期末考试复习

一、商分简介:

1、 商务分析是对以下要素的应用:

数据　信息技术　统计分析　量化方法　计算机模型

2、 商务分析的目的:

帮助经理人对商业运作有更好的理解，可以创造出更好的基于事实的决策。

3、 商务分析思维:

数据思维、业务思维（商业思维）

4、 商务分析范围:

描述性分析 利用数据理解过去和现在企业的表现来提供决策借鉴

预测性分析 对历史数据进行分析，发现数据中的存在的模式与规律，并投射至未来，从而完成对未来的预测。

决策性分析（规范性分析）指定目标函数，并基于此制定最优决策。

eg 大部分零售店在季末都进行降价清库存活动

• 核心问题: 何时开始降价，降价幅度多高可
以最大化收益?

• 商务分析的潜在应用:

– 描述性分析: 检查类似产品（价格、销售量、广告）的历史数据

– 预测性分析: 预测价格变动一定量的时候的销量变化

– 决策性分析:找到最优的降价幅度和广告策略以最大化销售收入或利润

用商务分析解决问题的思路步骤:

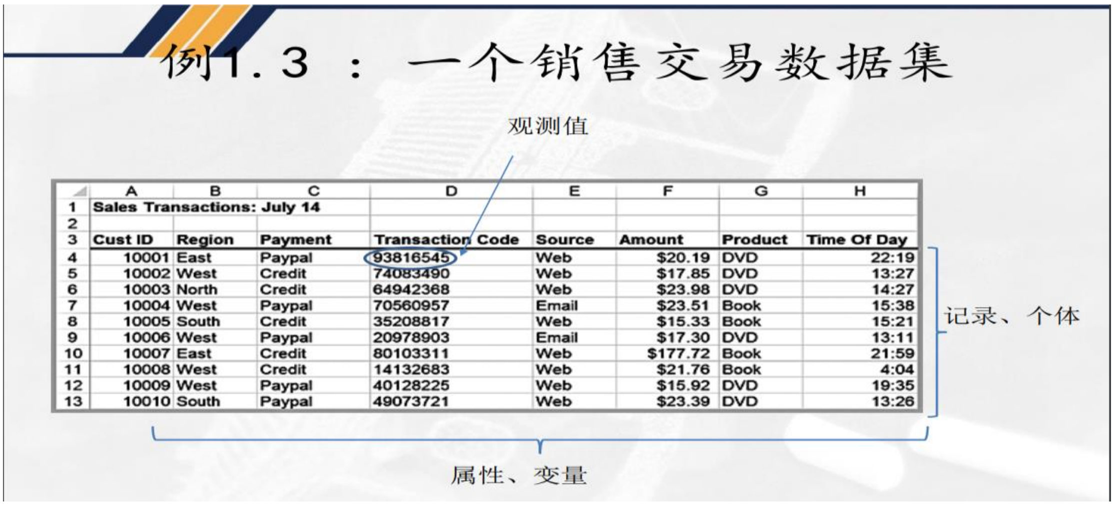
认识问题（产品成本过高）-定义问题（选址不合理）-结构化问题（目标是要最小化成本）-分析问题（利用实验等找到答案）-对结果进行解读并做政策（了解假设、理想）-实施决策

二、数据

数据: 通过测量过程得到的数字、文字、图表结果

信息: 分析数据的结果，从数据中抽取出的有意义的部分

数据集: 数据的集合



	A	B	C	D	E	F	G	H
1	Sales Transactions: July 14							
2								
3	Cust ID	Region	Payment	Transaction Code	Source	Amount	Product	Time Of Day
4	10001	East	Paypal	93816545	Web	$20.19	DVD	22:19
5	10002	West	Credit	74083490	Web	$17.85	DVD	13:27
6	10003	North	Credit	64942368	Web	$23.98	DVD	14:27
7	10004	West	Paypal	70560957	Email	$23.51	Book	15:38
8	10005	South	Credit	35208817	Web	$15.33	Book	15:21
9	10006	West	Paypal	20978903	Email	$17.30	DVD	13:11
10	10007	East	Credit	80103311	Web	$177.72	Book	21:59
11	10008	West	Credit	14132683	Web	$21.76	Book	4:04
12	10009	West	Paypal	40128225	Web	$15.92	DVD	19:35
13	10010	South	Paypal	49073721	Web	$23.39	DVD	13:26

也就是说：一行一行的数据即为一条记录，一列一列则为变量，每一个小方格即为观测值

大数据：传统数据处理应用软件不足以处理的大或者复杂的数据集，体现在：
数据量大——记录和变量特别多的数据集
数据种类多
数据价值密度低
数据产生和处理速度快

数据测量尺度
名义尺度 -数据只展示类别信息：商品名称、编号
顺序尺度 –数据展示了顺序等级：
间隔尺度- 数值间的距离按某一固定度量单位显示，可比较（最常见的类型：日期
比率尺度 –距离可比较，此外还有绝对零点的定距数据，数值之间的比率也有意义：价格、成本、重量、雇佣时长、周期

一手数据：自己调查的来的
包括了
探索性数据收集——形成最初的预见和洞察，例如销量下降的原因 方法：焦点小组、深度访谈
描述性数据收集——产生相关顾客群的特征的数据。如顾客花了多少钱等 方法：问卷调查
有关于问卷调查展开看看：
• 量表：测量工具
– 定类量表（收集的数据类型为名义数据） 下面饮料中你喜欢哪些？
– 定序量表（收集的数据类型为顺序数据） 请根据你的喜好程度给下列饮料排序
– 定距量表（收集的数据类型为间隔数据） 请你再下面的表中选出你对每种饮料的喜好程度 表中有 很喜欢，喜欢，一般等；请拖动鼠标表明你对可乐的喜好程度左边是很喜欢，右边是很讨厌
– 定比量表（收集的数据类型为比率数据）请把一百分分配给下面的饮料来表示对他们的喜好程度
Ps 分辨定比定距的根本点在于：定比才有“0”这个东西！
问卷调查优点：低成本、容易实施
缺点：难以获得无偏额回答，如何选择回答问卷的适合人群（可以结合焦点小组）
因果性数据收集——验证因果关系 方法：AB 测试（一部分人用 A，另一部分人用 B）
如果 AB 测试不可行时（单元太少，测试的项目不合伦理……）只能进行观察性因果研究
因果关系推断前提：相关性 时间顺序 没有共同的驱动因素。

小结上述一手数据：
• 探索性数据收集– 研究问题偏探索性（ambiguous problems）
• 产品的销量降低了，为什么？
• 描述性数据收集– 研究问题偏描述性（aware of problem）
• 什么样的顾客在买我们的产品？哪些人在买竞争对手的产品？
• 因果性数据收集– 研究问题的因果性较强（problem clearly defined）
• 如果我改变登录页面的设置，是否会有更多的人购买产品？

二手数据：从统计局或别人那里下载出来的

一手 or 二手：一手数据获取成本高，如果有二手数据可以使用时，优先使用二手数据

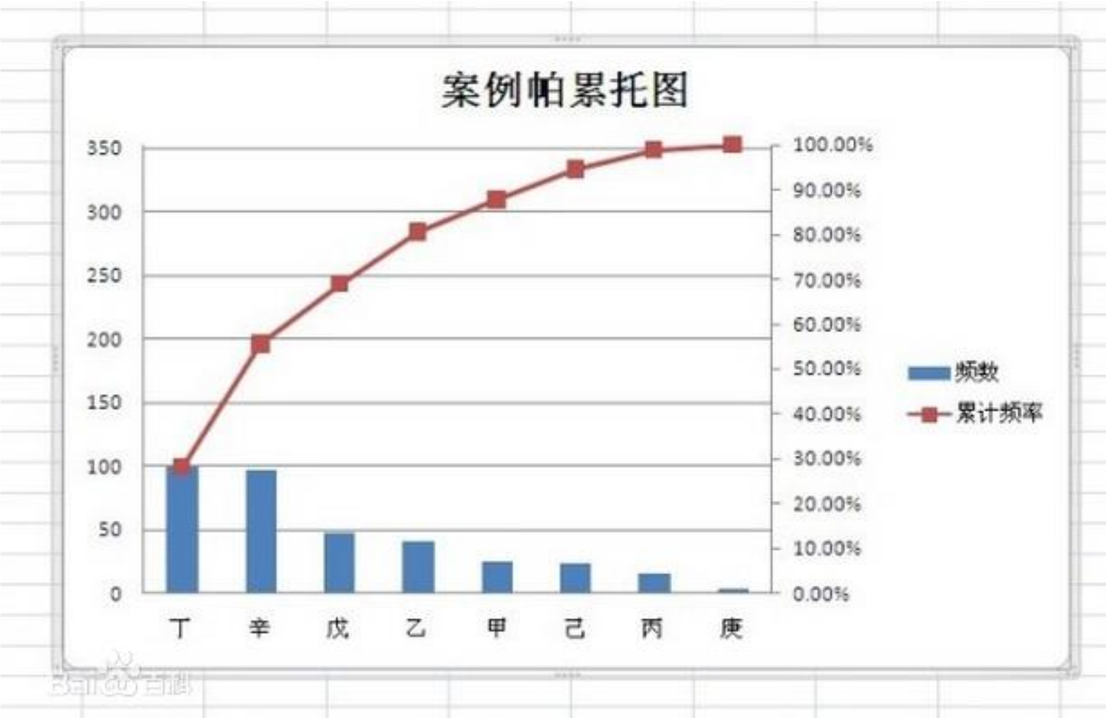
数据可视化：几种“图”

1、柱状图和条形图：比较不同类别的数量
• 柱状图：纵向；条形图：横向
– 簇状柱形图：比较不同类型的数量
– 堆积柱形图：比较不同类型的数量并查看其对总和的贡献
– 百分比堆积柱形图：比较不同类型的数量并显示其在总体中的百分比
2、饼图：显示比例，但不适用于类别太多的数据

数据透视表：对原始数据进行处理，得到汇总数据。直接随便点一个，然后点插入 插入数据透视表

频数分布图在数据分析中用直方图功能即可，注意接收区域的确定。要一开始给所有数据按顺序排列，找出最大最小值

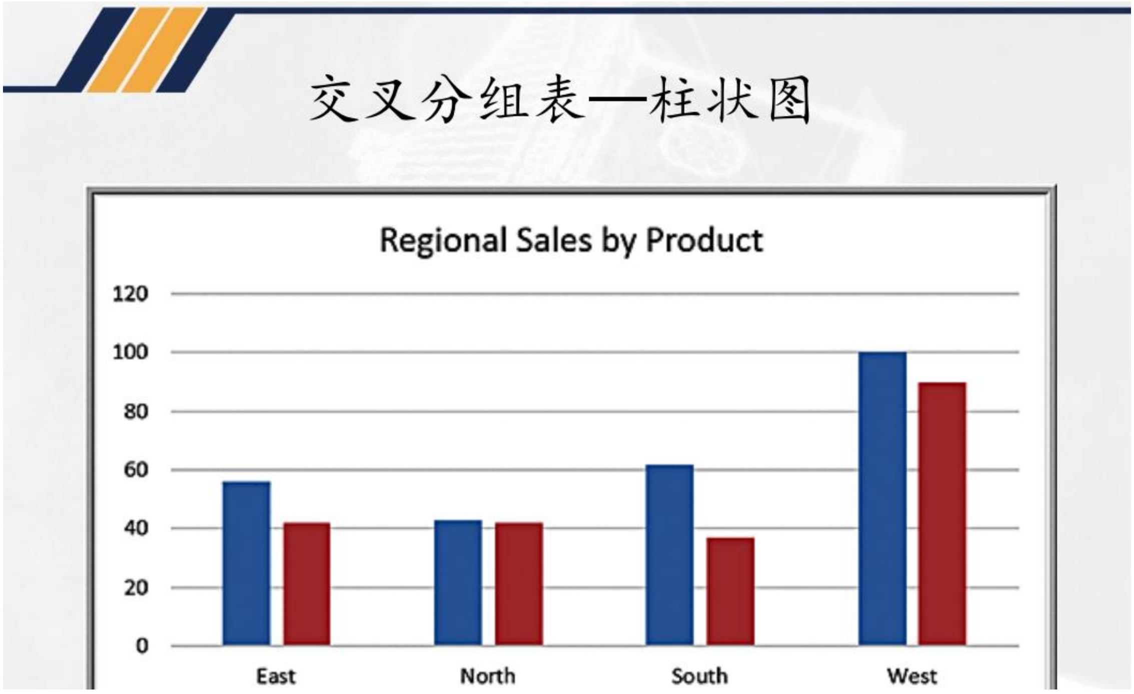
帕累托分析：二八定律



如：大多数财富都聚集在了少数人手上

而针对变量相关性的可视化：
折线图：展示了数据的趋势（多呈现时间尺度上的发展趋势）
散点图：直观展现了两个变量之间的关系，是分析两个变量关系的初步探索
交叉分组表：分析两个类别变量之间关系的有效工具

Region	Book	DVD	Total
East	57.1%	42.9%	100.0%
North	50.6%	49.4%	100.0%
South	62.6%	37.4%	100.0%
West	52.6%	47.4%	100.0%



应当指出的是，可以通过数据透视表做出交叉分组表
辛普森悖论告诉我们：根据汇总数据表和未汇总数据表得出的结论可能会完全不同。
因为数据汇总表在汇总过程中可能隐去了部分数据，如抽取出判罚的案件总数却忽略了判罚的法律类型

描述性统计指标：
厘清区别：数据可视化是用表格和图形的方法初步展示了数据中的信息。而描述性统计指标是指用数值方法展示数据的汇总信息，即用一些列运算得到什么平均数之类的东西
数据中心位置：
(算数)平均数，缺点，易受极端值影响。函数为 AVERAGE
几何平均数 Xg：即对 n 个数相乘开 n 此根（数据列表中不能有 0） 函数为：GEOMEAN
常常用分析财务数据的增长率。如基金回报率。
中位数：如果有奇数个观测值，中位数即为中间数，如果有偶数个观测值，中位数是中间两个位置值的平均值。 函数为 MEDIAN 中位数不容易受极端值影响。
众数（适用于包含不同数值较少的数据集）用统计频数分布可以找出众数 函数为：
=MODE.MULT(data range),ctrl+shift+enter

补充，百分位数 函数为 PERCENTILE.EXC（数据集，p%）
四分位数：第一四分位数=下四分位数=25%

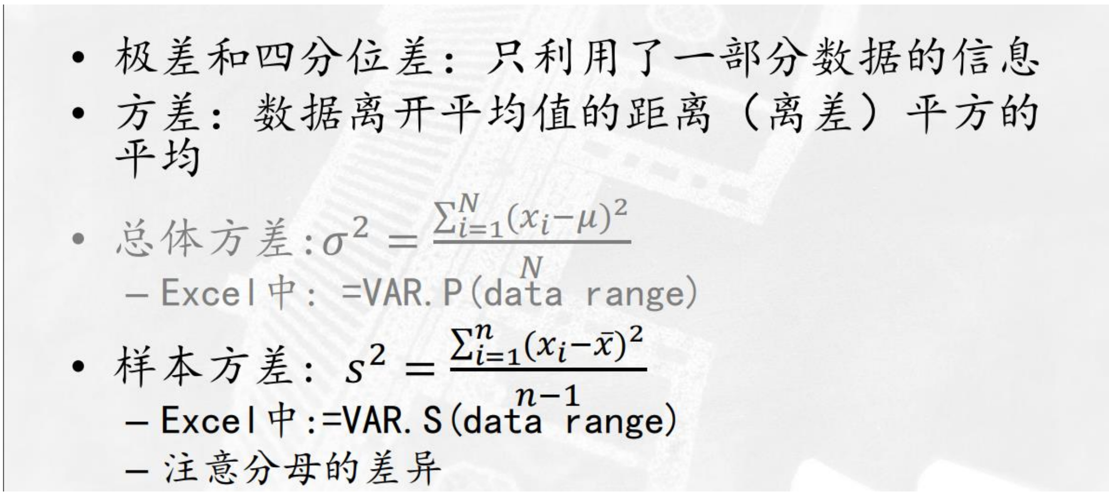

测量量表

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

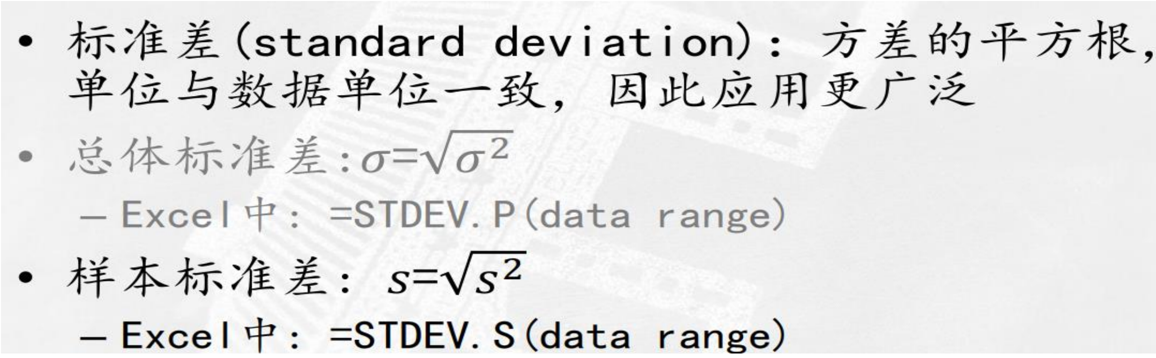
结合我们上面已经掌握的测量量表，不同的测量量表可以得到不同的描述性统计数据。

数据还可能存在偏离，故我们再引入新的
极差（range）函数为 MAX-MIN 极差易受极端值影响
四分位差（IQR）=Q3-Q1
方差：注意，我们这里的方差由于选取的部分数据，而不是总体 数据，所以我们的方差是样本方差而不是总体方差，而经过科学统计的结果，样本方差÷的是（n+1）而不是 n!



函数为：VAR.S

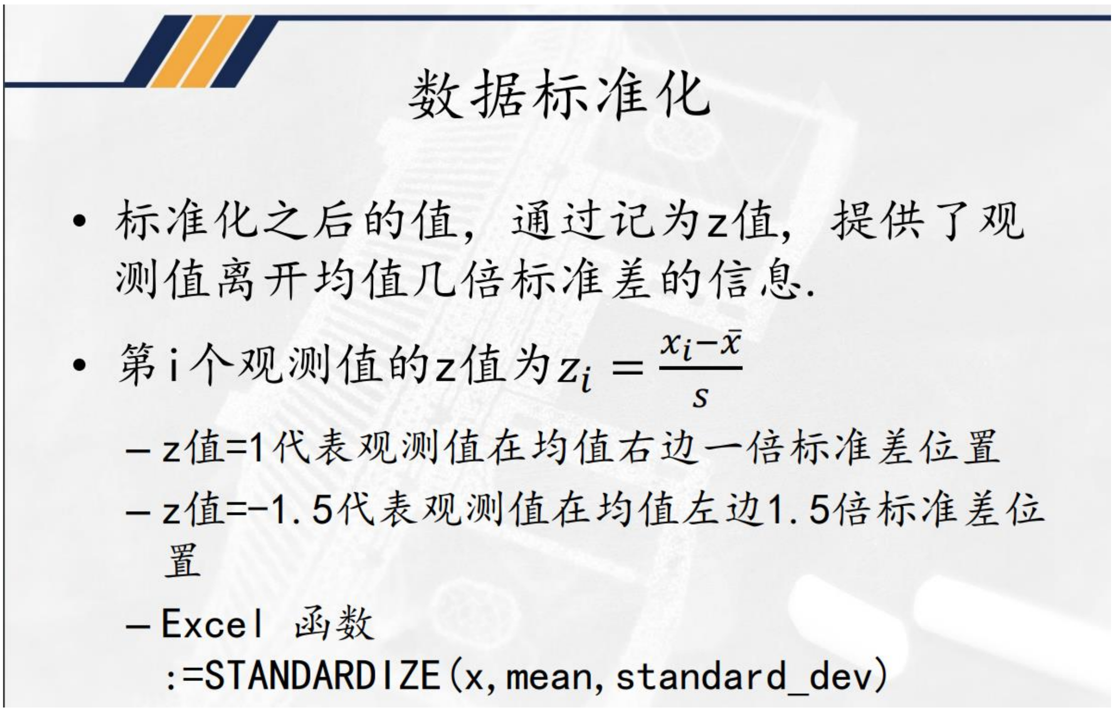
为了单位的统一，我们一般使用的标准差 函数为 STDEV.S



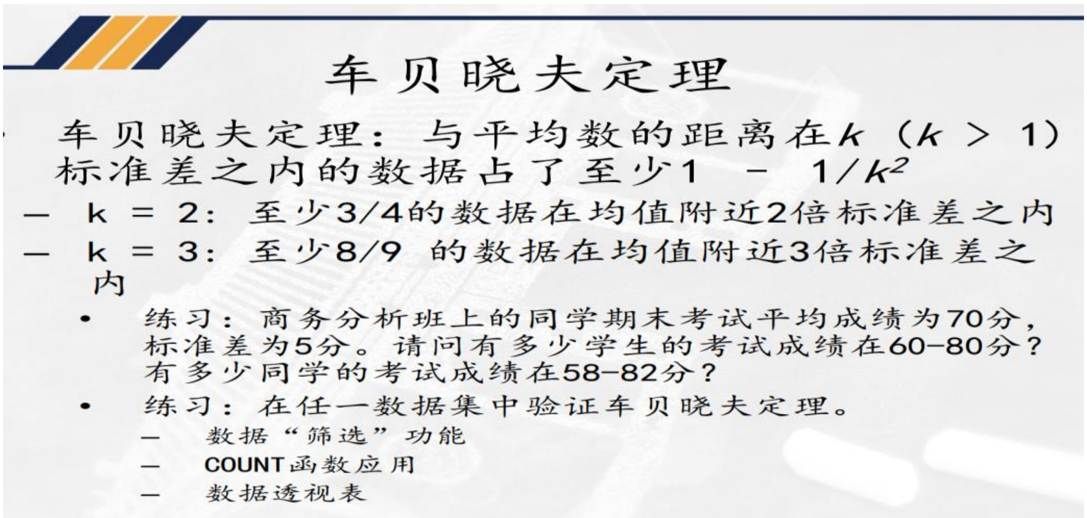
引入新的概念：标准差系数 CV=标准差/平均值*100%

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）

数据标准化：即求距离样本均值几倍标准差 函数：STANDARDIZE



引入标准化后的数据是为了运用车贝晓夫定理!



Ps 注意，可能需要调整单元格格式来显示负数。

而在车贝晓夫定理的基础上，经验法则又告诉我们。现实中数据往往成钟形。，此时有：

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

!! 应当指出!!!：以两倍标准差情况为例，我们可以说大约 95%的观测值在两倍标准差内（“钟形”经验法则），也可以说有至少 75%观测值在两倍标准差内（车贝晓夫定理），但我们决不能说至少 90%观测值在两倍标准差内！见例如下：

商务分析班上的同学期末考试平均成绩为 80 分，标准差为 5 分。请问有多少学生的考试成绩在 70-90 分之间？

大约 95%

至少 90%

大约 68%

至少 75%

我们在得到了标准化的数据（Z 值）后，可以考虑逆用车贝晓夫定理进行异常值分析。

异常值分析：
法一：Z 值小于-3 或者 Z 值大于 3 的
法二：利用 IQR（四分位差）：数据值超出 Q3+1.5IQR 和 低于 Q1-1.5IQR 的。
Pay attention：分析得到的异常值不一定是错误值！所以如果要进行后续分析时不可以把异常值舍去！

我们之前在用图表显示数据，即数据可视化时提到了我们分析数据，有用柱状图，饼状图分析不同类别或其他具有“独立性”数据的方式，也有如折线图，交叉分组表等用来分析数据之间联系的方式。同样，我们在进行描述性分析时也遵循这个思路，同样也要分析具有相关性的数据。

引入协方差。

注意到，由于是根据样本数据得出的协方差，我们的分母同样除的是 n-1

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

函数为 CORREL 实际上，这个协方差实际我们在高中阶段就已经有所接触，在高中概率统计中，这个东西常常出现在回归方程的相关计算中。

因为 L_{xy} 的大小与数据的度量单位有关，所以不宜直接用它度量成对样本数据相关程度的大小. 例如，在研究体重与身高之间的相关程度时，如果体重的单位不变，把身高的单位由米改为厘米，则相应的 L_{xy} 将变为原来的 100 倍，但单位的改变并不会导致体重与身高之间相关程度的改变.

为了消除度量单位的影响，需要对数据作进一步的“标准化”处理. 我们用

$$s_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2},\quad s_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

分别除 $x_i-\overline{x}$ 和 $y_i-\overline{y}(i=1,2,\cdots,n)$，得

$$\left(\frac{x_1-\overline{x}}{s_x},\frac{y_1-\overline{y}}{s_y}\right),\left(\frac{x_2-\overline{x}}{s_x},\frac{y_2-\overline{y}}{s_y}\right),\cdots,\left(\frac{x_n-\overline{x}}{s_x},\frac{y_n-\overline{y}}{s_y}\right).$$

为简单起见，把上述“标准化”处理后的成对数据分别记为

$$(x'_1,y'_1),(x'_2,y'_2),\cdots,(x'_n,y'_n),$$

仿照 L_{xy} 的构造，可以得到

$$r=\frac{1}{n}(x'_1y'_1+x'_2y'_2+\cdots+x'_ny'_n)$$

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

我们称 r 为变量 x 和变量 y 的**样本相关系数**（sample correlation coefficient）.

这样，我们利用成对样本数据构造了样本相关系数 r. 样本相关系数 r 是一个描述成对样本数据的数字特征，它的正负性可以反映成对样本数据的变化特征：

当 $r>0$ 时，称成对样本数据正相关. 这时，当其中一个数据的值变小时，另一个数据的值通常也变小；当其中一个数据的值变大时，另一个数据的值通常也变大.

当 $r<0$ 时，称成对样本数据负相关. 这时，当其中一个数据的值变小时，另一个数据的值通常会变大；当其中一个数据的值变大时，另一个数据的值通常会变小.

我们现在尝试给出相关系数 r 在-1 到 1 之间的证明（协方差可以为负）:
法一，高中课本证明方法:

观察 r 的结构，联想到二维（平面）向量、三维（空间）向量数量积的坐标表示，我们将向量的维数推广到 n 维，n 维向量 $\boldsymbol{a}$，$\boldsymbol{b}$ 的数量积仍然定义为

$$\boldsymbol{a}\cdot\boldsymbol{b}=|\boldsymbol{a}||\boldsymbol{b}|\cos\theta,$$

其中 θ 为向量 $\boldsymbol{a}$，$\boldsymbol{b}$ 的夹角. 类似于平面或空间向量的坐标表示，对于向量 $\boldsymbol{a}=(a_1, a_2,$

98 第八章 成对数据的统计分析

人民教育出版社

$\cdots, a_n)$ 和 $\boldsymbol{b}=(b_1, b_2, \cdots, b_n)$，我们有

$$\boldsymbol{a}\cdot\boldsymbol{b}=a_1b_1+a_2b_2+\cdots+a_nb_n.$$

设"标准化"处理后的成对数据 (x'_1, y'_1)，(x'_2, y'_2)，$\cdots$，(x'_n, y'_n) 的第一分量构成 n 维向量

$$\boldsymbol{x}'=(x'_1, x'_2, \cdots, x'_n),$$

第二分量构成 n 维向量

$$\boldsymbol{y}'=(y'_1, y'_2, \cdots, y'_n),$$

则有

$$r=\frac{1}{n}\boldsymbol{x}'\cdot\boldsymbol{y}'=\frac{1}{n}|\boldsymbol{x}'||\boldsymbol{y}'|\cos\theta.$$

因为 $|\boldsymbol{x}'|=|\boldsymbol{y}'|=\sqrt{n}$，所以样本相关系数

$$r=\cos\theta,$$

其中 θ 为向量 $\boldsymbol{x}'$ 和向量 $\boldsymbol{y}'$ 的夹角.

由 $-1\leqslant\cos\theta\leqslant1$，可知

$$-1\leqslant r\leqslant1.$$

法二：$r_{xy}=\frac{Sxy}{SxSy}$ 对这个式子左右两边同时平方，即证明 RHS（右手边）在 0 到 1 之间。式子大于零显然，即证式子小于 1，我们把分母移到另一边，可以发现新的式子就是 n 阶柯西不等式，得证。(n 阶柯西不等式可以用根的判别式 detail 来证明)

$$\sum_{k=1}^{n}{a_k}^2\sum_{k=1}^{n}{b_k}^2\geq\left(\sum_{k=1}^{n}a_kb_k\right)^2$$

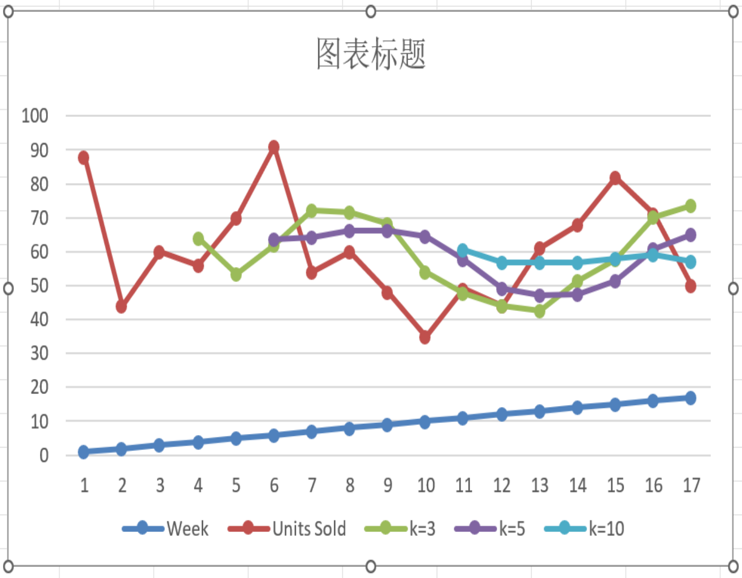
r 中的 Sx 与 Sy 均为两条数据的标准差。
应该指出，我们之所以要弄出一个样本相关系数，是因为协方差这个东西很容易收到量纲影响，不方便我们比较不同量纲（即数量级）之间的数据相关程度，因此我们使用样本相关系数。
r 离 1 越近代表数据正相关性越强，离-1 越近代表负相关性越强，而离 0 越近一般也就没有相关性。
但是，但是，但是！如果 r 趋向于 0 并不一定代表数据没有相关性，例如 x 的绝对值函数，如果取左右两端一样的距离，那么就会得到相关系数为零的结论。（正负相关性互相抵消了。）因此，我们进行相关性分析前要做的第一件事是生产散点图。

预测方法：
三大类预测方法：
- 主观预测
- 基于时间序列的预测模型
- 基于特征的预测模型

我们在这章中主要讨论基于时间序列的预测模型。
主要利用历史数据，如每周的销量。
时间序列可能包括了趋势性、季节性或者平稳的时间序列（只存在随机波动，如果是波浪形上下变化，是可以看作平稳时间序列的）
对于平稳时间序列，我们给出两种预测模型：移动平均法和指数平滑法。
应该指出，模型预测只能进行短期预测，而且使用前提是这是平稳的时间序列。

1、移动平均法：对下一期的预测等于最近 k 期观测值的平均
数学表达式：
$F_t=1/3A_{t-1} +1/3A_{t-2} +1/3A_{t-3}$ （以 k 取 3 为例）

Week	Units Sold	k=3	k=5	k=10
1	88			
2	44			
3	60			
4	56	64		
5	70	53.3333		
6	91	62	63.6	
7	54	72.3333	64.2	
8	60	71.6667	66.2	
9	48	68.3333	66.2	
10	35	54	64.6	
11	49	47.6667	57.6	60.6
12	44	44	49.2	56.7
13	61	42.6667	47.2	56.7
14	68	51.3333	47.4	56.8
15	82	57.6667	51.4	58
16	71	70.3333	60.8	59.2
17	50	73.6667	65.2	57.2



2 k 值取得越大，预测越“平滑”。原因是 k 值越大，相当于两个数据之间使用的数据重合度

越高，受极端值影响小。

1、 移动平均法的思想：通过多期的加总，把随机波动的影响去除，得到数据真正的水平。即把所有的误差进行了求和然后平均化。

总而言之，这个方法在对数据波动大的数据时建议取较小的 k，而在数据波动小时则取较大的 k。因为本身就比较平稳，那么肯定我取更多的数据进行预测就好，而如果本来就有起伏趋势，取过大的 k 可能会使得这种向上趋势被磨损掉。

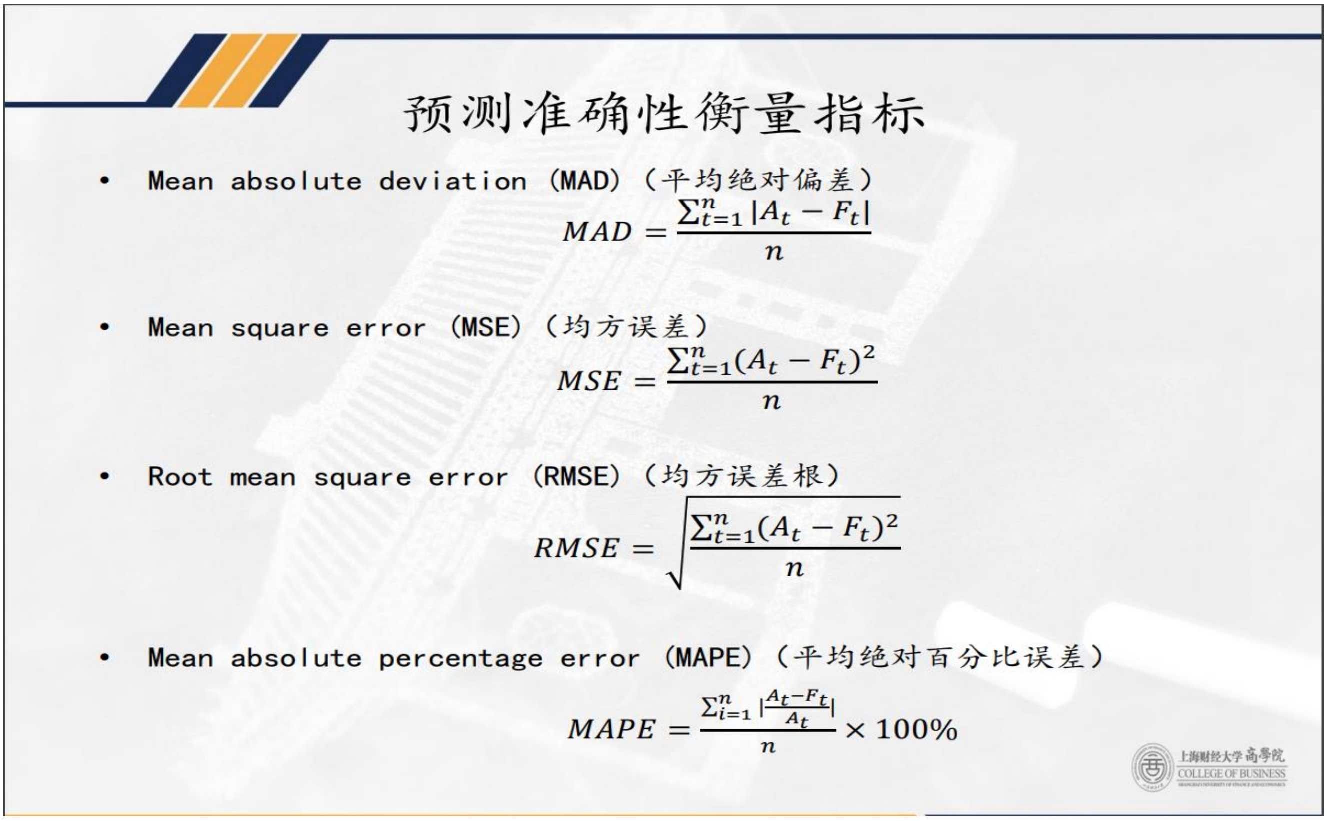
接下来我们尝试看看预测是否准确。

我们可以引入一下几个数据

MAD\MSE\RMSE

但是有时候我们还想要比较不同数量级的数据，我们就可以使用 MAPE，来达到消除数量级的影响

求绝对值：ABS （absolute）



注意这个东西不是方差，因为他每项是预测减去观测的计算处理，而不是数据与平均数的计算处理

2、简单指数平滑模型

我们想要预测时，离我预测数近的数据影响大，离我数据远的数据影响小：

指数平滑法即加权移动平均法。我们改写一下以上的式子，保证全部系数加起来仍然为 1，

且离 t 越远，数越小。 α求和为 1，由等比数列求和算出。

$F_t = \alpha A(t-1) + \alpha(1-\alpha)A(t-2) + \alpha(1-\alpha)^2A(t-3) + \cdots\cdots$

$= \alpha A(t-1) + (1-\alpha)\,[\alpha A(t-2) + \alpha(1-\alpha)A(t-3) + \cdots\cdots]$

$= \alpha A(t-1) + (1-\alpha)F_{t-1}$

所以我们用我们最后推的的式子进行计算即可，由于 F1 无法得出，我们一般可以直接取第一项值作为 F1。 但是最后统计平均误差时记得不能使用第一项数据。

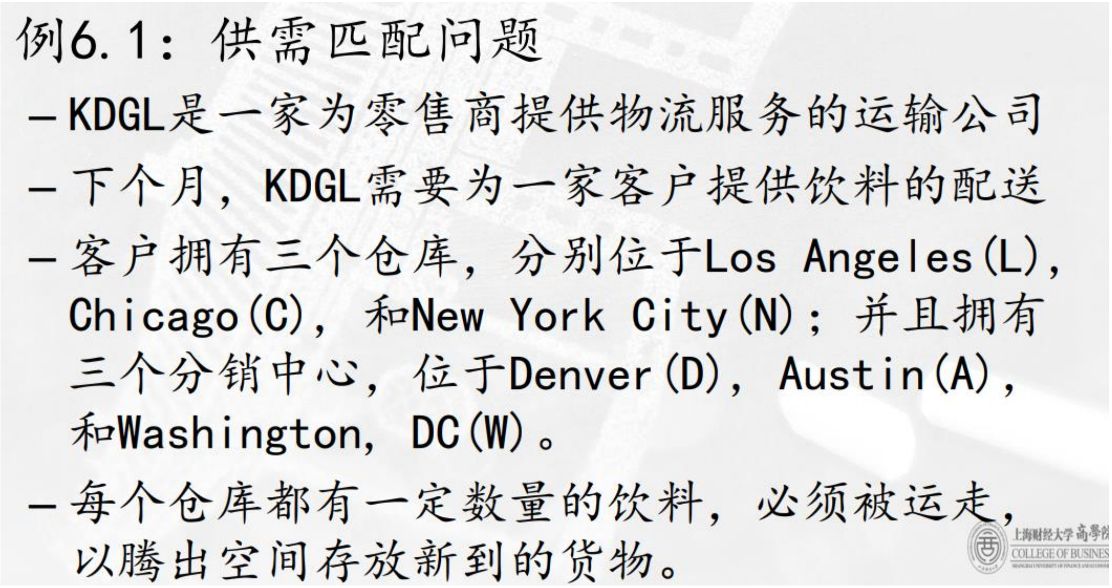
我们对 α 做出理解：

1、 这个方法的好处就在于随着 t 值越大，越后面的估计值就越少受前面估值的影响。而α的值越大，这种受前面估值影响递减的速度就越大，越快摆脱前面数据的影响。
2、 我们再改写一下我们的式子，得到了 $F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$这个式子可以这么来理解：$A_{t-1} - F_{t-1}$ 即为第 t-1 位数据的预测值与实际值的差值，而这个式子整体含义即补上这部分误差，且α越大，代表这份误差弥补的程度就越大。

决策性分析：

1、 决策模型：确定性模型 不确定性模型

对于确定性模型：见下图例子，即有特特别要求的模型。



例6.1：供需匹配问题

– KDGL是一家为零售商提供物流服务的运输公司
– 下个月，KDGL需要为一家客户提供饮料的配送
– 客户拥有三个仓库，分别位于Los Angeles(L)，Chicago(C)，和New York City(N)；并且拥有三个分销中心，位于Denver(D)，Austin(A)，和Washington，DC(W)。
– 每个仓库都有一定数量的饮料，必须被运走，以腾出空间存放新到的货物。

每个仓库数量

Warehouse	To Be Shipped Out (tons)
Los Angeles (L)	15
Chicago (C)	20
New York City (N)	30

每个分销中心都有需求必须被满足

Distribution Center	Minimum To Be Shipped In (tons)
Denver	10
Austin	13
Washington, DC	20

不同地点之间的单位运输成本不同

From / To	Denver (D)	Austin (A)	Washington, DC (W)
Los Angeles (L)	$105.00	$135.00	$153.00
Chicago (C)	$110.00	$140.00	$137.00
New York City (N)	$130.00	$132.00	$115.00

我们整理一下已经获得的信息，可以得到一张下列的表格:

KDGL的决策、目标和约束？

Minimize the Total Shipping Cost

$$105^*X_{LD} + 135^*X_{LA} + 153^*X_{LW} + 110^*X_{CD} + 140^*X_{CA} + 137^*X_{CW} + 130^*X_{ND} + 132^*X_{NA} + 115^*X_{NW}$$

$X_{LD} + X_{LA} + X_{LW} = 15$ (Los Angeles supply)

$X_{CD} + X_{CA} + X_{CW} = 20$ (Chicago supply)

$X_{ND} + X_{NA} + X_{NW} = 30$ (New York supply)

$X_{LD} + X_{CD} + X_{ND} \geq 10$ (Minimum Denver demand)

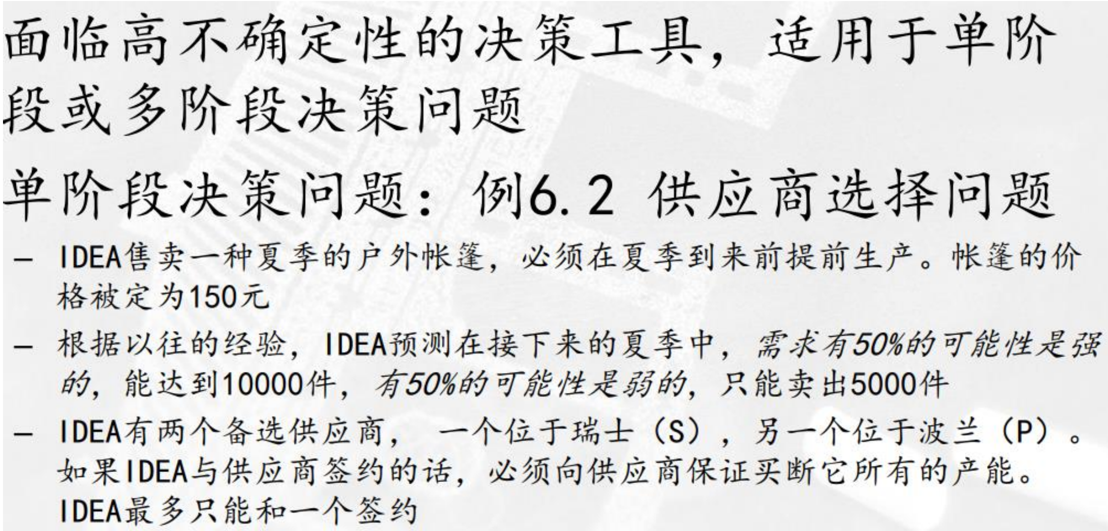
$X_{LA} + X_{CA} + X_{NA} \geq 13$ (Minimum Austin demand)

$X_{LW} + X_{CW} + X_{NW} \geq 20$ (Minimum Washington, DC demand)

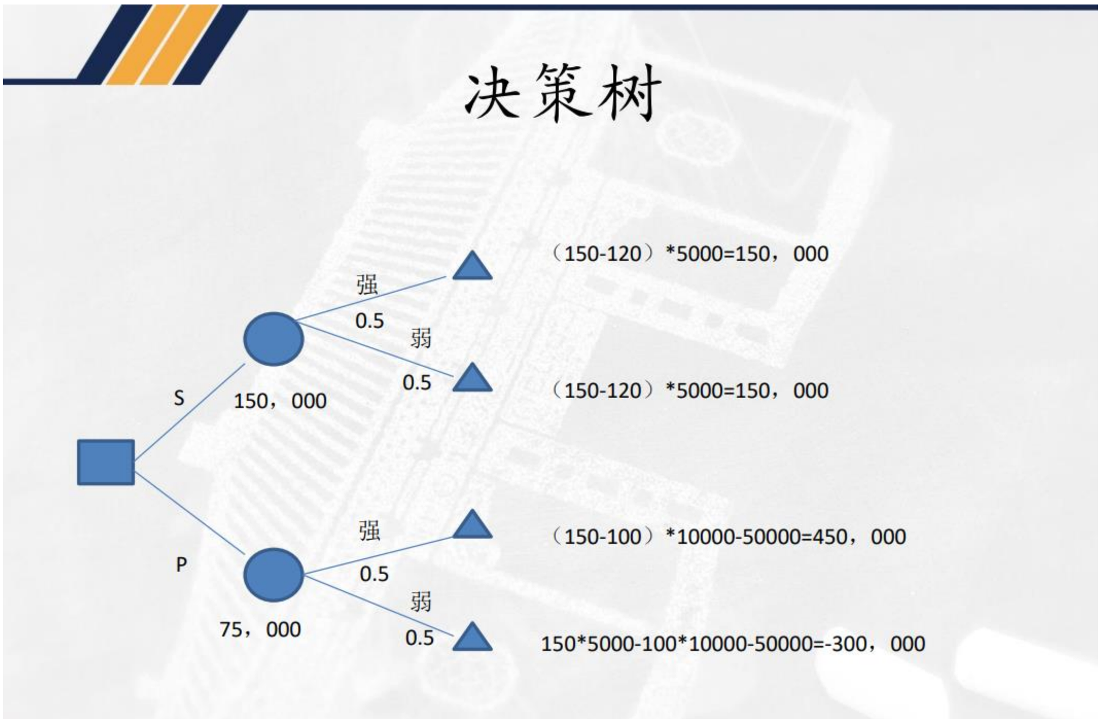
$X_{LD}, \ldots, X_{NW} \geq 0$ (Non-negativity)

• 确定性模型的好处
– 可以解决较大规模的问题，多产品，多资源
– 当决策者对环境有较大的控制时确定性假设有合理的地方，例如:
  • 制定短期计划
  • 制定带长期合同（价格、需求、供给）的长期计划
• 确定性模型的缺点
– 当现实中不确定性较高时，提供的决策不可靠

而生活中往往都是不确定性模型，我们主要有仿真模拟，决策树等方法。
我们重点谈谈决策树模型。
决策树模型常常用于概率性发生的情况，下面以一个问题为例，重要的是掌握理解决策树，画决策树:



正方形表示决策点，即决策者必须要做出决策的点。从该点延申出的线表示不同的决策方案。
圆形表示事件点，或者机会点。从这点引出的线表示不同的事件（不会同时发生），边下的数字表示该事件出现的概率。而且我们要在圆形下算出期望收益。
三角形表示结果点，即决策者做出该决策发生该事件后出现的结果，并在这类节点旁注明各种结果的收益损失。



对于结果的评价标准有：
风险规避——“maxi-min 最大化最差结果”
风险追逐（赌狗）——“maxi-max”最大化最好结果
中性——最大化期望收益（计算平均收益）