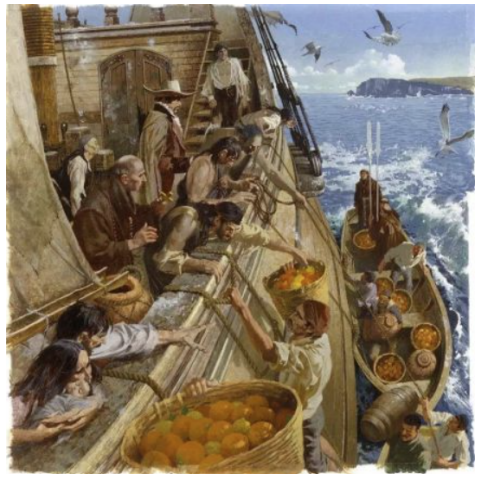
BUSS2302: Business Statistics
Lab 3: A/B Testing

成功案例: "海上瘟疫"坏血病是如何被战胜的?

柑橘⇒坏血病



- 从1500年到1800年这300年间，坏血病杀死了三百万名船员。
- 英国海军医生詹姆斯.林德发现了一种神奇的巧合:

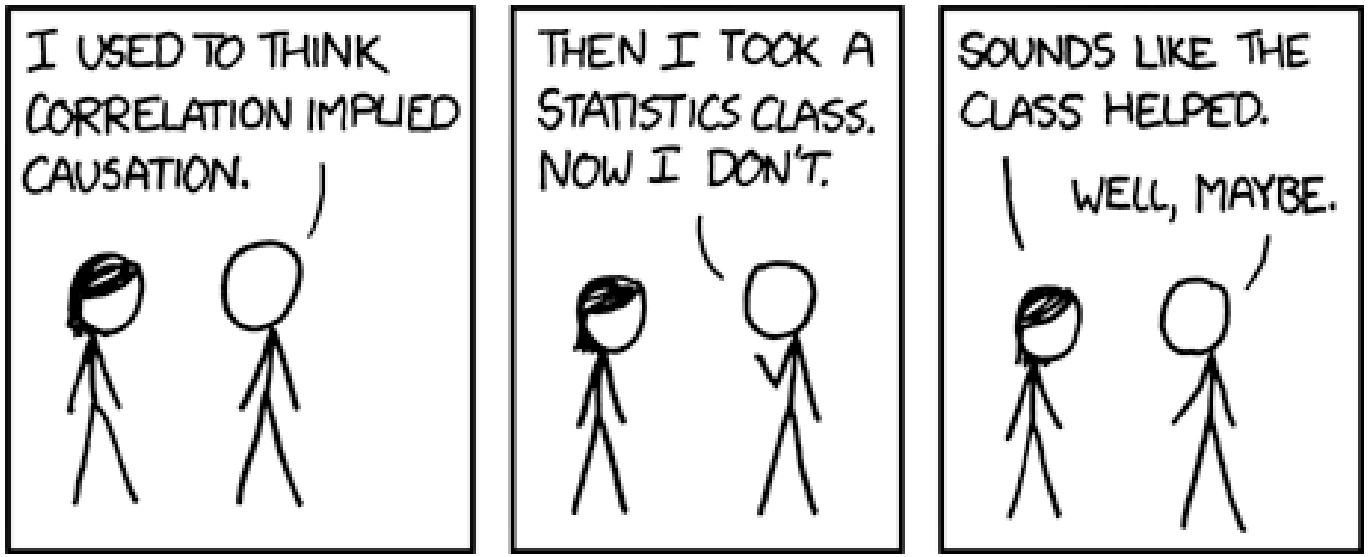
  食谱中有柑橘类的水果的船员患坏血病的几率更低。
- 船员的救命药难道就是再普通不过的青柠(Lime)吗?
- 学过商务统计课程后，你会怎么验证这一猜想?

成功案例: "海上瘟疫"坏血病是如何被战胜的?

柑橘⇒维生素C⇒坏血病(发现因果规律)

- 林德选择今天看来已经非常普通且广泛运用在互联网的方法来证明自己的猜想——A/B测试。
  - 他把患病的12名船员们分成六组。
  - 确保病人们的基本食物一样，所处环境也相同。
  - 唯一的变量是，给每组开出了不同的疗法：一组船员的饮食中加入青柠、一组加入橘子、一组加苹果、一组加醋、一组加酏剂、一组加海水。
  - 结果是摄入橘子和青柠的很快痊愈!
- 这是两百多年前的坏血病实验，其采纳的A/B测试已经成为医学领域最为常见的实验方法。
- 在医学和经济学实验中，我们更愿意称为双盲对照组实验。

相关性 $\xrightarrow{?}$ 因果关系



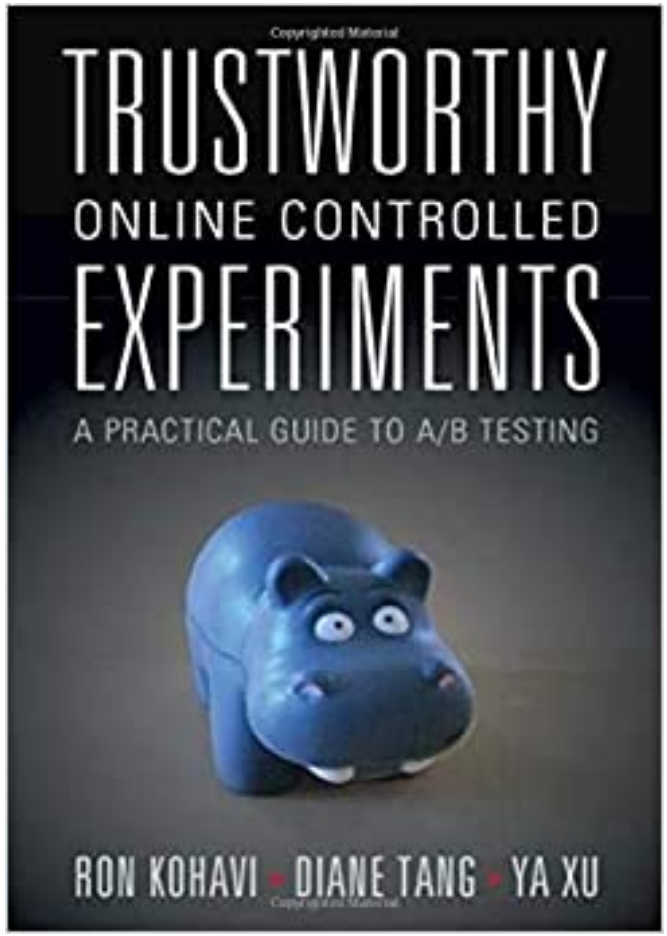
- 柑橘是治疗坏血病的原因吗?

What is A/B Testing?

- The A/B test origins from biomedical double-blind testing
  - Patients are randomly divided into two groups, and given placebos or test drugs without knowing prior.
  - The researchers then determine whether the test drug is effective according to the performance of two groups
  - Question: If you are the researchers, how to determine?
- A/B Tests: a controlled experiment that compares two variants
  - *A* (*Control*): 酸性物质，海水，苹果⇒安慰剂
  - *B* (*Treatment*)：青柠，橘子⇒新药
- An A/B test usually consist of:
  1. Randomly split experiment units between two versions
  2. Collect sample results of interest, and analyze
- The terms controlled experiments (控制对照实验) and A/B tests are used interchangeably.

Reference Book

- Trustworthy Online Controlled Experiments : A Practical Guide to A/B Testing.
- 关键迭代：可信赖的线上对照实验 ▸ 关键迭代
- **Question 1:** Data-Driven V.S. Highest Paid Person's Opinion (HiPPO)-Driven
- **Question 2:** Correlation V.S. Causation



Why A/B Tests?

- A/B 测试被广泛的应用于各大公司

- 字节跳动三大B端业务，巨量引擎、飞书和火山引擎 ▸ Volcengine
- 字节跳动认为，A/B测试产品是最能够体现数据驱动价值的产品，也最能代表其想传达给其他企业的理念：让数据驱动成为习惯。
- A/B 测试以及数据驱动的决策是互联网与大数据时代基础方法论的重要基石。—连乔，快手副总裁

Why A/B Testing?

Bing

- 在2012年，Bing的工程师建议通过将广告标题行与标题页下方第一行的文本结合的方式来加长广告标题行。
- 会有用吗？ ▸ Bing

Why A/B Testing?

Bing

- 在2012年，Bing的工程师建议通过将广告标题行与标题页下方第一行的文本结合的方式来加长广告标题行。
- 会有用吗？ ▸ Bing
- 在积压了半年之后，一个软件开发者实现了这个想法。
- 随机向一些用户展示了新的标题布局，其他的则是旧的
- 开始测试几个小时后，触发了收入过高警报。
- Bing 的收入增长了惊人的12%, 没有显着影响关键的用户体验指标

Why A/B Testing?

Summary

- In Bing, key themes in A/B testing:
  - Sometimes it's hard to assess the value of an idea
  - Small changes can have a big impact
  - The overhead of running an experiment must be small
    适用场景：难评价，低成本，大影响

Why A/B Testing?

Summary

- In Bing, key themes in A/B testing:
  - Sometimes it's hard to assess the value of an idea
  - Small changes can have a big impact
  - The overhead of running an experiment must be small
    适用场景：难评价，低成本，大影响
- RCTs (A/B testings) are:
  - the best scientific way to establishing causality with high probability
  - detect small changes such as changes over time (sensitivity)
  - detect unexpected changes
    优点：因果，敏感，意想不到

Terminology of A/B tests

- Terminology of A/B tests:
  - **Overall Evaluation Criterion (OEC)** (综合评价标准): A quantitative measure of the experiment's objective, also called the response or dependent variable (结果变量)
  - **Parameter** (参数): A controllable experimental variable that is thought to influence the OEC or other metrics of interest, sometimes are also called factors or variables (因子)
  - **Variant** (变体) : A user experience being tested, typically by assigning values to parameters (Treatments 处理)
  - **Randomization Unit** (随机化单元): A pseudo-randomization process is applied to units (e.g., users or pages) to map them to variants (Experiment units 实验对象)

Necessary Ingredients

- Necessary ingredients for running useful A/B tests:
  - 互不干扰： There are experimental units (e.g., users) that can be assigned to different variants with no interference (or little interference) $\Rightarrow$ Independence 独立性
  - 有足够多的实验单元： There are enough experimental units (e.g., users)
  - 关键指标: Key metrics, ideally an OEC, are agreed upon and can be practically evaluated.
  - 改动容易实现： Changes are easy to make

A/B 测试的具体步骤

0. 假设检验和统计推断(在课件三、四、五、六覆盖的核心内容！)
   - A/B实验分析的主要理论依据
1. 设计A/B实验
   - 明确实验目的：提高点击率？增加营收？提升品牌影响？ $\Rightarrow$ OEC
   - 确定随机化单元：
     - 用户（互联网公司）
     - 病人（医学实验）
     - 上市公司（经济学实验）
   - 实验需要多大样本量？
     - 预期的显著性水平和检验效力(课件三第一类错误和第二类错误的权衡！)
     - 更多的用户一定更好吗？ $\Rightarrow$ 如果依靠时间累计更多的用户
       $\Rightarrow$ 周内效应，季节性，初始效应
   - 实验需要运行多长时间？

A/B 测试的具体步骤

2. 运行实验获得数据(工具化日志记录)
   - OEC的样本数据(结果变量)
   - 分组信息
3. 分析结果
   - 对均值的检验：单均值，双均值，多均值(ANOVA)
   - 对方差的检验: 单方差，双方差
   - 对相关性的检验(因果)：方差分析
4. 从结果到决策：运行A/B测试的目标是以收集数据，分析数据，进而驱动决策。
   - 在获得统计显著性的结果后，进一步思考方案的成本
   - 统计显著$\Rightarrow$实际（经济显著）显著

An Academic Example

- Breza E, Stanford F C, Alsan M, et al. Effects of a large-scale social media advertising campaign on holiday travel and COVID-19 infections: a cluster randomized controlled trial[J]. *Nature medicine*, 2021, 27(9): 1622-1628.
- Some facts: During the COVID-19 crisis, many healthcare professionals used social media to spread public health messages
- Question: But will these messages really influence people's behavior at scale?

An Academic Example

But will these messages really influence people's behavior at scale?

An Academic Example

- Experiments: To figure it out, some nurses and doctors had recorded some videos to encourage people staying at home for the coming Thanksgiving holiday (26, Nov) and then disseminated to Facebook users as ads, see: ▸ Covid19Video
- It's known that holiday travel would lead to a surge in the epidemic
- Now we want to design an experiment to see whether these short videos would cause a decline in COVID-19 cases after the holidays

An Academic Example

Design of Experiment

We will make the following 5 decisions to finalize the design

1. What is the Overall Evaluation Criterion (OEC)?
2. What is the randomization unit?
3. What population of randomization units do we want to target?
4. How large (size) does our experiment need to be?
5. How long do we run the experiment?

Design of experiment

OEC (Response Variable)

- Q1: How to choose a proper OEC?
  - Basic principles:
    - using short-term metrics that predict long-term value (and hard to game);
    - can be practically evaluated;
    - reliable data can be collected, ideally cheap

Design of experiment

OEC (Response Variable)

- Q1: How to choose a proper OEC?
  - Basic principles:
    - using short-term metrics that predict long-term value (and hard to game);
    - can be practically evaluated;
    - reliable data can be collected, ideally cheap
  - We may choose zip code-level COVID-19 infections reported to state health authorities as OEC
- Q2: What is the randomization unit?
  - The randomization unit will be Facebook users in experimental areas

Design of Experiment

Randomization Units

- Q3: What population of randomization units do we want to target? How should we randomize them?

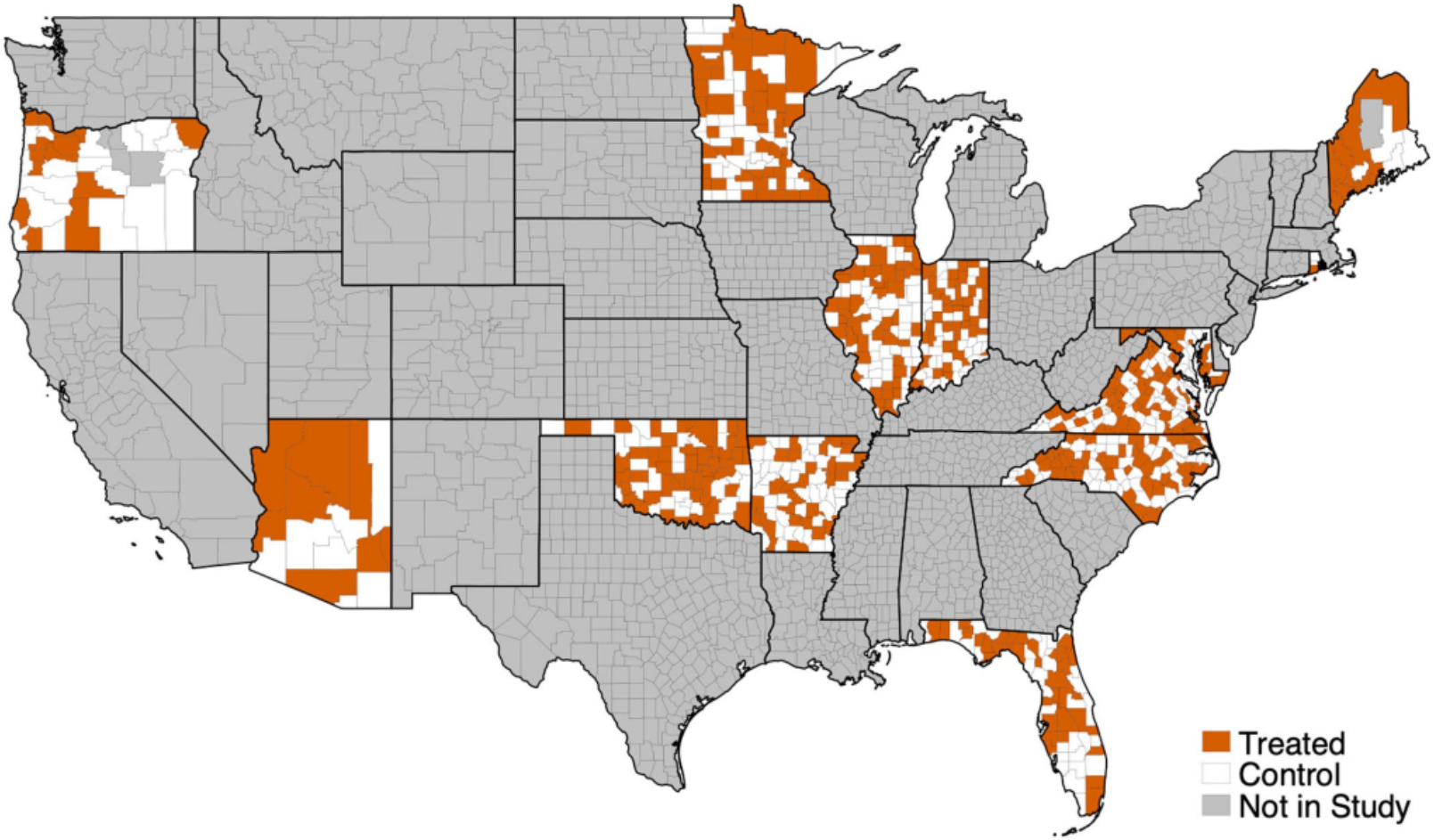
Design of Experiment

Randomization Units

- Q3: What population of randomization units do we want to target? How should we randomize them?
  - The researchers had selected 13 states where the data needed are available, then:
    - Randomly allocated counties to be 'high-intensity (H) or 'low-intensity' (L) with probability $\frac{1}{2}$ each
    - In H/L counties, randomized zip codes into intervention with probability $\frac{3}{4}$ / $\frac{1}{4}$ and control with probability $\frac{1}{4}$ / $\frac{3}{4}$
    - The videos were only sent to the users in intervention areas
    - 随机化单元：县级层次，邮编层次。
    - 实验组有两个层次：高强度和低强度处理！

Design of Experiment

Randomization Units



Design of Experiment

Sample Size

- Q4: How large (size) does our experiment need to be?
  - The dataset, a sample of 4925 zip codes can provide 80% power and 5% significance to detect standardized difference 0.057.
  - In the paper, they have 6998 zip-codes.
  - Recall Lecture 3:
    - Type I error: α; Type II error: β; Power: $1-\beta$
    - μ_0 and μ_1: Means under the null and alternative hypotheses
    - n_0 and n_1: sample sizes in two groups (control and treatment) (may be the same)
  - Given the above information, what is the desirable sample size?

Design of Experiment

Sample Size

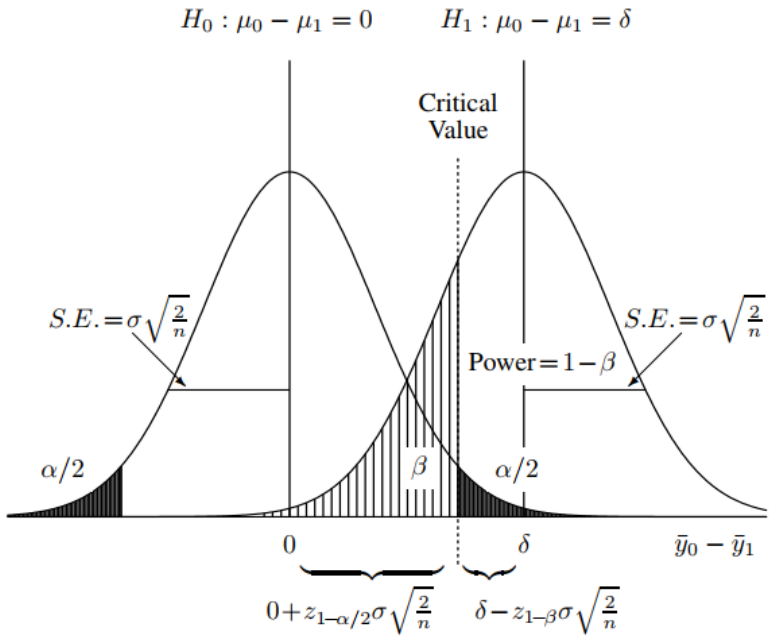


Fig. 2.1 Sampling model for two independent sample case. Two-sided alternative, equal variances under null and alternative hypotheses.

Two-tailed Tests: : $$n = \frac{2\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \sigma^2}{\left(\mu_0 - \mu_1\right)^2}$$

Design of Experiment

Sample Size

Type II Error β	Power $1-\beta$ Power	Numerator for Sample Size Equation (2.3)	
		One Sample	Two Sample
0.50	0.50	4	8
0.20	0.80	8	16
0.10	0.90	11	21
0.05	0.95	13	26
0.025	0.975	16	31

- Rule of Thumb: $n = \frac{16\sigma^2}{(\mu_0 - \mu_1)^2}$
- For $\alpha = 0.05$, $Z_{1-\alpha/2} = Z_{0.975} = 1.96$.
- For $\beta = 0.2$, $Z_{1-\beta} = Z_{0.8} = 0.84$.
- $2\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \approx 16$

Design of Experiment

Experiment Duration

- Q5: How long do we run the experiment?
  - The videos had been disseminated since 13 Nov.
  - To avoid seasonality and novelty effect, we would collect the data from 5 days after Thanksgiving (26, Nov)
  - We will collect 2 weeks data, which is COVID-19 daily new cases from 1 to 14 Dec.
  - You can download the dataset from Canvas
- Now let us play with the data

Python Implementation: Overview of the dataset

- We will provide you a modified dataset with 6024 zip-codes. The structure of the dataset is shown as below:

```
# load the dataset
data = pd.read_csv('data&analysis.csv')
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6024 entries, 0 to 6023
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   user_loc         6024 non-null   int64
 1   zip              6024 non-null   int64
 2   state            6024 non-null   object
 3   high_county_T1   6024 non-null   int64
 4   treated_T1       6024 non-null   int64
 5   urban            6024 non-null   int64
 6   two_weeks_cases  6024 non-null   int64
 7   majority_gop     6024 non-null   int64
dtypes: int64(7), object(1)
memory usage: 376.6+ KB
```

```
data.head()
```

	user_loc	zip	state	high_county_T1	treated_T1	urban	two_weeks_cases	majority_gop
0	4001	85925	AZ	0	0	1	0	0
1	4001	85924	AZ	0	0	0	21	0
2	4001	85936	AZ	0	0	1	14	0
3	4001	85940	AZ	0	0	0	0	0
4	4001	86505	AZ	0	0	0	11	0

- The summary of data from high and low intensity counties are like:

```
data.loc[data['high_county_T1']==0].describe()
```

	user_loc	zip	high_county_T1	treated_T1	urban	two_weeks_cases	majority_gop
count	2910.000000	2910.000000	2910.0	2910.000000	2910.000000	2910.000000	2910.000000
mean	24848.143643	48236.434708	0.0	0.249828	0.555326	92.438832	0.674227
std	14723.517922	22235.212970	0.0	0.432988	0.497015	199.255925	0.468744
min	4001.000000	2802.000000	0.0	0.000000	0.000000	0.000000	0.000000
25%	12103.000000	27959.750000	0.0	0.000000	0.000000	4.000000	0.000000
50%	18179.000000	47849.000000	0.0	0.000000	1.000000	22.000000	1.000000
75%	37158.500000	62659.500000	0.0	0.000000	1.000000	93.000000	1.000000
max	51810.000000	97630.000000	0.0	1.000000	1.000000	2744.000000	1.000000

(a) Data of low intensity counties

```
data.loc[data['high_county_T1']==1].describe()
```

	user_loc	zip	high_county_T1	treated_T1	urban	two_weeks_cases	majority_gop
count	3114.000000	3114.000000	3114.0	3114.000000	3114.000000	3114.000000	3114.000000
mean	26146.884393	48330.790623	1.0	0.745022	0.578356	90.239563	0.656069
std	13818.149251	22095.737032	0.0	0.435919	0.493902	172.544960	0.475095
min	4005.000000	2804.000000	1.0	0.000000	0.000000	0.000000	0.000000
25%	17031.000000	28450.250000	1.0	0.000000	0.000000	5.000000	0.000000
50%	24031.000000	47959.500000	1.0	1.000000	1.000000	24.000000	1.000000
75%	39300.500000	62546.750000	1.0	1.000000	1.000000	98.000000	1.000000
max	51840.000000	97701.000000	1.0	1.000000	1.000000	2026.000000	1.000000

(b) Data of high intensity counties

Recall: Two Population Means Testing

- Step 1: state the null and alternative hypotheses
  - H_0: $\mu_t \geq \mu_c$; H_1:$\mu_t < \mu_c$
- Step 2: select the level of significant α
  - We may choose $\alpha = 0.1$
- Step 3: specify the test statistic
  - $t = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$
- Step 4: Two approaches to determine the rejection rule
  - Critical-value approach
  - p-value approach
- Step 5: Decide to either reject H_0 or not to according to the rejection rule

Python Implementation - Two Population Means Testing

- We may deliver hypothesis testing procedures as what we have learned in Lab 2. e.g.

```
'''
H_0: the average COVID-19 infections in areas with intervention is greater or equal to that in the areas without intervention
H_a: the average COVID-19 infections in areas with intervention is less than that in the areas without intervention
'''
# Notice that the p-value output by function 'scipy.stats.ttest_ind' is two-tailed probability
t, pval = stats.ttest_ind(treatment['two_weeks_cases'], control['two_weeks_cases'], equal_var=False)
print('t-statistic:', t, 'p-value:', pval / 2)
print('p > 0.05, we cannot reject H_0')
```

```
t-statistic: -0.6758279130183211 p-value: 0.2495881006264814
p > 0.05, we cannot reject H_0
```

- We may also analyze the high and low intensity counties separately:

County treatment	Mean		
	Treatment	Control	p-value
All	89.70	92.94	0.25
Low intensity	98.24	90.51	0.195
High intensity	87.03	99.63	0.052

- Are these results reasonable?
- Are there any other factors may influence the result?

Further Analysis: ANOVA

- We can partition all counties into urban or rural.
- Recall:The goal of ANOVA is to split the **total variation** of a set of observations into different sources, such as
  - Treatment adjustments α_j (One-way ANOVA, Two-way ANOVA)
  - Error terms ε_{ij} (Benchmark under the H_0)
- Between-group Variance: Treatment adjustments α_j+Error term ϵ_{ij}
- Within-group Variance: Error terms ϵ_{ij}
- It motivates us to construct a test statistic as a variability ratio

$$\frac{\textbf{Between-group Variance}\ (\text{组间方差})}{\textbf{Within-group Variance}\ (\text{组内方差})}$$

and reject H_0 if this ratio is sufficiently large.

Further Analysis: ANOVA

- We want to find whether these videos had the same effect in urban and rural counties, thus we may define:
  - Factor: urban or rural counties
  - Treatment:video intervention
  - Experiment units: users in experimental areas
  - Response variable: zip-code level COVID-19 infections

Recall: Test for the Equality of k Population Means

- **Step 1:** Hypotheses: either for urban or rural counties,

$$H_0 : \mu_{u,t} = \mu_{u,c} = \mu_{r,t} = \mu_{r,c}$$
$$H_a : \text{Not all population means are equal}$$

  where $u/r, t/c$ represent the urban/rural and treatment/control respectively
- **Step 3:** Test Statistic

$$F = \text{MSTR}/\text{MSE}$$

- **Step 4:** Rejection Rule
  - p-value Approach: Reject H_0 if p-value$\leq \alpha$
  - Critical Value Approach: Reject H_0 if $F \geq F_\alpha$, where the value of F_α is based on an F distribution with $k-1$ numerator d.f. and $n_T - k$ denominator d.f.
- **Step 5:** Compare F-test statistic with rejection rule and make a conclusion.

Python Implementation: Two way ANOVA

- Package needed: `Statsmodels`.
- We first specify the model through the `formula` module, and fit the model using the `OLS` (ordinary least square) method.
- We use the function `statsmodels.stats.anova.anova_lm(*args, **kwargs)` to obtain the ANOVA table:
  - *args*: fitted linear model results instance.
  - *type*: "{1,2,3}". The type of ANOVA test to perform.
  - For more information, see the online document ▸ Document
- For low-intensity counties:

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

formula = 'two_weeks_cases ~ C(treated_T1) + C(urban) + C(treated_T1) : C(urban)'
data_low = data.loc[data['high_county_T1']==0]
model_low = ols(formula, data_low).fit()
anova_table = anova_lm(model_low, type=2)
anova_table
```

	df	sum_sq	mean_sq	F	PR(>F)
C(treated_T1)	1.0	3.256024e+04	3.256024e+04	0.928022	3.354575e-01
C(urban)	1.0	1.345125e+07	1.345125e+07	383.383722	2.735060e-80
C(treated_T1):C(urban)	1.0	5.319260e+04	5.319260e+04	1.516080	2.183133e-01
Residual	2906.0	1.019588e+08	3.508562e+04	NaN	NaN

Python Implementation: Two way ANOVA

- For high-intensity counties:

```
data_high = data.loc[data['high_county_T1']==1]
model_high = ols(formula, data_high).fit()
anova_table = anova_lm(model_high, type=2)
anova_table
```

	df	sum_sq	mean_sq	F	PR(>F)
C(treated_T1)	1.0	9.399691e+04	9.399691e+04	3.646061	5.629390e-02
C(urban)	1.0	1.234389e+07	1.234389e+07	478.809044	7.702414e-99
C(treated_T1):C(urban)	1.0	6.457622e+04	6.457622e+04	2.504857	1.135974e-01
Residual	3110.0	8.017704e+07	2.578040e+04	NaN	NaN

- Since in all cases the p-value is larger than 0.1, we cannot reject H_0.
- We may conclude that no significant difference in the effects of the intervention on COVID-19 cases between counties with different economic conditions either in high-intensity counties or low-intensity counties.
- **Questions (Homework Assignment)**:
  - What else may influence the result?
  - Can we further improve the experiment design? (e.g. use the proportion of infection instead of infections number . . .)