What is Simpson's paradox?

> “ If an experiment goes through ramp-up that is, two or more periods with different percentages assigned to the variants, combining the results can result in directionally incorrect estimates of the Treatment effects, that is, Treatment may be better than Control in the first phase and in the second phase, but worse overall when the two periods are combined. This phenomenon is called Simpson’s paradox” - *Crook et al*

> *A trend or result that is present when data is put into groups that reverses or disappears when the data is combined.*

辛普森悖论指如果我们将分组后的数据进行合并，则之前得出的数据结果会呈现相反的趋势。换句话说，作为数据分析师，我们可能会用某个特征对数据进行分组聚合计算相关数据指标，这时我们不可以根据增加了详细级别后的分组数据贸然做决定，需要考虑总体数据指标。

How Simpson's Paradox Works?

Now I will list some examples to demonstrate how Simpson paradox impacts the data decision-making.

Example 1: Online Controlled Experiment

下表显示一个网站在周五和周六两天每天都有1百万的访问者。然后实验者开始实施A/B Testing，并进行随机抽样，将抽取到的访问者随机分配给Two Variants, *Treatment and Control*。在周五的时候，实验者将1百万的流量中1%分配给Treatment，而在周日的时候，实验者决定5/5划分，即Treatment里有50%的流量。这两天的实验结果也被公布：周五Treatment和Control的转化率为分别为2.3%和2.02%，周六的时候分别为1.2%和1%。

Note: “Conversion Rate for two days. Each day has 1M customers, and the Treatment (T) is better than Control (C) on each day

		Firday	Saturday
Control	Conversion	20,000	5,000
	Sample Size	990,000	500,000
	Conversion Rate	$\frac{20{,}000}{990{,}000} = 2.02\%$	$\frac{5{,}000}{500{,}000} = 1\%$
Treatment	Conversion	230	6,000
	Sample Size	10,000	500,000
	Conversion Rate	$\frac{230}{10{,}000} = 2.3\%$	$\frac{6{,}000}{500{,}000} = 1.2\%$

如果我们不加思考，根据转化率会直接得出结论: $\text{Treatment is better than Control}$

主要原因是无论是哪一天，Treatment的转化率都要高于Control。这样看起来Treatment似乎效果很好。但是现在我们将数据进行总计，求出周五和周六的两个组别的总转化率进行对比。如下表：

		Firday	Saturday	Total
Control	Conversion	20,000	5,000	25,000
	Sample Size	990,000	500,000	1,490,000
	Conversion Rate	$\frac{20{,}000}{990{,}000} = 2.02\%$	$\frac{5{,}000}{500{,}000} = 1\%$	$\frac{25{,}000}{1{,}490{,}000} = 1.68\%$
Treatment	Conversion	230	6,000	6,230
	Sample Size	10,000	500,000	510,000
	Conversion Rate	$\frac{230}{10{,}000} = 2.3\%$	$\frac{6{,}000}{500{,}000} = 1.2\%$	$\frac{6{,}230}{510{,}000} = 1.2\%$

数据显示，总计后的Control Group的总转化率反超Treatment Group的总转化率。前面*Treatment is better than control*的结论在这里直接被推翻。而这个现象就被称为***Simpson's Paradox.***

我们现在开始研究其背后的原因。Treatment组的周六的转化影响对Treatment总体转化率的影响要更大，因为周六的时候Treatment Users数量更多。这种情况下，转化率低的这一天影响更大，自然总体转化率更低。

Example 2: UC Berkley’s suspected gender-bias

UC伯克利分校1973年研究生院的录取数据显示"女生的录取率为35%，而男生的录取率为44%"，所以有人得出结论认为**学校录取存在性别歧视。**为了确认是否存在性别歧视，按照学校学院分组的录取数据被公布如下：

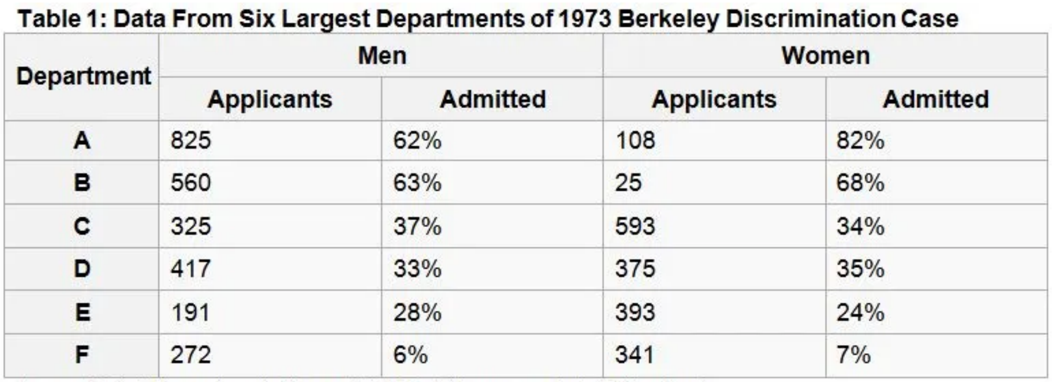


Table 1: Data From Six Largest Departments of 1973 Berkeley Discrimination Case

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Source: Bickel, Hammel, and O'Connell (1975); table accessed via Wikipedia at https://en.wikipedia.org/wiki/Simpson%27s_paradox.

从上方数据可以看出，A学院、B学院、D学院以及F学院中女生的录取率都要高于男生的录取率，其中A学院和B学院的录取率最高但是申请人数也最少，而与之相比其他学院录取率低并且申请人数更多。所以造成总体女生录取率低的原因不是**"Gender Discrimination"**，而是**大量的女生都去申请本身录取比率就很低的学院，而只有少量的女生去申请了录取率高更容易进的学院，所以导致总体录取的女生人数更少，降低了总体的录取率。**

Example 3: Strawberry Vs Peach

假设我们是一家软饮料公司。现在有两种味道的饮料草莓味和桃子味，现在需要决定到底投入哪一款进入市场。对两种味道都搜集了1000个样本，经过市场调查后得出以下数据：

Flavour	Like	Not Like	Rate
Strawberry	800	200	80%
Peach	750	250	75%

数据显示，人群中喜欢草莓的总比例高于喜欢桃子的比例，所以我们应该投放草莓味的饮料进入市场。

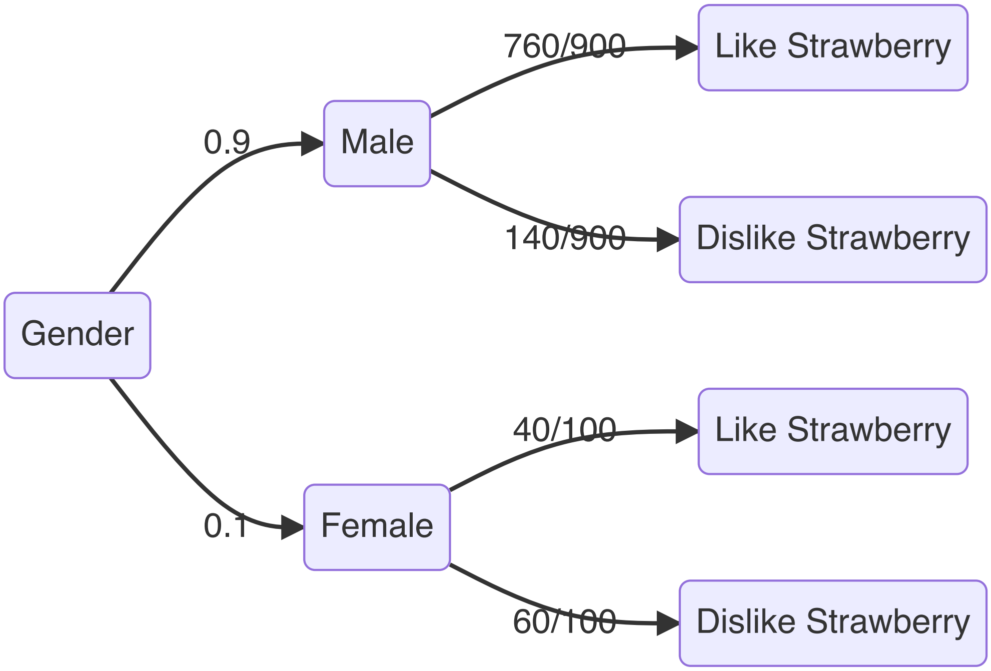
而现在市场部门又获取了被调查者的性别信息，从性别的维度对数据进行进一步的细分，数据如下：

Gender	Flavour	Like	Not Like	Total
Male	Strawberry	760	140	900 (84.4%)
Female	Strawberry	40	60	100 (40%)
Male	Peach	600	100	700 (85.71%)
Female	Peach	150	150	300 (50%)

基于最新的数据，无论是男性还是女性，都是喜欢桃子味道的比例要高于喜欢草莓味道的比例，得出了截然相反的结论。所以市场人员正在遭受**"辛普森悖论。"**

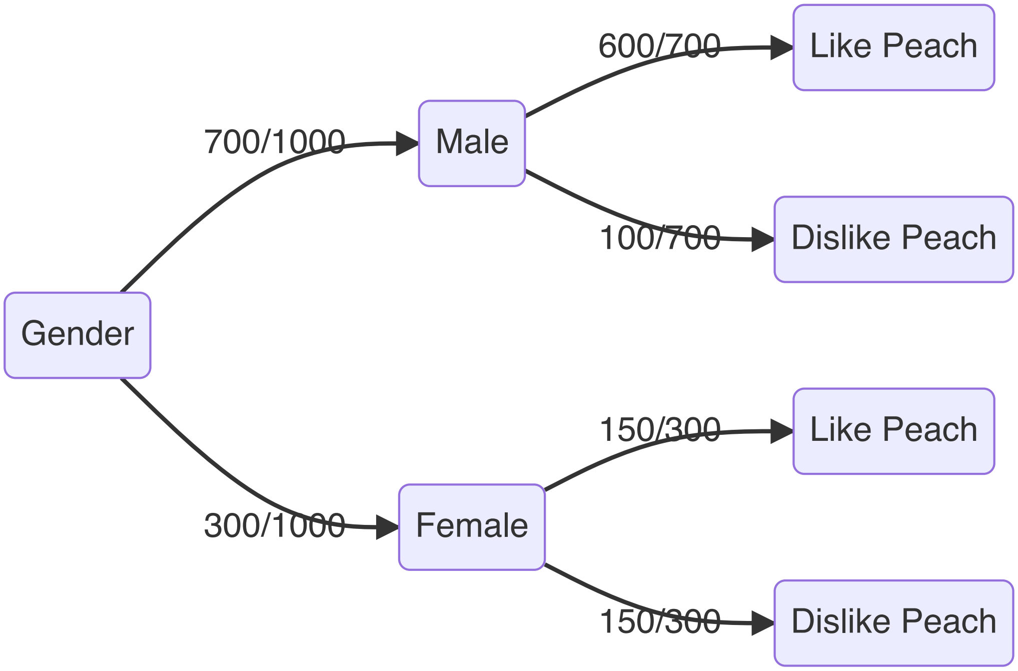
Conclusion:

Simpson's paradox arises when certain variables segment populations into segments. Such variables are referred to as **lurking variable,** and they are often difficult to be identified. We can explain the case of soft drink by Probabilities:



$$P(\text{Strawberry}) = P(\text{Male})P(\text{Strawberry}|\text{Male}) + P(\text{Female})P(\text{Strawberry}|\text{Female})$$
$$= \frac{900}{1000} \times \frac{760}{900} + \frac{100}{1000} \times \frac{40}{100}$$
$$= 0.8$$

Similarly, we can calculate the total probability for peach



$$
\begin{aligned}
P(\text{Peach}) &= P(\text{Male})P(\text{Peach}|\text{Male}) + P(\text{Female})P(\text{Peach}|\text{Female}) \\
&= \frac{700}{1000} \times \frac{600}{700} + \frac{300}{1000} \times \frac{150}{300} \\
&= 0.75
\end{aligned}
$$

We can consider $P(Male)\&P(Female)$ as weights that make the total probability shift towards the side of male. Hence, the marginal probability of Strawberry is higher than that of Peach in general.

Other Examples of Simpson's paradox in Online Controlled Experiment

- Users are sampled. Because there is concern about getting a representative sample from all browser types, the sampling is not uniform, and users in some browsers (such as, Opera or Firefox) are sampled at higher rates. It is possible that the overall results will show that the Treatment is better, but once the users are segmented into the browser types, the Treatment is worse for all browser types.
- An experiment is run at 50/50% for Control/Treatment, but an advocate for the most valuable customers (say top 1% in spending) is concerned and convinces the business that this customer segment be kept stable and only 1% participate in the experiment. Similar to the example above, it is possible that the experiment will be positive overall, yet it will be worse.
- An upgrade of the website is done for customers in data center DC1 and customer satisfaction improves. A second upgrade is done for customers in data center DC2, and customer satisfaction there also improves. It is possible that the auditors looking at the combined data from the upgrade will see that overall customer satisfaction decreased.

> “While occurrences of Simpson’s paradox are unintuitive, they are not uncommon. We have seen them happen multiple times in real experiments (Xu, Chen and Fernandez et al. 2015, Kohavi and Longbotham 2010). One must be careful when aggregating data collected at different percentages.”