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 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 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?

通过数据分析，我们发现参加建行《消费者分析》课程的员工在工作中表现更好。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 firms have better credit access, they also tend to achieve better financial performance. Does this mean better credit access leads to better financial performance?

在分析数据时，我们发现当企业获得更好的信贷准入时，它们往往也会实现更好的财务表现。这是否意味着更好的信贷准入会导致更好的财务表现?

Some people analyze the relationship between the interest rate and economic growth, and find that high interest rates lead to strong economic growth. Does this mean we should increase the interest rate to stimulate the economy?

有些人分析利率与经济增长之间的关系，发现高利率会导致强劲的经济增长。这是否意味着我们应该提高利率来刺激经济?

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

Suppose that you want to see how credit access affect financial performance, 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?

假如你想看看信贷准入是如何影响企业财务表现的，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，因为两所学校的学生是非常不同的。只有当两个学校的学生足够相似的时候，我们才可以公平的对比他们的学生收入。

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。

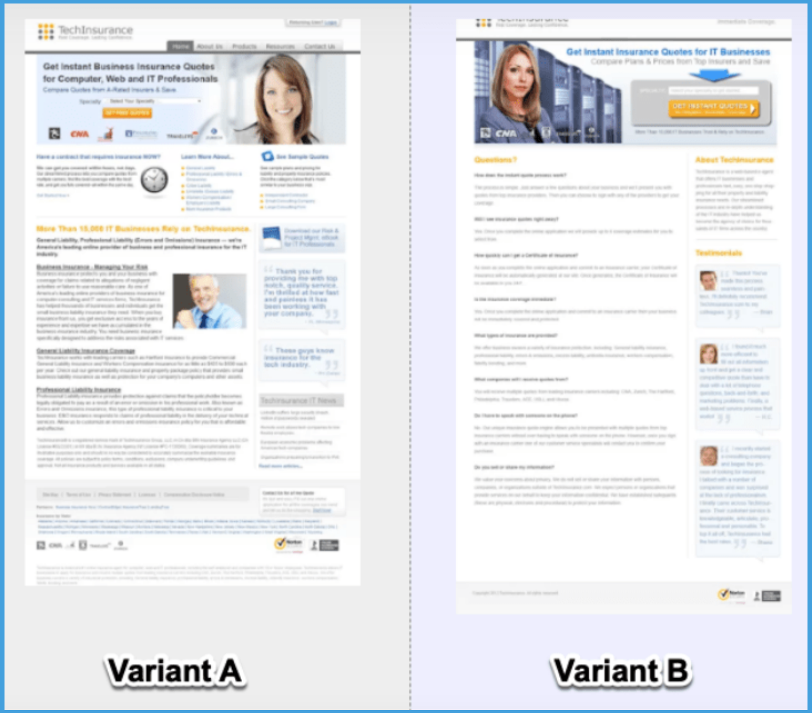
This is the basic idea of AB testing.

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。

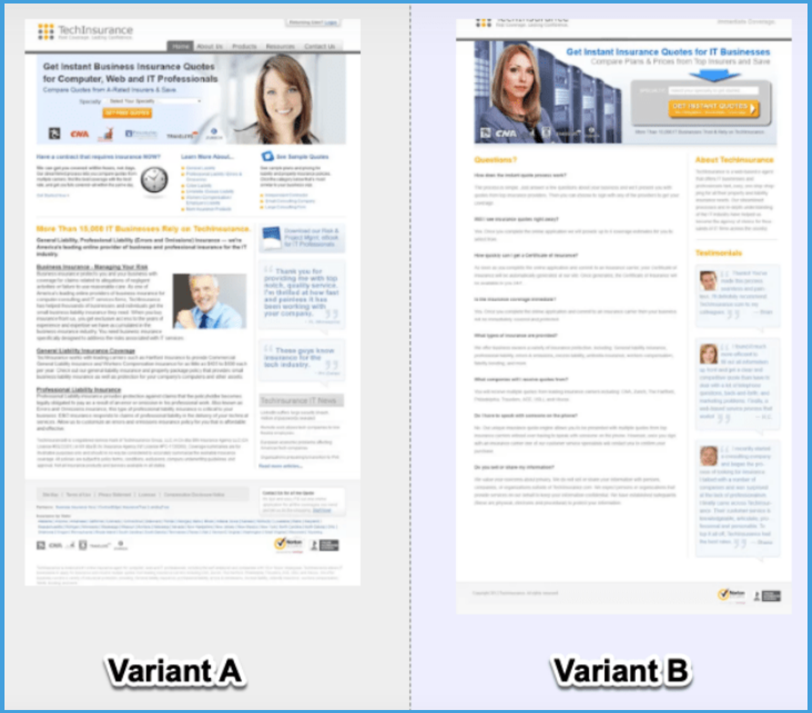
当我们需要比较两个以上的选项时，我们可以把受试者随机分配到不同的分组中，例如A组和B组。因为这两组随机分配，他们之间应该没有任何本质的不同。唯一的不同就是这两组的条件不尽相同。

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.


TechInsurance
Get Instant Business Insurance Quotes for Computer, Web and IT Professionals
Get Instant Insurance Quotes for IT Businesses
Compare Plans & Prices from Top Insurers and Save
Variant A
Variant B

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?



TechInsurance

Home About Us Products Resources Contact Us

Get Instant Business Insurance Quotes for Computer, Web and IT Professionals

Compare Quotes from A-Rated Insurers & Save

More Than 15,000 IT Businesses Rely on TechInsurance.

General Liability, Professional Liability (Errors and Omissions) Insurance — we're America's leading online provider of business and professional insurance for the IT industry.

Business Insurance - Managing Your Risk

Thank you for providing me with top notch, quality service. I'm thrilled at how fast and painless it has been working with your company.

Variant A

Get Instant Insurance Quotes for IT Businesses

Compare Plans & Prices from Top Insurers and Save

GET INSTANT QUOTES

Questions?

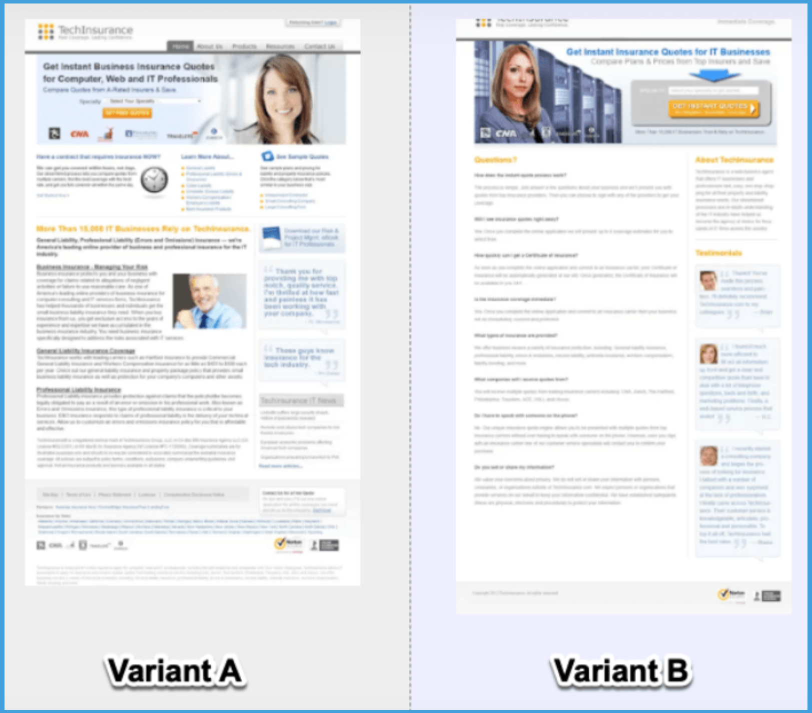
About TechInsurance

Testimonials

Variant B

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！


TechInsurance
Get Instant Business Insurance Quotes for Computer, Web and IT Professionals
Get Instant Insurance Quotes for IT Businesses
Compare Plans & Prices from Top Insurers and Save
GET INSTANT QUOTES
Questions?
About TechInsurance
Testimonials
Variant A
Variant B

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，我们也不能确定他们的收入差距到底是性别偏见导致的还是大学教育导致的。

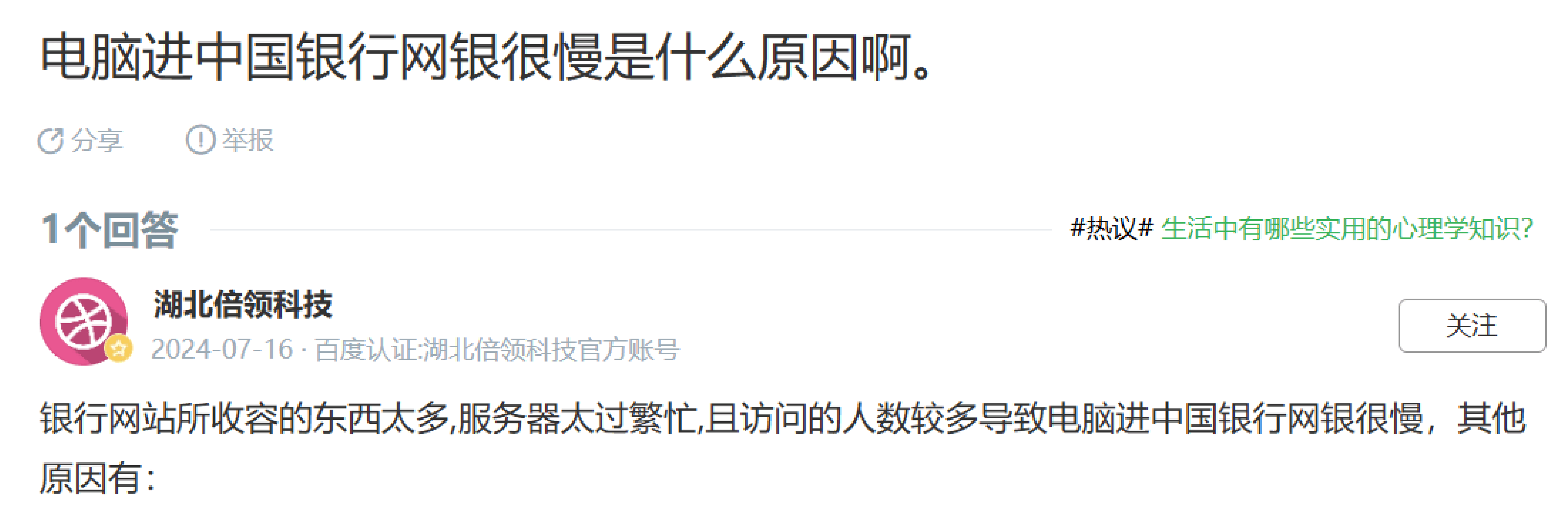
Example of A/B test: Speed matters.

> “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



电脑进中国银行网银很慢是什么原因啊。

1个回答

#热议# 生活中有哪些实用的心理学知识?

湖北倍领科技

2024-07-16 · 百度认证:湖北倍领科技官方账号

银行网站所收容的东西太多,服务器太过繁忙,且访问的人数较多导致电脑进中国银行网银很慢，其他原因有:

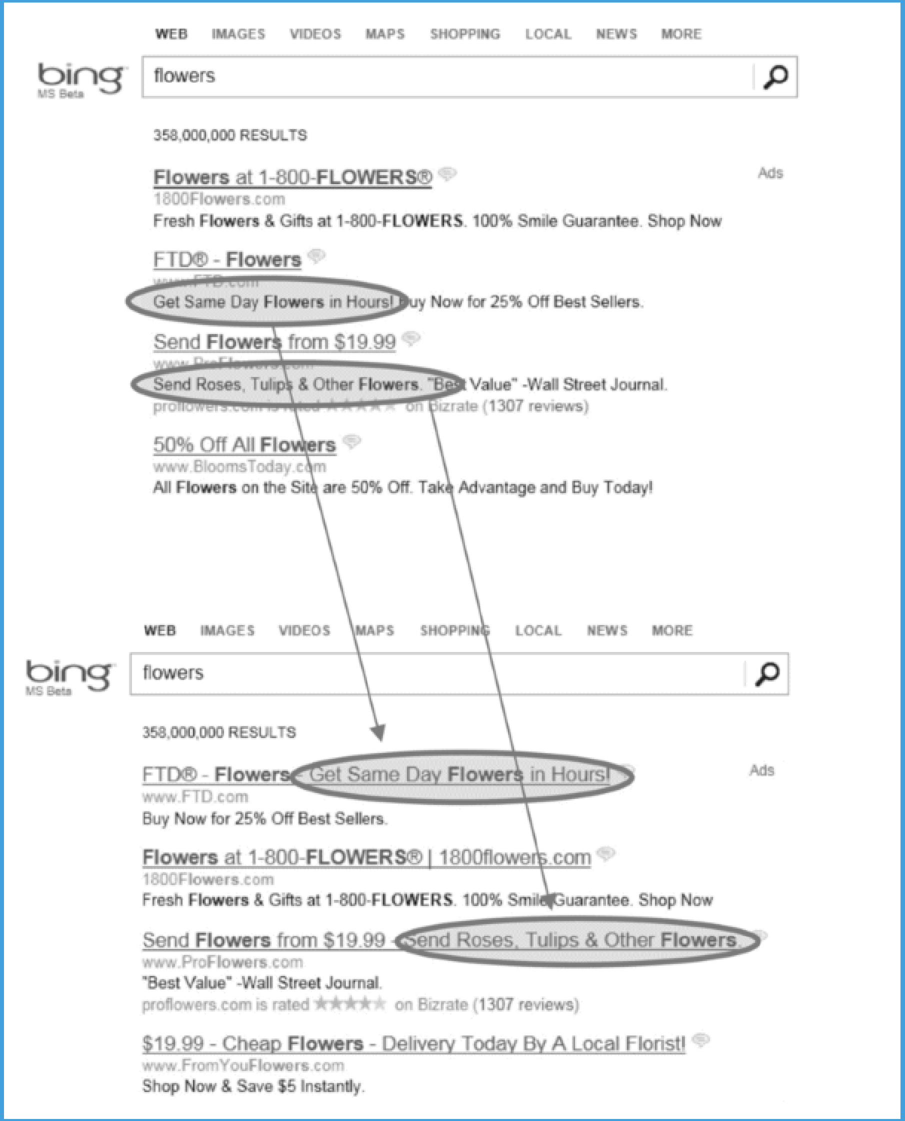
Example of A/B test: Speed matters.

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



flowers

358,000,000 RESULTS

Ads

Flowers at 1-800-FLOWERS®
1800Flowers.com
Fresh Flowers & Gifts at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now

FTD® - Flowers
www.FTD.com
Get Same Day Flowers in Hours! Buy Now for 25% Off Best Sellers.

Send Flowers from $19.99
www.ProFlowers.com
Send Roses, Tulips & Other Flowers. "Best Value" -Wall Street Journal.
proflowers.com is rated ★★★★★ on Bizrate (1307 reviews)

50% Off All Flowers
www.BloomsToday.com
All Flowers on the Site are 50% Off. Take Advantage and Buy Today!

flowers

358,000,000 RESULTS

Ads

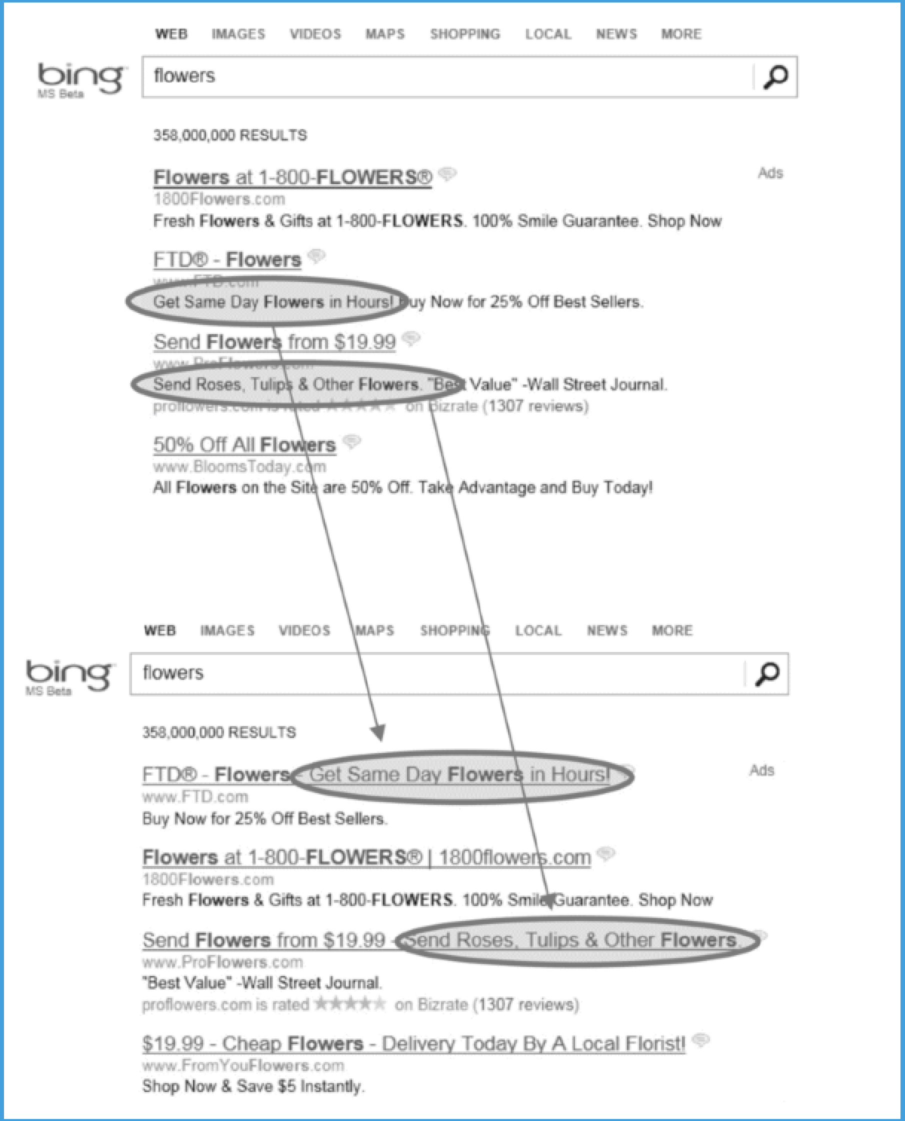
FTD® - Flowers - Get Same Day Flowers in Hours!
www.FTD.com
Buy Now for 25% Off Best Sellers.

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.

AB 测试的例子



flowers

358,000,000 RESULTS

Ads

Flowers at 1-800-FLOWERS®
1800Flowers.com
Fresh Flowers & Gifts at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now

FTD® - Flowers
www.FTD.com
Get Same Day Flowers in Hours! Buy Now for 25% Off Best Sellers.

Send Flowers from $19.99
www.ProFlowers.com
Send Roses, Tulips & Other Flowers. "Best Value" -Wall Street Journal.
proflowers.com is rated ★★★★★ on Bizrate (1307 reviews)

50% Off All Flowers
www.BloomsToday.com
All Flowers on the Site are 50% Off. Take Advantage and Buy Today!

flowers

358,000,000 RESULTS

Ads

FTD® - Flowers - Get Same Day Flowers in Hours!
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推广，用户使用信用卡在亚马逊购物时可以享受一定优惠。

虽然这个信用卡的利润很高，但广告的点击率[CTR]却非常低。你会怎么设计这个广告提高其效果?

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: Up to 4% p.a. savings rate

B: Enjoy HSBC Premier Elite Welcome Rewards worth over HKD88,000

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网络广告的有效性：

A: 新客戶尊享高達4%活期存款年利率

B: 尊享滙豐Premier Elite迎新禮遇價值超過港幣88,000元

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:

5 out of 80 [6.25%] users clicked through on ad A;
6 out of 80 [7.50%] users clicked through on ad B.

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

假设：

80名用户中有5人点击了广告A [6.25%]；
80名用户中有6人点击了广告B [7.50%]。

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：

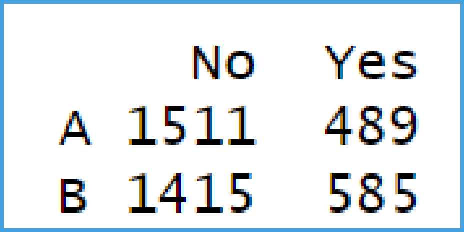
8,560名用户中有450人点击了广告A [5.25%]；
12,980名用户中有710人点击了广告B [5.47%]。

你能说广告B比广告A更有效吗？

The χ-Squared Test

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

Treated: Which ad consumers are exposed to.



```
     No  Yes
A  1511  489
B  1415  585
```

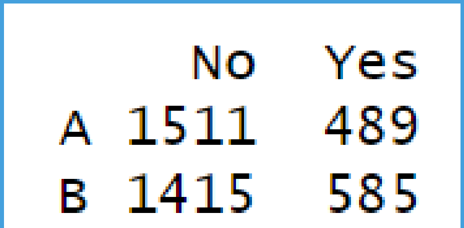
Among consumers who saw ad A, 489 clicked through and 1,511 did not click. Among consumers who saw ad B, 585 clicked through and 1,415 did not click.

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

卡方检验

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

实验内容: 消费者看到的是哪一种广告.



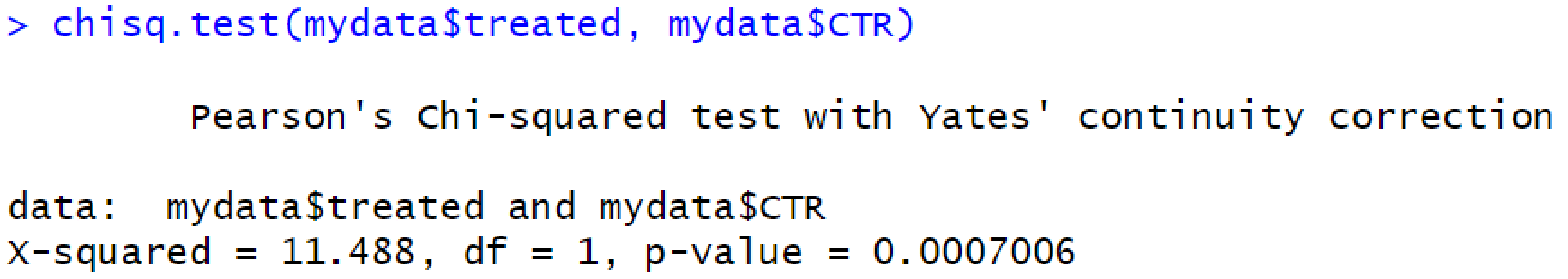
```
    No   Yes
A 1511   489
B 1415   585
```

在看到广告A的用户中, 489个人点击了广告，1,511个人没有点击广告. 在看到广告B的用户中, 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

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

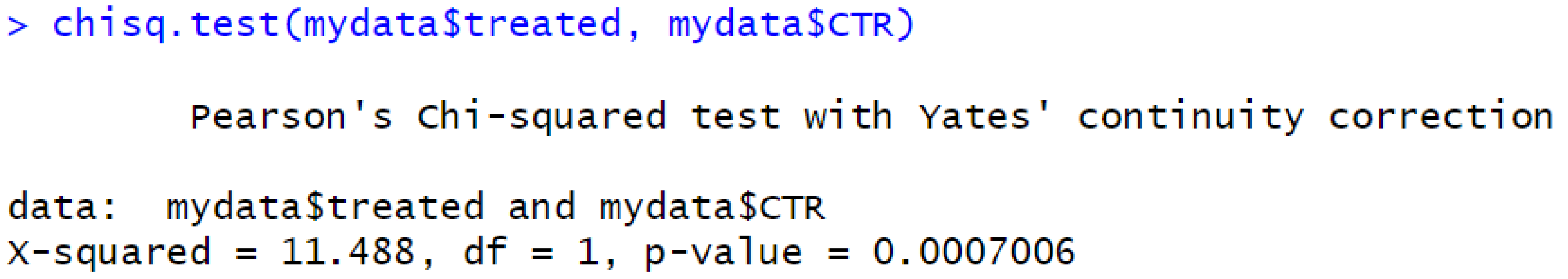


```
> chisq.test(mydata$treated, mydata$CTR)

        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

```
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.

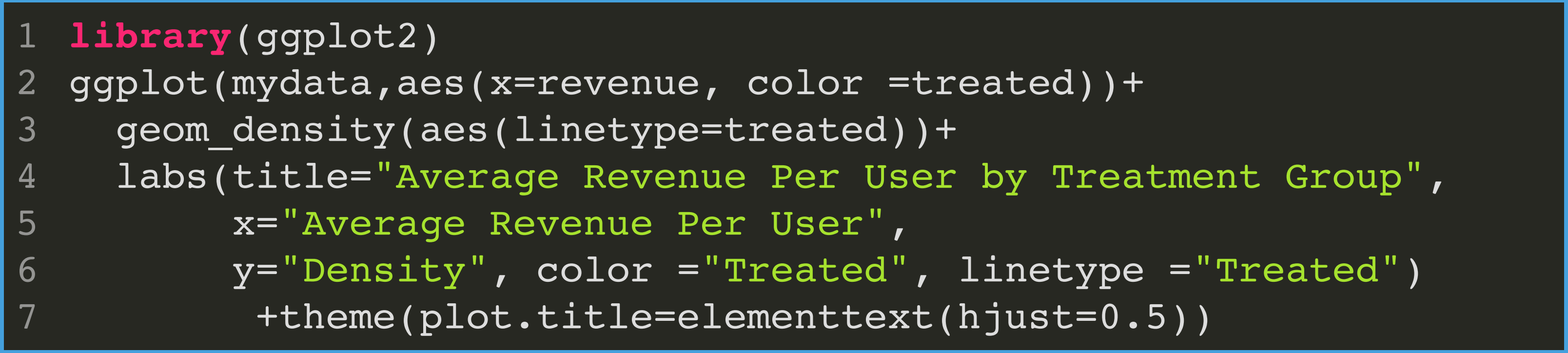
The complete code is here.

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

Revenue
利润

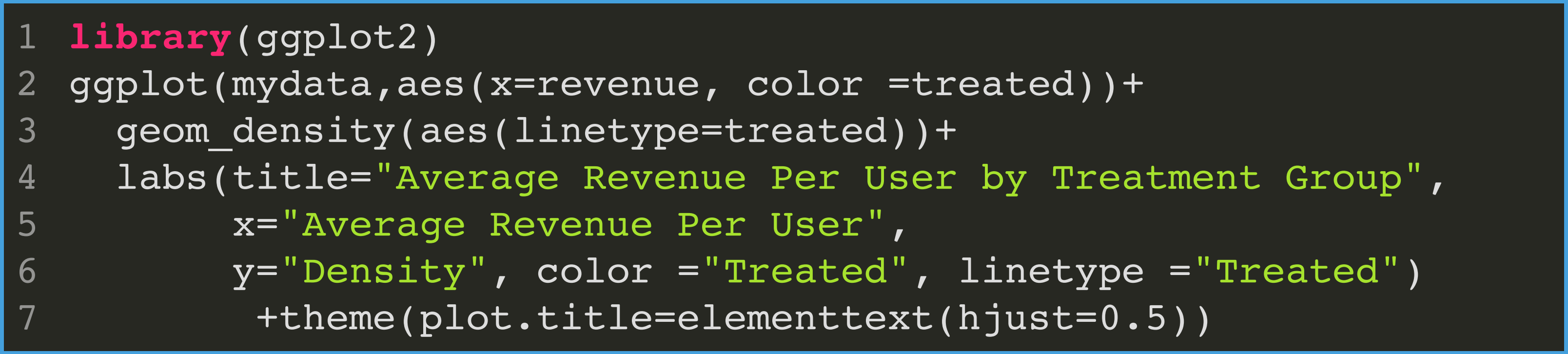
Next, we compare the revenue per consumer in the two conditions using the same dataset.

We first visualize the distribution.

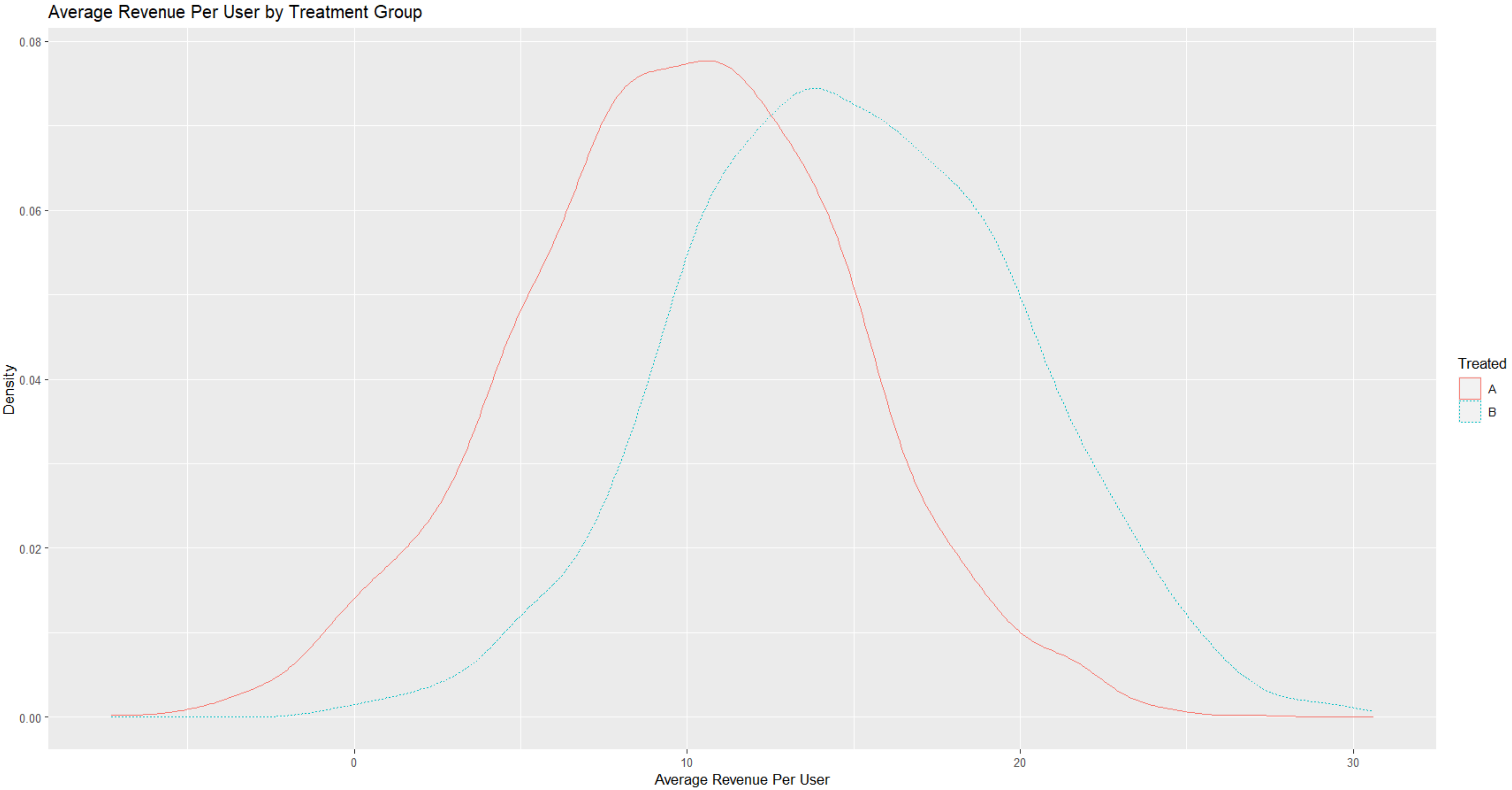


```r
library(ggplot2)
ggplot(mydata,aes(x=revenue, color =treated))+
  geom_density(aes(linetype=treated))+
  labs(title="Average Revenue Per User by Treatment Group",
       x="Average Revenue Per User",
       y="Density", color ="Treated", linetype ="Treated")
        +theme(plot.title=elementtext(hjust=0.5))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

现在我们观察同一数据中的每名用户的利润。我们首先观察两组用户利润对应的分布。



```r
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))
```

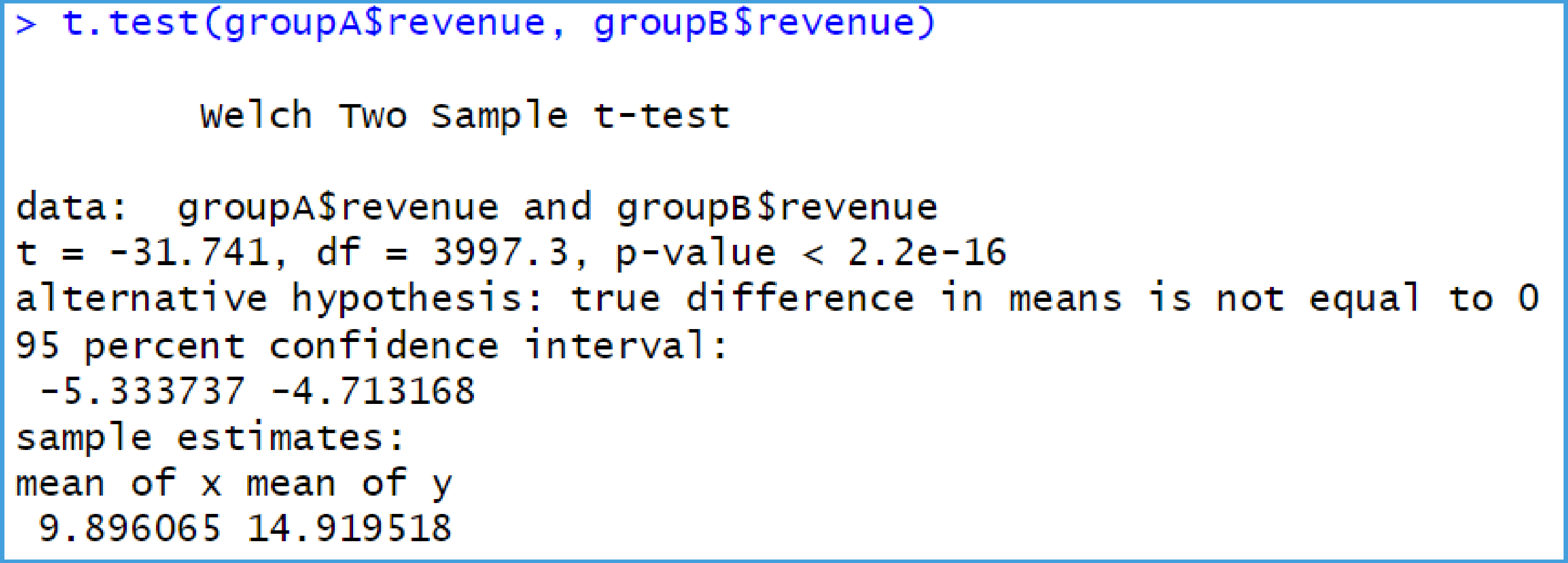

Average Revenue Per User by Treatment Group
Density
0.08
0.06
0.04
0.02
0.00
0
10
20
30
Average Revenue Per User
Treated
A
B

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.

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

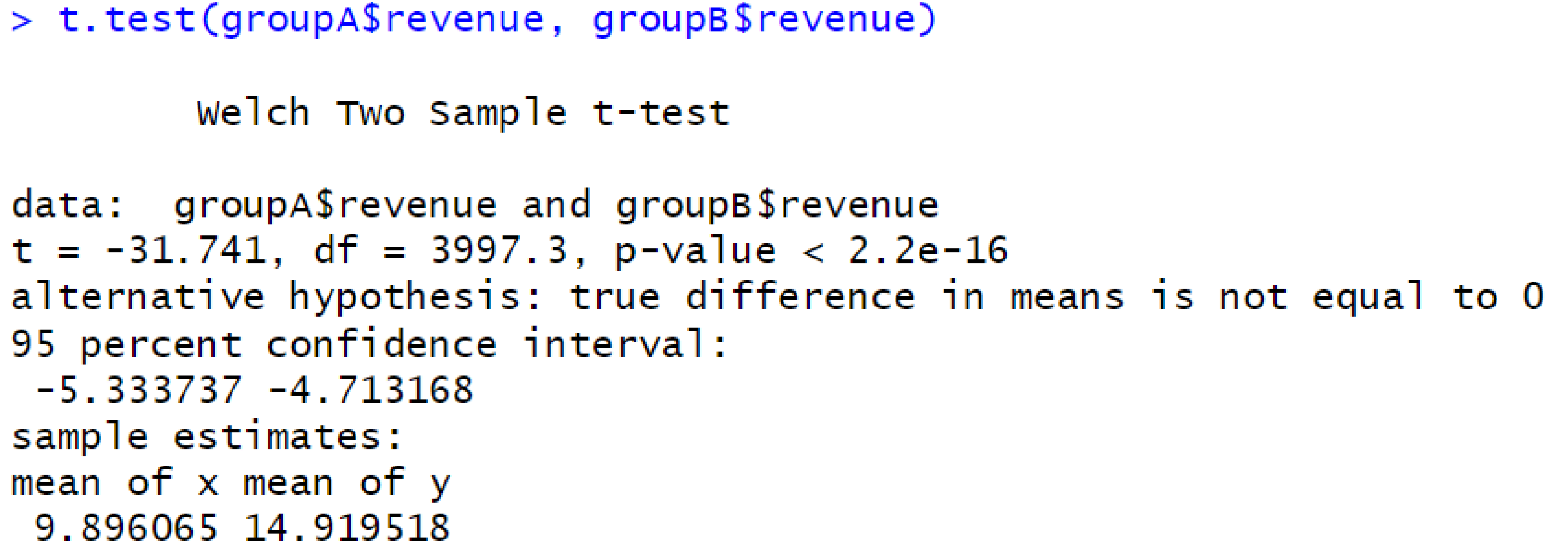
之前，我们用卡方检验判断了两组用户的点击率是否不同。而接下来，我们将使用 t 检验判断两组用户的利润有何不同。

```r
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组用户的平均利润更高 (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组用户有着较高的平均利润.

The complete code is here.

```r
library(readr)
library(ggplot2)
mydata = read_csv("https://ximarketing.github.io/data/AB.csv")
groupA = subset(mydata, treated == "A")
groupB = subset(mydata, treated == "B")
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)。

Some Caveats
一些注意事项

Question

A bank is testing a new loan product by randomly assigning customers to either a control group (existing loan product) or a treatment group (new loan product).

What is wrong with this A/B test?

问题

你的银行正在通过随机将客户分配到对照组(他们拿到现有贷款产品)或实验组(他们拿到新贷款产品)。通过这个实验，你希望比较哪一款贷款产品效果更好。

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课程。你随机选择一半员工并为他们提供参加培训的机会，另一半不员工则没有参加培训的机会，然后观察两组员工表现有没有显著差异。

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 司机数量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，如果某个用户发现自己拿到的存款利率低于其他人的存款利率，他或她可能会对你感到愤怒，觉得自己被歧视了。

你如何在不惹恼用户的情况下进行实验？

课后讨论问题：

你想观察信用卡手续费是如何影响用户消费的。为了回答这一问题，你进行了如下的AB测试：你随机选择一部分用户，他们还是原来的的手续费，你随机选择另一部分用户，他们的手续费减少到原来的一半。实验进行一个月后，你比较这两组消费者的消费情况有何不同。这个实验有哪些问题？