《生物统计》 预习

1. Chap 3 变量关系与探索

案例：

高血压影响因素研究

样本量：11812

多省市流行病调查：

北京 广东 黑龙江 江苏 内蒙古 山西 陕西 四川 西藏 新疆 云南 浙江

研究目的：

量化理化指标与是否罹患高血压之间的关系

筛选高血压患病危险因素

潜在因素：

生活习惯

理化指标

家族遗传信息

研究思路

• 交叉表

• 相关性探索（//correlation 相互关系 重相互 cor合作。 看两者之间的关系）

• 关联性探索（//relevance Association Relationship 关联性 重联 关系 rele 亲戚。复杂的多者关联。）

• 模型构建与分析

列联表 contingency table 质量管理相依表

（//contingency 未来可能发生的事;不测之事;突发事件 con共 ting接触 ency行为）

图：两张调查表，统计得高血压的和未得的，是否患过脑卒中

相关性 （//correlation 相互关系 重相互 cor合作。看两者之间的关系)

有两个数值型变量

有两个分类变量

看数值型变量与分类变量的关系

两个数值变量

例1：假设我们有兴趣研究男性的体脂(body fat)百分比和腹部周长之间的关系。

用散点图来显示两个数值变量之间的关系是一种简单的方法。

散点图

例2：我们研究了美国各州恶性黑色素瘤年死亡率与其地理中心纬度之间的关系。

利用散点图，我们可以发现两个数值变量之间可能的关系。

在上面的例子中，我们可以看到一个变量的变化与另一个变量的实质性系统变化（增加或减少）是一致的(//体脂图上两个值交点坐标散点分布在一条斜线附近，两个变量线性相关)。

因为整体关系可以用直线表示，所以我们说这两个变量是线性关系。

我们认为体脂百分比与腹围呈正线性关系（positive linear relationship）。

（//正相关 斜线 向上，值是变大的）

相比下，第二例图，我们认为恶性黑色素瘤的年死亡率与纬度呈负线性关系（negative linear relationship）。

（//负相关 斜线 向下，值是变小的）

散点图

在某些情况下，这两个变量是相关的，但关系不是线性的（//图 散点围 在斜线周围波动上升）

在其他一些情况下，两个变量之间没有关系（线性或非线性）（//图 无规则散布的点）

（//注意，点是x,y两个变量的交点，点的规律就是它们的关系，比如上面线性的， 根据线的走向就可以预测下次x为什么值时 y的值大概是多少）

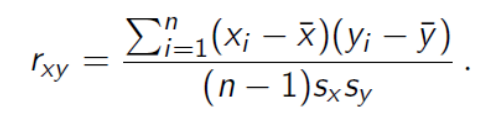
相关性 （//Correlation 两个变量的相互关系）

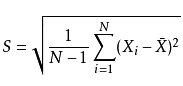
量化两个数值变量之间线性关系的强度（力度）和方向。我们可以用皮尔逊（Pearson’s）相关系数(correlation coefficient) ---- r ，作为汇总统计。

r 值永远在 -1 到 1 之间

当r 越接近 -1 或 1 时 两变量关系越强

r表示线性关系的走向（正相关或负相关）



S为样本标准偏差

代表x的平均值

//公式理解：上式意为 每个x与x均值之差 乘上 y与y 均差 累加 /（样本数-1） 乘 x的标准偏差 乘y的标准偏差（即两方均值差累加后再用两方的标准的偏差中合一下，取个平衡）

//至于标准偏差：单个变量样本与变量样本均值差的平方累加/ (样本数-1) 再开方（开方是因为里面加平方了，加上去的要还原，这个洗下牌就是标准平均差了）

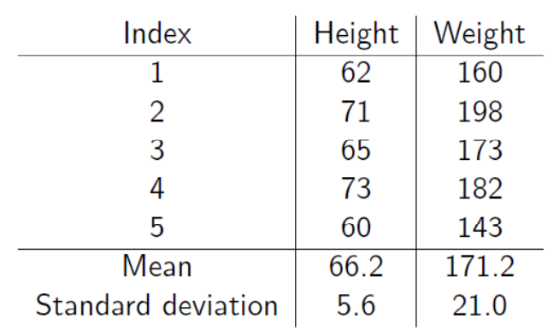
//公式中的除以n-1 就相当于 除以n，样本个数，也是为了每个样本平均下来， -1 是为了不让分母为0的情况出现。有些公式里会有+1，都是一个用意。1很小，在统计学宏观视野里面产生的值影响可以忽略。

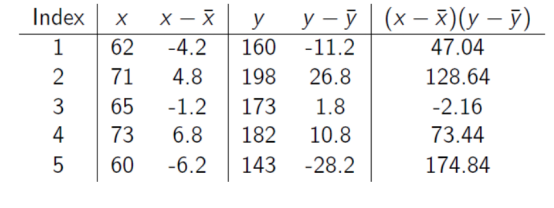
//统计的公式要在宏观上理解代表的含义，必竟公式是人总结出来的，为了表示一种现象的。常人正常应该都能理解才对[捂脸]。

补充，r代表线性关系的走向，即两股势力的相互拉扯

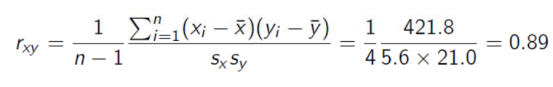
公式图

计算示例：





给定以上两组计算中间表，代入值，用上式计算r



两个分类变量

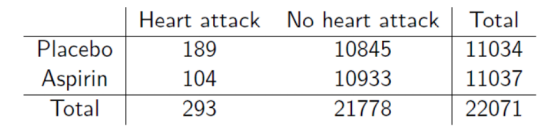


图 表格

在本例中，我们考虑了一项为期五年的研究，研究定期服用阿司匹林是否能降低心血管疾病的风险。

我们通常用列联表（contingency table）来汇总这些数据。

每个格子显示一种可能的疾病状态（心脏病发作或无心脏病发作）和实验组（安慰剂或阿司匹林）组合的频率。

利用这些频率，我们可以分别计算每个实验组中心脏病患者的样本比例。

安慰剂组有11034人，其中189人心脏病发作。因此，安慰剂组心脏病发作患者的比例为pl=189/11034=0:0171。

阿司匹林组心脏病患者比例（proportion）为p2=104/11037=0:0094。

我们将其（p）称为心脏病发作的风险（此处，样本比例用于测量风险）。

两个实验组心脏病发作的样本比例之间的显著差异让我们相信治疗和疾病状况是相关的。

测量关系强度的一种方法是计算【比例差】p2-p1。

与安慰剂组相比，阿司匹林组患心脏病的比例降低了0.0077。

比较样本比例的另一个常见汇总统计是【相对比例】p2/p1。

由于本例中的样本比例与心脏病发作的风险有关，我们将相对比例称为相对风险。

在这里，心脏病发作的相对风险是p2/p1 = 0:0094/0:0171 = 0:55。

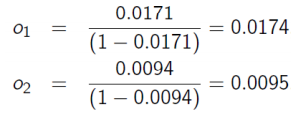
这意味着阿司匹林组心脏病发作的风险是安慰剂组的0.55倍。

--------

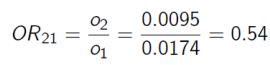
比较【样本概率】(odds 可能性，概率 )更常见

（即患者比例/非换患者比例）

安慰剂组O1和阿思匹林组O2心脏病发作的几率分别是：

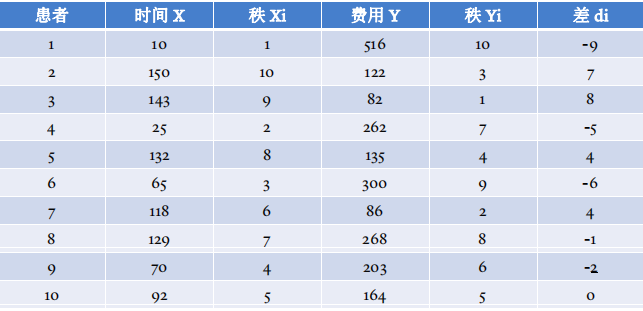


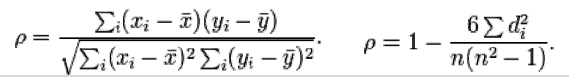
通常，比较样本概率，我们用【样本概率比】(sample odds ratio)



即o1/o2 两个概率比

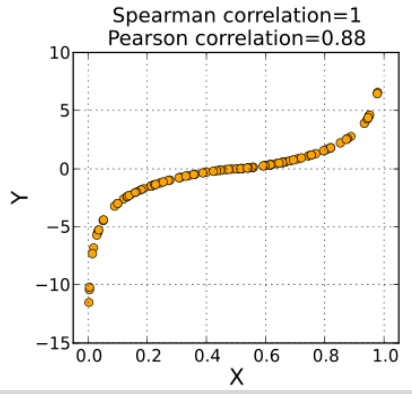
斯皮尔曼秩(rank 等级)相关系数 Spearman's rank correlation coefficient

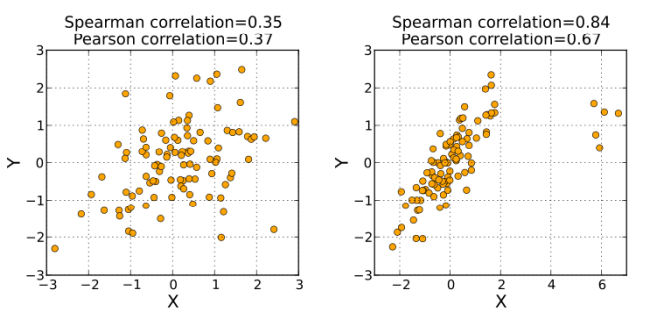




公式图

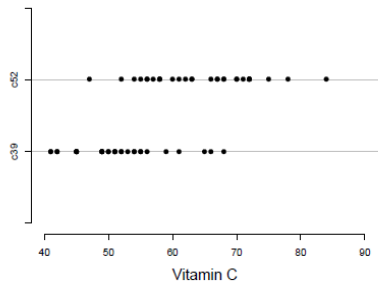
斯皮尔曼vs皮尔逊 （Spearman vs.Pearson）





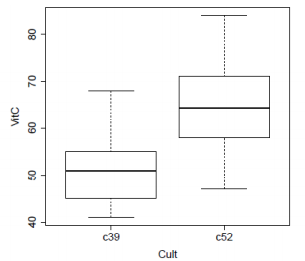
数值型变量和分类型变量的关系

常见的情况是，我们常常对分类变量和数值随机变量之间的关系感兴趣。



上图为从大包装中收集的卷心菜数据按品种（分类）划分的维生素C含量点图（数字）

一种更常见的展示 数值变量和分类变量关系的方法 是创建方框图。



不同品种维生素C含量盒图

一般来说，如果其中一个变量的分布随另一个变量的变化而变化，那么这两个变量是相关的。

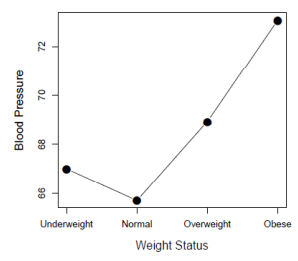
我们可以通过获得分类变量不同等级的汇总统计来测量数值变量分布的变化。

在检验一个数值变量和一个分类变量之间的关系时，通常使用平均差。

在上述例子中，两个品种之间的维生素C含量平均值（箱图中方盒的中线）的差异为64.4-51.5=12.9。

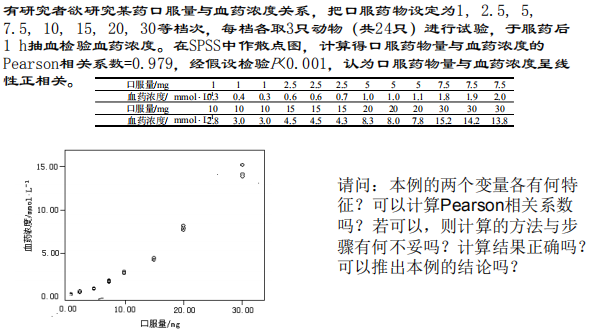
------

当分类变量具有多个等级（类别）时，使用平均值点图更容易比较不同层次的平均值。

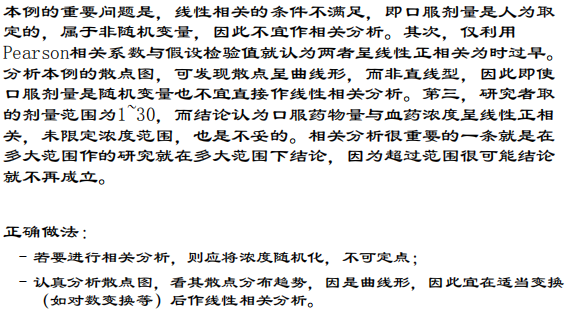


上图绘制了不同体重组（根据体重指数定义）的血压（bp, blood pressure）均值

思考题



参考答案：



关联性规则Association Rules：（多者相关）

起源:大型零售组织所面临的决策支持问题

- R .Agrawal等人在1993年首先提出了在交易数据项目之间挖掘关联规则的问题( Mining Association Rules)

内涵:分析数据库中不同变量间的关系程度

- 并不是要作分类/预测;更像是一种描述性分析

- 挖掘教据中的某类模式,事件与事件之间的依赖性或关联性

- 首先是应用于二分変量,后来才发展到多分类变量和连续变量

应用:

- 交叉销售

- 商品摆放

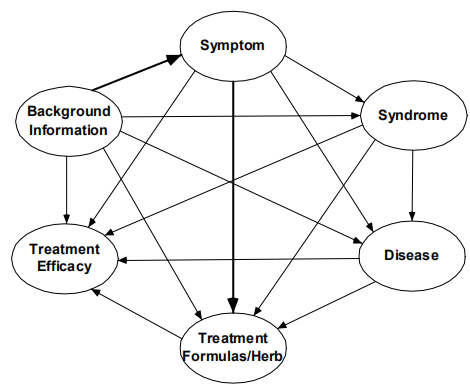
- 流失户分析

- 临床经验结

商业应用：

网上的商品推荐APP等

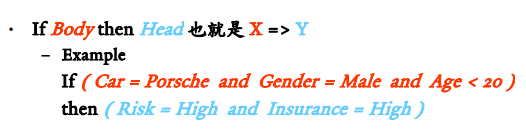
名医经验探索



关联规则

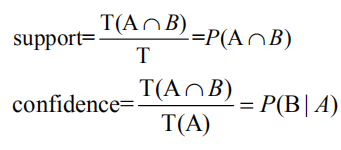
通过事务数据库，通过关联规则算法计算，得到规则





支持度（support）: A和B均发生的比例；反映规则的可靠程度。

置信度（confidence confidence）： A发生的事务中（//情况下)，B发生的比例；反映规则的把握程度。



 比例 概率

 A和B同时发生

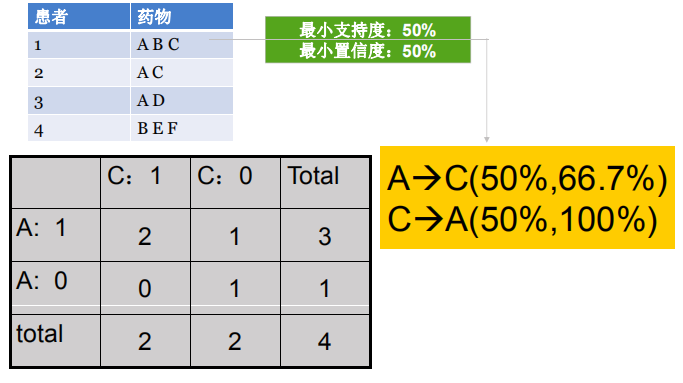
 A发生的个数里，B发生的占多少

 A为真的个数

T 全部样本数量

----------------

例



//看灰色表： 发生T真值用1表示，不发生用0表示，

支持度 = AC同时发生 两个1交叉格子2个/全部数量4, 50%，

置信度 = 同时发生的2个/A为真时的个数3 = 66.7%

// 看原始彩色表 A->c S 支持度 AC同时发生2个/总4个 = 50%; c置信度 AC同时发生的2个/A发生3个 = 66.7%

//S支持度 = CA同时发生 50%; c 置信度= CA时时发生2个/C发生2个 =100%

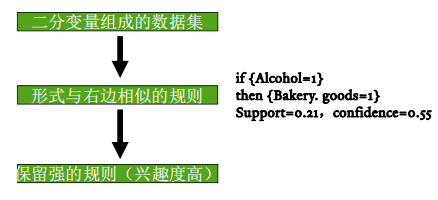
规则很多，都要吗？

- 每两个两组商品关系都可用这种规则表示

- 没必要把所有的规则都找出来，只找感兴趣的

- 最小支持度，最小置信度

- 人为定的，多次尝试，调整阈值，规则少时阈值（门槛）降低，多时门槛提升



评价指标

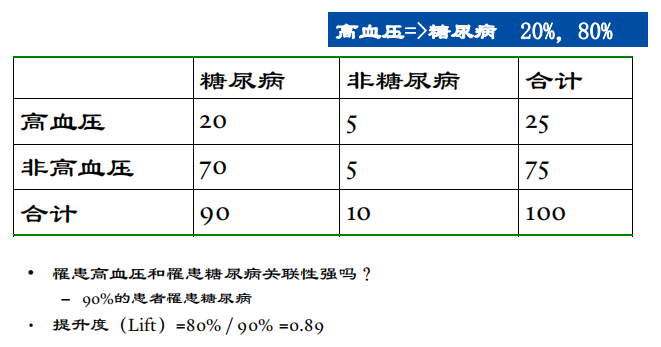
支持度

置信度

提升度

时效支持度

时效置信度



上式

s 支持度 = 20 高血压为糖尿病都为真（第二行二列格子）/100（总样本数） =20%

C 置信度（放置信心） = 20 都为真(都发生)/ 高血压为真的样本个数 25 （第二行四列格子） = 80%

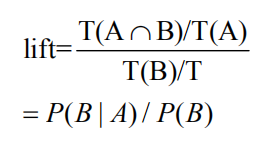
//C 和S不是相反的，c是S的递进和优化，因为s是在总样本中的发生率，c是在真样本（过滤掉了假的，本来就不可信的部分）中的发生率，

Lift 提升度 = 推导置信度/ 糖尿病（推导结果）成立的概率

//Lift是C进一步的递进和优化，中和了推导结果的发生率

即，下式

提升度Lift：



Lift=1，前项和后项经验独立

Lift>1，正相关的 （//分子大，分母小，即置信度在推导结果中的占比大，可置信，正相关）

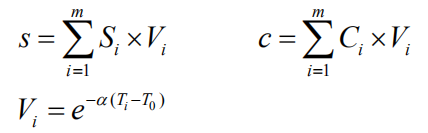
Lift<1，负相关的 （//分母大分子小， 即置信度在推导结果中的占比大小，不可置信，负相关）

时效支持度，时效置信度

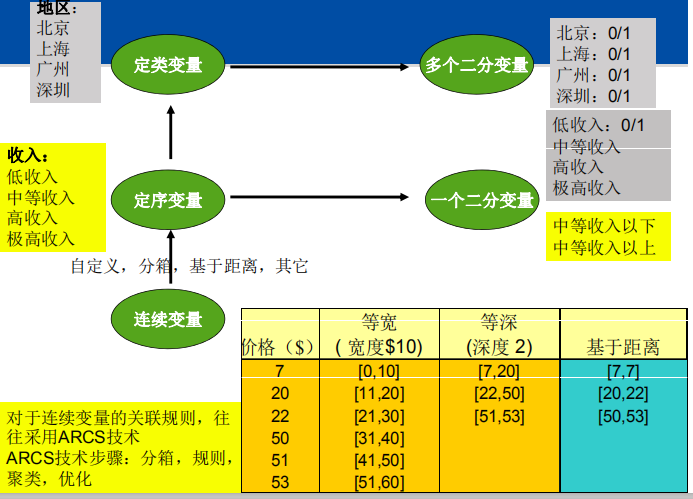
• 数据库越来越大，时间跨度很长，不同时间的事务，其价值不同

• 越近于考察点，价值越大。

• 加权思想（对每笔事务根据其与考察点的远近，作为权重，权重指数衰减//时间多重要性小，所以是负相关，衰减的）



连续性数据处理



算法简述

最有影响的方法是Apriori先验法；（//比如先假定一个概率，再去验证？）

规则可以分为两步：

– 先找频繁项集（支持度s大于最小支持度的所有项集）；

– 基于频繁项集生成规则；

– 效率问题主要在第1步，针对第1步的效率提出了一些改进方法。

统计案例：

拓展阅读

纸上得来终觉浅，绝知此事要躬行！