小红书

REDtech

分享主题：实用型因果推断方法在互联网中的实践

分享人：李少斌

欢迎关注
小红书技术 REDtech

目录

欢迎关注
小红书技术 REDtech

Insight Vs Science

- **Insight**
- 是指通过观察、分析、经验、直觉等方式，获得对某个问题、现象、情况或事物本质的深入理解和领悟。

- **Science**
- 是科学是一种基于实证和逻辑推理的知识体系，以系统化、规范化和可重复性的方式来研究自然现象、社会现象和人类思维等方面的知识。

---- From ChatGPT

欢迎关注
小红书技术 REDtech

从新用户留存分析看Insight Vs Science

Question：如何提升小红书新用户的留存率？

- **Insight**
- 访问过美妆品类的用户留存率高
- 访问类目数越多留存率越高
- 有内流播放的用户留存率高

Data

- **Science**
- ABtest
- 匹配
- PSM\PSM-DID
- DML\DRL\……

欢迎关注
小红书技术 REDtech

仅靠Insight和AB-test存在的问题

- **相关 Vs 因果**

相关性 ≠ 因果性

访问美妆留存率高 ？ 高留存的用户群体访问了美妆 or 美妆作品提升了用户留存率

- **预测 Vs 决策**

一个用户访问了美妆、访问了10个类目、使用了内流播放 -高概率留存用户

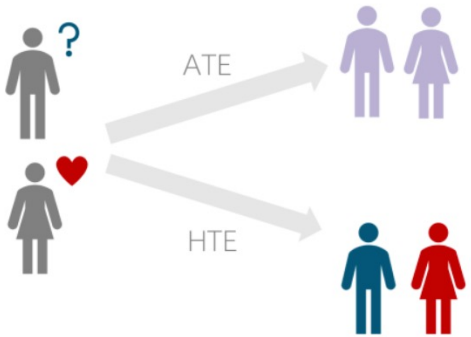
要提升用户留存率:增加美妆作品曝光占比？内容多样性？内流功能入口前置？ -？？？

- **ATE Vs HTE**

Average Treatment Effect

Heterogeneous Treatment Effect

增加美妆作品曝光占比？



欢迎关注
小红书技术 REDtech

目录

欢迎关注
小红书技术 REDtech

解决因果问题的科学框架\流派

REDtech

- Judea Pearl
- **Computer Scientist**
- Causal Graph Model、Backdoor Criterion、Frontdoor Criterion、Do-calculus、 Pearl Causal Hierarchy (Association,Intervention,Counterfactuals)

- Joshua D.Angrist
- **Economist**
- Double machine learning、Instrumental Variables、Panel Data and Fixed Effects、 Regression Discontinuity Design、2SLS

- Donald B.Rubin
- **Statistician**
- Potential Outcome Model\ Rubin Causal Model\ IPW\ABtest

欢迎关注
小红书技术 REDtech

解决因果问题的科学框架\流派

- Judea Pearl
- **Computer Scientist**

- Joshua D.Angrist
- **Economist**

- Donald B.Rubin
- **Statistician**
- Potential Outcome Model\ Rubin Causal Model\ IPW\ABtest

欢迎关注
小红书技术 REDtech

计算机科学（Causal Graph Model、DAG）

REDtech

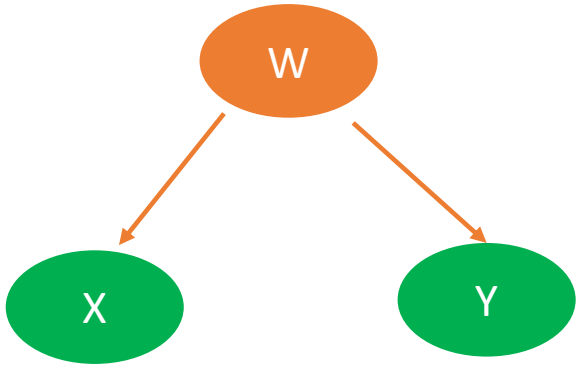
Chain

X ⫫ Y

X ⊥ Y | Z

X、Y 既有相关性也有因果性

Fork



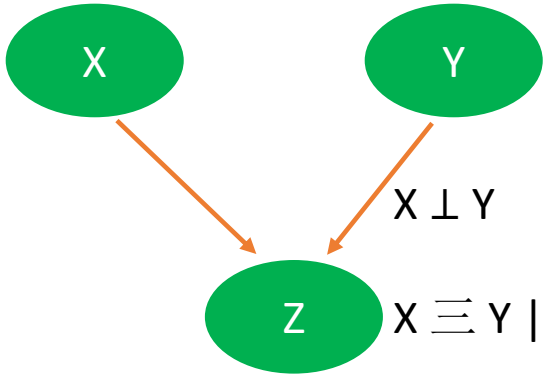
X ⫫ Y

X ⊥ Y | W

X、Y 有相关性但无因果性

示例：W为天气，X为溺水率，Y为冰激凌销量

Collide



X ⊥ Y

X ⫫ Y | Z

X、Y不相关，但在condition z的情况下，X、Y相关

select bias

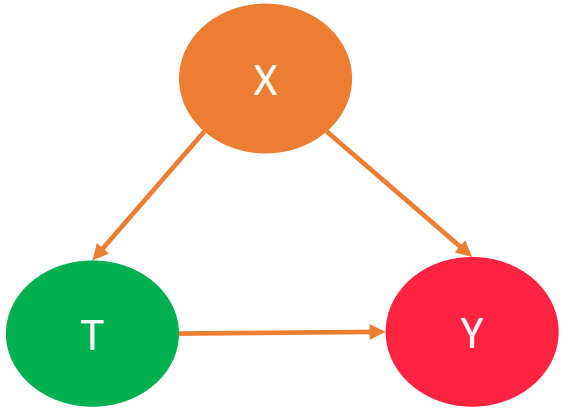
欢迎关注
小红书技术 REDtech

计算机科学（Causal Graph Model、DAG 与 ABtest）

Question：如何提升小红书新用户的留存率？

Data

- **DAG**

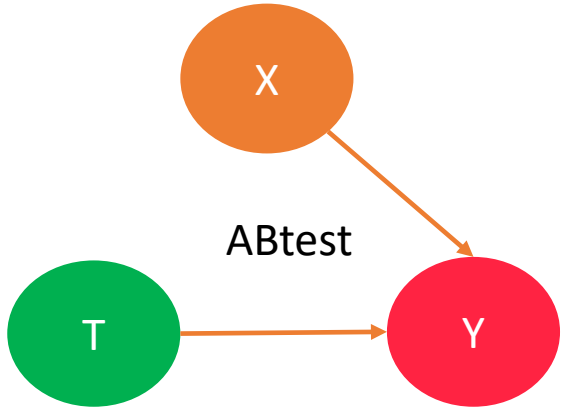


X：性别

T：是否进入内流

Y:留存率

- **ABtest**



μ :X→Y的留存率, θ :进入内流影响

T1实验组留存率： μ + θ

T0对照组留存率： μ

ATE = T1-T0 = θ

欢迎关注
小红书技术 REDtech

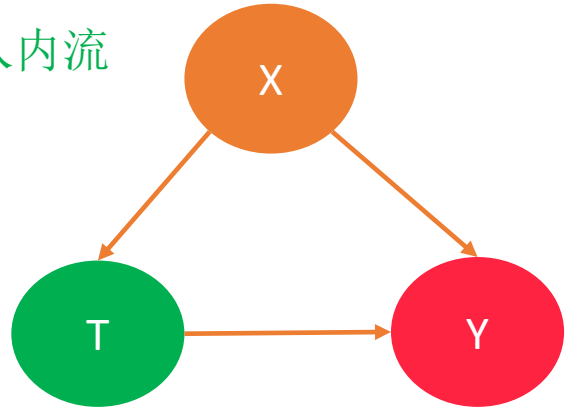
计算机科学（Do-Calculus\ Backdoor Criterion ）

X：性别

T：是否进入内流

Y:留存率

- 干预前



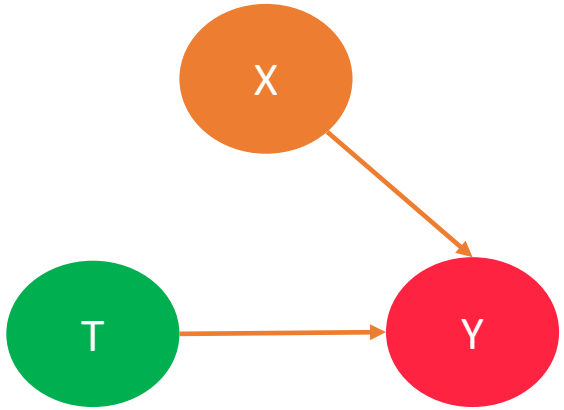
Do-calculus推导：

1.$P(Y=y|do(T=t)) = P_m(Y=y|T=t)$ (definition)

2.$P_m(Y=y|X=x,T=t) = P(Y=y|X=x,T=t)$

3.$P_m(X=x)=P(X=x)$

- 干预后



T=t

$P(Y=y|do(T=t)) = P_m(Y=y|T=t)$

$= \sum_x P_m(Y=y,X=x|T=t)$ 贝叶斯全概率公式

$= \sum_x P_m(Y=y|T=t,X=x)P_m(X=x|T=t)$ 条件概率

$= \sum_x P_m(Y=y|T=t,X=x)P_m(X=x)$

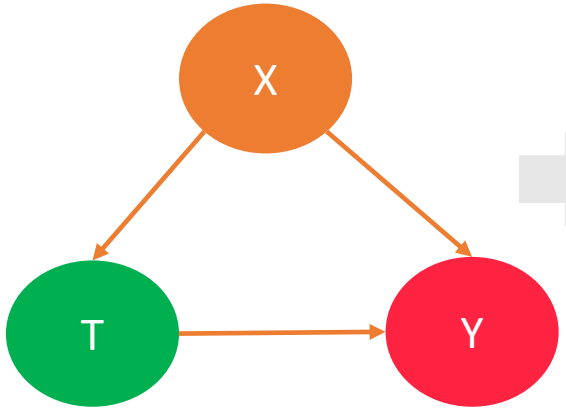
$= \sum_x P(Y=y|T=t,X=x)P(X=x)$ 观测数据获取因果

计算机科学（Causal Graph Model、DAG\Do-Calculus）

Question：如何提升小红书新用户的留存率？

Data

- **DAG**



X：性别

T：是否进入内流

Y:留存率

- **Do-Calculus**

性别	进入内流(T=1)		未进入内流(T=0)	
	留存数	留存率	留存数	留存率
女(X=1)	81（87）	93%	234（270）	87%
男(X=0)	192（263）	73%	55（80）	69%
合计	273（350）	78%	289（350）	83%

$P(Y=1|do(T=1)) = P(Y=1|T=1,X=1)P(X=1)+P(Y=1|T=1,X=0)P(X=0)$

$= 0.93*(87+270)/700+0.73*(263+80)/700$

$= 0.832$

$P(Y=1|do(T=0)) = P(Y=1|T=0,X=1)P(X=1)+P(Y=1|T=0,X=0)P(X=0)$

$= 0.87*(87+270)/700+0.69*(263+80)/700$

$= 0.7818$

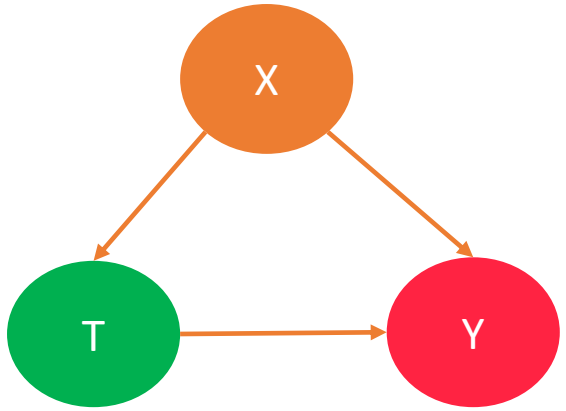
示例数据，与真实业务无关

$P(Y=1|do(T=1))-P(Y=1|do(T=0)) = 0.0502$　$0.0502 > 0$　进入内流有效，会使留存率上升5个百分点

欢迎关注
小红书技术 REDtech

计量经济学（Select Bias）

- **DAG**



X：性别

T：是否进入内流

Y:留存率

- **Select Bias**

$$\text{潜在结果} = \begin{cases} Y_{1i} \text{ if } T_i = 1 \\ Y_{0i} \text{ if } T_i = 0 \end{cases}$$

观察结果Y_i = $Y_{0i} + (Y_{1i}-Y_{0i})T_i$ --Yi潜在结果的线性组合

$$E[Y_i|T_i = 1]-E[Y_i|T_i = 0] = \underbrace{E[Y_{1i} \mid T_i = 1] - E[Y_{0i} \mid T_i = 1]}_{\text{处理的平均因果效应}} + \underbrace{E[Y_{0i} \mid T_i = 1] - E[Y_{0i} \mid T_i = 0]}_{\text{选择性偏误}}$$

如果选择性偏误的绝对值可能会很大，可能会影响我们相要寻找的**因果关系符号**！

随机实验的情况下Y_{0i} 和T_i 之间独立，$E[Y_{0i} \mid T_i = 1] = E[Y_{0i} \mid T_i = 0]$