

# Financial Econometric Analysis

## Lecture 1 - Introduction

Fang-Chang Kuo

National Chung Cheng University

September 12, 2020

# Outline

- Introduction
  - Syllabus
  - Business Question and Solution: A/B testing
  - Python
  - Integrated Development Environment

# Syllabus

# Overview of this Course

- **Overview:**

- This is an advanced applied economics course at graduate level.
- I will introduce machine learning tools used in **marketing**, **economics**, and **data science**.
- We will apply these tools to analyzing various real and fake data-sets.

- **Goals:**

- provide students with knowledge and toolkit to
  1. conduct empirical analyses that involves inference and prediction,
  2. evaluate economic/business policies,
  3. read quantitative analyses by other practitioners
- learn at least one programming language: **Python**

- **Office Hour:**

- Office: 331
- Office hour: Sat. 11:00-13:00, or by appointment

# Course Structure

- **Grading Policy:**

- 60% Assignments
- 20% Midterm
- 20% Final Project (Presentation + Paper)

- **Assignments:**

- Please bring your laptops to class.
- Assignments are extremely important in this course.
- Students are expected to complete assignments during classes.

- **Midterm:**

- We will disclose more details about midterm later.

- **Final Project:**

- Just come up with a research agenda and to analyze the data using the tools and concepts discussed in class.

- **TA:**

- There will be few TAs to help students with their assignments in class.

# Course Materials

- **Slides:**

- This course does not have a required textbook. Course slides will be provided as main materials.

- **Reference:**

- J. M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 7th edition, 2019
- M. H. DeGroot and M. J. Schervish. *Probability and statistics*. Pearson Education, 4th edition, 2012
- J. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009
- P. Davis and E. Garcés. *Quantitative Techniques for Competition and Antitrust Analysis*. Princeton University Press, 2009

# Programming Language: Python

## ● Python References:

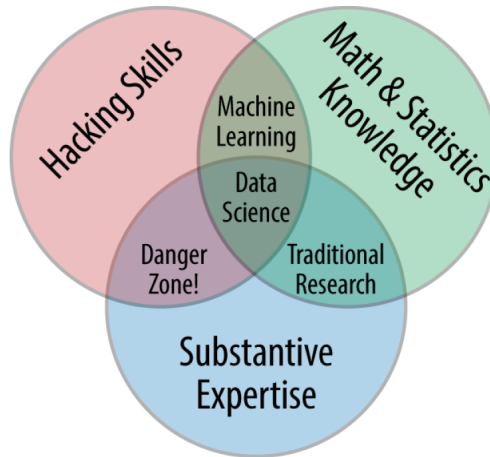
- W. Wesley. *Python for Data Analysis, 2nd Edition*.  
O'Reilly Media, Incorporated, 2017
- J. VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*.  
O'Reilly Media, Inc., 1st edition, 2016
- J. Grus. *Data Science from Scratch: First Principles with Python*.  
O'Reilly Media, 2019
- A. Müller and S. Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*.  
O'Reilly Media, 2016

# Course Plan

1. Introduction to Python (Numpy, Pandas, and Matplotlib)
2. Basic Data Structures, Plotting, and Visualization
3. Web Scraping (BeautifulSoup, Selenium)
4. Introduction to Machine Learning (Scikit-learn)
5. Supervised Learning: logit, KNN, Decision Tree, Random Forest, Support Vector Machine
6. Unsupervised Learning: K-means, PCA, Agglomerative Clustering



# Data Science?



# Business Question and Solution: A/B testing

# Example: SIMCITY

- Question:
  - How to display the promotional offer in order to drive more purchases and increase revenue?
- Which one do you prefer?
  - More likely to purchase the game.
- You can have many theories or ideas of how to design the layout.
- But we only care about real sales.
- So, directly getting consumers' feedbacks is the optimal strategy.

# Example: SIMCITY One Variation

The screenshot shows the EA SimCity website. At the top, there's a navigation bar with 'EA SIMCITY QUICK LINKS', 'REGISTER', 'LOGIN', 'HELP', and a language selector. The main header features the 'SIMCITY' logo, 'AVAILABLE NOW!', a 'NEWSLETTER SIGNUP' button, and social media links for Facebook, Twitter, and YouTube. Below the header is a secondary navigation bar with 'GAME INFO', 'NEWS', 'COMMUNITY', 'MEDIA', and a 'BUY NOW >' button. The main content area is divided into two columns. The left column is for the standard 'SIMCITY' edition, priced at \$59.99, with options for 'PC Download' and 'PC Physical', and a 'BUY NOW' button with an Origin icon. The right column is for the 'SIMCITY™ DIGITAL DELUXE EDITION', priced at \$79.99, with a 'PC Download' option and a 'BUY NOW' button with an Origin icon. Below the right column, a box titled 'DIGITAL DELUXE EDITION INCLUDES' lists four items: 'HEROES AND VILLAINS SET', 'FRENCH CITY SET', 'GERMAN CITY SET', and 'BRITISH CITY SET'. At the bottom of the main content area is a section titled 'Key Features'. The background of the website features a cityscape with a bridge and buildings.

EA SIMCITY QUICK LINKS REGISTER LOGIN HELP

# SIMCITY

 AVAILABLE NOW! NEWSLETTER SIGNUP f Like 28k Follow @SimCity

GAME INFO NEWS COMMUNITY MEDIA BUY NOW >

**SIMCITY™**



**\$59.99**

☒ PC Download  
☐ PC Physical

BUY NOW 

**SIMCITY™ DIGITAL DELUXE EDITION**



**\$79.99**

PC Download

BUY NOW 

**DIGITAL DELUXE EDITION INCLUDES**

-  HEROES AND VILLAINS SET
-  FRENCH CITY SET
-  GERMAN CITY SET
-  BRITISH CITY SET

## Key Features

# Example: SIMCITY The Other Variation

The screenshot shows the SIMCITY website with a dark header. The top navigation bar includes "SIMCITY QUICK LINKS", "REGISTER", "LOGIN", "HELP", and a language selector. The main header features the "SIMCITY" logo, the text "AVAILABLE NOW!", a "NEWSLETTER SIGNUP" button, and social media links for Facebook (Like), Twitter (25k), and a "Follow @SimCity" button. Below the header is a navigation menu with "GAME INFO", "NEWS", "COMMUNITY", "MEDIA", and a "BUY NOW >" button. The main content area features a large banner with the text "PRE-ORDER AND GET \$20 OFF YOUR NEXT PURCHASE" over a cityscape background with the Statue of Liberty. Below the banner are two product listings. The first listing is for "SIMCITY™" priced at "\$59.99", with options for "PC Download" (selected) and "PC Physical", and a "BUY NOW" button with a "Origins" logo. The second listing is for "SIMCITY™ DIGITAL DELUXE EDITION" priced at "\$79.99", with a "PC Download" option and a "BUY NOW" button with a "Origins" logo. At the bottom of the second listing, the text "DIGITAL DELUXE EDITION" is visible.

**SIMCITY™**

**\$59.99**

☒ PC Download  
☐ PC Physical

**BUY NOW**

**SIMCITY™ DIGITAL DELUXE EDITION**

**\$79.99**

PC Download

**BUY NOW**

**DIGITAL DELUXE EDITION**

# A/B testing

- **A/B testing** is a way to compare two versions of something to figure out which performs better.
- The aim of A/B testing is to enable you to make incremental improvements to your website or app.
- **A/B testing** is just a fancy terminology. Essentially, it is a randomized control trial.
- The outcome variable in a A/B test is binary and you can design specific feature, or layout for your own interest.
- It is like generating a almost ideal data-set for your own use.

## Example: SIMCITY Result

- Conversion Rate:
  - First Page: 10.2%
  - Second Page: 5.8%
- The difference is 43.4%.
- The promo banner actually hurt conversions.
- The results are somewhat counter-intuitive since the promo banner is meant to improve sales.
- However, it went the opposite way.
- It's good for them to run an A/B test just before full release.

# Notes

- To conduct a valuable A/B test, it's crucial that you limit changes to one variable.
- There are other considerations:
- How many traffics/observations do you need?
  - It is a trade-off of cost and accuracy.
- Is the difference statistically significant?
  - We will talk about this immediately.
- Is the difference economically significant?
  - Based on your estimates on sales difference



# Test of Differences in Proportion

- In statistics, A/B testing is just a two samples test in proportion/conversion.
- We can think of each variation as a Bernoulli trial where  $p_A$  is the probability of conversion for A.
- Let's say that  $N_A$  people see A, and that  $n_A$  of them click it.
- Then if  $N_A$  is large, we know that  $n_A/N_A$  is approximately a normal random variable with mean  $p_A$ .
- Similarly, for alternative B.

# Test of Differences in Proportion

- Let's review some basics of hypothesis testings.
- Hypothesis:  $H_0 : p_A = p_B$
- Standard deviation:

$$S_{p_A - p_B} = \sqrt{p(1-p) \left( \frac{1}{N_A} + \frac{1}{N_B} \right)}, \quad p = \frac{N_A p_A + N_B p_B}{N_A + N_B}$$

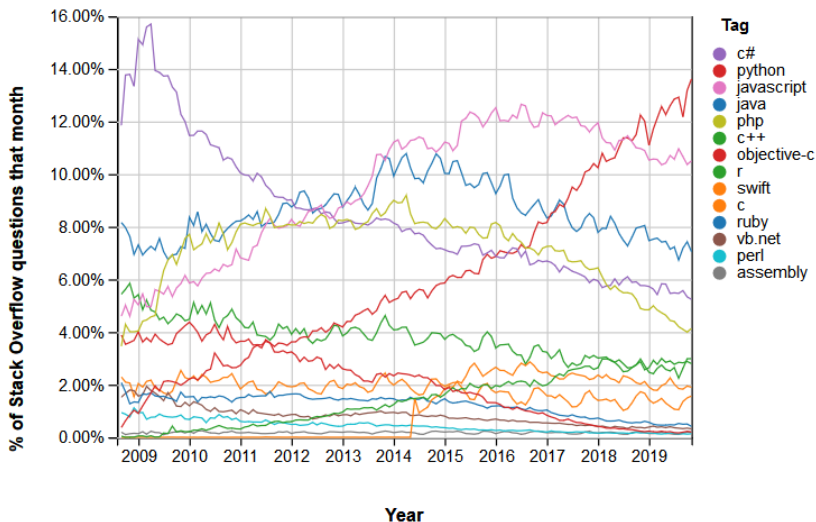
- We assume that null hypothesis is true and two samples are from the same distribution.

$$z = \frac{(p_A - p_B)}{S_{p_A - p_B}}$$

- We can reject the null hypothesis based on the z-statistic.
- I will upload **abtest.csv** for a A/B testing assignment after we talk more about Python and Pandas.

# Python

# Why Python? Stack Overflow Trend



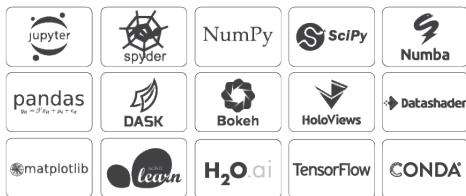
Sources: [Stack Overflow Trends](#)

# Python



- **Python** is an interpreted, high-level, general-purpose programming language first released 30 years ago.
- **Python** itself was not specifically designed with data analysis or scientific computing in mind.
- The usefulness of Python for data science stems primarily from the large and active ecosystem of **third-party packages**.
- There are different versions of Python. In this class, we will focus on Python 3.7 version.

# How to Install Python?



- <https://www.anaconda.com/distribution/#download-section>
- Important points:
  - Install the latest version. BTW, we use Python 3.8 version.
  - If you are asked during the installation process whether you'd like to make Anaconda your default Python installation, say yes.
  - Otherwise, you can accept all of the defaults.

# Integrated Development Environment

# Jupyter Notebooks



- **Jupyter notebooks** are one of the many possible ways to interact with Python and the scientific libraries.
- A browser-based interface supporting **Julia**, **Python**, and **R**:
  - The ability to write and execute commands.
  - Formatted output in the browser, including tables, figures, animation, etc.
  - The option to mix in formatted text and mathematical expressions.
- Because of these possibilities, Jupyter is fast turning into a major player in the scientific computing ecosystem.



# Jupyter Notebooks Environment

Untitled 2

<https://notebooks.azureml-int.net/n/sVUasfucctw/notebooks/Untitled%20.ipynb#>

jupyter Untitled 2 Last Checkpoint: 22 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help Python 2

Cell Toolbar: None

**This is a markdown cell used for documentation**

The above cell was written as: "### This is a markdown cell", followed by Shift+Enter

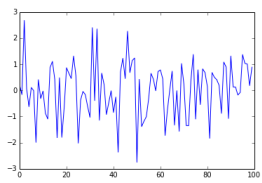
In [12]: `import math`  
`print "This is a code cell... and Pi is = ", math.pi`

This is a code cell... and Pi is = 3.14159265359

In [13]: `# to enable inline graphs, etc.`  
`%pylab inline`  
`plot(randn(100))`

Populating the interactive namespace from numpy and matplotlib

Out[13]: [matplotlib.lines.Line2D at 0x7fe2a2496cd0]



# Spyder



- If you are switching from **Matlab** or **Rstudio** to Python; **Spyder** is the way to go, It very intuitive for scientific computing.
- It is integrated into Anaconda. So, you don't need to install it separately.
- I like Spyder because of its **variable exploring** feature.
- Spyder feels more basic than other IDEs. It is not very fancy. But it can definitely get the job done.

# Spyder Environment

The screenshot displays the Spyder Python IDE environment. The main editor window shows a Python script named `2nd_panda.py` with the following code:

```

1 #!/usr/bin/env python3
2 #- coding: utf-8 -*-
3 ***
4 Created on Thu Jan 11 14:23:25 2018
5
6 @author: Shiv
7 ***
8 import csv
9 import statistics as st
10
11 subs_data = []
12 cust_data = []
13
14 def dominant_season_for_users(file):
15
16     # Define dictionary.
17     with open(file, 'r') as csvfile:
18
19         trip_reader = csv.DictReader(csvfile)
20
21         for row in trip_reader:
22             if row['user_type'] == 'Subscriber':
23                 subs_data.append(row['month'])
24             elif row['user_type'] == 'Customer':
25                 cust_data.append(row['month'])
26             else:
27                 'Not a valid user.'
28
29         subs_mode = st.mode(subs_data)
30         cust_mode = st.mode(cust_data)
31
32         return(subs_mode, cust_mode)

```

The Variable explorer on the right shows the following variables:

Name	Type	Size	Value
chicago	DataFrame	(72131, 5)	Column names: duration, month, hour, day_of_week, user_type
data	str	1	<a href="https://vega.github.io/vega-datasets/data/cars.json">https://vega.github.io/vega-datasets/data/cars.json</a>
df	DataFrame	(4, 3)	Column names: Genre, Rating, Gender
dur_df	dict	3	{'washington_dur':Series, 'chicago_dur':Series, nyc_dur':Series}
nyc	DataFrame	(276798, 5)	Column names: duration, month, hour, day_of_week, user_type
wash_mode	int64	1	6
washington	DataFrame	(66326, 5)	Column names: duration, month, hour, day_of_week, user_type

The IPython console shows the following execution:

```

In [39]: data
Out[39]: 'https://vega.github.io/vega-datasets/data/cars.json'

In [40]: chart = alt.Chart(data).mark_point().encode(
...:     x='Horsepower:Q',
...:     y='Miles_per_Gallon:Q',
...:     color='Origin:N',
...: )

In [41]: chart

In [42]:

```

The bottom status bar indicates: Permissions: RW, End-of-lines: LF, Encoding: UTF-8, Line: 7, Column: 1, Memory: 63 %.

# PyCharm



- PyCharm is an ideal IDE for professional developers working on large projects.
- Looks very fancy!!!
- There are tons of features. They are very helpful for debugging.
- However, PyCharm is a little bit resource-intensive. If have a computer with a small amount of RAM, you may want to use other lighter options.
- There two versions: community version is free.

# PyCharm Environment

The screenshot displays the PyCharm IDE interface with three main components:

- Editor (main.py):** Contains Python code for reading a CSV file and creating a DataFrame.
 

```

      usecols = [x.name for x in DATA_DICTIONARY if x.usecol]

      # dtypes should be a dict of 'col_name' : dtype
      dtypes = {x.name_: x.dtype for x in DATA_DICTIONARY if x.dtype}

      # same for converters
      converters = {x.name_: x.converter for x in DATA_DICTIONARY if x.converter}

      df = pd.read_csv('data/survey.csv',
                      header=0,
                      names=names,
                      dtype=dtypes,
                      converters=converters,
                      usecols=usecols)
      
```
- SciView: Data:** Shows a preview of the DataFrame 'df'.
 

	otherlang_js	otherlang_c	otherlang_php	otherlang_cs	otherlang_ruby	ot
0	True	False	True	False	False	
1	True	False	False	False	False	
2	True	False	False	False	False	
4	False	False	False	False	False	
5	True	False	False	False	False	
6	False	True	False	False	False	
- Python Console:** Shows the execution of the script.
 

```

      # Considering we now only have two categories left:
      # - Yes
      # - No, I use Python for secondary projects only
      # Let's turn it into a bool

      df['python_main'] = df['python_main'] == 'Yes'
      
```
- Special Variables:** A sidebar showing the state of variables in the current scope.
  - `DATA_DICTIONARY` = (list) <class 'list':>: [<survey\_c... View
  - `converters` = (dict) ('otherlang\_none': <function no... View
  - `df` = (DataFrame) python\_main ol... View as DataFrame
  - `dtypes` = (dict) ('python\_main': CategoricalDtype(cc... View
  - `names` = (list) <class 'list':>: ['python\_main', 'otherla... View
  - `usecols` = (list) <class 'list':>: ['python\_main', 'otherl... View