大数据分析基础

统计分析

- 统计分析理论基础
- 利用Python实现统计分析

- 概率
  - 亦称“或然率”，它是反映随机事件出现的可能性（likelihood)大小
- 条件概率
  - 条件概率是指事件A在事件B发生的条件下发生的概率，表示为：P（A|B）
  - 若只有两个事件A、B，那么$P(A|B) = \frac{P(AB)}{P(B)}$
    - P(AB)表示A、B两件事共同发生的概率，即联合概率
    - 如果A、B是相互独立的事件，则P(AB) = P(A)*P(B)
  - 全概率公式
    - 如果事件B_1、$B_2 \ldots B_n$构成一个完备事件组，即它们两两互不相容，其和为全集($\sum_{i=1}^{n} P(B_i)$=1)，且$P(B_i) > 0$，那么对于任意事件A都有$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$
  - 贝叶斯公式
    - $P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i)*P(A|B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$

贝叶斯公式

- 假设一个常规的吸毒检测结果的敏感度与可靠度均为99%，也就是说，当被检者吸毒时，每次检测呈阳性（+）的概率为99%；而被检者不吸毒时，每次检测呈阴性（-）的概率为99%。那么这个检测是否真的可靠呢？
- 假设某公司将对其全体雇员进行一次鸦片吸食情况的检测，已知0.5%的雇员吸毒。我们想知道，每位医学检测呈阳性的雇员吸毒的概率有多高？
  - 令“D”为雇员吸毒事件，“N”为雇员不吸毒事件，“+”为检测呈阳性事件
  - 通过描述可知：P(D) = 0.005（0.5%的人吸毒），显然P(N) = 1-P(D) = 0.995
  - P(+|D)代表雇员吸毒且检测为阳性的概率，为0.99；P(-|D) = 0.01（吸毒但阴性，即误诊）
  - P(+|N)代表雇员不吸毒但误诊为阳性的概率，为0.01；P(-|N) = 0.99
  - 检测呈阳性时确实吸毒的条件概率：$P(D|+) = \frac{P(D)P(+|D)}{P(+|D)P(D)+P(+|N)P(N)} = \frac{0.005*0.99}{0.99*0.005+0.01*0.995} = 0.332215$

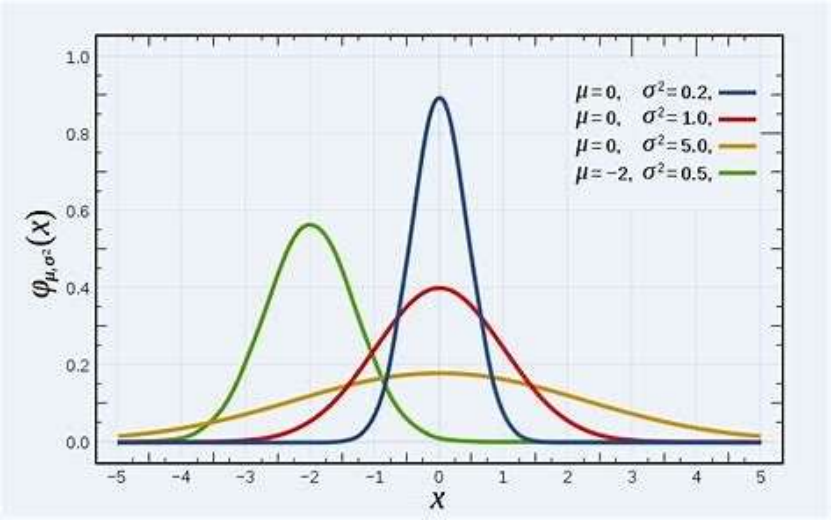
先验概率和后验概率

- 先验概率是指根据以往经验和分析得到的概率
  - 所谓的先验概率是我们人在未知条件下对事件发生可能性猜测的数学表示
- 后验概率是指事情已经发生，要求这件事情发生的原因是由某个因素引起的可能性的大小

后验概率 $P(D|+)$ = 先验概率 $\frac{P(D)P(+|D)}{P(+|D)P(D)+P(+|N)P(N)}$

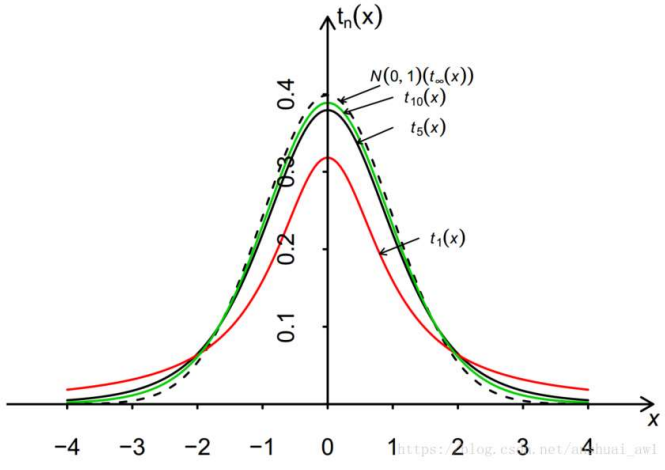
- 利用先验概率在事情发生后计算出后验概率，这个后验概率在后续的实践中就变成了先验概率，这就是贝叶斯学习的过程
  - 如果让此人再次复检（P(D)=33.2215%替换了原先的0.5%），将会得到此人吸毒的概率为98.01%。此人第三次复检，会得到此人吸毒的概率为99.98%，已经超过了检测的可靠度。

- 概率密度函数（*pdf*）
  - 如果对于随机变量X的分布函数*F(x)*，存在非负函数*f(x)*，使得对于任意实数有 $F(x)=\int_{-\infty}^{x} f(t)dt$，则称X为连续型随机变量，其中*f(x)*称为X的概率密度函数
  - *F(x)*则是累积分布函数（*cdf*）
  - 正态分布
    - 概率密度函数N(μ, σ^2)：$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
    - 累积分布函数：$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$
    - 标准正态分布：N(0, 1)



- 假设检验(Hypothesis Testing)
  - 又称统计假设检验，是用来判断样本与样本、样本与总体的差异是由抽样误差引起还是本质差别造成的统计推断方法
- 零假设（Null Hypothesis）
  - 在推论统计学中，零假设（符号H_0）是做统计检验时的一类假设
  - 零假设的内容一般是希望能证明为错误的假设，或者是需要着重考虑的假设
  - 在相关性检验中，一般会取“两者之间无关联”作为零假设
  - 与零假设相对立的是备择假设
    - 零假设H_0：A和B相等
    - 备择假设H_1：A和B不相等

- T检验，亦称student t检验（Student's t test），主要用于样本含量较小（例如n < 30），总体标准差σ未知的正态分布
  - T检验是用t分布理论来推论差异发生的概率，从而比较两个平均数的差异是否显著
    - H_0：没有显著差异
  - 适用条件
    - 已知一个总体均数
    - 可得到一个样本均数及该样本标准差
    - 样本来自正态或近似正态总体
  - 单总体检验
    - 单总体t检验是检验一个样本平均数与一个已知的总体平均数的差异是否显著
  - 双总体检验
    - 双总体t检验是检验两个样本平均数与其各自所代表的总体的差异是否显著
  - 配对样本检验
    - 配对样本t检验可视为单样本t检验的扩展，不过检验的对象由一群来自常态分配独立样本更改为二群配对样本之观测值之差



- Type I error & Type II error
  - Type I：P(拒绝H_0|H_0为真) = α
  - Type II：P(接受H_0|H_0不为真) = β

		根据研究结果的判断	
		拒绝H_0 （实际上拒绝零假设）	接受H_0 （实际上不拒绝零假设）
真实情况	H_0是真实的 （理论上不应该拒绝零假设）	错误判断 （阳性判断错误 Type I error）	正确判断
	H_0是错误的 （理论上应该拒绝零假设）	正确判断	错误判断 （阴性判断错误 Type II error）

Type I error & Type II error

- 我们希望α和β都尽可能小，但是当样本容量到一定的时候，α和β是互相制约的，一个变小之后另一个就会变大
  - Type I：药物无效却被认为有效；Type II：药物有效却被认为无效
- 通常人们的做法是控制犯第一类错误的概率不大于一个较小的数α，而使得犯第二类错误的概率β尽可能小
  - 因为一般来说H_0都是我们希望拒绝的假设
- 这种第一类错误的概率加以限制而不考虑第二类错误β的假设检验称为显著性检验，而将给定的犯第一类错误的概率称为显著性水平。
  - 显著性检验是假设检验中最常用的一种方法，也是一种最基本的统计推断形式
  - 其基本原理是先对总体的特征做出某种假设，然后通过抽样研究的统计推理，对此假设应该被拒绝还是接受做出推断。

P-value

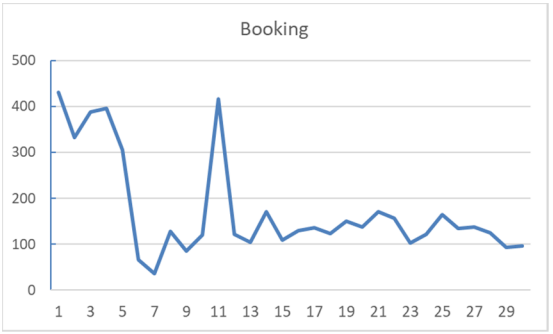
- P值即概率，反映某一事件发生的可能性大小
- P值的计算
  - 一般地，用X 表示检验的统计量，可由样本数据计算出该统计量的值C，根据检验统计量X的具体分布，可求出P值
    - 左侧检验的P值为检验统计量X 小于样本统计值C 的概率，即：P = P{ X < C}
    - 右侧检验的P值为检验统计量X 大于样本统计值C 的概率：P = P{ X > C}
    - 双侧检验的P值为检验统计量X 落在样本统计值C 为端点的尾部区域内的概率的2 倍：P = 2P{ X > C} (当C位于分布曲线的右端时) 或P = 2P{ X< C} (当C 位于分布曲线的左端时) 。若X 服从正态分布和t分布，其分布曲线是关于纵轴对称的，故其P 值可表示为P = P{| X| > C} 。
  - 计算出P值后，将给定的显著性水平α与P 值比较，就可作出检验的结论
    - 如果α > P值，则在显著性水平α下拒绝原假设
    - 如果α ≤ P值，则在显著性水平α下不拒绝原假设

不拒绝不等于接受原假设

数据的类型

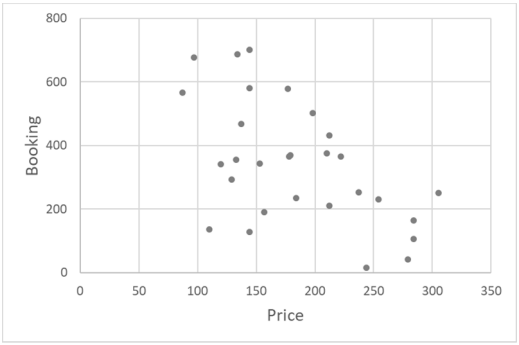
时间序列数据

- 在不同时间上收集到的数据，用于所描述现象随时间变化的情况
- 这类数据反映了某一事物、现象等随时间的变化状态或程度
  - 某一酒店在不同时期的销量



截面数据

- 不同主体在同一时间点或同一时间段的数据，也称静态数据
  - 某个时间一些酒店各自的销量



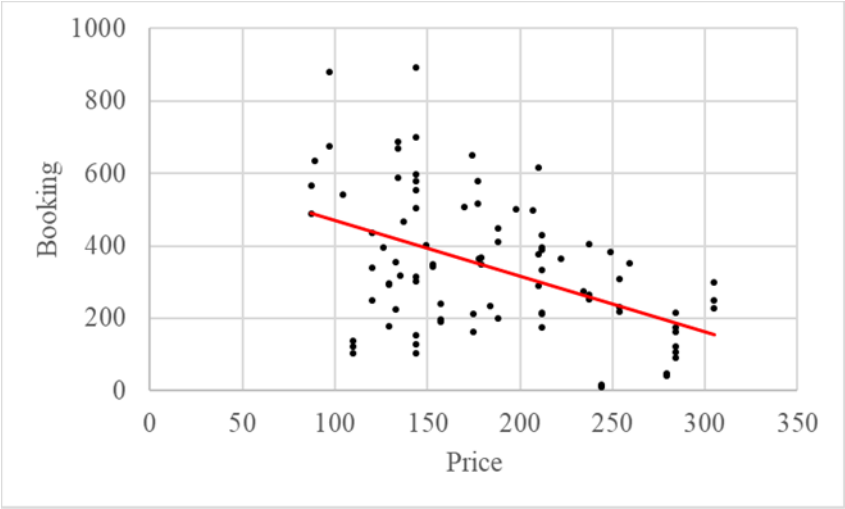
面板数据

- 面板数据是时间序列数据和截面数据的结合
- 是指在时间序列上取多个截面，在这些截面上同时选取样本观测值所构成的样本数据
  - 一些酒店在一段时间内各个时间点的销量

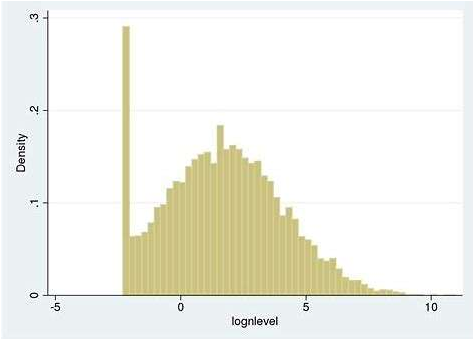
普通最小二乘法

OLS(Ordinary Least Squares)，通过最小化误差的平方和寻找数据的最佳函数匹配

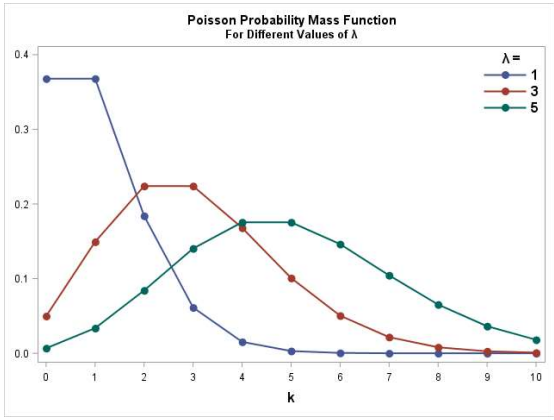
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$，其中$\varepsilon_i$称为随机扰动项
- 预测值$\hat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$，其中$\widehat{\beta_0}$、$\widehat{\beta_1}$均为参数的估计值
- 误差为$y_i - \hat{y_i}$，OLS的目标就是最小化误差平方和，即$min\{\sum_{i=1}^{n}(y_i - \hat{y_i})^2\}$
- OLS回归的五个假设：
  - 线性假设
  - 自变量外生，即$E[\varepsilon_i | \boldsymbol{x_i}] = 0$
  - 随机扰动项符合正态分布$\varepsilon_i \sim N(0, \sigma^2)$
  - 所有扰动项具有相同的方差（与自变量无关）且相互独立（与其它扰动项无关）
  - 观测结果的个数多于自变量的个数，同时不存在完全共线性



- Logistic（或Logit）回归
  - 当因变量y为0/1变量时（比如贷款是否违约），通常会设置一个潜变量y^*
  - 然后设置一个阈值，大于阈值则y=1，反之则为0
  - 因此，y^*就是因变量y是否为1的概率P(y=1|x)，即$0 \le y^* \le 1$
  - Logistic分布
    - 概率密度函数$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$，累积分布函数$F(x) = \frac{1}{1+e^{-x}}$
    - $y_i^* = \frac{1}{1+e^{-(z_i)}}$，其中$z_i = \boldsymbol{X_i\beta}$
- 极大似然估计（Maximum Likelihood Estimate）
  - 给定一堆数据，假如我们知道它是从某一种分布中随机取出来的，可是我们并不知道这个分布具体的参数，即“模型已定，参数未知”
  - MLE 的目标是找出一组参数，使得模型产生出观测数据的概率最大，理论上MLE可以估计任何形式的模型
  - 假设**y**是一维变量，θ是y服从的分布的参数，$f(y_1, y_2, \dots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = L(\theta|y)$似然函数，MLE就是求出使似然函数$L(\theta|y)$值最大的$\theta$
  - 对于有观测值的回归来说，就是使得$f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})$最大，其中$\boldsymbol{X}$为自变量，$\boldsymbol{\beta}$为待求解的一组参数（系数）



- Censored Data
  - Logit模型的因变量其实是一种被截断的分布
  - 另外一种常见的因变量被截断的情况是Tobit模型
  - Tobit模型本质上是普通线性回归的一种特殊情况
    - 因变量大致上连续并且符合正态分布，但是必须大于0，即有一部分0的理论值应当小于0
    - $y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i$，$y_i = \begin{cases} y_i^* & y_i > 0 \\ 0 & y_i \le 0 \end{cases}$
    - Tobit模型可以进一步推广，将阈值设定为任意数值，方向也可相反



- 计数模型
  - 当因变量非连续、且为自然数的时候，就需要用到计数模型
  - 最典型的计数模型为泊松（Poisson）模型
  - 泊松模型的因变量符合泊松分布$P(y=k)=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,\ldots$，其中$\ln\lambda = X\beta$

- 固定效应和随机效应
  - 当存在遗漏变量问题，且遗漏变量和已有变量相关，就可能会需要固定效应（Fixed effect）
  - $y_{i,t} = \boldsymbol{x_i\beta} + c_i + \varepsilon_{i,t}$，其中$c_i$表示遗漏变量的效应，$E[c_i|\boldsymbol{X_i}] = h(\boldsymbol{X_i})$
  - 如果遗漏变量和已有变量不相关，即$E[c_i|\boldsymbol{X_i}] = 0$，则是随机效应
  - 统计方法来判断是否存在固定效应
    - 用Hausman检验来比较使用固定效应得到的估计结果和使用随机效应得到的估计结果，如果存在显著的系统性差异，则使用固定效应

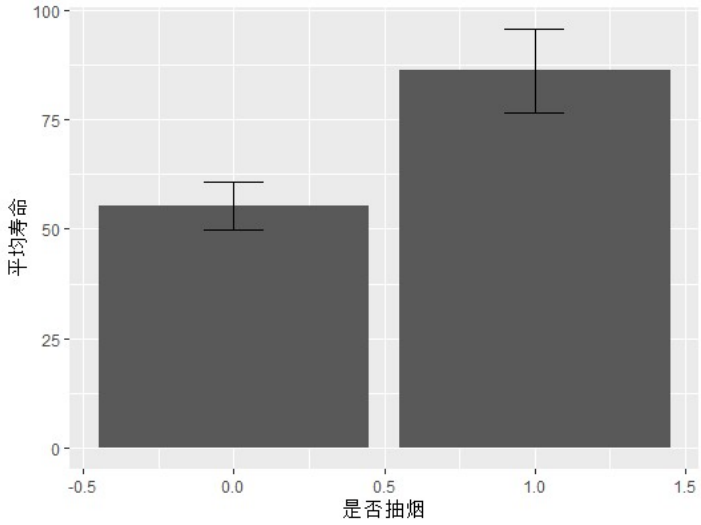
注意：对固定效应一个更简单（但是不完全准确）的理解是每个个体都有一个自己的截距项，且截距项之间的差异是系统性的

- 当我们面临一个复杂的非线性模型的时候
  - 比如$y = \alpha + \beta_1 e^{-\gamma_1 x_1} + \beta_2 x_1^2 + \beta_3 \sin(\gamma_2 x_2) + \beta_4 \cos(\gamma_3 x_1) \ln \gamma_4 x_2$
  - MLE的前提是连续可导，且求解过程非常复杂
  - 贝叶斯估计
    - 基本思想：后验概率最大化
    - $P(\text{参数}|\text{数据}) = \frac{P(\text{数据}|\text{参数})P(\text{参数})}{P(\text{数据})} \propto P(\text{数据}|\text{参数})P(\text{参数})$

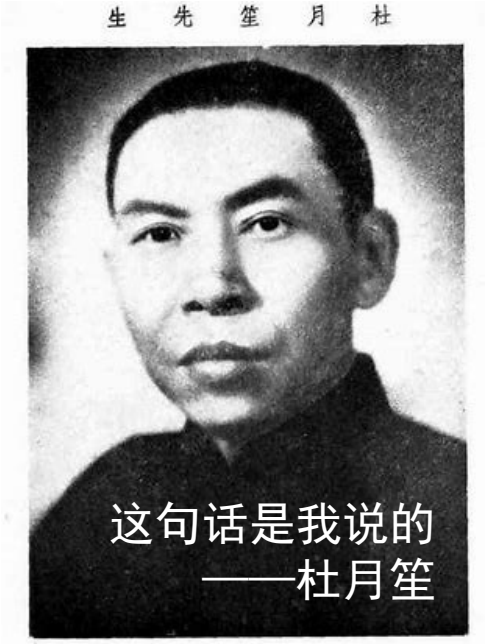
似然函数　　参数的先验概率

    - 根据先验概率自定义一组参数，计算后验概率
    - 随机生成一组参数的改变值，计算后验概率
    - 如果后验概率提升，则更新参数
    - 开始下一轮计算，反复迭代N轮
    - 根据最后M轮的参数值来计算参数的估计值

- 吸烟有害身体健康？
  - 吸烟：张学良(103)、纪晓岚(81)、梁实秋(84)、周作人(82)、丘吉尔(91)、爱因斯坦(76)
  - 不吸烟：列宁(54)、拿破仑(52)、希特勒(56)、杜甫(58)、苏轼(64)、曹雪芹(48)
  - *所以我们是否能得到吸烟有益身体健康的结论？*



不抽烟不喝酒的男人一般靠不住，不可托付终身



- 测量误差（Mearsurement Error Bias）
  - 信度（Reliability）
    - 即可靠性，它指的是采取同样的方法对同一对象重复进行测量时，其所得结果相一致的程度
  - 效度（Validity）
    - 即有效性，它是指测量工具或手段能够准确测出所需测量的事物的程度
  - 什么是健康？是否可靠、有效地对健康进行了测量？
  - 是否吸烟？中间有没有戒烟？吸了多久？

自选择偏差（Self-selection Bias/Sample-selection Bias）

降落伞的使用体验为什么都是好评？

- 因降落伞有问题而失事的人想给差评也给不了！

Heckit（Heckman两步法）

- 无偏的样本分成两部分：1）能观测到因变量的部分；2）无法观测到因变量的部分
  - 某个观测样本是否能够观测到因变量是一个选择过程
  - 直接对样本集1进行回归就会出现自选择偏差
- 寻找一些外生变量，使得我们能够使用Probit模型估计出每个观测样本进入到样本集1的概率
  - Probit模型和Logistic模型类似，都是因变量为0/1的模型，区别在于Probit模型的随机扰动项服从正态分布
  - 注意，这些外生变量可以部分和第二阶段回归的自变量重合，但是不能是第二阶段回归自变量的子集
- 然后计算每个观测样本的Inverse Miller Ratio，$IMR = {\phi(\hat{p})}/{\Phi(\hat{p})}$
  - ϕ表示正态分布的概率密度函数，Φ表示正态分布的累积分布函数，$\hat{p}$是第一阶段回归的因变量预测值
- 最后将*IMR*放到第二阶段回归中作为一个自变量

- 数据操纵
  - 刻意选择一些有利于证明观点的数据
    - 周作人的哥哥周树人55岁因病去世
- 遗漏变量误差（Omitted Variable）
  - 年代，生活条件，其它有害健康的习惯，基因等
- 联立性问题（Simultaneity）
  - 反向因果关系（Reverse Causality）
    - 研究发现平常不会被体罚的儿童比那些接受体罚的儿童平均智商高出了10个点，所以体罚降低智商？
  - 互为因果
    - 人以群分还是近朱者赤？
  - 相关而非因果
    - 游泳死亡人数越多，冰棍销量越高
      - 其实是因为温度升高
    - 有第三方因素同时影响了自变量和因变量

- 录取中的性别歧视
  - 谁在歧视？
  - Simpson's Paradox
    - 在某个条件下的两组数据，分别讨论时都会满足某种性质，可是一旦合并考虑，却可能导致相反的结论
  - 原因
    - 两组的录取率相差极大
    - 有另外的潜在因素影响
  - 解决办法（如果必须要聚合分析的话）
    - 恰当的设置各个组的权重消除影响

全校录取率统计			
	申请人数	录取人数	录取率
男性	120	25	21%
女性	120	50	42%
合计	240	75	31.3%

商学院录取率统计			
	申请人数	录取人数	录取率
男性	20	15	75%
女性	100	49	49%
合计	120	64	53.3%

法学院录取率统计			
	申请人数	录取人数	录取率
男性	100	10	10%
女性	20	1	5%
合计	120	11	9.2%

Simpson, E. H. (1951), "The Interpretation of Interaction in Contingency Tables," Journal of the Royal Statistical Society, 13, 238-241.