第四章
描述性统计指标

描述性统计指标

- 数据可视化：用表格和图形的方法初步展示了数据中的信息
- 描述性统计指标：数值方法展示数据的汇总信息
  - 单变量数据统计指标
  - 多变量相关性指标

数据中心位置：算术平均数

- （算术）平均值：$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
  - Excel 函数：AVERAGE(data range)
- 平均值的性质：$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- 平均值易受极端值影响
  - 练习：计算Purchase Orders数据集中cost per order的平均值
  - =AVERAGE(B2:B95)

数据中心位置：几何平均数

- 几何平均数：$\bar{x}_g = \sqrt[n]{x_1 x_2 \ldots x_n} = (x_1 x_2 \ldots x_n)^{1/n}$
- Excel 函数：=GEOMEAN(data range)
- 常常用于分析财务数据的增长率

例4.1 某基金的年回报率

年	回报率（%）	增长因子
1	-22.1	0.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	-37.0	0.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

$\bar{x}_g = 1.029$ $\bar{x} = 1.050$

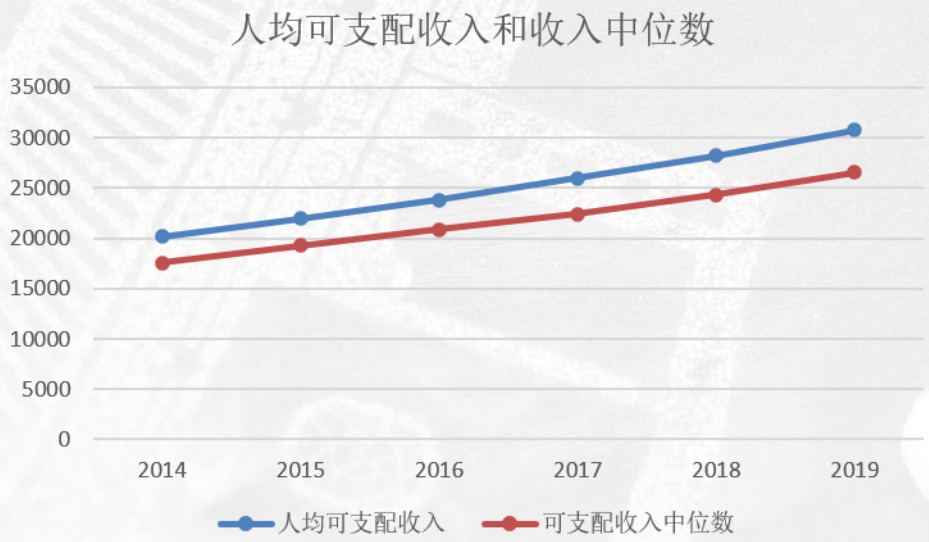
数据中心位置：中位数

- 中位数：数据从小到大排列，位于中间位置的值
  - 一半的数据比中位数小，一半数据比中位数大
  - 有奇数个观测值，中位数是中间位置的值
  - 有偶数个观测值，中位数是中间两个位置值的平均值
  - Excel 函数:=MEDIAN(data range)
- 中位数不易受到极端值的影响
  - 练习：Purchase Orders数据集中计算cost per order的中位数
  - =MEDIAN(B2:B94)

例4.2 居民可支配收入

居民人均可支配收入和收入中位数（元）

	2014	2015	2016	2017	2018	2019
均值	20167	21966	23821	25974	28228	30733
中位数	17570	19281	20883	22408	24336	26523



- 有效推进共同富裕

百分位数和四分位数

- 百分位数：第p百分位数把数据分割为两个部分，大约有p%的观测值比第p百分位数小；而大约有（100-p)%的观测值比第p百分位数大。
  - 将数据从小到大排序，第p百分数位置
  - $L_p = \frac{p}{100}(n+1)$
  - Excel函数：=PERCENTILE.EXC(data range, p%)
- 四分位数
  - Q_1--第一四分位数，第25百分位数
  - Q_2--第二四分位数，第50百分位数（中位数）
  - Q_3--第三四分位数，第75百分位数

例 4.3：电脑修理时长

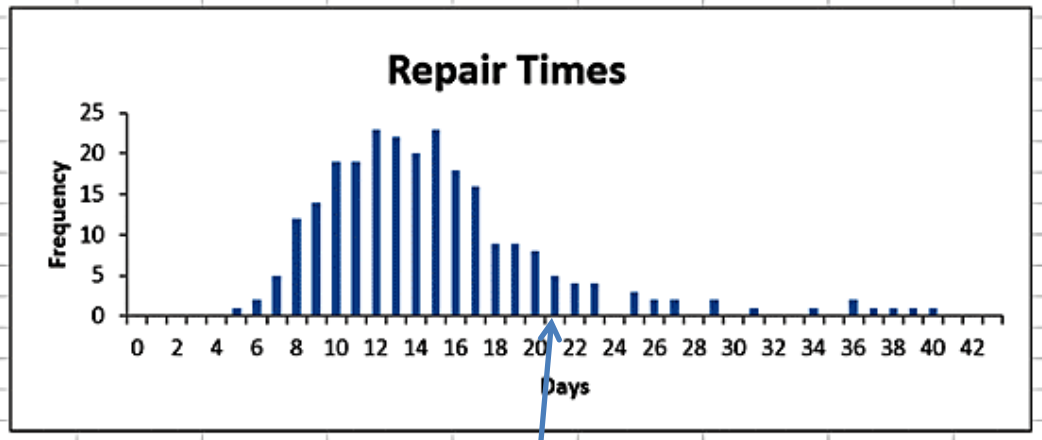
Computer Repair Times数据集 记录250个客户的电脑修理时间

- 当顾客询问大约需要等多久才能取到电脑时，如何回答？
- 中位数修理时间2周；均值大约15天.

	A	B
1	Computer Repair Times	
2		
3	Sample	Repair Time (Days)
4	1	18
5	2	15
6	3	17
250	247	31
251	248	6
252	249	17
253	250	13
254		
255	Mean	14.912
256	Median	14
257	Mode	15

例4.3 电脑修理时长

	A	B	C	D
1	Computer Repair Times			
2				
3			Relative	Cumulative
4	Days	Frequency	Frequency	Percentage
5	0	0	0.000	0.0%
6	1	0	0.000	0.0%
7	2	0	0.000	0.0%
8	3	0	0.000	0.0%
9	4	0	0.000	0.0%
10	5	1	0.004	0.4%
11	6	2	0.008	1.2%
12	7	5	0.020	3.2%
13	8	12	0.048	8.0%
14	9	14	0.056	13.6%
15	10	19	0.076	21.2%
16	11	19	0.076	28.8%
17	12	23	0.092	38.0%
18	13	22	0.088	46.8%



90% 在三周之内修理完

分类型数据位置信息

- 比例，记作p，是某一分类型数据（次品、错误）所占比例.
  - 练习：Purchase Orders数据集中计算供应商为Spacetime Technologies的订单比例
  - 数据透视表

	A	B	C	D	E	F	G	H	I	J
1	**Purchase Orders**									
2										
3	**Supplier**	**Order No.**	**Item No.**	**Item Description**	**Item Cost**	**Quantity**	**Cost per order**	**A/P Terms (Months)**	**Order Date**	**Arrival Date**
4	Spacetime Technologies	A0111	6489	O-Ring	$ 3.00	900	$ 2,700.00	25	10/10/11	10/18/11
5	Steelpin Inc.	A0115	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
6	Steelpin Inc.	A0123	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
7	Steelpin Inc.	A0204	5319	Shielded Cable/ft.	$ 1.10	16,500	$ 18,150.00	30	09/15/11	10/05/11
8	Steelpin Inc.	A0205	5677	Side Panel	$ 195.00	120	$ 23,400.00	30	11/02/11	11/13/11
9	Steelpin Inc.	A0207	4312	Bolt-nut package	$ 3.75	4,200	$ 15,750.00	30	09/01/11	09/10/11
10	Alum Sheeting	A0223	4224	Bolt-nut package	$ 3.95	4,500	$ 17,775.00	30	10/15/11	10/20/11

数据中心位置：众数

- 众数是发生频率最高的观测值
- 众数适用于这样的数据集：包含的不同数值较少
- 通过统计频数分布可以快速找出众数
- Excel 函数：
  - 单个众数：=MODE.SNGL(data range)
  - 多个众数:选中多个单元格（同一列),=MODE.MULT(data range),ctrl+shift+enter
  - 练习：Purchase orders数据集中计算A/P Terms众数

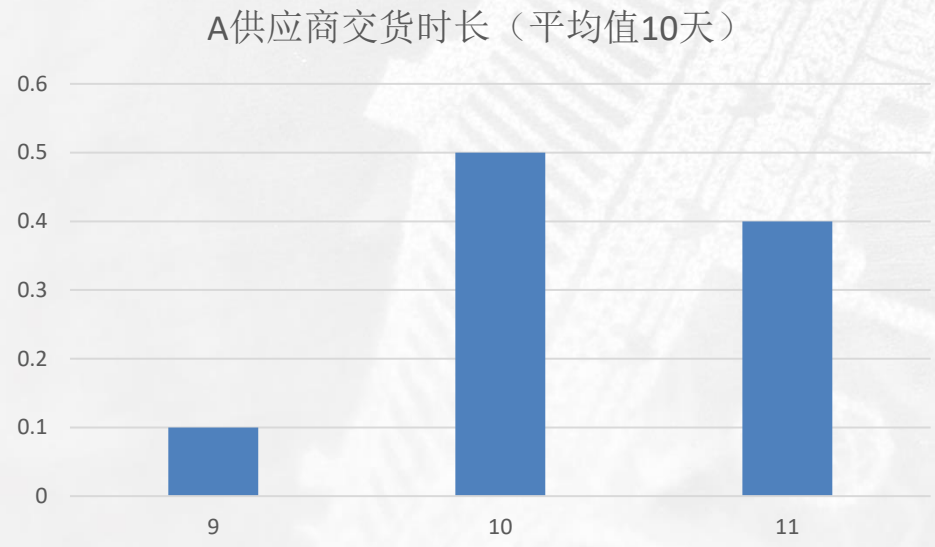
测量量表

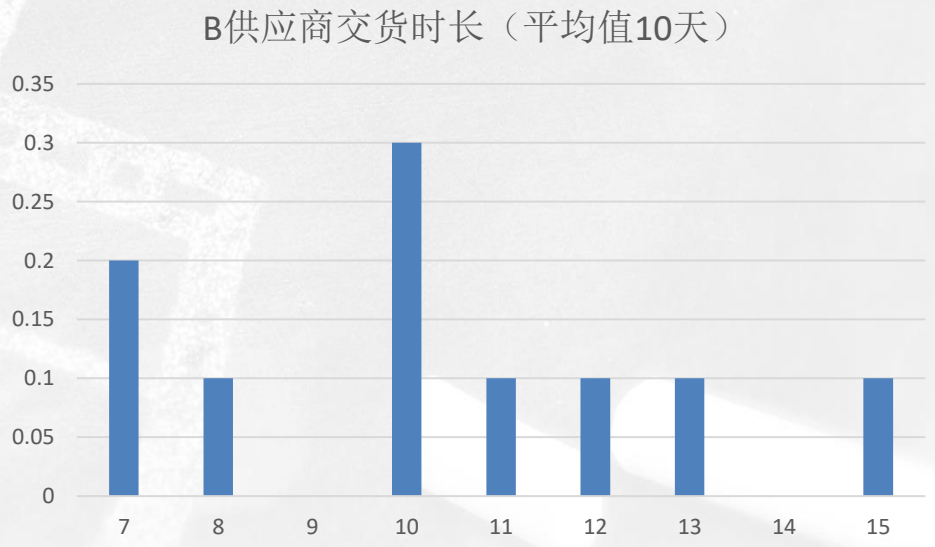
量表	基本比较	例子	平均测量
定类	同一性	男-女、使用-不使用	众数
定序	有序性	品牌偏好、质量等级	中位数
定距	定距比较	对品牌态度	均值
定比	绝对数量比较	单位销售量、购买数量	几何平均数

- 低级别的量表的平均测量指标也可以用于高级别量表

数据变异程度

- 除了位置的度量以外，人们往往还需要考虑变异程度，即数据离散程度。





极差

- 极差（range）：变量的最大观测值与最小观测值之差
- Excel函数：=MAX(*data range*) – MIN(*data range*).
- 极差易受极端值影响
  - 练习：Purchase Orders数据集中计算cost per oerder变量的极差

四分位差

- 四分位差(interquartile range, IQR):第三四分位数Q_3-第一四分位数Q_1
- 克服异常值的影响
  - 练习：Purchase Orders数据集中计算cost per oerder变量的四分位差

方差

- 极差和四分位差：只利用了一部分数据的信息
- 方差：数据离开平均值的距离（离差）平方的平均
- 总体方差：$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$
  - Excel中：=VAR.P(data range)
- 样本方差：$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
  - Excel中：=VAR.S(data range)
  - 注意分母的差异
  - 练习：计算purchase orders数据集中变量cost per order的方差，注意方差的单位

标准差

- 标准差(standard deviation)：方差的平方根，单位与数据单位一致，因此应用更广泛
- 总体标准差：$\sigma=\sqrt{\sigma^2}$
  - Excel中：=STDEV.P(data range)
- 样本标准差：$s=\sqrt{s^2}$
  - Excel中：=STDEV.S(data range)
  - 练习：计算purchase orders数据集中变量cost per order的标准差
- 财务分析中用标准差度量风险

标准差系数

- 标准差系数（coefficient of variation, CV）$=\frac{\text{标准差}}{\text{平均值}}*100\%$
- 适用于比较具有不同标准差和不同平均数的变量的变异程度
- 较好地衡量了相对投资回报的投资风险
  - 财务风险分析中常用1/CV衡量单位风险的投资回报（越高越好）
  - 股票的夏普值

例4.5 股票的投资风险

- *Closing Stock Prices*数据集
  - Intel (INTC)投资风险较高
  - 指数基金投资风险最低

	A	B	C	D	E	F
1	**Closing Stock Prices**					
2						
3	**Date**	**IBM**	**INTC**	**CSCO**	**GE**	**DJ Industrials Index**
4	9/3/2010	$127.58	$18.43	$21.04	$15.39	10447.93
5	9/7/2010	$125.95	$18.12	$20.58	$15.44	10340.69
6	9/8/2010	$126.08	$17.90	$20.64	$15.70	10387.01
22	9/30/2010	$134.14	$19.20	$21.90	$16.25	10788.05
23	10/1/2010	$135.64	$19.32	$21.91	$16.36	10829.68
24	**Mean**	$130.93	$18.81	$21.50	$16.20	$10,639.98
25	**Standard Deviation**	$3.22	$0.50	$0.52	$0.35	$171.94
26	**Coefficient of Variation**	0.025	0.027	0.024	0.022	0.016

数据标准化

- 标准化之后的值，通过记为z值，提供了观测值离开均值几倍标准差的信息.
- 第i个观测值的z值为$z_i = \frac{x_i - \bar{x}}{s}$
  - z值=1代表观测值在均值右边一倍标准差位置
  - z值=-1.5代表观测值在均值左边1.5倍标准差位置
  - Excel 函数 :=STANDARDIZE(x,mean,standard_dev)

例4.6 计算z值

- *Purchase Orders*数据集中将变量Cost per order进行标准化

	A	B	C
1	Observation	Cost per order	z-score
2	x1	$2,700.00	-0.79
3	x2	$19,250.00	-0.24
4	x3	$15,937.50	-0.35
5	x4	$18,150.00	-0.27
6	x5	$23,400.00	-0.10
91	x90	$6,750.00	-0.65
92	x91	$16,625.00	-0.32
93	x92	$74,375.00	1.61
94	x93	$72,250.00	1.54
95	x94	$6,562.50	-0.66
96			
97	Mean	$26,295.32	
98	Standard Deviation	$29,842.83	

=(B2 - B97)/B98, or
=STANDARDIZE(B2,B97,B98).

车贝晓夫定理

- 车贝晓夫定理：与平均数的距离在k（$k > 1$）标准差之内的数据占了至少$1 - 1/k^2$
  - k = 2：至少3/4的数据在均值附近2倍标准差之内
  - k = 3：至少8/9 的数据在均值附近3倍标准差之内
    - 练习：商务分析班上的同学期末考试平均成绩为70分，标准差为5分。请问有多少学生的考试成绩在60-80分？有多少同学的考试成绩在58-82分？
    - 练习：在任一数据集中验证车贝晓夫定理。
      - 数据“筛选”功能
      - COUNT函数应用
      - 数据透视表

经验法则

- 现实中很多数据的直方图呈现钟形
- 具有钟形分布的数据
  - 大约 68% 的观测值在均值附近一倍标准差内
  - 大约 95% 的观测值在均值附近两倍标准差内
  - 大约 99.7% 的观测值在均值附近三倍标准差内
- 注意，车贝晓夫“定理”适用于任一分布形状的数据

异常值检测

- 异常大或异常小的观测值称为异常值（outlier）
- 检查异常值的反常原因
  - 利用z值检测异常值：z值小于-3或大于3
  - 利用四位分数检测异常值：
    - 下限=$Q_1 - 1.5IQR$
    - 上限=$Q_3 + 1.5IQR$
  - 注意：异常值不一定是错误值；两种方法的上下限可以不同，可以用任一种或两种方法
    - 练习：在purchase orders数据集中用两种方法检测cost per order的异常值

描述性统计工具

- Purchase Orders 数据集

- 数据-数据分析-描述性统计
- 文件-选项-加载项-分析工具库-转到-确定

	A	B	C	D
1	*Cost per order*		*A/P Terms (Months)*	
2				
3	Mean	26295.31915	Mean	30.63829787
4	Standard Error	3078.053014	Standard Error	0.702294026
5	Median	15656.25	Median	30
6	Mode	14910	Mode	30
7	Standard Deviation	29842.8312	Standard Deviation	6.808993205
8	Sample Variance	890594573.8	Sample Variance	46.36238847
9	Kurtosis	2.079637302	Kurtosis	1.512188562
10	Skewness	1.664271519	Skewness	0.599265003
11	Range	127431.25	Range	30
12	Minimum	68.75	Minimum	15
13	Maximum	127500	Maximum	45
14	Sum	2471760	Sum	2880
15	Count	94	Count	94

描述性统计工具

- 标准误差
- 偏度
- 峰度

- 箱形图：中位数、第一四分位数、第三四分位数、上限、下限、异常值

变量相关性统计指标

- 协方差和相关系数
  - 样本协方差：$s_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$
    - Excel函数：=COVARIANCE.S(array1,array2)
  - 相关系数（不受数据量纲影响）：$r_{xy}=\frac{s_{xy}}{s_x s_y}$
    - Excel函数：=CORREL(array1,array2)
    - 相关系数的取值在-1到1之间。如何证明？
    - 正负性与协方差一致

协方差和相关系数计算

Colleges and Universities 数据集：数据-数据分析
相关系数

	A	B	C	D	E	F
1		*Median SAT*	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*	*Graduation %*
2	Median SAT	1				
3	Acceptance Rate	-0.601901959	1			
4	Expenditures/Student	0.572741729	-0.284254415	1		
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
6	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1

注意：输入的数据必须位于相邻列

相关系数解释

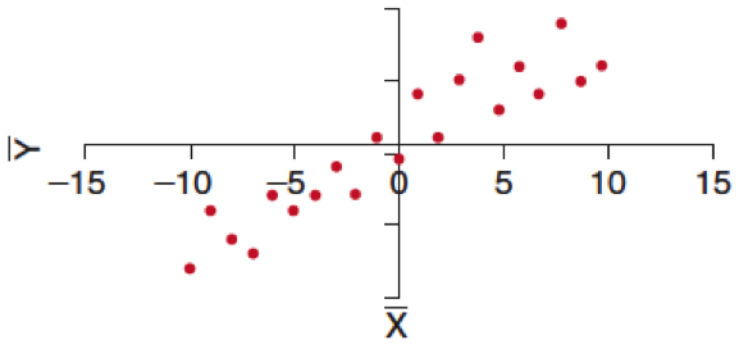
x_i	y_i
5	10
10	30
15	50

$$r_{xy} = 1$$

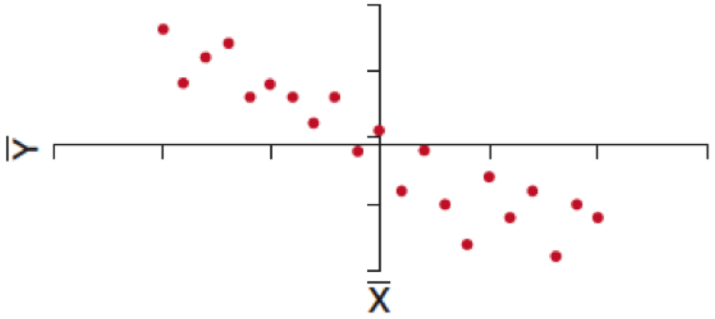
x_i	y_i
5	-10
10	-30
15	-50

$$r_{xy} = -1$$

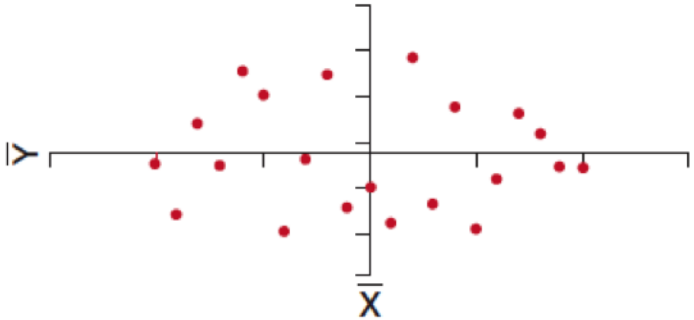
散点图与相关系数



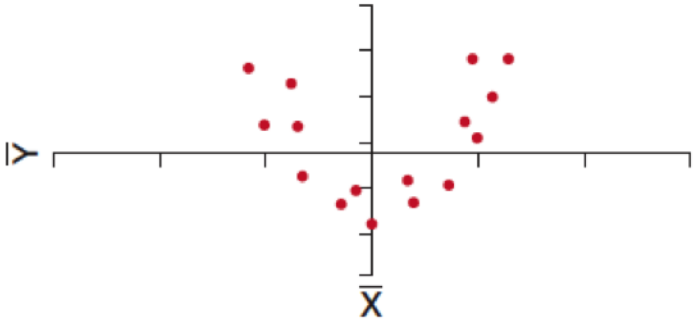
(a) Positive Correlation



(b) Negative Correlation



(c) No Correlation



(d) A Nonlinear Relationship with No Linear Correlation

相关系数展现了两个变量的线性相关程度，结合散点图更好地了解两个变量的相关性

数据挖掘*

- 数据挖掘：用统计方法、量化工具等从数据集中发现一些隐藏的模式的过程。
- 较大数据集中自动发现模式

数据挖掘*

- 聚类分析：
  - 把成千上万条记录（例如客户购买记录）根据相近程度归置成少数几类的过程，其中每一类里面的记录近似性高，不同类之间的记录近似性低。
  - 定义记录之间的距离、聚类算法（划分聚类、层次聚类……）

数据挖掘*

- 关联分析（购物篮分析）
  - 关联分析试图在一个大数据集中揭示哪些变量之间可能存在着有趣的关联。
  - 关联分析的结果是形成一些“由于某些事件的发生而引起另外一些事件的发生”之类的规则，即关联规则。假定存在规则X→Y，
    - 支持度：$s(X \to Y) = \frac{\sigma(X \cap Y)}{N}$
    - 置信度：$c(X \to Y) = \frac{\sigma(X \cap Y)}{\sigma(X)}$

数据挖掘*

- 关联分析
  - 假设超市拥有一个100000条客户购买记录的数据集，每条记录记录了客户购买A、B、C三种产品的情况。其中有2000条记录都购买了A和B产品，而在这2000条记录中，有800条记录同时购买了C产品。那么对“如果客户购买了A和B，那么他也会购买C”这条关联规则来说，
  - 支持度为？置信度为？