最中间的值的平均值为中位数。即(52+56)/2=54。

假设我们只有该列表的前11个值。给定奇数个值，中位数是最中间的值。这是列表的第6个值，其值为52。

中心性度量：中位数

- 当数据集很大时，计算中值代价很高。对于数值型属性，比较容易计算其近似值。
- 如果将数据根据值以区间分组，每个区间的频率已知。比如，雇员按照年薪间隔$10,000-20,000, $20,000-30,000等分组。称包含中位数频率的区间为中值区间。可以通过下面的公式近似计算整个数据集的中位数

$$\text{median} = L_1 + \left(\frac{\frac{N}{2} - \left(\sum freq\right)_l}{freq_{median}}\right) width$$

L_1是中值区间的最低值，N是数据值的个数，$(\sum \text{freq})_l$是所有低于中值区间间距的频率之和。$freq_{median}$ 是中值区间的频率，$width$是中值区间的宽度

中心性度量：众数

- 众数（mode）是另一个衡量中心性的测量。众数是一系列数据中出现频率最高的值。
- 众数可以是定性的也可以是定量的属性。有可能好几个不同的值都出现大量的频率，导致众数不止一个。众数有1个、2个、3个的分别称为unimodal（单峰值）, bimodal（二峰值）, trimodal（三峰值）。
- 一个极端的例子，如果每个数据值都仅出现一次，则没有众数。

中心性度量：众数

举例:

假设我们有salary的如下值（以千美元为单位），按递增次序显示：30,31,47,50,52,52,56,60,63,70,70,110。求其众数

- 2个众数：52和70

中心性度量：中列数

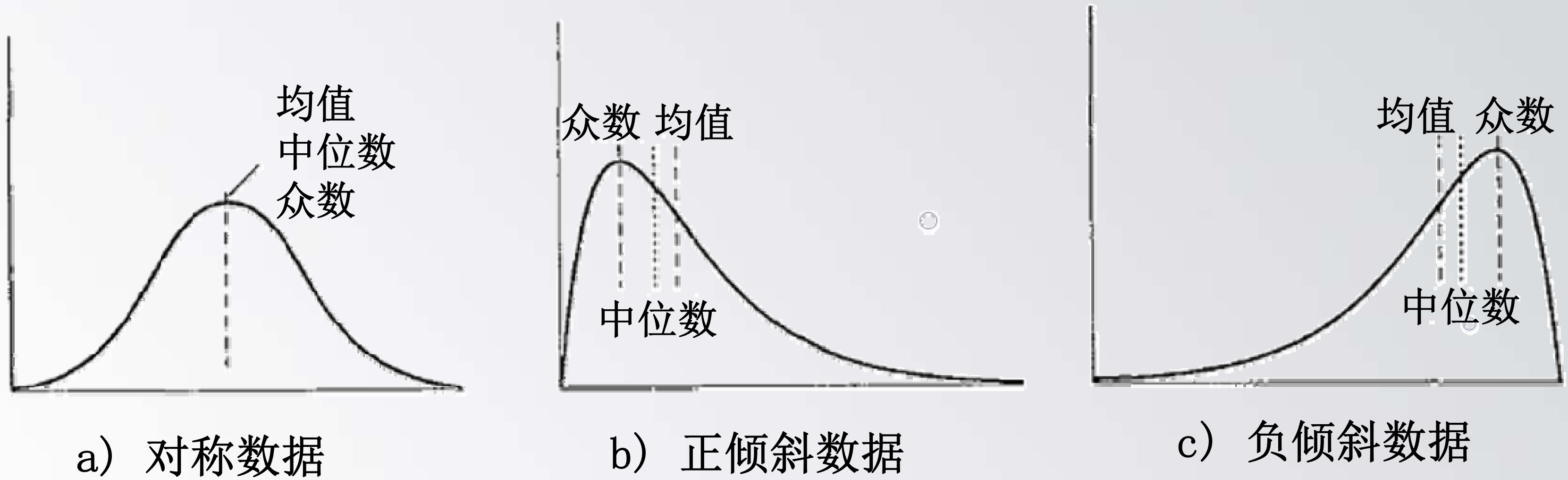
- **中列数**(midrange)是数据集中最大值和最小值的平均值。可以用来评估数值型数据的中心性趋势。
- 举例:

  假设我们有salary的如下值（以千美元为单位），按递增次序显示：30,31,47,50,52,52,56,60,63,70,70,110

  求其中中列数
- 中列数是：30+110/2=70.

数据的对称和偏斜

- 在对称的单峰频率曲线数据分布中，平均数，中值和众数都在同样的中点值上。
- 实际应用中，绝大部分都不是对称的。如果众数的值小于中值，称为正偏斜；如果众数的值大于中值，称为负偏斜



对称、正倾斜和负倾斜数据的中位数、均值和众数

数据分散性度量：方差和和标准差

- 方差和标准差是测量数据分散度的。
- N个观察$x_1,x_2,\ldots x_N$的方差：

$$\sigma^2 = \frac{1}{N}\sum_{\mathrm{i}=1}^{N}(x_i - \overline{x})^2 = (\frac{1}{N}\sum_{i=1}^{N}x_i^2) - \overline{x}^2$$

- 其中，$\overline{x}$ 是均值，σ 是标准差。
- 标准差测量的是数据偏离均值的发散程度，因此只有在均值接近数据中心的时候才考虑。
- 标准差为0只有在所有数据值都相等时才发生。
- 根据Chebyshev’s 不等式，至少 $(1-1/k^2)*100\%$的数据不会远离均值的K个标准差的范围。所以，标准差是一个很好的衡量数据分散度的指标。

数据分散性度量：方差和和标准差

例　在（30,31,47,50,52,52,56,60,63,70,70,110）中

我们计算均值得到 $\overline{\mathrm{x}}$=58 000美元。

为了确定该例子数据集的方差和标准差，我们置N=12，得到：

$$\sigma^2 = \frac{1}{12}(30^2 + 36^2 + 47^2 + ... + 110^2) - 58^2 \approx 379.17$$

$$\sigma \approx \sqrt{379.17} \approx 19.14$$

数据分散性度量：极差

- 令x_1, x_2, … x_N是某个数值属性X的一系列观察，数据集的极差表示的是最大值和最小值的差。

数据分散性度量：极差

- 偏度 （Skewness）

是统计数据分布偏斜方向和程度的度量

$$Skewness = S = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)s^3}$$

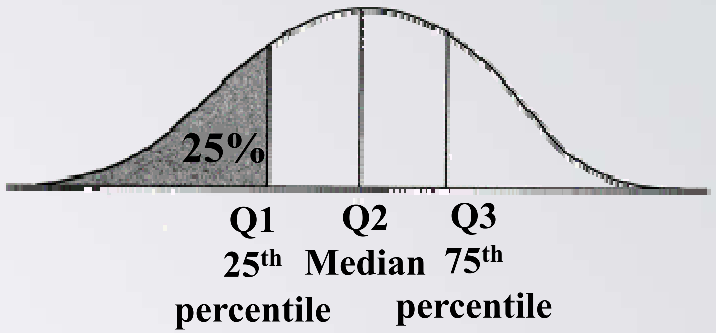
- 偏度 （Skewness）

描述总体中所有取值分布形态陡缓程度的统计量

$$Kurtosis = K = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{ns^4} - 3$$

数据分散性度量：分位数

- 假设数据按照属性X升序排列。我们可以挑选特定的数据点把数据分割成大小相等的连续数据集



- 分位数是数据分布上有一定间隔的数据点，将数据分成基本相等大小的连续数据集。
  - 2-分位点把数据划分为高低两半。即中位数。
  - 4-分位点（quartile）是把数据分布分成4个等量大小的3个数据点，每一个部分表示数据分布的1/4。
  - 100-分位数（percentile，百分位数）将数据集分成100个大小相等的连续集合。

数据分散性度量：分位数

- 给定第k个q分位点x, 至多k/q的数据值小于x，至多q-k/q的数据值大于x 。k是大于0小于q的整数。 共有q-1个q-分位点。
- 分位数反应了分布的中心，散布以及形状。
- 第1个四分位数，表示为Q1, 是第25个百分位点。它把数据值最低的25%切断。第3个四分位数，表示为Q3，是第75个百分位数。它切断了数据值低的75%。
- Q1和Q3的距离，简单反应了数据中心的一半数据的范围。这个距离被称为四分位数极差(IQR)。被定义为:

$$IQR = Q_3 - Q_1$$

数据分散性度量：分位数

例 假设我们有salary的如下值（以千美元为单位），按递增次序显示：30,31,47,50,52,52,56,60,63,70,70,110 让我们找出其中的四分位数。

四分位数是3个值，把排好的数据集划分成4个相等的部分。上述12个数据已经按递增序排序。这样，该数据集的四分位数分别是该有序表的第3、第6和第9个值。因此Q1=47 000美元，而Q3=63 000美元。

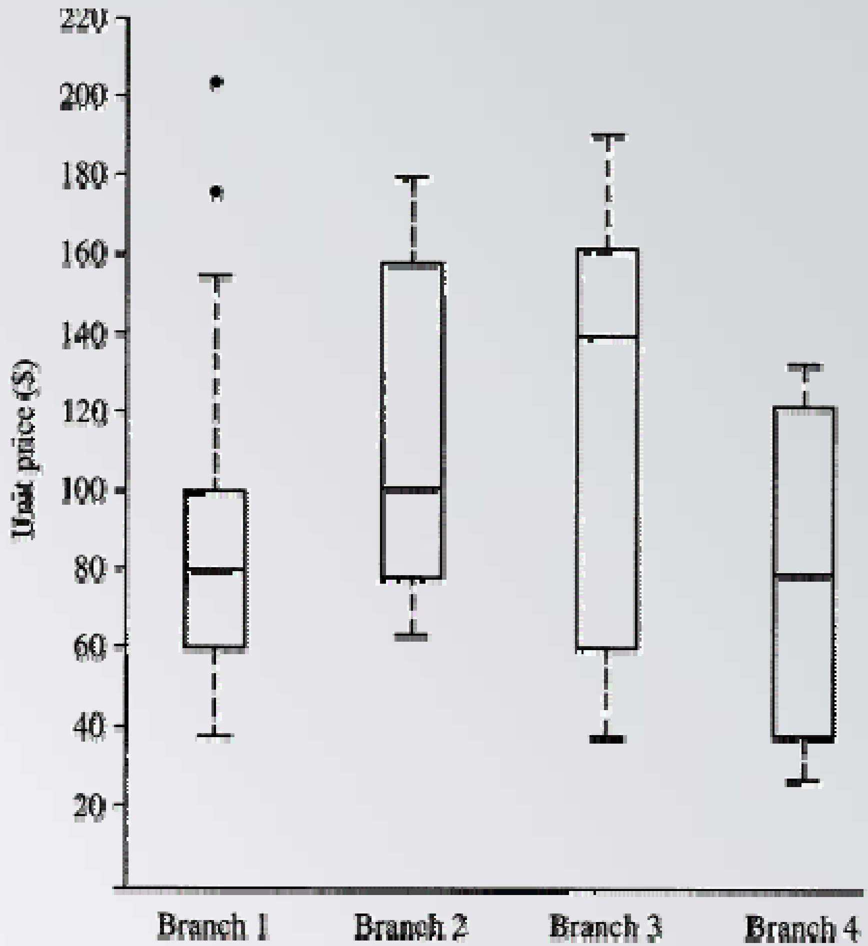
于是，四分位数极差为IQR=63 000-47 000=16 000美元。

数据分散性度量：五数概括

- 五数概括（Five-number summary）由中值，Q1， Q3，最小值和最大值组成，按次序表示为：Minimum, Q1, Median, Q3, Maximum.
- 注意去除离群点：第一分位数之下或第三分位数之上1.5*IQR的值

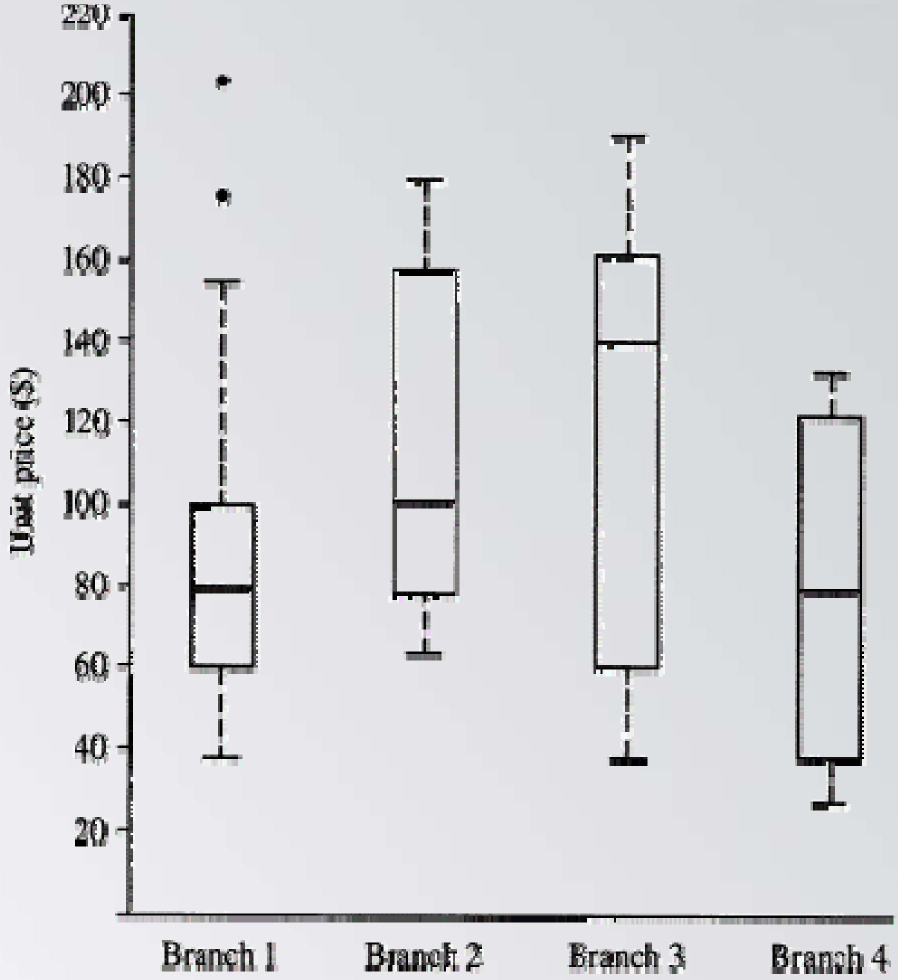
数据分散性度量：盒图

- **盒图**
  - 盒图体现了五数概括。
  - 盒子的端点在四分位数上，盒的长度是四分位数极差(IQR)
  - 中位数是箱子中间的线
  - 盒子外面的两根须是观察的最大值和最小值
  - 箱线图的计算时间复杂度是o(nlogn).



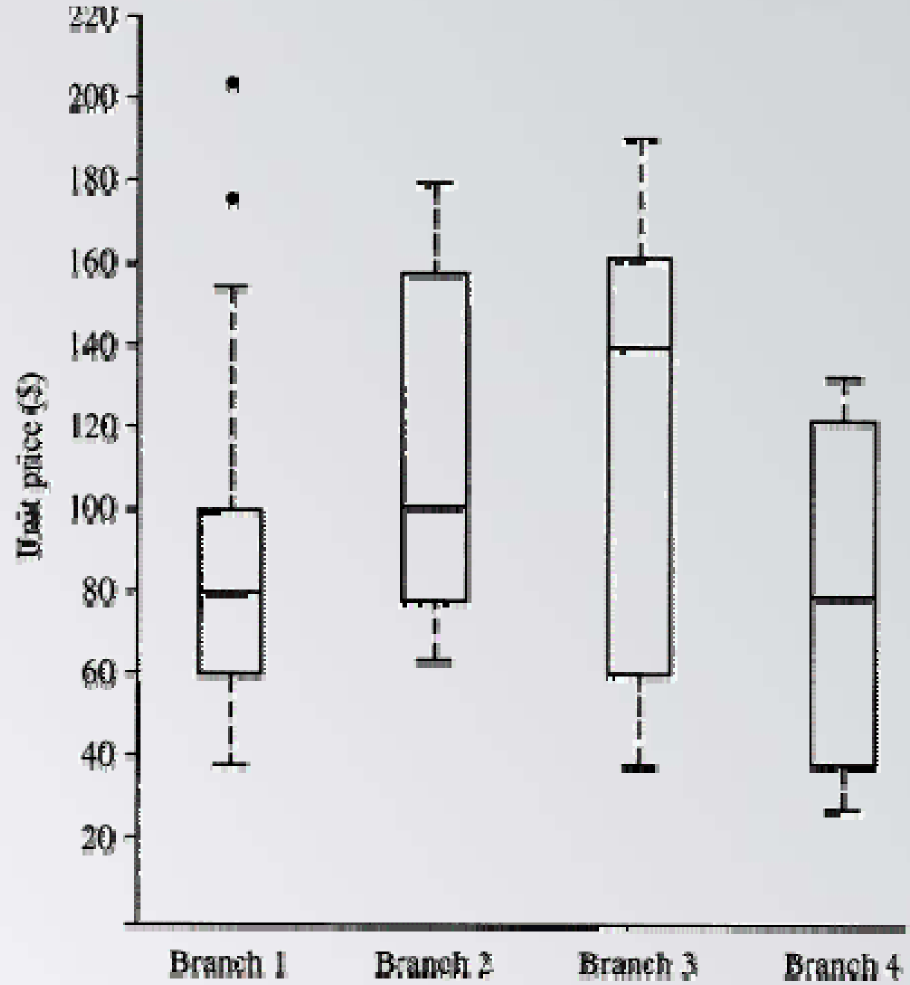
数据分散性度量：盒图

例 右图给出在给定的时间段ALLElectronics的4个部门销售的商品单价数据的盒图。对于branch1，我们看到销售商品单价的中位数是80美元，Q3是100美元。注意，该部门的两个边缘的观测值被个别地绘出，因为它们的值175和202都超过IQR的1.5倍，这里 IQR=40。



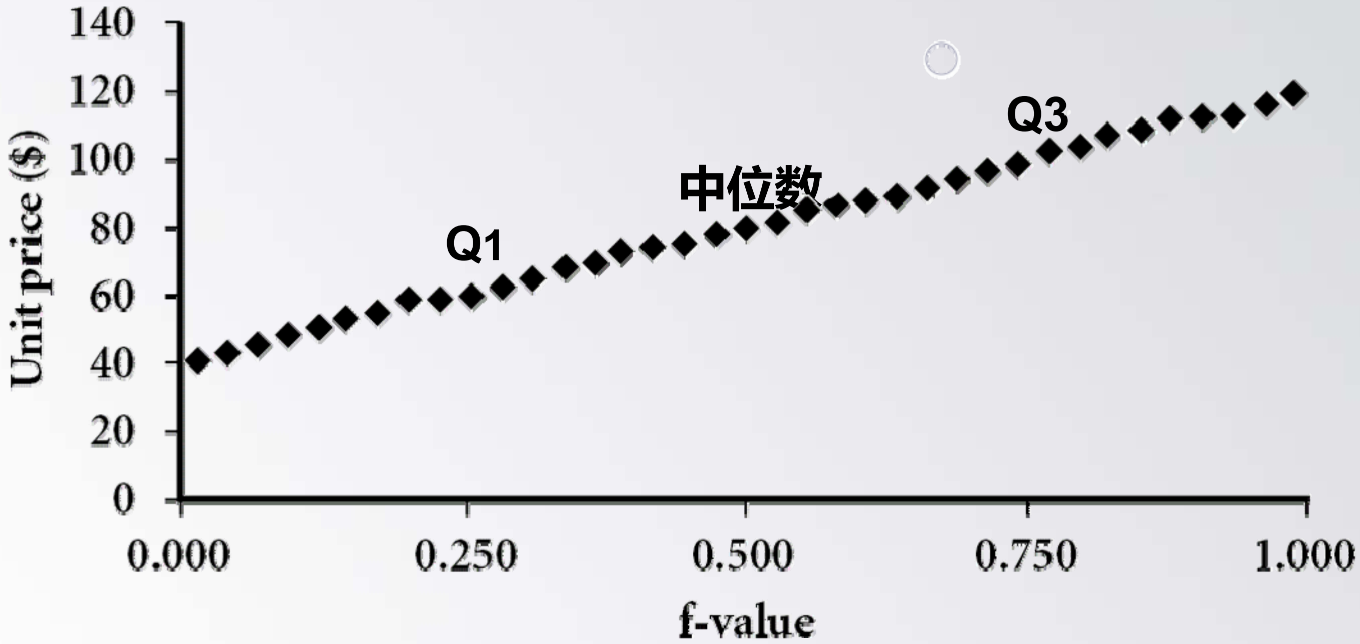
数据分散性度量：盒图

练习：
请分析右图所示的四个branch中哪个利润率最高？（假设每个branch 售出的商品数目相同，成本也相同）



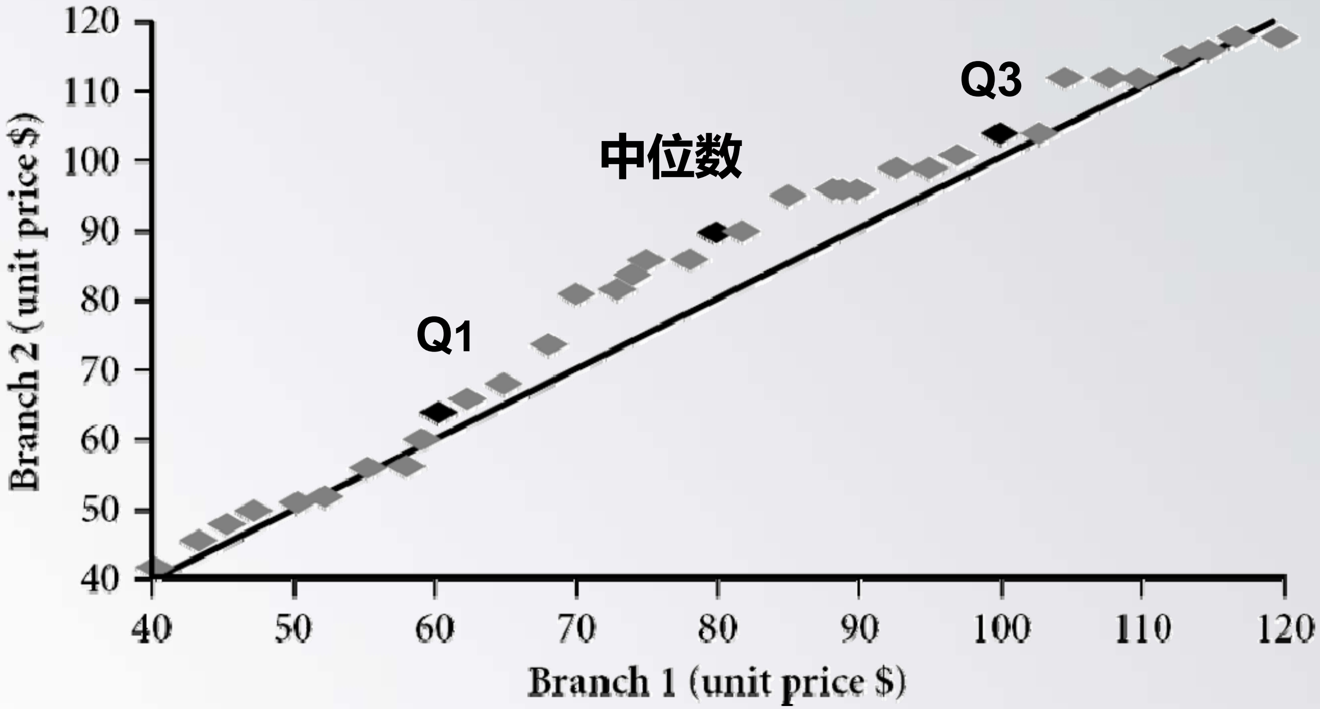
数据基本统计描述的图形显示

- **分位数图**（quantile plot)
  - 显示了给定属性的所有数据
  - 描述了分位数信息
  - 每个观测值x_i与一个百分数f_i配对，指出大约fi×100%的数据小于x_i



数据基本统计描述的图形显示

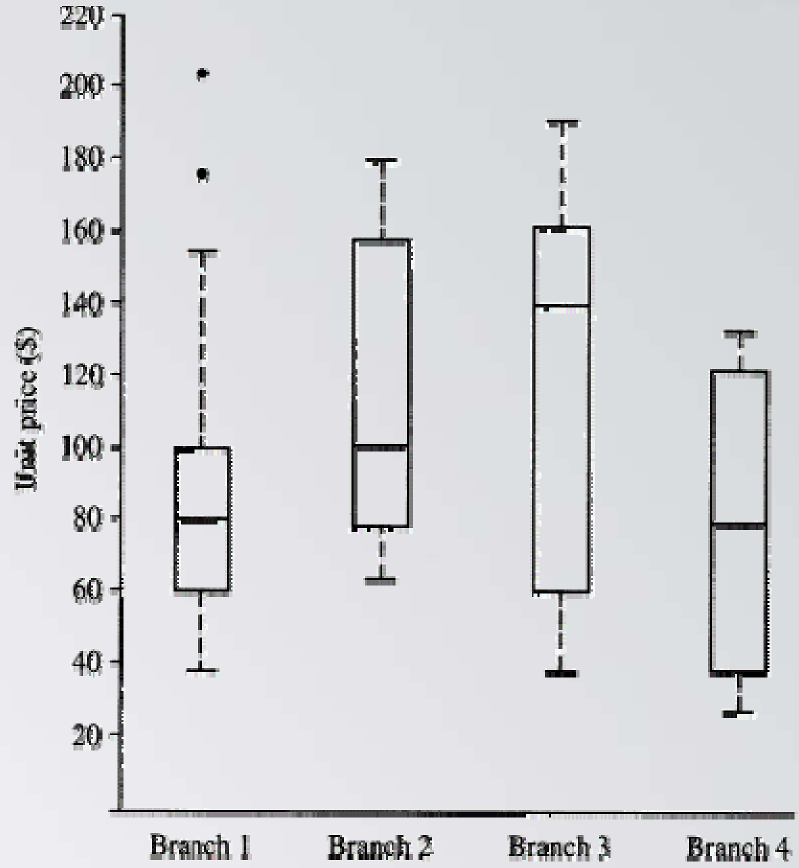
- **分位数-分位数图**（quantile-quantile plot，q-q图）
  - 截取相同长度的数据，根据对应数据作为横纵坐标画出
  - 直线表示两个数据相同的情况，黑点表示分位数点



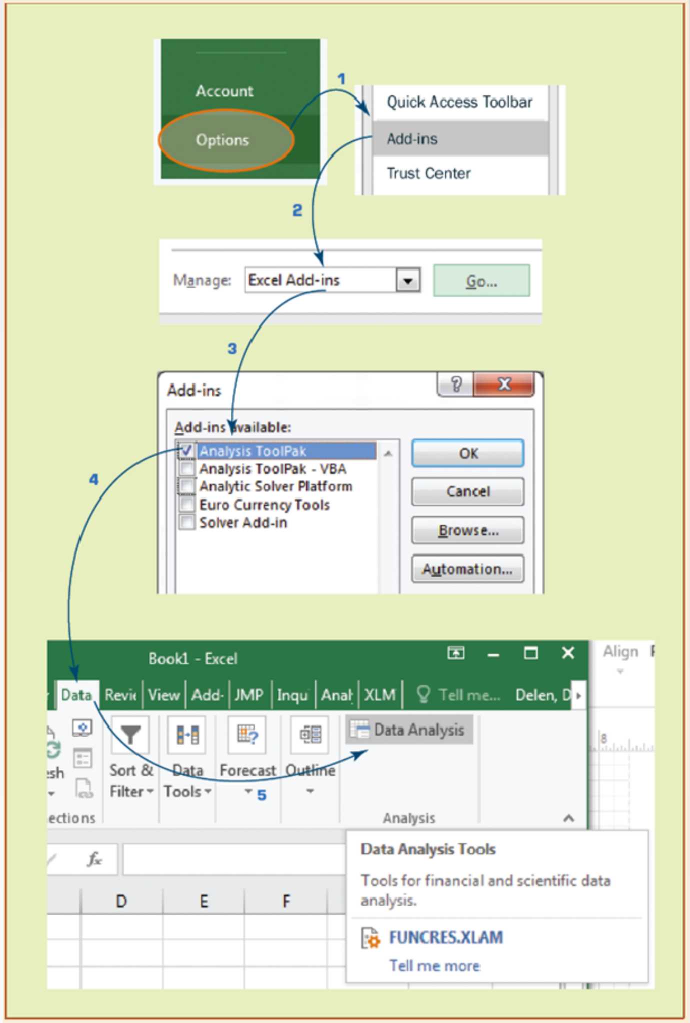
- 该图表示部门1销售的商品单价比部门2低

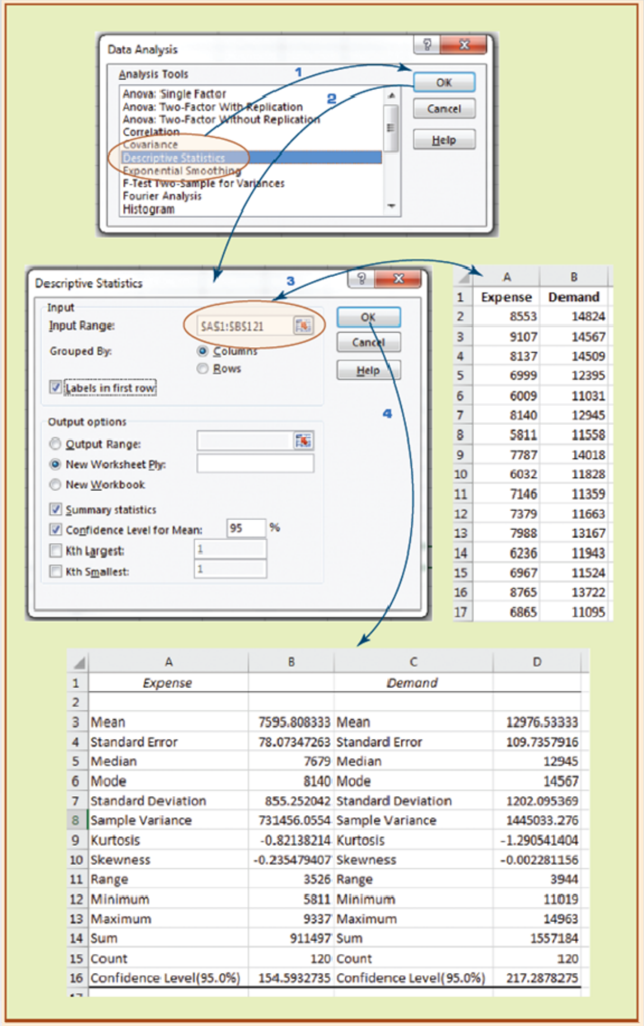
数据基本统计描述的图形显示

练习：
请请右边盒图转化为分位数图以及分位数-分位数图，并比较四个branch中哪个利润率最高？



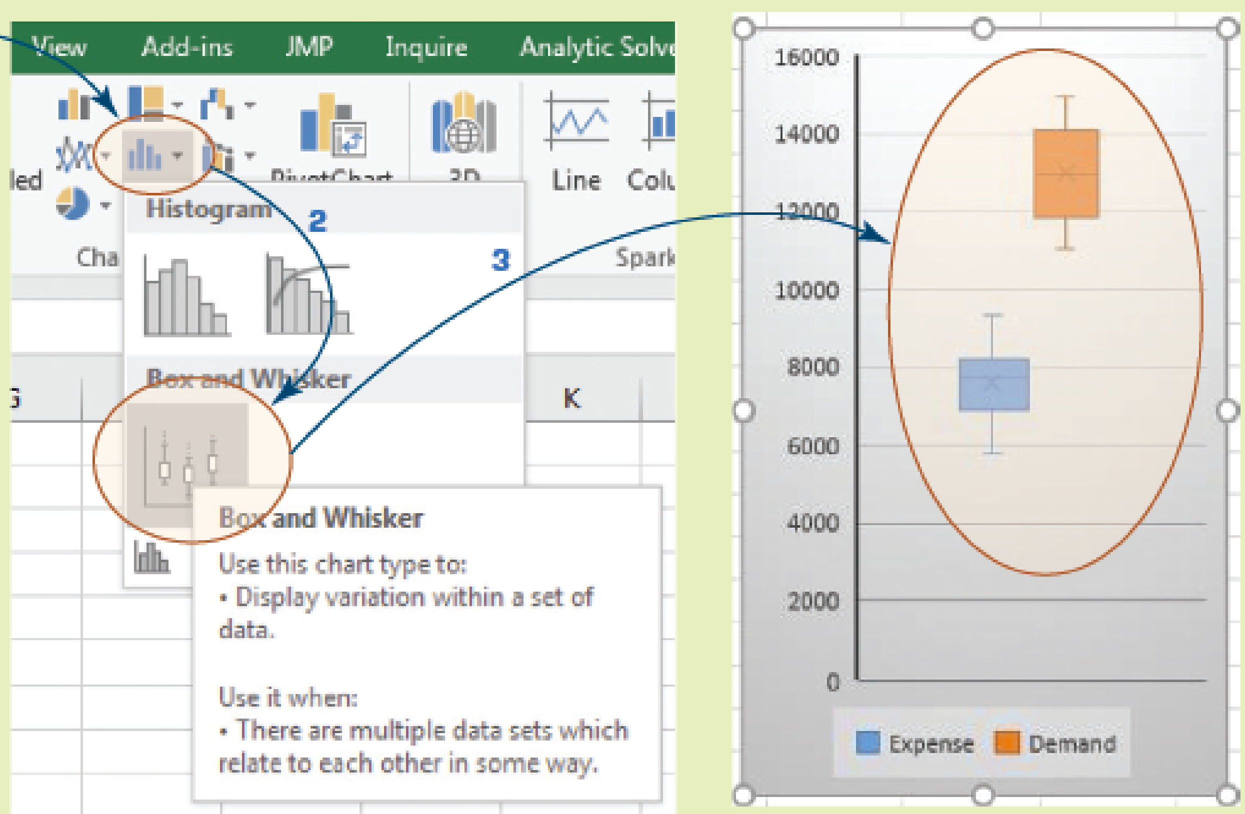
使用Excel进行描述性分析





https://jingyan.baidu.com/article/fedf073708eeb75ad89776e.html excel中的描述性数据统计

使用Excel进行描述性分析



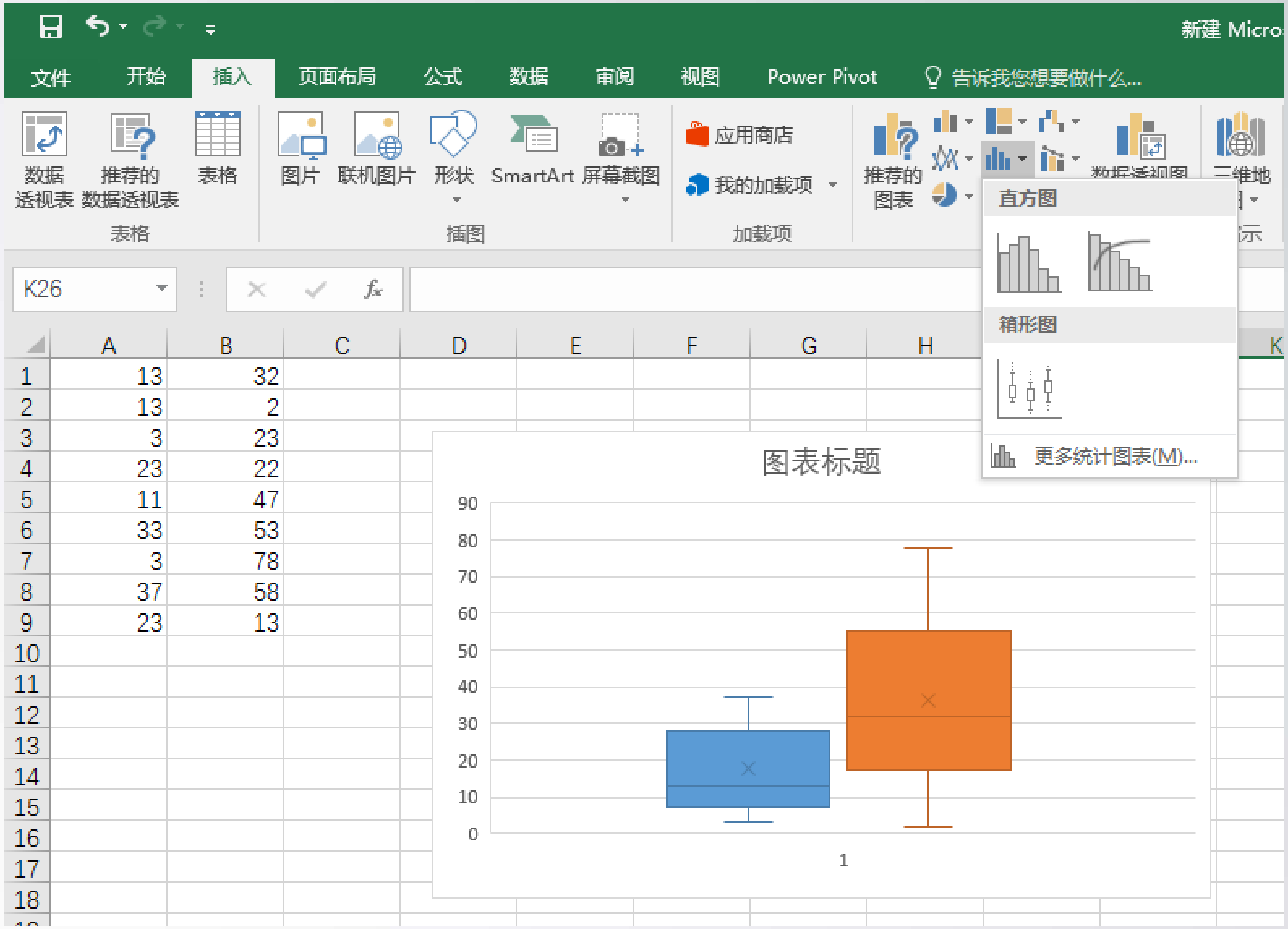
	A	B
1	Expense	Demand
2	8553	14824
3	9107	14567
4	8137	14509
5	6999	12395
6	6009	11031
7	8140	12945
8	5811	11558
9	7787	14018
10	6032	11828
11	7146	11359
12	7379	11663
13	7988	13167
14	6236	11943
15	6967	11524
16	8765	13722
17	6865	11095

https://jingyan.baidu.com/article/27fa7326be83d046f8271fc6.html excel如何画箱图

使用Excel进行描述性分析

列1		列2	
平均	363.2222	平均	15.33333
标准误差	356.2423	标准误差	8.449195
中位数	3	中位数	3
众数	3	众数	3
标准差	1068.727	标准差	25.34758
方差	1142177	方差	642.5
峰度	8.997088	峰度	5.576687
偏度	2.99935	偏度	2.362353
区域	3212	区域	76
最小值	1	最小值	2
最大值	3213	最大值	78
求和	3269	求和	138
观测数	9	观测数	9
置信度(95.0%)	821.4963	置信度(95.0%)	19.48388

使用Excel进行描述性分析



练习

- 练习使用EXCEL进行描述性分析
- 练习在对称、正偏斜、负偏斜情况下的偏度（自行准备数据）

Thank You!

Q&A