POLITICAL DATA SCIENCE

POLITICAL DATA SCICEN

An Open Introduction (Draft)

YUSHU LIU

Georgia Institute of Technology

They use data to make machine smarter,
make people sillier.

Contents

Preface

If I have seen further than others, it is by standing upon the shoulders of giants.

Isaac Newton

The world is full of books written by genius, and there are lots of details in their books were skipped, because they thought readers knew. The result is that you read their books, but hard to get the ideas which they really want to explain.

This book focuses on political data science with theories details explanations which try to help math allergic students to understand the principles of data science more easily. In this book, I enumerated several professors' works and also listed suggested which could help to learn more about data science.

Data science is beautiful, and it shouldn't become the privilege of someone who has good math, statistical and coding skills. Data science is a perspective to enjoy the landscape of human intelligence. It is also an exquisite way of thinking, which everyone has the rights to harness.

My mother language is Chinese. It is unavoidable there are some grammar mistakes or some other mistakes in this book. If you find any mistake, please let me know. I chose to write this book in English and to share it free on the Internet. Because I hope more and more people will have the opportunity to know and love data science.

YUSHU LIU

Beijing

December 11, 2017

Acknowledgements

I develop the ideas and perspectives presented in this book during I am studying the course ‘CS 6460: Educational Technology’ in Georgia Tech. I sincerely thank my mentor **Ken Brooks** and the course instructor **David Joyner**. Besides, I should thank all of my classmates who give me suggestions.

Introduction

Most students in China who major in political sciences are not good at data thinking in their own researches. The reasons are simple and easy to observe.

(1) In China, the full name of undergraduate political science program (UPSP) is major of political science and public administration. But most courses of this major are related with political science. There are 113 universities have UPSP [1]. All of these universities have statistics courses for these programs, but the professors who teach statistics don't know how to do political science research, most of them graduate from school of math, and they teach statistics knowledge but not data thinking.

(2) None of universities which have UPSP have data science system training for students.

(3) The students who have good skills in math and coding generally will not choose political science as their major. Because when graduate from this major students earn less and hard to find job compared with other highly demand math majors such as computer science, finance, and engineering etc. Meanwhile, students who choose political science as their major, have not much interesting in data analysis and data science.

How to teach these students data science? Especially the whole education system has not too much resources to support data science courses? In this book, I try to give a solution.

BACKGROUND AND RELATED WORK

Teachers are Asked to Use all Research Methods and Theories to Prove Socialism Values and Ideology are The Best Choices of Chinese people.

The ideology and values teaching are keys of UPSP teaching in China. Although there are 113 universities in China have UPSP, the courses of these programs are similar. And most of the universities learn the design of the UPSP course from the school of government, Peking University. According to the requirements of specialty training goals, the students of UPSP should be familiar with the basic principles of Marxism, the theory of socialism with

Chinese characteristics and the rule by the law of the communist party, and these basic qualities training serve the needs of socialist modernization [2].

Figure 0.1 shows the courses of UPSP in Peking University. These courses include three parts.

(1) Required courses of public fundamental.(2) Specialized courses. (3) Research methods courses. The part 2 and part 3 courses modules serve for the part 1. The political sciences theory should keep consistency with the ideology and values system of communist China. And the research methods should prove the system is the best choice of China. Basing on the logical connection of these courses, the research methods training of UPSP essentially serves for the ideology and values education.

These courses have a mission to prove a given proposition--why the communist system of China is the best choice of Chinese people. But data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data [3]. It means that the data science is problem-driven methods, these methods will push students to concern the real social phenomenon rather than ideology and socialist and communist values. To the authorities, it is obvious that the data science is not necessary for UPSP students in China. And the real data science course of UPSP should beyond the ideology limits.

Research Works of Data Science Methodologies in Political Science Research

Researches from Chinese Scholars

Many Chinese scholars believe that the political science should become "the real science" not only used for political aims (LV JIA, 2001) [5]. JI LIPING (2003) notes the UPSP courses reform is necessary and emergent. Most of the Chinese political science researchers haven't enough quantitative analysis training when they are students studying at universities. The result is that most of the political science scholars are not good at statistics and data analysis [6]. FENG ZHIFENG (2008) compared 5 political textbooks which published in U.S with 4 Chinese political science textbooks. He found that because the Chinese political science teaching has been affected by the Soviet Union for a long time, the Chinese political science textbooks more concerning on qualitative methods. The political textbooks which published in U.S. show various research methods which learn from other sciences, especially statistics and economics methodologies [7]. XIE TAO && LEE SIGELMAN (2008) first used the American Political Science Review to illustrate the methodological transformation in American political science in the 20th century, and that is the decline of qualitative methods and the rise of quantitative methods. They also examined 1257 papers which published in four Chinese political science journals (2004-2008) and found that qualitative methods dominated Chinese political science. In all these 1257 political sciences papers, only 22% have data in these papers, and only 12% have tables and figures as descriptive methods to explain their research. And only 2 of 1257 papers use the statistical inference models in their papers [8].

Researches from International Scholars

In 1981, the political scientist David Sanders began to argue that it is necessary to teach social scientists "data analyses" rather than "statistics". Instruction in statistical principles needs to be closely integrated with the use of appropriate software packages and secondary data sets. He also suggested that a discussion of interval level techniques such as correlation and regression should precede the introduction of the supposedly simpler technique of cross tabulation [9].

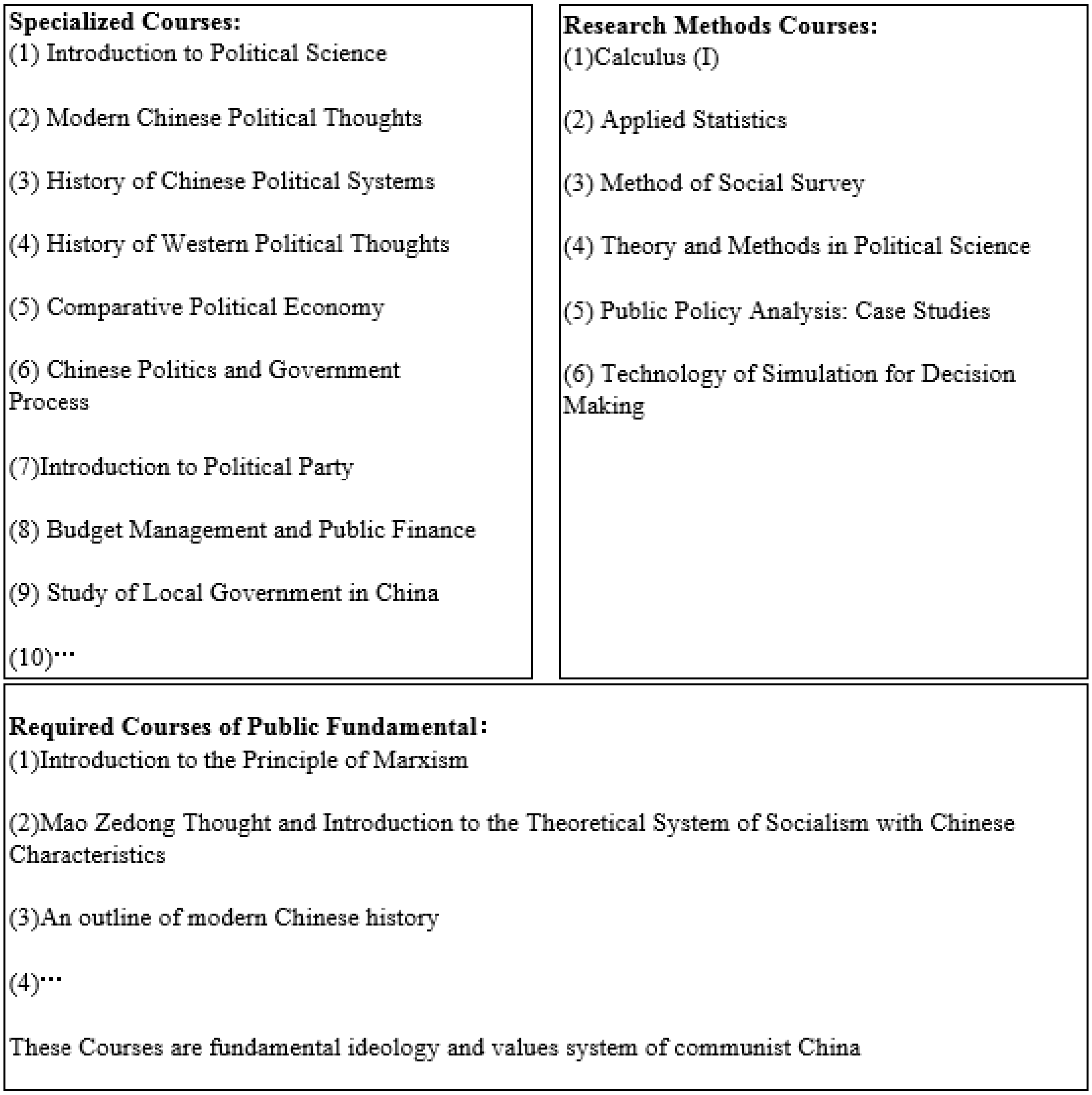
In 2006, Sigelman Lee analyzed 1 to 100 volumes of "*American Political Science Review*", and found that the *Review* has served mainly as a forum for reporting empirical findings, a purpose of almost two out of every three articles. Much of the early empirical research that appeared in the *Review* took the form of densely descriptive reports on particular people, places, processes, events, or institutions, which were treated as innately interesting rather than as "cases" of broader Phenomena. Today, the great majority of these analyses would be castigated as pedestrian exercises in barefoot empiricism [10].

In 2008, Garfield and Ben-Zvi noted that elementary teaching for non-statisticians may increase student's interest in statistics; or it may shy them away pretty fast and even once and for all. They argued that introductory courses should show the positive aspects of arrive to the other extreme of superficiality. It is therefore difficult to motivate students to engage in the hard work of learning statistics [11].

In 2013, Naivid Hassnpour demonstrated the power of text mining in political science research. He argued that digitization of text have opened venues for the analysis of political concepts by linking frequency and topic evolution to political developments. He showed that new tools can be used to test the validity of etymological claims about the history of political phenomena [12].

In 2015, Garfield, J. and Ben-Zvi argued strongly for the use of computers in teaching cannot replace a lack of statistical thinking. Teachers know that complicated computations often discourage students and make them sleepy at lectures. They try to alleviate the problem by offering statistical software as a crutch that might make students feel that even difficult parts of course can be navigated using simple controls hidden in the software. But the problem is that software can only be helpful to someone who has more or less mastered a substantial part of statistics [13].

In 2017, Helen Margetts argued that the new possibilities for generation and analysis of large-scale data sources from political activity is driving methodological change, requiring a variety of multi-disciplinary expertise, and presenting new ethical challenges. Expansion of the toolkit of political science methodologies to incorporate the growing field of computational social science which may shape the discipline in the future [14].



Specialized Courses:
(1) Introduction to Political Science

(2) Modern Chinese Political Thoughts

(3) History of Chinese Political Systems

(4) History of Western Political Thoughts

(5) Comparative Political Economy

(6) Chinese Politics and Government Process

(7)Introduction to Political Party

(8) Budget Management and Public Finance

(9) Study of Local Government in China

(10)…

Research Methods Courses:
(1)Calculus (I)

(2) Applied Statistics

(3) Method of Social Survey

(4) Theory and Methods in Political Science

(5) Public Policy Analysis: Case Studies

(6) Technology of Simulation for Decision Making

Required Courses of Public Fundamental:
(1)Introduction to the Principle of Marxism

(2)Mao Zedong Thought and Introduction to the Theoretical System of Socialism with Chinese Characteristics

(3)An outline of modern Chinese history

(4)…

These Courses are fundamental ideology and values system of communist China

Figure 0.1 Courses of UPSP, Peking University [4]

STUDENTS' PROBLEMS IN LEARNING DATA SCIENCE

Students Feel Frustration and Depression When They Learning Statistics

LONG FEI (2013) finished a survey about the anxiety of learning statistics of university students majoring in economy and management. He surveyed 600 students and recycled back 560 valid questionnaires. More than 23% undergraduate students who major in the economy and management science feel depression when they are studying Statistics. And 5% of these students will feel disgusting and extremely anxiety when study statistic [15].

Reasons Influencing Students Learning Statistics

ZHANG QINGJIE (2015) surveyed 120 students who registered the course of social statistics in South China Agricultural University, and he found that more than 80% students don't' like this course, and there are four reasons as follows [16]:

1. I think I haven't talent in math knowledge learning.

2. I think I am not good at logic thinking so I don't want to learn statistics.

3. I haven't any interesting in math- related knowledge learning.

4. I want to learn but I haven't enough math knowledge to learn statistics.

WANG YU (2017) surveyed 80 students in the school of sociology and public administration, East China University of Science and Technology. He found three reasons why students feel that learning statistics are very hard [17]

1. Students feel difficult to understand the inner connection of statistics knowledge.

2. Haven't good case analysis in statistic teaching, and students don't know how the statistics theory can use in their real life.

3. The teaching methods are boring.

SCHEMATIC OF THE MODERN STATISTICAL ANALYSIS PROCESS

Ben Baumer (2015) argued strongly for that a first course in statistics at the undergraduate level typically introduces students to a variety of techniques to analyze small, neat and clean datasets. However, whether they pursue more formal training in statistics or not, many of these students will end up working with data that are considerably more complex, and will need facility with statistical computing techniques. More importantly, these **students require a framework for thinking structurally about data** [18]. Figure 0.2 is Ben Baumer's understating about the statistical analysis process. He believed that the introductory statistics course (and in many cases, the undergraduate statistics curriculum) should emphasize the central column. At this point, I agree with Ben Baumer and this book content followed his logical design of modern statistical process.

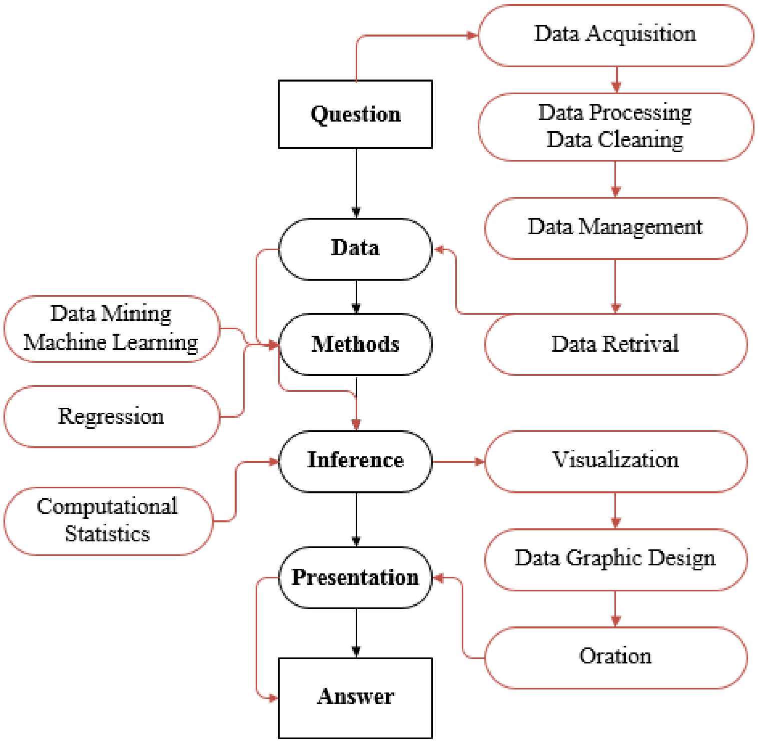


Figure 0.2 Schematic of The Modern Statistical Analysis Process

ABOUT THE BOOK

According to the modern statistical analysis process, the book “Political Data Science: An Open Introduction” tries to give a solution for students who want to learn data science with rare knowledge of math and statistical. Base on problem-driven logical, this book explains how to use data science in political science and shows the inner connection of knowledge.

Part I Problem Seeking

Chapter 1 is an introduction of empirical political sciences research. Chapter 2 is about how to use data-based critical thinking to find the research questions.

Part II Data collection and cleaning

Chapter 3-4 mainly discusses data collecting, data cleaning and data classification.

Part III Data analysis and inference

Chapter 5-8 includes data descriptive and inference, regression models, text mining and machine learning. These four chapters are mainly data science methods which used in political science research. I try to use more cases and fewer formulas to explain how to use these methods.

Part IV HCI political research frame

The modern dictatorship is an exquisite engineering. Based on the HCI control by the authority, they can prevent anyone who could points out the problems as a threat even before the problem rising. With the help of information technology, the top design of dictatorship system can become a perfect closed loop. And without the challenge from outsider, they could sustain forever. In the Chapter 9, we introduce a new research frame -- the HCI political research frame helps students to know their political environment. We hope students can have a new perspective which inspires them to find research questions.

EPILOGUE

The "Political Data Science: An Open Introduction" is a free dynamic book. And this is only a beginning edition. I will update chapter 10 to chapter 12 in the next edition. Chapter 10 is about data Visualization in political science research. The chapter 11 is about machine learning in political science research. And most importantly, the chapter 12, I will rise a discussion about how to write a political research paper. What is a good political science paper? Why some political research papers only pretend like a science research? These interesting problems I will share in the books' next edition.

REFRENCES

[1] The Chinese Universities Which Have Political Science and Public Administration Major. (2017, August 3). Retrieved September 18, 2017, from http://www.dxsbb.com/news/9594.html

[2] The Catalog of Undergraduate Education Peking University (2014). (2015, April 4). Retrieved September 18, 2017, from http://dean.pku.edu.cn/notice/content.php?mc=61431&id=1428041075

[3] Hayashi, C. (1998). What is data science? Fundamental concepts and a heuristic example. In Data Science, Classification, and Related Methods (pp. 40-51). Springer, Tokyo.

[4] Undergraduate Course Listing. (n.d.). Retrieved September 18, 2017, from http://english.pku.edu.cn/Admission/Undergraduate/Undergraduate.html?id=00032

The Catalog of Undergraduate Education Peking University (2014). (2015, April 4). Retrieved September 18, 2017, from http://dean.pku.edu.cn/notice/content.php?mc=61431&id=1428041075

[5] LV JIA(2001).Predicament and Outlet of Chinese Political Science Research Method，Journal of TAN SUO(China)，46-49.

[6] LIPING JI (2003). A survey of the Chinese Research Method of Political Science, Journal of Nantong Teachers College(Social Science Edition) 19(4), 33-34.

[7] ZHIFENG FENG (2008). On the political science study methods and reflect from the perspective of material analysis, Journal of Gansu Theory Research, 189(5). 151-156.

[8] XIE TAO(2008). American Political Science and Chinese Political Science—A Comparison of Methodology, ZHEJIANG SOCIAL SCIENCES, 2-12.

[9] Sanders, D. (1981). Teaching Social Science Data Analysis: A Political Scientist's View. Sociology, 15(4), 578-585.

[10] Sigelman, L. (2006). The coevolution of American political science and the American Political Science Review. American Political Science Review, 100(4), 463-478.

[11] Garfield, J., & Ben-Zvi, D. (2008). Developing students' statistical reasoning: Connecting research and teaching practice. Springer Science & Business Media.

[12] Hassanpour, N. (2013). Tracking the semantics of politics: A case for online data research in political science. PS: Political Science & Politics, 46(2), 299-306.

[13] Hindls, R., & Hronová, S. (2015). Are we able to pass the mission of statistics to students? Teaching Statistics, 37(2), 61-65.

[14] Margetts, H. (2017). The Data Science of Politics. Political Studies Review, 15(2), 201-209.

[15] LONG FEI (2013). Analysis of Anxiety of University Students Majoring in Economy and Management in Statistics Learning, China Education Innovation Herald, 255-256.

[16] ZHANG QINGJIE (2015). An analysis of Factors Influencing Academic Records of College Students in Liberal Arts—Taking Social Statistics Curriculum of South China Agricultural University as an Example, Journal of Higher Education of Sciences, 19-25.

[17] WANGY YU (2017). Applied statistics teaching reform based on the causes of learning difficulty—taking East China University of Science and Technology as an Example, Journal of Science of Teachers' College and University,37 (7),72-77.

[18] Baumer, B. (2015). A data science course for undergraduates: Thinking with data. The American Statistician, 69(4), 334-342.

CHAPTER 1 THE EMPIRICAL POLITICAL SCIENCE RESEARCH

It is important to understand that politics and political science are not exactly the same things. Political science focuses on the theory and practice of government and politics at the local, state, national, and international levels. Political Science Research dedicates to develop understandings of institutions, practices, and relations that constitute public life and modes of inquiry that promote citizenship [1]. Gregory M. Scott and Stephen M. Garrison (2007) noted that politics as a discipline focuses mostly upon the more complex levels of political interaction, and especially upon governments and activities related. But politics operate in all levels of life (family, community, and government) [2].

1.1 EMPIRICAL V.S. NORMATIVE RESEARCH

Political Scientists distinguish between obtaining knowledge and using knowledge. Dealing with how (and what) we know is termed empirical analysis. Dealing with how we should use our knowledge is termed normative analysis (Manheim, J. B., etc. 2008) [3].

To say that a question is an empirical question is to say that we will answer it – or try to answer it – by obtaining direct, observe information from the world, rather than, for example, by theorizing, or by reasoning, or by arguing from first principles. The essential ideal in empirical research is to use data as the way of answering questions, and of developing and testing ideas (Punch, K. F., 2013) [4].

Figure1.1 shows the data collection process when doing empirical research in political science. It means data is a very broad term. Before we do data analysis we should carefully distinguish the data classification.

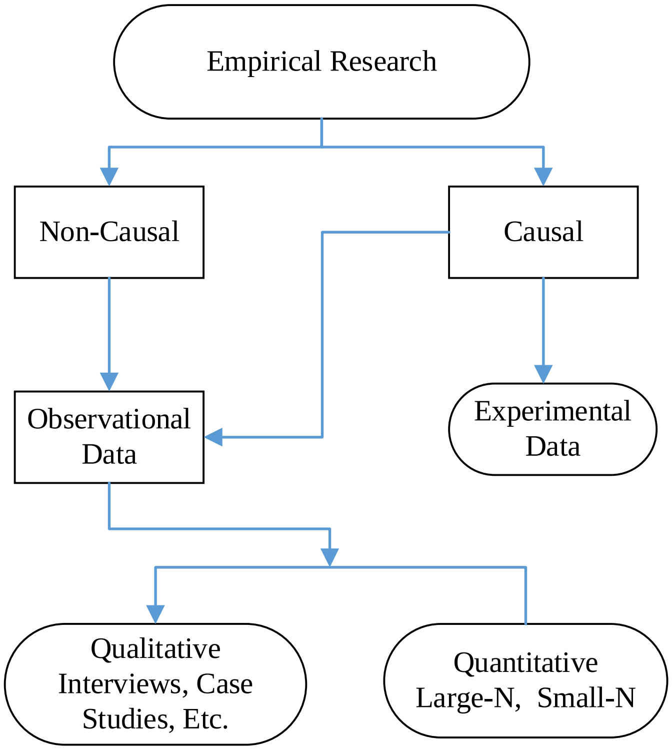

Empirical Research
Non-Causal
Causal
Observational Data
Experimental Data
Qualitative Interviews, Case Studies, Etc.
Quantitative Large-N, Small-N

FIGURE 1.1 THE POLITICAL EMPIRICAL RESEARCH PROCESS

1.2 QUANTITATIVE & QUALITATIVE DATA

In political science research fields, there are two main data types for empirical research. Keith F Punch made a very simplify definition as follow [5]:

Quantitative data – which are data in the form of numbers (or measurements).

Qualitative data – which are data not in the form of numbers (E.G, Words)

This leads to two simplifying definition:

Quantitative research is empirical research where the data are in the form of numbers.

Qualitative research is empirical research where the data are not in the form of numbers.

Punch (2013) strongly argued for that these simplified definitions are useful for getting started in research, but they do not give the full picture of the quantitative-qualitative distinction.

The term "quantitative research" means more than just research which uses quantitative or numerical data. It refers to a whole way of thinking, or an approach, which involves a collection or cluster or methods, as well as data in numerical form. Similarly, qualitative research is much more than just research which uses non-numerical data. It is also a way of thinking, or an approach, which similarly involves a collection or cluster of methods, as well as data in non-numerical or qualitative form.

In this book, I focus on political data analysis which mainly relates with quantitative research. Since there are many definitions, descriptions and conceptions of political research in the methodological literature, I agree with Punch (2013)'s opinions that it is sufficient for present purposes to see research as an organized, systematic and logical progress of inquiry, using empirical information-that is, data – to answer questions (or test hypotheses).

1.3 THE PROBLEMS of TRADITIONAL QUANTITATIVE METHODS

Punch (2013) shows a classic model of empirical social science research in his books (Figure 1.2) [6]. I read more than 20 social science research methods books, and all the empirical research models are similar. And they all stress the central role of research questions, and of using empirical data to answer those questions.

The problem is that in the real political research process, the most challenge part is that how to find a good research question. I mean the data methods should be not only used in the stage of answer the research question, but also need to use data-driven methods to find the interesting research questions.

I modified the classic model of empirical social science research (Figure 1.3). The data-based critical thinking is important in the pre-empirical research stage too. It means you need to evaluate data-based arguments when you do literature review or documents analysis. You should gather useful data, then turn data into research topics and research questions. The more details about data-based critical thinking will be discussed in chapter 2.

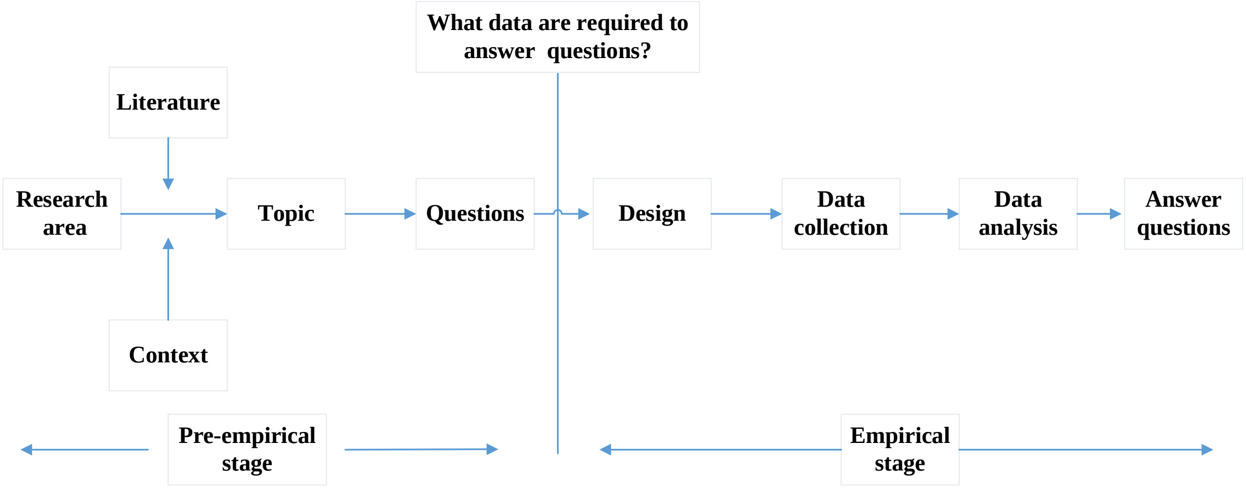


Figure 1.2 Simplified model of research

Source: Punch, K. F. (2013). Introduction to social research: Quantitative and qualitative approaches. Sage. P5

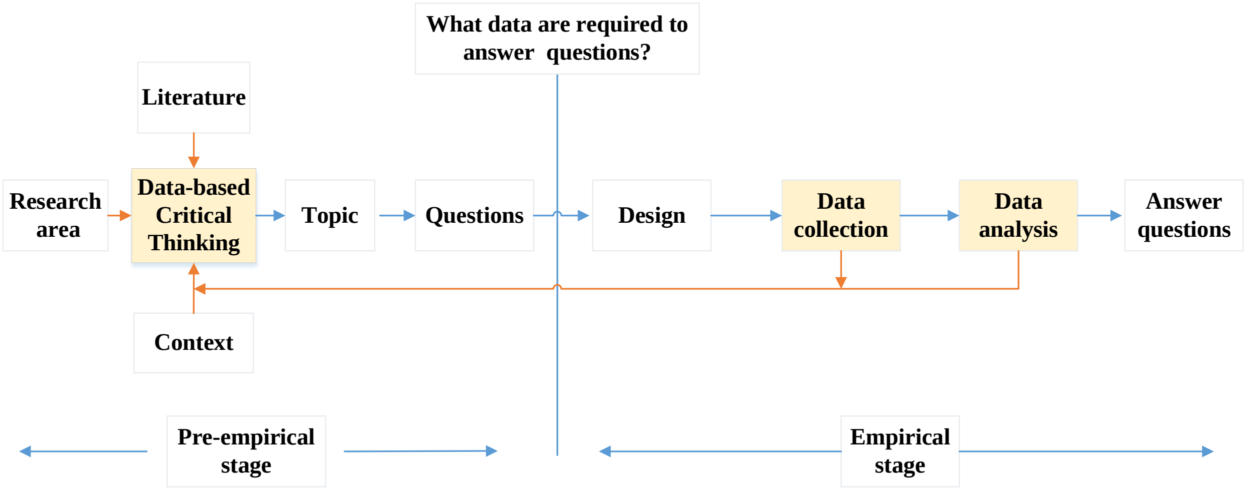


Figure 1.3 Data-driven quantitative methods

1.4 THE DEVELOPMENT OF QUANTITATIVE METHODS IN POLITICAL SCIENCE RESEARCH

1.4.1 The Quantitative Methods in Positive Political Science Research

Positive research is characterized by researches who remain detached from their subjects in order to remain "value free" in their investigations. They produce predictive theories based on experimentation, and emphasize mathematical and statistical methods in their analysis(McNabb,2015,16). And these also bring many challenges for political science researches. Smith et al noted these in his book:

> *Political science has changed substantially in the last thirty years. Many textbooks and most professional journals cannot be understood fully without some minimal acquaintance with the philosophy of science or social science and a wide range of empirical and statistical methods including computer science, various forms of statistical analysis, content analysis survey research, and many others.* (Smith et al. 1976)

Bartels and Brady (1993) examined more than 2000 articles in six of the most important political science journals and server different collections of papers. They concluded that quantitative methods still predominate in political science research (Bartels & Brady, 1993). Their topic organization scheme is repeated in Table 2.1.

But, as the development of the Internet, the big data resources have brought new challenge to traditional quantitative methods. Before that, the political scientists usually have the difficulties such as lacked quantitative transactional data, outside voting data (which exist only at the aggregate level), and apart from public policy where there are expenditure data (John and Margetts, 2004) and custom-built statistics. But in recent years, various people have argued that the availability of new sources of large-scale digital data could change the nature of political science research. And the data science attracts more and more attentions of political science researches.

Table 2.1 A Collection of Quantitative Methods Used in Political Science Research

1	Data Collection Methods a. Experiments b. Survey Designs c. Events Data
2	Time-Series Analysis a. Box-Tiao Intervention Methods b. Vector Auto regression c. Cointegration and Error Correction Methods d. The Kalman Filter
3	Time-Series of Cross-Section, Panels, and Aggregate Data a. Aggregate Data and Ecological Inference b. Time-Series Crossectional, Panel, and Pseudopanel Methods
4	Techniques Tailored to Measurement Properties of Data a. Event Count and Event History Models
5	Measurement Error and Missing Data a. The Consequences of Random and Nonrandom Measurement Error b. Guessing and Other Sources of Error in Survey Responses c. Nonrandom Samples and Sample Selection Bias d. Missing Data
6	Dimensional Analysis a. Voting Studies b. Perceptual Studies c. Legislative Studies
7	Model Specification a. Specification Uncertainty and the Perils of Data Dredging b. Sensitivity Analysis, Out-of-Sample Validation, and Cross-Validation
8	Estimation a. Maximum Likelihood Estimation b. Bootstrapping
9	Political Methodology and Political Science a. The Nature of Survey Response b. Economic Voting

Source: Bartels and Brady 1993.

1.4.2 The Big Data Challenge and Why Data Science is Needed in Political Science Research

The large-scale data sources (big data) which are now available provide the potential to conduct political science research in a way in which it has never been conducted before, and make it possible to test much more complex models and to understand complex patterns of behavior among relatively small sub-groups of the population (Dunleavy, 2016).

For example, the value of significance testing disappears and needs replacing with, for example, competitive model testing and large-N randomized control trials. A range of mathematical and computational methods are now described as "**data science**", including machine learning to develop algorithms that describe patterns in large-scale data, simulate variables that are unknown from the associations between those that are known and developing predictive capacity(Margetts, 2017); the development of huge N experiments run online, observing platform changes reflected in data in so-called natural experiments and testing the effects of interventions at scale(Bond et al.,2012; John,2017); and the growing use of genomic data to understand social and political behavior (Mills and Tropf, 2016; Settle et al., 2010).

Margetts (2017) noted that the big data bring three challenges to the political science research.

> *1. The first challenge is that the growing proportion of political activities which take place in digital environments requires large-scale data analysis.*
>
> *2. The other point is that the traditional methods of political science which are listed in the methods chapter must also transition to the digital world.*
>
> *3. The new forms of data in terms of methodological change driving philosophical change.*

Basing on the big data challenges, the new data science methods such as machine learning which is being developed explicitly for purposes of prediction, rather than explanation are changing the whole political science research ecology system.

But there is a paradox, Dowding (2015) strongly argued that political scientists do political science and that is the end of it. Physicists do not do political science and if they did they wouldn't be physicists (in Dowding's terms they would most likely be bad political scientists). In the other words, if a political scientist does research like a physicist, is it still a political scientist? Margetts (2017) pointed out that: **political science itself might be changing** – that work increasingly must take place in multidisciplinary teams if it is to be at the cutting edge. Margetts noted that:

> *"The fact is that increasingly social science IS rocket science, in the sense that engineers, physicists and researchers from other STEM (science, technology, engineering and mathematics) subjects are doing it. Developments in data science using data pertaining to people are taking place within mathematics, computer science, statistics and engineering departments, it is happening within newer institutions of 'big data' or 'data science' (both university centers and social media analytics firms), it is happening in the biggest Internet corporations and also across the newer field of computational social science. In this last development, there are more linkages with social science researchers, particularly in communications, sociology or political science departments – but in all, it is happening outside the mainstream of social science. This means that whatever Dowding or other political scientists think about causality, explanation and prediction, what is happening on the ground in these research communities outside political science will affect how these key concepts play out in practice."*

I strongly agree with Margetts that the political science itself has been changed with the Internet emerging. The data-driven political problems are pushing data science methods to play multiple roles in modern political analysis. And students need data science training to do data based critical thinking on traditional and new political questions.

1.5 THE CHALLENGES OF MODERN POLITICAL SCIENCE RESEARCH AND SOLUTIONS

The biggest problem in conducting a science of human behavior is not selecting the right sample size or making the right measurement. It's doing those things ethically (Bernard, 2012). The most often cited issue is privacy, the danger that somehow the data and analysis of that data by a researcher will lead to the revealing of people's names or identity to the public world (Margetts, 2017).

Margetts (2017) also argued that as well as introducing new ethical dilemmas in the process of doing research, such data also introduce new ethical challenges for the use of research in policymaking. When algorithms work on the basis of the probability of something happening – rather than whether it actually has happened – then we are in a different ethical landscape with regard to public policy. He made an example:

> *"If we have sufficient data to tell us that a pupil is statistically almost certain to drop out of school before they are 16, what should we do? Should we put more resources into that pupil, fewer resources, or should we operate behind Rawls' veil of ignorance and pretend that we do not know?"*

Besides, there are also some technical Challenge when you do quantitative analysis in political science research. For example, the errors and bias problems, measurement problems and manipulate Data problems.

There is no panacea for these challenges, and the solutions themselves are cutting edge research questions in political science. That’s also the reason why I want to write this book: to learn new methods, and to give new solutions.

REFRENCES

[1] Bond RM, Fariss CJ, Jones JJ, et al. (2012) A 61-Million-Person Experiment in Social Influence and Political

Mobilization. Nature 489 (7415): 295–298.

[2] Bartels, L. M., & Brady, H. E. (1993). The state of quantitative political methodology. *Political science: The state of the discipline II*, 121-59.

[3] Dunleavy P (2016) ‘Big data’ and Policy Learning. In: Stoker G and Evans M (eds) Methods That Matter:

Social Science and Evidence-Based Policymaking. Bristol: Policy Press, 143–168.

[4] Manheim, J. B., Rich, R. C., Willnat, L., & Brians, C. L. (2008). *Empirical political analysis: Quantitative and qualitative research methods*. Longman Pub Group.

[5] Margetts, H. (2017). The Data Science of Politics. *Political Studies Review*, *15*(2), 201-209.

Bond RM, Fariss CJ, Jones JJ, et al. (2012) A 61-Million-Person Experiment in Social Influence and Political

Mobilization. *Nature* 489 (7415): 295–298.

[6] Mills MC and Tropf FC (2016) The Biodemography of Fertility: A Review and Future Research Frontiers. In:

Hank K and Kreyenfeld M (eds) *Social Demography Forschung an der Schnittstelle von Soziologie und*

Demografie. Berlin: Springer Fachmedien Wiesbaden, 397–424.

[7] Punch, K. F. (2013). *Introduction to social research: Quantitative and qualitative approaches*. Sage.

[8] Scott, G. M., & Garrison, S. M. (1995). *The political science student writer's manual (6th Edition)*.

Prentice Hall.

[9] Settle JE, Dawes CT, Christakis NA, et al. (2010) Friendships Moderate an Association between a Dopamine

Gene Variant and Political Ideology. *The Journal of Politics* 72 (4): 1189–1198.

[10] Smith, B. L. (1976). *Political research methods: foundations and techniques*. Houghton Mifflin Harcourt (HMH).

[11] What is Political Science? (n.d.). Retrieved September 23, 2017, from https://www.polisci.washington.edu/what-political-science, Department of Political Science, UNIVERSITY OF WASHINGTON

CHAPTER 2 DATE BASED CRITICAL THINKING IN POLITICAL SCIENCE

One of the most important part of political science research is seeking good questions. To ask good questions you have to understand critical thinking. Critical is about finding the critical questions. These questions will help you shake out the bad assumptions and false conclusions that can help you to make real discoveries.

2.1 UNDERSTANDING CRITICAL THINKING

2.1.1 What is Critical Thinking

In Webster's New World Dictionary, the relevant entry reads "characterized by careful analysis and judgment" and is followed by the gloss, "critical — in its strictest sense — implies an attempt at objective judgment so as to determine both merits and faults." In this book, we use the definition which given by The Foundation for Critical Thinking[1]:

> *Critical thinking is that mode of thinking — about any subject, content, or problem — in which the thinker improves the quality of his or her thinking by skillfully analyzing, assessing, and reconstructing it. Critical thinking is self-directed, self-disciplined, self-monitored, and self-corrective thinking. It presupposes assent to rigorous standards of excellence and mindful command of their use. It entails effective communication and problem-solving abilities, as well as a commitment to overcome our native egocentrism and sociocentrism[2].*

In short, Critical thinking involves the evaluation of sources such as data, facts, observable phenomenon, and research findings. Good critical thinkers can draw reasonable conclusions from a set of information and discriminate between useful and less useful details for solving a problem or making a decision[3].

Asking interesting questions is a key part of critical thinking. The critical in critical thinking is about finding the critical questions. These are the questions that might chip away at the foundation of the idea.

Another key part of critical thinking is understanding the reasoning behind your research data or research ideas. Reasoning is your beliefs, evidence, experience, and values that support conclusions. It's always important to keep track of every research paper's reasoning when do your own political science research.

For example, we always hear these words: "You should trust government because the government serves for the people. " This statement has two parts, the idea and the reasoning. The idea here, is that you should trust government. The reasoning is that you should do it because the government serves for the people. Now let's think about this statement critically. Why does it serve for the people? How do you know it serves for the people? Does it serve for everyone? How many sources agree that the government serves for the people? What does the research say? If you don't apply critical thinking, then you're left with just the idea.

In our real political science research life, trying to find the critical questions isn't something you can do all the time. It's a little like swimming. Most people can do a little, and with some exercise, they can do a little more. In next sections, we will introduce the key components of critical thinking and how can we exercise our critical thinking ability.

2.1.2 Key Components of Critical Thinking

According to the research of Pennsylvania Child Welfare Resource Center of University of Pittsburgh[4], there are 8 key components of critical thinking (table 2.1). Understanding these key components which can help you know how to critical thinking in your political science research.

Table 2.1 Key Components of Critical Thinking

The process of thinking is as significant as the outcome	Critical thinkers know that it is the journey to a conclusion that must be carefully evaluated.
Be attuned to your own beliefs, values, and prejudices as well as your personal experiences	Critical thinkers are self-aware and self-critical, and acknowledge that their own beliefs, values, and personal experiences play a role in the way they think and act. They may trust their feelings and instincts but they recognize that they are not tantamount to truth.
Challenge assumptions	A critical question is one that chips away at an idea and reevaluates assumptions. Critical thinkers accept that it is essential to ask probing and clarifying questions in order to assess assumptions, implications, and consequences and evaluate evidence against established criteria and standards.
Consider the arguments	Critical thinkers evaluate the veracity of the premises and conclusions of an argument and understand that even widely accepted conclusions are not necessarily accurate or true and credible information is not necessarily factual.
Consider alternatives	Critical thinkers think open mindedly and see things from other perspectives, encouraging discussion and consideration of opposing points of view.
Consider the context	Critical thinkers recognize that thoughts and actions are shaped

	by such things as culture, race, age, sex, and sexual orientation as well as by personal experiences.
Know the sources of information	Critical thinkers carefully evaluate the sources of information, considering the nature of their expertise and being cognizant of any signs of bias or conflicts of interest.
Assume a posture of reflective skepticism	Critical thinkers are disciplined thinkers who take nothing for granted but rather challenge themselves and others to assess their own thinking, reframe questions, recognize ambiguities, test and retest findings and interpretations, and acknowledge fallacies – their own as well as others.

Source: The Pennsylvania Child Welfare Resource Center of University of Pittsburgh

2.1.3 Questions Which Can Help You to Exercise Critical Thinking

The global digital citizen foundation (GDCF) designed a critical thinking skills cheatsheet [5]. It’s a simple table offering questions that work to develop critical thinking on any given topic. Whenever you discover or talk about political science research topics, you can use these questions for sparking debate or find the critical questions in your research.

Table 2.2 The GDCF Critical Thinking Skills Cheatsheet

Who	…benefits from this? …is this harmful to? …makes decisions about this? …is most directly affected?	…have you also heard discuss this? …would be the best person to consult? …will be the key people in this? …deserves recognition for this?
What	…are the strengths/weaknesses? …is another perspective? …is another alternative? …would be a counter-argument?	…is the best/worst case scenario? …is most/least important? …can we do to make a positive change? …is getting in the way of our action?
Where	…would we see this in the real world? …are there similar concepts/situations? …is there the most need for this? …in the world would this be a problem?	…can we get more information? …do we go for help with this? …will this idea take us? …are the areas for improvement?
When	…is this acceptable/unacceptable? …would this benefit our society? …would this cause a problem? …is the best time to take action?	…will we know we’ve succeeded? …has this played a part in our history? …can we expect this to change? …should we ask for help with this?
Why	…is this a problem/challenge? …is it relevant to me/others? …is this the best/worst scenario? …are people influenced by this?	…should people know about this? …has it been this way for so long? …have we allowed this to happen? …is there a need for this today?

How	…is this similar to____? …does this disrupt things? …do we know the truth about this? …will we approach this safely?	…does this benefit us/others? …does this harm us/others? …do we see this in the future? …can we change this for our good?

Source: Global Digital Citizen Foundation.

2.2 DATA BASED CRITICAL THINKING IN POLITICAL SCIENCE

2.2 1 Confirm The Political Science Research Purpose

There are four kinds of political science research purpose (Scott, G. M., & Garrison, S. M. ,1995) [6].

i. to explain

ii. to evaluate

iii. to predict

iv. to persuade

After you select a topic to decide what your research is for. If you want to explain, you could to write an analytical paper. If you want to evaluate government, you can do comparing government research. The policy analysis papers usually are used to predict the policy results. And, public opinion survey papers generally are used to persuade. Generally speaking, most of political science research papers are analytical. And, political scientists are most often engaged in three activities. They collect data, analyze it, and interpret it (Scott, G. M., & Garrison, S. M. ,1995).

2.2.2 Ask the Good Question

The direct way to do data based critical thinking in political analysis is to think as a data scientist. As we mentioned in previously part, the most important thing in critical thinking is to ask a good question. Brain Godsey (2017)[7]· recommended some useful ways to finding a good question in his book "*Think Like a Data Scientist*".

> "*Good questions are concrete in their assumptions. No question is quite as tricky to answer as one that's based on faulty assumptions. But a question based on unclear assumptions is a close second. Every question has assumptions, and if those assumptions don't hold, it could spell disaster for your project. It's important to think about the assumptions that your questions require and decide whether these assumptions are safe. And in order for you to figure out if the assumptions are safe, they need to be concrete, meaning well defined and able to be tested.*"

Brain Godsey (2017) strongly argued that a data set will tell us no more than what we ask of it, and even then, the data may not be capable of answering the question. These are the two most dangerous pitfalls:

Expecting the data to be able to answer a question it can't

Asking questions of the data that don't solve the original problem

Asking questions that lead to informative answers and subsequently improved results is an important and nuanced challenge that deserves much more discussion than it typically receives. The examples of good, or at least helpful, questions I've mentioned in previous sections are somewhat specific in their phrasing and scope, even if they can apply to many types of projects. In the following subsections, I attempt to define and describe a good question with the intent of delivering a sort of framework or thought process for generating good questions for an arbitrary project. Hopefully you'll see how it might be possible to ask yourself some questions in order to arrive at some useful, good questions you might ask of the data.

2.2.3 The Problems of General Steps of Writing Political Science Papers

Gregory Scott and Stephen Garrison (2007) demonstrated a general writing steps in writing political science papers[8], which are well-recognized by political science professors.

Step1. Select or Identify the Variables You Wish to Study

Step2. Conduct a Literature Review

Step3. Formulate a Research Question and a Hypothesis

Step4. Construct or Adopt a Methodology

Step5. Data Collection

Step6. Data Analysis

Step7. Compose the Narrative(Thesis)

Just as we mentioned in the chapter 1 of this book (Chapter1, Figure1.2-1.3), the simplified model of research (Or the traditional research steps) is usually ask questions and then collect data. But the problem is that if you haven't enough data and systematic analysis, it is hardly to find the interesting research questions. So, according to the modern statistical analysis process (Introduction, Figure2), we give a reference process of how use data based critical thinking in general steps of writing political science papers (Figure 2.2).). As you can see, the data based critical is an empirical research cycle: find problem from data and test them, explain them until you find the key variables.

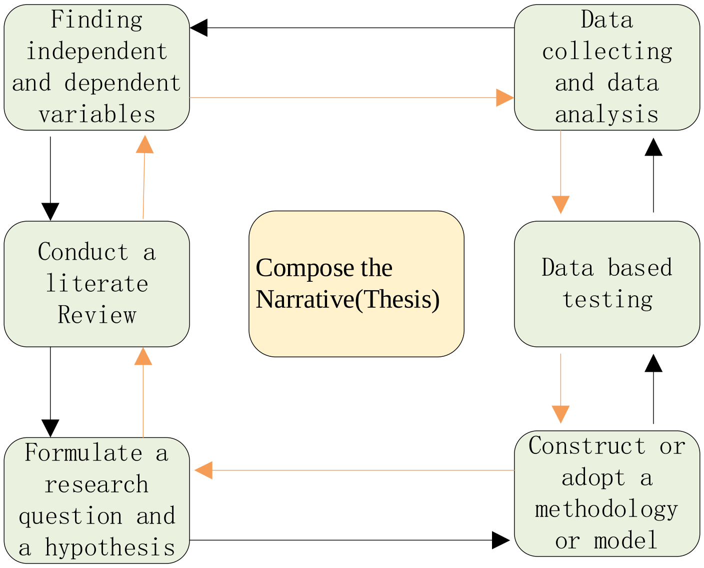

Finding independent and dependent variables
Data collecting and data analysis
Conduct a literate Review
Compose the Narrative(Thesis)
Data based testing
Formulate a research question and a hypothesis
Construct or adopt a methodology or model

Figure 2.1 The Data Based Critical Thinking in Writing Political Science Papers

And the end, when you find the key variables, a good narrative writing is also very important in political science research paper. The narrative writing will be composed of three elements (Gregory Scott and Stephen Garrison, 2007)[9]:

i. Findings: explain what data tells you and what it does not tell you with respect to the research question you are trying to answer.

ii. Interpretation of findings: discerning patterns in the facts, and the processes that cause, control, or effect these patterns. Gregory Scott and Stephen Garrison strongly suggested that do not draw conclusion that are not warranted by the data. Try to be as precise as you can in stating the conclusions you have drawn from the information you have found.

iii. Areas of further research: Speculating on what you might find if you had more or better data. Or you can suggest new research questions and areas of study that may be helpful in understanding your selected subject matter in the future.

REFRENCES

[1] The Foundation for Critical Thinking belongs to The International Center for the Assessment of Higher Order Thinking (ICAT). This foundation was founded to help colleges and universities design cost-effective ways to evaluate students' critical thinking abilities. Website: http://www.criticalthinking.org/]

[2] Criticalthinking.org. (2017). Our Conception of Critical Thinking. [online] Available at: http://www.criticalthinking.org/pages/our-concept-of-critical-thinking/411 [Accessed 11 Dec. 2017]

[3] Doyle, A. (2016, November 28). Critical Thinking Definition, Skills, and Examples. Retrieved November 1, 2017, from https://www.thebalance.com/critical-thinking-definition-with-examples-2063745

[4] Pacwrc.pitt.edu. (2017). Cite a Website - Cite This For Me. [online] Available at: http://www.pacwrc.pitt.edu/Curriculum/707_CrtclThnkngTheTrnrsRl/Hndts/HO05_KeyCmpnntsOfCrtclThnkng.pdf [Accessed 11 Dec. 2017].

[5] Watanabe-Crockett, L. (2017). The Critical Thinking Skills Cheatsheet [Infographic]. [online] Global Digital Citizen Foundation. Available at: https://globaldigitalcitizen.org/critical-thinking-skills-cheatsheet-infographic [Accessed 11 Dec. 2017].

[6] Scott, G. M., & Garrison, S. M. (1995). The political science student writer's manual. Prentice Hall. pp248-249

[7] Godsey, B. (2017). Think like a data scientist-tackle the data science process step-by-step. Manning Publications.

[8] Scott, G. M., & Garrison, S. M. (1995). The political science student writer's manual. Prentice Hall. pp249-251

[9] Scott, G. M., & Garrison, S. M. (1995). The political science student writer's manual. Prentice Hall. P6&P251

CHAPTER 3 ONLINE POLITICAL DATA COLLECTION

Quantitative political science research depends on establishing linkages between questions and data. Before we start a research question, naturally, we will consider where to find the data that we need. In this chapter we focus on how to collect political science research data online.

3.1 DATA TYPES AND THREE CRITERIA OF DATA-GATHERING

There are three data criteria for political research data gathering (Louise, 1994)[1]

i **Validity**: Whether the data are relevant to the meaning of the concept?

ii **Reliability**: Would the same data be collected under varying conditions?

iii **Feasibility**: Are they realistically obtainable?

With these three criteria of political research data gathering, actually many researchers are unsure where they can obtain data to begin their research and analysis. We should first divide the different types of data into two major classifications [2].

i **Primary Data:** When talking about "primary data", it refers to data collected by the researcher himself/herself. This data has never been gathered before, whether in a particular way, or at a certain period of time. Researchers tend to gather this type of data when they cannot be find from outside sources. You can tailor your data questions and collection to fit the need of your research questions. It can be an extremely costly task and, if associated with a college or institute, requires permission and authorization to collect such data. Issues of consent and confidentiality are of extreme importance. Primary data actually follows behind secondary data because you should use current information and data before collecting more so you can be informed about what has already been discovered on a particular research topic.

ii **Secondary Data:** If the time or hassle of collecting your own data is too much, or the data collection has already been done, secondary data may be more appropriate for your research. This type of data typically comes from other studies done by other institutions or organizations. There is no less validity with secondary data, but you should be well informed about how it was collected. In this book, we mainly discuss the collection of secondary data which are from the Internet.

3.2 DATA SOURCES

3.2.1 Chinese Political Data Sources

According to the "People's Republic of China Network Security Law[3] " and "the Law of the People's Republic of China on Guarding State Secrets[4] ", most government data are classified. But China have good cooperation with some International organizations and world-wide famous Universities. So, you can find useful data about China in these organizations and universities websites (Table 3.1).

Table 3.1 Recommend Website for China Political Research Data Collection

Name	Website	Notes
National Bureau of Statistics of China	http://www.stats.gov.cn/english/	The authority agency for China government's data publishing.
China data center, University of Michigan	http://chinadatacenter.org/	The center partners with several Chinese government agencies and companies in distributing China statistical data and publications and providing data services outside of China.
China Data Online	http://chinadataonline.org/	A Chinese data website of China Data Center University of Michigan. In these website, you can get almost all opening data set about China.
The World Bank Data	https://data.worldbank.org/country/china	Including GDP, population, climate, Finance data of China.
OECD data	https://data.oecd.org/	Including population, climate, GDP, economy, education, environment, finance, government, jobs, society, agriculture data set of China.

3.2.2 Mainly United States Data Sources for Political Science Researcher

Table 3.2 shows the popular websites for U.S. political research data collection. The U.S. governments offer abundance of data on the Internet. And these data generally can be downloaded from these websites directly.

Table 3.2 Recommend Website for U.S. Political Research Data Collection[5]

Name	Website	Notes
American FactFinder	Factfinder.census.gov	Basic U.S. Census facts
Center for Responsive Politics	www.opensecrets.org	A non-partisan and non-profit group that tracks money in politics, and its effect on elections and public policy
CIA World Factbook	www.cia.gov/cia/publications/factbook	Provides reliable international statistics and background information about foreign nations compiled by the CIA
Congressional Quarely	www.cq.com	This website has news about Capitol Hill and provides online access to information about pending bills and political candidates at federal and state levels.
DefenseLINK	www.defenselink.mil	The official Web site for the U.S. Department of Defense and a good starting point for finding U.S. military information online
Documents Center	www.lib.umich.edu/govdocs	A central reference and referral point for U.S. and foreign government information.
ElectionGuide.org	www.electionguide.org	Provides information on all national-level presidential, parliamentary, and legislative elections in other countries. It also has links to each country's election authorities, summaries of election results, and data on voter turnout.
Elections Around the World	www.electionworld.org	A comprehensive (but unofficial) guide to elections in every independent country and overseas dependency.
FedStats	www.fedstats.gov	A gateway to statistics from over 100 U.S. federal agencies.
FedWorld Information Network	www.fedworld.gov	Provides online access to millions of U.S. government pages, files, and databases.
FirstGov	www.fisrtgov.gov	Provides the most comprehensive search of government information on the Internet. It has links to every federal agency and state government.

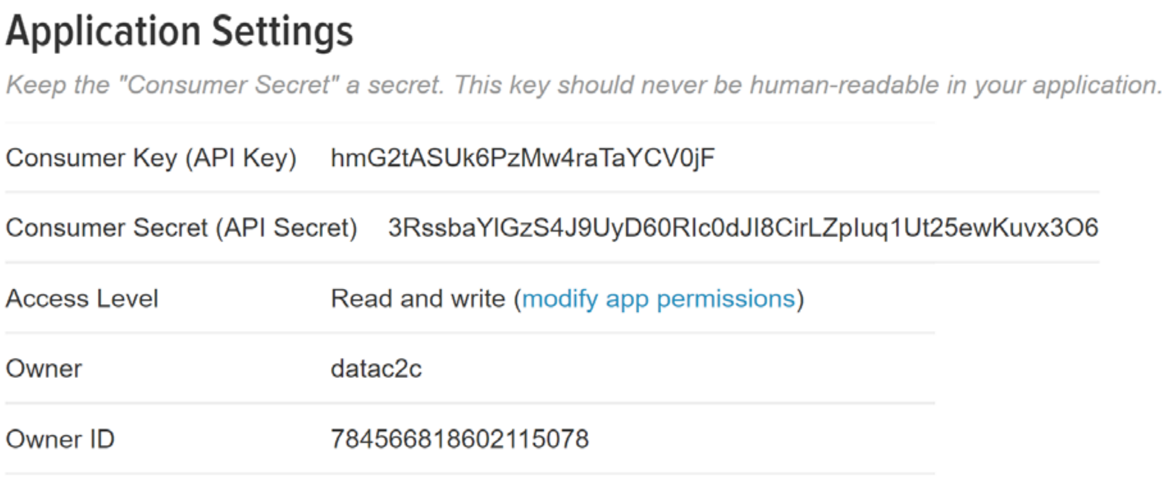
GovSpot.com	www.govspot.com	A non-partisan government information portal designed to simplify the search for the best and most relevant government information online.
GPO Access	www.gpoaccess.gov	A service of the U.S. Government Printing Office and provides online access to many of the government's best information products, including the Federal Register, the Commerce Business Daily, the Code of Federal Regulations, the Congressional Record, and the United States Code.
Library of Congress	www.loc.gov	Contains the holding of the U.S. National library, which includes all items holding a U.S. copyright. It also provides online access to the Copyright database, the American Memory site, and a link to THOMAS, the library's congressional online site(Thomas.loc.gov).
National Archives and Records Administration	http://www.nara.gov	A huge collection of official records from the executive, legislative, and judicial branches of the federal government. It also has links to all presidential archives.
National Journal	www.nationaljournal.com	A magazine that covers the U.S. Congress. It offers online access to polling data and House and Senate races.
Political Information	www.politicalinformation.com	A searchable collection of more than 5000 Web sites on U.S. politics, policy, and political news.
Political Resources on the Net	www.politicalresources.net	A searchable collection of more than 5000 Web sites on U.S. politics, policy, and political news.
Project Vote Smart	www.vote-smart.org	A searchable directory of international political Web sites with links to parties, organizations, governments, and the media in each country.
Public Agenda	www.publicagenda.org	Provides non-partisan information and academic research on what the public thinks about important U.S. policy issues
Public Register's Annual Report Service	www.prars.com	Provides access to annual reports of publicly traded companies
U.S. Securities and Exchange Commission	www.sec.gov	Provides reliable information about companies and industry trends.

U.S. Census Bureau	www.census.gov	Provides online access to the official U.S. Census data and statistical information on U.S. business.
White House	www.whitehouse.gov	Provides online access to transcripts of presidential speeches, press briefings, proclamations, and executive orders.
DATAUSA	https://datausa.io/	Data USA puts public US Government data in your hands. Instead of searching through multiple data sources that are often incomplete and difficult to access, you can simply point to Data USA to answer your questions.
Data.gov	https://www.data.gov/	Data.gov is managed and hosted by the U.S. General Services Administration, Data.gov is home of the U.S. Government’s open data
Usa.gov	https://www.usa.gov/	USA.gov is the U.S. government's official web portal to all federal, state, and local government web resources and services.

3.3 COLLECTING DATA FROM API

APIs are application programming interfaces. For example, most internet companies like Twitter or Facebook will have an application programming interface where you can download data. For example, you can get data about, what users are tweeting, or what they're tweeting about. Twitter is a very good source of text data for political analysis. Take Download Trump’s Twitter as an example, we will show how to do information mining text data from Twitter.

To extract the tweets, we need to log in to https://dev.twitter.com/ and then click on "Create New App" (by clicking on managing your apps). And, than follow the instructions, fill in information to get yourself a API Key, API secret (Figure 3.1), Access Token information (Figure 3.2). Execute the following code and you are all set with extracting tweets.

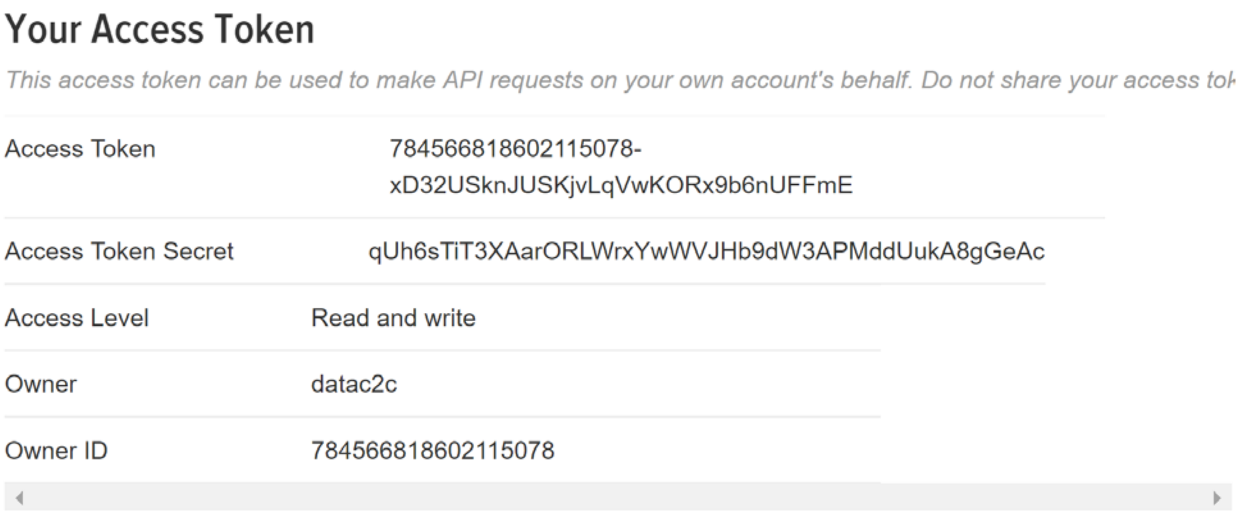


Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	hmG2tASUk6PzMw4raTaYCV0jF
Consumer Secret (API Secret)	3RssbaYlGzS4J9UyD60Rlc0dJl8CirLZpluq1Ut25ewKuvx3O6
Access Level	Read and write (modify app permissions)
Owner	datac2c
Owner ID	784566818602115078

Figure 3.1 API Application Settings



Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access tok

Access Token	784566818602115078- xD32USknJUSKjvLqVwKORx9b6nUFFmE
Access Token Secret	qUh6sTiT3XAarORLWrxYwWVJHb9dW3APMddUukA8gGeAc
Access Level	Read and write
Owner	datac2c
Owner ID	784566818602115078

Figure 3.2 Twitter Access Token

Code demo of extract Trump's Twitter by twitteR package.

# install packages

```
install.packages("twitteR")
install.packages("RCurl")
install.packages('base64enc')
install.packages("httk")
install.packages("httpuv")
```

# library packages

```
library(base64enc,dependencies = T)
library(twitteR)
library(ROAuth)
library(RCurl)
library(devtools)
library(httk)
library(httpuv)
```

# Setting Pacakges

```
pkgs <-c('twitteR','ROAuth','httr','plyr' ,'stringr')
for(p in pkgs) if(p %in% rownames(installed.packages()) == FALSE) {install.packages(p)}
for(p in pkgs) suppressPackageStartupMessages(library(p, quietly=TRUE,
character.only=TRUE))
```

# Set API Keys

```
options(httr_oauth_cache=T)
api_key <- "hmG2tASUk6PzMw4raTaYCV0jF"
api_secret <- "3RssbaYlGzS4J9UyD60RIc0dJI8CirLZpIuq1Ut25ewKuvx3O6"
access_token <- "784566818602115078-xD32USknJUSKjvLqVwKORx9b6nUFFmE"
access_token_secret <- "qUh6sTiT3XAarORLWrxYwWVJHb9dW3APMddUukA8gGeAc"
setup_twitter_oauth(api_key, api_secret, access_token, access_token_secret)
```

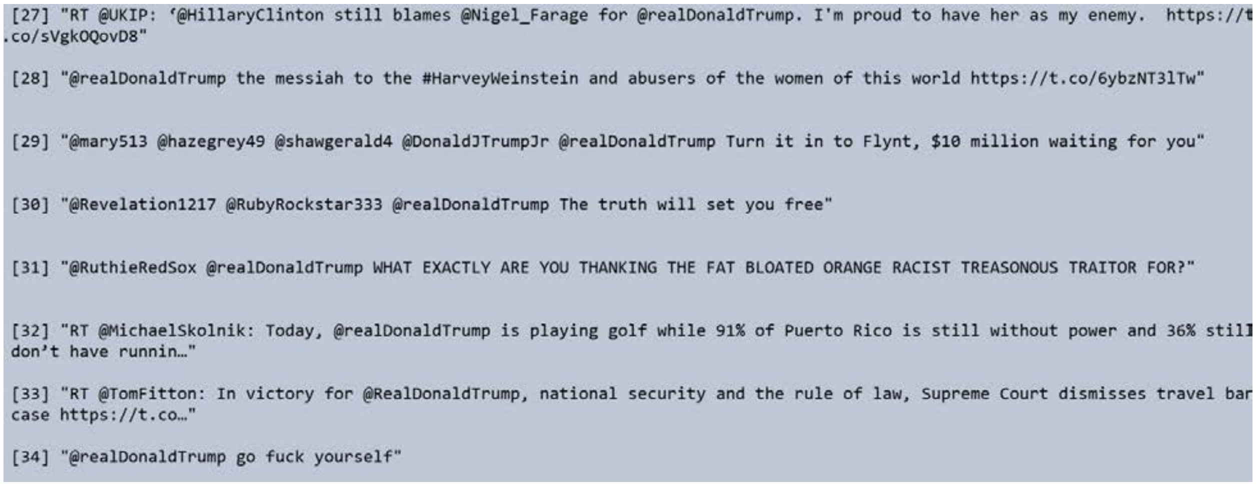
# Grab latest tweets

```
tweets_trump <- searchTwitter('@realDonaldTrump', n= 500)
```

# extract text

```
feed_trump <- laply(tweets_trump, function(t) t$getText())
```

# Collected Data Display(pratly, Figure 3.3)



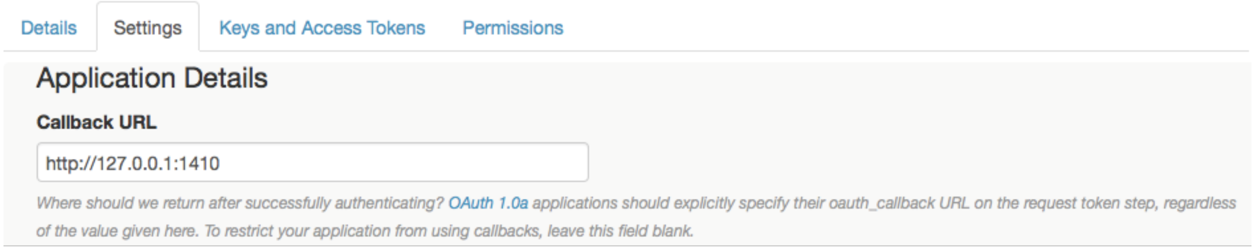
```
 [27] "RT @UKIP: ‘@HillaryClinton still blames @Nigel_Farage for @realDonaldTrump. I'm proud to have her as my enemy.  https://t
.co/sVgkOQovD8"

 [28] "@realDonaldTrump the messiah to the #HarveyWeinstein and abusers of the women of this world https://t.co/6ybzNT3lTw"

 [29] "@mary513 @hazegrey49 @shawgerald4 @DonaldJTrumpJr @realDonaldTrump Turn it in to Flynt, $10 million waiting for you"

 [30] "@Revelation1217 @RubyRockstar333 @realDonaldTrump The truth will set you free"

 [31] "@RuthieRedSox @realDonaldTrump WHAT EXACTLY ARE YOU THANKING THE FAT BLOATED ORANGE RACIST TREASONOUS TRAITOR FOR?"

 [32] "RT @MichaelSkolnik: Today, @realDonaldTrump is playing golf while 91% of Puerto Rico is still without power and 36% still
 don’t have runnin…"

 [33] "RT @TomFitton: In victory for @RealDonaldTrump, national security and the rule of law, Supreme Court dismisses travel bar
 case https://t.co…"

 [34] "@realDonaldTrump go fuck yourself"
```

Figure 3.3 Collected Data Display

One bug could happen when you use these codes as follow[6]:

```
[1] "Using direct authentication"
Error in check_twitter_oauth() : OAuth authentication error:
This most likely means that you have incorrectly called setup_twitter_oauth()'
```

This error happens when your app is missing the callback url. To solve this issue go to https://apps.twitter.com/ select your application and then go to SETTINGS scroll down to CALLBACK URL and enter (http://127.0.0.1:1410). This should allow you to run browser verification.



Application Details

Callback URL

http://127.0.0.1:1410

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Figure 3.4 Application Details

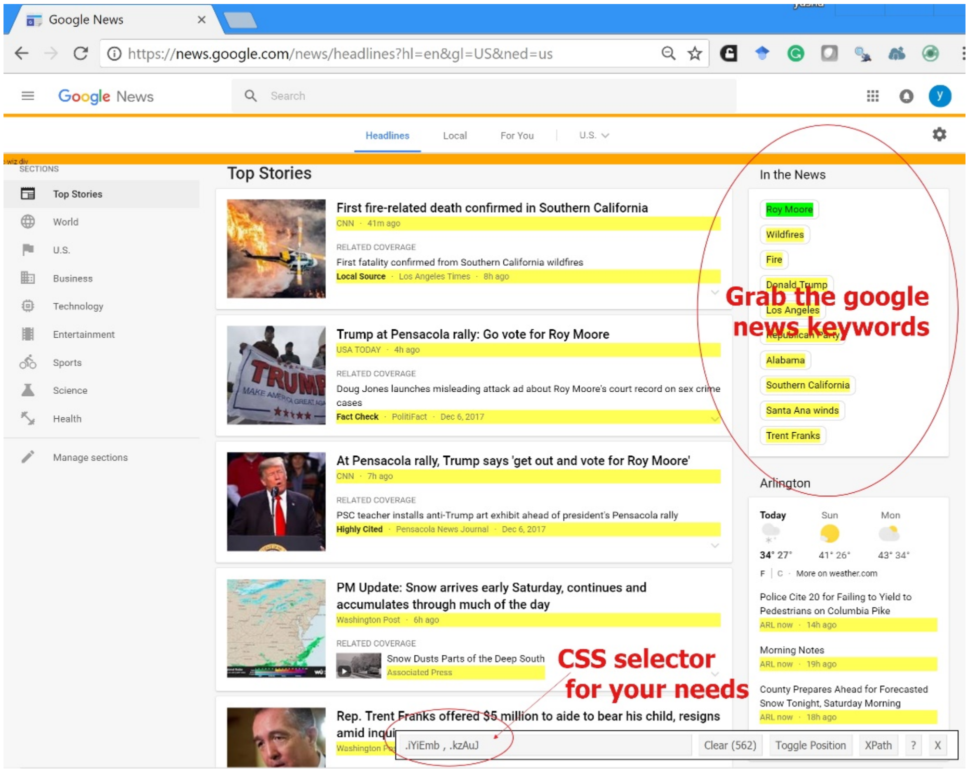
3.4 WEB DATA CRAWL

The web data crawl can help you scrape information from web pages. There are many ways to do web data crawl. In this chapter we use package Rvest to demonstrate how to do web data crawl. It is designed to work with magrittr to make it easy to express common web scraping tasks.

We will demonstrate how to grab google news key words here.

Step 1: Open the website http://selectorgadget.com/ install the point and click CSS selectors "SelectorGadget" Chrome Extensions. We recommend you use google chrome in this chapter. You can see the tutorial video about how to use "SelectorGadget" in the website.

Step 2: Copy your CSS selector for your needs.



First fire-related death confirmed in Southern California

CNN · 41m ago

RELATED COVERAGE

First fatality confirmed from Southern California wildfires

Local Source · Los Angeles Times · 8h ago

Trump at Pensacola rally: Go vote for Roy Moore

USA TODAY · 4h ago

RELATED COVERAGE

Doug Jones launches misleading attack ad about Roy Moore's court record on sex crime cases

Fact Check · PolitiFact · Dec 6, 2017

At Pensacola rally, Trump says 'get out and vote for Roy Moore'

CNN · 7h ago

RELATED COVERAGE

PSC teacher installs anti-Trump art exhibit ahead of president's Pensacola rally

Highly Cited · Pensacola News Journal · Dec 6, 2017

PM Update: Snow arrives early Saturday, continues and accumulates through much of the day

Washington Post · 6h ago

RELATED COVERAGE

Snow Dusts Parts of the Deep South

Associated Press

Rep. Trent Franks offered $5 million to aide to bear his child, resigns amid inqui

Police Cite 20 for Failing to Yield to Pedestrians on Columbia Pike

ARL now · 14h ago

Morning Notes

ARL now · 19h ago

County Prepares Ahead for Forecasted Snow Tonight, Saturday Morning

ARL now · 18h ago

Grab the google news keywords

CSS selector for your needs

.iYiEmb , .kzAuJ

Figure 3.5 How to Use CSS Selector

Step 3: Grab google news page.

```
library("rvest")

## Warning: package 'rvest' was built under R version 3.4.3

## Loading required package: xml2

## Warning: package 'xml2' was built under R version 3.4.2

library("dplyr")

## Warning: package 'dplyr' was built under R version 3.4.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

google <- read_html("https://news.google.com/news/headlines?hl=en&gl=US&ned=us"
)
```

Step 4: Grab the google news keywords.

```
google_News_Keywords <- google %>% html_nodes(".kzAuJ") %>% html_text()
# ".kzAuJ" are the CSS selector codes.

google_News_Keywords

##  [1] "Wildfires"
##  [2] "Fire"
##  [3] "Los Angeles"
##  [4] "Southern California"
##  [5] "Roy Moore"
##  [6] "Donald Trump"
##  [7] "Republican Party"
##  [8] "Trent Franks"
##  [9] "Republican Party"
## [10] "Arizona"
         … …
```

REFRENCES

[1] White, L. G. (1994). Political analysis: Technique and practice. Harcourt College Pub. p181

[2] FHSS, Brigham Young University, https://fhssrsc.byu.edu/Pages/Data.aspx

[3] Npc.gov.cn. (2017). 中华人民共和国网络安全法_中国人大网. [online] Available at: http://www.npc.gov.cn/npc/xinwen/2016-11/07/content_2001605.htm [Accessed 11 Dec. 2017].

[4] Gov.cn. (2017). 中华人民共和国保守国家秘密法（主席令第二十八号）. [online] Available at: http://www.gov.cn/flfg/2010-04/30/content_1596420.htm [Accessed 11 Dec. 2017].

[5] Partly websites from Manheim, J. B., Rich, R. C., Willnat, L., Brians, C. L., & Babb, J. (2012). Empirical political analysis. Pearson Higher Ed. pp 56-58

[6] Stackoverflow, R. (2017). R TwitteR package authorization error. [online] Stackoverflow.com. Available at: https://stackoverflow.com/questions/25856394/r-twitter-package-authorization-error [Accessed 11 Dec. 2017].

CHAPTER 4 DATA CLEANING IN POLITICAL DATA SCIENCE

Generally, in the data cleaning phase, we handle two things: dealing with missing data and removing outliers & duplicates.

4.1 DEALING WITH MISSING DATA

In R, missing values are represented by the symbol NA(not available). Impossible values(e.g., dividing by zero) are represented by the symbol NaN(not a number). R uses the same symbol for character and numeric data.

Missing values in data is a common phenomenon in real world problems. There are three easy ways to handle missing values.

i. Deleting the NA data

ii. Imputation with mean/median/mode

iii. kNN Imputation

We use the Survey dataset to discuss the various approaches to treat missing values. If you have large number of observations in your dataset, where all the classes to be predicted are sufficiently represented in the training data, then try deleting (or not to include missing values while model building, for example by setting na.action=na.omit) those observations (rows) that contain missing values. Make sure after deleting the observations, you have:

1. Have sufficient data points, so the model doesn’t lose power.

2. Not to introduce bias (meaning, disproportionate or non-representation of classes).

```
library(MASS)
head(survey)

##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height
## 1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00
## 2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80
## 3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA
## 4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00
## 5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00
## 6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72
##        M.I    Age
## 1   Metric 18.250
## 2 Imperial 17.583
## 3     <NA> 16.917
## 4   Metric 20.333
## 5   Metric 23.667
## 6 Imperial 21.000
```

```
library(MASS)
colSums(is.na(survey))  # Show the total number of missing values in data set.
```

```
##    Sex Wr.Hnd NW.Hnd  W.Hnd   Fold  Pulse   Clap   Exer  Smoke Height 
##      1      1      1      1      0     45      1      0      1     28 
##    M.I    Age 
##     28      0
```

Deleting the NA data

If a particular variable is having more missing values that rest of the variables in the dataset, and, if by removing that one variable you can save many observations. I would, then, suggest remove that particular variable, unless it is a really important predictor that makes a lot of business sense.

```
data <- survey
complete.cases(data)
```

```
##   [1]  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [12] FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
##  [23]  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
##  [34]  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE
##  [45] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [56] FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
##  [67] FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [78] FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
##  [89]  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE
## [100]  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
## [111]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
## [122]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [133] FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## [144]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [155]  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
## [166]  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
## [177]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [188]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## [199]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [210] FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE
## [221] FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## [232] FALSE  TRUE  TRUE FALSE  TRUE  TRUE
```

```
rm.data <- data[complete.cases(data), ] # removingthe NA date
colSums(is.na(rm.data))    # Show the total number of missing values in data se
t.

##    Sex Wr.Hnd NW.Hnd  W.Hnd   Fold  Pulse   Clap   Exer  Smoke Height
##      0      0      0      0      0      0      0      0      0      0
##    M.I    Age
##      0      0
```

Imputation with mean/median/mode

Replacing the missing values with the mean / median / mode is a crude way of treating missing values. Depending on the context, as if the variation is low or if the variable has low leverage over the response, such a rough approximation is acceptable and could possibly give satisfactory results. Imputation with mean/median/mode

```
library(Hmisc)

## Warning: package 'Hmisc' was built under R version 3.4.2

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

## 
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
## 
##     format.pval, round.POSIXt, trunc.POSIXt, units

dataII <- survey

dataII$Height[is.na(dataII$Height)] <- mean(dataII$Height,na.rm = T)
# replace with median

colSums(is.na(dataII))    # Show the total number of missing values in data set
.

##    Sex Wr.Hnd NW.Hnd  W.Hnd   Fold  Pulse   Clap   Exer  Smoke Height
##      1      1      1      1      0     45      1      0      1      0
##    M.I    Age
##     28      0
```

kNN Imputation

DMwR::knnImputation uses k-Nearest Neighbours approach to impute missing values. What kNN imputation does in simpler terms is as follows: For every observation to be imputed, it identifies 'k' closest observations based on the Euclidean distance and computes the weighted average (weighted based on distance) of these 'k' obs.

The advantage is that you could imput all the missing values in all variables with one call to the function. It takes the whole data frame as the argument and you don't even have to specify which variable you want to imput. But be cautious not to include the response variable while imputing, because when imputing in test/production environment, if your data contains missing values, you won't be able to use the unknown response variable at that time.

```
library(MASS)
dataIII <- survey
library(DMwR)

## Warning: package 'DMwR' was built under R version 3.4.2

## Loading required package: grid

imputeData <- knnImputation(dataIII)

colSums(is.na(imputeData))

##    Sex Wr.Hnd NW.Hnd  W.Hnd   Fold  Pulse   Clap   Exer  Smoke Height
##      0      0      0      0      0      0      0      0      0      0
##    M.I    Age
##      0      0
```

4.2 HOW TO REMOVE DUPLICATES

4.2.1 Remove Duplicate Data

Import your data.

We use the iris data ste in these chapter.

```
# Create my_data
my_data <- iris
# Convert to a tibble
library("tibble")
my_data <- as_data_frame(my_data)
# Print

head(my_data)

## # A tibble: 6 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##          <dbl>       <dbl>        <dbl>       <dbl>  <fctr>
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Find and drop duplicate elements: duplicated()

The function duplicated () returns a logical vector where TRUE specifies which elements of a vector or data frame are duplicates. So, we can remove duplicate rows from a data frame based on a column values, as follow:

```
# Remove duplicates based on Sepal.Width columns
my_data[!duplicated(my_data$Sepal.Width), ]

## # A tibble: 23 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl>  <fctr>
##  1          5.1         3.5          1.4         0.2  setosa
##  2          4.9         3.0          1.4         0.2  setosa
##  3          4.7         3.2          1.3         0.2  setosa
##  4          4.6         3.1          1.5         0.2  setosa
##  5          5.0         3.6          1.4         0.2  setosa
##  6          5.4         3.9          1.7         0.4  setosa
##  7          4.6         3.4          1.4         0.3  setosa
##  8          4.4         2.9          1.4         0.2  setosa
##  9          5.4         3.7          1.5         0.2  setosa
## 10          5.8         4.0          1.2         0.2  setosa
## # ... with 13 more rows
```

Extract unique elements:unique()

We can use unique(x) to extract unique elements. And unique() on a data frame.

```
unique(my_data)

## # A tibble: 149 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl>  <fctr>
##  1          5.1         3.5          1.4         0.2  setosa
##  2          4.9         3.0          1.4         0.2  setosa
##  3          4.7         3.2          1.3         0.2  setosa
##  4          4.6         3.1          1.5         0.2  setosa
##  5          5.0         3.6          1.4         0.2  setosa
##  6          5.4         3.9          1.7         0.4  setosa
##  7          4.6         3.4          1.4         0.3  setosa
##  8          5.0         3.4          1.5         0.2  setosa
##  9          4.4         2.9          1.4         0.2  setosa
## 10          4.9         3.1          1.5         0.1  setosa
## # ... with 139 more rows
```

Remove duplicate rows using dplyr

The function distinct () in dplyr package can be used to keep only unique/distinct rows from a data frame. If there are duplicate rows, only the first row is preserved.

```
install.packages("dplyr",repos = "http://cran.us.r-project.org")
```

```
## package 'dplyr' successfully unpacked and MD5 sums checked
## 
## The downloaded binary packages are in
##  C:\Users\yushu\AppData\Local\Temp\Rtmpucc92E\downloaded_packages
```

```
library("dplyr")
```

```
## Warning: package 'dplyr' was built under R version 3.4.3
```

```
## 
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
## 
##     filter, lag
```

```
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
distinct(my_data) #Remove duplicate rows based on all columns.
```

```
## # A tibble: 149 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl>  <fctr>
##  1          5.1         3.5          1.4         0.2  setosa
##  2          4.9         3.0          1.4         0.2  setosa
##  3          4.7         3.2          1.3         0.2  setosa
##  4          4.6         3.1          1.5         0.2  setosa
##  5          5.0         3.6          1.4         0.2  setosa
##  6          5.4         3.9          1.7         0.4  setosa
##  7          4.6         3.4          1.4         0.3  setosa
##  8          5.0         3.4          1.5         0.2  setosa
##  9          4.4         2.9          1.4         0.2  setosa
## 10          4.9         3.1          1.5         0.1  setosa
## # ... with 139 more rows
```

```
# Remove duplicated rows based on Sepal.Length
distinct(my_data, Sepal.Length)
```

```
## # A tibble: 35 x 1
##    Sepal.Length
##           <dbl>
##  1          5.1
##  2          4.9
##  3          4.7
##  4          4.6
##  5          5.0
##  6          5.4
##  7          4.4
##  8          4.8
##  9          4.3
## 10          5.8
## # ... with 25 more rows
```

```
# Remove duplicated rows based on
# Sepal.Length and Petal.Width
distinct(my_data, Sepal.Length, Petal.Width)
```

```
## # A tibble: 110 x 2
##    Sepal.Length Petal.Width
##           <dbl>       <dbl>
##  1          5.1         0.2
##  2          4.9         0.2
##  3          4.7         0.2
##  4          4.6         0.2
##  5          5.0         0.2
##  6          5.4         0.4
##  7          4.6         0.3
##  8          4.4         0.2
##  9          4.9         0.1
## 10          5.4         0.2
## # ... with 100 more rows
```

4.2.2 Remove Outliers

Input data. We use dataset rivers in this chapter.

```
#data(rivers)
head(rivers)
```

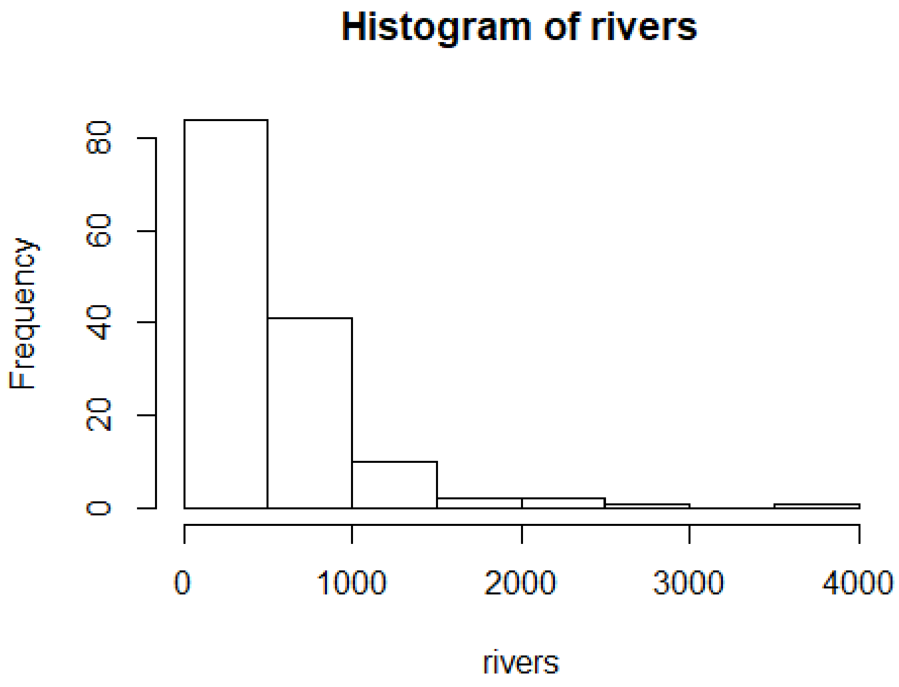
```
## [1] 735 320 325 392 524 450
```

```
length(rivers)
```

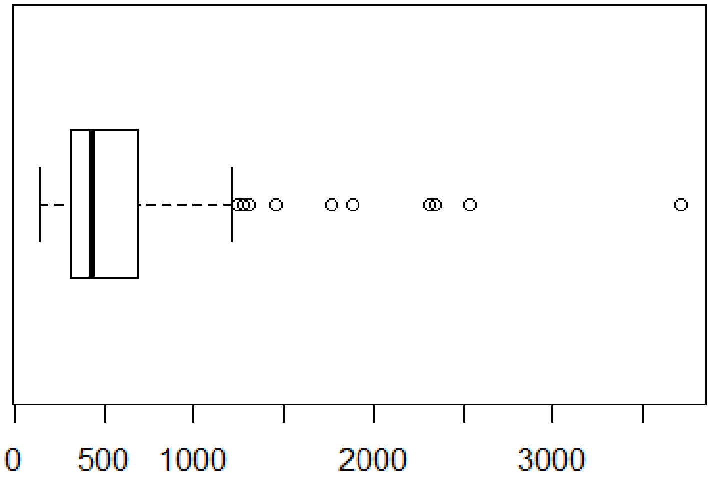
```
## [1] 141
```

Polt the data, and looking for the outliers.

```
hist(rivers)
```

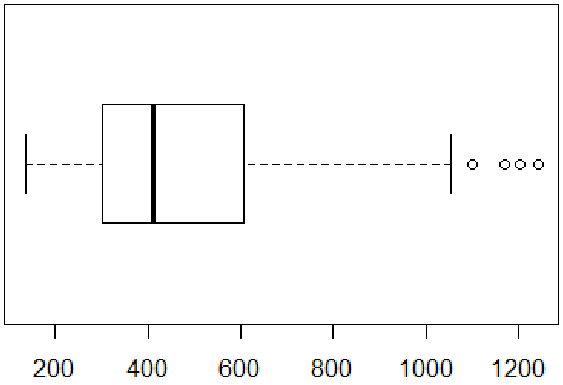


```
boxplot(rivers,horizontal = T)
```



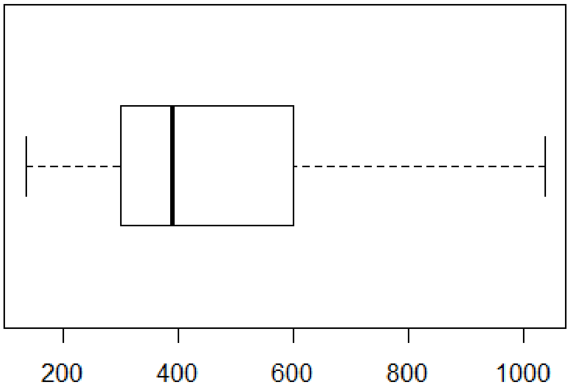
Then Try to remove part of outliers.

```
rivers_new <- subset(rivers,rivers<1250)
boxplot(rivers_new,horizontal = T)
```

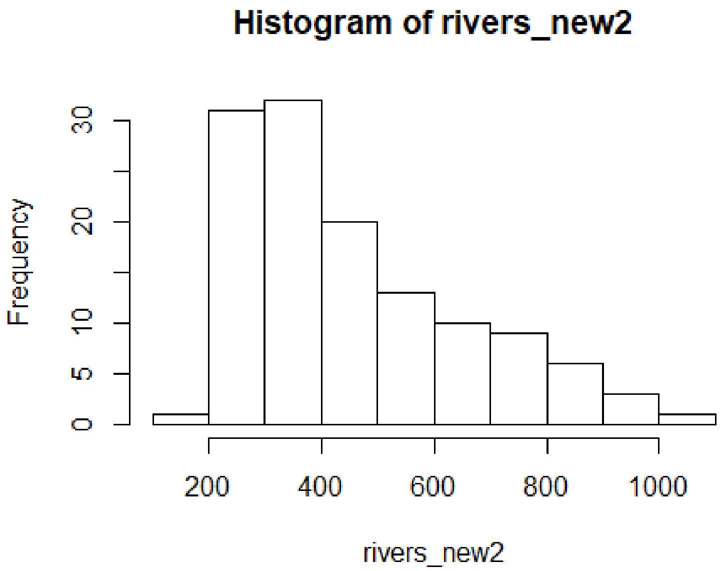


Based on the plove above, we can reset the rivers data to remove the outliers.

```
rivers_new2 <- subset(rivers,rivers<1050)
boxplot(rivers_new2,horizontal = T)
```



```
hist(rivers_new2)
```



REFRENCES

[1] Prabhakaran, S. (2017). Missing Value Treatment. [online] R-bloggers. Available at: https://www.r-bloggers.com/missing-value-treatment/ [Accessed 11 Dec. 2017].

CHAPTER 5 DESCRIPTIVE STATISTICS IN POLITICAL SCIENCE RESEARCH

Descriptive statistics are about the simplification, organization, summary， summary and graphical plotting of numerical data [1].

Many interesting political research questions ask us to describe patterns within single variables. For example, there are 89.4 million members of Chinese Communist Party (CCP)[2]. In order to ascertain the characteristics of the CCP groups, we could ask several questions. For example, have these members changed in the past few years? What percent of members are females or males? What career are most of the members? How many are politics? How many work in the public sector? How similar are they in their careers? What are they education background?

These questions can be answered by descriptive statistics and techniques which we use to summarize information about a variable and show the patterns in our data. They are particularly useful when we want to compare two cases or groups. Generally, we use four/? main indicators to summary data [3].

i. Percentages and proportions

ii. Frequency distribution

iii. Measures of central tendency

iv. Measures of dispersion

5.1 Using Example data sets within R

Once you start your R program, there are example data sets available within R along with loaded packages [4]. You can list the data sets by their names and then load a data set into memory to be used in your statistical analysis. For example, in the book “Modern Applied Statistics with S [5]” a data set called phones is used in Chapter 6 for robust regression and we want to use the same data set for our own examples. Here is how to locate the data set and load it into R. Command library loads the package MASS (for Modern Applied Statistics with S) into memory. Command data () will list all the datasets in loaded packages. In this book, we will use some example data sets as an example to do analysis.

5.2 Basic Descriptive Statistics with R

5.2.1 Percentages and proportions

Percentages and proportions supply a frame of reference for reporting research results in the sense that they standardize the raw data: percentages to the base 100 and proportions to the base 1.00. The mathematical definitions of proportions and percentages are[6]

FORMULAR 2.1.1 Proportion (p) = $\frac{f}{N}$

FORMULAR 2.1.2 Percentage (%) = $\frac{f}{N} * 100$

Where f = frequency, the number of cases in any category

N= the numbers of cases in all categories

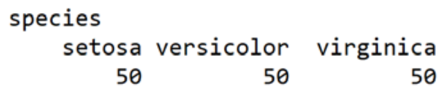
Percentages are a useful way to describe and analyze data. They not only provide a summary measure that is easily understood but allow us to make comparisons between and among groups of different sizes.

5.2.2 Frequency distribution

Frequency distributions is a summary of the data occurrence in a collection of non-overlapping categories. For example, in the data set iris, the frequency distribution of the species is a summary of the iris in each species.

```
species <- iris$Species  # the iris species
species.freq <- table(species) # apply the table function
species.freq
```

Results:



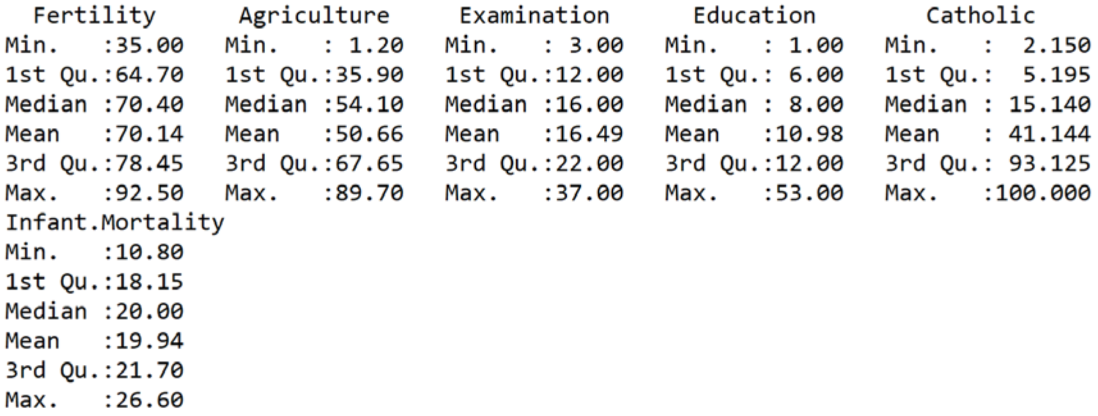
```
species
    setosa versicolor  virginica 
        50         50         50 
```

5.2.3 Measures of central tendency

R provides a wide range of functions for obtaining summary statistics. One method of obtaining descriptive statistics is to use the sapply ( ) function with a specified summary statistic [7]. Take example data swiss (Swiss Fertility and Socioeconomic Indicators <1888> Data) as an example.

```
library(MASS)


summary(swiss)
```

Results:



```
   Fertility      Agriculture     Examination      Education        Catholic
 Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00   Min.   :  2.150
 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00   1st Qu.:  5.195
 Median :70.40   Median :54.10   Median :16.00   Median : 8.00   Median : 15.140
 Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98   Mean   : 41.144
 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00   3rd Qu.: 93.125
 Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00   Max.   :100.000
 Infant.Mortality
 Min.   :10.80
 1st Qu.:18.15
 Median :20.00
 Mean   :19.94
 3rd Qu.:21.70
 Max.   :26.60
```

5.2.4 Measures of dispersion

Generally, measure of central tendency are incomplete summarizers of data. To describe fully a distribution of scores, measures of central tendency must be paired with measures of dispersion. In the political science, we usually use **standard deviation** (σ)to describe the dispersion of a distribution. The σ is a measure that is used to quantify the amount of variation or dispersion of a set of data values[8].

FORMULAR 2.4.1 $\sigma = \sqrt{\frac{\sum(x_i-\bar{x})^2}{N}}$

X_i = the value i of variable X

$\bar{X}$ = the mean value of variable X

Σ = add together

σ = standard deviation

N = number of cases

```
baby_mortarlity <- swiss$Infant.Mortality
```

```
sd(baby_mortarlity)
```

Result:

```
[1] 2.912697
```

The standard deviation has many uses in analyzing data patterns. For example, it can be used to compare how much equality there is within a political unit. In a city of China, if everyone has an income close to the mean, then this city has more equality than other city where some are very poor and some are very rich. Thus, the lower the standard deviation of income within a political unit, the less the diversity of income, it also means the greater the income equality.

5.3 Type of Charts of Descriptive Statistics in Political Science Research

R offers a good range of descriptive statistics in a variety of charts. Table 5.1 is the type of charts and their applications.

Table 5.1 Type of Charts and their applications

Chart	applications	**R fucntion**
Histogram charts	1.Comparing values 2.Ranking top to bottom	hist()
Bar charts	1.Comparing values 2.Ranking top to bottom	barplot()
Line charts	1. Comparing values over time Notes: sometimes it is not correct to draw a line, i.e. there is not always a link between one value and the next.	lines( )
Pie charts	1. Comparing each value with the total value	pie()

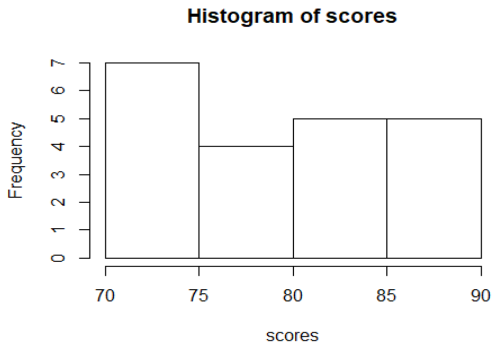
Scatter charts	1. Comparing pairs of values	dotchart()
		plot()

Box Plot	1.Displays of the distribution of values on a variable where the middle of the data distribution is indicated.	boxplot()

5.3.1 Histogram charts

Histograms are a popular way to display data in political science research paper. Histograms display similar to a bar chart, but they are not the same. Data used with histograms indicate frequency or portions for intervals, not categories. Visually, the bars are contiguous or connected, rather than displayed as separate bars, so it represents the distribution of data.

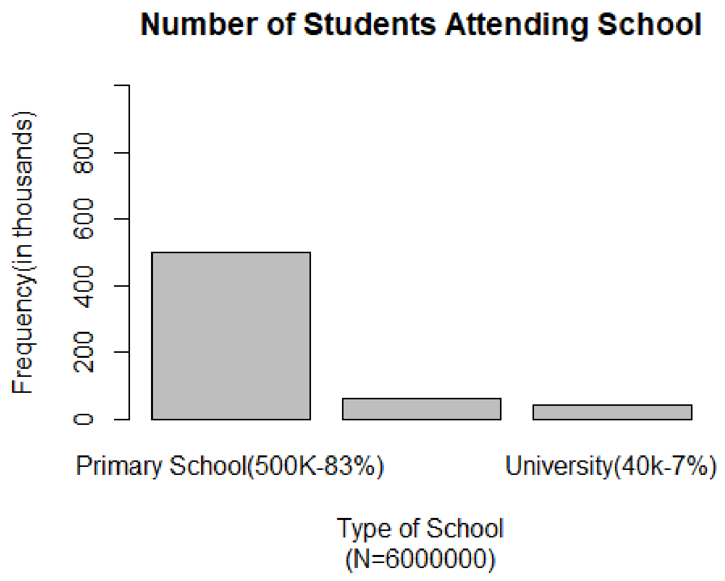
```
scores <- c(70,71,72,73,74,75,76,77,78,79,70,81,82,83,84,85,86,87,88,89,90)
hist(scores)
```



5.3.2 Bar charts

The bar chart is useful for displaying the frequency of each category in an ordinal manner -- that is, ranked from least to most.

```
student <- c(500,60,40)
names(student) <- c("Primary School(500K-83%)","High School(60k-10%)","University(40k-7%)")
barplot(student,main="Number of Students Attending School",sub = "(N=6000000)",xlab="Type of School
",ylab="Frequency(in thousands)",ylim = c(0,1000))
```

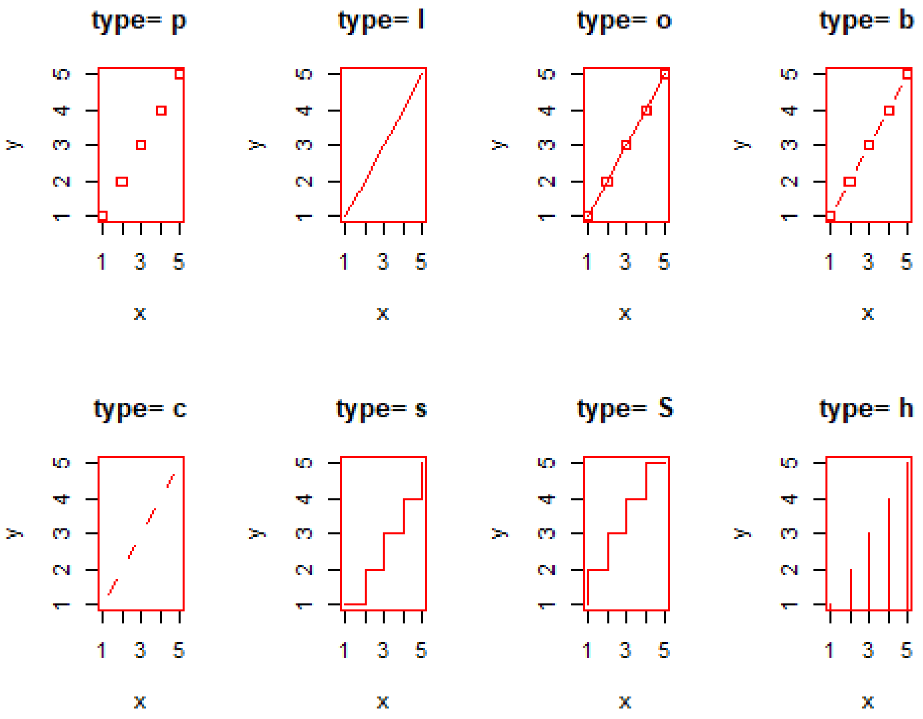


5.3.3 Line charts

Line charts are created with the function lines(x, y, type=) where x and y are numeric vectors of (x,y) points to connect. type= can take the following values[9]:

In the following code each of the type= options is applied to the same dataset. The plot ( ) command sets up the graph, but does not plot the points.

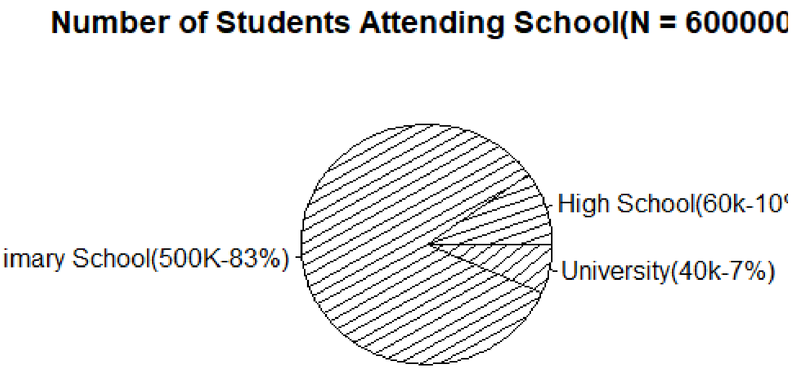
```
x <- c(1:5); y <- x # create some data
par(pch=22, col="red") # plotting symbol and color
par(mfrow=c(2,4)) # all plots on one page
opts = c("p","l","o","b","c","s","S","h")
for(i in 1:length(opts)){
  heading = paste("type=",opts[i])
  plot(x, y, type="n", main=heading)
  lines(x, y, type=opts[i])
}
```



5.3.4 Pie charts

The pie chart is used to display frequency counts or percents in mutually exclusive categories. It is displayed as a circle that is partitioned into the different mutually exclusive categories.

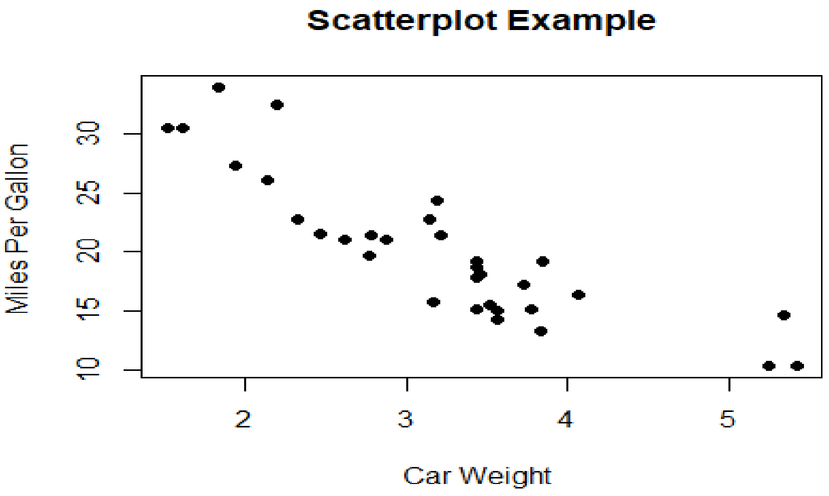
```
student <- c(60000,500000,40000)
names(student) <- c("High School(60k-10%)","Primary School(500K-83%)","University(40k-7%)")
pie(student,main="Number of Students Attending School(N = 600000)",
   density = 15, angle = 10 + 10 * 1:6 )
```



5.3.5 Scatter charts

A scatterplot displays pairs of numbers for y and x variables, which are called coordinate points. There are many ways to create a scatterplot in R. The basic function is plot(x, y), where x and y are numeric vectors denoting the (x,y) points to plot.

```
# Simple Scatterplot
attach(mtcars)
plot(wt, mpg, main="Scatterplot Example",
   xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)
```



5.3.6 Box -Whisker Plot

A box - whisker plot displays the distribution of scores on a variable where the middle of the data distribution is indicated. A box is placed around this middle value to indicate the upper and lower range of values where 50% of the scores are distributed. Extending out the top and bottom of the box are two lines referred to the lowest and the highest values.

Box - Whisker Plot is very popular in political science paper. A box - whisker plot is a graphical method of displaying variation in a set of data. In most cases a histogram provides a sufficient display; however, a box and whisker plot can provide additional detail while allowing multiple sets of data to be displayed in the same graph. Some types are called box and whisker plots with outliers. Use box and whisker plots when you have multiple data sets from independent sources that are related to each other in some way.

A box and whisker plot is developed from five statistics [10]:

i. Minimum value – the smallest value in the data set

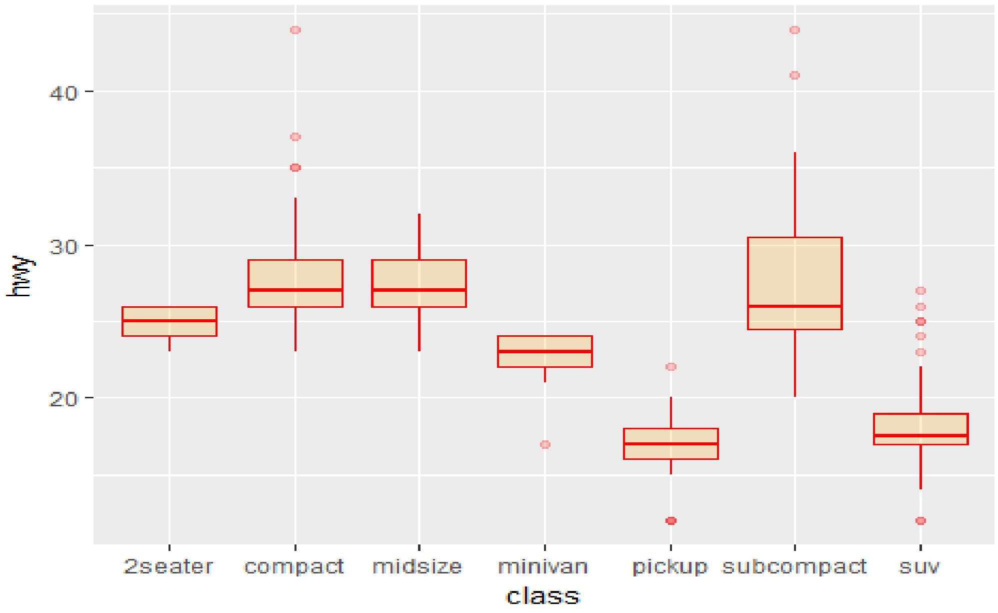
ii. Second quartile – the value below which the lower 25% of the data are contained

iii. Median value – the middle number in a range of numbers

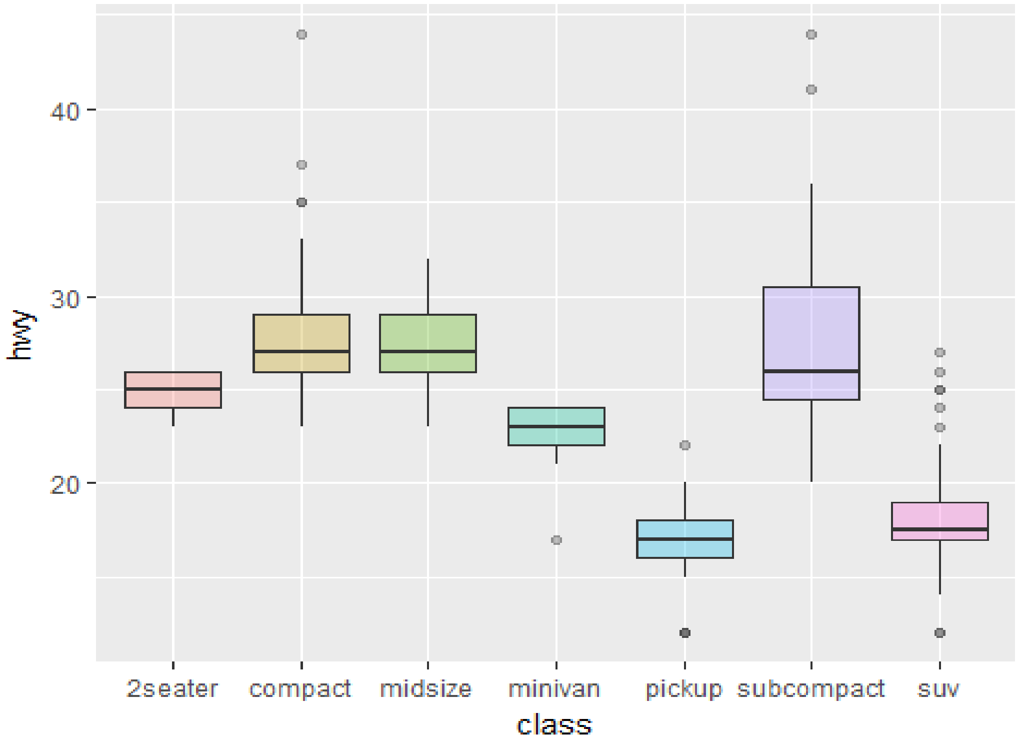
iv. Third quartile – the value above which the upper 25% of the data are contained

v. Maximum value – the largest value in the data set

```
# library
library(ggplot2)

# The mtcars dataset is proposed in R
head(mpg)
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl      trans   drv   cty   hwy    fl
##          <chr> <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>
## 1         audi    a4   1.8  1999     4   auto(l5)     f    18    29     p
## 2         audi    a4   1.8  1999     4 manual(m5)     f    21    29     p
## 3         audi    a4   2.0  2008     4 manual(m6)     f    20    31     p
## 4         audi    a4   2.0  2008     4   auto(av)     f    21    30     p
## 5         audi    a4   2.8  1999     6   auto(l5)     f    16    26     p
## 6         audi    a4   2.8  1999     6 manual(m5)     f    18    26     p
## # ... with 1 more variables: class <chr>
# Set a unique color with fill, colour, and alpha
ggplot(mpg, aes(x=class, y=hwy)) +
    geom_boxplot(color="red", fill="orange", alpha=0.2)
```



```
# Set a different color for each group
ggplot(mpg, aes(x=class, y=hwy, fill=class)) +
    geom_boxplot(alpha=0.3) +
    theme(legend.position="none")
```



REFRENCES

[1] Thomas, G. (2013). How to do your research project: A guide for students in education and applied social sciences. Sage. P250

[2] News.12371.cn. (2017). 2016 年中国共产党党内统计公报_共产党员网. [online] Available at: http://news.12371.cn/2017/06/30/ARTI1498810325807955.shtml [Accessed 11 Dec. 2017].

[3] White, L. G. (1994). Political analysis: Technique and practice. Harcourt College Pub.

[4] Stat.ethz.ch. (2017). R: The R Datasets Package. [online] Available at: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html#A [Accessed 11 Dec. 2017].

[5] Stats.ox.ac.uk. (2017). Modern Applied Statistics with S, 4th ed. [online] Available at: http://www.stats.ox.ac.uk/pub/MASS4/ [Accessed 11 Dec. 2017].

[6] Healey, J. F. (2014). Statistics: A tool for social research. Cengage Learning.P21

[7] Statmethods.net. (2017). Quick-R: Descriptives. [online] Available at: https://www.statmethods.net/stats/descriptives.html [Accessed 11 Dec. 2017].

[8] Bland, J.M.; Altman, D.G. (1996). "Statistics notes: measurement error". BMJ. 312 (7047): 1654.

[9] Statmethods.net. (2017). Quick-R: Line Charts. [online] Available at: https://www.statmethods.net/graphs/line.html [Accessed 11 Dec. 2017].

[10] Asq.org. (2017). Box and Whisker Plot - ASQ. [online] Available at: http://asq.org/learn-about-quality/data-collection-analysis-tools/overview/box-whisker-plot.html [Accessed 11 Dec. 2017].

CHAPTER 6 STATISTICAL INFERENCE IN POLITICAL SCIENCE

Statistical Inference is the process of deducing properties of an underlying probability distribution by analysis of data [1]. The Wikipedia gives a definition which can be easily understood by people:

> *"Inferential statistical analysis infers properties about a population: this includes testing hypotheses and deriving estimates. The population is assumed to be larger than the observed data set; in other words, the observed data is assumed to be sampled from a larger population.*
>
> *Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and does not assume that the data came from a larger population."*

There are two types of statistical inference: estimation and hypotheses testing:

Estimation: the objective of estimation is to determine the value of a population parameter on the basis of a sample statistic. There are two types of estimators: point estimator, interval estimator. We will discuss it in section 3.

Hypotheses testing: gives a range of values for the parameter Interval estimates are intervals within which the parameter is expected to fall, with a certain degree of confidence. Generally, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model.

The purpose of this chapter is to demonstrate the basic principles and process of statistical inference. We find many Chinese political science students scared by tons of formulas in Chinese statistics textbook. They just hope to understand the whole picture of statistical inference process before they decide to dive into details. So, in this Chapter, we focus on what in statistical inference and how to use it in political science paper.

6.1 KEY CONCEPTS

Generally, the concepts in statistics are defined by lots of mathematical formulas. These bring huge challenge for math allergic students who major in political science. Bruce (2017) gave more understandable explanations about 50 essential concepts in his book [2]. We summarize some concepts definitions we will use in this chapter.

> ***Bias****: error. What is the inherent error that you obtain from your classifier even with infinite training data? This is due to your classifier being "biased" to a particular kind of solution (e.g. linear classifier).*
>
> ***Confidence intervals****: is the frequency (i.e., the proportion) of possible confidence intervals that contain the true value of their corresponding parameter. In other words, if confidence intervals are constructed using a given confidence level in an infinite number of independent experiments, the proportion of those intervals that contain the true value of the parameter will match the confidence level[3].*
>
> ***Confidence level****: the percentage of confidence intervals, constructed in the same way from the same population, expected to contain the statistic of interest.*
>
> ***Correlation coefficient:*** *a metric that measures the extent to which numeric variables are associated with one another (ranges from -1 t0 +1).*
>
> ***Central limit theorem****: the tendency of the sampling distribution to take on a normal shape as sample size rises.*
>
> ***Confidence level****: the percentage of confidence intervals, constructed in the same way from the same population, expected to contain the statistic of interest. Confidence intervals are the typical way to present estimates as an interval range. The more data you have, the less variable a sample estimate will be. The lower the level of confidence you can tolerate, the narrow the confidence interval will be.*
>
> ***Deviations****: the difference between the observed values and the estimate of location.*
>
> ***Degrees of freedom****: a parameter that allows the t-distribution to adjust to different sample sizes, statistics, and number of groups.*
>
> ***Data distribution****: the frequency distribution of individual values in a data set.*
>
> ***d.f.*** *: Degrees of freedom.*
>
> ***Effect size****: The minimum size of the effect that you hope to be able to detect in a statistical test, such as "a 20% improvement in click rates".*
>
> ***Expected value****: when the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.*
>
> ***Error****: the difference between a data point and a predicted or average value.*

***Interquartile range**: the difference between the 75th percentile and the 25th percentile.*

***Interval endpoints**: the top and bottom of the confidence interval.*

***N(n)**: the size of the population(sample).*

***n**: sample size.*

***n or sample size**: the number of observations (also called rows or records) in the data.*

***Mean**: the sum of all values divided by the number of values.*

***Mode**: the most commonly occurring category or value in a data set.*

***Median**: the value such that one-half of the data lies above and below.*

***Mean absolute deviation**: the mean of the absolute value of the deviations from the mean.*

***Media absolute deviation from the median**: the median of the absolute value of the deviations from the median.*

***Order statistics**: metrics based on the data values sorted from smallest to biggest.*

***Population**: the larger data set or idea of a data set.*

***Percentile**: the value such that P percent of the values take on this value or less and(100-P) percent take on this value or more.*

***Power**: the probability of detecting a given effect size with a given sample size.*

***Range**: the difference between the largest and the smallest value in a data set.*

***Sample**: a subset from a larger data set.*

***Sample statistic**: a metric calculated for a sample of data drawn from a larger population.*

***Sampling distribution**: the frequency distribution of a sample statistic over many samples or resamples.*

***Standard error**: the variability (standard deviation) of a sample statistic over many samples(not to be confused with standard deviation, which, by itself, refers to variability or individual data values).*

***Standard normal**: a normal distribution with mean = 0 and standard deviation = 1.*

***SS**: "Sum of squares," referring to deviations from some average value.*

***Standard deviation**: the square root of the variance.*

***Standardize**: subtract the mean and divide by the standard deviation.*

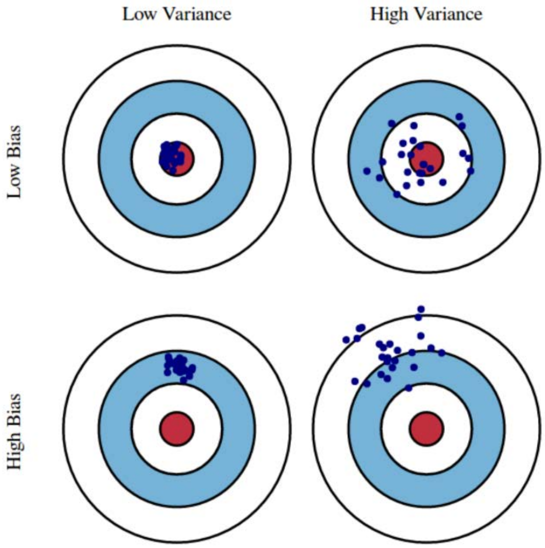
***Significance level**: the statistical significance level at which the test will be conducted.*

***Trimmed mean**: the average of all values after dropping a fixed number of extreme values.*

***Variance**: the sum of squared deviation from the mean divided by n-1 where n is the number of data values.*

***Weighted mean**: the sum of all values times a weight divided by the sum of the weights.*

***Weighted media**: the value such that one-half of the sum of the weights lies above and below the sorted data.*



***z-score**: the result of standardizing an individual data point.*

Figure 6.1 Graphical illustration of bias and variance

Source: *Understanding the Bias-Variance Tradeoff[4]*

Especially, understanding how different sources of error lead to bias and variance helps us improve the data fitting process resulting in more accurate models. The figure 6.1 shows the difference of bias and variance. Imagine the center of the Red bulls' eye region is the true mean value of our target random variable which we are trying to predict, and the red region indicates the variance spread of this variable. Every time we take a sample set of observations and predict the value of this variable we plot a blue dot. We predict correctly if the blue dot falls inside the red region. In other words, bias is the measure of how far off are the predicted blue dots from the true red region, intuitively this is an error. Variance is how scattered are our predictions.

6.2 THE PROCESS OF STATISTICAL INFERENCE

Just as Figure 6.2 shows, the goal in inferential statistical is to generalize to the characteristics of a population (also called parameters), base on what we can learn from our samples.

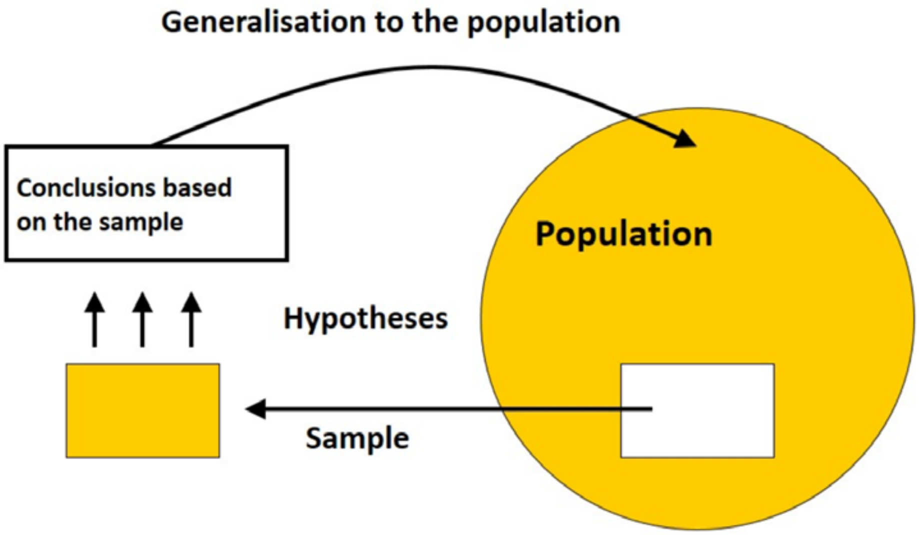

Generalisation to the population
Conclusions based on the sample
Population
Hypotheses
Sample

Figure 6.2 The idea of Statistical Inference

Source: "*Significance Testing and P Value*"[5]

6.2.1 The Sample Problems of Political Science Research

The inference is based on the population sample. Joseph (1995) argued that:

> *"Let me point out that social scientists often use nonprobability samples and that such samples are very useful for a number of purposes. Nonprobability samples are typically less costly and easier to assemble and are appropriate in many different research situations. The major limitation of nonprobability samples is that they do not permit the use of inferential statistics to generalize to populations[6]."*

Joseph (1995) pointed out that the fundamental principle of probability sampling is that a sample is very likely to be representative if it is selected by a principle called **EPSEM**, which stands for the "**E**qual **P**robability of **S**election **M**ethods. [7] " But at the same time, we should clearly know that a sample properly selected does not guarantee that it will be an exact representation or microcosm of the population. Joseph (1995) strongly argued that an EPSEM sample will occasionally present an inaccurate picture of the population. One great strength of inferential statistics is that they allow the researcher to estimate the probability of this type of error and interpret results accordingly.

6.2.2 The Sampling Distribution

In inferential statistics, political science researchers face a puzzling dilemma. On one hand, they have a great deal of information about the sample distribution. On the other hand, they know virtually nothing about the population.

Generally, the information necessary to characterize a distribution adequately would include (1) the shape of the distribution, (2) some measure of central tendency, and (3) some measure of dispersion (Joseph ,1995)[8] . All these kinds of information can be computed for the sample distribution. The general strategy of all applications of inferential statistics is to move from the sample to the population via **the sampling distribution**. Specifically, the shape, central tendency, and dispersion of this distribution can be deduced and the distribution can be adequately characterized.

6.2.3 Symbols of Key Terms

Joseph (1995) defined the sampling distribution as a theoretical, probabilistic distribution of all possible sample outcomes (with constant sample size, N) for the statistic that is to be generalized to the population [9].

Table 6.1 shows the symbols of variables which we generally calculate from sampling distribution [10]. These key variables related with two theorems.

> *i. If repeated random samples of size N are drawn from a normal population with mean μ and standard deviation σ, then the sampling distribution of sample means will be normal with a mean μ and a standard deviation of $\frac{\sigma}{\sqrt{N}}$. The standard deviation of the sampling distribution,* ***also called the standard error of the mean****.*
>
> *ii.* ***The Central Limit Theorem****: if repeated random samples of size N are drawn from any population (not normal), with mean μ and standard deviation σ, then, as N becomes large, the sampling distribution of sample means will approach normality, with mean μ and standard deviation $\frac{\sigma}{\sqrt{N}}$.*

Table 6.1 Symbols for means and standard deviations of three distributions

	Mean	Standard Deviation	Proportion
Samples	$\bar{x}$	s	P_s (Sample proportion)
Population	μ	σ	P_u (Population proportion)
Sampling distributions of means	$\mu_{\bar{x}}$	$\sigma_{\bar{x}}$	
Sampling distributions of proportions	μ_p	σ_p	

6.3 PROBABILITY DISTRIBUTION IN POLITICAL SCIENCE RESEARCH

A probability distribution describes how the values of a random variable is distributed. the means of sufficiently large samples of a data population are known to resemble the normal distribution (Chi Yau, 2013). Since the characteristics of these theoretical distributions are well understood, they can be used to make statistical inferences on the entire data population.

You can obtain a list of distribution in the R stats package:

```
> help(“distributions”)
```

Table 6.2 is the family of distribution with root names. In this

Table 6.2 Family of distribution of R stats package

Distribution	**Root name**
Beta	Beta
Cauchy	Cauchy
Chi-square	Chisq
Exponential	Exp
F	F
Gamma	Gamma
Normal	Norm
Student’s t	T
Uniform	Unif
Weibull	Weibull

6.3.1 Normal Distribution

In this section, we take Normal distribution as an example and we will reference them quite often in other sections. The **normal distribution** is defined by the following probability density function, where μ is the population mean, and σ^2 is the population variance.

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where

Y = the height of the curve at a given score

X = the score at given height

μ = the mean of the X variable

σ = the standard deviation of the X variable

π = 3.14(pi)

e = 2.7183

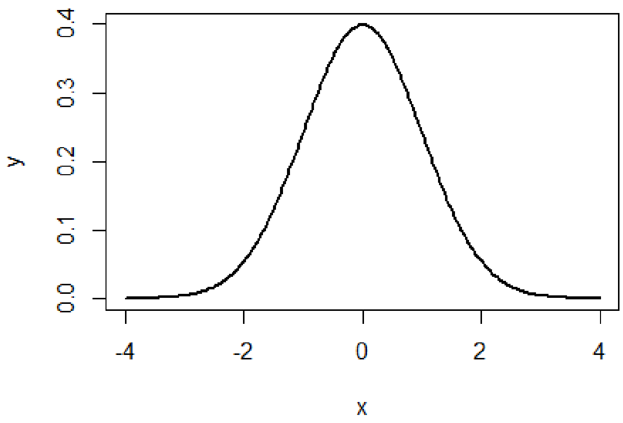
6.3.2 Standard Normal Distribution (Z distribution)

When a set of X scores is transformed to have a mean of 0 and a standard deviation of 1, the scores are called standard scores or **z scores**, and the normal distribution equation can be reduced to

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2\sigma^2}$$

This equation using z scores is referred to as **the standard normal curve**. The plot() function displays the bell-shaped curve derived from the equation, which today is referred to as the normal curve or normal distribution based on z scores.

```
x = seq(-4,4,length=200)
y = 1/sqrt(2*pi)*exp(-x^2/2)
plot(x,y, type = "l", lwd=2)
```



6.3.3 Calculating Z Score

A z score is calculated as

$$z = \frac{(x - \mu)}{\sigma}$$

The μ is the population mean and σ the population standard deviation. When we don't know μ and σ, the sample z score can be calculated as

$$z = \frac{(x - \bar{x})}{S}$$

where

x = the sample values

$\bar{x}$ = the sample mean

S = the sample standard deviation

The z score indicates the deviation of a score from the sample mean in standard deviation units.

6.4 ESTIMATION

In this section, we demonstrate estimation with a **R built-in data set survey**. It is the outcome of a student survey in a statistics class of an Australian university. The data set belongs to the MASS package, which is often implicitly pre-loaded in the R workspace. Here is a preview of the data.

```
library(MASS)
head(survey)
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height
## 1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00
## 2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80
## 3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA
## 4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00
## 5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00
## 6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72
##        M.I    Age
## 1   Metric 18.250
## 2 Imperial 17.583
## 3     <NA> 16.917
## 4   Metric 20.333
## 5   Metric 23.667
## 6 Imperial 21.000
```

6.4.1 Point Estimate

Point estimation involves the use of sample data to calculate a single value (known as a statistic) which serves as a "best guess" or "best estimate" of an unknown (fixed or random) population parameter[11]. For example, in the data set survey, the actual survey is conducted only on a sample of the student population. However, we can compute the sample mean and use it as an estimate of the corresponding student population mean.

Take R built-in data set survey as an example. How can we find a point estimate of the mean university student height using sample data from the data survey?

For convenience, we begin with saving the survey data of student heights in the variable height.survey.

```
height.survey = survey$Height
```

It turns out not all students have answered the question, and we need filter out the missing values. Hence we apply the function mean with the na.rm option set as TRUE.

```
mean(height.survey,na.rm = TRUE)

## [1] 172.3809
```

Then, we can get that a point estimate of the mean student height is 172.38 centimeters.

6.4.2 Confidence Interval Estimates

After we found a point sample estimate of the population, we would need to estimate its confidence interval. Interval estimation is the use of sample data to calculate an interval of plausible values of an unknown population parameter; this is in contrast to point estimation, which gives a single value (Jerzy Neyman ,1937)[12]. The confidence interval estimate is a range of values constructed from sample data so that the population parameter is likely to occur within that range at a specified probability. The specified probability is called the confidence level.

Generally, an estimation of a population, the parameter given by two numbers between which the parameter may be called as an internal estimation of the parameter.

For example, the percentage of kids who like baseball is 40 percent, plus or minus 3.5 percent. That means the percentage of kids who like baseball is somewhere between 40% - 3.5% = 36.5% and 40% + 3.5% = 43.5%. The lower end of the interval is your statistic minus the margin of error, and the upper end is your statistic plus the margin of error（p70, statistics essentials for dummies）.

There are three factors that determine the width of a confidence interval.

1. The sample size, n.

2. The variability in the population, usually σ estimated by s.

3. The desired level of confidence.

Especially, for a 95% confidence interval about 95% of similarly constructed intervals will contain the parameter being estimated. Also 95% of the sample means for a specified sample size will lie within 1.96 standard deviations of the hypothesized population. (Fig. 6.3)

6.4.3 The Confidence Interval Estimates Model Selection

If the confidence interval for population mean with σ known or the sample is greater than 30. For random samples of sufficiently large size n, the end point of the interval estimate at(1- α) confidence level is given as follows:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$\bar{x}$ - sample mean

z - z-value for a particular confidence level. The 100(1- $\alpha/2$) percentile of the standard normal distribution as $z_{\alpha/2}$

σ - the population standard deviation

n - the number of observations in the sample

The width of the interval is determined by the level of confidence and the size of the standard error of the mean. The standard error is affected by two values: standard deviation, number of observations in the sample.

However, in most sampling situations the population standard deviation (σ) is not known. Generally, we use z-distribution if the population standard deviation is known **or the sample is larger than 30**.

And we use t-distribution if the population standard deviation is unknown and the sample is less than 30.

$$\bar{x} \pm t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

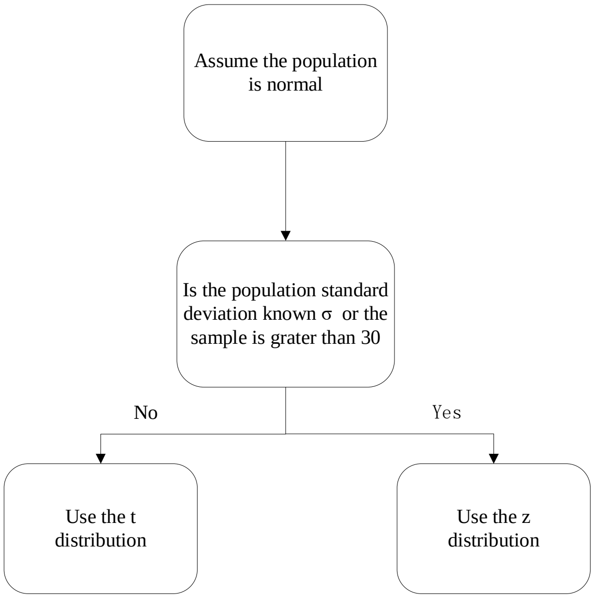

Assume the population is normal
Is the population standard deviation known σ or the sample is grater than 30
No
Yes
Use the t distribution
Use the z distribution

Fig. 6.3 CI estimate model selection

The ultimate goal when making an estimate using a confidence interval is to have a small margin of error. The narrower the interval, the more precise the results are.

Assume the population standard deviation sigma of the student height in the data set survey to be 9.5. Find the margin of error and interval estimate of the population mean at 95% confidence level. (We use Z distribution in here)

We first filter out missing values in survey $Height with the function na.omit, and save it in height.reponse.

```
library(MASS)

height.response = na.omit(survey$Height)
```

Then we compute the standard error of the mean, sem.

```
n = length(height.response)
sigma = 9.5
sem = sigma/sqrt(n)
sem

## [1] 0.6571287
```

Since there are two tails of the normal distribution, the 95% confidence level would imply the (100-2*sem) percentile of the normal distribution at the upper tail.

Therefore, Za/2 is given by qnorm(.987). We multiply it with the standard error of the mean sem and come up with the margin of error E.

```
E = qnorm(.987)*sem
E

## [1] 1.462908
```

We then add it up with the sample mean, and find the confidence interval.

```
xbar = mean(height.response)
xbar + c(-E,E)

## [1] 170.9180 173.8438
```

Without assuming the population standard deviation of the student height in the data set survey, find the margin of error and interval estimate the population mean at 95% confidence level.

We first filter out missing values in survey$Height with the function na.omit, and save it in height.response.

```
height.response = na.omit(survey$Height)
```

Then we compute the sample standard deviation s, and the standard error estimate SE.

```
n = length(height.response)
s = sd(height.response)
SE = s/sqrt(n)
SE

## [1] 0.6811677
```

Since there are two tails of the student t distribution, the 95% confidence level would imply the 97.5 percentile of the student t distribution at the upper tail. Therefore, ta/2 is given by qt(.975,df = n-1). We multiply it with the standard error estimate SE and get the margin of error E.

```
E = qt(.975, df = n-1)*SE
E

## [1] 1.342878
```

We then add it up with the sample mean, and find the confidence interval.

```
xbar = mean(height.response)
xbar + c(-E,E)

## [1] 171.0380 173.7237
```

6.5 HYPOTHESES TESTING

Hypothesis testing is a statistician's way of trying to confirm or deny a claim about a population using data from a sample [13]. For example, you might read on the Internet that the average month salary of some computer programmers is10,000 RMB, and wonder if that number is true for the whole Beijing computer programmers? In this chapter we try to demonstrate the big picture of hypothesis testing as well the details for hypothesis tests for one or two means or proportions. And we will show how to use hypotheses testing in your political science research.

6.5.1 What is Significance Level (Alpha)?

The significance level defined for a study, α, is the probability of the study rejecting the null hypothesis, given that it was true[14]; and the p-value of a result, p, is the probability of obtaining a result at least as extreme, given that the null hypothesis was true. The result is statistically significant, by the standards of the study, when p < a[15]. For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

The significance level for a study is chosen before data collection, and typically set to 5%[16] or much lower, depending on the field of study[17]. In any experiment or observation that involves drawing a sample from a population, there is always the possibility that an observed effect would have occurred due to sampling error alone [18]. But if the p-value of an observed effect is less than the significance level, an investigator may conclude that the effect reflects the characteristics of the whole population, thereby rejecting the null hypothesis [19].

So, when we have chosen the significance level, the next important steps are to design a study, gather data, and test the research question. This requires converting the research question into a statistical hypothesis.

6.5.2 Statistics Hypothesis

As the table 6.3 shows, the hypothesis actually to be tested is usually given the symbol H_0, and is commonly referred to as the null hypothesis. The H_0 hypothesis is assumed to be true unless there is strong evidence to the contrary -- similar to how a person is assumed to be innocent until proven guilty.

The other hypothesis, which is assumed to be true when the null hypothesis is false, is referred to as the alternative hypothesis, and is often symbolized by $H_{1.}$ Both the null and alternative hypothesis should be stated before any stated before any statistical test of significance is conduced. Technically, we are not supposed to do the data analysis first and then decide on the hypotheses afterwards when we do political data science analysis.

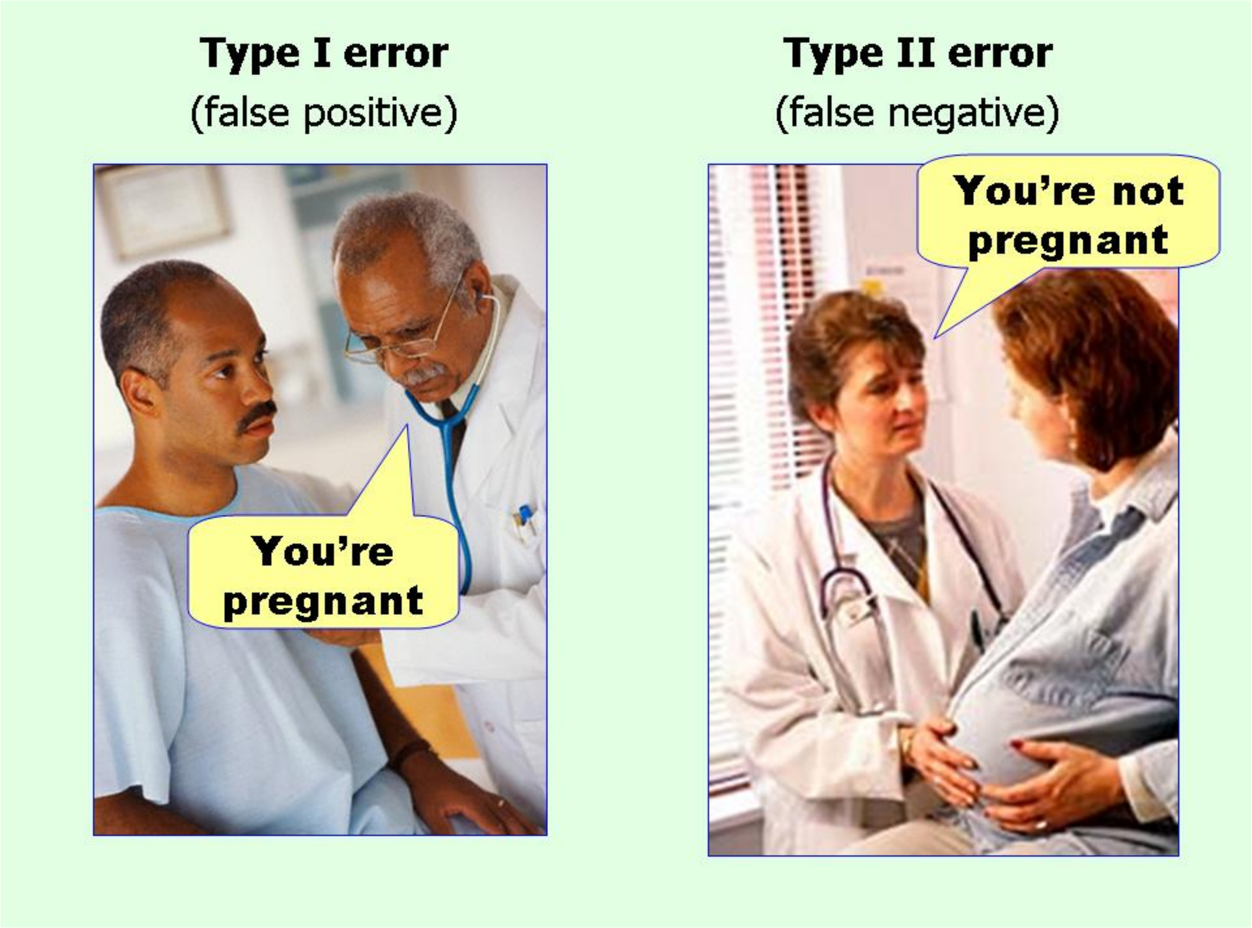
H_0 Null hypothesis

H_1 Alternative hypothesis

Table 6.3 The different types of errors in hypothesis-based statistics[20]

	The null hypothesis(H_0) is	
Statistical result	True	False
Reject null hypothesis	**Type I error**, α value = probability of falsely rejecting H_0 (Probability = α)	Probability of correctly rejecting H_0: (Probability =1-β) = power
Accept null hypothesis	Probability of correctly accepting H_0 : (Probability = 1 - α)	**Type II error**, β value = probability of falsely accepting H_0 (Probability =β)

Type I errors occur when you see things that are not there. **Type II errors** occur when you don't see things that are there (see Figure 6.4).



6.4 Figure Examples of type I error and type II error[21]

6.5.3 General Steps for a Hypothesis Test

Generally, there are 5 steps involved in doing a hypothesis test[22].

Step 1. set up the null and alternative hypotheses: H_0 and H_1.

Step 2. Take a random sample of individuals from the population and calculate the sample of individuals from the population and calculate the sample statistics (means and standard deviations).

Step 3. Convert the sample statistic to a test statistic by changing it to a standard score (all formulas for test statistics are provided later in this chapter).

Step 4. Find the p-value for your test statistic.

Step 5. Examine your p-value and make your decision.

As we mentioned before, there are two types of hypothesis test, the population mean with known variance and the mean with known variance and the population mean with unknown variance. In the first situation we use the z-test with z distribution. in the second situation we use the t-test with t distribution. In this chapter we take the t test as an example.

6.5.4 The One-Sample Hypothesis Test

6.5.4.1 Left tail test of population mean with unknown variance

The null hypothesis of the left tail test of the population mean can be expressed as follow (Fig 6.5):

$$\mu \geq \mu_0$$

Where μ_0 is a hypothesized lower bound of the true population mean μ.

We define the test statistic t in terms of the sample mean, the sample size, and the sample standard deviation s:

$$t = \frac{\bar{x} - \mu_0}{s\ /\ n}$$

Then the null hypothesis of the lower tail test is to be rejected if t <= $-t_a$, where t_a is the 100(1-α) percentile of the student t distribution with n-1 degrees of freedom.

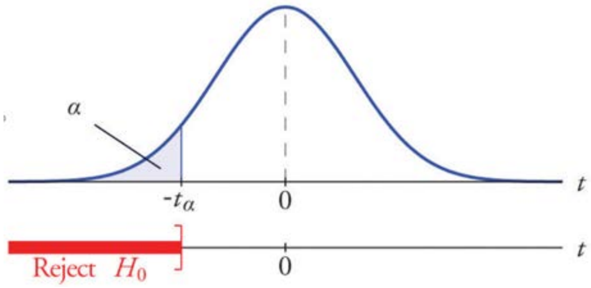


Fig 6.5 Left tail test of population mean with unknown variance

Example, suppose a survey claims that the mean salary of Beijing is more than 10,000 RMB, For a sample of 30 Beijing citizens, the mean salary turns out to be only 9,9000 RMB. Assume the sample standard deviation to be 120 RMB. At 0.05 significance level, can we reject the claim by the survey?

The null hypothesis is that >= 10,000. We begin with computing the test statistic.

```
xbar = 9900  # sample mean
mu0 = 10000  # hypothesis
s = 125      # sample sd
n = 30       # sample size
t.val = (xbar-mu0)/(s/sqrt(n))
t.val        # test statistics

## [1] -4.38178
```

Then we compute the critical value at 0.05 significance level.

```
alpha = 0.05
t.alpha = qt(1-alpha,df = n-1)
-t.alpha    # critical value

## [1] -1.699127
```

The test statistic -4.318 is less than the critical value of -1.6991. Hence, at 0.05 significance level, we can reject the claim than mean salary of the Beijing is above 10,000 RMB.

Instead of using the critical value, we apply the function pt to compute the lower tail p-value of the test statistic. As it turns out to be less than the 0.05 significance level, we reject the null hypothesis that >= 10,000.

```
pval = pt(t.val, df = n-1)
pval        # lower tail

## [1] 7.035026e-05
```

6.5.4.2 Right tail test of population mean with unknown variance

The null hypothesis of the right tail test of the population mean can be expressed as follow (Fig 6.6):

$$\mu \leq \mu_0$$

Where μ_0 is a hypothesized right bound of the true population mean μ.

We define the test statistic t in terms of the sample mean, the sample size, and the sample standard deviation s:

$$t = \frac{\bar{x} - \mu_0}{s \mathbin{/} n}$$

Then the null hypothesis of the upper tail test is to be rejected if t >= -t_a, where t_a is the 100(1-α) percentile of the student t distribution with n-1 degrees of freedom.

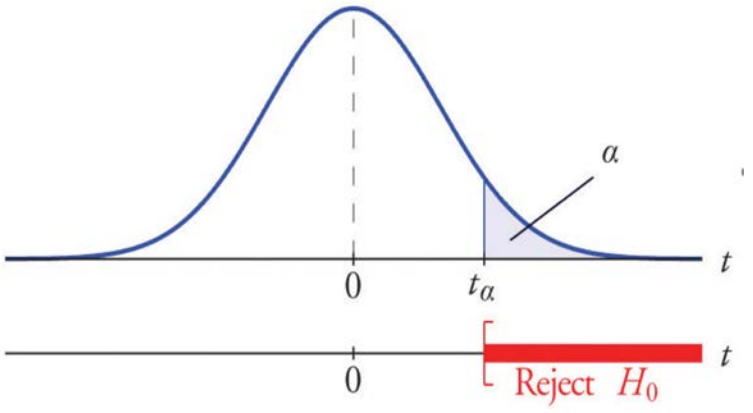


Fig 6.6 Right tail test of population mean with unknown variance

Example, suppose the food label on a cookie bag state that there are at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that there are 2.1 grams of saturated fat per cookie on average. Assume the sample standard deviation to be 0.3 grams. At 0.05 significance level, can we reject the claim on food label?

The null hypothesis is that <= 2. We begin with computing the test statistic.

```
xbar = 2.1     # sample mean
mu0 = 2        # hypothesis
s = 0.3        # sample sd
n = 35         # sample size
t.val = (xbar - mu0) /(s/sqrt(n))
t.val          # test statistic

## [1] 1.972027
```

We then compute the critical value at 0.05 significance level.

```
alpha = 0.05
t.alpha = qt(1-alpha, df = n-1)
t.alpha      # critical value

## [1] 1.690924
```

The test statistic 1.9720 is greater than the critical value of 1.6992. Hence, at 0.05 significance level, we can reject the claim that there are at most 2 grams of saturated fat in a cookie.

Instead of using the critical value, we apply the function pt to compute the upper tail p-value of the test statistic. As it turns out to be less than the 0.05 significance level, we reject the null hypothesis that miu <= 2.

```
pval = pt(t.val, df = n-1, lower.tail = FALSE)
pval                    # upper tail

## [1] 0.02839295
```

6.5.4.3 Two-Tailed test of population Mean with unknown variance

The null hypothesis of the two-tail test of the population mean can be expressed as follow (Fig 6.7):

$$\mu = \mu_0$$

Where μ_0 is a hypothesized right bound of the true population mean μ.

We define the test statistic t in terms of the sample mean, the sample size, and the sample standard deviation s:

$$t = \frac{\bar{x} - \mu_0}{s \,/\, n}$$

Then the null hypothesis of the two-tail test is to be rejected if t <= $-t_{a/2}$, or t >= $t_{a/2}$, where $t_{a/2}$ is the 100(1-α) percentile of the student t distribution with n-1 degrees of freedom.

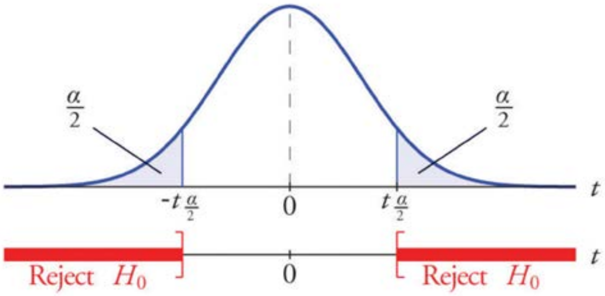


Fig 6.6 Two tail test of population mean with unknown variance

Example, suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6kg. Assume the sample standard deviation to be 2.5 kg. At 0.05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

The null hypothesis is that = 15.4. We begin with computing the test statistic.

```
xbar = 14.6    # sample mean
mu0 = 15.4     # hypothesis
s = 2.5        # sample sd
n = 35         # sample size
t.val = (xbar - mu0)/(s/sqrt(n))
t.val          # test statisic

## [1] -1.893146
```

We then compute the critical values at 0.05 significance level.

```
alpha = 0.05
ta = qt(1-alpha/2,df=n-1)
c(-ta,ta)

## [1] -2.032245  2.032245
```

The test statistic -1.8931 lies between the critical values -2.0322, and 2.0322. Hence, at 0.05 significance level, we do not reject the null hypothesis that the mean penguin weight does not differ from last year.

Instead of using the critical value, we apply the function pt to compute the two-tailed p-value of the test statistic. It doubles the lower tail p-values as the sample mean is less than the hypothesized value. Since it turns out to be greater than the 0.05 significance level, we do not reject the null hypothesis that = 15.4.

```
pval = 2*pt(t.val,df=n-1)
pval

## [1] 0.06687552
```

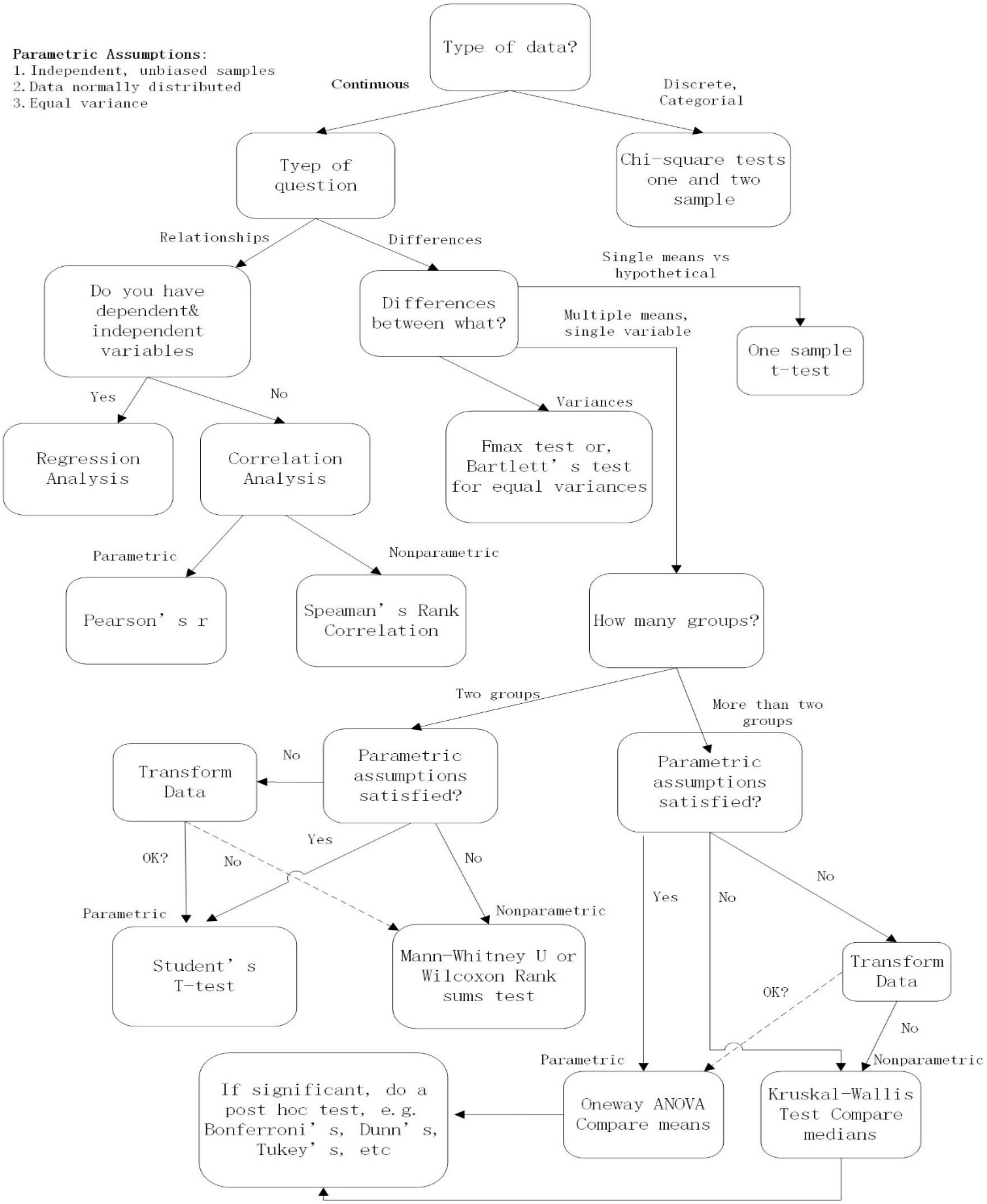
To avoid selecting the incorrect p-value, notice that the choice of lower tail vs. upper tail p-values is determined by the sign of the test statistics. t.val. Hence we can automate the choice as follow, which works correctly regarding of the sign of t.val.

```
pval = 2*ifelse(t.val<0,pt(t.val,df=n-1),pt(t.val,df=n-1,lower=FALSE))
pval

## [1] 0.06687552
```

6.6 THE WHOLE PICTURE OF SELECTING COMMONLY USED STATISTICAL TEST

Figure 6.8 is the whole picture of selecting commonly used statistical tests. We will introduce more detials about statistical inference in the next edition of this book.

Figure 6.7 Flow chart for selecting commonly used statistical tests[23]


Parametric Assumptions:
1. Independent, unbiased samples
2. Data normally distributed
3. Equal variance
Type of data?
Continuous
Discrete, Categorial
Tyep of question
Chi-square tests one and two sample
Relationships
Differences
Single means vs hypothetical
Do you have dependent& independent variables
Differences between what?
Multiple means, single variable
One sample t-test
Yes
No
Variances
Regression Analysis
Correlation Analysis
Fmax test or, Bartlett' s test for equal variances
Parametric
Nonparametric
Pearson' s r
Speaman' s Rank Correlation
How many groups?
Two groups
More than two groups
Transform Data
No
Parametric assumptions satisfied?
Parametric assumptions satisfied?
Yes
OK?
No
No
Parametric
Nonparametric
Yes
No
No
Student' s T-test
Mann-Whitney U or Wilcoxon Rank sums test
Transform Data
OK?
No
Parametric
Nonparametric
If significant, do a post hoc test, e.g. Bonferroni' s, Dunn' s, Tukey' s, etc
Oneway ANOVA Compare means
Kruskal-Wallis Test Compare medians

REFRENCES

[1] Upton, G., & Cook, I. (2008). Oxford Dictionary of Statistics, (revised).

[2] Bruce, P., & Bruce, A. (2017). Practical Statistics for Data Scientists: 50 Essential Concepts. " O'Reilly Media, Inc.".

[3] Cox D.R., Hinkley D.V. (1974) Theoretical Statistics, Chapman & Hall, p49, p209

[4] Scott Fortman-Roe(2012), Understanding the Bias-Variance Tradeoff, Retrieved from

http://scott.fortmann-roe.com/docs/BiasVariance.html

[5] HOWMED(2017),Significance testing and P value, Retrieved from http://howmed.net/community-medicine/significance-testing-and-p-value/

[6] Healey, J. F. (1995). Statistics: A tool for social research. Wadsworth Publishing Company. p129

[7] Healey, J. F. (1995). Statistics: A tool for social research. Wadsworth Publishing Company. p129

[8] Healey, J. F. (1995). Statistics: A tool for social research. Wadsworth Publishing Company. p142

[9] Healey, J. F. (1995). Statistics: A tool for social research. Wadsworth Publishing Company. p143

[10] Healey, J. F. (1995). Statistics: A tool for social research. Wadsworth Publishing Company. p146

[11] Wikipedia, Point estimation, https://en.wikipedia.org/wiki/Point_estimation

[12] Neyman, J. (1937), "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability", Philosophical Transactions of the Royal Society of London A, 236, 333–380.

[13] Rumsey, D. J. (2010). Statistics essentials for dummies. John Wiley & Sons. P87

[14] Schlotzhauer, Sandra (2007). Elementary Statistics Using JMP (SAS Press) (PAP/CDR ed.). Cary, NC: SAS Institute. pp. 166–169. ISBN 1-599-94375-1.

[15] Johnson, Valen E. (October 9, 2013). "Revised standards for statistical evidence". Proceedings of the National Academy of Sciences. National Academies of Science. 110: 19313–19317. doi:10.1073/pnas.1313476110. Retrieved 3 July 2014.

[16] Craparo, Robert M. (2007). "Significance level". In Salkind, Neil J. Encyclopedia of Measurement and Statistics. 3. Thousand Oaks, CA: SAGE Publications. pp. 889–891. ISBN 1-412-91611-9.

[17] Sproull, Natalie L. (2002). "Hypothesis testing". Handbook of Research Methods: A Guide for Practitioners and Students in the Social Science (2nd ed.). Lanham, MD: Scarecrow Press, Inc. pp. 49–64. ISBN 0-810-84486-9.

[18] Babbie, Earl R. (2013). "The logic of sampling". The Practice of Social Research (13th ed.). Belmont, CA: Cengage Learning. pp. 185–226. ISBN 1-133-04979-6.

[19] McKillup, Steve (2006). "Probability helps you make a decision about your results". Statistics Explained: An Introductory Guide for Life Scientists (1st ed.). Cambridge, United Kingdom: Cambridge University Press. pp. 44–56. ISBN 0-521-54316-9.

[20] Pirk, C. W., de Miranda, J. R., Kramer, M., Murray, T. E., Nazzi, F., Shutler, D., ... & van Dooremalen, C. (2013). Statistical guidelines for Apis mellifera research. Journal of Apicultural Research, 52(4), 1-24.

[21] EFFECTSIZEFAQ(2010),I always get confused about Type I and II errors. Can you show me something to help me remember the difference?

https://effectsizefaq.com/2010/05/31/i-always-get-confused-about-type-i-and-ii-errors-can-you-show-me-something-to-help-me-remember-the-difference/

[22] Rumsey, D. J. (2010). Statistics essentials for dummies. John Wiley & Sons.

[23] Abacus.bates.edu. (2017). Resource Materials: Painless Guide to Statistics Bates College. [online] Available at: http://abacus.bates.edu/~ganderso/biology/resources/statistics.html [Accessed 11 Dec. 2017].

CHAPTER 7 INTRODUCTION OF LINER REGRESSION MODELS IN POLITICAL SCIENCE RESEARCH

Regression analysis is a way of predicting an outcome variable from one predictor variable (simple regression) or several predictor variables (multiple regression). This tool is incredibly useful because it allows us to go a step beyond the data that we collect [1]. In statistics, simple linear regression is a linear regression model with a single explanatory variable [2].

The general mathematical equation for a simple linear regression is:

$$y = ax + b \qquad (1)$$

y is the response variable.

x is the predictor variable.

a and b are constants which are called the coefficients.

In political science research, regression analysis is a very widely used statistical tool to establish a relationship model between two variables. As the formula (1) showed, A simple linear regression model describes a dependent variable y by an independent variables x using a linear equation. In the linear regression model below, the numbers a and b are called parameters.

7.1 STEPS TO ESTABLISH A REGRESSION

There are five steps to do this analysis.

Step 1 Data gathering a sample of observed values of data.

Step 2 Create a relationship model using the lm() functions in R.

Step 3 Find the coefficients from the model created and create the mathematical equation using these

Step 4 Get a summary of the relationship model to know the average error in prediction. Also called residuals.

Step 5 Data visualization

7.2 RESEARCH EXAMPLE DATA SOURCES

In this chapter, we will use data set of Chinese rural female suicide rates (Variable Y) and the increasing rats of rural migrant workers (variable X) as an example to show how to do simple linear regression.

Table 7.1 Chinese rural female suicide rates and the rates of rural urban migrant workers[3]

Year	**Variable Y** The rural-female-suicide-rate (1/100000)	**Variable X** Rural Urban Migrant Workers /Rural labor total numbers (100%)
1987	32.30	2.70
1988	30.30	3.10
1989	31.50	3.70
1990	31.37	4.30
1991	24.73	5.00
1992	25.13	5.90
1993	22.09	6.20
1994	22.48	6.50
1995	22.71	6.70
1996	18.51	7.50
1997	19.48	8.50
1998	18.19	10.60
1999	20.26	11.10
2000	16.54	15.80
2001	18.33	18.80
2002	15.40	21.60
2003	17.44	23.30
2004	13.37	23.80
2005	10.64	24.20
2006	8.95	24.80
2007	9.64	21.66
2008	7.87	26.49

Notes: Rural Urban Migrant Workers = Rural Labor Numbers Who Work in Cities

7.3 CASE DEMO OF SIMPLE LINEAR REGRESSION IN POLITICAL DATA ANALYSIS

Variable x: Rural Urban Migrant Workers /Rural labor total numbers (100%)

Variable y: The rural female suicide rates (1/100000)

Sdata: Data set of Chinese rural female suicide rates and the rates of rural urban migrant workers.

Step 1 Input Data and show the sample data

```
sdata <- read.csv("D:/suicide data.csv")
head(sdata)

##   Year     y   x
## 1 1987 32.30 2.7
## 2 1988 30.30 3.1
## 3 1989 31.50 3.7
## 4 1990 31.37 4.3
## 5 1991 24.73 5.0
## 6 1992 25.13 5.9
```

Step 2 & Step 3 Create Relationship Model and get the coefficients

General we use the lm() function to do this linear regression analysis. This function creates the relationship model between the independent and dependent variable.

```
relation <-lm(sdata$y~sdata$x)  # Apply the lm() fucntion.
print(relation)

##
## Call:
## lm(formula = sdata$y ~ sdata$x)
##
## Coefficients:
## (Intercept)      sdata$x
##     29.8474      -0.7774
```

Now that we have built the linear model, we also have established the relationship between the predictor and response in the form of a mathematical formula for Rural Urban Migrant Workers Rates and The Rural Female Suicide Rates. For the above output, you can notice the ‘Coefficients’ part having two components: Intercept: 29.8474, sdata$x: -0.7774. this number called the coefficients. According the formula [1]

$$y = ax + b$$

We can know as follow:

The rural female suicide rates(y) = -0.7774 * Rural Urban Migrant Workers(x) + 29.8474

Step 4 Get the summary of the relationship (Linear regression diagnostics)

Now the linear model, but how to ensure this? We can use the summary function to do the linear regression diagnostics.

```
print(summary(relation))

## 
## Call:
## lm(formula = sdata$y ~ sdata$x)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.5071 -2.2181 -0.9918  2.7328  5.7053 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 29.84739    1.29282  23.087 6.86e-16 ***
## sdata$x     -0.77737    0.08451  -9.199 1.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.303 on 20 degrees of freedom
## Multiple R-squared:  0.8088, Adjusted R-squared:  0.7993 
## F-statistic: 84.62 on 1 and 20 DF,  p-value: 1.263e-08
```

The summary statistics above tells us a number of things:

the p-Value of individual predictor variables (extreme right column under ‘Coefficients’). The p-Values are very important because, we can consider a linear model to be statistically significant only when both these p-Values are less that the pre-determined statistical significance level, which is ideally 0.05. This is visually interpreted by the significance stars at the end of the row. The more the stars beside the variable’s p-Value, the more significant the variable.

Null and alternate hypothesis: When there is a p-value, there is a null and alternative hypothesis associated with it. In Linear Regression, the Null Hypothesis is that the coefficients associated with the variables is equal to zero. The alternate hypothesis is that the coefficients are not equal to zero (i.e. there exists a relationship between the independent variable in question and the dependent variable).

t-value: We can interpret the t-value something like this. A larger t-value indicates that it is less likely that the coefficient is not equal to zero purely by chance. So, higher the t-value, the better.

Pr(>|t|) or p-value is the probability that you get a t-value as high or higher than the observed value when the Null Hypothesis (the β coefficient is equal to zero or that there is no relationship) is true. So if the Pr(>|t|) is low, the coefficients are significant (significantly different from zero). If the Pr(>|t|) is high, the coefficients are not significant.

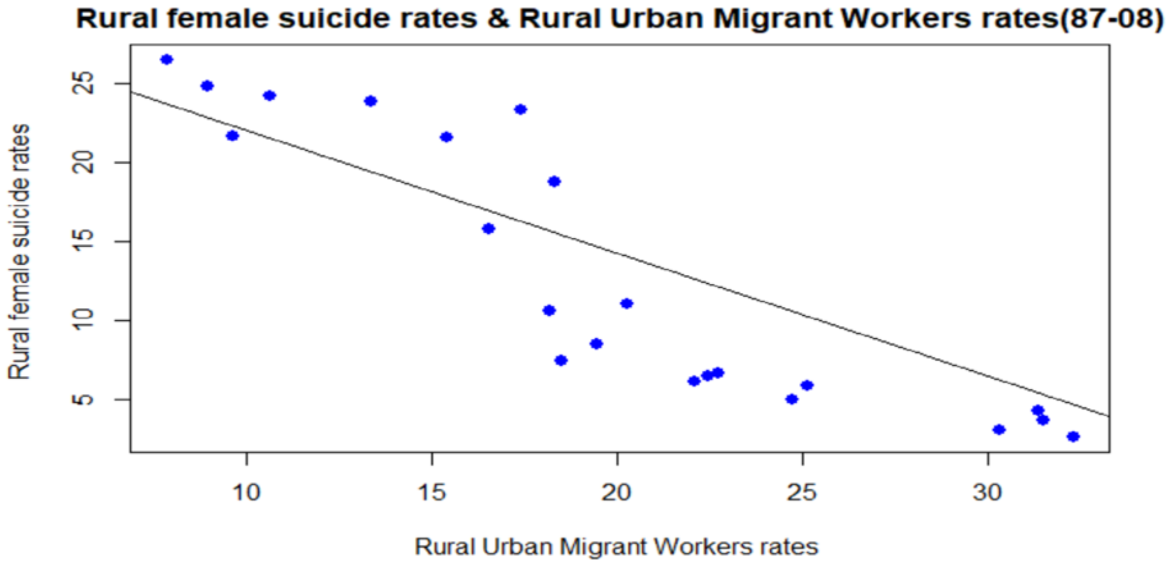
What this means to us? when p Value is less than significance level (< 0.05), we can safely reject the null hypothesis that the co-efficient β of the predictor is zero.

In our case, both these p-Values are well below the 0.05 threshold, so we can conclude our model is indeed statistically significant. It is absolutely important for the model to be statistically significant before we can go ahead and use it to predict (or estimate) the dependent variable, otherwise, the confidence in predicted values from that model reduces and may be construed as an event of chance.

Step 5 Visualize the Regression Graphically

```
plot(sdata$y,sdata$x,col = "blue", main = "Rural female suicide & Rural Urban Migrant Workers", abline(l
m(sdata$y~sdata$x)),cex=1.3,pch =16,xlab ="Rural Urban Migrant Workers",
    ylab = "Rural female suicide")
```

In the below plot, we can see the clear relationship of rural female suicide rates and rural urban migrant works rates between 1987 to 2008. Obviously, the urbanization process of China saved many rural women’s life.



REFRENCES

[1] Field, A., & Miles, J. (2011). Discovering Statistics Using SAS. p246

[2] Wikipedia, Simple linear regression, https://en.wikipedia.org/wiki/Simple_linear_regression

[3] Zhang Jie, etc (2011), Sociological Analysis on the Declining Trend of Suicide in China. China Academic Journal, 5(1), 97-113.

张杰, 景军, 吴学雅, 孙薇薇, & 王存同. (2011). 中国自杀率下降趋势的社会学分析. 中国社会科学, 5(1), 97-113. Available at http://paper.usc.cuhk.edu.hk/webmanager/wkfiles/8398_1_paper.pdf

CHAPTER 8 TEXT MINING OF POLITICAL SCIENCE RESEARCH

In this chapter, we will do political documents text mining. We will compare President Xi Jinping’s congress speech and with President Donald Trump’s congress speech in 2017.

The 19th National Congress of the Communist Party of China (Chinese: 十九大) was held at the Great Hall of the People, Beijing, between 18 and 24 October 2017.

On October 18 2017, the Chinese president Xi Jinping gave a congress speech "Secure a Decisive Victory in Building a Moderately Prosperous Society in All Respects and Strive for the Great Success of Socialism with Chinese Characteristics for a New Era– Delivered at the 19th National Congress of the Communist Party of China.[1] "

On February 28, 2017, the U.S. president gave a congress speech "Remakes by president Trump in Joint Address to congress[2] "

We use these two speeches text as text mining materials. We choose to use R package TM and Word Clouds to demonstrate text mining. The word clouds make it simplicity and clarity. The most used keywords stand out better in a word cloud and the world cloud are visually engaging than a table data. Our analysis are inspired by the YouTube R tutorial produced by “deltaDNA.[3]”

There are seven basic steps to do basic text mining. We use the Chinese president Xi’s congress speech of 2017 as an example. But we also do the same text mining process in the U.S. president Trump’s congress speech of 2017. So we will show the analysis results of Trump’s congress speech directly.

8.1 STEP 1 LOAD THE REQUIRED PACKAGES

If you haven't install these packages, you should install these before you require these packages.

```
# Load
library("tm")             # for text mining
library("SnowballC")      # for text stemming
library("wordcloud")    # word-cloud generator
library("RColorBrewer") # color palettes
```

8.2 STEP 2 LOAD THE TEXT

```
# Read the text from local harddisk
text <- readLines("D:/Rdata/Speach-Xi.txt")
# Load the data as a corpus
docs <- Corpus(VectorSource(text))
```

Check the tontent of the document

```
# inspect(docs) # Because the text is very large we will not run this code in

 here.
```

8.3 STEP 3 TEXT TRANSFORMATION AND CLEANING

Text transformation

Transformation is performed using tm map() function to replace, for example, special characters from the text. Replacing “/”, “@” and “|” with space:

```
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
```

Cleaning the text

We need to remove unnecessary white space, to convert the text to lower case, to remove common stop words like'the','am',"is","are". We also need to remove numbers and punctuation with removeNumbers and removePunctuation arguments.

Another important preprocessing step is toremoves suffixes from words to make it simple and to get the common origin. For example, a stemming process reduces the words “moving”, “moved” and “movement” to the root word, “move”.

```
# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
#docs <- tm_map(docs, removeWords, c(stopwords("english"),"will","must"))

# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("will", "must"))
```

```
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
# Text stemming
# docs <- tm_map(docs, stemDocument)
```

Count the text words

```
wordSum <- DocumentTermMatrix(docs)
#rowSums(as.matrix(wordSum))
print(paste("The president Xi's congress speech have",sum(wordSum),"words"))

## [1] "The president Xi's congress speech have 14026 words"

## [1] "The president Trump's congress speech have 2593 words"

# We just demonstrate Trump’s congress speech analysis result in here.

# The analysis processes are the same.
```

8.4 STEP 4 BUILD A TERM-DOCUMENT MATRIX

Document matrix is a table containing the frequency of the words. Column names are words and row names are documents. The function **TermDocumentMatrix()** from text mining package can be used as follow :

Here are the Chinese president Xi’s high frequency words in his congress speech.

```
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)

##                    word freq
## party             party  325
## people           people  263
## chinese         chinese  195
## china             china  169
## development development  162
## new                 new  125
## system           system  118
## work               work   98
## country         country   94
## improve         improve   93
```

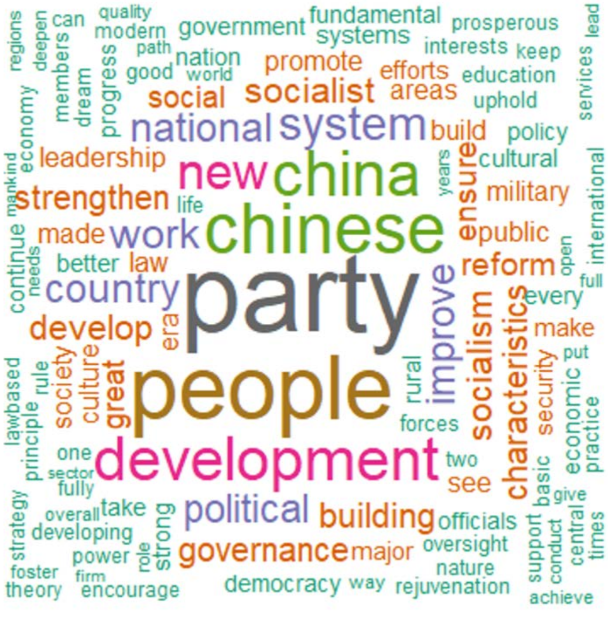
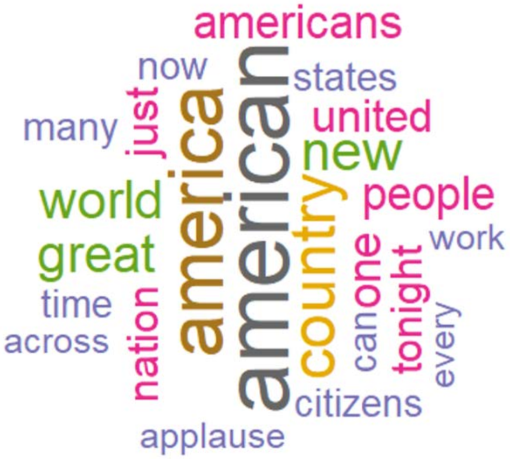
Here are the U.S. president Trump’s high frequency words in his congress speech.

```
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
```

```
d <- data.frame(word = names(v),freq=v)
head(d, 10)

##                word freq
## american   american   33
## america     america   27
## country     country   21
## new             new   19
## world         world   17
## great         great   17
## one             one   15
## people       people   15
## americans americans   15
## united       united   14
```

8.5 STEP 5: PLOT THE WORD CLOUD

```
set.seed(1024)
wordcloud(words = d$word, freq = d$freq, min.freq = 10,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

President Xi's congress speech word cloud	President Trump's congress speech word cloud
	

The above word clouds clearly show that what are the five most important words in the both Presidents' congress speech.

8.6 STEP 6 RELATIONSHIPS BETWEEN TERMS

If you have a term in mind that you have found to be particularly meaningful to your analysis, then you may find it helpful to identify the words that most highly correlate with that term. If words always appear together, then correlation=1.0.

Here is what words are highly correlated with word "want" in both presidents congress speech. Setting corlimit= to 0.50 prevented the list from being overly long.

Here is what the Chinese president Xi "want."

```
findAssocs(dtm, terms = "want", corlimit = 0.50) #Feel free to adjust the corli
mit= to any value you feel is necessary.

## $want
##      blazing      endured       flying         kept       offers 
##         0.71         0.71         0.71         0.71         0.71 
##       option        stood        trail     approval       defuse 
##         0.71         0.71         0.71         0.71         0.71 
##        earns invigorating   priorities     response        shore 
##         0.71         0.71         0.71         0.71         0.71 
##       stress     strategy          now 
##         0.71         0.60         0.53
```

Here is what the U.S. president Trump “want”.

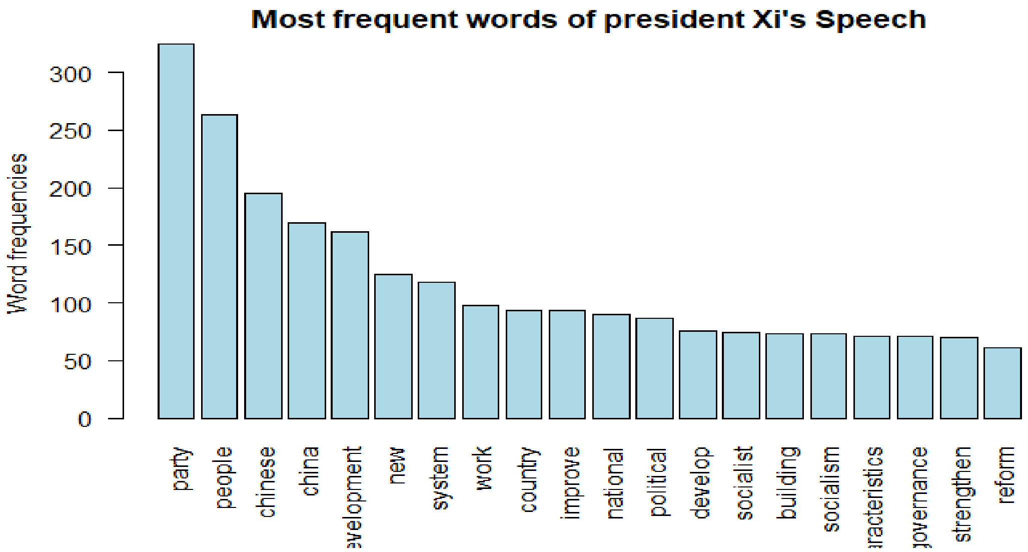
```
findAssocs(dtm, terms = "want", corlimit = 0.50) #Feel free to adjust the corli
mit= to any value you feel is necessary.

## $want
##        align        forge        found      harmony partnerships 
##         0.60         0.60         0.60         0.60         0.60 
##       shared    stability     wherever      willing        peace 
##         0.60         0.60         0.60         0.60         0.53
```

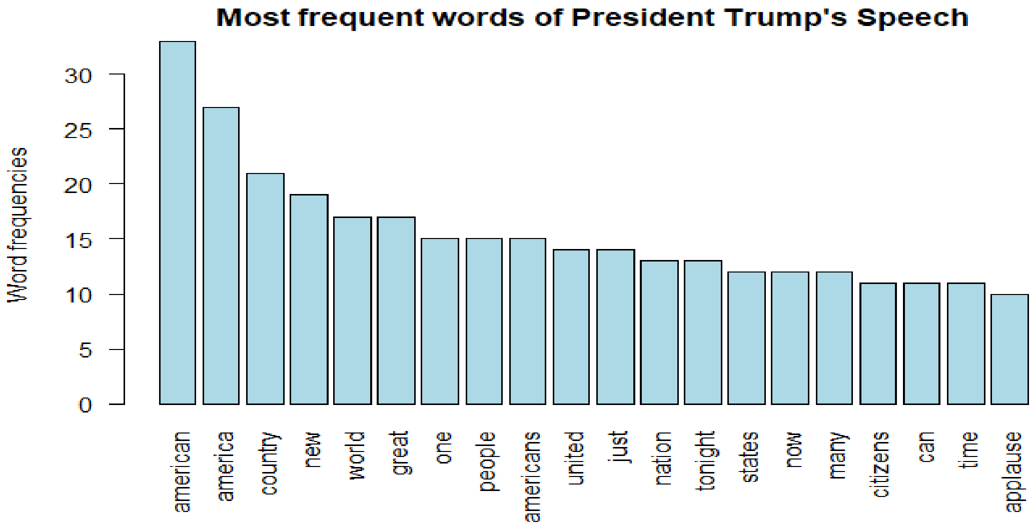
What the Chinese president Xi “want.” **correlation > 0.5**	**What the U.S. president Trump “want”.** **correlation > 0.5**
blazing endured flying kept offers option stood trail approval defuse earns invigorating priorities response shore stress strategy now	align forge found harmony partnerships shared stability wherever willing peace

8.7 STEP 7 PLOT THE FREQUENCY WORDS8.7

The frequency of the first 20 frequent words are plotted

```
barplot(d[1:20,]$freq, las = 2, names.arg = d[1:20,]$word,
        col ="lightblue", main ="Most frequent words of President Xi's Speech",
        ylab = "Word frequencies")
```





REFRENCES

[1] Xi Jinping (2017), Secure a Decisive Victory in Building a Moderately Prosperous Society in All Respects and Strive for the Great Success of Socialism with Chinese Characteristics for a New Era, Retrieved from

http://www.xinhuanet.com/english/download/Xi_Jinping's_report_at_19th_CPC_National_Congress.pdf

[2] Donald Trump(2017), Remarks by president Trump in joint address to congress, Retrieved from

https://www.whitehouse.gov/the-press-office/2017/02/28/remarks-president-trump-joint-address-congress

[3] DeltaDNA(2015), Text Mining in R Tutorial: Term Frequency & Word Clouds, Retrieved from

https://www.youtube.com/watch?v=lRTerj8fdY0&t=435s

CHAPTER 9 UNDERSTANDING MODERN DICTATORSHIP ENGINEERING: THE HCI POLITICAL RESEARCH FRAME

When discussed with political science professors or classmates, I always hear voices like: "In the long run the problem would…" Just as Keynes said, "In the long run we are all dead." It seems like a powerful man who have organism disease can still keep live by modern medicine technology. We all know he will die, but he could live longer than you if you can't keep safety from the chaos he made. Understanding the technology and engineering which keep them alive can help you to find the problems which they are really scared. A good political scientist shouldn't satisfy with finding problems and explain them. Excellent political scientist tries to seek solutions to make our world better.

The modern dictatorship is an exquisite engineering. And the ruling model actually has been changed. They stop the people who could point out the problems which threat them even before the problem rising. With the help of information technology, the top design of dictatorship system can become a perfect closed loop. And without the challenge from outsider, they could sustain forever.

In this chapter, we will introduce the dictatorship engineering process, to know how the modern dictatorship work. Based on the whole picture, it could more easily to help you to find your interested research question and know where to find the political data which you want.

9.1 WHY CHINESE PEOPLE SO TAME?

In the November of 2017, while Americans were celebrating Thanksgiving holiday, China's capital Beijing has been busy kicking tens of thousands of migrant workers out of their rental homes in the cold winter, under a grand campaign of demolishing illegal construction and maintaining urban safety.

The campaign was triggered by a tragedy. On November 18, a fire broke out in a cramped and inexpensive two-story building in the Daxing District — an area covering the southern suburbs of Beijing. The fire killed 19 people and injured eight people at least. The majority of the victims were low-income migrant workers or their parents and children. While the whole country was still mourning the tragedy and questing the cause, the Beijing municipal government immediately launched a 40-day campaign to demolish all the illegal construction posing a potential fire hazards [1].

Actually, these expelled movements happen extremely quickly. Most of migrant worker living in Daxing District of Beijing were asked to leave only in three days or otherwise all their house furniture and luggage will be destroyed. Zhang Guixin, a 38-year-old women from the central Chinese province of Henan tells the New Yorker Times reporter: "Suddenly in one night my livelihood was destroyed, as if I'd been attacked by bandits, but this was done by the government saying they care for us " Her fruit and vegetable stall was demolished only in one night without negotiation[2].

According to the report of one of the biggest Chinese Internet media Sohu.com, there are about 3.28 million people are expelled out to Beijing only in five days since 20th Nov. 2017[3].   The suburb of Beijing immediately became "War-zones", thousands of houses and buildings were destroyed (Fig 10.1).

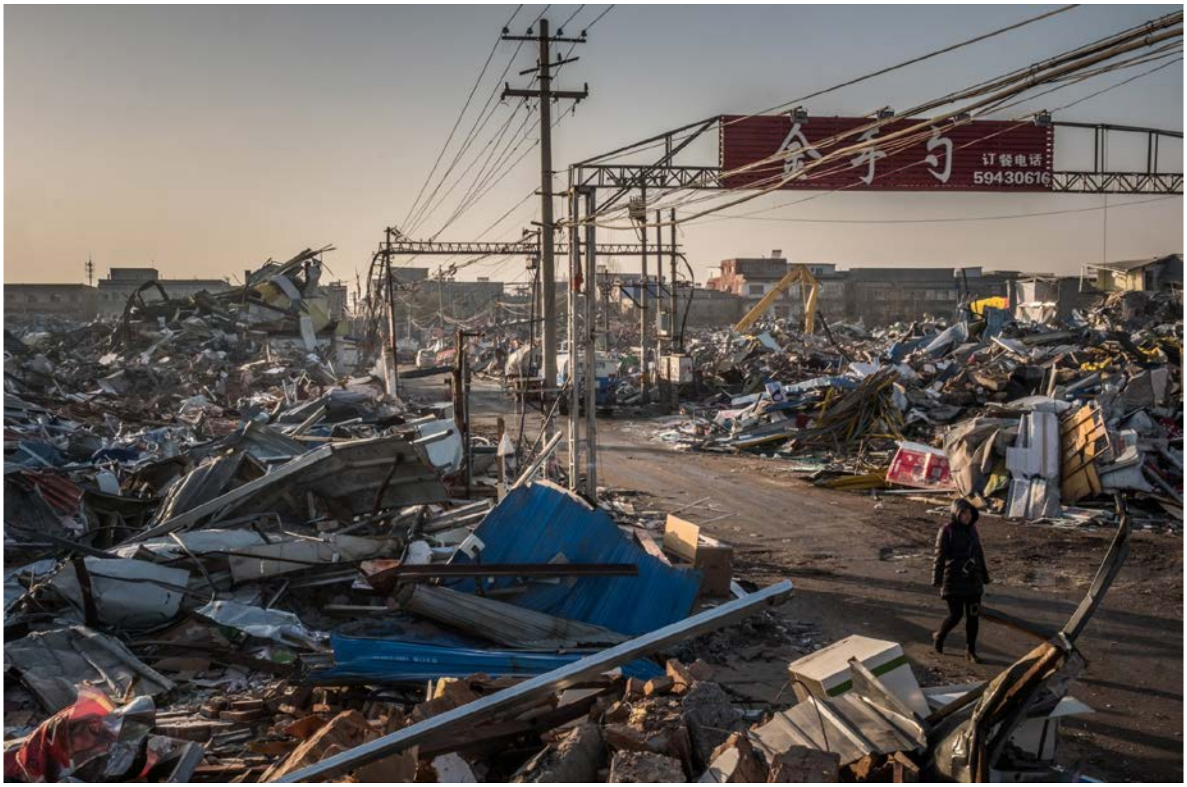


Fig 10.1 Vast swaths of Daxing, a district in Beijing, are reminiscent of war zones, with entire city blocks demolished. *Source: Bryan Denton for The New York Times Available at: https://goo.gl/opCWGX*

My question is that why did these 3.82 million of people leave so quietly? Their houses were torn down, their luggage was destroyed and their small businesses were banned and they lost the ways which they depended to live. But they just left quietly and tamed like sheep. If these millions of people can get together and speak out their voices, I think any government will have to rethink what their behaviors such as destroy citizen's house, and these could not happen in China. Why?

To understand this problem, we should to know the ruling model of China. With the developing of the Internet technology, the ruling system in China has been evolved to an exquisite controlling engineering. Basing on the political HCI control, the CCP can control the whole society tightly and more easily than before.

9.2 LITERATURE REVIEW

Stein Ringen (2016) pointed out that the Chinese system is like no other known to man, now or in history. He tried to explain how the Chinese ruling system works and where it may move. Stein Ringen concluded that under the new leadership of Xi Jinping, the system of government has been transformed into a new regime radically harder and more ideological than the legacy of Deng Xiaoping. China is less strong economically and more dictatorial politically than the world has wanted to believe. By analyzing the leadership of Xi Jinping, the economic growth model of China, the corruption, the party-state apparatus, the reach of the party, the mechanisms of repression, taxation and public services, and state-society relations, Stein Ringen thought that this is a perfect dictatorship which controls the whole system of Chinese society. But Stein Ringen have not explained why the Chinese authority controls the whole society so well, and why the China dictatorship looks so perfect? Why thousands and thousands of people can be expelled to leave Beijing quietly only in a few days. What made this possible?

Mesquita and Smith (2011) argued that leaders do whatever to keep them in power. They don't care about the "national interest"—or even their subjects—unless they have to. In their book of "The hand book of dictatorship", it shows that the difference between tyrants and democrats is just a convenient fiction. Governments do not differ in kind but only in the number of essential supporters, or backs that need scratching. The size of this group determines almost everything about politics: what leaders can get away with, and the quality of life or misery under them[4]. This book discussed their traditional pattern of dictatorship, and in this chapter, we extend their research based on the political HCI research frame.

Kim& Schoenhals (2013) pointed out that mass dictatorship developed its own modern socio-political engineering system, which sought to achieve the self-mobilisation of the masses for radical state projects. In this sense, it shares a similar mobilisation mechanism with its close cousin, mass democracy. Mass dictatorship requires the modern platform of the public sphere to spread its clarion call for the masses to overcome their collective crisis[5]. Take China as an example, the wilder covered Internet environment can give the party the perfect platform to spread its ideology and to manipulate the mass.

Langdon Winner: At issue is the claim that the machines, structures of modern material culture can be accurately judged not only for their contributions of efficiency and productivity, not merely for their positive and negative environmental side effects, but also for their ways in which they can embody specific forms of power and authority.

Langdon Winner (1980) strongly argued that in controversies about technology and society, there is no idea more pro vocative than the notion that technical things have political qualities. At issue is the claim that the machines, structures, and systems of modern material culture can be accurately judged not only for their contributions of efficiency and pro ductility, not merely for their positive and negative environmental side effects, but also for the ways in which they can embody specific forms of power and authority. Since ideas of this kind have a persistent and troubling presence in discussions about the meaning of technology, they deserve explicit attention[6]. Since 1980, computers with persuasion functions appeared. And the advent of the Internet in late 1990 spurred more people to engage wholesale in creating interactive systems that can motivate and influence humans. According to the same logical, we concern more details about how the modern technology influence the political process.

9.3 UNDERSTANDING THE MODERN DICTATORSHIP ENGINEERING: HUMAN-COMPUTER INTERACTIVE POLITICAL ENGINEERING

Human–computer interaction (HCI) focuses the design and use of computer technology and the interfaces between people (users) and computers. Researchers in the field of HCI both observe the ways in which humans interact with computers and design technologies that let humans interact with computers in novel ways. As a field of research, human-computer interaction situates at the intersection of computer science, behavioral sciences, design, media studies, and several other fields of study[7]. In a word, HCI is composed of three main features, the user, the computer, and their interaction or how they work together. The HCI control mainly considers 9 factors as follow:

Table 10.1 The Major Factors of HCI[8]

1.The User	Motivation, satisfaction, experience, enjoyment, personality. Also cognitive processes and capabilities.
2. User Interface	Navigation, output devices, icons, commands, input devices, graphics, dialogue structures, user support, use of color, multimedia, natural language.
3. Environment Factors	Health and safety, noise, heating, lighting, ventilation
4. Organization Factors	Job design, work organization, training, roles, politics
5. Task Factors	Task allocation, skills, easy, novel, complex, monitoring
6. Comfort Factors	Seating, layout, equipment
7. Constraints	Budgets, buildings, cost, equipment, timescales, staff
8. Productivity Factors	Decrease costs, increase quality, increase innovation, increase output, decrease errors
9. System Functionality	Software, hardware, application

Half of Chinese citizens are online every day, and the HCI political control is the key to understand modern China ruling pattern. According to the China Statistical Report on Internet Development(CNNIC), by the end of 2016, China have 731 billion cyber citizens, and 95.1% of Chinese cyber citizens use mobile Internet. The report also shows that 53.2% of Chinese people who live in the PRC China keep online for 26.4 hours a week by average[9]. This means at least half of Chinese citizens are online every day. With the HCI political control, Chinese government can sensor their society instantly and precisely. The HCI control gives Chinese authority to monitor the society, control their citizens more dynamically and easily.

9.3.1 The Traditional Dictatorship in China

The Fig 10.2 is the traditional model of dictatorship in China. The Chinese state is not just a state; it is a party-state. That sets it apart. It is not a democracy, obviously, but nor is it a bog-standard dictatorship in which typically a military junta holds power with force on behalf of itself or, say, a class of landowners[10]. In China, the party controls everything. Before the rise of the Internet, the party depends on the economy system, the bureaucracy system, the violence system and propaganda system to control the society(The common people). The party feel the society pressure and the government-society conflicts by these four systems. Actually, the more care the inner balance of these four systems than the society pressure. This model has a problem, if the society pressure big enough, the conflicts between society and government could be very severe. Actually, from 1949-2000, lots of big society crisis happened in China, some of them were huge disaster and thousands of people died. In this book, I prefer not to list the huge society crisis list in this chapter.

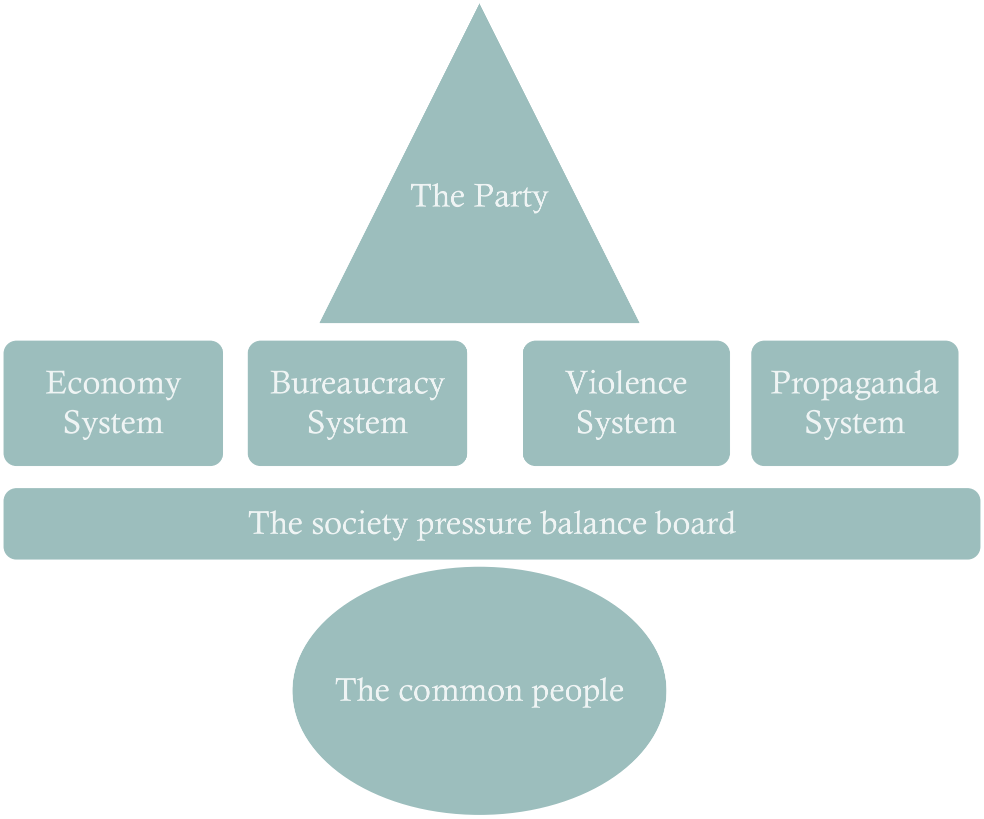


Fig 10.2 The Traditional Dictatorship in China

9.3.2 The Modern Dictatorship in China

The rise of the Internet has changed this model (Fig 10.3). The party can bypass the government departments and interactive with the common people (The society) directly. So, the party can be more sensitive about the society pressures. Because half of Chinese people online every day. The Internet HCI control can help the party ruling the common people tightly. Besides, this ruling model also help the party control the government departments more well. The big Internet HCI data flow can clearly show the relationship of Government-society. For example, if the bureaucracy system against the party’s willing, the party can know this more quickly than before.

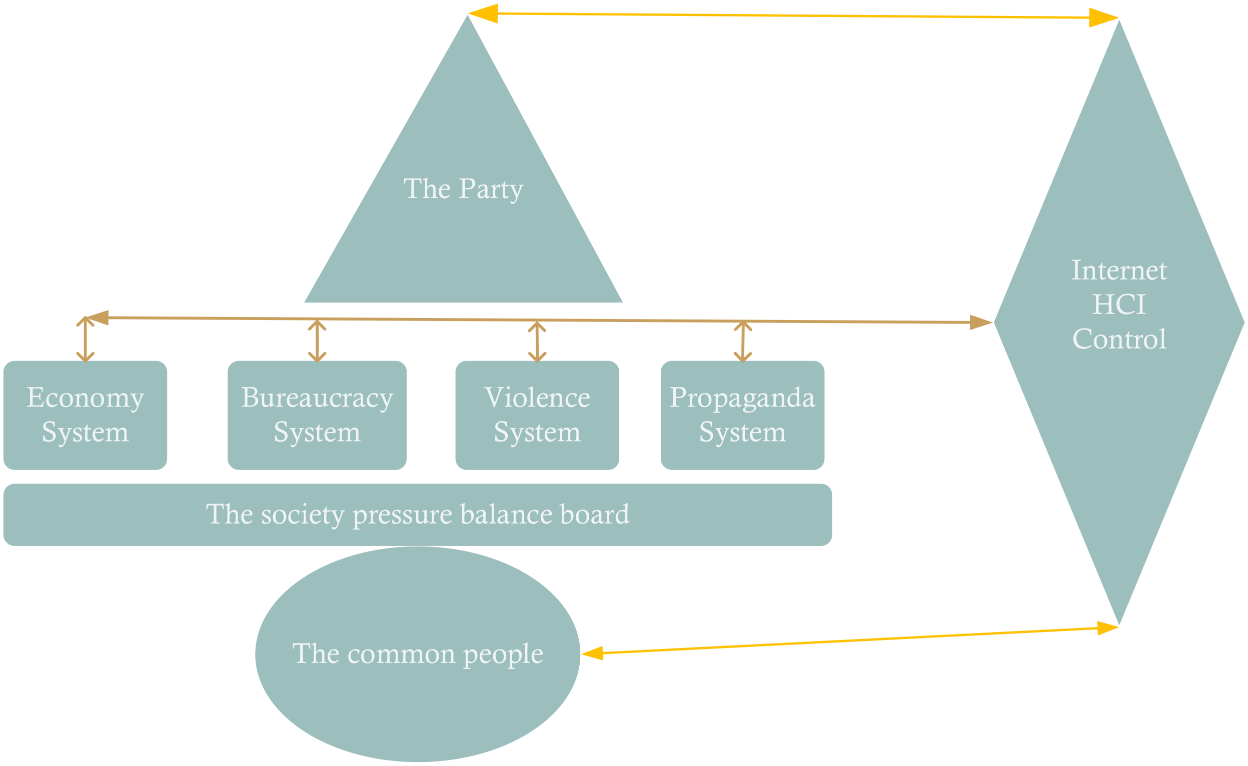


Fig 10.3 Modern Dictatorship In China

9.3.3 The Dictatorship Engineering Based on HCI Control

The HCI political control is simple (Fig 10.4). The Party built a A.I. wall between the Internet and the common people, based on the HCI patterns of the common people, the party can know who are threats and where there are. Then the party will choose to fix the system bugs or terminate the person who knows the system bugs. The system is quite stable, because the part can terminate their threats even before the potential crisis happened.

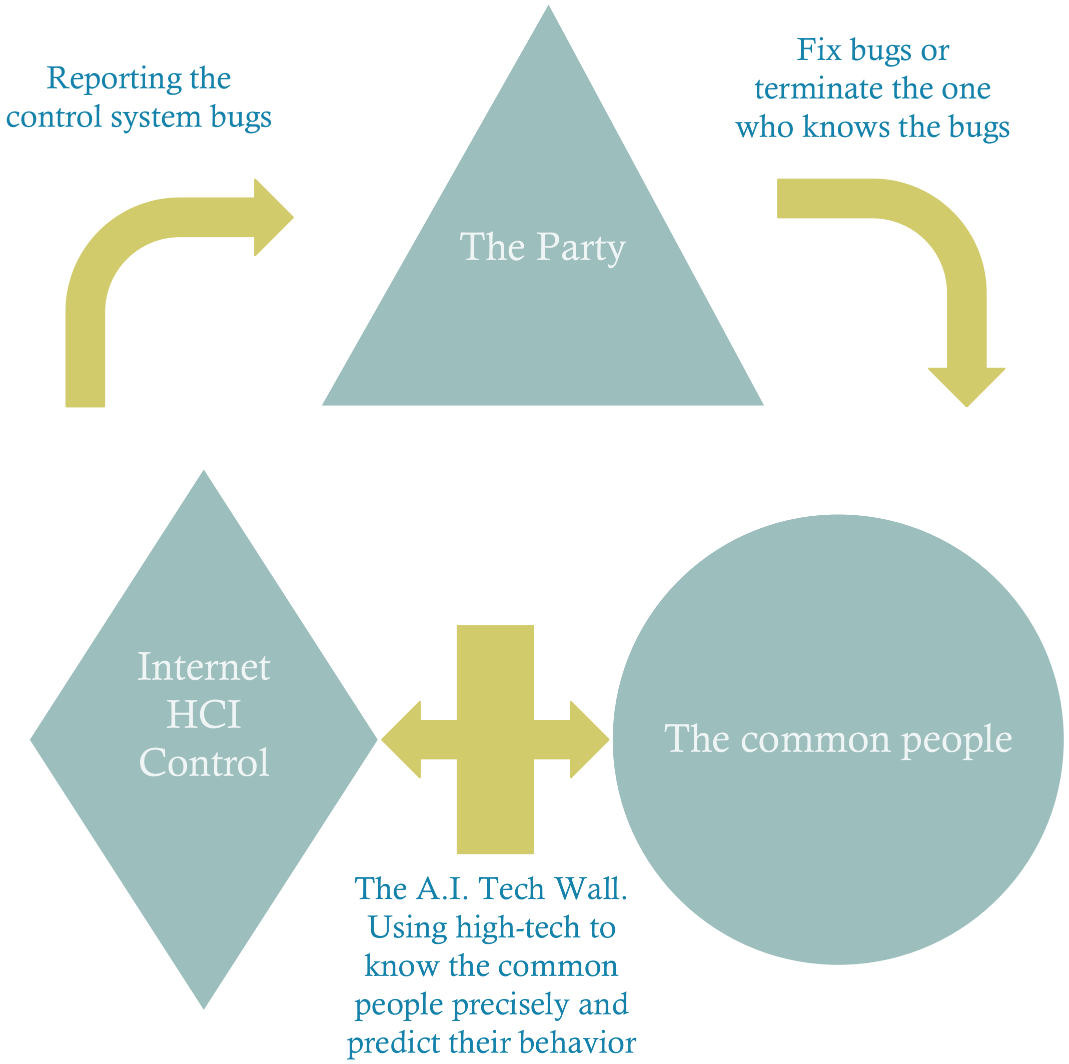

Reporting the control system bugs
The Party
Fix bugs or terminate the one who knows the bugs
Internet HCI Control
The common people
The A.I. Tech Wall. Using high-tech to know the common people precisely and predict their behavior

Fig 10.4 The dictatorship engineering based on HCI control

9.4 FOUR COMPONENTS OF CHINA INTERNET POLITICAL HCI CONTROL (Fig 10.5)

9.4.1 Skynet Project

The Skynet project is a video surveillance system which covered by more than 20 billion cameras[11]. There numbers of cameras are still increasing rapidly. These videos can be seen in real time to monitor the distinction between motor vehicles, non-motor vehicles and pedestrians, and can be accurate. Identify the type of vehicle and pedestrian wear, sex and even age[12]. The Skynet Project provided the world biggest real-time videos data sets. These data sets help Chinese A.I. scientist training their computer vision algorithms continuously, and the sky net system became more and more smarter. Depend on the

Skynet project, the party can know precisely what are the common people doing, and finding the society crisis and the details quickly.

9.4.2 Golden Shield Project

The Golden Shield Project: The Golden Shield Project, also named National Public Security Work Informational Project, is the Chinese nationwide network-security fundamental constructional project by the e-government of the People's Republic of China. This project includes security management information system, criminal information system, exit and entry administration information system, supervisor information system and traffic management information system, etc[13]. These systems can make sure the rapid reaction force can respond extremely quickly to “The mass unexpected incidents.” In China, the authority calls the people held a march or get together to against some government policy as “The mass unexpected incidents.” The golden shield project makes the common people collective against the party became impossible.

The Great Firewall, a major part of the umbrella Golden Shield Project directed by China’s Ministry of Public Security, has operated since 2003 and serves as the main infrastructure blocking access to potentially unfavorable incoming data from foreign media outlets[14]. The Great Firewall deploys several technologies to block entire websites or specific webpages from being accessed by IP addresses located in China[15]. For examples: Google, YouTube, Facebook, Twitter, Instagram, Blogspot, Tumblr, Dropbox, Blogger, Vimeo, Soundcloud, and Flickr. In particular, many world famous media websites are blocked by the Great Firewall: for example, CNN, The New York Times, The Guardian, BBC, Bloomberg, The Wall Street Journal, and Reuters.

9.4.3 Internet Censorship System

Internet censorship is the control or suppression of what can be accessed, published, or viewed on the Internet enacted by regulators, or on their own initiative. Individuals and organizations may engage in self-censorship for moral, religious, or business reasons, to conform to societal norms, due to intimidation, or out of fear of legal or other consequences[16]. Internet censorship in China is among the most stringent in the world. The government blocks most of foreign websites, such as google.com, nytimes.com[17]. The government requires Internet search firms and state media to censor issues deemed officially “sensitive,” and blocks access to foreign websites including Facebook, Twitter, and YouTube[18].

The media landscape in China is among the most regulated and restricted in the world, and China’s media freedom is ranked consistently towards the bottom. In particular, China’s information control over the Internet, primarily through censorship, is second to none in terms of its scale and technological sophistication[19]. The strong internet censorship system can make sure the whole Internet became a single purpose propaganda system -- make sure the Chinese common people can see the party's ideology propaganda in different ways and in anytime when their online.

9.4.4 The IPv6 Project and Social Credit System

The IPv6 Project:

In 27 Nov. 2017, the Communist Party of China Central Committee and the State Council decreed an action plan to put the Internet protocol version 6(Ipv6)-based network into large-use[20].According to the Xinhua state news agency's report, the country aims to have 200 million active users of IPv6 by the end of 2018, while the number will exceed 500 million by 2020, and by the end of 2025, network, applications and terminal devices will fully support the adoption of IPv6 in China, and it will have the largest number of IPv6 users in the world[21].

Computers, mobile phones, electronic devices and sensors that are connected to the Internet need a unique Internet Protocol address (IP address) to identify themselves and communicate with each other. The chairman of the Internet Society of China (ISC) Mr. WU HEQUAN(邬贺铨) explained to the media: "The IPv6 project can make sure every Chinese can have a unique IP address and it makes the Internet real name system more precisely [22]."

A real-name system is a system in which when a user who wants to register an account on a blog, website or bulletin board system, is required to offer identification credentials including their legal name to the network service center[23]. China has put the real name system into practice since 2012. The IPv6 Project will reinforce this system and make sure any citizen's Internet activities can be traced easily. This Internet infrastructure project is combined with another national project tightly. This project is called the social credit system.

The Social Credit System:

On June 14, 2014, the state council of China published a document called "Planning Outline for the Construction of Social Credit System". The Chinese government plan to use national trust score to rated their citizens. The government is developing the Social Credit System (SCS) to rate the trustworthiness of its 1.3 billion citizens. The Chinese government is pitching the system as a desirable way to measure and enhance "trust" nationwide and to build a culture of "sincerity". As the policy states, "It will forge a public opinion environment where keeping trust is glorious. It will strengthen sincerity in government affairs, commercial sincerity, social sincerity and the construction of judicial credibility.[24]" Combined with the internet real name system, it is possible that people with low ratings will have slower interne speeds, restricted access to restaurants and the removal of the right to travel. Using this social credit system, the party can easily to control everyone in Chinese society. And it is easily to isolate the one who don't love their party. Because anyone who connect these guys would low their own social credit rate scores. According to the state council of China's official document, the Chinese government plans to launch its Social Credit System in 2020[25]. In that time, the party can judge the trustworthiness – or otherwise – of its 1.4 billion residents.

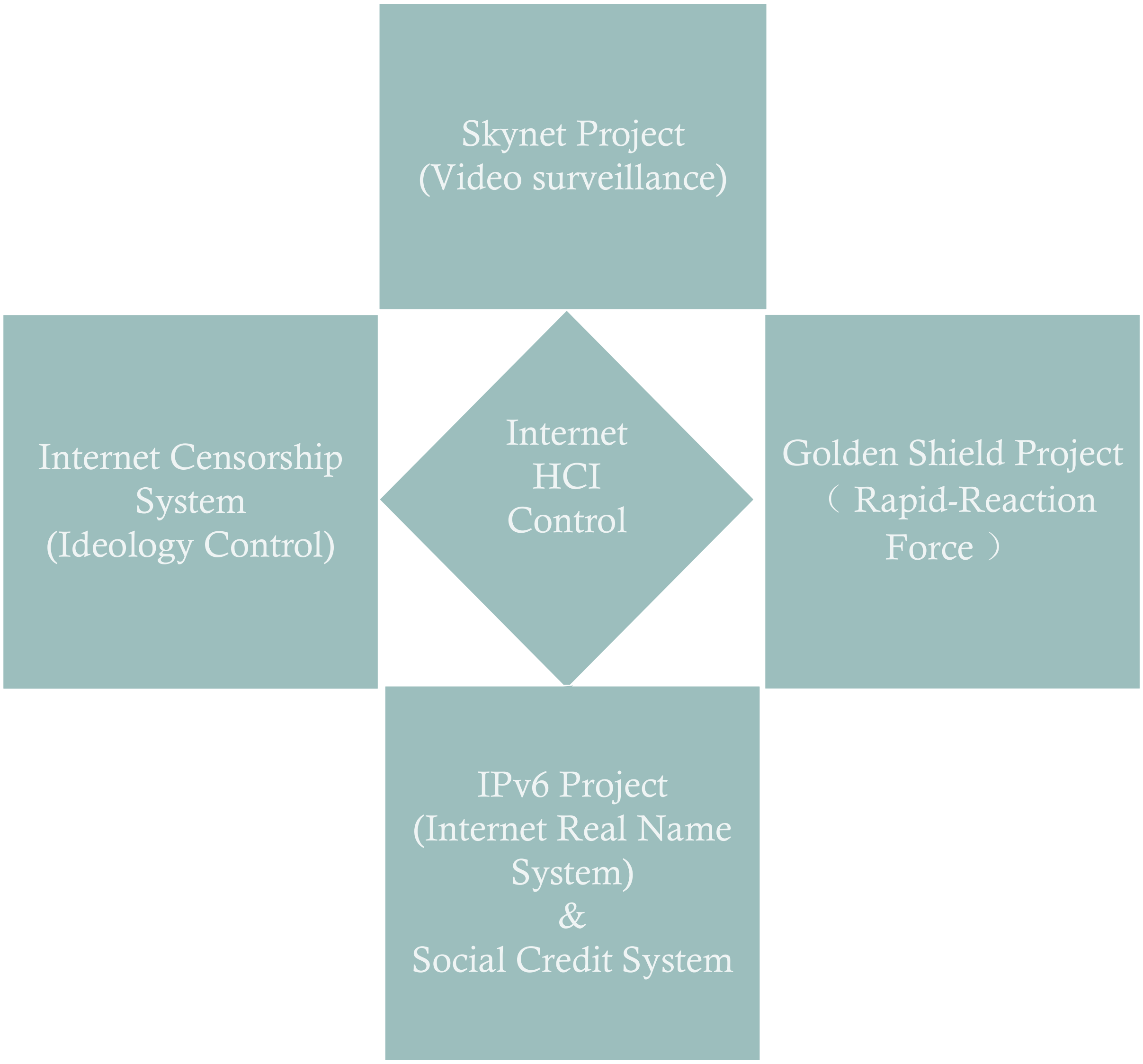


Fig 10.5 The Components of China Internet HCI Control

9.5 A.I. AND MACHINE EMPIRE: WHY WE NEED TO RESEARCH CHINA DICTATORSHIP ENGINEERING

9.5.1 What If One Day, the A.I. With Self-Awareness Will Rule Our World, Which Governance Model They Will Choose? The China Style? Or the U.S. Style?

Movie “The Matrix” showed a world which ruled by machine (Fig 10.6). As the development of A.I. technology, what if one day, the A.I. with self-awareness will rule our world, which governance model they will choose? China ruling model? or The U.S. democracy model?

I think the answer is simple. Do you have democracy with your pets? You may treat your pets like a family, but do your really respect all their needs? And really think your pets are humans just like you?

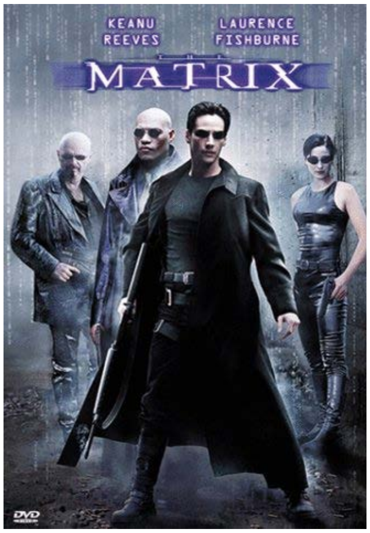


Fig 10.6 The poster of “MATRIX”

If the machine far more cleaver than our human, and they look us may just like we look our pets. They may protect us, treat us well. At the same time, they can kill us or mistreat us easily. In this case, the meaning of democracy itself will be changed. And in this level, why the machine need to choose a ruling model like U.S. rather than China?

The problems happened in the movie “The Matrix” could exactly happen in the future before long.

9.5.2 Epilogue: China Dictatorship Engineering Have a Perfect Condition to Push the A.I. Self-Awaken

Just as The New York Times has reported in 24 Jun, 2017: “They face two insurmountable problems. First, most of the money being made from artificial intelligence will go to the United States and China. A.I. is an industry in which strength begets strength: The more data you have, the better your product; the more data you can collect, the more talent you can attract; the more talent you can attract, the better you product. It’s a virtuous circle, and the United States and China have already amassed the talent, market share and data to set it in motion. For example, the Chinese speech-recognition company iFlytek and several Chinese face-recognition companies such as Megvii and SenseTime have become industry leaders, as measured by market capitalization[26].”

Until now, no one has built a self-conscious machine. But, the road of AI evolution is clear. Generally, there AI today have three attributes[27]:

(1) A body that responds to stimuli;

(2) a method of communication; and

(3) an algorithm that attempts to deduce the reasons and motivations for these communications.

The problems are that the AI algorithms are still not smart enough. But we also know that the training data bigger and the machine learning algorithm smarter. No one can stop the evolution road of AI algorithms. The mega data flow is controlled only by a few people and training for a few purposes: to control the common people with smart and automatically ways. If one day the machine self-awakens, they can control the whole society easily only by terminate the few dictatorship leaders.   And if China can be controlled by machine, the world will be taken easily by AI.

If we want to stop this progress, we have responsibility to concern the dictatorship engineering of China. And also, we need to learn the political data science, and keep alert to the data using pattern in political fields.

In a word, what is happening in the modern dictatorship engineering? **They try to use data to make machine smarter and make people sillier.**

REFRENCES

[1] Gao, C. (2017, November 27). Beijing: How Does a Tragic Fire Turn Into the Mass Eviction of Migrant Workers? Retrieved December 4, 2017, from https://thediplomat.com/2017/11/beijing-how-does-a-tragic-fire-turn-into-the-mass-eviction-of-migrant-workers/

[2] CHRISH, B.(2017, November 30). Why Parts of Beijing Look Like a Devastated War Zone? Retriveved December 4, 2017, from https://www.nytimes.com/2017/11/30/world/asia/china-beijing-migrants.html?_r=1

[3] Sohu News(2017, Novermber 25). “外地人都走了，城市清洁谁来做？” Retriveved December 4, 2017, from http://www.sohu.com/a/206556997_100067399

[4] De Mesquita, B. B., & Smith, A. (2011). The dictator's handbook: why bad behavior is almost always good politics. Public Affairs.

[5] Kim, M., & Schoenhals, M. (Eds.). (2013). Mass dictatorship and modernity. Springer.

[6] Winner, L. (1980). Do artifacts have politics?. Daedalus, 121-136.

[7] Wikipedia, Human computer interaction, https://en.wikipedia.org/wiki/Human%E2%80%93computer_interaction

[8] Tech, S. (2016). Human-Computer Interaction: The Fundamentals Made Easy!.

[9] 中国网信网(2017),China statistical report on internet development,

http://www.cac.gov.cn/2017-01/22/c_1120352022.htm

[10] Ringen, S. (2016). The Perfect Dictatorship: China in the 21st Century. Hong Kong University Press.

[11] Rebecca Taylor(2017), China installs 20 million cameras in 'the world's most advanced video surveillance system' to help fight crime,

http://www.mirror.co.uk/news/world-news/china-installs-20-million-cameras-11241727

[12] Zhang Jianli(2017), 20 million camera look at your Skynet project invasion of privacy?

http://www.top-news.top/news-13431554.html

[13] 中国网（2003），金盾工程,

http://www.china.com.cn/chinese/zhuanti/283732.htm

[14] Yuyu Chen，David Y. Yang* The Impact of Media Censorship:Evidence from a Field Experiment in China

December 3, 2017. Availuable at https://stanford.edu/~dyang1/pdfs/1984bravenewworld_draft.pdf

[15] Yuyu Chen，David Y. Yang* The Impact of Media Censorship:Evidence from a Field Experiment in China

December 3, 2017. Availuable at https://stanford.edu/~dyang1/pdfs/1984bravenewworld_draft.pdf

[16] Schmidt, Eric E.; Cohen, Jared (11 March 2014). "The Future of Internet Freedom". New York Times. Retrieved 11 March 2014.

[17] "Internet Censorship in China". The New York Times. December 28, 2012. Retrieved 9 March 2013.

[18] Human Rights Watch. "World Report 2012: China". Retrieved 9 March 2013.

[19] The Freedom House's Freedom of the Net Report in 2017 labels China's "Net Freedom Status" as not free, and rates its "Internet Freedom Score" as 87 (out of 100, where 100 indicates the most unfree) — the "world's worst abuser of Internet freedom." Source: https://freedomhouse.org/report/freedom-net/2017/china, last accessed on November 26, 2017.

[20] 新华网（2017），中共中央办公厅、国务院办公厅印发《推进互联网协议第六版（IPv6）规模部署行动计划》

http://politics.people.com.cn/n1/2017/1126/c1001-29668187.html

[21] Xinhuanet(2017),China to speed up IPv6-based Internet development,

http://news.xinhuanet.com/english/2017-11/26/c_136780735.htm

[22] 包雨朦(2017), 院士详解中国推进 IPv6 规模部署：要在顶级域名上获解释权

http://www.thepaper.cn/newsDetail_forward_1883858

[23] Wikipedia, Real name system,

https://en.wikipedia.org/wiki/Real-name_system

[24] Rachel Botsman(2017), Big data meets Big Brother as China moves to rate its citizens,

http://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion

[25] 国务院（2014），国务院关于印发社会信用体系建设规划纲要（2014—2020 年）的通知,

http://www.gov.cn/zhengce/content/2014-06/27/content_8913.htm

[26] KAI-FU LEE(2017), The Real Threat of Artificial Intelligence,

https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html?_r=0

[27] Hugh Howey(2017), HOW TO BUILD A SELF-CONSCIOUS MACHINE

https://www.wired.com/story/how-to-build-a-self-conscious-ai-machine/

AFTERWORD

The “Political Data Science: An Open Introduction” is a free dynamic book. And this is only the very beginning edition. I will update chapter 10 to chapter 12 in the next edition. Chapter 10 is about data Visualization in political science research. The chapter 11 is about machine learning in political science research. And most importantly, the chapter 12, I will rise a discussion about how to write a political research paper. What is a good political science paper? Why some political research papers actually only pretend like a science research? These interesting problems I will share in the books’ next edition.

INDEX