A / B Testing

AB测试

Correlation vs. Causation

相关性和因果关系

According to Financial Times, HKU Global MBA graduates, on average, make a salary of $129,149 per year three years after graduation.

By contrast, Harvard Global MBA graduates, on average, make a salary of $246,509 per year three years after graduation.

Question: Can we claim that Harvard helps MBA students make more money?

根据英国 Financial Times 报道, 香港大学国际 MBA 毕业生毕业三年后的平均收入是 $129,149 万美元每年.

与之相对的是, 哈佛大学国际 MBA 项目毕业生毕业三年后的平均收入水平是 $246,509 万美元每年.

问题: 与香港大学相比，哈佛大学是不是更能帮学生赚钱?

According to data analysis, we find that those who take marketing analytics class perform better in their job. Does this mean marketing analytics makes people perform better?

根据数据分析，我们发现参加消费者分析课程的人在工作中表现更好。这是否意味着消费者分析可以提高大家的工作表现？

We call the above problem
"omitted variable bias."

我们把上述现象称为“遗漏变量偏差”

When analyzing data, we find that when a bank offers a lower interest rate, more consumers are willing to deposit money at the bank. Does this mean we should offer lower interest rate to encourage more consumers to deposit their money?

在分析数据时，我们发现当银行提供较低的利率时，更多消费者愿意在银行存款。这是否意味着我们应该提供较低的利率，以鼓励更多消费者存款?

Some people analyze the relationship between the size of police force (e.g., number of police per capita) and the local crime rate, and find out that when there are more police force, the crime rate is also higher.

Does this mean police officers are trouble-makers who make the city more dangerous?

有人研究了警察数量(例如平均每千人对应的警察人数)和当地犯罪率之前的关系,发现警力越充足的地方，反而当地犯罪率越高。

这是不是说明警察会制造犯罪，提升当地犯罪率?

We call the above problem “reserved causality.”

我们称之为“反向因果关系”

Suppose that you want to see whether Harvard is better than HKU, what would you do?

Suppose that you want to see how smartphone sales change with price, what would you do?

Suppose that you want to see how police force affects local crime rates, what would you do?

假如你想知道哈佛大学是不是比香港大学更好，你应该怎么做?

假如你想看看手机销量是如何受到价格影响的，你应该怎么做?

假如你想知道存款利率如何影响存款意愿，你应该怎么做?

Solution: A/B tests

解决方案：AB测试

Intuition for the Harvard vs. HKU example:

Recall that we cannot direct compare Harvard graduates' salary with HKU graduates' salary because the students are so different. How about making the student background very similar?

How to achieve that? We can use random assignment.

我们再去考虑哈佛 vs. 港大的例子:

我们刚刚讨论到，我们不应该直接对比哈佛和香港大学毕业生的收入，因为他们面对的学生是不同的。我们可以考虑让这两个学校的学生更加类似。

怎么做到这点呢？我们考虑随机分配。

How to achieve that? We can use random assignment.

Suppose that there are 10,000 students applying to Harvard or HKU MBA program. Then, we randomly admit 100 to Harvard, and randomly admit another 100 to HKU. So, whether you are admitted to Harvard only depends on your luck, not your age, ability, IQ, talent, family background...

If there is difference in salary, the difference can only be driven by school education.

怎么做到这点呢？我们考虑随机分配。

假设有 10,000 个申请哈佛或者香港大学 MBA 项目的学生. 我们 随机 将其中 100 个人录取到哈佛大学, 并 随机 将另外 100 个人录取到香港大学。这样，你被录取到哪所学校以及你录取与否只取决于你的运气，而跟你的年龄，能力，智商，天赋，家庭条件等等都没有任何的关系。

如果这两组学生的收入不同，这个差异只能用教育来解释。

This is the basic idea of AB testing.

When we want to compare two (or more) conditions to see which one works better, we can randomly assign participants into two (or more) groups, namely group A and group B. Since there are no other differences between the two groups, any difference in the outcome is driven by the difference in the conditions.

这就是AB测试的基本思路。

当我们需要比较两个以上的选项时，我们可以把受试者随机分配到不同的分组中，例如A组和B组。因为这两组随机分配，他们之间应该没有任何本质的不同。唯一的不同就是这两组的条件不尽相同。

The key for successful A/B testing is random assignment. You must make sure that people in group A and group B are similar enough, ruling out other potential causes of the effects. AB testing is the gold standard for finding causal relationship. It is commonly adopted by big tech firms.

AB测试成功的关键就在于随机分组。你必须保证A组的成员和B组的成员足够相似，这样可以排除其他潜在的原因。AB测试时发现因果关系的黄金标准，也被科技公司大量使用。

这是某个保险公司的页面，猜猜哪个页面效果更好？

AB测试成功的关键就在于随机分组。你必须保证A组的成员和B组的成员足够相似，这样可以排除其他潜在的原因。AB测试时发现因果关系的黄金标准，也被科技公司大量使用。

页面 B 的效果更好！

Why is randomization important?

Consider an example in which you assign male MBAs to Harvard and female MBAs to HKU. If one university performs better, you don't know whether this is caused by gender difference or by difference in the schools.

为什么随机分配这么重要?

假设我们没有随机分配，而是把男性MBA分配到哈佛大学，而把女性MBA分配到香港大学。如果哈佛大学或者香港大学的学生收入更高，我们也不能确定这是性别偏见导致的还是大学教育导致的。

Example of A/B test: Speed matters.

"The dangers of a slow web site: frustrated users, negative brand perception, increased operating expenses, and loss of revenue."

——Steve Souders

AB测试的例子：速度的重要性

“网站网速过慢的危险: 兴致索然的用户，负面品牌形象，更高的运营成本，丢失的收入。”

——谷歌专家 Steve Souders

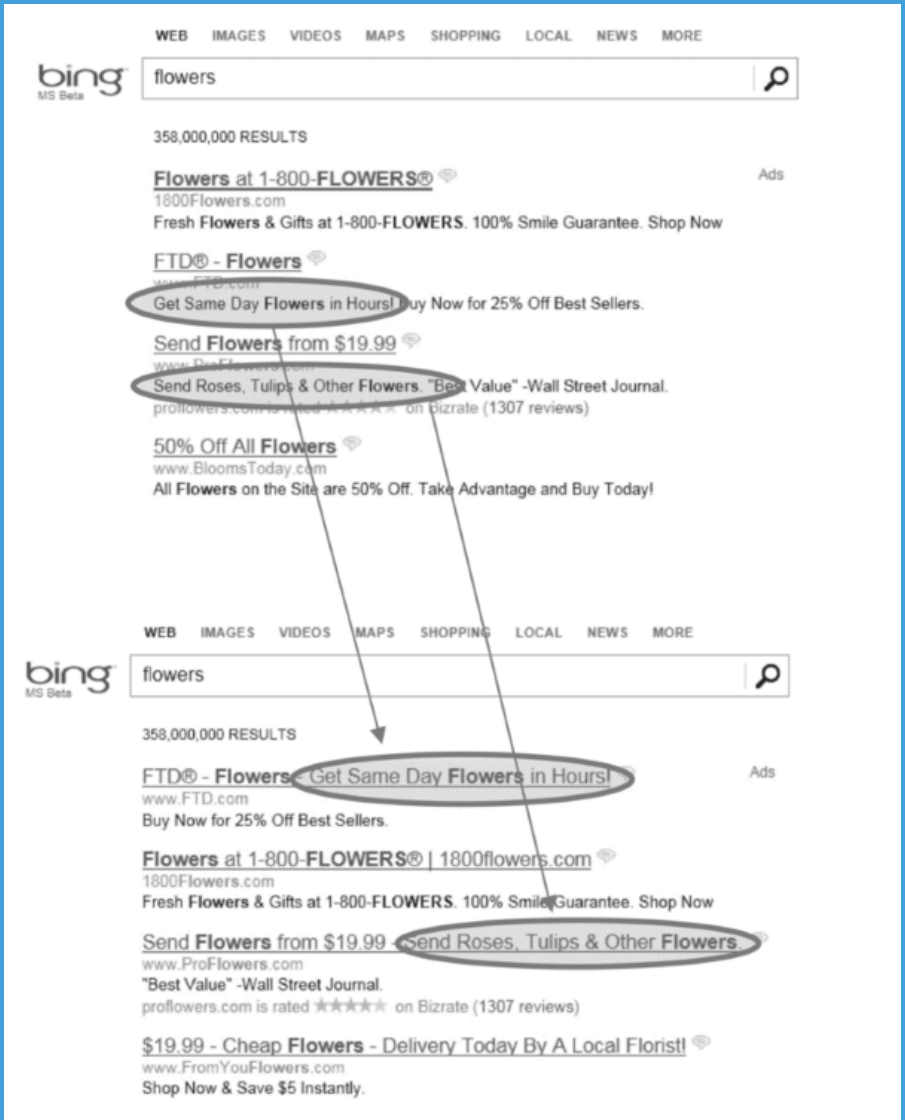
Example of A/B test: Speed matters.

Of course, faster is better, but how important is it to improve performance by 0.1 second? Should you have a person focused on performance? Maybe a team of five? The return-on-investment (ROI) of such efforts can be quantified by running a simple experiment.

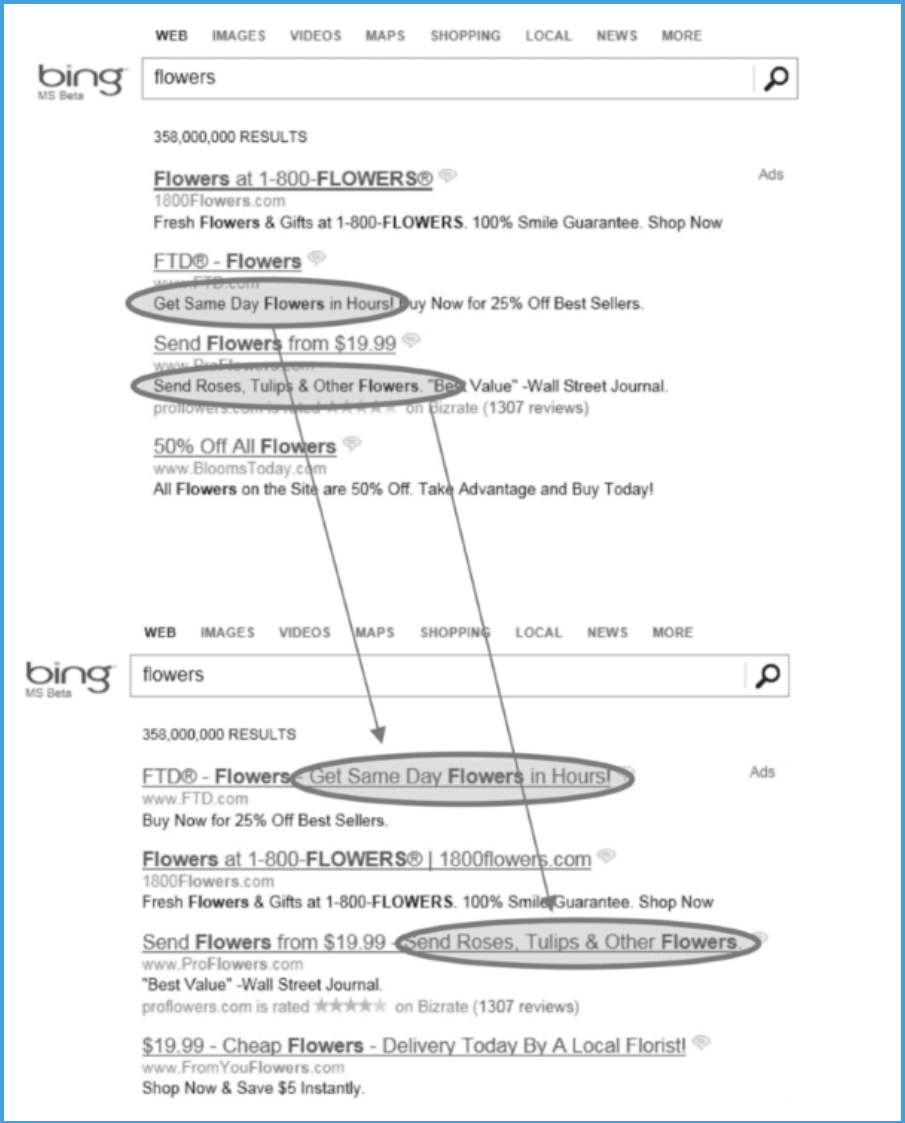
AB测试的例子：速度的重要性

网速当然是越快越好，但是把网速提高0.1秒到底能带来多大的收益呢？我们需要一个专家还是一个五人团队来帮助我们提高网速？为了回答这个问题，我们需要一个简单的实验来测试网速和收益之间的关系。

Example of A/B test



AB 测试的例子


WEB IMAGES VIDEOS MAPS SHOPPING LOCAL NEWS MORE
bing MS Beta
flowers
358,000,000 RESULTS
Flowers at 1-800-FLOWERS®
Ads
1800Flowers.com
Fresh Flowers & Gifts at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now
FTD® - Flowers
Get Same Day Flowers in Hours! Buy Now for 25% Off Best Sellers.
Send Flowers from $19.99
Send Roses, Tulips & Other Flowers. "Best Value" -Wall Street Journal.
on Bizrate (1307 reviews)
50% Off All Flowers
www.BloomsToday.com
All Flowers on the Site are 50% Off. Take Advantage and Buy Today!
WEB IMAGES VIDEOS MAPS SHOPPING LOCAL NEWS MORE
bing MS Beta
flowers
358,000,000 RESULTS
FTD® - Flowers - Get Same Day Flowers in Hours!
Ads
www.FTD.com
Buy Now for 25% Off Best Sellers.
Flowers at 1-800-FLOWERS® | 1800flowers.com
1800Flowers.com
Fresh Flowers & Gifts at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now
Send Flowers from $19.99 - Send Roses, Tulips & Other Flowers.
www.ProFlowers.com
"Best Value" -Wall Street Journal.
proflowers.com is rated on Bizrate (1307 reviews)
$19.99 - Cheap Flowers - Delivery Today By A Local Florist!
www.FromYouFlowers.com
Shop Now & Save $5 Instantly.

Example of A/B test

Nobody thought this simple change, among the hundreds suggested, would be the best revenue-generating idea in Bing's history! The feature was prioritized low and languished in the backlog for more than six months until a software developer decided to try the change, given how easy it was to code. An engineer implemented the idea and began evaluating the idea on real users, randomly showing some of them the new title layout and others the old one.

AB 测试的例子

没有人认为在数百个建议中，这个简单的改变会成为Bing历史上最佳的盈利点子！这个功能被优先级排得很低，在积压工作中被搁置了六个多月，直到一位软件开发人员决定尝试这个改变，因为编码非常容易。一位工程师实施了这个想法，并开始在真实用户中评估这个想法，随机向一些用户展示新的标题布局，而向其他用户展示旧的标题布局。

Example of A/B test

A few hours after starting the test, a revenue-too-high alert triggered, indicating that something was wrong with the experiment. The Treatment, that is, the new title layout, was generating too much money from ads.

Bing's revenue increased by a whopping 12%, which at the time translated to over $100M annually in the US, without hurting key user-experience metrics. The experiment was replicated multiple times over a long period.

AB 测试的例子

测试开始几个小时后，一个收入过高的警报被触发，表明实验出现了问题。实验组，即新的标题布局，从广告中赚取了太多的钱。

Bing 的收入惊人地增加了12%，这当时在美国每年转化为超过1亿美元，而不会损害关键的用户体验指标。这个实验在长时间内多次复制。

Example of A/B test

Amazon placed a credit-card offer on the home page. It was highly profitable but had a very low click-through rate (CTR). What would you do to make it more effective?

AB 测试的例子

亚马逊在主页上放置了一项信用卡优惠。虽然利润很高，但点击率（CTR）非常低。你会怎么做才能提高其效果?

Example of A/B test

The controlled experiment demonstrated that this simple change increased Amazon's annual profit by tens of millions of dollars.

对照实验表明，这一简单的变化使亚马逊的年利润增加了数千万美元。

Click-Through Rates

点击率

Suppose that we want to test the effectiveness of two banner ads:

A: **Enjoy 15% for your car insurance!**

B: **Last-minute deals for your car insurance!**

Our outcome is whether a user clicks through with ad A versus ad B. How do we tell if one ad is more effective than the other?

假设我们想测试两个横幅广告的有效性：

A: 享受您的汽车保险15%优惠！

B: 汽车保险特惠马上截止！

我们可以观察到用户是否点击了广告A或广告B。我们如何判断哪个广告更有效呢？

Suppose that:

45 out of 856 [5.25%] users clicked through on ad A;
99 out of 1,298 [7.62%] users clicked through on ad B.

Can you say that ad B is more effective than ad A?

假设：

856名用户中有45人点击了广告A [5.25%]；
1298名用户中有99人点击了广告B [7.62%]。

你能说广告B比广告A更有效吗？

Suppose that:

45 out of 856 [5.25%] users clicked through on ad A;
71 out of 1,298 [5.47%] users clicked through on ad B.

Can you say that ad B is more effective than ad A?

假设：

856名用户中有45人点击了广告A [5.25%]；
1298名用户中有71人点击了广告B [5.47%]。

你能说广告B比广告A更有效吗？

Suppose that:

450 out of 8,560 [5.25%] users clicked through on ad A;
710 out of 12,980 [5.47%] users clicked through on ad B.

Can you say that ad B is more effective than ad A?

假设：

8,560名用户中有450人点击了广告A [5.25%]；
12,980名用户中有710人点击了广告B [5.47%]。

你能说广告B比广告A更有效吗？

The χ-Squared Test

```
1 library(readr)
2 mydata = read_csv("https://ximarketing.github.io/data/AB.csv")
3 head(mydata)
4 table(mydata$treated, mydata$CTR)
```

Treated: Which ad consumers are exposed to.

	No	Yes
A	1511	489
B	1415	585

Among consumers who saw ad A, 489 clicked through and 1,511 did not click. Among consumers who saw ad B, 585 clicked through and 1,415 did not click.

It seems that ad B is more effective than ad A.

卡方检验

```
1 library(readr)
2 mydata = read_csv("https://ximarketing.github.io/data/AB.csv")
3 head(mydata)
4 table(mydata$treated, mydata$CTR)
```

实验内容: 消费者看到的是哪一种广告.

	No	Yes
A	1511	489
B	1415	585

在看到广告A的消费者中, 489个人点击了广告，1,511个人没有点击广告. 在看到广告B的消费者中, 585个人点击了广告，1,415个人没有点击广告.

似乎广告B比广告A有效一点。

The χ-Squared Test

```
1 chisq.test(mydata$treated, mydata$CTR)
```

```
> chisq.test(mydata$treated, mydata$CTR)

        Pearson's Chi-squared test with Yates' continuity correction

data:  mydata$treated and mydata$CTR
X-squared = 11.488, df = 1, p-value = 0.0007006
```

Here, we focus on the p-value. Typically, when $p < 0.05$, we claim the two conditions lead to significantly different outcomes; and in our case, $p < 0.001$, meaning that ad B is more effective than ad A.

卡方检验

```
1 chisq.test(mydata$treated, mydata$CTR)
```

```
> chisq.test(mydata$treated, mydata$CTR)

        Pearson's Chi-squared test with Yates' continuity correction

data:  mydata$treated and mydata$CTR
X-squared = 11.488, df = 1, p-value = 0.0007006
```

这里，我们重点关注 p 值. 一般而言, 当 $p < 0.05$, 我们认为这两组是显著不同的; 在我们的例子中, $p < 0.001$, 这说明广告 B 比广告 A 更加有效.

The complete code is here.

```
1 library(readr)
2 mydata = read_csv("https://ximarketing.github.io/data/AB.csv")
3 head(mydata)
4 table(mydata$treated, mydata$CTR)
5 chisq.test(mydata$treated, mydata$CTR)
```

Revenue
收入

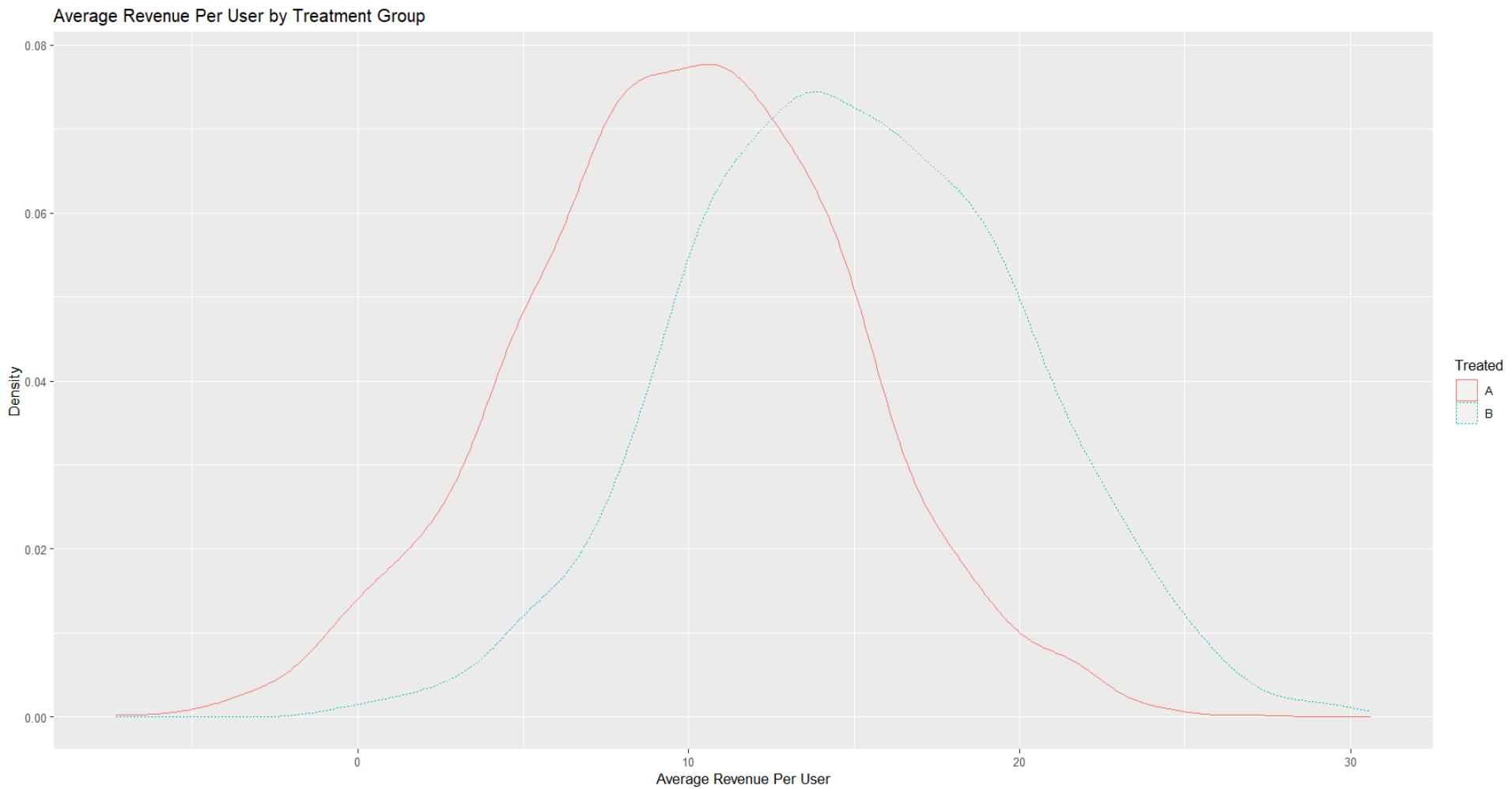
Next, we compare the revenue per consumer in the two conditions using the same dataset.

We first visualize the distribution.

```
library(ggplot2)
ggplot(mydata,aes(x=revenue, color =treated))+
  geom_density(aes(linetype=treated))+
  labs(title="Average Revenue Per User by Treatment Group",
       x="Average Revenue Per User",
       y="Density", color ="Treated", linetype ="Treated")
        +theme(plot.title=elementtext(hjust=0.5))
```

现在我们观察同一数据中的每名消费者收入。我们首先观察两组消费者收入对应的分布。

```
library(ggplot2)
ggplot(mydata,aes(x=revenue, color =treated))+
  geom_density(aes(linetype=treated))+
  labs(title="Average Revenue Per User by Treatment Group",
       x="Average Revenue Per User",
       y="Density", color ="Treated", linetype ="Treated")
        +theme(plot.title=elementtext(hjust=0.5))
```


Average Revenue Per User by Treatment Group
0.08
0.06
0.04
0.02
0.00
Density
0
10
20
30
Average Revenue Per User
Treated
A
B

While we use χ-Square test to compare the click-through rates in the two groups, we now use t-test to compare the revenue per users in the two groups.

```
1 groupA = subset(mydata, treated == "A")
2 groupB = subset(mydata, treated == "B")
3 t.test(groupA$revenue, groupB$revenue)
```

之前，我们用卡方检验判断了两组消费者的点击率是否不同。而接下来，我们将使用 t 检验判断两组消费者的收入有何不同。

```
groupA = subset(mydata, treated == "A")
groupB = subset(mydata, treated == "B")
t.test(groupA$revenue, groupB$revenue)
```

```
> t.test(groupA$revenue, groupB$revenue)

        Welch Two Sample t-test

data:  groupA$revenue and groupB$revenue
t = -31.741, df = 3997.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.333737 -4.713168
sample estimates:
mean of x mean of y
 9.896065 14.919518
```

The mean for group B is greater (14.91 vs. 9.89). Also, the p-value is highly significant (because $2.2 \times 10^{-16} \ll 0.05$), we can confidently claim that individuals in group B contribute a much higher revenue on average.

```
> t.test(groupA$revenue, groupB$revenue)

        Welch Two Sample t-test

data:  groupA$revenue and groupB$revenue
t = -31.741, df = 3997.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.333737 -4.713168
sample estimates:
mean of x mean of y
 9.896065 14.919518
```

我们发现B组消费者的平均收入更高 (14.91 vs. 9.89). 此外，我们发现 p 值也非常显著 (注意到 $2.2 \times 10^{-16} \ll 0.05$), 这是，我们可以肯定B组消费者有着较高的平均收入.

The complete code is here.

```
1 library(readr)
2 library(ggplot2)
3 mydata = read_csv("https://ximarketing.github.io/data/AB.csv")
4 groupA = subset(mydata, treated == "A")
5 groupB = subset(mydata, treated == "B")
6 t.test(groupA$revenue, groupB$revenue)
```

Question: What is the difference between the χ-squared test and the t-test?

t-test is used to compare the means of two continuous variables. χ-squared test, by contrast, demonstrates whether there is an association between two categorical variables.

问题: 我们应该如何选择卡方检验和 t 检验?

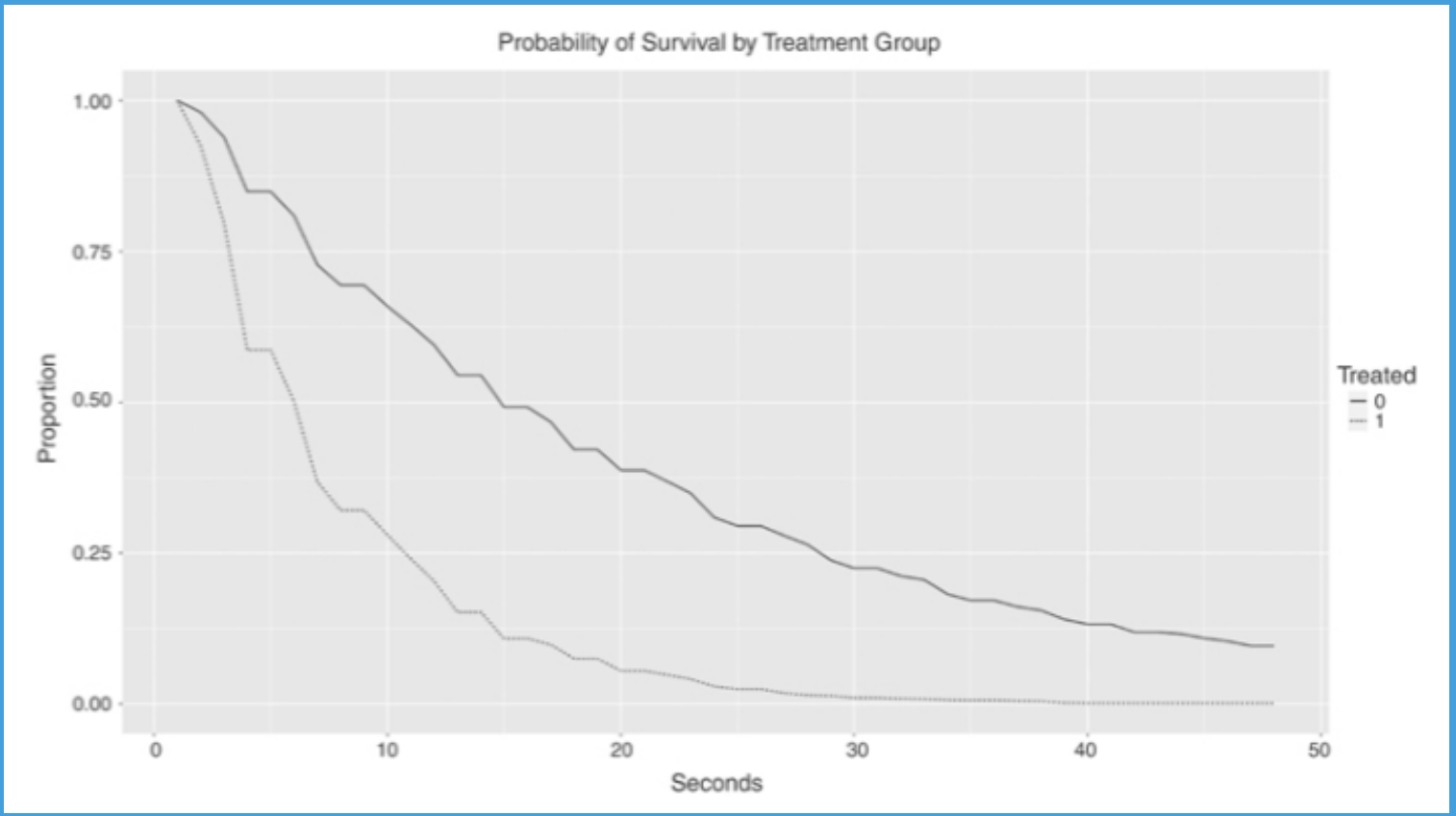
t 检验用于检验两个连续变量(收入是一个连续变量)。卡方检测侧用于检验两个分类变量(买或不买，红色或蓝色)。

Time in Product

停留时间

Time in product measures how much time an individual consumer spends on the product page. While you can direct run a *t*-test to compare their means, there is another commonly used test, the log-rank test, which compares whether the distributions in the two conditions are the same.

停留时间指个别消费者在产品页面上花费的时间。虽然我们可以直接进行 t 检验来比较它们的均值，但还有另一种常用的检验方法，即对数秩检验 (log-rank test)，它比较两个条件下的分布是否相同。



It seems that individuals in group B (the upper line) behaves differently from individuals in group A.

似乎 B 组（上线）中的个体与 A 组中的个体表现得不同。

We run the log-rank test 我们进行对数秩检验:

```
library(readr)
library(survival)
mydata =
read_csv("https://ximarketing.github.io/data/AB.csv")
fit = survdiff(Surv(as.numeric
(as.character(mydata$time_in_product)))
                    ~ mydata$treated)
fit
```

```
Call:
survdiff(formula = Surv(as.numeric(as.character(mydata$time_in_product))) ~
    mydata$treated)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
mydata$treated=A 2000     2000     1121       688      1089
mydata$treated=B 2000     2000     2879       268      1089

 Chisq= 1089  on 1 degrees of freedom, p= <2e-16
```

As before, we focus on the p-value. Here, $p < 2 \times 10^{-16} \ll 5\%$, suggesting that the two distributions are significantly different.

和之前一样，我们关注 p 值。这里，$p < 2 \times 10^{-16} \ll 5\%$，说明两个分部明显不同。

Some Caveats

一些注意事项

Question

You want to investigate the effect of "buy-one-get-one" promotion on supermarket sales. You randomly assign half of the products for promotion (group A), and half of them without promotion (group B), and see the difference?

What is wrong with this A/B test?

问题

你想调查“买一送一”促销对超市销售的影响。你随机将一半产品分为促销组（A 组），另一半不进行促销（B 组），然后观察差异。

这个 AB 测试有什么问题?

Question

You help CCB design a training program to help employees learn marketing. You randomly select half of the employees for the opportunity to join the training program and the other half without training, and see the difference.

What is wrong with the A/B test?

问题

你帮助建设银行设计一个培训项目，以帮助员工学习市场营销。你随机选择一半员工提供参加培训的机会，另一半不参加培训，然后观察差异。

这个 AB 测试有什么问题?

Question

You want to study the effect of Uber driver supply on the consumer demand. You want to change the number of Uber drivers to see how the number of orders change. In some (randomly assigned) conditions you have more drivers and in some (randomly assigned) conditions you have fewer drivers.

But you cannot force drivers to work in certain hours. What could you do in this case?

问题

你想研究 Uber 司机供应对消费者需求的影响。你想改变 Uber 司机的数量，看看订单数量如何变化。在一些（随机分配的）条件下，你有更多司机，而在另一些（随机分配的）条件下，你有更少司机。

但你无法强制司机在特定时间工作。在这种情况下，你可以怎么办?

Question

You want to study how the interest rates affect users' willingness to deposit. However, if a user finds out her interest rate is lower than others, he or she may get angry with you.

How can you run the experiment without annoying your users?

问题

你想研究利率如何影响用户的存款意愿。然而，如果用户发现自己的利率低于其他人，他或她可能会对你感到愤怒。

你如何在不惹恼用户的情况下进行实验？